中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

学　　校	长春理工大学
参赛队号	**20101860020**
队员姓名	**1.张聆铭**
	2.刘阳
	3.谷晓雁

中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

题 目　　基于多元回归的辛烷值损失模型构建及参数优化

摘　　要：

本文研究了汽油辛烷值损失预测模型。主要创新点在于**结合因子分析和多元回归分析**，建立了汽油辛烷值损失预测模型，针对操作变量优化问题建立了**整数规划**模型，通过**遗传算法**对模型进行寻优，**在保证硫含量不大于 5μg/g 的前提下，199 个样本辛烷值损失降低在 30%以上，平均降低辛烷值损失 44.46%.**

问题一：我们对原始数据和样本数据中可能影响后续建模分析的因素进行了处理，具体包括：（1）对数据存在部分残缺的位点进行插值，**共计插值 181 条数据**；（2）删除数据残缺较多以至于无法补全或者数据全部为空值的位点，**共计删除 17 个异常位点**；（3）对操作范围不符合操作经验和工艺要求的变量予以筛选，**筛选出 6 条数据**；（4）按照拉依达准则对离群的奇异值进行检测。经过数据处理后，**保留了 325 个样本的 343 个变量，共计剩余有效数据 128210 项**。

问题二：我们对数据中 343 个变量进行因子分析，将高维空间中 367 个变量降维到低维空间的 26 个因子，**累计解释方差 87.09%**，每个因子**高内聚**，各因子之间**低耦合**。将 26 个因子作为自变量建立基于多元回归辛烷值损失预测模型，并**使用未参与拟合的样本数据进行了模型验证，94.15%预测值相对误差小于 0.3，97.85%预测值残差在±0.5 区间内**。

问题三：我们对操作变量优化**建立了整数线性规划模型**，使用**逐步回归模型建立产品硫含量的约束函数**，通过**遗传算法对整数规划模型进行求解**，在保证硫含量不大于 5μg/g 的前提下，**325 个样本中有 199 个辛烷值降低幅度在 30%以上**。

最后，对于汽油辛烷值损失预测模型的优缺点进行了评价，提出了未来改进方向和推广中可能遇到的问题。

关键字：因子分析；多元回归；整数规划；遗传算法；辛烷值损失

目 录

1. 问题重述

1.1 问题背景

汽油辛烷值（RON）是车用汽油最重要的品质指标之一，辛烷值越高表示汽油的抗爆性越好，提高辛烷值对汽车动力经济性能有十分重要的意义。辛烷值与汽油中的硫、烯烃等含量有关。根据 GB17930-2016《车用汽油》要求，车用汽油（V）的烯烃含量体积分数不大于 24%，硫含量不大于 10mg/kg，在符合相关国家标准降低汽油中硫、烯烃含量的同时，也要尽量保持较高的辛烷值。据测算，辛烷值损失量每降低 0.1 mg/kg 可直接增加经济效益 1211.6 万元/年。

过去采用数据关联和机理分析的方法对化工过程建模，但由于工艺过程的复杂性、设备的多样性以及操作变量（控制变量）之间具有高度非线性和相互强耦联的关系，而且辛烷值测定具有滞后性，对数据处理有较高的要求。辛烷值和操作变量之间的数学模型分析工作较少，如何控制操作变量、建立损失预测模型对于提高汽油精制的经济效益具有重要意义。

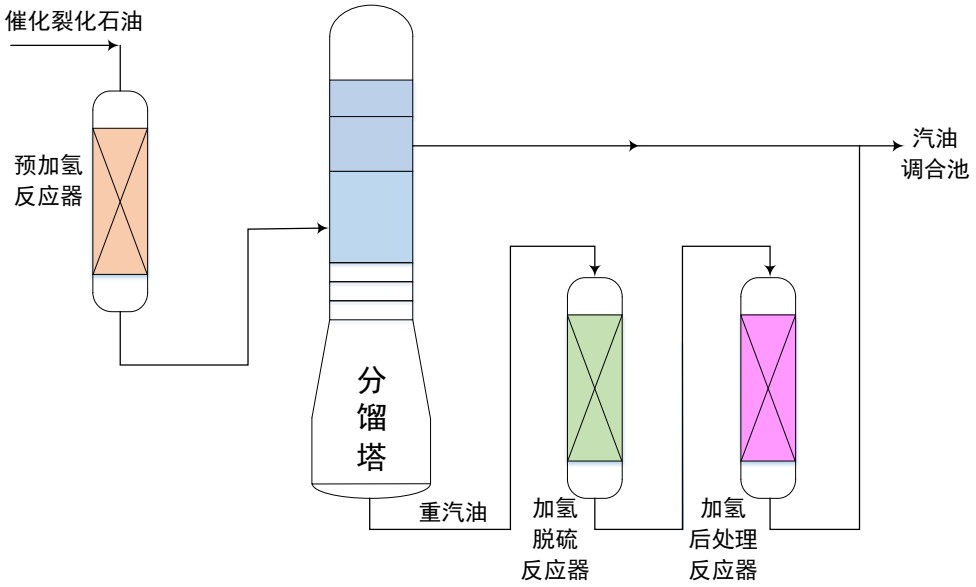

图 1.1 催化裂化汽油精制脱硫装置

1.2 问题重述

某石化企业催化裂化汽油精制脱硫装置运行 4 年，积累了大量历史数据， 附件一是该石化企业 2017 年 4 月 17 日至 2020 年 5 月 26 日采集的 325 个样本数据以及与每个样本对应的 367 个变量，附件三提供了其中 285 号和 313 号样本的原始数据，附件二是确定样本数据的方法，能够为原始数据和样本数据处理提供指导，附件四是 354 个操作变量的具体信息，要求我们利用以上数据和信息通过数据挖掘技术建立汽油辛烷值（RON）损失的预测模型，给出每个样本优化操作条件，在保证脱硫效果的前提下尽量降低辛烷值损失 30%以上。

具体需要解决以下问题：

1) 数据处理
   a) 整定和筛选附件三中的原始数据；
   b) 预处理附件一中的样本数据；
2) 提取主要变量建立辛烷值损失预测模型
   a) 选取合适的降维方法筛选建模主要变量；
   b) 构造辛烷值损失预测模型并验证；
3) 优化主要操作变量方案并进行可视化展示
   a) 根据预测模型优化主要变量操作方案；
   b) 以图形展示辛烷值和硫含量在优化调整中的变化情况。

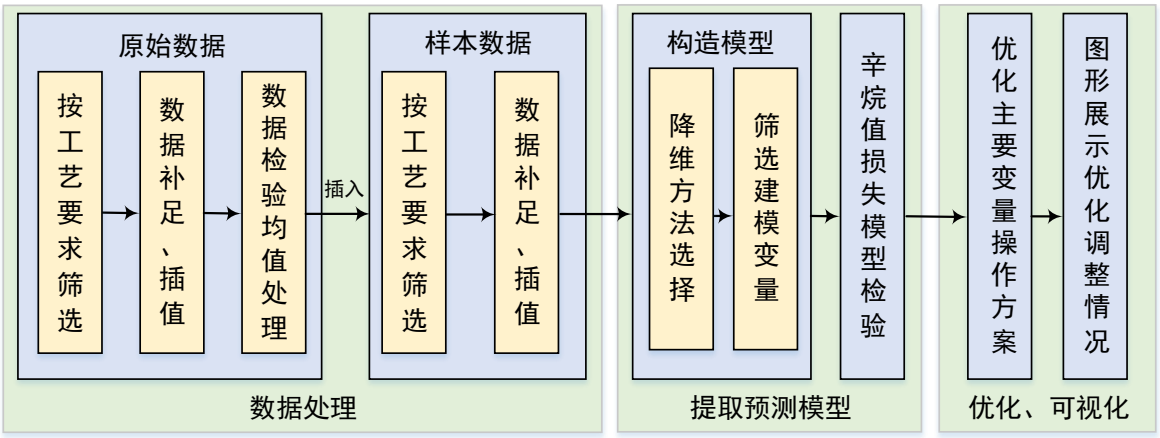

图 1.2 问题处理流程

2. 模型假设

考虑到实际情况，本文做出如下假设：

1) 假设题目中所提供的数据都是真实有效的。
2) 假设未统计的其他操作变量和原料信息与辛烷值的相关性较小，可忽略不计。
3) 假设去除异常位点和数据，不影响数据的整体分布。
4) 假设所有数据都是同一套实验装置采集的。

3. 符号系统

符号	符号说明	单位
$X(x_i)$	原始数据变量	-
$F(f_i)$	原始数据变量的公共因子	-
σ	标准误差	-
$A(a_{ij})$	因子载荷矩阵	-
$\varepsilon(\varepsilon_i)$	特殊因子	-
b_{ji}	第i个原始变量对第j个因子的得分	-
β_j	回归参数	-
μ_i	随机扰动	-
e	残差	-
E	相对误差	-
R_l	辛烷值损失量	-
$x_{i,j}$	第j个样本第i个变量	-
$x_{i\min}$	第i个变量的最小取值范围	-
$x_{i\max}$	第i个变量的最大取值范围	-
n_i	第i个变量的步进次数	-
Δ_i	第i个变量的步进值	-
p	变量的个数	-
m	主要变量的个数	-
k	样本的个数	-
S	产品性质的硫含量	ug/g
r_{sp}	Spearman 相关系数	-
d_i	两组数据的等级之差	-

4. 数据处理

4.1 问题分析

285 号、313 号样本原始数据包含样本的原料性质、产品性质、待生吸附剂性质、再生吸附剂性质和操作变量等 367 个变量，问题要求对这 367 个变量进行数据整定和筛选，然后加入到附件一所提供的 325 个数据样本中，供后续建模分析使用。

表 4.1 两个附件文件的数据情况

附件名	数据情况
附件一：325 个样本数据	325 个样本的 367 个变量数据 （其中包括 354 个操作变量）
附件三：285 号和 313 号样本原始数据	285 号 13 个样本产品、原料、吸附剂性质 313 号 13 个样本产品、原料、吸附剂性质 285 号样本 354 个操作变量 313 号样本 354 个操作变量

在样本数据和原始数据中可能存在影响后续建模和分析的因素有：

1) 数据存在部分残缺；
2) 位点的数据残缺较多以至于无法补全；
3) 位点的数据全部为空值；
4) 变量的操作范围不符合操作经验和工艺要求；
5) 存在部分异常离群值。

对此，我们首先采用最大最小限幅法对原始数据操作变量进行筛选，删除超出工艺和操作范围的数据，并将残缺较少的数据使用前后两小时数据平均值进行插值补全，删除残缺较多无法补全的数据。在此基础上对数据项按照拉依达准则（3σ 准则）进行异常值检测，将操作变量求取平均加入数据样本中，再对样本数据进行预处理，具体包括最大最小限幅法筛选、删除异常位点数据和缺失项插值。数据处理的流程图如下图 4.1 所示：

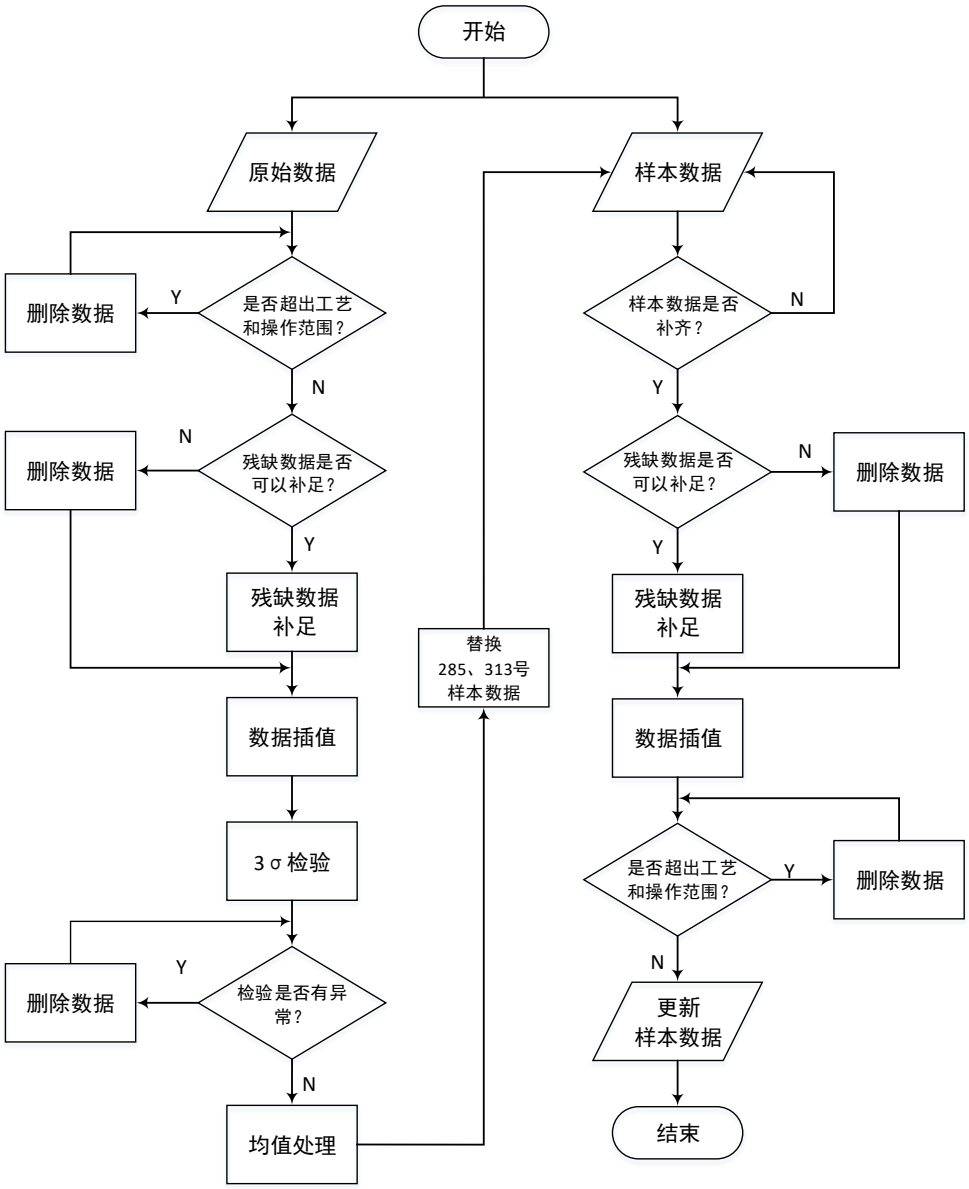

图 4.1 数据处理的流程

4.2 数据处理

4.2.1 原始数据整定与筛选

数据处理的对象包括样本数据和原始数据，由于需要将原始数据处理结果加入到样本数据中，所以我们先对原始数据进行处理，其处理步骤如下：

步骤 1：最大最小限幅法筛选：根据操作变量信息提供的范围，对原始数据进行筛选；

步骤 2：删除超出工艺和操作范围的样本；

步骤 3：删除缺值：删除残缺数据较多而无法补充的原始数据；

步骤 4：插值：对能够补全的部分数据缺失使用前后两小时数据平均值进行代替；

步骤 5：异常值检测：根据拉依达准则（3σ 准则）进行异常值检测；

步骤 6：均值处理：使用 2 小时内操作变量平均值作为辛烷值操作变量的最终数据。

在实际生产过程中，受到实际生产工艺和条件的限制，操作变量只能在一定范围内变化，所以我们根据操作变量信息所提供的范围，将超出范围的样本数据删除。例如，非净化风进装置流量的取值范围为 0-900 N m^3/h，但是在原始数据中的第 7 次采样存在超出该范围的数据，我们通过最大最小限幅筛选法将原始数据中超出范围的数据删除，如图 4.2 所示：

表 4.2 原始数据删除前后对比

删除前	删除后
543.3043	543.3043
595.8862	595.8862
648.4681	648.4681
511.6857	511.6857
446.7246	446.7246
463.6430	463.6430
910.5219*	-
784.4914	784.4914
514.8187	514.8187
473.3051	473.3051
463.4757	463.4757
422.0896	422.0896
472.5419	472.5419
866.6923	866.6923
420.2141	420.2141

注：灰底上标*为删除值

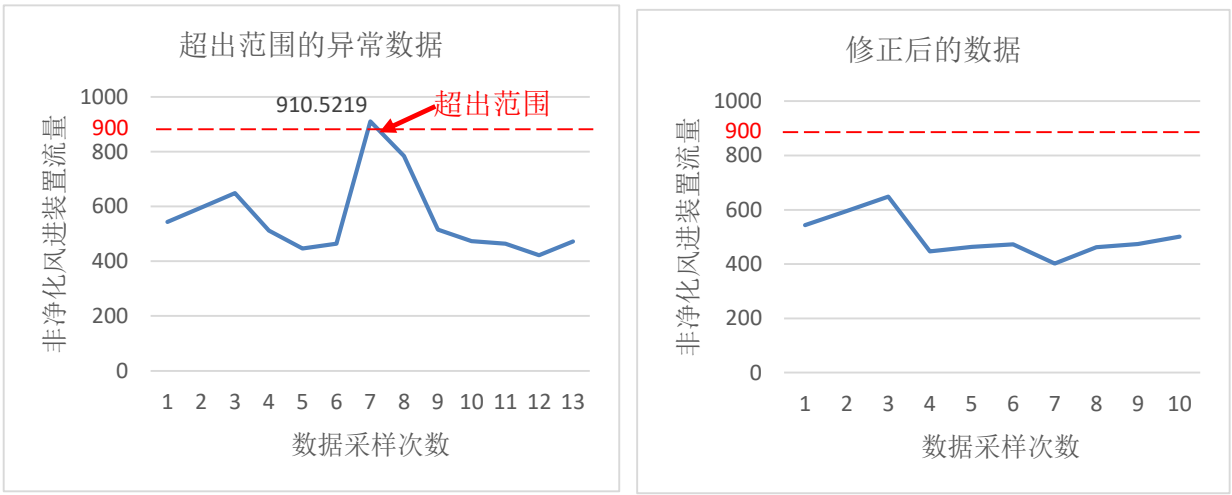

图 4.2 数据删除前后

按照同样方法对其他原始数据进行筛选，共计删除了 31 条样本原始数据，保留了 285 号样本原始数据 40 条，313 号样本原始数据 9 条。

在保留的 49 条原始数据中，有 5 处数据项存在缺失，使用前后两小时数据平均值对缺失数据进行插值填补，以 2#催化汽油进装置流量的缺失数据为例，图 4.3 是插值前后的对比图，填补完成后数据整体呈平稳趋势，保持在 40-60 之间，无异常值，插值效果良好，

数据值可以采用。

表 4.3 插值前后数据对比

插值前	插值后
53.86215	53.86215
46.35702	46.35702
44.51688	44.51688
0^*	52.38405
0^*	52.38405
0^*	52.38405
0^*	52.38405
48.84124	48.84124
51.11994	51.11994
51.83154	51.83154
46.70276	46.70276

注：灰底上标*的为缺失值

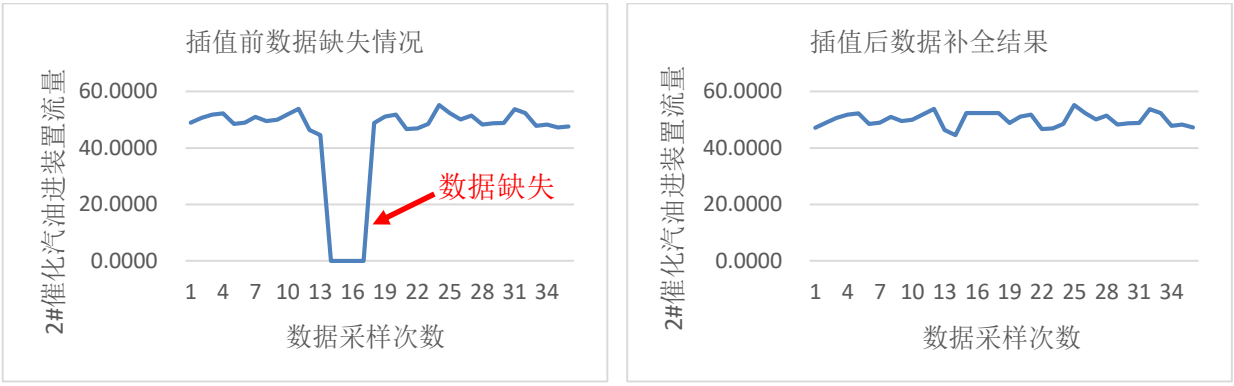

图 4.3 插值前后对比

对于原始数据的异常值检测依据的是拉伊达准则（3σ 准则），其适用于数据量大且呈现正态或近似正态分布的样本。首先假设一组只含有随机误差的检测数据，对其进行计算处理得到标准偏差σ，按概率确定一个区间，认为超过这个区间的误差，就不属于随机误差而是粗大误差，含有该误差的数据应予以剔除。其数学模型如下：

设对被测量变量进行等精度测量，得到$x_1,x_2.\cdots,x_n$，算出其算术平均值$\bar{x}$即剩余误差$v_i=x_i-\bar{x}(i=1,2,\cdots,n)$，按贝塞尔公式算出标准误差$\sigma$，若某个测量值$x_b$的剩余误差$v_b(1\leqslant b\leqslant n)$，满足$|v_b|=|x_b-x|>3\sigma$，则认为$x_b$是含有粗大误差值的坏值，应予剔除，其中贝塞尔公式为

$$\sigma=\left[\frac{1}{n-1}\sum_{i=1}^{n}v_i^2\right]^{1/2}=\left\{\left[\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n\right]/(n-1)\right\}^{1/2} \tag{4.1}$$

4.2.2 样本数据预处理

随后我们对样本数据进行预处理：

步骤 1：覆盖：将原始数据处理得到的两条原始数据最终结果插入到样本数据中；

步骤 2：最大最小限幅筛选法：根据操作变量信息提供的范围，对样本数据进行筛选，删除超出工艺和操作范围的样本；

步骤 3：删除缺值：删除残缺数据较多而无法补充的样本数据；

步骤 4：插值：对能够补全的部分数据缺失使用临近点的拟合值进行了代替。

使用原始数据 2 小时内操作变量的平均值作为辛烷值操作变量值插入样本数据中相应样本号的位置中。有 17 个位点仅仅含有部分时间位点数据，无法使用插值补充，直接将这些位点删除，用拟合插值填补了剩下的 176 项缺失数据。

由于篇幅所限，我们无法将全部数据以图表的形式在本文中展示出来，以上仅是例示性地说明我们在数据处理过程中所采用的方法及取得的效果，完整的数据处理结果将以附件形式提交。我们对原始数据的异常和缺失情况进行了统计，具体情况如下图所示：

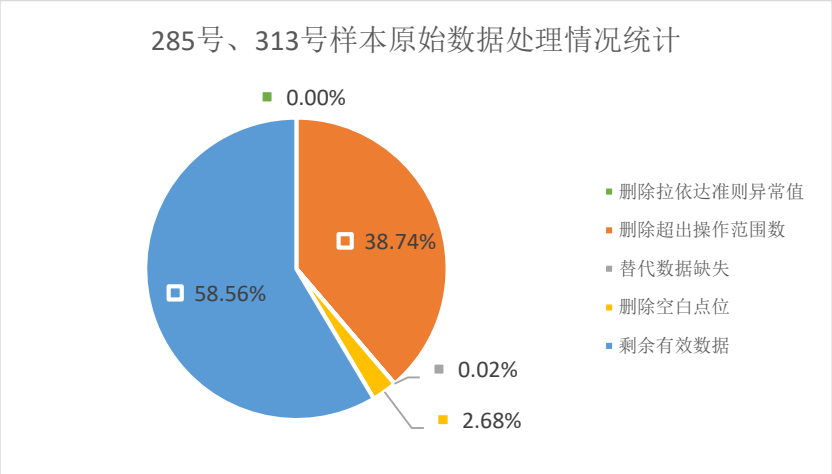

图 4.4 第 285 号、313 号样本原始数据处理情况

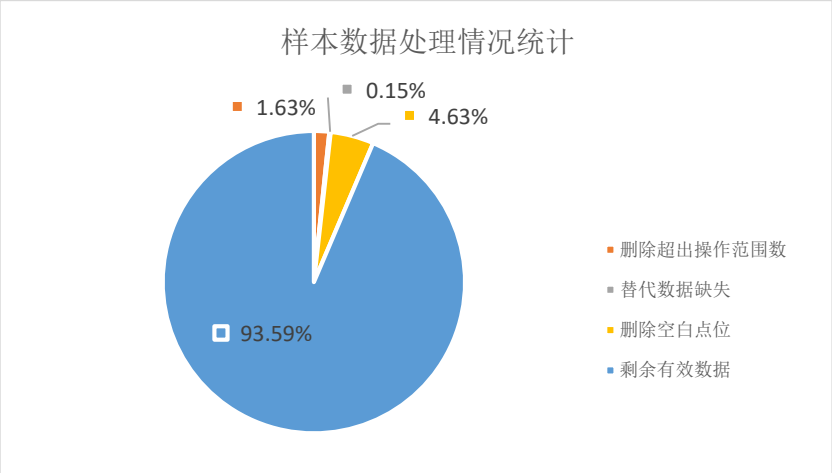

图 4.5 全部样本数据处理情况

4.3 数据进一步分析

完成对样本数据预处理后，为了进一步了解各个变量与辛烷值损失量之间的相互关系，为后续建模提供更有意义的指导信息，我们将辛烷值损失量与其他 342 个变量进行相关性分析。Spearman 相关系数是描述两组变量之间是否存在着相同或相反趋同性的一种指标，该检验不需要假定服从正态分布，在两组数据都没有重复观测值的情况下，Spearman 等级相关系数的公式为

$$r_{sp}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{4.2}$$

式中d_i表示两组数据的等级之差；n为样本量。

相关系数的大小代表着相关性的强弱：当$r_{sp}\geq 0.3$时，一般可以认为两者具有相关性。辛烷值损失量与其他 342 个变量之间的相关性如下表所示：

表 4.4 与辛烷值损失量存在相关性的前 20 个变量

相关性Top20			
点位号	中文名称	相关性系数	系数绝对值
S-ZORB.TE_5007.DACA	E-203壳程出口管温度	-0.347265515	0.347265515
S-ZORB.TE_1504.DACA	E-106管程入口管温度	-0.339832671	0.339832671
S-ZORB.AT-0009.DACA.PV	S_ZORB_AT-0009	-0.332403286	0.332403286
S-ZORB.AT-0008.DACA.PV	S_ZORB_AT-0008	0.324609841	0.324609841
S-ZORB.TE_5201.DACA	A-202A/B出口总管温度	0.324407134	0.324407134
S-ZORB.PDT_2104.PV	反应器顶底压差	0.324180067	0.324180067
S-ZORB.LC_1203.PIDA.PV	D-121含硫污水液位	0.322447069	0.322447069
S-ZORB.LC_1201.PV	D104液面	-0.321287053	0.321287053
S-ZORB.FT_9403.PV	氮气进装置流量	0.320655681	0.320655681
S-ZORB.TE_1601.PV	加热炉进口温度	-0.317681656	0.317681656
S-ZORB.PDT_1004.DACA	ME-104出入口	-0.315423918	0.315423918
S-ZORB.TE_1101.DACA	E-101壳程出口总管温度	0.31324996	0.31324996
S-ZORB.TE_6008.DACA.PV	预热器出口空气温度	0.312585803	0.312585803
	硫含量，μg/g	-0.311963175	0.311963175
S-ZORB.TE_5008.DACA	E-205壳程出口管温度	0.310423992	0.310423992
S-ZORB.FT_1204.DACA.PV	D-121含硫污水排量	0.308685846	0.308685846
S-ZORB.TE_7506B.DACA	K-103B进气温度	-0.308268437	0.308268437
S-ZORB.PT_2607.DACA	R-102底排放滑阀后氮气线压力	-0.303711824	0.303711824
S-ZORB.TE_1604.DACA	F-101出口支管#3温度	0.302927242	0.302927242
S-ZORB.FT_2433.DACA	D-106压力仪表管嘴反吹气流量	0.302322694	0.302322694

表 4.5 与辛烷值损失量存在正相关的前 10 个变量

正相关Top10			
点位号	中文名称	相关性系数	系数绝对值
S-ZORB.AT-0008.DACA.PV	S_ZORB_AT-0008	0.324609841	0.324609841
S-ZORB.TE_5201.DACA	A-202A/B出口总管温度	0.324407134	0.324407134
S-ZORB.PDT_2104.PV	反应器顶底压差	0.324180067	0.324180067
S-ZORB.LC_1203.PIDA.PV	D-121含硫污水液位	0.322447069	0.322447069
S-ZORB.FT_9403.PV	氮气进装置流量	0.320655681	0.320655681
S-ZORB.TE_1101.DACA	E-101壳程出口总管温度	0.31324996	0.31324996
S-ZORB.TE_6008.DACA.PV	预热器出口空气温度	0.312585803	0.312585803
S-ZORB.TE_5008.DACA	E-205壳程出口管温度	0.310423992	0.310423992
S-ZORB.FT_1204.DACA.PV	D-121含硫污水排量	0.308685846	0.308685846
S-ZORB.TE_1604.DACA	F-101出口支管#3温度	0.302927242	0.302927242

表 4.6 与辛烷值损失量存在负相关的前 10 个变量

负相关Top10			
点位号	中文名称	相关性系数	系数绝对值
S-ZORB.TE_5007.DACA	E-203壳程出口管温度	-0.347265515	0.347265515
S-ZORB.TE_1504.DACA	E-106管程入口管温度	-0.339832671	0.339832671
S-ZORB.AT-0009.DACA.PV	S_ZORB_AT-0009	-0.332403286	0.332403286
S-ZORB.LC_1201.PV	D104液面	-0.321287053	0.321287053
S-ZORB.TE_1601.PV	加热炉进口温度	-0.317681656	0.317681656
S-ZORB.PDT_1004.DACA	ME-104出入口	-0.315423918	0.315423918
	硫含量，μg/g	-0.311963175	0.311963175
S-ZORB.TE_7506B.DACA	K-103B进气温度	-0.308268437	0.308268437
S-ZORB.PT_2607.DACA	R-102底排放滑阀后氮气线压力	-0.303711824	0.303711824
S-ZORB.PT_7510.DACA	K-103A排气压力	-0.302111337	0.302111337

与辛烷值损失量相关性较强的前 20 个变量中，有 19 项属于操作变量，可见针对操作变量的优化对于调节辛烷值损失量具有重要意义；仅有的 1 项非操作变量为硫含量，与辛烷值损失量呈负相关，相关性系数为-0.312，这一关系也与本题的研究背景相符，硫含量与辛烷值损耗曲线如图 4.6 所示：

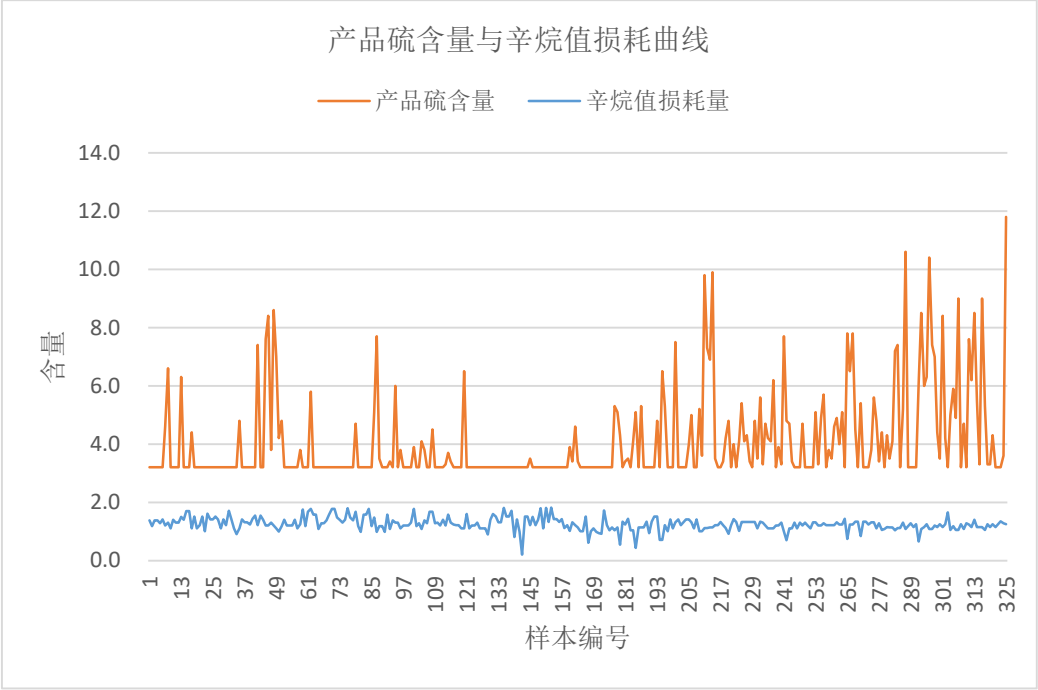

图 4.6 产品硫含量与辛烷值损耗曲线

5．建立辛烷值损失预测模型

5.1 问题分析

通过数据处理已经排除了数据中可能影响建模效果的因素，接下来要根据这些数据建立辛烷值损失预测模型。建立辛烷值损失预测模型是本题的核心，也是做好后续操作变量优化的基础。

首先要对数据变量进行降维，如果不对数据进行降维，直接将所有数据代入建立辛烷值损失预测模型，不仅对算法设计和模型构建造成困难，还可能会导致“维数灾难”，难以得出有意义的结果。数据降维是通过映射将样本从高维空间映射到低维空间，从而获得高维数据有意义的低维表示的过程。我们采用因子分析法，借助少数几个公共因子描述许多指标或因素之间的联系，将相关比较密切的几个变量归在同一类中，每一类变量就成为一个独立的公共因子，能够以较少的几个公共因子作为主要变量反映原数据的大部分信息，每个因子具有明确的解释性，能够最大程度地保留专业分析的作用。在此基础上采用多元回归的方法构建辛烷值损失预测模型，对辛烷值损失量进行预测。最后对模型进行误差分析和精度评估。具体解题流程图如下：

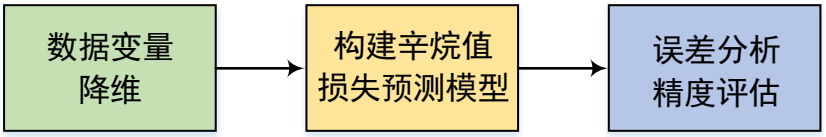

图 5.1 损失模型建模流程

5.2 模型建立

5.2.1 因子分析模型

与主成分分析法生成不同的是，因子分析法生成的变量具有实际的物理意义，有助于我们对模型后续操作变量的优化。本文在建立辛烷值损失预测模型过程中使用的因子分析数学模型为

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \cdots\cdots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases} \tag{5.1}$$

式中：$X=(x_1,x_2,\cdots x_p)$是p个原始数据变量；$F=(f_1,f_2,\cdots f_p)$是公共因子，即选择的主要变量；$A(a_{ij})$是公共因子F的系数，称为因子载荷矩阵，其中a_{ij}是因子载荷，即第i个原有变量在第j个因子上的载荷；$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots\varepsilon_p)$为$X$的特殊因子，是不能被公共因子包含的部分。

公共因子反映各原始变量的相关关系，使用公共因子代表原始变量，需要使用原始变量的观测值来计算各个公共因子的得分，其数学模型为

$$\begin{cases} f_1 = b_{11}x_1 + b_{12}x_2 + \cdots + b_{1p}x_p \\ f_2 = b_{21}x_1 + b_{22}x_2 + \cdots + b_{2p}x_p \\ \cdots\cdots \\ f_m = b_{m1}x_1 + b_{m2}x_2 + \cdots + b_{mp}x_p \end{cases} \tag{5.2}$$

式中：x_p 为标准化后的数据；b_{ji} 为第 i 个原始变量对第 j 个因子的得分。

5.2.2 多元线性回归模型

通过因子分析可以证明所选取的主要变量对辛烷值损失有很大的影响，采用多元线性回归模型进行预测。多元线性回归描述的是一个变量受到多个不同变量的影响的模型，建立多元线性回归模型时，为了保证回归模型具有优良的解释能力和预测效果，应首先注意自变量选择，其准则是：

1) 自变量对因变量必须具有显著的影响，并呈密切的线性相关；
2) 自变量与因变量之间的线性相关必须是真实的，而不是形式上的；
3) 自变量之间应具有一定的互斥性，即自变量之间的相关程度不应高于自变量与因变量之间的相关程度；
4) 自变量应具有完整的统计数据，其预测值容易确定。

结合本文自变量与因变量的关系，以 m 个自变量 X_{ij}，因变量 Y_i，建立多元线性回归方程

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_m X_{im} + u_i \\ &= \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + u_i \end{aligned} \tag{5.3}$$

式中：m 为解释变量的数目；β_j 为回归参数矩阵；u_i 为随机扰动。

5.3 模型的求解

5.3.1 因子分析模型求解

本文采用因子分析法对 367 个变量归类分析，提取公共因子，再以每个公共因子的方差贡献率作为权重与该公共因子的得分乘数之和构造得分函数。因子分析法的求解步骤如下：

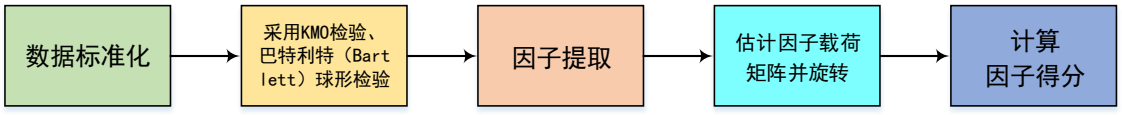

图 5.2 因子分解法步骤

步骤 1：数据标准化，将数据进行无量纲化处理，消除单位的影响，

数据标准化处理方法如下：

$$X_p = \frac{x_i - \bar{x}}{\sigma/\sqrt{n}} \tag{5.4}$$

式中：X_p 为标准化处理后的数据；x_i 为原始数据；$\bar{x}$ 为原始数据均值；σ 为原始变量数据的标准差；

步骤 2：确定待分析的原始变量是否适合进行因子分析，采用 KMO 检测和巴特利特（Bartlett）球形检测；

步骤 3：因子提取，确定所需公共因子的个数；

步骤 4：估计因子载荷矩阵，并将其旋转，便于公共因子的解释和命名；

步骤 5：计算每一个样本因子得分。

1) KMO 检验与巴特利特（Bartlett）球形检验

本文采用 KMO 检验和巴特利特（Bartlett）球形检验。当 KMO>0.5,巴特利特球形检验的相伴概率值小于 0.05 时，适合做因子分析。运用 SPSS 数据分析软件进行 KMO 和巴特利特检验，其结果如下：

表 5.1 KMO 和巴特利特检验

检验方法	结果
KMO 取样适切性量数	0.774
巴特利特球形度检验	近似卡方：20039.552 自由度：946.000 显著性：0.000

如表 5.1 所示，KMO 值为 0.774 大于 0.5，巴特利特球形检验的相伴概率值近似为 0.000，在 5%的显著性水平下拒绝原假设，认为原始变量间存在关联性，因此原有变量适合做因子分析。

2) 因子解释方差

建立 343 个原始变量的相关系数矩阵，求取特征值和特征向量，以碎石图的形式表示特征值，统计 343 个因子对应的特征值、方差百分比与累计百分比，要求选取特征值大于 1 的因子，且因子的方差百分比越大，表明该因子的比重越大，对公共因子的影响越大。根据特征值及方差百分比的选取原则选取因子 1 至因子 40 的部分数据，统计如下：

表 5.2 解释方差总和

因子	特征值	方差百分比	累计百分比	因子	特征值	方差百分比	累计百分比
1	109.497	31.923	31.923	21	2.344	0.684	84.363
2	38.524	11.232	43.155	22	2.216	0.646	85.009
3	23.198	6.763	49.918	23	1.906	0.556	85.565
4	19.608	5.717	55.635	24	1.845	0.538	86.103
5	14.133	4.121	59.756	25	1.761	0.513	86.616
6	11.711	3.414	63.170	26	1.628	0.475	87.091
7	9.658	2.816	65.986	27	1.516	0.442	87.533
8	8.260	2.408	68.394	28	1.443	0.421	87.953
9	7.245	2.112	70.506	29	1.365	0.398	88.351
10	6.545	1.908	72.414	30	1.309	0.382	88.733
11	6.027	1.757	74.171	31	1.301	0.379	89.112
12	5.049	1.472	75.643	32	1.241	0.362	89.474
13	4.820	1.405	77.049	33	1.227	0.358	89.832
14	4.037	1.177	78.226	34	1.128	0.329	90.161
15	3.971	1.158	79.383	35	1.096	0.320	90.480
16	3.502	1.021	80.405	36	1.063	0.310	90.790
17	3.103	0.905	81.309	37	1.041	0.303	91.094
18	2.867	0.836	82.145	38	0.980	0.286	91.379
19	2.756	0.804	82.949	39	0.952	0.278	91.657
20	2.506	0.731	83.680	40	0.912	0.266	91.923

表 5.2 显示，共有 37 个因子的特征值大于 1，基于过程中内定取特征值大于 1 的原则，初步筛选 1 至 37 个因子，方差积累量为 91.094%，表明前 37 个因子的信息可以代表原始数据。鉴于问题二要求选取主要变量的个数小于 30，通过对累计百分比的分析，发现第 22 个因子的方差累计量达到 85.009%，因此还需要对第 22 个因子到第 27 个因子的特征值以及方差累计量的变化趋势进行分析，确定最终选取的因子数量，如图 5.3 所示：

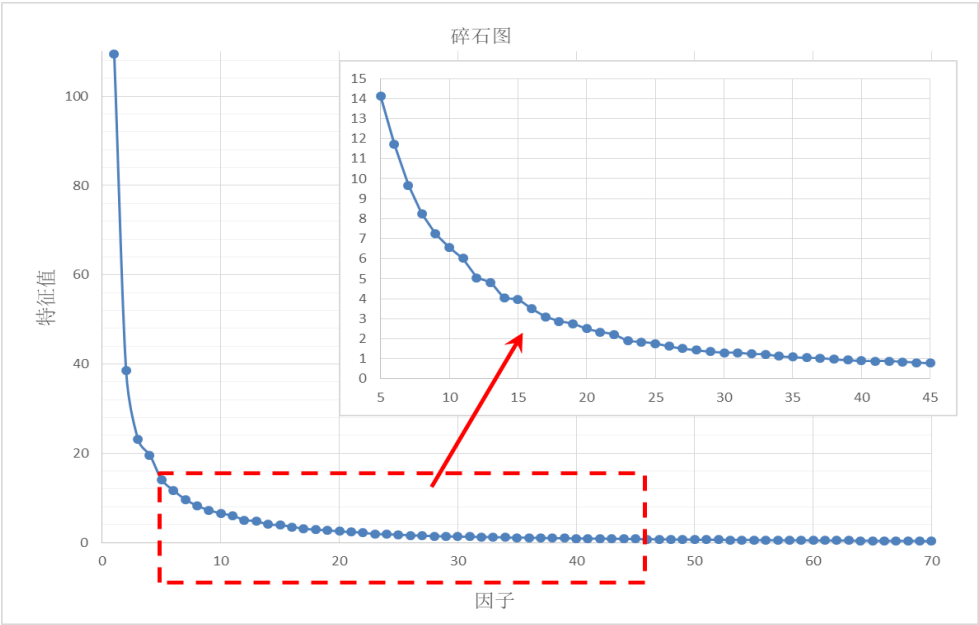

图 5.3 碎石图

由碎石图分析可知，因子 1 至因子 20 特征值变化剧烈，特征值由 109.497 迅速变化至 2.506，因子 20 至因子 40 特征值变化趋于平缓，由 2.506 缓慢降至 0.912，对因子 5 到因子 40 碎石图进行放大分析，观察特征值的变化趋势，结合表 5.1 的特征值及方差的百分比，选择 1 至 26 号因子作为主要变量，将 348 个原始变量降维保留至 26 个主要变量，即产生了 26 个公共因子。

3) 因子载荷矩阵

选取了合适的公共因子个数之后，需要根据变量之间的相关系数矩阵估计因子载荷矩阵。采用主成分分析法对因子载荷矩阵进行估计，并将得到的因子载荷矩阵旋转，便于了解每个公共因子的意义，采用最大方差法将因子载荷矩阵旋转，针对旋转后的因子载荷矩阵进行分析。表 5.3 是部分旋转因子载荷矩阵：

表 5.3 旋转因子载荷矩阵

	因 子							
	1	2	3	4	…	24	25	26
K-103A 排气压力	0.886	0.288	0.212	0.146	…	-0.020	0.004	-0.004
K-103A 排气温度	-0.885	-0.295	-0.206	-0.146	…	0.020	-0.004	0.004
D-123 压力	-0.480	-0.122	-0.330	-0.206	…	0.144	-0.096	-0.018
E-206 壳程出口管温度	-0.566	-0.095	-0.004	0.022	…	0.221	0.211	0.083
反吹氢气压力	0.984	0.028	0.029	0.047	…	-0.014	-0.004	-0.009
ME-103 进出口差压	0.882	0.116	-0.161	-0.142	…	-0.009	0.020	-0.019
过滤器 ME-101 出口温度	-0.230	-0.345	-0.490	0.014	…	-0.120	-0.019	0.023
…	…	…	…	…	…	…	…	…
D-102 温度	0.191	0.331	0.428	-0.154	…	-0.071	0.015	-0.042
原料进装置流量累计	0.936	-0.288	-0.119	-0.014	…	0.011	0.011	0.019
燃料气进装置压力	-0.162	-0.125	-0.117	-0.05	…	0.037	-0.04	0.008
反吹氢气温度	0.366	-0.168	0.183	0.051	…	0.013	-0.025	0.039
再生器下部温度	-0.135	0.011	0.051	0.000	…	-0.004	0.010	0.012
再生风流量	0.481	0.389	-0.093	0.078	…	0.080	0.032	0.022

表 5.3 共包含 26 个公共因子，343 个原始变量，分析每个公共因子对应的旋转后的因子系数，除去只包含一个原始指标的公共因子，公共因子 1 包含的原始指标个数最多，其包含原始指标的因子系数在 0.992 至 0.138 变化，有很明显地两极分化，也存在因子系数变化不大的公共因子，比如公共因子 7，其所包含的四个原始指标的因子系数最大为 0.838，最小为 0.447。

前 9 个公共因子对应的原始指标较多，其中公共因子 1 包含 96 个原始指标，占比最大，公共因子 24、公共因子 26 各对应三个原始指标，公共因子 18、公共因子 23 和公共因子 25 都只对应 1 个原始指标。下表是 26 个公共因子对应的具体原始指标及主要因素，主要因素指向每个公共因子的意义，解释公共因子对应的主要原始指标。

4) 公共因子得分

由于选取的公共因子和原始变量之间具有相关关系，公共因子反映了原始变量大部分信息，标准化后的原始变量用来表征 26 个公共因子得分，评价公共因子选取的合理性。下表选取了部分原始变量以及对应的 1 至 26 号公共因子得分系数矩阵：

表 5.4 因子得分系数矩阵

	因 子							
	1	2	3	4	…	24	25	26
硫含量,μg/g	-0.001	-0.014	0.008	-0.003	…	-0.005	0.028	-0.013
辛烷值 RON	0.001	-0.011	-0.015	-0.005	…	-0.024	-0.013	-0.070
饱和烃,v%（烷烃+环烷烃）	-0.005	0.004	0.000	0.001	…	-0.012	0.014	0.037
烯烃,v%	0.008	-0.012	-0.002	0.002	…	0.027	0.008	-0.014
芳烃,v%	-0.007	0.021	0.006	-0.008	…	-0.043	-0.056	-0.055
溴值,gBr/100g	0.003	-0.012	-0.003	-0.012	…	-0.013	-0.044	0.057
密度(20℃),kg/m³	-0.003	0.017	0.008	-0.001	…	-0.023	-0.061	-0.138
…	…	…	…	…	…	…	…	…
汽油产品去气分累积流量	0.014	-0.005	0.000	0.001	…	-0.005	0.001	-0.007
8.0MPa 氢气至循环氢压缩机入口	0.000	-0.011	-0.006	0.097	…	-0.026	0.029	-0.005
8.0MPa 氢气至循环氢压缩机入口	0.001	-0.028	0.000	0.006	…	0.018	0.007	0.003
8.0MPa 氢气至反吹氢压缩机出口	0.000	0.014	0.009	-0.098	…	0.011	-0.022	0.013
8.0MPa 氢气至反吹氢压缩机出口	0.000	-0.028	0.000	0.005	…	0.009	0.003	-0.008
D101 原料缓冲罐压力	-0.011	0.001	-0.001	-0.003	…	0.007	0.002	-0.013

将表 5.5 的系数带入（5.3）中可以得到对应的 26 个公共因子的函数，同时根据各个公共因子的方差百分比计算公共因子的综合得分：

$$F=(31.923f_1+11.232f_2+\cdots+0.475f_{26})/87.091 \qquad (5.5)$$

5) 公共因子的解释

下表是 26 个公共因子对应的具体原始变量及主要因素，主要因素指向每个公共因子的意义，解释公共因子对应的主要原始变量。

表 5.5 公共因子对应主要变量

因子	主要相关变量	主要因素	因子	主要相关变量	主要因素
1	K-101A 进气温度 K-101A 进气压力 … K-101A 排气压力 K-103A 排气压力	与反应器气体有关	14	稳定塔顶回流流量 稳定塔下部温度 … 蒸汽进装置流量 E-205 壳程出口管温度	与稳定塔有关
2	R102 再生器提升氮气流量 再生器顶底差压 … P-105A/B 出口总管流量 非净化风干燥后露点温度	与再生器有关	15	再生器温度 再生器下部温度（2605） … 再生器下部温度（2606） R-102 #1 通风挡板温度	与再生器有关
…	…	…	…	…	…

12	再生风流量部差压 再生器顶烟气温度 ... 再生器顶部/再生器接收器差压	与再生器有关	25	0.3MPa 凝结水出装置流量	与物料消耗有关
13	加热炉循环氢出口温度 加热炉炉膛压力 ... F-101 辐射室底部压力	与加热炉有关	26	循环氢至闭锁料斗料腿流量 闭锁料斗 H2 过滤器出口气流量 P-101A 入口过滤器差压	与闭锁料斗有关

5.3.2 多元回归模型求解及分析

1) 多元线性回归模型求解

我们将 325 组样本数据的前 300 组数据用于拟合回归系数，剩余的 25 组数据用于验证回归模型，得到的回归参数矩阵如下：

表 5.6 回归参数矩阵

回归系数β_i	第i项	回归系数β_i	第i项
1.26799	常数项	-0.00213	14
-0.04869	1	-0.00475	15
-0.04949	2	0.04170	16
-0.01827	3	-0.03140	17
-0.00964	4	-0.03129	18
-0.00117	5	0.00928	19
-0.00549	6	-0.01023	20
-0.00588	7	0.00946	21
-0.01374	8	-0.00364	22
0.02822	9	0.00492	23
0.02532	10	0.00032	24
0.02480	11	-0.00402	25
-0.01052	12	0.01135	26
-0.04664	13		

本模型的自变量是公共因子，因变量是辛烷值的损失值，共 26 个自变量，325 个因变量，根据回归参数矩阵，得到下面的多元参数方程：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{325} \end{bmatrix} = \begin{bmatrix} 1 & x_{1\,1} & \cdots & x_{1\,26} \\ 1 & x_{2\,1} & \cdots & x_{2\,26} \\ & & \vdots & \\ 1 & x_{325\,1} & \cdots & x_{325\,26} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{26} \end{bmatrix} \tag{5.6}$$

2) 模型分析验证

对辛烷值损失预测模型采用多元线性分析模型进行预测，预测目标是辛烷值的损失值，因子分析确定的 26 个公共因子作为辛烷值损失值的影响因素。通过对 325 条样本数据进行多元线性回归，其中最后 25 条为未使用过的测试数据，得到的部分预测数据如下：

表 5.7 多元线性回归

真实值	预测值	残差e	相对误差E
1.380000	1.347661	-0.032340	0.023434
1.180000	1.270395	0.090395	0.076606
1.380000	1.294605	-0.085390	0.061880
1.380000	1.377396	-0.002600	0.001887
1.280000	1.415353	0.135353	0.105745
1.410000	1.373815	-0.036184	0.025663
1.200000	1.310997	0.110997	0.092498
1.300000	1.292877	-0.007119	0.005479
1.100000	1.317519	0.217519	0.197745
1.400000	1.361321	-0.038680	0.027628
…	…	…	…
1.150000	1.174148	0.024148	0.020998
1.150000	1.170011	0.020011	0.017401
1.150000	1.168493	0.018493	0.016081
1.050000	1.217333	0.167333	0.159364
1.250000	1.215627	-0.034372	0.027498
1.150000	1.273137	0.123137	0.107076
1.250000	1.125662	-0.124340	0.099471
1.150000	1.288894	0.138894	0.120777
1.250000	1.251492	0.001492	0.001194
1.350000	1.268086	-0.081910	0.060677
1.280000	1.240768	-0.039230	0.030650
1.250000	1.186551	-0.063450	0.050759

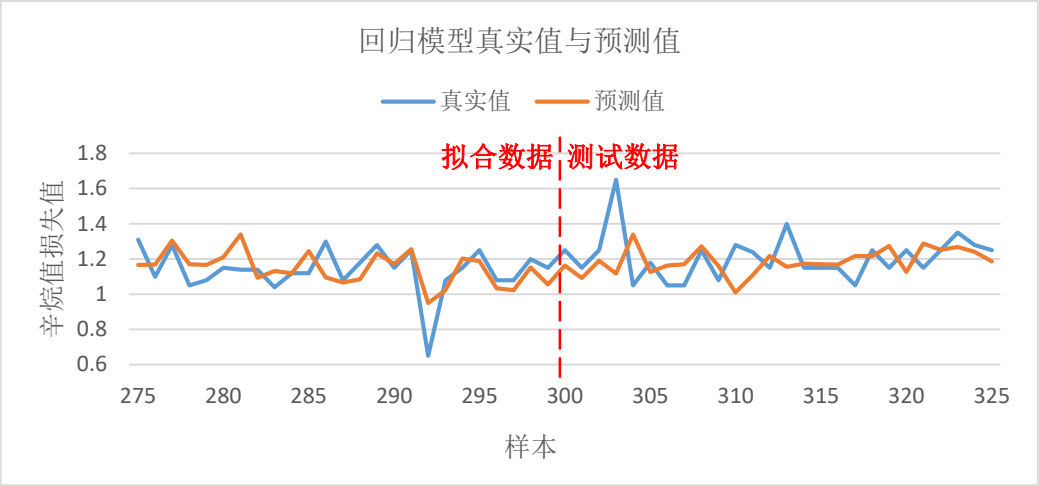

图 5.4 回归模型预测数据拟合曲线

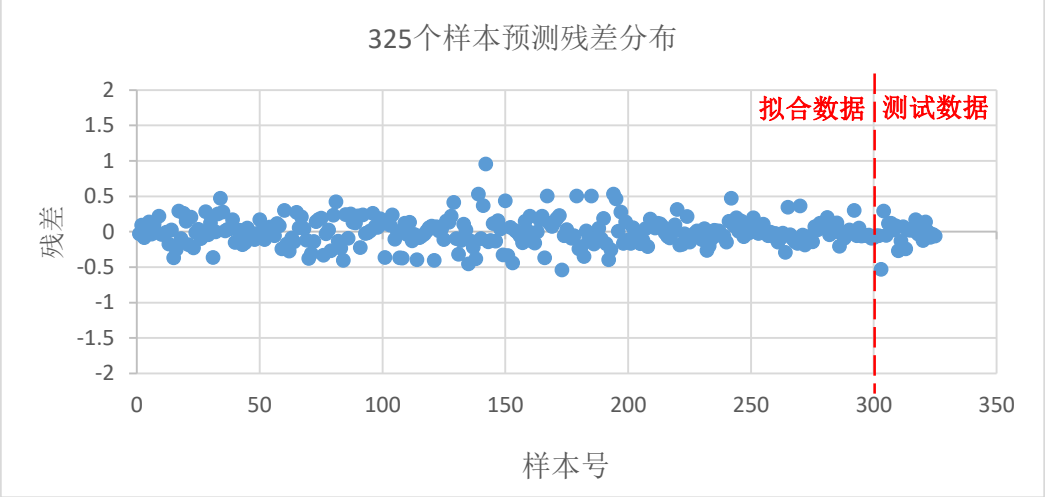

图 5.5 回归模型样本残差

残差e是真实值与预测值的差，残差图是以自变量为横坐标，残差e为纵坐标做出的散点图。通过残差所提供的信息，分析数据的可靠性、周期性或其他干扰，可用于分析关于误差项假定的合理性以及线性回归关系的假定的可能性。残差平方和反应除去自变量和因变量的线性关系外，所有其他因素影响的总和，残差平方和越小，说明预测越准确。通过对残差的分析，共有 321 个残差，其中大于 0.6 的有 3 个，大于 0.5 的有 7 个，大于 1 的只有一个，有 97.8462%的残差落在[-0.5 0.5]区间内，且此区间范围内所有散点随机分布，没有固定的趋势，不存在异方差的情况，说明模型选择存在合理性。

相对误差E定义为绝对误差ΔN与约定真值$N_{真}$的比值，即

$$E=\frac{\Delta N}{N_{真}}\times 100\% \tag{5.7}$$

对相对误差E的结果进行统计分析，发现相对误差大于 0.3 的有 19 个，相对误差在 0.3 以内的占比较重，为 94.1538%，可以认为多元线性回归模型具有一定合理性。

6．操作方案优化与可视化分析

6.1 问题分析

建立辛烷值损失预测模型的目标是求出优化辛烷值损失降幅大于 30%的样本，并得到相对应主要变量操作条件。首先在一定约束条件下求取辛烷值损失最小值，判断是否满足辛烷值损失降幅大于 30%的条件，最后找到满足优化目标对应样本的操作条件。该优化问题共有变量 343 个，样本数据 325 个，在优化过程中原料、待生吸附剂、再生吸附剂的性质将保持不变，即 325 个样本数据中 1 至 7 号、9 至 12 号变量数据保持不变。

优化的约束条件为：

1) 要求保证产品硫含量不大于 5μg/g，产品硫含量与原料、待生吸附剂、再生吸附剂以及 331 个操作变量可能存在一定相关关系，我们建立以产品中硫含量作为因变量、343 个变量作为自变量的逐步回归预测模型作为优化的约束条件之一；
2) 331 个操作变量数据需在各自的取值范围内优化，且优化过程只允许以每次步进 Δ 值的方式对操作变量进行调整。

以上述约束建立整数线性规划模型并通过遗传算法求解，问题求解流程图如下：

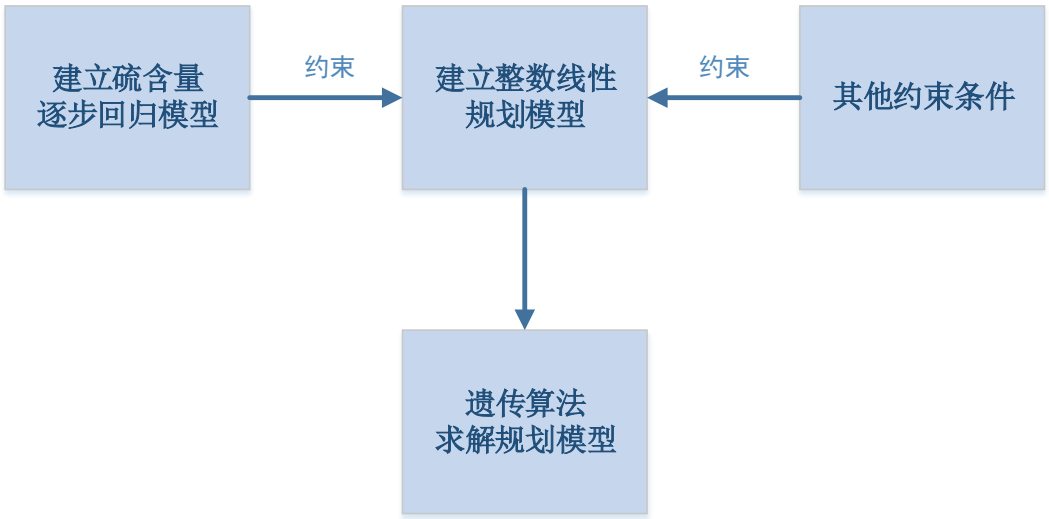

图 6.1 问题求解流程

6.2 建立优化模型

6.2.1 目标函数

首先求取辛烷值损失最小值，将因子得分系数矩阵带进多元回归模型中，得到如下模型：

$$\min\ R_l = c_0 + \sum_{i=1}^{p} c_i x_i = c_0 + \sum_{i=1}^{p}\sum_{j=1}^{m} \beta_j b_{ij} x_i \qquad (6.1)$$

式中，$c_i = \sum_{j=1}^{m} \beta_j b_{ij}$ 表示辛烷值损失量回归模型系数；R_l 表示辛烷值损失量；p 表示变量

个数；m表示多元回归公共因子的个数； x_i为第i个变量。在 343 个变量中，有 11 个涉及原料和产品性质的变量保持不变，分别是 7 个原料性质$(x_1 \sim x_7)$、2 个待生吸附性质$(x_9 \sim x_{10})$、2 个再生吸附性质$(x_{11} \sim x_{12})$，同时产品硫含量与其余变量之间存在一定相关关系，所以目标函数等价于

$$\min\ R_l = c_8 x_8 + \sum_{i=13}^{p} c_i x_i \tag{6.2}$$

式中，x_8表示产品硫含量；$x_i(i=13,\cdots,p)$表示 331 个操作变量。产品硫含量应满足约束条件$x_8 \leqslant 5$，本文采用逐步回归模型建立产品硫含量与其他变量之间的关系。

逐步回归的基本思想是逐个引入自变量，每次引入对因变量最显著的自变量，并对方程中的已存在变量逐个进行检验，把不显著的变量逐个从方程中剔除，最终的回归方程中包含对因变量具有显著影响的变量，又剔除了对因变量影响不显著的变量。本文中逐步回归的因变量是产品硫含量，逐步回归的基本步骤如下：

1) 求取全部自变量的偏回归平方大小，从大到小依次引入回归方程；
2) 对回归方程所含全部变量进行检验，剔除不显著因素，直到回归方程中所含的所有变量对因变量的影响都显著时，才考虑引入新的变量；
3) 在剩余未选因素中，选出对因变量作用最大者，检验其显著性：若存在显著性，则引入回归方程，否则不引入；
4) 最终没有显著性因素可以引入，也没有不显著因素需要剔除，得到回归方程。

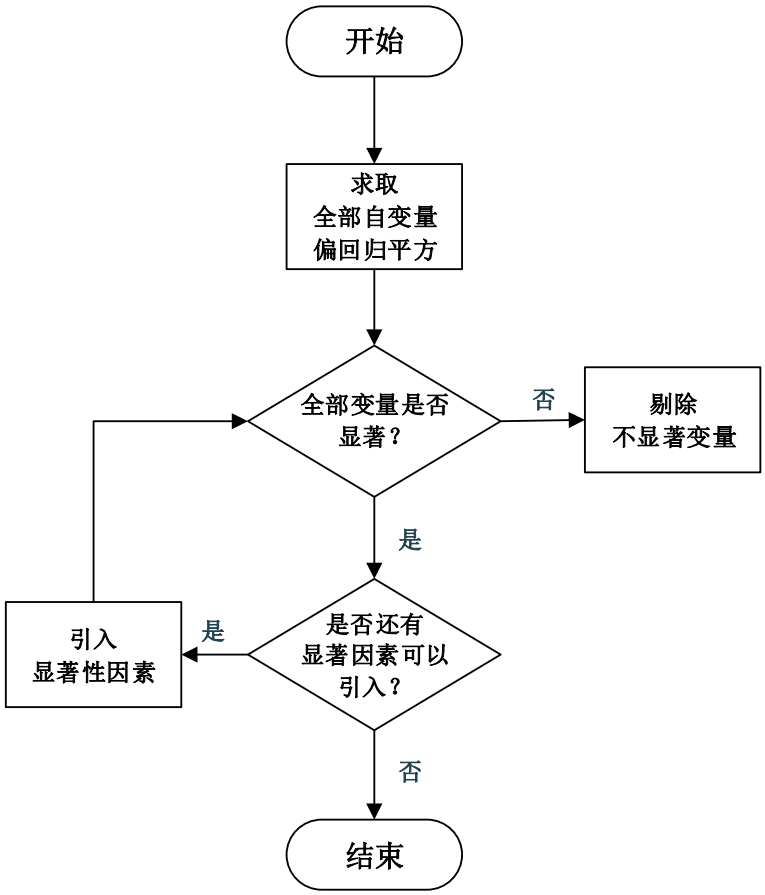

图 6.2 逐步回归流程

求解得到回归方程为

$$x_8 = m_0 + \sum_{k\epsilon A} m_k x_k \tag{6.3}$$

式中，x_8为预测的产品硫含量；m_o为回归方程的常数项；m_k为回归方程的系数。为了保持与目标函数变量的序号一致，上述回归方程的自变量集采用了不连续集合表示，$A=\{5,6,23,33,80,86,118,125,141,145,168,182,235,247,336\}$，即共有 15 个变量与产品硫含量有关，其中包含序号为5、6的变量是不变的原料性质，可划分为常数项，将回归方程的系数带入方程：

$$\begin{aligned} x_8 = {} & 39.0248 - 7.44286\times10^{-2}x_5 + 1.78557\times10^{-2}x_6 \\ & -8.2684\times10^{-3}x_{23} - 4.63709x_{33} + 4.62119\times10^{-2}x_{80} \\ & -1.92759\times10^{-2}x_{86} - 5.9552\times10^{-2}x_{118} + 1.22517x_{125} \\ & -7.9627\times10^{-3}x_{141} - 1.20777\times10^{-2}x_{145} - 5.15196\times10^{-2}x_{168} \\ & +4.5955\times10^{-3}x_{182} - 6.11329\times10^{-2}x_{235} + 3.03882\times10^{-2}x_{247} \\ & -9.05489e^{-5}x_{336} \end{aligned} \tag{6.4}$$

将产品硫含量的回归方程带入到目标函数中，去掉常数项后，目标函数可等价为

$$\min \quad c_8 \sum_{k\epsilon B} m_k x_k + \sum_{i=13}^{p} c_i x_i \tag{6.5}$$

式中，$B=\{23,33,80,86,118,125,141,145,168,182,235,247,336\}$；集合$B$为集合$A$的子集，即集合$A$去掉常数项5、6后的集合。

6.2.2 约束条件

题目给出了 2 个约束条件，分别是对产品硫含量和操作变量的约束。

1) 对产品硫含量的约束

目标函数的优化要求在产品硫含量不大于 5ug/g 的前提下，产品硫含量会受到其他变量的影响，根据逐步回归可知，产品硫含量受到 2 个原料变量、13 个操作变量影响，其约束条件表达式为

$$x_8 = \sum_{k\epsilon A} m_k x_k + m_0 \leqslant 5 \tag{6.6}$$

2) 对操作变量步进值的约束

根据操作变量信息，操作变量只能在各自的取值范围内变化，且为了防止避免实际操作过程中大幅度调整所带来的波动，每个操作变量都设有步进值Δ，即每个操作变量每次允许的调整幅度，优化的过程只允许以每次步进Δ值的方式对操作变量进行调整，其表达式为

$$x_{i\min} \leqslant x_i \leqslant x_{i\max},\ i=13,\cdots,\ p \tag{6.7}$$

$$x_i = x_{i,j} + n_i\Delta_i\ ,\ i=1,2,\cdots,p;\ j=13,\cdots,\ k \tag{6.8}$$

式中，x_8表示产品硫含量；$x_{i,j}$表示第j个样本的第i个变量；$x_{i\min}$表示第i个变量的最小取值范围；$x_{i\max}$表示第i个变量的最大取值范围；n_i表示第i个变量的步进次数；Δ_i表示第i个变量的步进值；k表示样本的个数。

6.2.3 得出模型

经过整理，得出优化模型：

$$\min \quad c_8 \sum_{k\epsilon B} m_k x_k + \sum_{i=13}^{p} c_i x_i$$

$$s.t. \begin{cases} m_0 + \sum_{k\epsilon A} m_k x_k \leqslant 5 \\ x_{i\min} \leqslant x_i \leqslant x_{i\max} \quad , i = 13, \cdots, p \\ x_i = x_{i,j} + n_i \Delta_i \end{cases} \tag{6.9}$$

分析该模型可知，共有 331 个连续型变量x_i，对应 331 个整数变量n_i，该模型为整数线性模型，约束条件有两种，共有 633 个不等式约束，331 个等式约束。

6.2.4 模型简化

为进一步简化模型，可将等式约束代入目标函数：

$$c_8 \sum_{k\epsilon A} m_k \left(x_{k,j} + n_k \Delta_k\right) + \sum_{i=13}^{p} c_i \left(x_{i,j} + n_i \Delta_i\right) \tag{6.10}$$

常数项不影响优化，将上式最小化，可等价为

$$\min \quad c_8 \sum_{k\epsilon A} n_k \Delta_k + \sum_{i=13}^{p} c_i n_i \Delta_i \tag{6.11}$$

代入约束条件可得

$$\begin{cases} m_0 + \sum_{k\epsilon A} m_k (x_{k,j} + n_k \Delta_k) \leqslant 5 \\ x_{i\min} - x_{i,j} \leqslant n_i \Delta_i \leqslant x_{i\max} - x_{i,j} \end{cases} \tag{6.12}$$

综上所述，模型可等价为

$$\min \quad c_8 \sum_{k\epsilon A} n_k \Delta_k + \sum_{i=13}^{p} c_i n_i \Delta_i$$

$$s.t. \begin{cases} m_0 + \sum_{k\epsilon A} m_k (x_{k,j} + n_k \Delta_k) \leqslant 5 \\ x_{i\min} - x_{i,j} \leqslant n_i \Delta_i \leqslant x_{i\max} - x_{i,j} \end{cases} , i = 13, \cdots, p \tag{6.13}$$

本模型为整数线性模型，简化后有 331 个整数变量n_i，662 个不等式约束。

6.3 遗传算法求解整数规划

我们采用遗传算法对整数线性规划模型进行求解，遗传算法作为一种搜索启发式算法，能够对可行解域的某个群体逐次演化出适应度高的近似解，并根据适应度的大小选择是否遗传，将所有个体全部筛选后，选择出满足条件且具有最大适应度的个体即为最优解。遗传算法的基本运算过程如下:

1) 编码：对实际问题进行编码处理；
2) 初始化：选择可行解域的某个群体作为问题的初始群体；
3) 适应度评价：计算所有个体的适应度，作为筛选依据；
4) 选择操作：根据个体的适应度判断选择交叉还说变异的遗传方式；
5) 依次进行交叉运算和变异运算；
6) 判断是否达到最优解：所有个体都运算过后，选择过程中具有最大适应度的个体作为最优解输出。

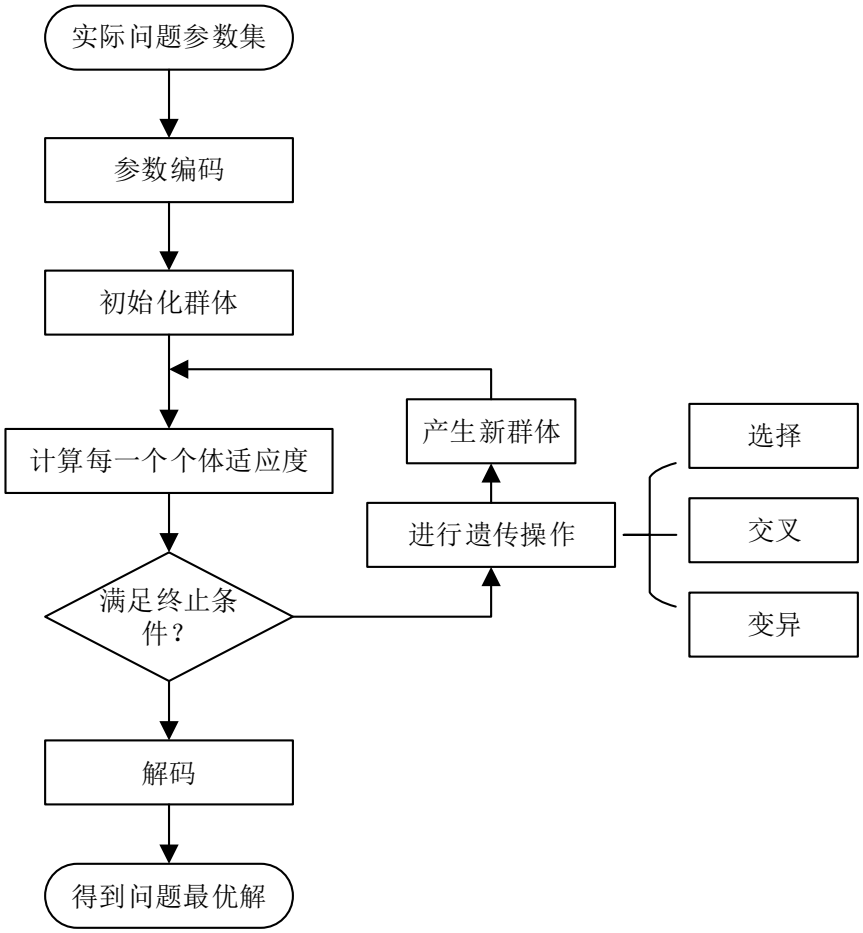

图 6.3　遗传算法流程框图

遗传算法伪代码如下：

Algorithm 1: The iterative algorithm of GA

Input: sample data, initial code
Output: Optimize result

```
1.  BEGIN
2.      I=0;
3.      Initialize P(I);
4.      Fitness P(I);
5.      While (not Terminate-Condition)
6.      {
7.          I ++;
8.          GA-Operation P(I);
9.          Fitness P(I);
10.     }
11. END
```

我们使用遗传算法对整数规划求解获得的 325 个样本的优化结果中，考虑到题目实际情况，我们将辛烷值损失量降低到 0.6 以下的 24 个数据标记为异常值，予以删除；在剩余的 301 个有效数据中，降幅达到 30%以上的样本数为 199 个，301 个样本的平均降幅为 44.46%，具体降幅分布情况如表 6.1 所示：

表 6.1 优化后辛烷值损失降幅分布情况

	数量	百分比
降幅 0-30%	102	31.38%
降幅 30-100%	178	54.78%
降幅 100%以上	21	6.46%
异常数据	24	7.38%
合计	325	

6.4 操作变量优化过程可视化分析

6.4.1 全体样本优化整体可视化分析

优化前后的辛烷值损失量变化对比如图 6.4 所示，在 301 个有效样本中，辛烷值损失量总体呈下降变化，平均下降值为 0.4，平均降幅为 44.46%；在辛烷值下降的同时，产物硫含量总体上升，平均硫含量为 4.04μg/g，均控制在 6μg/g 以下。

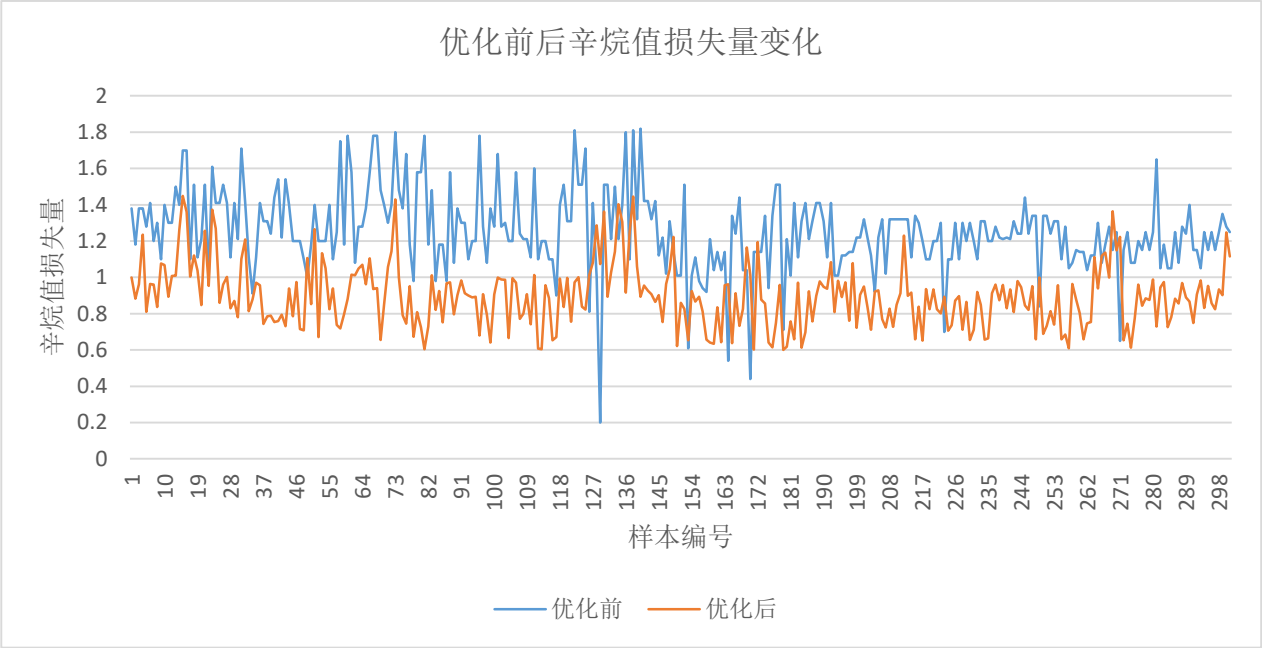

图 6.4 优化前后辛烷值损失量对比

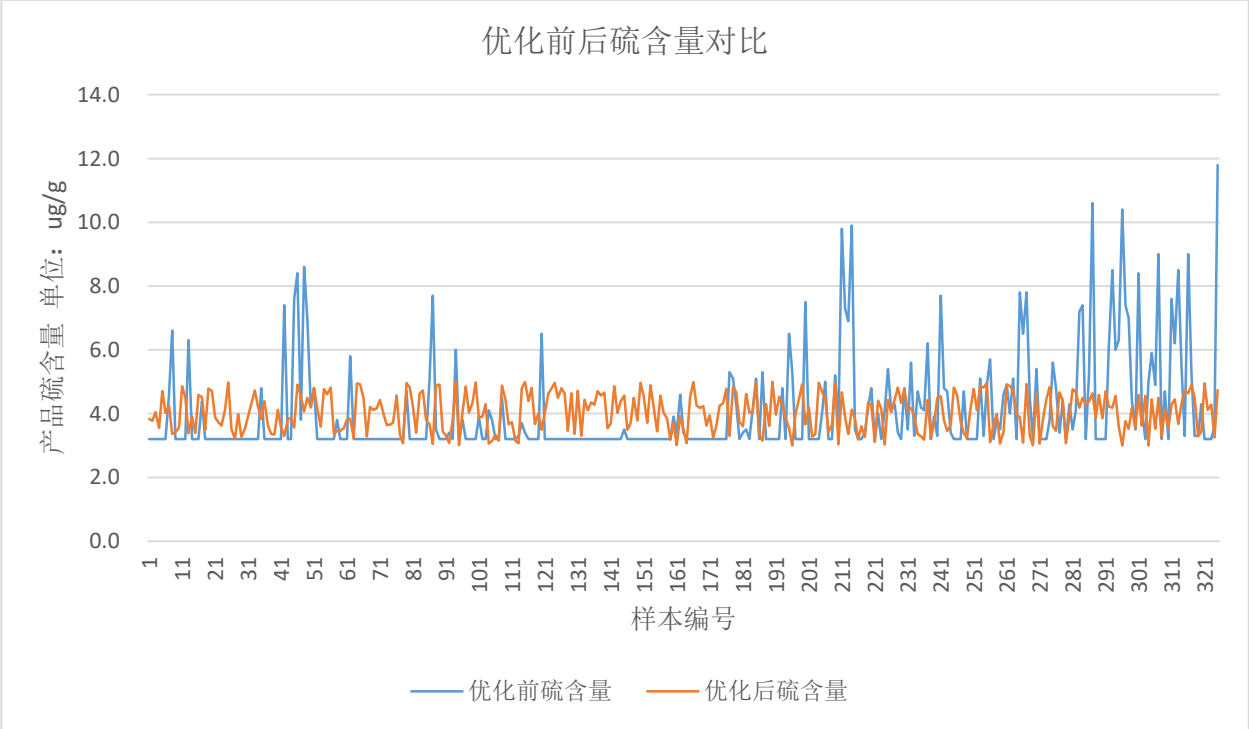

图 6.5 优化前后硫含量对比

6.4.2 133 号样本优化过程可视化分析

下图给出了主要操作变量优化调整过程中对应 133 号样本汽油辛烷值损失量和产品硫含量的收敛曲线，从图中可以看出，辛烷值损失量与产品硫含量的变化关系与总体基本一致。随着迭代次数的增加，两者同步收敛，其中辛烷值损失量在 0.89 时逐渐趋于平稳。

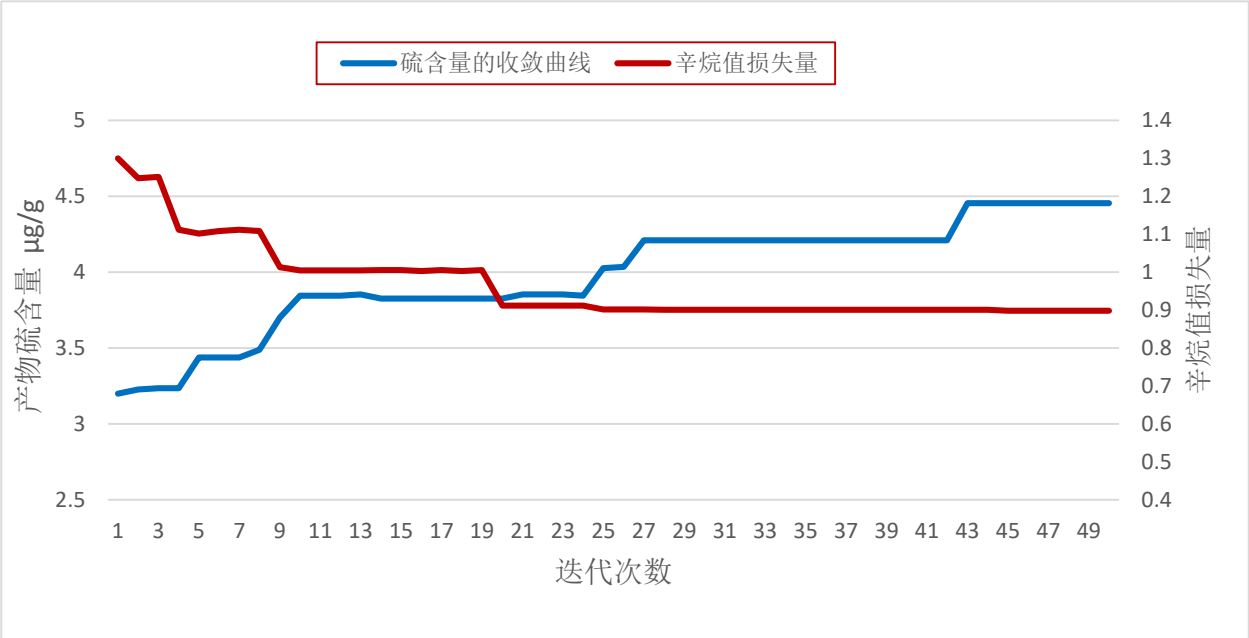

图 6.6 133 号样本辛烷值损失量和硫含量迭代变化曲线

7. 模型评价与改进

7.1 模型的优点

1. 问题一中我们对缺失的数据进行了删除、插补、上下限幅筛选和异常值检测等处理方法，一定程度上保证了数据的可靠性和完整性。

2. 问题二中采用了因子分析法对数据进行降维，生成的新因子具有代表性和独立性的同时，也保留了实际物理意义，具有更好的解释性，能在实际工程中进行专业分析。

3. 问题二中有效利用多元回归对模型进行拟合，具有适合大量数据、简单有效等特点，并划分了拟合数据集和测试数据集，通过独立的数据集对拟合方程进行验证，证明了模型的准确性和可行性。

4. 问题三中建立整数规划模型对辛烷值损失值进行寻优，使用逐步回归模型建立产品硫含量与其他自变量之间的关系并作为目标函数的约束条件之一，通过遗传算法对规划模型进行了求解。

7.2 模型的缺点

1. 由于汽油精制装置机制复杂、结构多样，模型只能在特定装置下对辛烷值进行有效预测，使用具有一定的局限性。

2. 由于时间和专业知识背景有限，建模仅仅将原料产物性质、操作工艺等纳入考虑，排除了化学反应物因素等影响，在实际工程实践过程中还需要对汽油精制装置工作机理做更加深入的了解，制定更加具体的方案。

7.3 模型的改进与推广

1. 未来可以使用核主成分分析进行非线性降维，通过神经网络模型、小波分析等方法进行曲线拟合。

2. 本文是对汽油精制过程中辛烷值损失进行预测和优化，建模思路和方法对于其他化学工艺制备过程的优化问题同样适用。

参考文献

[1] 左苏. 基于主成分回归模型的工程项目成本预测[D].扬州大学,2014.

[2] 毕达天,邱长波,张晗.数据降维技术研究现状及其进展[J].情报理论与实践,2013,36(02):125-128.

[3] 吴晓婷,闫德勤.数据降维方法分析与研究[J].计算机应用研究,2009,26(08):2832-2835.

[4] 杨于镭,张祥东,姜浩.基于 GA-BP 优化算法的 BP 网络及其在汽油调合辛烷值建模中应用[J].微计算机信息,2006(11):276-278.

[5] 孙自强,顾幸生,党晓恒,俞金寿.连续催化重整装置辛烷值软测量研究[J].系统仿真学报,2001(S1):171-172+175.

[6] 张祥东,钱锋. 汽油调合的辛烷值(RON)估算方法综述[A]. 中国自动化学会控制理论专业委员会.第二十四届中国控制会议论文集（上册）[C].中国自动化学会控制理论专业委员会:中国自动化学会控制理论专业委员会,2005:4.

[7] 王瑾,蒋书波.汽油辛烷值 NIR 数据处理与建模仿真[J].计算机与应用化学,2011,28(07):947-950.

附录

程序 1：3σ检测

```
clc
clear

%加载文件
filename1 = '附件三处理完成.xlsx';
xlRange_file1_285 = 'B4:MQ43';
xlRange_file1_313 = 'B45:MQ53';
A = xlsread(filename1,xlRange_file1_285);
B = xlsread(filename1,xlRange_file1_313);

%奇异值检测
CheckRes_file1_285 = isoutlier(A,'mean');
CheckRes_file1_313 = isoutlier(B,'mean');

find(CheckRes_file1_285);
find(CheckRes_file1_313);
```

程序 2：Spearman 相关性分析

```
%Spearman 斯皮尔曼相关性分析
clc;
clear;
close;

%加载文件
filename = '5 插值完成.xlsx';
xRange = 'C4:MI328';
xdata = xlsread(filename,xRange);

%X 为 n 行 P 列矩阵，表示有 P 个特征，每个特征有 n 个样本点
%coeff 为 P 行 P 列矩阵，coeff(i,j)表示第 i 个特征和第 j 个特征的相关系数
coeff = corr(xdata,'type','Spearman');
```

程序 3：因子分析

```
clc;clear;

%加载文件
A1=xlsread('附件一插值后.xlsx','Sheet1','C4:J324');
A2=xlsread('附件一插值后.xlsx','Sheet1','M4:MN324');
A=[A1,A2];

%标准化
[Astd,A_mean,A_std]=zscore(A);

%计算特征之间的协方差矩阵
r=cov(Astd);
[vec,val,con]=pcacov(r);
cum=cumsum(con);

%碎石图
x=1:348';
figure
plot(x,val,'r-','LineWidth',2);hold on

axis([-5,353,-5,120])
ylabel('特征值')
xlabel('主成分')
figure;
plot(x(7:45),val(7:45),'r-','LineWidth',2);hold on
plot(x(7:45),val(7:45),'rx-','LineWidth',1);hold on
plot(x(7:45),ones(39),'--','LineWidth',1);hold on
ylabel('特征值')
xlabel('主成分')

f1=repmat(sign(sum(vec)),size(vec,1),1);
vec=vec.*f1;%
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2;                                          %计算全部因子的载荷矩阵
num=26;                                             %num 为因子的个数
a1=a(:,1:num);                                      %提出 26 因子的载荷矩阵
tcha=diag(r-a1*a1');                                %因子的特殊方差
ccha=r-a1*a1'-diag(tcha);                           %求残差矩阵
[b,t]=rotatefactors(a(:,1:num),'method','varimax')  %对载荷矩阵进行旋转
coef=inv(r)*b;
score=B*coef;
```

程序 4：多元回归

```
clear ;clc;

%加载文件
B=xlsread('因子分析 0919 结果.xlsx','得分矩阵','D8:AC11');
X1=xlsread('样本数据处理后.xlsx','Sheet1','C4:J328');
X2=xlsread('样本数据处理后.xlsx','Sheet1','M4:MI328');
X=[X1,X2];

[Xstd,X_mean,X_std]=zscore(X);
F=Xstd*B;
y=xlsread('样本数据处理后.xlsx','Sheet1','L4:L328');
n=length(F); %n 表示样本个数
[b,bint,r,rint,s]=regress(y(1:300,:),[ones(300,1),F(1:300,:)]);
rcoplot(r,rint)    %残差及其执行区间作图

mask=find(rint(:,1).*rint(:,2)<0);
newy=y(mask);
newF=F(mask,:);
[a,aint,r,rint,s]=regress(newy,[ones(length(newF(:,1)),1),newF]);
rcoplot(r,rint)  %残差及其执行区间作图

%计算预测值
y_=[ones(325,1),F]*b;
figure;
plot(275:325,y(275:325));
hold on
plot(275:325,y_(275:325));
xlabel('样本')
ylabel('辛烷损失值')
legend('实际损失值','预测损失值')

%模型检验
epsilon=y_-y;            %残差
delta=abs(epsilon./y);  %相对误差
figure;
plot(301:325,epsilon(301:325),'o');hold on
plot(295:330,zeros(1,36),'--');
axis([295 330 -1 1])
xlabel('样本')
ylabel('残差')
```

程序 5：遗传算法求解整数规划主程序

```
clc;clear;
load Xmean;
load Xstd;
load c;
c8=c(8);
c13_343=c(13:end)';
c1_7_8_12=c([1:7,8:12])';
load delta
% delta=(delta-X_mean(13:end)')./X_std(13:end)';

load Xmin
load Xmax
% Xmin=(Xmin-X_mean(13:end)')./X_std(13:end)';
% Xmax=(Xmax-X_mean(13:end)')./X_std(13:end)';

xij=xlsread('数据.xlsx','Sheet2','B2:LT2')';
% xij=(xij-X_mean(13:end)')./X_std(13:end)';
xp=xlsread('数据.xlsx','Sheet3','B3:M3')';

m=zeros(343,1);
m([5 6 23 33 80 86 118 125 141 145 168 182 235 247 336])=[-0.0744286 0.0178557 -
0.0082684 -4.63709 0.0462119 -0.0192759 -0.059552 1.22517 -0.0079627 -0.0120777 -
0.0515196 0.0045955 -0.0611329 0.0303882 -9.05489e-5];
m0=39.0248;
L1=(Xmin-xij)./delta;
L2=(Xmax-xij)./delta;
Nmax=L2;
Nmin=L1;
Nmax=ceil(Nmax);
Nmin=floor(Nmin);
postion=find(L1>L2);
Nmax(postion)=L1(postion);
Nmin(postion)=L2(postion);
global Cmin;
varnum=331;      %变量个数
eps=1e-1;
popsize=20;      %群体大小，修改
Gene=50;         %迭代次数
pc=0.95;         %交叉概率
pm=0.05;         %变异概率
```

```
%计算每个变量编码所需要的的长度

for i=1:varnum
    L(i)=ceil(log2((Nmax(58)-Nmin(58))/eps));
end
L=real(L);
chromlength=sum(L);      %每个个体所需总位长
count=0;
spoint=cumsum([0 L]);
pop=round(rand(popsize,chromlength));
while 1
      tempn=round(rand(1,chromlength));                  %随机产生初始群体
      n=decodechrom(spoint,varnum,tempn,Nmax,Nmin);      %真实的十进制值
      n=round(n);
      ndelat=(n'.*delta);
      S=m0+m(5)*xp(5)+m(6)*xp(6)+sum(m(13:end).*(xij+ndelat));
      if S<=5&&S>=3
         count=count+1;
         pop(count,:)=tempn;
      end
       if count>=popsize
          break;
      end
end

for i=1:Gene %Gene 为迭代次数
    %将二进制转化为十进制
    real10=decodechrom(spoint,varnum,pop,Xmax,Xmin);         %真实的十进制值
    real10=round(real10);
    [objvalue]=calobjvalue(real10,c13_343,m,c8,delta);       %计算目标函数
    fitvalue=calfitvalue(objvalue);
    [newpop]=selection(pop,fitvalue,'roulette');             %选择
    [newpop]=crossover(newpop,pc,'singlepoint');             %交叉
    count=0;
    for j=1:2*popsize
        tempn=newpop(j,:);
        n=decodechrom(spoint,varnum,tempn,Xmax,Xmin);        %真实的十进制值
        n=round(n);
        ndelat=(n'.*delta);
        S=m0+m(5)*xp(5)+m(6)*xp(6)+sum(m(13:end).*(xij+ndelat));
        if S<=5&&S>=2
            count=count+1;
            newpop(count,:)=tempn;
```

```
        end
        if count>popsize
            newpop=newpop(1:20,:);
        else
            newpop=[newpop(1:count,:),count(1:popsize-count,:)];
        end
        [newpop]=mutation(newpop,pm,'binary');
    end
    [bestindividual,bestfit]=best(pop,fitvalue);
    bestindividual10(i,:)=decodechrom(spoint,varnum,bestindividual,Xmax,Xmin);
    y(i)=bestfit+Cmin;
    y_mean(i)=mean(fitvalue+Cmin);
    plot(i,y(i),'r.');
    hold on
    plot(i,y_mean(i),'b.');
    pop=newpop;
end
```

附件清单

	内容	文件名
附件 1	所有程序源代码	源代码.zip
附件 2	样本数据预处理结果	样本数据预处理结果.xlsx
附件 3	因子分析结果	因子分析结果.xlsx
附件 4	多元回归结果	多元回归结果.xlsx
附件 5	遗传算法求解结果	遗传算法求解结果.xlsx