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		CCM2301722

母亲的身心健康对婴儿身心发展的影响

摘要

维护婴儿成长是社会发展的重要任务之一。本文针对婴儿成长问题，建立分类模型及优化模型，对母亲的身心健康与婴儿成长进行分析，通过 MATLAB 软件求解，预测最后 20 组婴儿的行为特征及睡眠质量，制定出合理有效的治疗方案使得编号 138 位婴儿的行为特征从中等型转变为安静型，并在此基础上调整策略使该婴儿睡眠质量变为优。

首先对数据进行预处理。首先对问卷进行异常值处理，剔除婚姻状况为 3、6 的数据。又编号 180 位受访者的为 99:99，故也删去。因此总共剔除数据 10 个。接着，对婴儿行为特征、整晚睡眠时间进行量化处理。

针对问题一，根据题目要选取母亲的身体指标、心理指标、婴儿的行为特征和睡眠质量作为指标。考虑到母亲的身体指标有 4 个变量，心理指标有 3 个变量，且变量之间具有相关性，则先利用主成分分析法进行降维，再分析变量间是否存在相互关系，从而达到简化题目的效果。接着使用 Spearman 相关系数分析，得到：母亲的身体指标与婴儿行为特征、睡眠质量之间不存在显著影响，无线性相关性；母亲的心理指标与婴儿行为特征、整完睡眠时间、睡醒次数之间存在显著影响，存在线性关系。

针对问题二，首先针对婴儿行为特征的分布状况以进行描述，结果表明不需对样本进行采样。接着对模型进行似然检比卡方检验，P 值小于 0.05，可使用有序逻辑回归模型。然后计算回归系数，再根据因变量阈值表，得到最终的有序逻辑回归模型。结合混淆矩阵与准确率，对模型进行评估，结果表明模型质量较好。并通过建立的有序逻辑回归模型得到最后 20 组的分类结果。

针对问题三,首先根据给出的患病得分与治疗费用之间的关系，确定总治疗费用心理指标的关系，并引入上四分位数与下四分位数作为每种婴儿行为特征的边界范围，建立婴儿行为特征约束的单目标最值优化模型，最后通过智能算法计算出婴儿行为特征从矛盾型变为中等型所耗费的治疗费用最少为 28159 元，从矛盾型变为安静型所耗费的治疗费用最少为 33564.67 元。

针对问题四，首先查找参考文献，确定婴儿睡眠质量优良中差比例分别为 16.74%、16.74%、16.74、49.77%。采用灰色关联度模型对 380 个观测值综合排序。按照确定的比例划分即可得到婴儿睡眠质量的综合评判结果。再对睡眠质量的评判结果量化，将优、良、中、差分别量化为 3、2、1、0。接着建立随机森立分类模型，预测最后 20 组婴儿的综合睡眠质量的类别，并对随机森林分类模型进行评估。

针对问题五，首先确定婴儿优质睡眠质量边界条件，然后在问题三模型的基础上，将确定婴儿优质睡眠质量中 3 种心理指标的范围作为约束条件添加入模型中，建立最少治疗费用的单目标多约束优化模型，最后采用智能算法求解出问题三基础下婴儿优质睡眠质量时最少的治疗费用为 35134 元，而变为安静型是最少的治疗费用为 36004.67 元。

本文所采用的模型可以很好地根据实际情况解决分类问题和优化问题，具有较好地适用性和准确性。并且，本文对模型结果进行分析，对模型性能进行评估，同时提出了改进方向，完善模型，以扩大使用范围。

关键词：婴儿成长　有序逻辑回归分类模型　规划模型　随机森林分类模型　身心指标

一、问题重述

1.1 问题背景

婴儿是国家人口增长的重要组成部分，维护婴儿成长是中国社会发展的重要任务之一。而婴儿人生中最重要的人之一便是母亲，母亲既作为营养提供者和身体保护者，又作为情感支持者和安全提供者。故婴儿的生理和心理健康可能受到母亲身心健康状态的影响，导致婴儿睡眠质量的降低。

母亲的身体状态包含了年龄、婚姻状况、教育程度、妊娠时间、分娩方式，而心理状态则包括了分娩相关创伤后应激障碍情况（CBTS）、产后抑郁情况（EPDS）、住院期间的焦虑和抑郁水平情况（HADS），探究母亲身心健康情况与婴儿睡眠质量的关系成为观察影响婴儿成长的方法。

1.2 要解决的问题

要求根据已知的 390 名 3 至 12 个月婴儿以及其母亲相关数据，通过相关文献了解专业背景，进而回答下列问题：

对于问题一，请根据附件中的数据研究母亲的身体指标对婴儿的行为特征和睡眠质量是否有影响，以及母亲的心理指标对婴儿的行为特征和睡眠质量是否有影响。

对于问题二，安静型、中等型、矛盾型是婴儿的行为特征的三种类型，请根据已知的婴儿行为特征建立以婴儿的行为特征为因变量，母亲的身体指标与心理指标为自变量的关系模型，并根据建立的关系模型预测编号 391-410 号这 20 组的婴儿行为特征属于什么类型。

对于问题三，对母亲心理患病程度的治疗可以促进婴儿的认知、情感和社交发展，改善婴儿的行为特征，已知母亲三种心理病状的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，请根据表 1 中患病状况与治疗费用之间的关系，建立在 238 编号的婴儿行为特征从矛盾型变为中等型的条件下，使得花费治疗费用最少的模型，求出最少的治疗费用。而当 238 编号的婴儿行为特征变为安静型时，同样建立治疗费用最少模型，得到最少的治疗费用为多少。

对于问题四，已知整晚睡眠时间、睡醒次数、入睡方式是婴儿睡眠质量的指标，请将已知婴儿睡眠质量指标的数据进行评价得分，并将其分为优、良、中、差四个等级。再根据分类得到的婴儿综合睡眠质量，建立其与母亲的身体指标、心理指标之间的关联模型，并根据建立的模型预测编号 391-410 号 20 组婴儿的综合睡眠质量。

对于问题五，在问题三建立的治疗费用最少模型基础上，添加 238 号婴儿的综合睡眠质量为优的约束条件，求解出最少的治疗费用。

二、问题分析

2.1 问题一的分析

问题一要求分析母亲的身体指标和心理指标是否对婴儿的行为特征和睡眠质量存在影响。考虑到母亲的身体指标有 4 个变量，心理指标、睡眠质量均有 3 个变量，则先利用主成分分析法进行降维，再分析变量间是否存在相互关系，从而达到简化题目的效果。接着对数据进行正态性检验，又样本值为 380，远小于 5000，为小样本数据，因此选择 S-W 检验。若结果表明呈现正态分布，则使用 Pearson 相关系数；反之，则使用 Spearman 相关系数。最终通过相关性分析判断母亲的身体指标和心理指标是否对婴儿的行为特征和睡眠质量存在影响。

2.2 问题二的分析

问题二要求建立以婴儿的行为特征为因变量，母亲的身体指标与心理指标为自变量的关系模型，并根据建立的模型预测 20 组(编号 391-410 号)的婴儿行为特征类型。显然建立分类模型去预测婴儿特征类型。因此本文考虑使用有序逻辑回归模型进行预测。首先针对婴儿行为特征的分布状况以进行描述，判断是否需要对样本进行采样。接着对模型进行似然检比卡方检验，分析似然检比卡方显著性，若 $P<0.05$，说明模型有效，反之模型不成立。接着计算回归系数，再根据因变量阈值表，得到最终的有序逻辑回归模型。结合混淆矩阵与准确率，对模型进行评估。并通过建立的有序逻辑回归模型得到最后 20 组的分类结果。

2.3 问题三的分析

问题三要求根据表 1 中患病状况与治疗费用之间的关系，建立在 238 编号的婴儿行为特征从矛盾型变为中等型和安静型的条件下，使得花费治疗费用最少的模型，求出最少的治疗费用。首先患病程度的变化率与治疗费用的正比关系建立每个心理指标与治疗费用之间的关系，然后求和得到总治疗费用与 3 个心理指标间的关系，其次引入上四分位数与下四分位数作为每一种婴儿行为特征的边界范围，求解出每一种婴儿行为特征所对应心理指标的取值范围，接着以 *CBTS* 、 *EPDS* 、 *HADS* 的患病程度变化率为决策变量，每种婴儿行为特征的边界以及问题二中求得处于中等型和安静型行为特征的方程作为约束条件，总治疗费用为目标函数，建立婴儿行为特征约束的单目标最值优化模型，并运用智能算法进行求解，得到婴儿行为特征从矛盾型变为中等型和安静型的最少治疗费用。

2.4 问题四的分析

要求根据整晚睡眠时间、睡醒次数、入睡方式三种指标对婴儿的综合睡眠质量进行评判分类，以分类的结果数据建立与母亲的身体指标、心理指标的关联模型。由于睡眠质量的三个指标可以通过查阅文献以及直观的得出其为正向或负向指标，同时也能得到影响睡眠质量指标的三个最优值，在本题数据量较小，数据的质量程度不高的情况下，可以采用简单的评价模型对每一组编号进行评价，从高到低得到评价分数，并根据评价分数按照一定占比简单分类，即可得到婴儿睡眠质量的综合评判结果。接着对睡眠质量的评判结果量化，将优、良、中、差分别量化为 3、2、1、0。然后建立多种机器学习模型伯明并对其进行评估，最终选择建立最优的一个机器学习分类模型，预测最后 20 组婴儿的综合睡眠质量的类别。

2.5 问题五的分析

问题五要求在问题三建立的治疗费用最少模型基础上，添加 238 号婴儿的综合睡眠质量为优的约束条件，求解出最少的治疗费用。首先同样引入上四分位数与下四分位数作为婴儿优质睡眠质量的边界范围，其次在问题三模型的基础上，将引入的优质睡眠质量的边界范围作为约束条件加入到模型中，最后通过智能算法求解得到婴儿行为特征从矛盾型变为中等型和安静型，同时睡眠质量变为优时的最少治疗费用。

三、模型的假设

（1）假设每一类的行为特征中心理指标的边界可用上四分位数与下四分位数代替；

（2）假设母亲一旦进行心理治疗，无论有没有降低患病得分都必须缴纳三种病症的门槛费用 1000 元；

（3）假设**

四、符号说明

符号	符号说明	单位	符号	符号说明	单位
X_1	CBTS 得分数值	分	X_9	婴儿行为特征	/
X_2	EPDS 得分数值	分	X_{10}	婴儿性别	/
X_3	HADS 得分数值	分	X_{11}	婴儿年龄	月
X_4	母亲年龄	岁	X_{12}	整晚睡眠时间	点
X_5	婚姻状况	/	X_{13}	睡醒次数	次
X_6	教育程度	/	X_{14}	入睡方式	/
X_7	妊娠时间	周数	F_1	心理指标	/
X_8	分娩方式	/	F_2	身体指标	/

五、模型的建立与求解

本文先对数据进行预处理，再进行模型的建立与求解。

首先对问卷进行异常值处理。编号 43、编号 95、编号 134、编号 196、编号 301、编号 308、编号 355 位受访者的婚姻状况为 3；编号 231、编号 306 位受访者的婚姻状况为 6。而婚姻状况仅为两类(1 和 2)，即已婚和未婚，故删去编号 43、编号 95、编号 134 等 9 个观测值。又编号 180 位受访者的为 99:99，显然记录或手机有误，为无效问卷，故删去编号 180。因此总共剔除数据 10 个，有效问卷共 380 份，属于正常范畴。

接着，本文对变量进行量化处理。婴儿行为特征指标进行量化，分别将安静型、中等型、矛盾型赋值为 0、1 和 2。其次，将整晚睡眠时间（时：分：秒）进行量化，即其将转换为定量变量。如将 5:00:00 转换为数值 5，将 5:30:00 转换成数值 5.5。

5.1 问题一模型的建立与求解

问题一要求分析母亲的身体指标和心理指标是否对婴儿的行为特征和睡眠质量存在影响。母亲的身体指标包括年龄、婚姻状况、教育程度、妊娠时间、分娩方式；产妇心理指标通过三种不同的问卷得分表示，问卷分别为 CBTS、EPDS、HADS；婴儿睡眠质量指标包括整晚睡眠时间、睡醒次数、入睡方式。考虑到母亲的身体指标有 4 个变量，心理指标、睡眠质量均有 3 个变量，则先利用主成分分析法进行降维，再分析变量间是否存在相互关系，从而达到简化题目的效果。

相关分析是对变量两两之间的相关程度进行分析。Pearson 相关系数适用于定量数据，且数据需服从正态分布；Spearman 相关系数适用于定量变量有序变量，且当数据不满足正态分布时使用。因此本题先对数据进行正态性检验，又样本值为 380，远小于 5000，

为小样本数据，因此选择 S-W 检验。若结果表明呈现正态分布，则使用 Pearson 相关系数；反之，则使用 Spearman 相关系数。问题一思路图如下所示：

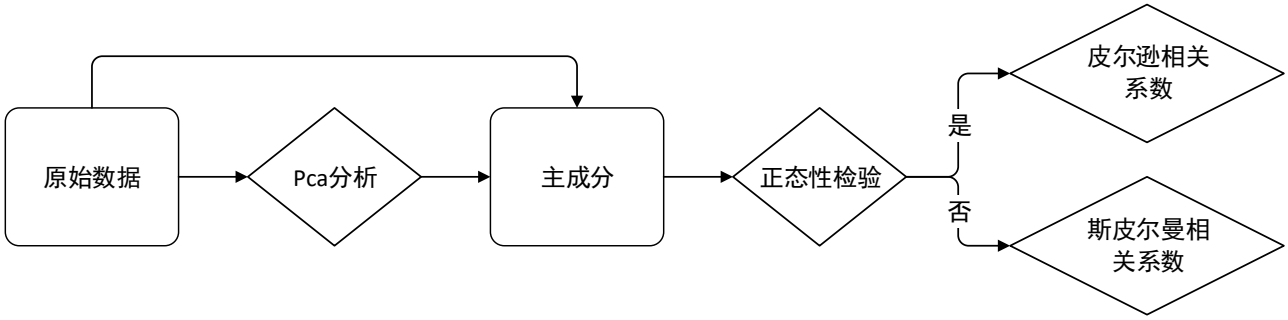


图 1 问题一求解思路图

5.1.1 基于主成分分析法降维

为了简化题目，本文先将母亲的身体指标即变量 CBTS、EPDS、HADS 进行降维处理，参考何晓群[1]的多元统计分析，使用 spsspro 软件进行主成分分析。首先需通过 KMO 和 Bartlett 的检验，判断本题数据是否适合进行主成分分析。若检验通过，分析方差解释表格及碎石图，确定主成分的数量。接着通过分析主成分载荷系数与热力图，可以分析到每个主成分中隐变量的重要性。基于主成分载荷图通过将多主成分降维成双主成分或者三主成分，通过象限图的方式呈现主成分的空间分布。最后通过分析成分矩阵，得出主成分成分公式与权重。

(1) 主成分分析法模型理论

首先对数据 CBTS、EPDS、HADS 进行标准化。标准化后的数据为一个三维总体 $X=\left(X_1,X_2,X_3\right)$。计算对标准化后数据的协方差矩阵 R：

$$R=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \tag{1}$$

计算协方差举证的特征值 $\lambda_1\geq\lambda_2\geq\lambda_3\geq 0$ ，以及对应的特征向量 a_1,a_2,a_3 ，且

$$a_1=\left(a_{1i},a_{2i},\cdots,a_{nj}\right) \tag{2}$$

其中 a_{ni} 表示第 i 个特征向量的第 n 个分量。由特征向量组成 n 个新的指标变量:

$$\begin{cases} Y_1=a_{11}X_1+a_{21}X_1+\cdots+a_{n1}X_1 \\ Y_2=a_{12}X_2+a_{22}X_2+\cdots+a_{n2}X_2 \\ Y_3=a_{13}X_3+a_{23}X_3+\cdots+a_{n3}X_3 \end{cases} \tag{3}$$

式中，Y_1 是第一主成分，Y_2 是第二主成分，Y_3 是第三主成分。

再分别计算主成分的贡献率 b_i ：

$$b_i=\frac{\lambda_i}{\sum_{k=1}^{n}\lambda_k} \qquad i=1,2,3 \tag{4}$$

以及主成分的累计贡献率 α_p ：

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \qquad p \le n \tag{5}$$

(2) 主成分分析具体步骤

Step1 进行 KMO 和 Bartlett 检验

为了确定是否适合进行主成分分析法，需先进行 KMO 与 Bartlett 球形检验。

表 1 KMO 与 Bartlett 球形检验表

KMO 值		0.73
Bartlett 球形度检验	近似卡方	742.69
	df	3
	P	0.000

由上表可知，KMO 的值为 0.73，大于 0.6，同时 Bartlett 球形检验的结果显示显著性 P 值为 0.000，小于 0.05，水平上呈现显著性，各变量间具有相关性，主成分分析有效，程度为一般。因此可进行主成分分析。

Step 2 确定主成分数量

本文通过方差解释率与碎石图结合判断主成分个数。

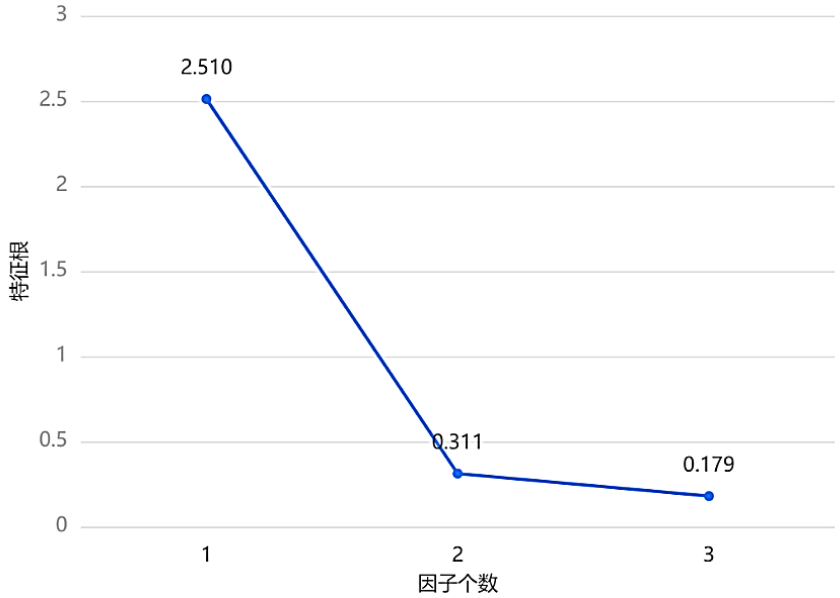


图 2 碎石图

由碎石图可知成分 1、成分 2、成分 3 的特征根分别为 2.510、0.311、0.179，仅成分 1 的特征根大于 1。

表 2 方差解释表

成分	方差解释率(%)	累积方差解释率(%)
1	83.679	83.679
2	10.356	94.034
3	5.966	100

由上表可知，成分 1 累计方差解释率已超过 80%。每个变量都能很好地被解释。结合表 2 以及碎石图坡度趋于平缓的趋势判断主成分的数量为 1。

Step 3 得到最终模型

表 3 因子载荷系数及得分系数表

变量	主成分 1 因子载荷系数	公因子方差	因子得分系数
CBTS	0.900	0.810	0.3584
EPDS	0.94	0.884	0.3745
HADS	0.904	0.817	0.3600

上表可分析每个主成分中隐变量的重要性。通过因子得分系数得出因子公式，得到最终模型的公式：

$$F_1 = 0.3584X_1 + 0.3745X_2 + 0.3600X_3 \tag{6}$$

同理，对母亲的 5 个身体指标年龄、婚姻状况、教育程度、妊娠时间、分娩方式进行 PCA 降维，具体过程详见附录一，得到一个主成分，最终模型公式为：

$$F_2 = 0.4740X_4 + 0.3069X_5 + 0.4199X_6 - 0.3830X_7 + 0.3733X_8 \tag{7}$$

通过主成分分析对指标进行降维，最终得到两组数据，分别为心理指标 F_1 和身体指标 F_2。两次 PCA 降维累积方差解释率均超过 80%， F_1 和 F_2 均可很好地解释。

5.1.2 Spearman相关系数分析

通过主成分分析，已经对数据进行降维，下文将进行相关分析。本题先对数据进行正态性检验，若呈现正态分布，则使用 Pearson 相关系数；反之，则使用 Spearman 相关系数。又由于样本值为 380，远小于 5000，为小样本数据，因此选择 S-W 检验。结果显示 P 值均为 0，小于 0.05，具体结果可见附录二。因此不服从正态分布，所以进行 Spearman 相关系数分析。

故对变量：心理指标、身体指标、行为特征、睡眠质量(整晚睡眠时间、睡醒次数、入睡方式)进行斯皮尔曼相关性分析。

(1) 显著性检验

首先对各个变量进行显著性检验，结果如下表所示：

表 4 显著性检验结果表

	身体指标	心理指标	婴儿行为特征	整晚睡眠时间	睡醒次数	入睡方式
身体指标	0.000	0.198	0.114	0.829	0.392	0.089
心理指标	0.198	0.000	0.006	0.002	0.044	0.490
婴儿行为特征	0.114	0.006	0.000	0.021	0.000	0.971
整晚睡眠时间	0.829	0.002	0.021	0.000	0.000	0.000
睡醒次数	0.392	0.044	0.000	0.000	0.000	0.000
入睡方式	0.089	0.490	0.971	0.000	0.000	0.000

有上表可知母亲的身体指标与婴儿行为特征、睡眠质量之间不存在显著影响，无线性相关性；母亲的心理指标与婴儿行为特征、整完睡眠时间、睡醒次数之间存在显著影响，存在线性关系。

(2) 计算相关系数

变量 X 与 Y 斯皮尔曼相关系数计算公式[2]如下：

$$p_{XY} = \frac{\sum_i \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_i \left(X_i - \bar{X}\right)^2 \sum_i \left(Y_i - \bar{Y}\right)^2}} \tag{8}$$

利用上式计算心理指标 F_1、身体指标 F_2 与行为特征 X_9、睡眠质量(整晚睡眠时间 X_{13}、睡醒次数 X_{14}、入睡方式 X_{15})之间关系。

得到相关系数矩阵热力图如下所示：

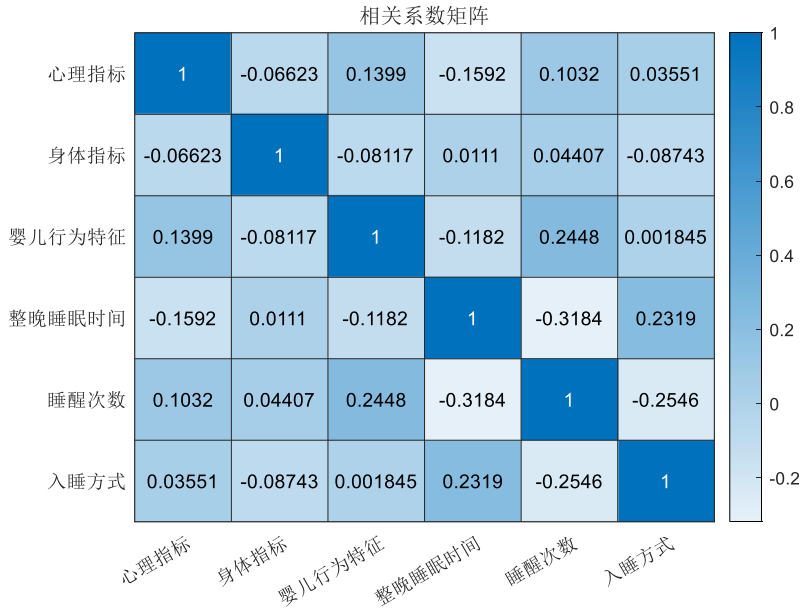


图 3 相关系数矩阵热力图

斯皮尔曼相关关系数值均在-1 到 1 之间，1 表示完全正相关，而-1 则是完全负相关。由上图可知母亲的身体指标与婴儿行为特征呈现负相关，而与睡眠质量均呈现正相关。母亲的心理指标与婴儿行为特征呈现正相关，与整晚睡眠时间呈现负相关，与睡醒次数呈现正相关。

5.1.3 结果分析

通过斯皮尔曼相关系数分析得到：母亲的身体指标与婴儿行为特征、睡眠质量之间不存在显著影响，无线性相关性；母亲的心理指标与婴儿行为特征、整完睡眠时间、睡醒次数之间存在显著影响，存在线性关系。并且母亲的身体指标与婴儿行为特征呈现负相关，而与睡眠质量均呈现正相关。母亲的心理指标与婴儿行为特征呈现正相关，与整晚睡眠时间呈现负相关，与睡醒次数呈现正相关。

5.2 问题二模型的建立与求解

首先针对婴儿行为特征的分布状况以进行描述，判断是否需要对样本进行采样。接着对模型进行似然检比卡方检验，分析似然检比卡方显著性，若 P<0.05，说明模型有效，反之模型不成立。再根据模型参数表，分析 X 是否呈显著性(P<0.05)，用于探究是否影响关系。结合预测分类混淆矩阵与模型评价中的分类指标，分析模型预测。并得到最后 20 组的分类结果。

5.2.1 建立有序逻辑回归模型

Step 1 因变量分布状况的简单描述统计

当婴儿行为特征分类水平的数据量出现严重不平衡时，则需对数据进行过采样或者欠采样。下表展示了因变量各分组的分布情况。类别总计 0、1、2 三种，分别代表安静型、中等型、矛盾型。安静型频数为 116，百分比为 30.526%。中等型频数为 220，占比为 57.895%。矛盾型频数为 44、占比为 11.579%。分布相对均衡，且数据集样本量较少，若再采样，样本量更少，影响模型质量。

表 5 有序分类因变量分布表

因变量	类别	频数	百分比(%)
婴儿行为特征	1	220	57.895
	0	116	30.526
	2	44	11.579
	总计	380	100.0

Step 2 似然比卡方检验

接着对模型进行似然检比卡方检验，分析似然检比卡方显著性，判断模型是否有效。

由于问题一的结果表明母亲的身体指标与婴儿行为特征之间不存在相关性。因此，建立母亲的心理指标与婴儿行为特征之间关系的模型。对 P 值进行分析，模型的模型的似然比卡方检验的结果显示，显著性 P 值 0.153，水平上不呈现显著性。检验不通过，所以考虑添加母亲的身体指标五个因素，再进行似然比卡方检验，并观察 P 值。

Step 3 得到有序逻辑回归模型[3]

通过 spsspro 得到有序逻辑回归结果如下表所示：

表 6 有序逻辑回归结果表

变量	回归系数	标准误差	OR95%置信区间	
			上限	下限
母亲年龄	-0.051	0.024	0.907	0.996
婚姻状况	0.038	0.531	0.367	2.938
教育程度	0.16	0.106	0.953	1.446
妊娠时间（周数）	-0.036	0.057	0.863	1.079
分娩方式	0.076	0.867	0.197	5.904
CBTS	-0.006	0.033	0.931	1.06
EPDS	0.027	0.029	0.97	1.088
HADS	0.021	0.04	0.944	1.104

上表展示了模型的结果，包括模型的系数、标准误差、OR 值、置信区间等。

表 7 因变量分类阈值表

婴儿行为特征	0	1	2
预测值	$\hat{X}_9 \le -2.598$	$-2.598 < \hat{X}_9 \le 0.329$	$0.329 < \hat{X}_9$

结合表 6、表 7，得到模型公式如下所示：

$$X_9 = \begin{cases} 0, & \hat{X}_9 \le -2.598 \\ 1, & -2.598 < \hat{X}_9 \le 0.329 \\ 2, & 0.329 < \hat{X}_9 \end{cases} \tag{9}$$

式中，

$$\begin{aligned} \hat{X}_9 = &-0.006X_1 + 0.027X_2 - 0.021X_3 - 0.051X_4 + 0.038X_5 \\ &+ 0.16X_6 - 0.036X_7 + 0.076X_8 \end{aligned} \tag{10}$$

5.2.2 有序逻辑回归模型的评估

使用准确率对模型进行初步评估。*TP* 表示正样本判断为正样本的样本数；*TN* 表示负样本判断为负样本的样本数；*FP* 表示负样本判断为正样本的样本数；*FN* 表示正

样本判断为负样本的样本数。准确率为

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

通过计算得准确率达到 0.6。

混淆矩阵左下斜对角线中的数值均是分类正确的结果。同样，明显观察出分类准确率较高，模型较为良好。由图可知，由 40 个安静型分类正确，28 个中等型分类正确，64 个矛盾型分类正确。

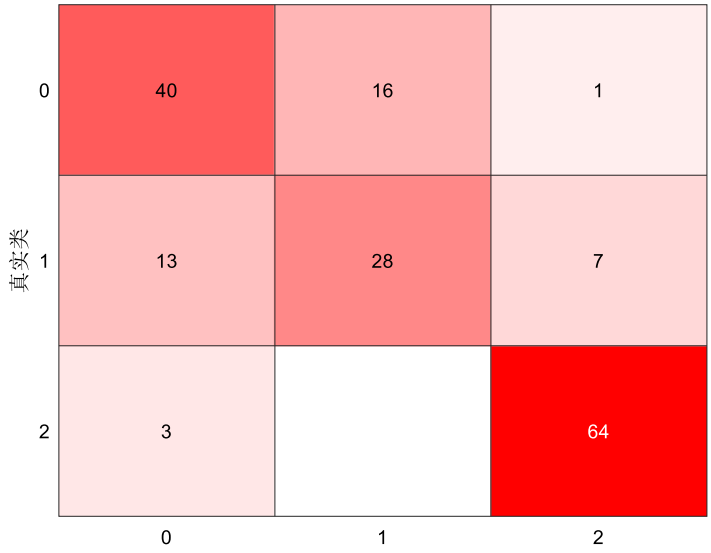


图 4 有序逻辑回归分类混淆矩阵热力图

5.2.3 结果分析

最终通过建立的有序逻辑回归分类模型模型预测 20 位婴儿的行为特征如下所示：

表 8 问题一分类结果表

编号	分类结果	编号	分类结果	编号	分类结果	编号	分类结果
1	1	6	1	11	1	16	0
2	0	7	1	12	1	17	1
3	1	8	1	13	1	18	1
4	1	9	1	14	2	19	0
5	1	10	0	15	1	20	1

表中 0 代表婴儿的行为特征为安静型、1 代表婴儿的行为特征为中等型、2 代表婴儿的行为特征为矛盾型。

5.3 问题三模型的建立与求解

题目要求在已知患病得分与治疗费用的关系基础上，求解使得 238 号婴儿行为特征从矛盾性变为中等型和安静型，治疗费用最少的情况。首先可以去表 1 的已知条件拟合出患病程度的变化率与治疗费用的关系，由拟合的方程求和得出总治疗费用，然后由每种婴儿行为特征中心理病状得分的上四分位数与下四分位数作为每种婴儿行为特征的边界，以总的治疗费用作为目标函数，*CBTS*、*EPDS*、*HADS* 三种心理病状的变化率作为决策变量，每种婴儿行为特征的边界以及问题二中求得处于中等型和安静型行为特征的方程作为约束条件，通过智能算法进行求解得到婴儿行为特征变为相应的情况下，所需的最少治疗费用，具体流程图如下所示：

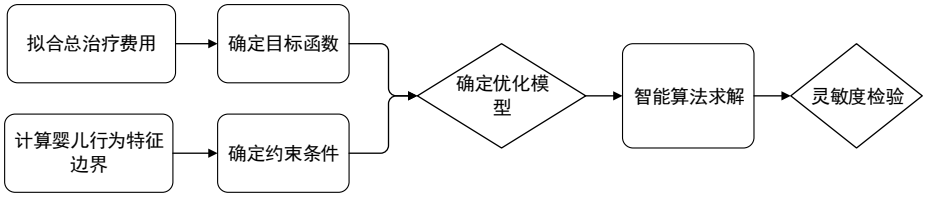


图 5 问题三思路流程图

5.3.1 患病程度变化率与治疗费用函数关系的确定

已知 *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，故构建以每种患病程度的治疗费用为因变量，每种患病程度的变化率为自变量的一次函数，通过简单的计算可以得到每个一次函数的相关系数如下表所示：

表 9 每种患病程度变化率与治疗费用关系系数表

	CBTS	EPDS	HADS
k	2612/3	695	2440
b	200	500	300

由求得的相关系数带入方程可得治疗费用与患病程度的变化率的方程如下：

$$\begin{cases} y_1 = \dfrac{2612}{3} x_1 + 200 \\ y_2 = 695\, x_2 + 500 \\ y_3 = 2440 x_3 + 300 \end{cases} \tag{12}$$

其中 y_1 表示用于治疗 *CBTS* 的治疗费用， x_1 表示 *CBTS* 的患病程度变化率，同理可得， y_2 和 y_3 分别表示用于治疗 *EPDS* 和 *HADS* 的治疗费用， x_2 和 x_3 分别表示 *EPDS* 和 *HADS* 两种患病程度的变化率。

将拟合得到的 3 个一次函数绘图观察三者之间的变化关系，作图如下：

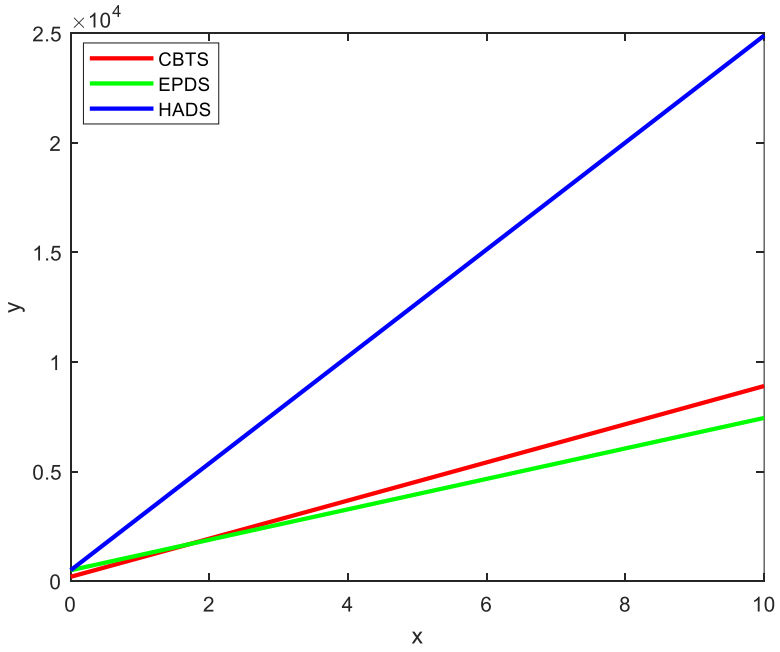


图 6 三种病状治疗费用与患病程度变化率函数比较图

从图中可以看出在相同的患病程度变化率情况下，*HADS* 的治疗费用明显高于其他两种病状，而 *CBTS* 和 *EPDS* 之间存在着交叉点，在波动大概小于 2 时，*EPDS* 的治疗费用大于 *CBTS*，而在波动大于 2 后，*CBTS* 的治疗费用则大于 *EPDS*。由此可以得出在之后求解最少治疗费用时，应该使得 *HADS* 的患病程度变化率为最小时，才有可能达到最小的治疗费用，为后续求解结果时提供一个参考依据。

由于假设一旦进行心理治疗，无论有没有降低患病得分都必须缴纳三种病症的门槛费用 1000 元，故总的治疗费用为三种患病的治疗费用的求和，同时也为建立优化模型的目标函数，即：

$$y = \frac{2612}{3}x_1 + 695x_2 + 2440x_3 + 1000 \tag{13}$$

其中，y 表示用于治疗三种心理病状的总费用，x_1、x_2 和 x_3 分别表示三种患病程度的变化率。

5.3.2 各类婴儿行为特征边界条件的确定

由于问题要求求解使得婴儿的行为特征从矛盾型变为中等型和安静型所花费的最少治疗费用，所以需要得出每一种婴儿行为特征所对应的取值范围，因为在问题二中仅通过有序逻辑回归拟合出了婴儿的行为特征与母亲的身体指标与心理指标之间的关系，并没有得出每一种心理指标的取值范围，同时每一组编号的心理指标的最大值与最小值并不能完全代表该组属于该行为特征，故为了使建立的模型更加准确，引入上四分位数与下四分位数作为每一种婴儿行为特征的边界范围

对于附件中的数据在 Excel 中按照婴儿的行为特征筛选出每一种行为特征所包含的编号及其三种心理指标，筛选后的结果详见“四分位数”Excel 表格，通过简单的 matlab 运算得到每一种婴儿行为特征的三种心理指标所对应的上四分位数与下四分位数，如下表所示：

表 10 每种婴儿行为特征对应的 3 种心理指标边界表

	安静型 CBTS	安静型 EPDS	安静型 HADS	安静型 CBTS	中等型 EPDS	中等型 HADS	矛盾型 CBTS	矛盾型 EPDS	矛盾型 HADS
上四分位数	8	12.5	10	9	13	11	10	14	11
下四分位数	1	3	4	2	4	5	3	6	5.5

从表中可以看出在三种婴儿行为特征中安静型行为的的上下边界最小而矛盾型行为的上下边界最大。安静型的 *CBTS* 取值范围为 1 到 8，*EPDS* 取值范围为 3 到 12.5，*HADS* 取值范围为 4 到 10；而中等型的 *CBTS* 取值范围为 2 到 9，*EPDS* 取值范围为 4 到 13，*HADS* 取值范围为 5 到 11；至于矛盾型的 *CBTS* 取值范围为 3 到 10，*EPDS* 取值范围为 6 到 14，*HADS* 取值范围为 5.5 到 11。通过求解的每种婴儿行为特征的上四分位数与下四分位数可以作为后续优化模型的约束条件。

5.3.3 婴儿行为特征约束的单目标最值优化模型的建立

（1）决策变量的确定

问题中患病的治疗费用仅与 *CBTS*、*EPDS*、*HADS* 这三种心理指标患病程度变化率的呈现函数关系，而与母亲的身体指标与婴儿的睡眠质量指标无关，故建立的优化模型的决策变量为 *CBTS*、*EPDS*、*HADS* 的患病程度变化率，即 x_1、x_2 和 x_3。

（2）目标函数的确定

问题要求在婴儿的行为特征从矛盾型变为中等型的条件下，求得最少的治疗费用，故构建的目标函数应为患病程度变化变化情况下的总治疗费用，即公式(13)所示为：

$y=\frac{2612}{3}x_1+695x_2+2440x_3+1000$，即只要进行治疗，则就需要 1000 元的治疗费用。

（3）约束条件的确定

问题要求的条件为婴儿的行为特征从矛盾型变为中等型，根据问题二利用有序逻辑回归建立的拟合方程得到在行为特征属于中等型时，y 的取值范围为-2.598 到 0.329，故可将拟合方程中的 3 个心理指标得分值自变量转化为以 3 个心理指标得分变化率的自变量，然后带入属于中等型行为特征中的取值范围中，并将 238 号的身体指标数据带入除心理指标得分变化率的其他自变量中，最终得到综合心理指标得分变化率的约束条件为：

$$-2.598\le 0.006x_1-0.027x_2-0.021x_3-0.0824\le 0.329 \tag{14}$$

其中 x_1、x_2 和 x_3 分别表示三种患病程度的变化率。

同时根据表 10 所的 3 种心理指标边界条件可以得到处于中等型行为特征时，*CBTS*、*EPDS*、*HADS* 这三种心理指标患病程度变化率约束条件如下：

其中 x_1，即 *CBTS* 患病程度变化率应该属于 6 到 13 区间

$$2\le 15-x_1\le 9 \tag{15}$$

其中 x_2，即 *EPDS* 患病程度变化率应该属于 9 到 18 区间

$$4\le 22-x_2\le 13 \tag{16}$$

其中 x_3，即 *HADS* 患病程度变化率应该属于 7 到 13 区间

$$5\le 18-x_3\le 11 \tag{17}$$

（4）优化模型的确定

综上，通过确定的决策变量、建立的目标函数和计算得到的约束条件，可以建立当婴儿行为特征由矛盾型转变为中等型时的单目标最值优化模型如下所示：

$$\min y=\frac{2612}{3}x_1+695x_2+2440x_3+1000$$
$$\text{s.t.}\begin{cases}-2.598\le 0.006x_1-0.027x_2-0.021x_3-0.0824\le 0.329\\ 2\le 15-x_1\le 9\\ 4\le 22-x_2\le 13\\ 5\le 18-x_3\le 11\end{cases} \tag{18}$$

而当婴儿行为特征由矛盾型转变为安静型时，根据上式建立的模型同理可得，将约束条件中的综合心理指标得分变化率 y 的取值范围由-2.598 到 0.329 转变为大于等于 0.329；*CBTS* 患病程度变化率的区间转变为 7 到 14；*EPDS* 患病程度变化率的区间转变为 9.5 到 19；*HADS* 患病程度变化率的区间转变为 8 到 14，目标函数不变可得婴儿行为特征由矛盾型转变为安静型时的单目标最值优化模型如下所示：

$$\min y=\frac{2612}{3}x_1+695x_2+2440x_3+1000$$
$$\text{s.t.}\begin{cases}0.329\le 0.006x_1-0.027x_2-0.021x_3-0.0824\\ 1\le 15-x_1\le 8\\ 3\le 22-x_2\le 12.5\\ 4\le 18-x_3\le 10\end{cases}\tag{19}$$

5.3.4 单目标最值优化模型的智能算法求解及结果分析

对于建立的婴儿行为特征约束的单目标最值优化模型，采取智能算法-模拟退火算法进行求解，在 MATLAB 中通过 simulannealbnd 函数进行，首先随机生成一个初始解，通过函数的自动温度初始化过程来确定初始温度，并选取迭代次数为 200 次，约束条件作为定义变量的上下限，进行迭代求解，输出最优极值以及所对应的最优位置，得到全局最优解。

利用模拟退火算法的迭代求解由矛盾型转变为中等型时的优化模型过程如下图所示：

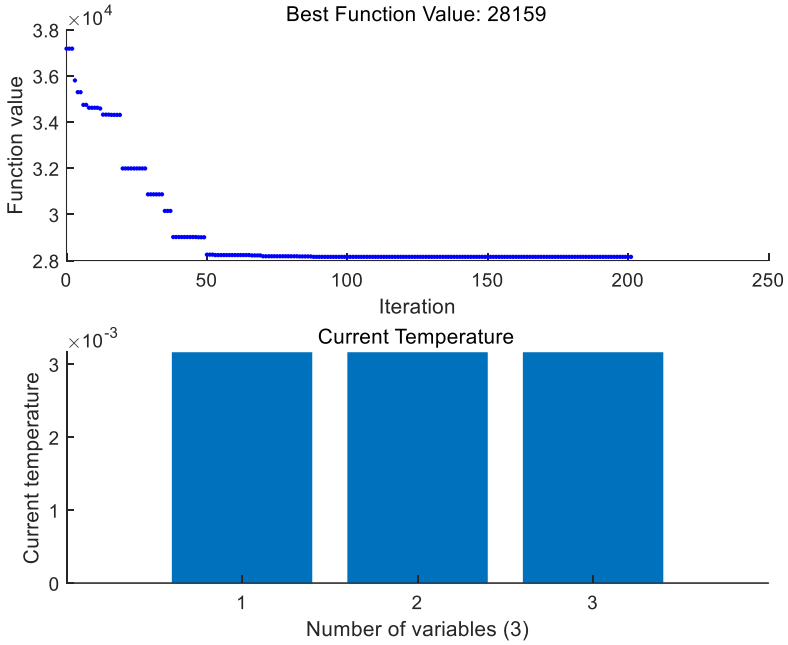


图 7 婴儿行为特征变为中等型优化模型迭代求解过程图

从图中可以看出在经过 50 次左右的迭代后，基本上就已经找到了最优结果，即最少的治疗费用为 28159 元，此时的退火温度已经降到温度接近于 0 左右。

表 11 婴儿行为特征变为中等型治疗费用最少结果数据表

	y	x_1	x_2	x_3
最优值	28159.00	6.00	9.00	7.00

从表中可以看出在 x_1 即 *CBTS* 患病程度得分下降 6 分； x_2 即 *EPDS* 患病程度得分下降 9 分；x_3 即 *HADS* 患病程度得分下降 7 分时，婴儿行为特征从矛盾型变为中等型所耗费的治疗费用最少为 28159 元，这一结果与开始时治疗费用与患病程度的变化率拟合函数关系的图像所做的分析相差并不太大，说明计算结果具有一定的有效性。

对于当婴儿行为特征从矛盾型变为安静型时最少的治疗费用，将约束条件根据建立的优化模型进行更行继续带入算法中求解即可，得到的迭代求解过程如下图所示：

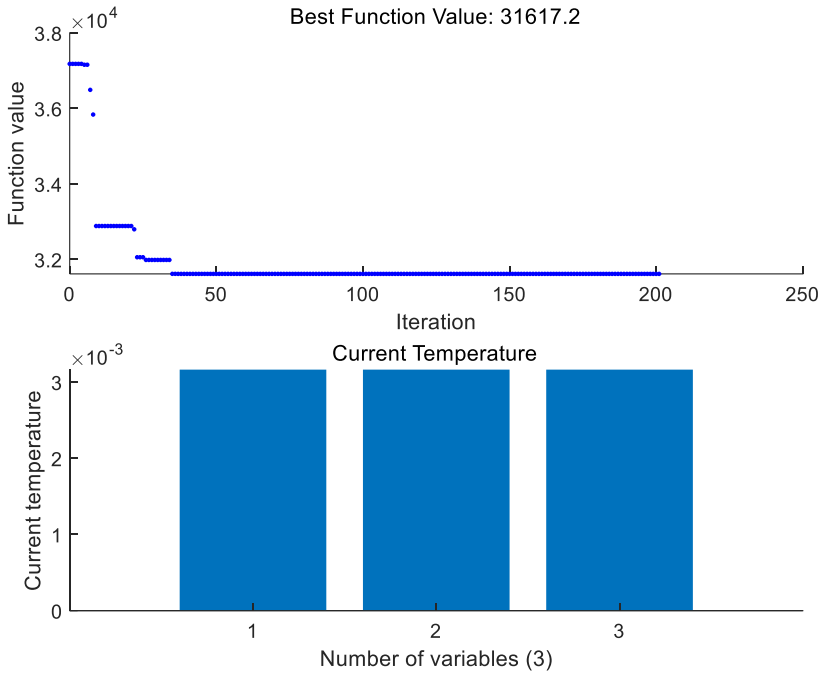


图 8 婴儿行为特征变为安静型优化模型迭代求解过程图

从图中可以看出在经过 30 次左右的迭代后，基本上就已经找到了最优结果，即最少的治疗费用为 31617.2 元，此时的退火温度已经降到温度接近于 0 左右。

计算得到的准确结果及 *CBTS* 、 *EPDS* 、 *HADS* 三种心理指标患病程度变化率的取值见下表：

表 12 婴儿行为特征变为安静型治疗费用最少结果数据表

	y	x_1	x_2	x_3
最优值	31617.17	7.00	9.50	8.00

从表中可以看出在即 *CBTS* 患病程度得分下降7分；即 *EPDS* 患病程度得分下降9.5分；即 *HADS* 患病程度得分下降8分时，婴儿行为特征从矛盾型变为安静型所耗费的治疗费用最少为31617.17元，由于患病程度得分一般下降程度为整数，而 x_2 的取值范围为9.5到19，故取 x_2 为10替换之前的9.5，带入目标函数中求解最终得到最少的治疗费用为33564.67元。

对比婴儿行为特征变为中等型的数据结果表，从中可以看出三种心理指标患病程度变化率的取值基本只变化了1左右，得到的最少的治疗费用从28159元上升到33564.67元。

5.3.5 优化模型的灵敏度检验

针对建立的优化模型，需要检测其面对波动值时的适应程度，观察模型的对 *CBTS* 、 *EPDS* 、 *HADS* 三种心理指标患病程度变化率的接受程度，对其进行灵敏度检验。对所建立的优化模型每一项系数波动 1%并计算结果，将得到的结果与起始值进行比较并绘图如下：

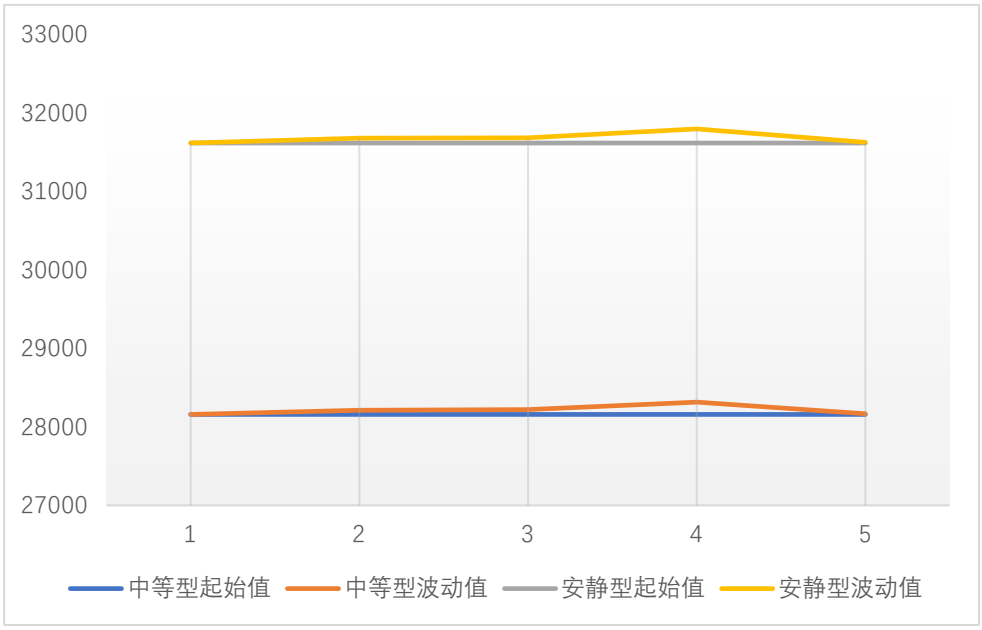


图 9 单目标最值优化模型的灵敏度检验图

从图中可以观察到，在各个系数项增加 1%后，波动后与波动前计算得到的最少治疗费用波动并不算太大，变化为中等型条件求得的最少治疗费用波动基本在 200 以内，变化为安静型条件求得的最少的治疗费用波动基本在 100 以内，即整体的波动程度在 1%以内，说明所建立的模型灵敏度好，不易受特殊条件的干扰。

计算得到波动后最少的治疗费用数据见下表：

表 6 行为特征约束条件下最少治疗费用结果数据

系数增加 1%	初始值	x_1	x_2	x_3	常数项
中等型	28159.02	28211.25	28221.55	28315.8	28169
安静型	31617.17	31678.11	31683.2	31796.37	31627.17

5.4 问题四模型的建立与求解

问题要求根据整晚睡眠时间、睡醒次数、入睡方式三种指标对婴儿的综合睡眠质量进行评判分类，以分类的结果数据建立与母亲的身体指标、心理指标的关联模型。由于睡眠质量的三个指标可以通过查阅文献以及直观的得出其为正向或负向指标，同时也能得到影响睡眠质量指标的三个最优值，在本题数据量较小，数据的质量程度不高的情况下，可以采用简单的评价模型对每一组编号进行评价，从高到低得到评价分数，并根据评价分数按照一定占比简单分类，即可得到婴儿睡眠质量的综合评判结果。

题目还需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，预测最后 20 组婴儿的综合睡眠质量的类别。因此本文尝试建立机器学习分类模型，对多种分类模型结果进行评估，比较准确率，选择建立随机森立分类模型。本文先对睡眠质量的评判结果量化，将优、良、中、差分别量化为 3、2、1、0。接着切分数据，70%分为训练集，30%分为测试集。通过训练集数据来建立随机森林分类模型。接着通过建立的随机森林来计算特征重要性。并对模型的分类结果进行评估。

5.4.1 基于灰色关联度对婴儿综合睡眠质量的评判分类

通过查阅文献可知整晚睡眠时间越长说明睡眠质量越好，睡醒次数越少说明睡眠质量越好，入睡方式越靠近自行入睡说明睡眠质量越好，越靠近陪伴入睡的说明睡眠质量

越差[4]。则最优的睡眠质量其整晚睡眠时间、睡醒次数、入睡方式可得分别为 10,0,5，并依此为母序列，每组编号的睡眠质量指标为子序列，建立灰色关联度模型，得到每组编号与最优睡眠质量的关联度，即得分，并排序。然后根据参考文献可知婴儿睡眠问题发生率为 49.77%[5]，可将睡眠质量为“差”的占比定为 49.77%，对于其余的优、良、中等类别分别各占据剩余 50.23%的三分之一，即各部分占比为 16.74%，最后将关联得分的顺序按照各类别的百分占比进行分类，得到最终的四种类别。

算法具体步骤如下：

Step 1 对负向指标睡醒次数的正向化处理。本文使用取值转换法，其转换公式如下：

$$f(x_s(2)) = \max x_s(2) - x_s(2), s \in (1:380) \tag{20}$$

式中 $x_s(2)$ 表示每一组编号中的第二个指标数据，即睡醒次数，用睡醒次数的最大值减去该组的睡醒次数进行正向化处理。

Step 2 对进行数据无量纲化处理。本文使用 MinMax 标准化，其转换函数如下：

$$f(x) = \frac{x_s(t) - \min x_s(t)}{\max x_s(t) - \min x_s(t)} \times (a-b) + b \tag{21}$$

x' = (x - min(x)) / (max(x) - min(x)) * (max_value - min_value) + min_value

其中，$x_s(t)$ 是原始数据，$f(x)$ 是标准化后的数据，$\min x_s(t)$ 和 $\max x_s(t)$ 分别是原始数据的最小值和最大值，a 和 b 是指定的最小值和最大值,本文中指定的最小值和最大值分别取 0.002 和 1。

Step 3 确立查找的最优睡眠质量为母序列，记为：

$$x_0 = \{x_0(1), x_0(2), x_0(3)\} \tag{22}$$

Step 4 将每组编号的睡眠质量指标确定为子序列，分别记为：

$$\begin{aligned} x_1 &= \{x_1(1), x_1(2), x_1(3)\} \\ x_2 &= \{x_2(1), x_2(2), x_2(3)\} \\ &\vdots \\ x_{380} &= \{x_{380}(1), x_{380}(2), x_{380}(3)\} \end{aligned} \tag{23}$$

Step 5 k 为指标，分辨系数 $\rho \in (0, +\infty)$。当 $\rho < 0.5463$ 时，分辨力最好，本文选取 $\rho = 0.5$。求解母序列与子序列之间的灰色关联系数值。

$$\xi_i(k) = \frac{\min\limits_s \min\limits_t |x_0(t) - x_s(t)| + 0.5 \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}{|x_0(t) - x_s(t)| + 0.5 \max\limits_s \max\limits_t |x_0(t) - x_s(t)|} \tag{24}$$

$|x_0(t) - x_s(t)|$ 代表母序列与子序列对应差值的绝对值，而 $\min\limits_s \min\limits_t |x_0(t) - x_s(t)|$ 表示所有所求绝对值中的最大值，$\max\limits_s \max\limits_t |x_0(t) - x_s(t)|$ 表示所有所求绝对值中的最小值。

Step 6 使用下式求解灰色关联度值。

$$r_i = \frac{1}{n}\sum_{k=1}^{n}\xi_i(k) \quad k=1,2,3 \tag{25}$$

n 为每组编号的睡眠质量指标数目，即公式为对每一行求平均值。

最后根据上式使用 matlab 计算结果，并根据优、良、中、差各自所占的比例，从得分由大到小的顺序中分类出四类所对应的编号，最后根据计算的得分与分类情况绘制表格如下所示：

表 13 灰色关联度得分分类表

编号	得分情况	睡眠质量
1	0.553787879	差
2	0.814814815	良
3	0.744444444	中
4	0.608465608	差
5	0.733333333	中
……	……	……
386	0.590598291	差
387	0.71957672	中
388	0.582539683	差
389	0.472013367	差

注：由于表格过大，具体 380 行编号的分类结果详见附件 q4 改。

5.4.2 基于随机森林分类模型预测

题目要求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，预测最后 20 组婴儿的综合睡眠质量的类别。因此本文尝试建立机器学习分类模型，对多种分类模型结果进行评估，比较准确率，选择建立随机森立分类模型。本文先对睡眠质量的评判结果量化，将优、良、中、差分别量化为 3、2、1、0。接着切分数据，70%分为训练集，30%分为测试集。通过训练集数据来建立随机森林分类模型[6]。接着通过建立的随机森林来计算特征重要性。并对模型的分类结果进行评估。

首先将数据划分为训练集、测试集，分别占比 70%和 30%。进行有放回地采样，进行袋外数据测试，但不进行洗牌与交叉验证。其他参数如下表所示：

表 14 随机森林模型参数表

参数名	参数值
节点分裂评价准则	gini
决策树数量	100
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阀值	0

通过建立的随机森林模型得到特征重要性，展示了各自变量的重要性比例。

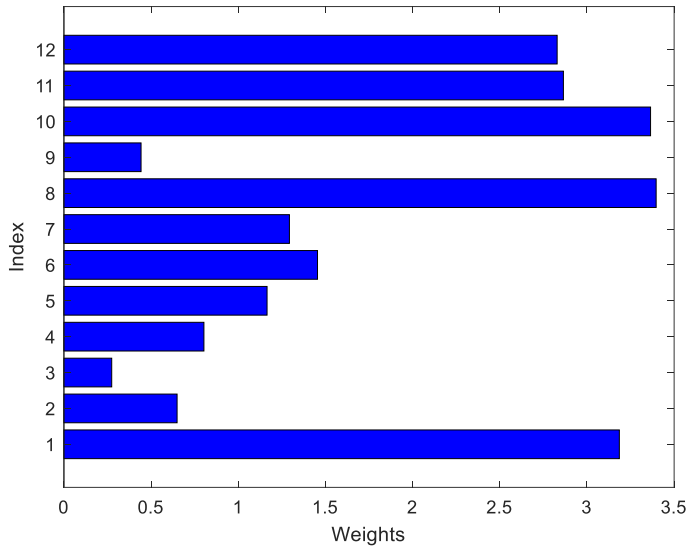


图 10 随机森林模型特征重要性图

5.3.3 随机森林分类模型评估

对于所建立的随机森林分类模型结果是否准确，可通过混淆明显观察出来。

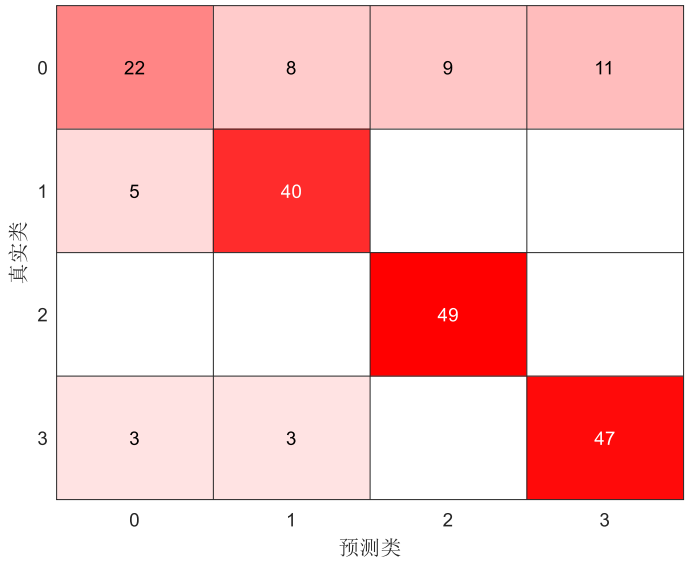


图 11 混淆矩阵热力图

由上图可知，左下斜对角线中的数值均是分类正确的结果。明显观察出分类准确率较高，模型较为良好。

接着使用准确率对模型进行初步评估。*TP* 表示正样本判断为正样本的样本数；*TN* 表示负样本判断为负样本的样本数；*FP* 表示负样本判断为正样本的样本数；*FN* 表示正样本判断为负样本的样本数。准确率为

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{26}$$

通过matlab软件计算得准确率为0.74与0.81之间，模型较为良好。

精确率为

$$precision = \frac{TP}{TP + FP} \tag{27}$$

召回率为

$$recall = \frac{TP}{TP + FN} \tag{28}$$

并绘制出 ROC 曲线如下图所示：

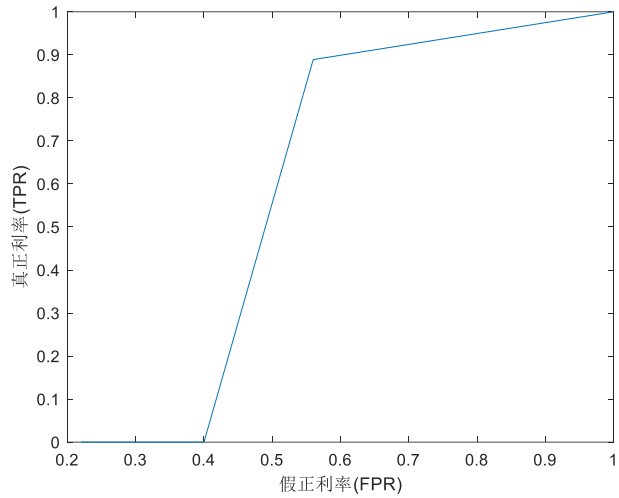


图 12 ROC 曲线图

ROC 曲线通常用来评估模型的质量。横纵坐标分别为真正例率(召回率)和假正例率。曲线越接近左上角，模型效果越好。由图可知，该模型效果较好。

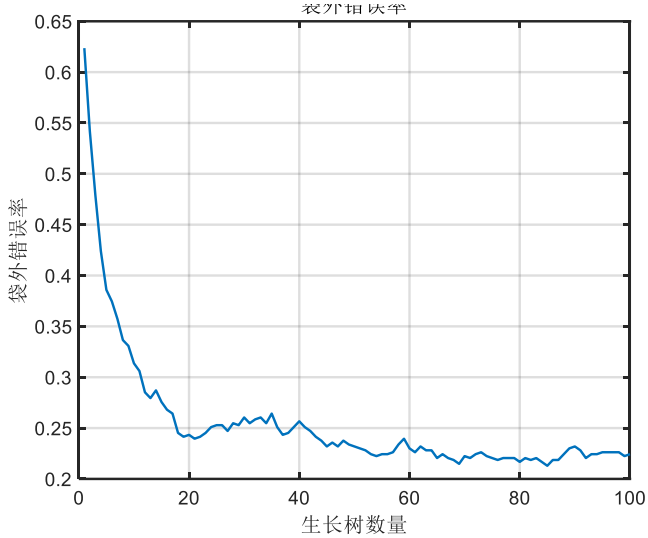


图 13 带外错误率图

袋外错误率是指在构建随机森林时，对于每个训练样本，将其未被抽到的决策树用于评估该样本，得到的分类结果与真实标签进行比较，统计分类错误的样本所占的比例。由上图可知泛化能力较好，分类效果较好。

5.4.4 结果分析

首先通过灰色关联度进行综合排序，并依据文献将婴儿的睡眠质量按照一定比例划分为优良中差四个类别。具体结果见附件 q4 改。

最终通过建立的随机森林模型预测 20 位婴儿的睡眠质量如下所示：

表 15 问题四分类结果表

编号	分类结果	编号	分类结果	编号	分类结果	编号	分类结果
1	2	6	2	11	3	16	0
2	1	7	2	12	0	17	0
3	0	8	0	13	0	18	3
4	3	9	0	14	1	19	3
5	1	10	1	15	2	20	2

表中 0 代表婴儿的睡眠质量差、1 代表婴儿的睡眠质量中、2 代表婴儿的睡眠质量良、3 代表婴儿的睡眠质量优。

5.5 问题五模型的建立与求解

题目要求在问题三的基础上，求解使得 238 号让 238 号婴儿的综合睡眠质量评级为优，治疗费用最少的情况。由于在第四问中使用随机森林分类并没有与有序逻辑回归相同的约束条件，故采用与问题三相同的方法，计算婴儿的综合睡眠质量为优的数据中三种心理病状得分的上四分位数与下四分位数作为婴儿综合睡眠质量的边界，然后目标函数与决策变量都与问题三中的优化模型相同，将婴儿综合睡眠质量的边界作为约束补充增加到问题三的约束条件中，最后通过智能算法进行求解得到在问题三基础下，婴儿综合睡眠质量变为优的情况下，所需的最少治疗费用，简要思路流程图如下所示：

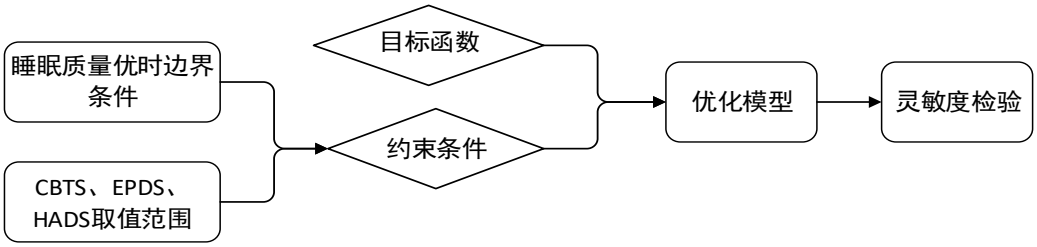


图 14 问题五思路流程图

5.5.1 婴儿优质睡眠质量边界条件的确定

同问题三相似，引入上四分位数与下四分位数作为婴儿综合睡眠质量为优的边界范围，以此来作为约束条件用于约束婴儿的综合睡眠质量评级为优的条件，从而能够建立优化模型求解出最少的治疗费用。

对于问题四已经给出的每组编号所对应的综合睡眠质量，在 Excel 中按照婴儿的综合睡眠质量筛选出综合睡眠质量为优所包含的编号及其三种心理指标，筛选后的结果详见“四分位数”Excel 表格，通过简单的 MATLAB 运算得到每一种婴儿综合睡眠质量的三种心理指标所对应的上四分位数与下四分位数，如下表所示：

表 16 婴儿综合睡眠质量为优对应的 3 种心理指标边界表

	优睡眠质量 CBTS	优睡眠质量 EPDS	优睡眠质量 HADS
上四分位数	9	12	9.25
下四分位数	1	3	4

从表中可以看出在婴儿睡眠质量为优的数据中 *CBTS* 取值范围为 1 到 9，*EPDS* 取值范围为 3 到 12，*HADS* 取值范围为 4 到 9.25。

5.3.3 最少治疗费用的单目标多约束优化模型的建立

由于问题基于问题三的基础上，仅增加一个使得婴儿综合睡眠质量评级为优的约束条件，故建立的优化模型中的决策变量与目标函数均与问题三的模型相同。

（1）约束条件的增加

约束条件则在问题三的基础上增加了 *CBTS* 、 *EPDS* 、 *HADS* 的取值范围，即：

$$1 \le 15 - x_1 \le 9 \tag{29}$$

其中 x_1，即 *CBTS* 患病程度变化率应该属于 6 到 14 区间

$$3 \le 22 - x_2 \le 12 \tag{30}$$

其中 x_2，即 *EPDS* 患病程度变化率应该属于 10 到 19 区间

$$4 \le 18 - x_3 \le 9.25 \tag{31}$$

其中 x_3，即 *HADS* 患病程度变化率应该属于 8.75 到 14 区间

（2）优化模型的确定

综上，将新增的约束条件加入到问题三建立的模型中，可以得到婴儿行为特征由矛盾型转变为中等型和综合睡眠质量为优时的最少治疗费用优化模型如下所示：

$$\begin{aligned} &y = \frac{2612}{3}x_1 + 695x_2 + 2440x_3 + 1000 \\ &\text{s.t.}\begin{cases} -2.598 \le 0.006x_1 - 0.027x_2 - 0.021x_3 - 0.0824 \le 0.329 \\ 2 \le 15 - x_1 \le 9 \\ 4 \le 22 - x_2 \le 12 \\ 5 \le 18 - x_3 \le 9.25 \end{cases} \end{aligned} \tag{32}$$

当求解在婴儿行为特征由矛盾型转变为安静型和综合睡眠质量为优时的最少治疗费用时，同样将婴儿优质睡眠质量边界条件再入问题三所建立的模型中，可以得到最少治疗费用的单目标多约束优化模型如下：

$$\begin{aligned} &y = \frac{2612}{3}x_1 + 695x_2 + 2440x_3 + 1000 \\ &\text{s.t.}\begin{cases} 0.329 \le 0.006x_1 - 0.027x_2 - 0.021x_3 - 0.0824 \\ 1 \le 15 - x_1 \le 8 \\ 3 \le 22 - x_2 \le 12 \\ 4 \le 18 - x_3 \le 9.25 \end{cases} \end{aligned} \tag{33}$$

5.3.4 单目标最值优化模型的智能算法求解及结果分析

对于建立的婴儿行为特征与综合睡眠质量为约束的单目标最值优化模型，采取智能算法-模拟退火算法进行求解，在 MATLAB 中通过 simulannealbnd 函数进行，首先随机生成一个初始解，通过函数的自动温度初始化过程来确定初始温度，并选取迭代次数为 200 次，约束条件作为定义变量的上下限，进行迭代求解，输出最优极值以及所对应的最优位置，得到全局最优解。

利用模拟退火算法的迭代求解由矛盾型转变为中等型及综合睡眠质量为优时的优化模型过程如下图所示：

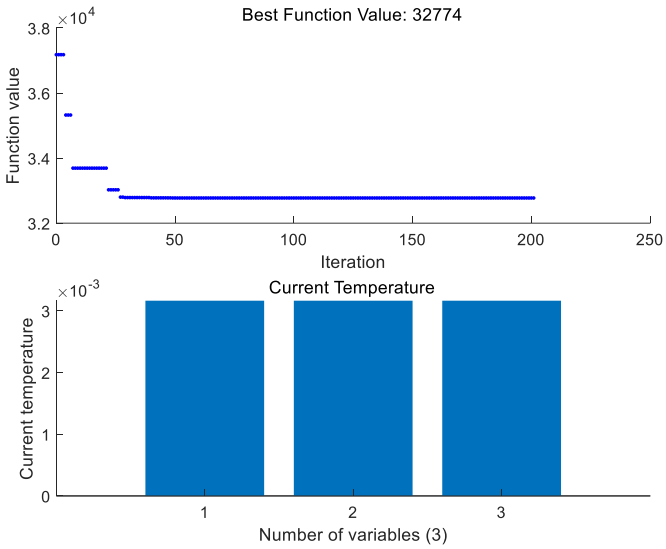


图 15 中等型特征+优质睡眠质量优化模型迭代求解过程图

从图中可以看出在经过 25 次左右的迭代后，基本上就已经找到了最优结果，即最少的治疗费用为 32774 元，此时的退火温度已经降到温度接近于 0 左右。

表 17 中等型特征+优质睡眠质量治疗费用最少结果数据表

	y	x_1	x_2	x_3
最优值	32774.00	6.00	10.00	8.75

从表中可以看出在 x_1 即 *CBTS* 患病程度得分下降 6 分； x_2 即 *EPDS* 患病程度得分下降 10 分； x_3 即 *HADS* 患病程度得分下降 8.75 分时，婴儿行为特征从矛盾型变为中等型同时为优质睡眠质量所耗费的治疗费用最少为 32774 元。

由于患病程度得分一般下降程度为整数，而 x_3 的取值范围为 8.75 到 14，故 x_3 取为 9 替换之前求解得到的 8.75，带入目标函数中求解最终得到最少的治疗费用为 35134 元。

对于当婴儿行为特征从矛盾型变为安静型同时为优质睡眠质量时最少的治疗费用，将约束条件根据建立的优化模型进行更行继续带入算法中求解即可，得到的迭代求解过程如下图所示：

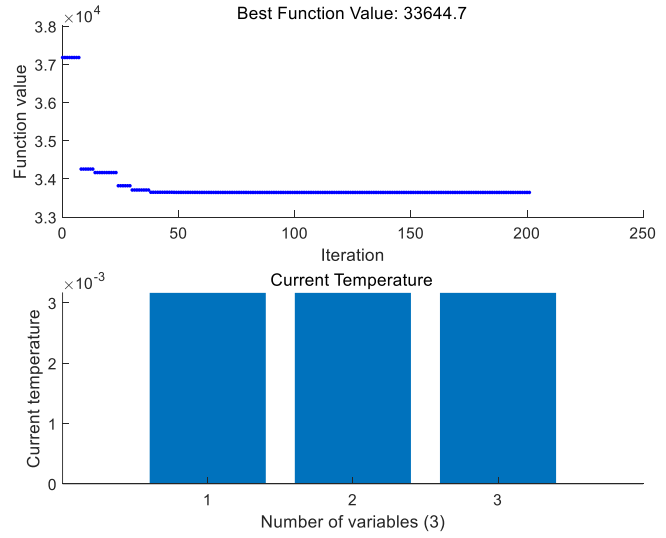


图 16 安静型特征+优质睡眠质量优化模型迭代求解过程图

从图中可以看出在经过 30 次左右的迭代后，基本上就已经找到了最优结果，即最少的治疗费用为 33644.7 元，此时的退火温度已经降到温度接近于 0 左右。

计算得到的准确结果及 *CBTS* 、 *EPDS* 、 *HADS* 三种心理指标患病程度变化率的取值见下表：

表 18 安静型特征+优质睡眠质量治疗费用最少结果数据表

	y	x_1	x_2	x_3
最优值	33644.7	7.00	10	8.75

从表中可以看出在即 *CBTS* 患病程度得分下降7分；即 *EPDS* 患病程度得分下降10分；即 *HADS* 患病程度得分下降8.75分时，婴儿行为特征从矛盾型变为安静型同时为优质睡眠质量所耗费的治疗费用最少为33644.7元，由于患病程度得分一般下降程度为整数，而 x_3 的取值范围为8.75到14，故取 x_3 为9替换之前的8.75，带入目标函数中求解最终得到最少的治疗费用为36004.67元。

对比婴儿行为特征变为中等型同时为优质睡眠质量的数据结果表，从中可以看出三种心理指标患病程度变化率的取值基本没有什么变化，得到的最少的治疗费用从35134元上升到36004.67元，可以得出，在婴儿睡眠质量为优时，婴儿的行为特征对治疗费用的影响不大。

5.3.5 优化模型的灵敏度检验

针对建立的优化模型，需要检测其面对波动值时的适应程度，观察模型的对 *CBTS* 、 *EPDS* 、 *HADS* 三种心理指标患病程度变化率的接受程度，对其进行灵敏度检验。对所建立的优化模型每一项系数波动 1%并计算结果，将得到的结果与起始值进行比较并绘图如下：

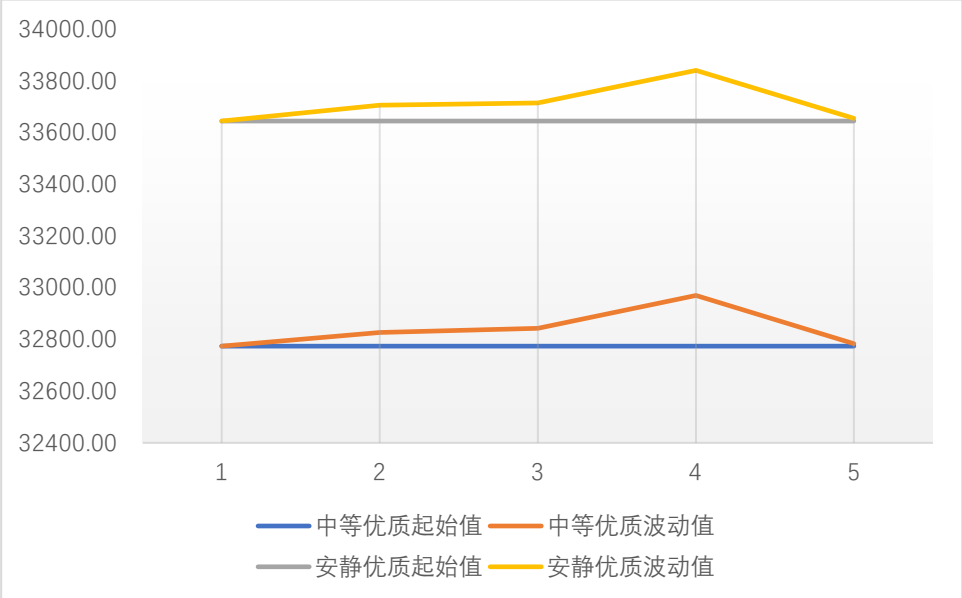


图 17 单目标最值优化模型的灵敏度检验图

从图中可以观察到，在各个系数项增加 1%后，波动后与波动前计算得到的最少治疗费用波动并不算太大，变化为中等型同时优质睡眠质量条件求得的最少治疗费用波动基本在 200 以内，变化为安静型同时优质睡眠质量条件求得的最少的治疗费用波动同样基本也在 200 以内，即整体的波动程度在 1%以内，说明所建立的模型灵敏度好，不易

受特殊条件的干扰。

计算得到波动后最少的治疗费用数据见下表：

表 6 行为特征与睡眠质量约束条件下最少治疗费用结果数据

系数增加 1%	初始值	x_1	x_2	x_3	常数项
中等+优质睡眠	32774.00	32826.24	32843.50	32970.00	32748.00
安静+优质睡眠	33644.67	33705.61	33714.17	33840.67	33654.67

六、模型的评价与推广

6.1 模型的优缺点

6.1.1 优点

(1) 婴儿行为特征有一定的顺序性，有序逻辑回归模型适合处理有序分类变量，能够更好地利用不同类别之间的顺序信息。简单直观、预测能力强，可解释性强。并且该算法相对简单。

(2) 灰色关联度模型具有将强的适应性、灵活性，结果的直观性和解释性都较强。

(3) 随机森林分类模型，适用于处理复杂非线性问题，适应各种数据分布。对大规模数据集和高维特征有良好的扩展性，在不平衡数据集上同样也具有较好的性能。

(4) 规划模型精确性高、效率高、可复用性强：数学建模规划模型可以精确地描述实际问题，能够得到较为精确的结果；规划模型可以通过数学优化算法高效地求解问题，节省人力和时间成本；已经建立的数学建模规划模型可以在类似问题中复用，减少重复工作。

6.1.2 缺点

有序逻辑回归模型：需要大量的训练数据。

规划模型：对问题的复杂性有一定要求，可能无法完全覆盖实际情况，求解过程相对困难。需要大量数据进行参数估计和模型验证。

6.2 模型的改进与推广

6.2.1 模型的改进

分类模型可利用特征工程或设计损失函数进行模型改进。针对具体问题，选择不同的改进方法。也可以进行模型融合，将有序逻辑回归与其他分类模型进行组合，从而提高分类质量。

规划模型可以通过增加数据进行模型改进，从而提高模型的精确性、可解释性。也可通过建立完善的评估指标和评估方法，对模型进行全面的评估和调优，确保模型在各种情况下都能产生高质量的规划结果。

6.2.2模型的推广

分类模型可解决许多不同领域的问题，如信用评分和风险预测、市场营销和消费行为、车辆识别、医学图像分析、文本分类和请按发现等。

规划模型也主要用于建立和预测变量之间的关系，可解决生产和物流规划、项目管理和资源分配、资源的分配和任务的优先级、排班和员工调度、能源管理和优化、城市和交通规划、市场需求预测和库存管理、基础设施规划和优化。从而最大程度地满足人们的需求，最优化项目进度和资源利用，并提高社会效益。

七、参考文献

[1] 何晓群.多元统计分析.北京：中国人民大学出版社，2012.
[2] 徐维超. 相关系数研究综述[J]. 广东工业大学学报,2012,29(3):12-17.
[3] 张俊艳,余敏. 基于有序逻辑回归的标准必要专利价值影响因素研究[J]. 电子科技大学学报（社会科学版）,2018,20(1):15-19.
[4] 刘卓娅, 郭玉琴, 宋娟娟, 等. 婴幼儿入睡方式及其对睡眠质量的影响[J]. 中国当代儿科杂志, 2022, 24(3): 297-302.
[5] 陈孝颖. 某三甲医院1～11月龄婴儿睡眠现况及影响因素研究[D]. 南昌大学, 2022.
[6] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016.

八、附录

附录清单

附录一 问题一PCA降维5个母亲的身体指标

1、KMO 检验和 Bartlett 的检验

表 19 KMO 检验和 Bartlett 检验表

KMO 值		0.505
Bartlett 球形度检验	近似卡方	742.69
	df	3
	P	0.000

通过 KMO 检验（KMO>0.5），说明了题项变量之间是存在相关性的，符合主成分分析要求。Bartlett 检验：P<0.05，呈显著性，则可以进行主成分分析。

2、方差解释表

表 20 方差解释表

成分	特征根	方差解释率(%)	累积方差解释率(%)
1	1.28	25.599	25.599
2	1.211	24.215	49.815
3	0.97	19.4	69.215
4	0.794	15.871	85.086
5	0.746	14.914	100

方差解释表中，在主成分 3 时，总方差解释的特征根低于 1.0，变量解释的贡献率达到 69.215。

3、因子载荷系数表

表 21 因子载荷系数及得分系数表

变量	主成分 1 因子载荷系数	公因子方差	因子得分系数
母亲年龄	0.607	0.368	0.474
婚姻状况	0.393	0.154	0.307
教育程度	0.537	0.289	0.42
妊娠时间（周数）	-0.49	0.24	-0.383
分娩方式	0.478	0.228	0.373

4、碎石图

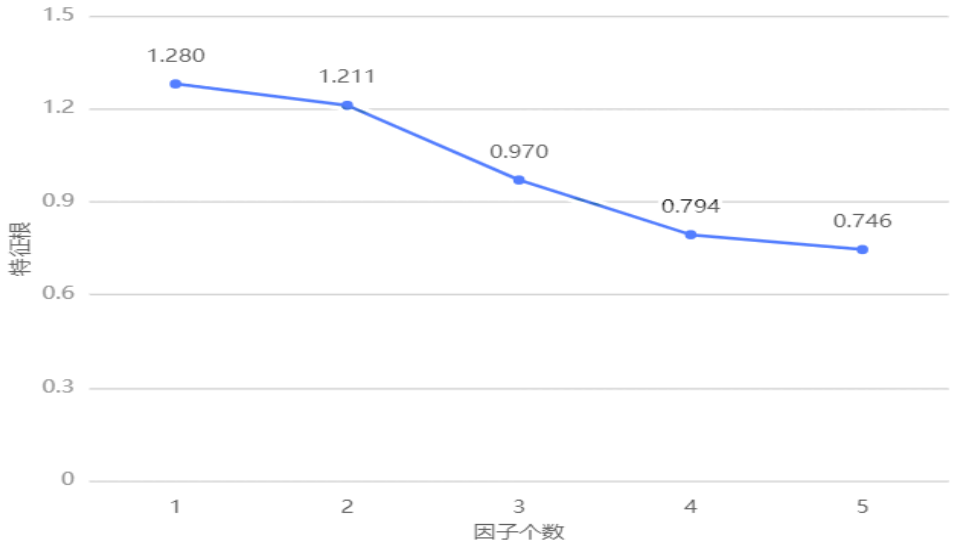


图 18 碎石图

附录二 问题一正态性检验

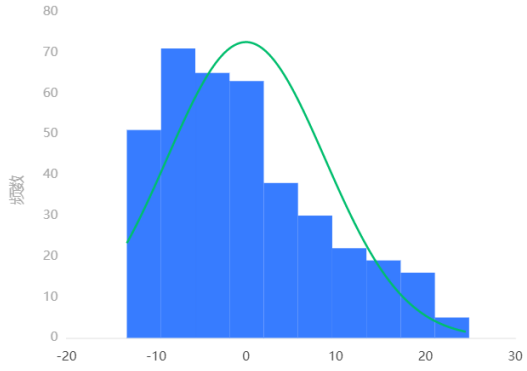
1、总体描述结果

表 22 正态性检验表

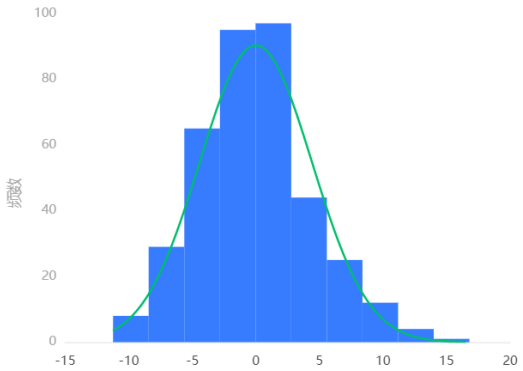
变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验
心理指标	380	-1.469	0	8.796	0.695	-0.297	0.945(0.000***)
身体指标	380	-0.168	0	4.419	0.285	0.486	0.99(0.009***)
婴儿行为特征	380	2	1.811	0.621	0.152	-0.536	0.772(0.000***)
整晚睡眠时间	380	10	10.172	1.448	-1.003	1.148	0.9(0.000***)
睡醒次数	380	1	1.471	1.622	1.718	4.326	0.808(0.000***)
入睡方式	380	4	3.045	1.403	-0.305	-1.404	0.831(0.000***)

P<0.05，该数据不满足正态分布。样本峰度绝对值小于 10 并且偏度绝对值小于 3，所以需结合正态分布直方图、PP 图或者 QQ 图可以描述为基本符合正态分布。

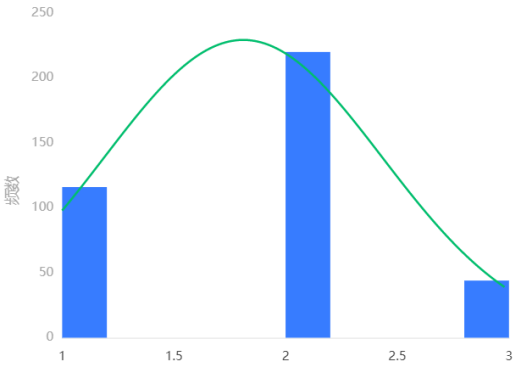
2、正态性检验直方图



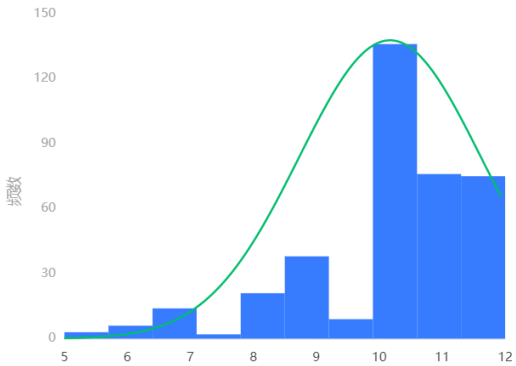
心理指标



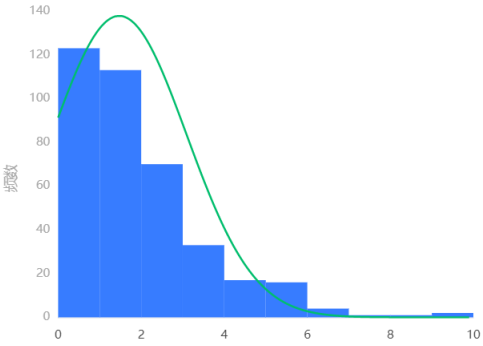
身体指标



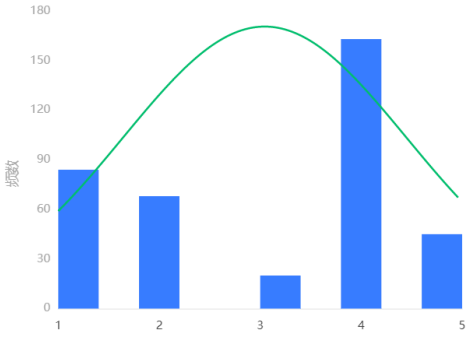
婴儿行为特征



整晚睡眠时间



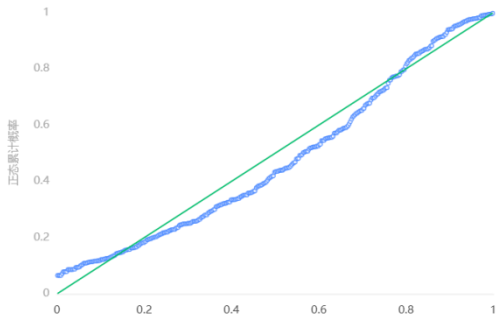
睡醒次数



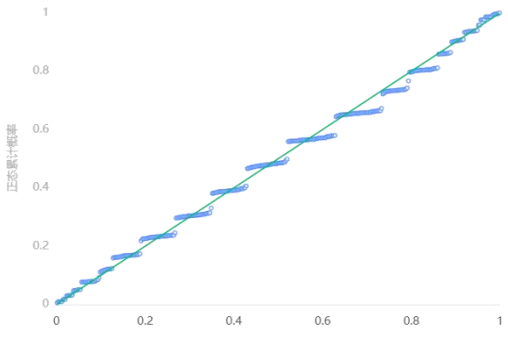
入睡方式

图 19 正态性检验直方图

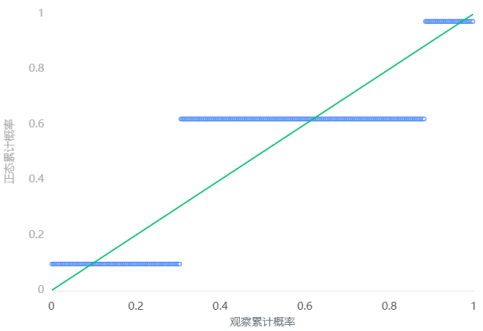
3、正态性检验 P-P 图



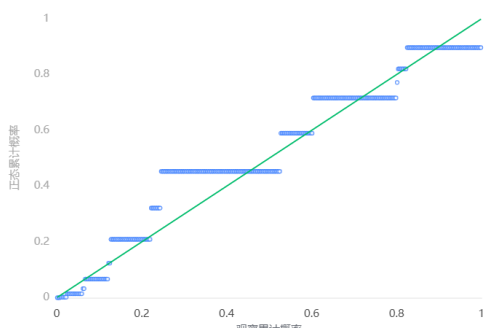
心理指标



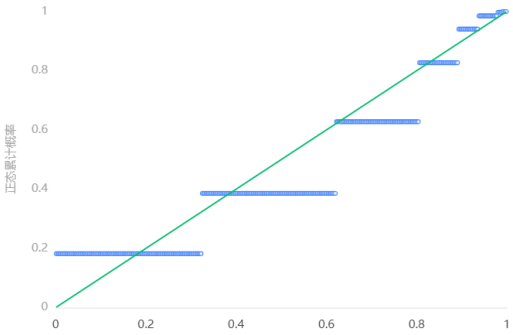
身体指标



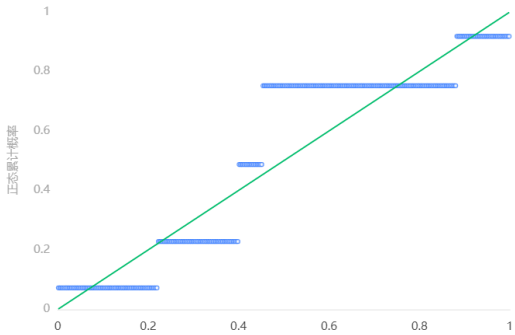
婴儿行为特征



整晚睡眠时间



睡醒次数



入睡方式

图 20 正态性检验 P-P 图

附录三 问题一斯皮尔曼相关系数部分代码（matlab）

```
var1 = data(:, 1);
var2 = data(:, 2);
var3 = data(:, 3);
var4 = data(:, 4);
var5 = data(:, 5);
var6 = data(:, 6);
correlation_matrix = corr([var1, var2, var3, var4, var5, var6], 'Type', 'Spearman');
figure;
heatmap({'心理指标', '身体指标', '婴儿行为特征', '整晚睡眠时间', '睡醒次数', '入睡方式'}, {'心理指标', '身体指标', '婴儿行为特征', '整晚睡眠时间', '睡醒次数', '入睡方式'}, correlation_matrix);
title('相关系数矩阵');
```

附录四 问题三规划问题部分代码（matlab）

```
fitnessFunc = @(x) (2612/3*x(1) + 695*x(2) + 2240*x(3) + 1000);
constraintFunc = @(x) [ 0.006 x(1) - 0.027 x(2) - 0.0824 x(3) - 0.0824 + 2.598 ; 0.006 x(1) - 0.027 x(2) - 0.0824 x(3) - 0.0824 -0.329];
x0 = [9.777; 10.8; 9]; % 初始解
lb = [6; 9; 7]; % 下限
ub = [13; 18; 13]; % 上限
options = optimoptions('simulannealbnd', 'MaxIterations', 200, 'Display', 'off', 'PlotFcn', {@saplotbestf, @saplottemperature});
[x, fval] = simulannealbnd(fitnessFunc, x0, lb, ub, options);
fprintf('y 的最小值是：%.2f\n', fval);
fprintf('x1 的值为：%.2f\n', x(1));
fprintf('x2 的值为：%.2f\n', x(2));
fprintf('x3 的值为：%.2f\n', x(3));
%% 模拟退火算法检验
fitnessFunc = @(x) (((2612/3)* 1.01)*x(1) + 695*x(2) + 2240*x(3) + 1000); %X1 系数增加 1%
%fitnessFunc = @(x) (2612/3*x(1) + (695* 1.01)*x(2) + 2240*x(3) + 1000); %X2 系数增加 1%
%fitnessFunc = @(x) (2612/3*x(1) + 695*x(2) + (2240* 1.01)*x(3) + 1000); %X3 系数增加 1%
%fitnessFunc = @(x) (2612/3*x(1) + 695*x(2) + 2240*x(3) + (1000 * 1.01)); %常数项系数增加 1%
constraintFunc = @(x) [ 0.006 x(1) - 0.027 x(2) - 0.0824 x(3) - 0.0824 + 2.598 ; 0.006 x(1) - 0.027 x(2) -
```

```
0.0824 x(3) - 0.0824 -0.329];
x0 = [9.777; 10.8; 9]; %  初始解
lb = [6; 9; 7]; %  下限
ub = [13; 18; 13]; %  上限
options = optimoptions('simulannealbnd', 'MaxIterations', 200, 'Display', 'off', 'PlotFcn', {@saplotbestf, @saplottemperature});
[x, fval] = simulannealbnd(fitnessFunc, x0, lb, ub, options);
fprintf('y 的最小值是：%.2f\n', fval);
fprintf('x1 的值为：%.2f\n', x(1));
fprintf('x2 的值为：%.2f\n', x(2));
fprintf('x3 的值为：%.2f\n', x(3));
```

附录五 问题四灰色关联度模型代码（**matlab**）

```
x=xlsread('q4.xlsx');
x=x(:,13:end)
%负向指标的处理
x(:,2)=max(x(:,2))-x(:,2);
x=mapminmax(x',0.002,1)
data=x';
ck=data(1,:); %提出参考数列
bj=data(2:end,:); %提出比较数列
n=size(data,2); %求矩阵的列数，即观测时刻的个数
m=size(bj,1); %求比较数列的行数
for i=1:m
    t(i,:)=bj(i,:)-ck;
end
mn=min(min(abs(t')));
mx=max(max(abs(t')));
rho=0.5;
ksi=(mn+rho*mx)./(abs(t)+rho*mx);
r=ksi'；r=sum(ksi')/n ；
c = size(ksi,2);
plot(ksi(:,1:c),'-'
```

附录六 问题四随机森林分类树图

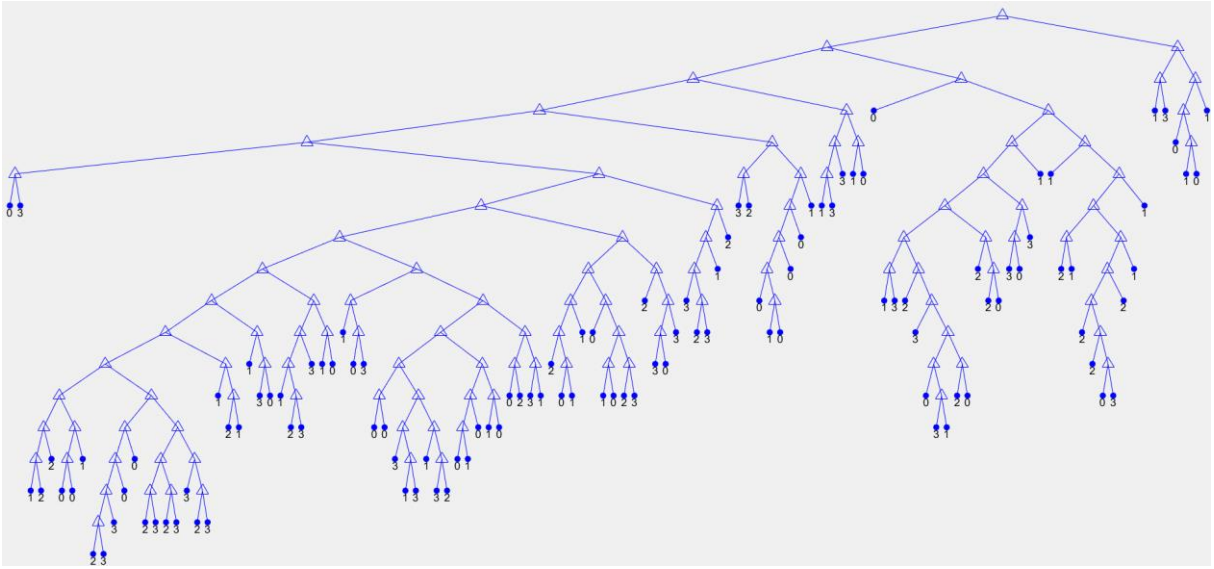


图 21 随机森林分类树图

附录七 问题四随机森林分类模型部分代码（matlab）

```
%% 随机森林分类
load data2
unique_labels = unique(data(:, end));
num_classes = numel(unique_labels);
class_counts = histcounts(data(:, end), [unique_labels; max(unique_labels)+1]);
max_count = max(class_counts);
oversampled_data = [];
for i = 1:num_classes
    class_data = data(data(:, end) == unique_labels(i), :);
    num_oversample = max_count - class_counts(i);
    oversampled_samples = randsample(size(class_data, 1), num_oversample, true);
    oversampled_class_data = class_data(oversampled_samples, :);
    oversampled_data = [oversampled_data; class_data; oversampled_class_data];
end
data=[oversampled_data(:,1:8);x_vali]
feature_column = 3;
feature_data = data(:, feature_column);
unique_values = unique(feature_data);
one_hot_encoded = zeros(size(data, 1), numel(unique_values));
for i = 1:size(data, 1)
    value = feature_data(i);
    index = find(unique_values == value);
    one_hot_encoded(i, index) = 1;
end
data_with_one_hot = [data(:, 1:feature_column-1) one_hot_encoded data(:, feature_column+1:end)];
data_num=[data_with_one_hot(:,1:2),data_with_one_hot(:,8:12)]
mean_values = mean(data_num);
```

```
std_values = std(data_num);
normalized_data = (data_num - mean_values) ./ std_values;
data=[normalized_data(1:756,1:2),data_with_one_hot(1:756,3:7),normalized_data(1:756,3:7),oversampled
_data(1:756,9)]
x_vali=[normalized_data(757:776,1:2),data_with_one_hot(757:776,3:7),normalized_data(757:776,3:7)]
X = data(:,1:12); % 特征矩阵
y = data(:,13); % 标签向量
trainRatio = 0.7; % 训练集比例
testRatio = 0.3; % 测试集比例
[trainInd,testInd] = dividerand(size(X,1), trainRatio, testRatio);
X_train = X(trainInd,:);
y_train = y(trainInd);
X_test = X(testInd,:);
y_test = y(testInd);
numTrees = 100;
classificationTreeEnsemble = TreeBagger(numTrees,X_train,y_train,'OOBpredictorImportance','on');
y_pred = predict(classificationTreeEnsemble, X_test);
accuracy = sum(str2double(y_pred) == y_test) / numel(y_test);
disp(['准确度：' num2str(accuracy)]);
predicted_labels=str2double(y_pred);
true_labels=y_test;
confusion = confusionmat(true_labels, predicted_labels);
accuracy = sum(diag(confusion)) / sum(confusion(:));
precision = confusion(2, 2) / sum(confusion(:, 2));
recall = confusion(2, 2) / sum(confusion(2, :));
f1_score = (2 * precision * recall) / (precision + recall);
scores =predicted_labels; % 假设有预测的概率得分向量
labels = true_labels; % 假设有真实的类别标签向量
[~, ~, ~, AUC] = perfcurve(labels, scores, 2);
predictions =predicted_labels; % 模型的预测结果
labels = true_labels; % 真实标签，0 表示负例，1 表示正例
positive_count = sum(labels == 1);
negative_count = sum(labels == 0);
thresholds = unique(predictions); % 不同的阈值
tpr = zeros(size(thresholds));
fpr = zeros(size(thresholds));
for i = 1:numel(thresholds)
    threshold = thresholds(i);
    predicted_labels = predictions >= threshold;
    true_positive = sum(predicted_labels(labels == 1) == 1);
    false_positive = sum(predicted_labels(labels == 0) == 1);
    tpr(i) = true_positive / positive_count;
    fpr(i) = false_positive / negative_count;
end
```

附录八 问题五规划求解部分代码（matlab）

```
fitnessFunc = @(x) (2612/3*x(1) + 695*x(2) + 2240*x(3) + 1000);
constraintFunc = @(x) [ 0.006 x(1) - 0.027 x(2) - 0.021 x(3) - 0.0824 + 2.598 ; 0.006 x(1) - 0.027 x(2) - 0.021 x(3) - 0.0824 -0.329];
x0 = [9.777; 10.8; 9]; % 初始解
lb = [6; 10; 8.75]; % 下限
ub = [13; 18; 13]; % 上限
options = optimoptions('simulannealbnd', 'MaxIterations', 200, 'Display', 'off', 'PlotFcn', {@saplotbestf, @saplottemperature});
[x, fval] = simulannealbnd(fitnessFunc, x0, lb, ub, options);
fprintf('y 的最小值是：%.2f\n', fval);
fprintf('x1 的值为：%.2f\n', x(1));
fprintf('x2 的值为：%.2f\n', x(2));
fprintf('x3 的值为：%.2f\n', x(3));
```