**汽油精制过程中的数据挖掘**

**一、背景**

汽油是小型车辆的主要燃料，汽油燃烧产生的尾气排放对大气环境有重要影响。汽油清洁化重点是降低汽油中的硫、烯烃含量，同时尽量保持其辛烷值。

辛烷值（以RON表示）是反映汽油燃烧性能的最重要指标，并作为汽油的商品牌号（例如89#、92#、95#）。现有技术在对催化裂化汽油进行脱硫和降烯烃过程中，普遍降低了汽油辛烷值。辛烷值每降低1个单位，相当于损失约150元/吨。以一个100万吨/年催化裂化汽油精制装置为例，若能降低RON损失0.3个单位，其经济效益将达到四千五百万元。

由于炼油工艺过程的复杂性以及设备的多样性，炼油过程的操作变量（控制变量）之间具有高度非线性和相互强耦联的关系，如何优化炼油过程是一个巨大的挑战。

**二、目标**

依据从催化裂化汽油精制装置采集的325个数据样本（每个数据样本都有354个操作变量），在保证汽油产品脱硫效果（这里要求产品硫含量不大于5μg/g），尽量降低汽油辛烷值损失在30%以上的前提下，通过数据挖掘技术，建立汽油辛烷值（RON）在石油炼制过程中的损失函数关系（与操作变量）。

**三、问题**

1. 数据处理

请参考4年的工业数据(见附件一“325个样本数据.xlsx”)的预处理结果，依“样本确定方法”（附件二）对285号和313号数据样本进行预处理（原始数据见附件三“285号和313号样本原始数据.xlsx”）并将处理后的数据分别加入到附件一中相应的样本号中，供下面研究使用。

2. 寻找建模主要变量

由于催化裂化汽油精制过程是连续的，虽然操作变量每3 分钟就采样一次，但辛烷值（因变量）的测量比较麻烦，一周仅2次无法对应。但根据实际情况可以认为辛烷值的测量值是测量时刻前两小时内操作变量的综合效果，因此预处理中取操作变量两小时内的平均值与辛烷值的测量值对应。这样产生了325个样本（见附件一）。

建立降低辛烷值损失模型涉及包括7个原料性质、2个待生吸附剂性质、2个再生吸附剂性质、2个产品性质等变量以及另外354个操作变量（共计367个变量），工程技术应用中经常使用先降维后建模的方法，这有利于忽略次要因素，发现并分析影响模型的主要变量与因素。因此，请你们根据提供的325个样本数据（见附件一），通过降维的方法筛选出建模主要变量（操作变量），使之尽可能具有代表性、独立性，并请详细说明建模主要变量的筛选过程及其合理性。

3. 数据挖掘

在保证汽油产品脱硫效果（这里要求产品硫含量不大于5μg/g），尽量降低汽油辛烷值损失在30%以上的前提下，通过数据挖掘技术，建立在石油炼制过程中汽油辛烷值（RON）与操作变量之间的函数关系。

附件：

附件一：325个样本数据.xlsx

附件二：样本确定方法.docx

附件三：285号和313号样本原始数据.xlsx

附件四：354个操作变量信息.xlsx