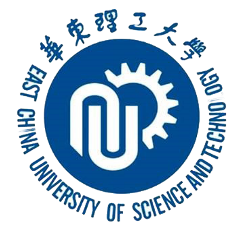
****  ** **

**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.张耹铭** |
| **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 （方法+对象+目的）

摘 要：

本文研究了XXX问题。主要创新点在于（创新点1解决了XX问题，取得了XX有益效果），（创新点2解决了XX问题，取得了XX有益效果）以及（创新点3解决了XX问题，取得了XX有益效果）。

问题一：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

问题二：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

问题三：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

最后，对于XXX模型的优缺点进行了评价，本文提出了以下几点（未来的改进措施？）：（1）（2）（3）。

关键字：关键词1；关键词2；关键词3；关键词4；关键词5

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[1]。

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，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
2. 整定和筛选附件三中的原始数据；
3. 预处理附件一中的样本数据；
4. 提取主要变量建立辛烷值损失预测模型
5. 选取合适的降维方法筛选建模主要变量；
6. 构造辛烷值损失预测模型并验证；
7. 优化主要操作变量方案并进行可视化展示
8. 根据预测模型优化主要变量操作方案；
9. 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. 模型假设，如果不够可以根据题目条件进行补充。
2. 模型假设，如果不够可以根据题目条件进行补充。
3. 模型假设，如果不够可以根据题目条件进行补充。
4. 模型假设，如果不够可以根据题目条件进行补充。
5. 模型假设，如果不够可以根据题目条件进行补充。
6. 模型假设，如果不够可以根据题目条件进行补充。

3. 符号系统

|  |  |  |
| --- | --- | --- |
| 符号 | 符号说明 | 单位 |
|  | 原始观测变量 | - |
|  | 原始观测变量的公共因子 | - |
|  | 因子载荷矩阵 | - |
|  | 原始观测变量的特殊因子 | - |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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（3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：

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，删除超出工艺和操作范围的样本；
2. 删除缺值：删除残缺数据较多而无法补充的原始数据；
3. 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；
4. 异常值检测：根据拉依达准则（3准则）进行异常值检测；
5. 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约束，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·m3/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 |
| 463.643 | 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，插值效果好，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（3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，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，算出其算术平均值即剩余误差，按贝塞尔公式算出标准误差，若某个测量值的剩余误差，满足，则认为是含有粗大误差值的坏值，应予剔除，其中贝塞尔公式为   |  |  | | --- | --- | |  | （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：

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；
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；
3. 删除缺值：删除残缺数据较多而无法补充的样本数据；
4. 插值：对能够补全的部分数据缺失使用整体数据平均值进行了代替；
5. 异常值检测：根据拉依达准则（3准则）进行异常值检测。

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，用拟合插方法填补了剩下的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*（？）*

预处理后的统计分析

增加拓展内容1：辛烷值的主要相关因素

5. 问题二：寻找建模主要变量

6.1问题分析

问题二（问题的具体分析），（我们拟采用的方法），其建模求解流程框图如下：

问题二要求从367个变量中选取30个以下具有代表性、独立性的的主要变量。附件一共给出了7个原料性质、2个再生吸附性质、2个产品性质以及354个操作变量共367个建立辛烷值损失模型，采用先降维后建模的方法，忽略次要因素，分析影响模型的主要因素，拟采用因子分析进行降维求解。

因子分析法的核心就是对若干综合指标进行因子分析并提取公共因子，再以每个因子的方差贡献率作为权术与该因子的得分乘数之和构造得分函数。

图6.1 问题二建模求解流程

6.2因子分析

本文在建立辛烷值损失模型过程中采用的因子分析的数学模型如下



上式中，向量是可观测随机向量，即原始观测变量，是的公共因子，即各个原观测变量的表达式中共同出现的因子，是相互独立的不可观测的理论变量，是公共因子的系数，称为因子载荷矩阵，称为因子载荷，是第个原有变量在第个因子的载荷，或者可以将看作第个变量在第公共因子上的权重。同时是与的协方差，也是相关系数，表示对的依赖程度或相关程度。是的特殊因子，是不能被前个公共因子包含的部分，这种因子也是不可观测的。

6.3模型求解

因子分析法的求解步骤

1. 数据标准化，将数据进行无量纲化处理，消除单位的影响
2. 确定待分析的原观测指标是否适合进行因子分析

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6.1 KMO和巴特利特检验

|  |  |  |
| --- | --- | --- |
| 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 |
| 显著性 | 0.000 |

如表6.1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，认为原有变量间存在关联性，因此原有指标适合做因子分析。

1. 确定因子数

使用Matlab建立原观测指标间的相关系数矩阵，求取特征值和特征向量，以碎石图的形式表示

6. （有具体写“问题三：具体问题”，否则“问题三的建模与求解”）

问题三（问题的具体分析），（我们拟采用的方法），其建模求解流程框图如下：

图7.1 问题三建模求解流程

据图对建模求解的具体方法进行说明。

表7.1（问题三的数据处理结果）

|  |  |  |  |
| --- | --- | --- | --- |
|  | 项目1 | 项目2 | 项目3 |
| 种类1 |  |  |  |
| 种类2 |  |  |  |
| 种类3 |  |  |  |

对问题求解的结果进行分析。

8. 模型评价与改进

8.1 模型的优点

1. 这里写模型的优点一。

2. 这里写模型的优点二。

3. 这里写模型的优点三。

8.2 模型的缺点

1. 这里写模型的缺点一。

2. 这里写模型的缺点二。

8.3 模型的改进与推广

本文提出的模型和方法可推广应用于XXXXXXXXX：

1. 本文模型在推广过程的普遍适用性（属优点）。
2. 本文模型在对同一领域的其他研究的参考意义和价值（属优点）。
3. 本文模型在推广过程中可能遇到的困难和不足之处（属缺点）。

参考文献

[1] 刘昕.1100kt/a催化汽油精制装置降低辛烷值损失的措施[J].中氮肥,2020(01):69-71.

附录

程序1：XXXXX程序

程序2：XXXXX程序

程序3：XXXXX程序

程序4：XXXXX程序

0