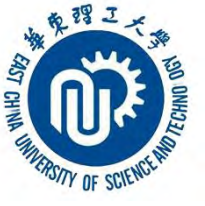



中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

学　　校		北京交通大学
参赛队号		**20100040057**
队员姓名	**1.**	赵飞飞
	2.	卢亚菡
	3.	童　磊

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　降低汽油精制过程中的辛烷值损失模型

摘　　要：

在汽油燃烧的尾气排放对大气环境污染有重要影响的今天，汽油清洁化技术十分关键。我国目前大规模采用的精制汽油的工艺为 S zorb 技术，该方法能够在保证较低的辛烷值损耗的前提下生产在 10ppm 以下的低硫汽油，辛烷值每降低 1 个单位，相当于损失约 150 元/吨。因此，如果能够通过相关工艺技术的调整或改进，降低辛烷值的损失，则可以为整个催化裂化重油化工工业乃至国民经济的发展带来巨大的效益。

对于问题 1，将 285 号和 313 号样本数据采集记录并主要针对非操作变量进行预处理，采取处理如下方式：变量值缺失值取为前后一段时间内平均；若超出变量范围则取为对应的最小值或最大值；利用 **3σ 准则**将异常数据剔除。两样本筛出不符合变量范围的数据分别为 120 和 290 个，剔除 3σ 区间外数据为 0 和 65 个。

对于问题 2，基于数据预处理的 366 个变量（包括 12 个性质变量及 354 个操作变量），根据其属性首先利用基于 **EM 算法**的 **GMM 聚类**对 354 个操作变量进行聚类，将其分为 30 类，并通过 **t－SNE 降维可视化技术**对聚类结果进行可视化；其次对 12 维性质变量及 30 类中每一聚类簇中根据信息增益理论计算其关于辛烷值损失的**信息增益值**，在信息增益计算结果排序的基础上结合**斯皮尔曼相关系数**，对 42 维变量进行筛选，剔除相关性较大及信息增益值较低的变量，最终筛选出 **28 维具有代表性和独立性的变量**（包括 3 维性质变量及 25 维操作变量）

对于问题 3，根据前问筛选出来的 28 个具有代表性及独立性的主要变量，建立了**基于慢特征分析的即时学习（SFA-JITL）辛烷值损失预测**框架。该模型首先对 28 个主要变量进行了慢特征分析筛选出 **7 维慢特征**，其次搭建即时学习框架对样本序列依次输入完成每个样本值辛烷值损失的预测，并在加入更新样本的过程中利用**局部加权慢特征方法**对慢特征权重不断调整，该模型适用于与工艺流程数据更新的预测建模。此外，SFA-JITL 辛烷值预测模型结果与真实值对比并于其他模型相比较验证了模型的有效性。

对于问题 4，根据问题 3 的辛烷值损失预测值，总结 325 个样本的辛烷值减损方案。在保证产品硫含量处于汽油标准范围的前提下，挑出**降幅大于 30%的样本共 56 个**，并详细分析所对应的主要变量经过优化后的操作条件，采取求均值的方法，获得最大可能适用所有样本的操作条件，进而为石化企业的长期、高效运营提供科学稳健的理论支撑。

对于问题 5，根据问题 4 中的优化方案，对 133 号样本的各主要操作变量进行逐步的迭代，最终**迭代 49 步**到达最终的优化方案。经过模型的分析预测，得到了整个迭代过程中产品的辛烷值、辛烷值损失和硫含量值数据，对整个迭代优化过程中对应的成品汽油的辛烷值和硫含量的变化轨迹进行可视化。

关键词：辛烷值损失，GMM 聚类，信息增益特征筛选，SFA-JITL 框架，操作变量优化

1. 问题重述

1.1 问题背景

石油炼制工业是国民经济的重要支柱产业，其中，催化裂化工艺是石化工业中最重要的工艺之一，其兴起源于现代运输行业对高辛烷值汽油和甲苯的需求。该工艺既可以实现原油深度加工，又可以提高轻质油收率、品质和经济效益[1]。目前，随着家用汽车的日渐普及，中国已逐步发展为世界上最大的燃料油品市场，且汽油消费量仍在逐年高速增长。

图 1 所示为世界各大经济体的汽油池组成情况，纵观全球平均水平，催化裂化汽油和重整汽油占比近乎持平；反观我国，催化裂化汽油占汽油池的 74%，是重油轻质化工艺技术的核心。截至 2019 年 12 月，我国成品汽油中 95%以上的硫和烯烃来自催化裂化汽油[2]。因此，今后催化裂化汽油还有很大的发展空间。

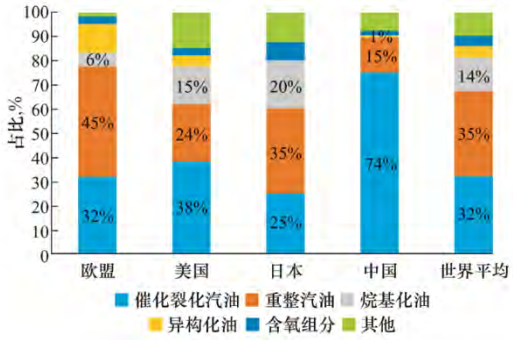


图 1 世界各地汽油池组成情况

此外，重整汽油占比 15%，我国对催化工艺的研究开发始于 20 世纪 50 年代，根据催化剂再生方式的类型，国内重整技术主要分为半再生重整技术、连续重整技术两种。截至 2019 年 6 月底，我国的连续重整规模达到了 93.32Mt/a，是半再生重整技术的 11.5 倍。连续重整技术的过程由许多单元组成，还伴随着复杂的化学反应，如图 2 所示[3]。然而，催化裂化技术比连续重整技术更为繁杂，这导致过程中操作变量（控制变量）规模相当可观，且各操作变量之间具有高度非线性和相互强耦联的关系。因此，对催化裂化技术进行控制和优化仍然是学术界和工业界面临的巨大挑战。

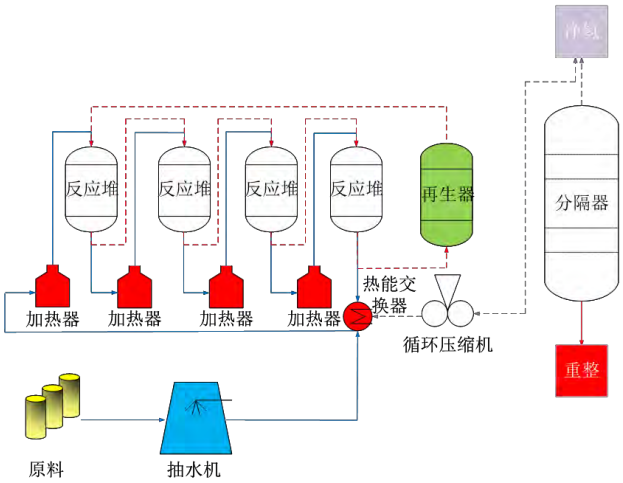


图 2 连续重整技术流程示意图

为了评估催化裂化过程的运行状态，需要监测一些关键的质量相关指标（比如辛烷值、电机辛烷值、蒸气压），以供工程师参考。在这些指标中，辛烷值（RON）被广泛

用于评估汽油质量、反映燃烧性能。在汽油燃烧的尾气排放对大气环境污染有重要影响的今天，汽油清洁化技术十分关键。汽油清洁化目的是降低汽油中的硫、烯烃含量，同时尽量保持产品中的辛烷值

我国原油对外依存度超过 70%，且大部分是中东地区的含硫和高硫原油，因此在催化裂化重油精制为轻质化油的过程中，若能改善相关工艺技术，探索出符合操作要求条件的辛烷值损失与各项操作变量及原料、产品性质之间的关系，则可以导向性地对于相关操作进行调整进而降低辛烷值的损失，从实际化工流程中产生的经验、数据等完成减少辛烷值损失工艺的改进。

综上所述，利用数据挖掘技术，分析非线性特征，采用即时学习框架，进一步进行建模预测，对有效降低汽油辛烷值，提高经济效益具有重要意义。以上研究内容的实现将为石化企业的高效运营提供理论支撑。

1.2 需要完成的任务

任务一：数据预处理

根据催化裂化汽油精制装置采集到的实际工艺原始数据，需要对 285 及 313 号样本原始的 40 个不同时间测得的数据整合确定样本的性质、操作变量性质及最后产品的辛烷值等。原始数据中包括一定程度的不良数据，例如部分位点的数据异常、超出变量实际范围、无法通过数据筛选准则等等。最终需根据相应数据预处理方法，整定数据并筛选样本，为后续建模任务奠定基础。

任务二：建模主要变量选取

在上述数据预处理的基础上需要对影响辛烷值损失的主要变量进行筛选，从而为后续的辛烷值损失预测模型及主要操作方案的优化奠定基础。这一任务中，根据具体的工程实际，需要筛选出尽可能具有代表性、独立性的 30 个以内的主要影响变量。由于控制变量之间具有高度非线性和相互强耦联的关系，需要采用适当的方法筛选主要代表性变量。

任务三：辛烷值损失预测模型

通过任务二代表性独立性变量的选取，需要将该些变量与辛烷值损失之间建立相互关系，从而为后续主要操作方案的优化做铺垫。预测模型通过数据挖掘技术发现筛选变量与辛烷值损失之间潜在的模型关系，并进行模型验证以论证模型有效性。该模型在一定程度上解释了各变量对辛烷值损失值造成的影响，为后续调整变量的优化操作奠定基础。

任务四：主要变量操作方案的优化

通过主要变量操作方案的优化，实现辛烷值损失的降低，是降低汽油精制过程中的辛烷值损失建模的终极任务，优化过程需要利用数据样本对辛烷值损失降幅大于 30%的样本并给出每个样本对应的主要变量优化后的操作条件，在保证汽油产品脱硫效果的基础上对主要变量的操作方案进行说明，为企业装置操作降低辛烷值损失提供参考。

任务五：优化调整过程可视化展示

这一任务主要利用上述模型及变量操作方法，对 133 号样本操作过程进行调整并展示主要操作变量优化调整过程中对应的汽油辛烷值和硫含量的变化轨迹。各操作变量需在取值范围内逐步调整到位，以符合实际工艺流程并以图形方式对调整路径实现清晰明了的可视化展示。

2. 模型假设及关键性符号说明

（1）假设题目所采集数据可以良好反映工厂实际化工过程即汽油精制过程中的一般运行情况；

（2）假设汽油精制过程中都属于正常运行状态，没有异常情况出现；

（3）假设通过数据找出的主要操作变量可以反映出辛烷值损失和硫含量变化的一般规律；

（4）假设每改变一次主要操作变量的值，产品汽油的辛烷值和硫含量都会相应及时的发生变化。

符号说明：

符号	意义
C_n	第 n 号操作变量原始数据向量序列
$C_n^{\,2}$	第 n 号操作变量预处理后数据向量序列
$\left\|C_n^{\,2}\right\|$	预处理后数据向量序列集合中元素的个数
$X_n^{\,i}$	第 n 号操作变量的第 i 个数据向量
$X_n^{\,it}, X_n^{\,id}$	数据向量中操作变量的采集时间和具体数据值
T	数据连续缺失时长上阈值
S	相邻数据向量之间的时间差
$Z_n^{(i)}$	第 n 号主要操作变量第 i 次迭代后的数值
$Z_n^{\,o}$	第 n 号主要操作变量最优值即目标数值
$\left\|\Delta_n\right\|$	第 n 号主要操作变量单次可调整幅值
$step_n$	第 n 号主要操作变量需要的迭代步数
$step$	所有主要操作变量需要的最大迭代步数

3.问题一分析与求解

3.1 问题一分析

本文数据为来自于中石化高桥石化实时数据库（霍尼韦尔 PHD）及 LIMS 实验数据库。其中操作变量数据来自于实时数据库，采集时间为 2017 年 4 月至 2020 年 5 月，采集操作位点数共 354 个，即采集**操作变量共 354 个**。原料、产品和催化剂数据来自于 LIMS 实验数据库，数据时间范围为 2017 年 4 月至 2020 年 5 月，数据包括了原料、产品、待生吸附剂、再生吸附剂的相应性质参数共 14 个，其中**非操作变量共 13 个，非变量参数（辛烷值损失）1 个**。同时还应注意到，原料及产品的辛烷值均为建模变量，即两者的差值为辛烷值的损失值，因此在建模辛烷值损失预测模型时，考虑到原料的辛烷值已知，则只将原料的辛烷值纳入变量考虑范围之内。于是本文中考虑将 354 个操作变量和 12 个非操作变量**共 366 个变量**（去除产品辛烷值）共同作为辛烷值损失模型构建的相关变量。

辛烷值损失模型的构建需要符合实际情况且误差控制在一定范围内的、相对准确的相关变量测量数据作为支撑。但是在实际的操作装置在进行数据测量、数据记录、数据导出等过程中难免会存在一定问题，使得最终得到的数据（原始数据）与希望得到的良好数据存在一定差距。例如，操作装置在进行数据测量过程中会遇到各种不同条件的较为复杂的实际工况，最终会导致采集的原始数据中存在着或多或少的不良数据，包括连续或间断性的数值缺失、数值漂移（偏大或偏小）等情况。因此，对不良数据进行科学有效地预处理，对于辛烷值损失模型的构建有着决定性意义。

对于实时数据库采集的不同位点的数据（操作变量数据）来说，采集数据的不同装置的数据均有部分位点存在问题，即部分变量只含有部分时间段的数据，部分变量的数据全部为空值或部分数据为空值，同时存在部分变量的数据超出了因此对原始数据进行处理后才可以使用；对于 LIMS 实验数据库采集的原料、产品、待生吸附剂、再生吸附剂的相应性质参数数据（非操作变量数据）来说，由于这些性质参数通常情况下是相应物质的固有属性，在一定时间范围内可以认为不发生变化，这从该数据采集频次为每周 2 次这一点上也可以看出。同时，根据观察发现，这 14 个性质参数在 285 号和 313 号中不存在空值缺失等问题，因此**以下只考虑对原始数据表中操作变量的处理**。

综合以上分析，现针对 285 号和 313 号样本的原始数据中各种类型的不良数据分别进行相应分析及处理。

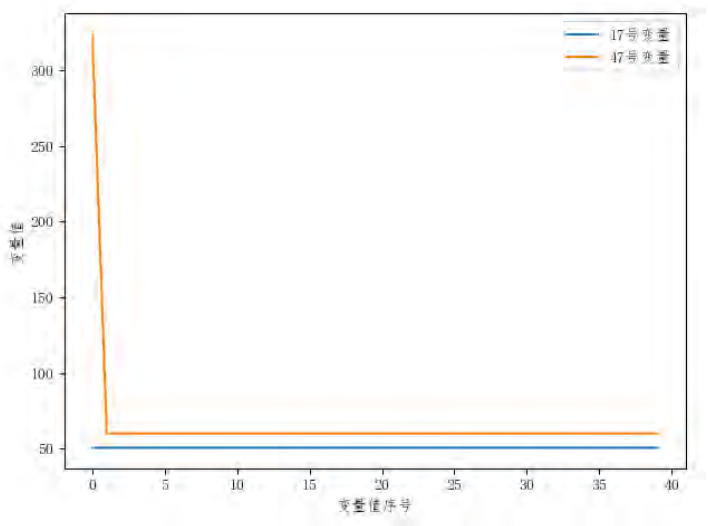


图 3-1 第 17、47 号变量数据折线图

对于 285 号和 313 号样本的原始数据，在采集数据的这两个小时过程中，操作变量基本保持不变，这一点观察数据便很容易发现，本文中以 285 号样本的某两个操作变量的数据情况为例，见图 3-1 所示。由图可知，所示两个操作变量的数据基本维持在某个数值附近，采集到的数值上微小的差异可能只是由于数据采集装置受到噪声干扰所致。而其中存在的**操作变量数值突变**的情况，应将其作为异常值剔除掉，剔除的标准则可以采用拉依达准则（3 σ 准则）。除这两个操作变量的原始数据之外，我们同时分析了其他操作变量的具体情况，确认了该思路的正确性。综上，针对该原始数据可能存在的异常情况作如下方面的预处理：

(1) 变量值缺失为空值

操作装置数据采集过程中导致部分点位出现连续的或者间断性的变量值缺失为空值，对于这种情况主要采取利用缺失值前后两小时内的数值取平均的方式来处理。然而在实际处理过程中，原始数据总时间跨度为 2 小时，因此根据实际情况可以采用缺失值前后半小时或一小时内的数据的平均值来代替。

(2) 变量值超出可调控范围（在要求的最小值到最大值之间）

在实际操作过程中，可能会使得操作变量的实际控制超出附件四所要求的最小最大值范围，对这种情况则考虑将该操作变量超出范围区间的值剔除出去，即在对不同位点的操作变量数据取平均得出最终的操作变量确定值过程中，对这些超出最小最大值范围的数据不予考虑。

(3) 变量值超出 3 σ 区间范围

有理由认为操作装置调整操作变量的过程为连续的，不会出现跳跃的情况。但是可能由于操作装置采集数据过程可能存在问题，在实际的 285 和 313 号样本原始数据中，可以发现该类型的“跳跃”。因此对于每个位点采集到的数据，考虑将这样的跳跃值剔除出去，即在对不同位点的操作变量数据取平均得出最终的操作变量确定值过程中，对这些在 3 σ 范围外的数据不予考虑。

(4) 不同表格中数据单位不一致的情况

同时可以发现，在所给的几个不同的表格数据中，存在不同操作变量数据的单位不一致的且无法辨认识别，导致无法对其进行有效的处理，对于此种情况，考虑在后续的处理中放弃对该操作变量的处理。

3.2 数据预处理

3.2.1 不良数据处理

(1) 变值缺失为空值

某个点位的操作装置数据采集过程中导致部分点位出现连续的或者间断性的变量值缺失为空值。

针对上述情况，假设连续数据采样过程下第 n 个点位的操作装置采集第 n 个操作变量数据序列为 $C_n=\left\{X_n^{\ 1},X_n^{\ 2},\cdots,X_n^{\ i},\cdots\right\}\left(i=1,2,3,\cdots\right)$，且 $X_n^{\ it},X_n^{\ id}$ 分别表示数据向量中的时间值和操作变量数据值。由原始数据可知操作装置每次的采样间隔实际时长为 3 分钟。由此可得如下存在空缺值的判定条件：

$$\left|X_n^{\ it}-X_n^{\ (i-1)t}\right|\geq 3 \tag{3-1}$$

同时，作如下两方面考虑：

① 如果空缺值过多，那么该位点采集到的数据的可信任度便越低，我们这里将连续空缺 10 个数据，即连续空缺数据 1 小时设定为一个时间上限阈值 T，超出此阈值的数据不予信任，即不予考虑，

② 前述已经提到，由于认为操作装置在一定时间内可以认为操作变量保持不变（变化很微小），因此对于一定时间范围内连续性的空缺值，可以认为这些空缺值保持一致。这里设定为连续空缺数据 1h(T)内，可以考虑空缺前 0.5h 和空缺后 0.5h 共 1h（为了满足连续空缺数据时长与用来平均的数据的时长之和在 2h 之内）的数据的平均值作为空缺值来处理。

综合以上考虑，假设在时间$i-1$到i之间补全的空缺值个数与数值为$\left(k, A_n^{(i-1)i}\right)$，并令时间间隔$S=\left|X_n^{it}-X_n^{(i-1)t}\right|$，则对于空缺值的处理如下：

$$\left(k, A_n^{(i-1)i}\right)=\begin{cases}(0,\ \sim),\ S<3\\ \left(S/3,\ ave\left(\sum_{j=i-10}^{i-1}X_n^{jd}+\sum_{j=i}^{i+9}X_n^{jd}\right)\right),\ 3\le S\le T\\ \text{舍弃},\ S>T\end{cases} \tag{3-2}$$

上述已经提及，在实际处理过程中，我们令$T=60$。

(2) 变量值超出可调控范围

假设第 n 个操作变量的取值范围为$\left(l_n,\ u_n\right)$，$\forall X_n^{\ i}\in C_n$,对于序列C_n作如下处理：

$$C_n\leftarrow\begin{cases}C_n,\ if\ l_n\le X_n^{id}\le u_n\\ C_n-\left\{X_n^{\ i}\right\},\ if\ X_n^{id}\le l_n\ or\ X_n^{id}\ge u_n\end{cases} \tag{3-3}$$

进一步地，去除序列C_n中的最大值和最小值，以获取更加准确的操作变量值，即分别作如下处理：

$$C_n\leftarrow C_n-\left\{X_n^{\ p}\right\}\ if\ X_n^{pd}=\max_i\left\{X_n^{id}\right\} \tag{3-4}$$

$$C_n\leftarrow C_n-\left\{X_n^{\ q}\right\}\ if\ X_n^{qd}=\min_i\left\{X_n^{id}\right\} \tag{3-5}$$

(3) 变量值超出 3σ 区间范围

现在假设经过(1)和(2)处理后新的操作变量序列为$C_n^{\ 1}$，且设$\left|C_n^{\ 1}\right|=N_1$，现在根据 3σ 准则来对$C_n^{\ 1}$分别依次进行如下处理：

$$X_n^{\ d}=\frac{1}{N_1}\sum_{i=1}^{N_1}X_n^{id} \tag{3-6}$$

$$\sigma=\sqrt{\frac{1}{N_1-1}\sum_{i=1}^{N_1}\left(X_n^{id}-X_n^{\ d}\right)^2} \tag{3-7}$$

$$C_n^1 \leftarrow \begin{cases} C_n^1,\ if\ X_n^d - 3\sigma \le X_n^{id} \le X_n^d + 3\sigma \\ C_n - \left\{X_n^i\right\},\ otherwise \end{cases} \tag{3-8}$$

我们把经过(3)处理后的新的操作变量序列记为C_n^2，且设$\left|C_n^2\right| = N_2$，那么综合(1)~(3)，现在我们可以得到第 n 个操作位点的最终确定的操作变量值为如下：

$$X_n = \sum_{i=1}^{N_2} X_n^{id},\ X_n^i \in C_2 \tag{3-9}$$

(4) 不同表格中数据单位不一致的情况

我们同时发现提供的数据表格中存在同一操作变量的数据单位不一致的情况，并且无法辨认识别，导致无法对其进行有效的处理，对于此种情况，在后续的处理中直接放弃对该操作变量的处理，如第 49 号操作变量原料缓冲罐液位。

3.2.2 算法及处理过程

现制定如下的数据处理程序框图（图 3-2），以表明数据数据流向，及上述问题一分析的具体处理流程。

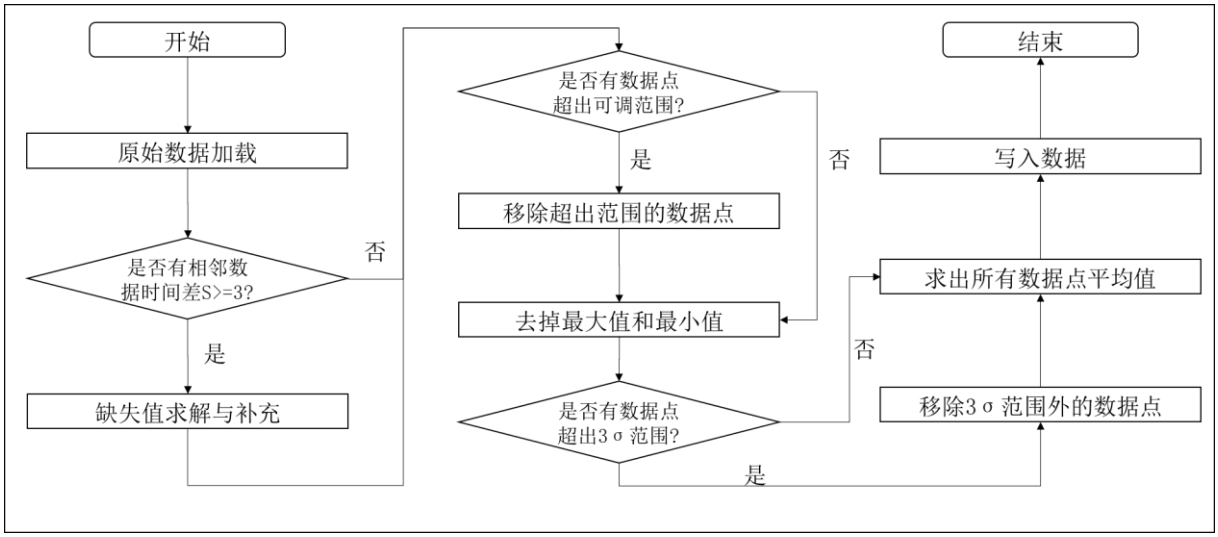


图 3-2：数据预处理流程图

3.3 数据预处理结果

题目给出了 285 号和 313 号样本时长共为 2h 的原始数据，原始数据给出了原料、产品和催化剂数据以及所有操作变量的测量数据。要求对原始 Excel 文件中原始数据信息进行数据预处理。

本文采用 python 程序对两个样本原始数据信息进行数据预处理，具体的步骤如下：

(1) 对原始数据中所有可能存在的空缺值进行检查，经过检查发现该数据中不存在任何的数据缺失（连续性或间断性的）。

(2) 检查原始数据中所有位点处采集到的数据是否符合附件四规定的最大最小值要求。经过检查，发现在这**两个样本原始数据中均存在同样的变量超出范围值的情况**，超出范围的这三个操作变量见表 3-1 所示，超出范围的数据存在包括了过大或过小两种情况。在此之后，去除了每个操作变量中最大值与最小值。

表 3-1 操作变量超出范围情况

编号	位号	中文名称	超范围情况
49	S-ZORB.SIS_LT_1001.PV	原料缓冲罐液位	过大
84	S-ZORB.AI_2903.PV	再生烟气氧含量	过小
111	S-ZORB.FT_1204.TOTAL	-	过大

(3) 在(2)的基础上，继续对数据进行 3σ 检验筛选，将其中可能偏离正常范围的值筛选出去。

经过上面的处理可以发现，步骤(2)中检查出来的第 49 号操作变量的远远超出要求的范围，该异常可能是由于不同表格中对该数据采用了不同的单位，而由于该单位根据提供的表格又无法进行还原，故在后续的处理中考虑将超出范围的变量值直接设定为该变量的最大值或最小值。

我们这里列出了 285 号样本原始数据经过处理后的结果，见表 3-2 所示。

表 3-2：285 号样本原始数据处理结果

编号	变量值	编号	变量值	编号	变量值	编号	变量值	编号	变量值
1	0.2733991	72	49.981307	143	-1.359208	214	4.0124291	285	0.1452189
2	24.208241	73	10.62899	144	105.45045	215	0.2834373	286	391.35238
3	2.5288704	74	0	145	21.807941	216	0	287	-0.253295
4	855.88252	75	2.3324642	146	47080.568	217	12.347814	288	-0.141465
5	421.50933	76	304.28485	147	34.413484	218	0.0404216	289	2.5337748
6	421.19624	77	289.96294	148	246.17465	219	1.0408825	290	411.70608
7	2.4270934	78	14.74304	149	0.1103821	220	0.2288788	291	418.92724
8	59.703011	79	78.503659	150	40.879962	221	1.0884328	292	415.58526
9	1108.2854	80	0.1135614	151	0	222	27.60731	293	413.03457
10	244.12175	81	37.388848	152	38.097822	223	24.58094	294	206.6559
11	320.42618	82	278.74076	153	35.890968	224	0.1503955	295	401.54985
12	2.4378081	83	493.88708	154	8.7953466	225	63.301385	296	446.23932
13	322.94328	84	0.5	155	223.48495	226	4.5152548	297	312.04005
14	5.8014656	85	38.567096	156	0.1690422	227	0.1319319	298	94.066782
15	0.6498192	86	10.925054	157	0.194687	228	8.1048482	299	0
16	126.62636	87	6175535.9	158	-0.380915	229	41.576178	300	335351.6
17	50.76226	88	1.7919679	159	94.288893	230	404.5059	301	87.750468
18	591.49149	89	607.0255	160	4.8277871	231	435.38922	302	373.2428
19	35.607198	90	1779805.8	161	285.05843	232	491.11406	303	373.38824
20	36.803967	91	10678623	162	25.018065	233	51.67095	304	61.72605
21	136.35251	92	6440.3578	163	2.8134454	234	4.5786288	305	42.184609
22	3.2030152	93	3061459.2	164	47.008992	235	422.47681	306	1528.542
23	0.8423927	94	19925674	165	-1.347855	236	33.641588	307	143.37015
24	766.08381	95	75820.439	166	294.6726	237	421.7869	308	34.941429
25	0	96	0	167	-4.751134	238	116.27937	309	34.536572
26	2360.5178	97	1164934.9	168	126.18964	239	421.19057	310	228.01807
27	1184.395	98	18024.439	169	66.912495	240	418.26273	311	422.30374
28	4.8709879	99	21893763	170	23.121955	241	1.0684636	312	0.5215696

29	35.793521	100	11203641	171	137.61311	242	20	313	0.0563587
30	0.3819439	101	66542.466	172	72.934466	243	82.266836	314	3.430658
31	486.53838	102	102409.39	173	43.949174	244	0.9655794	315	0.6023524
32	591.49149	103	2811449.2	174	54.526079	245	933.2332	316	0.4444199
33	0.9966278	104	48937211	175	577.34253	246	9.0791885	317	0.5482048
34	201.42632	105	24607699	176	-131010.1	247	0.0034067	318	0.745978
35	457.97091	106	-455671.5	177	10	248	-0.380884	319	0.7090541
36	456.97494	107	109.9855	178	674.44259	249	-0.381825	320	0.6622907
37	1772.128	108	2979984.5	179	115.63998	250	12.125632	321	1.1617232
38	277.83588	109	30.699441	180	490.77504	251	-0.03438	322	0.6550613
39	355.1831	110	0	181	492.37594	252	8.2658499	323	1.7346927
40	0.6411561	111	2500000	182	39.376913	253	23.225083	324	0.3815607
41	112.49933	112	1018502	183	-0.908778	254	23.097861	325	422.47681
42	0.5310105	113	413788.45	184	29.375149	255	569.66544	326	25.080279
43	44.300125	114	2973652	185	-0.899307	256	5.8070506	327	5.8070506
44	44.701206	115	2884400.6	186	241.3761	257	1276.4008	328	256.5611
45	101.05652	116	133.72461	187	1.0019082	258	0.0341815	329	345.3932
46	0	117	2.7211023	188	4.5786288	259	34.028421	330	30.699441
47	60.133328	118	2.6208263	189	22.274519	260	0.0300918	331	39.82213
48	43.86586	119	362.1262	190	730.74976	261	34.514926	332	0
49	80	120	119.92202	191	345.3932	262	330139.3	333	44.80332
50	65.938626	121	360.9564	192	3.1764869	263	-1511.205	334	360.9564
51	134.20282	122	149.80376	193	-0.014209	264	-273304.5	335	362.1262
52	129.44934	123	61.336831	194	0.0501739	265	-13121.4	336	422.30374
53	6527.7696	124	35.938946	195	6770.3099	266	0.1168448	337	149.80376
54	244.19839	125	42.26952	196	68.904645	267	34.271117	338	119.92202
55	65.795085	126	448.40089	197	319.52683	268	0.1153644	339	133.72461
56	0.0995019	127	0.3601841	198	-0.003077	269	34.34783	340	0.2830471
57	362.75292	128	-1.401113	199	0.3110278	270	15.62476	341	22.287637
58	139.80728	129	49.697883	200	-0.813467	271	19.24878	342	5.8515561
59	-0.121037	130	44.80332	201	279.8194	272	17.27226	343	82.288434
60	2.1958531	131	-0.647146	202	-1142.294	273	22.19806	344	92.659055
61	433.19237	132	-1.561129	203	256.5611	274	47.357668	345	2.3585593
62	412.87214	133	0.4502131	204	0.5528859	275	46.108543	346	3321.5832
63	-0.267372	134	34.530876	205	0	276	1.3969563	347	190.6942
64	0.1561951	135	258.63206	206	0.2648783	277	1.0285338	348	98944917
65	0.3596669	136	50.691596	207	0.0332799	278	35.820237	349	2433448
66	3.1801015	137	0.3042989	208	0	279	16.58084	350	2200.7891
67	2.3648251	138	-7740.235	209	-4.024113	280	3349.8923	351	5149259
68	49.601571	139	-1.257131	210	1.4605631	281	3349.9184	352	2846.8966
69	116.64519	140	21.823702	211	-0.073147	282	3358.323	353	5984749.3
70	132.32323	141	35.064482	212	-40.25356	283	100	354	-97.2107
71	38.131321	142	39.822829	213	1.3473398	284	22.852549		

4.问题二分析与求解

4.1 问题二分析

在本题中，需要根据样本的处理结果，对影响辛烷值损失的主要变量进行筛选，并使变量之间尽可能具有代表性、独立性。在此基础上才能构建良好的辛烷值损失预测模型。

为了使得汽油精制工程中的应用方便，问题二要求需要将包括 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）筛选降维，并且筛选后的主要变量需要控制在 30 个以下。另外，由于辛烷值损失是由原料性质中的辛烷值变量以及汽油产品中的辛烷值变量计算的结果，此外在工程实际应用中，原料的辛烷值是可以提前预知测量的，可视为已知变量，因此在筛选主要变量之前，需要剔除产品性质中的辛烷值变量，即需要在 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、1 个产品性质变量（即产品硫含量）以及 354 个操作变量（共计 366 个变量）中选取对辛烷值损失构成影响因素较大的性质及变量等。

由于变量维数众多，直接使用单一方法对变量特征进行筛选降维虽然简便，但忽视了很多变量之间的关联性及区别性。为了对变量的区分度及重要程度进行分析，我们采用基于 EM 算法的高斯混合模型聚类方法对除性质因素外的操作变量进行聚类，将数据归类可以避免筛选过程的盲目性，在该聚类基础上利用对每一类及性质变量利用信息增益理论对影响辛烷值的混合信息增益做计算，变量筛选的技术路线如下图 4-1 所示：

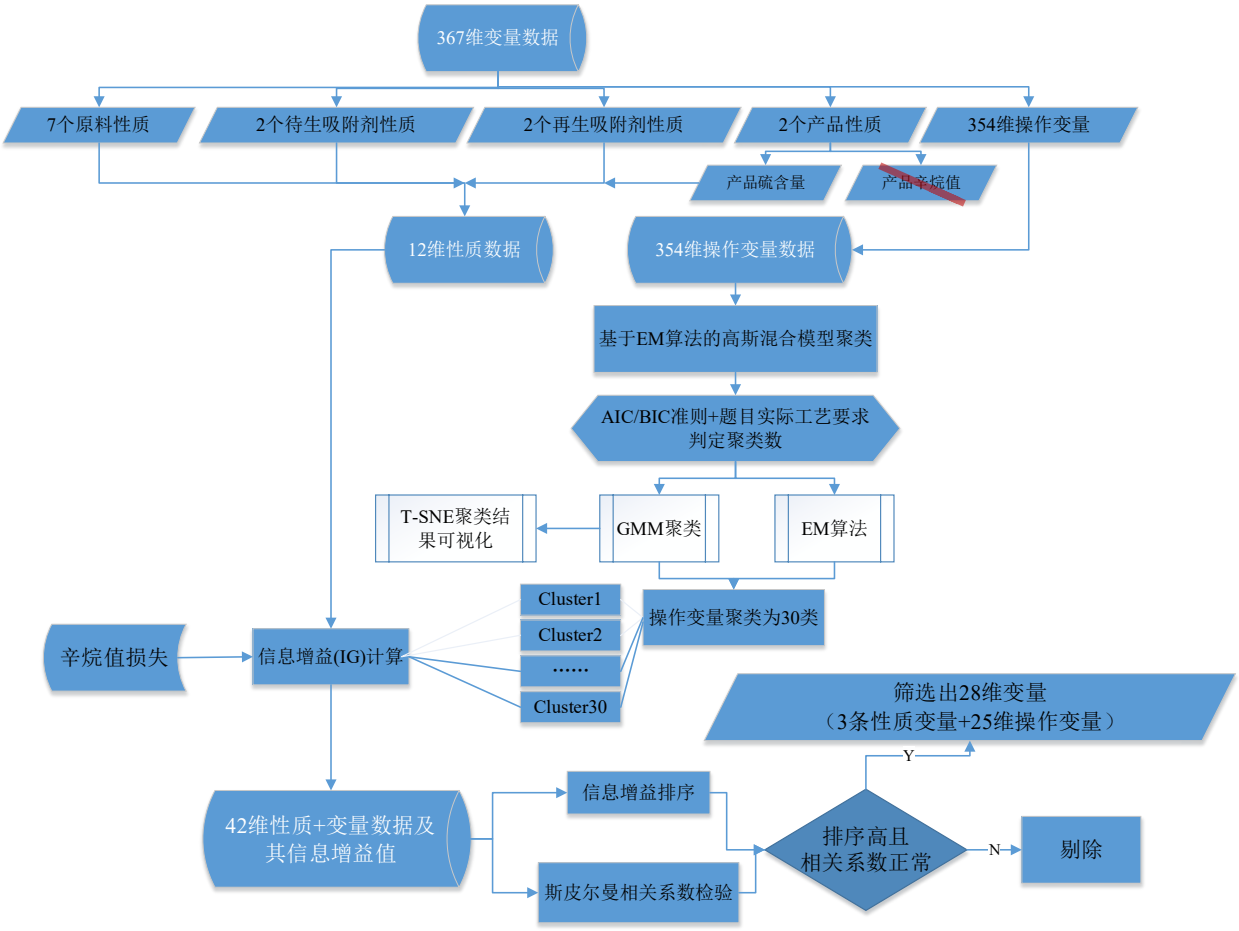


图 4-1 变量筛选技术路线

首先，在剔除产品辛烷值后的 366 维变量中将性质变量与操作变量单独分出(12 维性质变量及 354 维操作变量)，只对 354 维操作变量进行聚类，认为其在每一类中的数据表现特征及分布存在相似性，该过程利用 EM 算法求解高斯混合模型(GMM)聚类，用高维的高斯分布近似拟合每个操作变量在样本中的数据分布。聚类数有 AIC 及 BIC 准则判定并结合题目要求的变量筛选数选择，聚类结果可以为后续特征筛选提供基础，使得后续特征筛选可以直接在各类中操作，另外也利用了 t-SNE 降维可视化方法展示了聚类效果，表明了聚类结果的优越性。

其次，将原本的 12 维性质变量，以及聚类后的 30 类 354 维操作变量，对性质变量及每一类中的操作变量利用信息增益理论计算其对于辛烷值损失的信息增益，从而得出性质变量及每一类中操作变量的信息增益值。

最后，从每类中筛选出信息增益排序最高的操作变量，与性质变量组合成 42 维总变量，利用信息增益排序及斯皮尔曼相关系数的检验，剔除出信息增益最低且与其他变量相关系数较高的变量，完成总变量的筛选。最终筛选出 28 维变量（包括 25 个操作变量，产品性质硫含量，原料性质辛烷值，待生吸附剂性质 S 的 3 个性质变量）。

4.2 操作变量聚类

对于变量维数众多的特征筛选，直接进行筛选或降维方法的运用会忽略不同特征的已知属性。若某几个操作变量在数据上表现相近，且这些变量对于辛烷值损失的影响都比较巨大，直接在此众多变量中筛选降维会出现难以趋势的状况。对于因此需要根据变量在汽油催化裂化化工制造工艺流程中的属性并在此基础上对变量进行聚类。

首先对于 366 个变量，即 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、1 个产品性质变量（产品硫含量），根据各变量属性，将原料性质、待生吸附剂性质、再生吸附剂性质、产品性质变量单独分出，即只对 354 维可操作变量进行聚类分析。

4.2.1 高斯混合模型聚类

高斯混合模型(Gaussian mixed model, GMM)是多个高斯分布函数的线性组合，通过多个高斯分布函数的组合，理论上随着模型中高斯分布函数的增多，它可以精确地拟合任意复杂类型的分布[4]，它是学习速度最快的概率模型， GMM 试图找到能最佳模拟输入数据集的多维高斯分布的混合。

对于所有的 354 维操作变量，即 $d=354$，在汽油催化裂化精制过程中所有的操作变量 $x=\left(x^1,x^2,\ldots,x^d\right)^T$，包含 K 个组件的高斯混合模型可采用下述公式表示：

$$\begin{cases} p(x)=\sum_{k=1}^{K}\omega_k N\left(x|\ \mu_k,\Sigma_k\right) \\ N(x|\ \mu_k,\sum k)=\dfrac{1}{\sqrt{\left|\sum_k\right|}(2\pi)^{\frac{d}{2}}}\exp\left\{-\dfrac{1}{2}\left(x-\mu_k\right)^T\sum_k{}^{-1}\left(x-\mu_k\right)\right\} \\ \sum_{k=1}^{K}\omega_k=1,0\le\omega_k\le 1 \end{cases} \tag{4-1}$$

其中，$N(x|\ \mu_k,\sum k)$ 高斯概率密度函数，ω_k、μ_k、$\sum k$ 分别为混合模型中的第 k 个组件的权重、 均值、协方差矩阵。

根据实际经验，许多采集的数据集的分布服从实际上的高斯分布，并且在原始数据集中某变量在极个别数据条目中不符合高斯分布，根据中心极限定理，当样本越来越多时， 其变量分布会趋向于高斯分布。因此 GMM 在数据集规模较大的情形下可以拟合地更好，在聚类过程中也可以适用于更多灵活的簇的形状。另外，理论上通过增加模型的个数，GMM 可以逼近任何连续的概率密度分布以灵活的聚类的需求。

期望最大化算法(expectation-maximum, EM)可以用来迭代是一种 GMM 聚类优化策略，该算法经常被用来求解聚类问题。每一次迭代分为两步，分别为期望步(E 步)和极大步(M 步)。区别于 Kmeans 的依照具体的值来聚类，GMM 在聚类的过程中的输出是一系列的概率值(表示对应于各个不同的类的概率)，Kmeans 聚类效果与 GMM 聚类效果区别图如 4-2 所示，这种依概率值所携带的信息量比简单分配到某一类要多，因此依据此概率值所得的聚类结果更为精确。

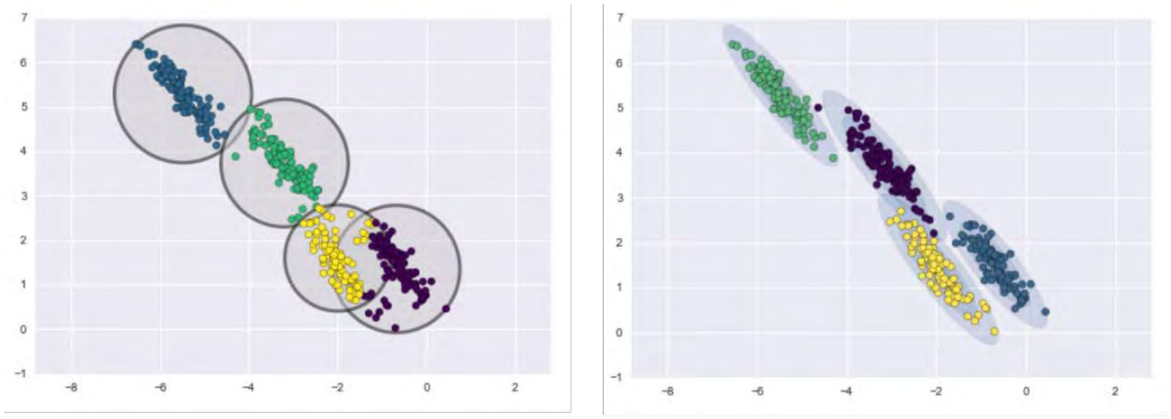


图 4-2 Kmeans 与 GMM 聚类效果区别

将高斯混合分布的参数集合记为：$\theta=\{\phi_1,\cdots,\phi_K,\mu_1,\cdots,\mu_K,\Sigma_1,\cdots,\Sigma_K\}$，将

$z_i\in\{1,2,\cdots,K\}$表示生成高斯混合分量，可由贝叶斯公式计算出 z_i：

$$\gamma_{ik}=p_M\left(z_i=k|\ x_i\right)=\frac{P\left(z_i=k\right)\cdot p_M\left(x_i|\ z_i=k\right)}{p_M\left(x_i\right)}=\frac{\bar{\phi}_k\varphi\left(x_i|\ \bar{\mu}_k,\bar{\Sigma}_k\right)}{\sum_{k=1}^{K}\bar{\phi}_k\varphi\left(x_i|\ \bar{\mu}_k,\bar{\Sigma}_k\right)} \tag{4-2}$$

其中，$\bar{\phi}_k,\bar{\mu}_k,\bar{\Sigma}_k$ 为模型参数 θ 的估计值，由以下式子给出：

$$L(X)=\ln^n\left(\prod_{i=1}\ p_M\left(x_i\right)\right)=\sum_{i=1}^{n}\ln\left(\sum_{k=1}^{K}\phi_k\cdot p_M\left(x_i|\ \mu_k,\Sigma_k\right)\right) \tag{4-3}$$

至此高斯混合分布已知，并可根据 $\lambda_i=\underset{k\in\{1,2,\cdots,K\}}{\operatorname{argmax}}\gamma_{ik}$，对样本集 X 中每个样本 i 所属的簇标记为 λ_i。

4.2.2 聚类结果

根据上述 EM 算法及 GMM 聚类方法的应用，首先去除汽油原材料性质、产品性质、待生吸附剂性质、再生吸附剂性质等，对 354 个操作变量应用上述 EM 算法及 GMM

聚类方法，并依据 AIC 及 BIC 准则，去 AIC 及 BIC 的较小值的分类个数并根据题目要求操作变量控制在 30 个一下的要求，综合上述所有依据，选择 GMM 聚类个数为 30，其 AIC 准则与 BIC 准则根据 GMM 聚类变化情况如下图 4-3 所示：

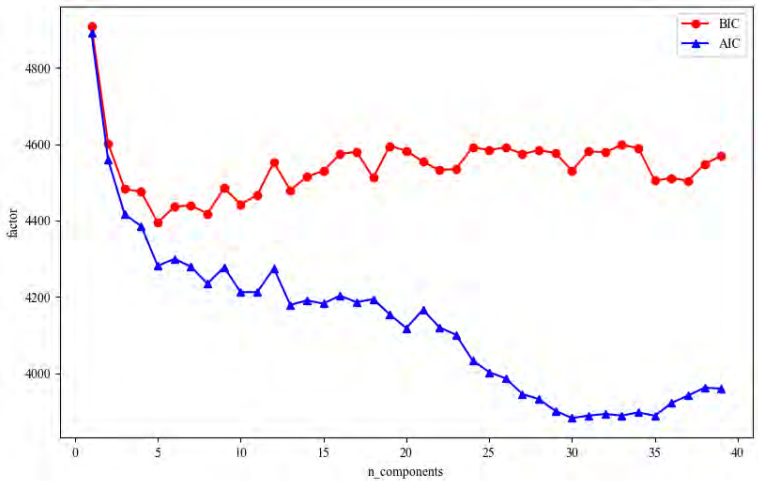


图 4-3 GMM 聚类 AIC/BIC 准则选取聚类数

通过上述方法的确定，对 354 个操作变量进行聚类，利用 GMM 及 EM 算法聚类成 30 类，由于聚类之后，操作变量的维数为 325 维，即 325 的样本特征，因此对于聚类结果的展示需要用到 t-SNE 降维可视化技术。t-SNE 是一种降维方法，主要用于高维数据的降维可视化展示，也可用于变量的降维，但在此处利用 t-SNE 是将变量的聚类效果及结果加以展现，由于每个变量已成为 325 维样本高维变量，需通过降维可视化技术加以站点，其聚类结果 2 维展示及 3 维展示如图 4-4 所示：

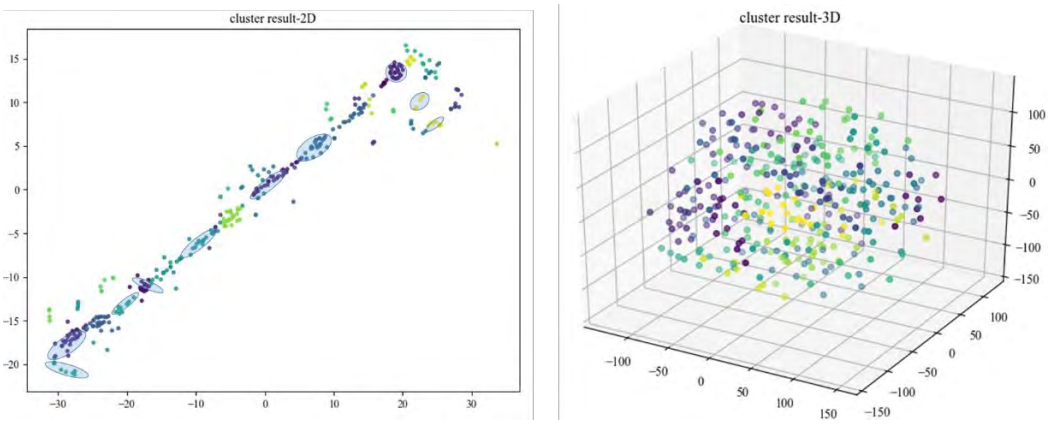


图 4-4 聚类结果可视化二维/三维展示

如下表 4-1 为 354 个操作变量各变量的聚类编号(0-29)，表示每个变量所属的聚类簇类别，其中操作变量均以位点名称中部英文字母代替：

表 4-1 354 个操作变量的聚类结果

位点	聚类簇	位点	聚类簇	位点	聚类簇	位点	聚类簇
FC_2801	0	AT-0009	4	PDI_1102	12	FC_2301	22
FT_9301	0	AT-0010	4	PC_6001	12	FT_1501	22
FT_5101	0	AT-0011	4	PT_6002	12	FT_9401	22
FT_9302	0	AT-0012	4	PC_1603	12	FT_5102	22
TE_6002	0	AT-0013	4	PC_2601	12	FT_2302	22
FT_3302	0	CAL.LINE	4	LT_3801	12	FT_1502	22

FT_1301	0	FT_1204	5	PT_2905	12	FT_5102	22
PT_7107	0	FT_1204	5	PT_2501	12	PC_2401B	23
PT_7103	0	TE_1601	6	PT_2502	12	TE_7508B	23
TE_7102	0	FT_3301	6	LT_1301	12	TE_7504B	23
FC_5001	0	TE_1103	6	FT_3501	12	TE_1504	23
FT_1006	0	TE_6001	6	PC_3001	12	PT_9301	24
FT_1504	0	FT_9102	7	PDT_3601	12	PT_9403	24
PDI_2102	1	FT_1004	7	FC_1104	12	PT_1501	24
PC_2105	1	FT_5201	7	FT_2803	12	PT_6009	24
PC_1301	1	FT_1501	7	PT_6003	12	PDI_2801	24
FC_1203	1	FT_1003	7	BS_AT_2401	12	PDI_2301	24
PT_1102	1	FT_9401	7	PT_2603	12	BS_AT_2402	24
TE_1501	1	FC_1101	7	PDT_2605	12	PC_2401	24
FT_2502	1	FT_5102	7	PT_2607	12	PC_2401B	24
FT_2433	1	FT_9102	7	PDT_3503	12	PC_2401B	24
TC_2201.OP	1	FT_1001	7	PDT_2906	12	PC_2401	24
LI_2107	1	FT_5204IZERA	7	PT_1604	12	TE_2103	25
PT_1101	1	TE_5102	8	PT_6005	12	TE_2005	25
PC_3501	1	TE_9001	8	PT_6008	12	FT_9202	25
PDI_2105	1	PDC_2502	8	FT_1202	13	TE_1608	25
FT_3702	1	TE_9003	8	FT_3304	13	TC_1606	25
PDT_2606	1	TE_9002	8	TC_2201	13	TE_2608	25
DT_2107	1	TE_3101	8	SIS_PT_6007	13	TE_2603	25
PDT_1003	1	TE_1502	8	TE_7508	13	TE_2104	25
PDT_1002	1	FT_2901	8	PT_7508	13	TE_2002	25
PDT_2409	1	TE_2901	8	TE_7506	13	TE_2004	25
PDT_3502	1	TE_2902	8	LC_5101	14	TE_2003	25
PDT_3002	1	FC_2432	8	FT_9001	14	TC_1607	25
PDI_2903	1	PDC_2702	8	FT_9403	14	TE_1605	25
PT_2106	1	LI_2104	8	FT_9201	14	TE_1604	25
PT_7505	1	BS_LT_2401	8	FT_9402	14	TE_1603	25
TE_7504	1	PDT_2001	8	TC_2607	14	TE_1602	25
PT_7502	1	TE_7506B	8	FT_9002	14	TXE_3202A	25
SIS_PDT_2103 A	1	TE_7502B	8	SIS_TE_2606	14	TXE_3201A	25
PT_2106	1	TE_7102B	8	SIS_TE_2605	14	TXE_2203A	25
CAL.SPEED	1	FC_2432	8	TE_2604	14	TXE_2202A	25
LT_9001	2	TE_5009	8	PT_2901	14	TE_1102	25
LT_3101	2	TE_1503	8	FT_5204	14	TE_2104	25
LT_1501	2	CAL.CANGLIANG	8	SIS_PDT_210 3B	15	TE_1102	25
LT_2101	2	TC_2101	9	FT_9101	15	FC_1202	26
LT_1002	2	TE_2301	9	FT_9301	15	FC_2702	26
PDT_2703B	2	TE_9301	9	FT_9302	15	FT_2303	27

PDI_2501	2	FT_3301	9	FT_2002	15	FT_2001	27
PT_6006	2	AT_1001	9	TE_6001	15	PDT_1004	27
PC_1001A	2	TC_2801	9	TE_1103	15	TEX_3103A	27
TC_5005	3	FC_3101	9	TE_1104	15	AT-0001	27
FC_5202	3	TE_2601	9	FT_1503	15	AT-0002	27
FT_1001	3	TE_1104	9	PDT_2503	16	PDT_2104	28
FC_1005	3	FT_3303	9	FT_1002	17	LC_5001	28
FC_1101	3	TC_2702	9	CAL.LEVEL	17	TE_5202	28
SIS_TE_6010	3	TE_2501	9	FT_9001	18	FC_2501	28
FC_1201	3	TE_2401	9	FT_5101	18	FT_1003	28
TE_1201	3	SIS_TE_2802	9	FT_9201	18	FT_1004	28
FT_5201	3	FT_2431	9	FT_9202	18	TE_1001	28
TE_1101	3	FT_3201	9	FT_9402	18	TE_1105	28
TE_1107	3	AT_1001	9	FT_9403	18	LC_1201	28
TE_1106	3	SIS_TE_6009	9	TE_2001	18	TE_1203	28
TE_5002	3	TE_6008	9	FT_1006IZERA	18	LC_1202	28
TE_5006	3	SIS_FT_3202	9	FT_1503IZERA	18	FC_2601	28
HIC_2533	3	TC_3203	9	FT_1504IZERA	18	PDT_2604	28
TE_5007	3	TC_3102	9	SIS_LT_1001	19	LC_5002	28
TE_1106	3	TE_6008	9	FC_1102	19	LC_3301	28
TE_1107	3	PT_2801	10	FT_1204	19	LC_1203	28
TE_1101	3	PT_2101	10	FC_1202	19	FT_3001	28
PC_5101	4	PT_2301	10	SIS_PT_2703	19	TE_5004	28
AT_5201	4	AC_6001	10	FC_5203	19	TE_5003	28
PT_9001	4	PT_1201	10	FC_5103	19	TE_5101	28
PT_9402	4	PC_1202	10	PT_7510	19	SIS_PT_2602	28
PT_9401	4	PT_1103	10	TE_3111	19	TE_5001	28
PT_1602A	4	LT_9101	10	PDC_2607	20	LC_2601	28
FC_3103	4	PT_7107B	10	LT_2901	20	DT_2001	28
PC_9002	4	PT_7103B	10	AT_6201	20	SIS_TEX_3103B	28
PC_3101	4	PT_1601	10	PT_7503	20	TE_3112	28
PC_3301	4	CAL_1.CANGLIANG	10	TE_7106	20	TE_5008	28
PDT_3602	4	LI_9102	11	FT_5104	21	LC_1203	28
PT_5201	4	FT_2701	11	FT_1002	21	RXL_0001	28
PC_2902	4	ZT_2533	11	PDI_2703A	21	AI_2903	29
AT-0003	4	TE_5201	11	FT_5104	21	PT_7510B	29
AT-0004	4	FT_3701	11	LC_5102	21	PT_7508B	29
AT-0005	4	ZT_2634	11	PDT_2704	21	PT_7505B	29
AT-0006	4	PC_2401	11	TE_7106B	21	PT_7503B	29

AT-0007	4	CAL_H2	12	TE_7108B	21
AT-0008	4	FT_9101	12	LC_5102	21

4.3 基于信息增益特征筛选

在上述通过高斯混合模型聚类的结果中，可认为在一类中的变量在样本数据上的表现是相似的，因此在此基础上，对已分好的类别中，从每类中筛选具有代表性的实际特征。

4.3.1 信息增益

信息增益采用决策树实现，具有良好的分类性能。该方法是一种基于熵的度量方法，并广泛应用于特征筛选领域。作为一种统计方法，信息增益根据特征和类别之间的相关性来分配特征权重。对于一个数据集， $S(s_1,s_2,...,s_n)$ 是 n 个实例的集合，$A(A_1,A_2,...,A_p)$ 是 p 个属性的集合，对于分类系统来说，$C(c_1,c_2,...,c_m)$ 是 m 个类别标签的集合， $p(c_i)$ 表示 S 中第 i 个类别标签 $c_i(i=1,2,...,\ m)$ 所占的比例，数据集的熵由以下公式给出：

$$\mathrm{H}(C)=-\sum_{i=1}^{\mathrm{m}}p\left(c_i\right)\log_2\left(p\left(c_i\right)\right) \tag{4-4}$$

信息增益(IG)是针对数据分类系统中的每一个特征的， $A_q\left(a_{q1},a_{q2},\ldots a_{\mathrm{qk}}\right)$ 表示数据集第 $q(q=1,2,...,p)$ 个属性，对于属性 $A_q\left(a_{q1},a_{q2},\ldots a_{\mathrm{qk}}\right)$，对应的条件熵为:

$$H\left(C|\ \mathrm{A}_q\right)=-\sum_{j=1}^{k}p\left(a_{qj}\right)\sum_{i=1}^{\mathrm{m}}p\left(c_i|\ a_{qj}\right)\log_2\left(p\left(c_i|\ a_{qj}\right)\right) \tag{4-5}$$

a_{qj} 是 A_q 属性的值，具有 k 种取值， $p\left(a_{qj}\right)$ 表示分类变量 C 的先验概率， $p\left(c_i|\ a_{qj}\right)$p 表示属性 A_q 确定条件下变量 C 的条件概率。另外， $H(C)$、$H(C|A_q)$ 之间的差值根据下面的公式来计算属性 A_q 的信息增益值。

$$IG\left(A_q\right)=\mathrm{H}(C)-H\left(C|\ \mathrm{A}_q\right) \tag{4-6}$$

4.3.2 基于信息增益排序及斯皮尔曼相关系数的变量筛选

在对该汽油精制催化裂化过程中，根据辛烷值的损失可以计算每个类别中的对于辛烷值损失影响的信息增益，由于辛烷值损失的变化幅度较小，在 325 个样本中，辛烷值的损失处于 0.2-1.82 的范围内，以 0.01 为间隔对辛烷值损失进行离散化则可用于计算全部变量（包括原料性质、产品性质、待生吸附剂性质、再生吸附剂性质 12 个变量及 354 个操作变量进行）信息增益的计算。在筛选特征过程中，由于通过前述的 GMM 模型将 354 个操作变量聚类，每一类中的数据分布较为相似，可以认为在每一类中对于辛烷值损失具有近似的信息增益，在此基础上对每一类中所有变量的信息增益做计算，由此，得到原料性质、产品性质、待生吸附剂性质、再生吸附剂性质 12 个变量及 30 聚类中每类

的最高信息增益操作变量共 42 个变量进行筛选，在此基础上进行信息增益排序。另外对该 42 个操作变量的斯皮尔曼相关系数矩阵及信息增益排序结果两个标准进行筛选，剔除相关性较大的变量。

本文采用斯皮尔曼相关系数对两个变量之间的关联程度进行表示，斯皮尔曼相关系数计算公式如下式所示：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{4-7}$$

其中，x_i，y_i 代表秩次，则根据上式 42 个变量的筛选结果进行计算如下图 4-5，为方便观察，计算出的相关系数均取绝对值：

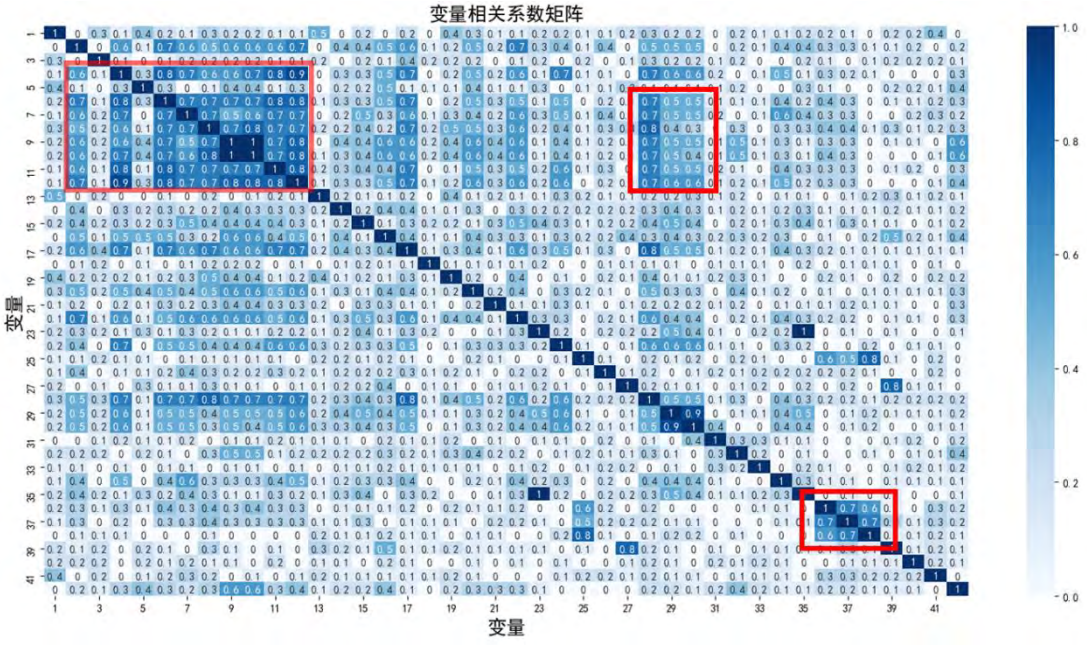


图 4-5 综合 42 各变量的相关系数矩阵

可以看到变量间基本的相关关系颜色较浅，大部分变量线性相关关系不大，几个变量间的相关关系较大颜色突出，主要为 R102 再生器提升氮气流量、D104 去稳定塔流量、蒸汽进装置压力等指标，综合信息增益排序及相关系数的指标，筛选了 28 个变量，其中有 3 个位性质变量，分别为产品性质-硫含量、原料性质辛烷值 RON

最终得到的 28 个主要变量及其信息增益值如下表所示，限于篇幅只展示最终筛选出的 28 个变量的信息增益值，为明确变量名称，筛选的主要变量以其中文名称表示：

表 4-2 28 个主要变量及其信息增益值

变量名称	信息增益值	变量名称	信息增益值
E-101D 壳程出口管温度	0.036858	产品性质-硫含量 μg/g	0.014713
还原器温度	0.031642	过滤器 ME-101 出口温度	0.014478
精制汽油出装置温度	0.030301	A-202A/B 出口总管温度	0.013757
K-103A 进气温度	0.025092	D-101 脱水包液位	0.012878
D121 去稳定塔流量	0.024861	再生器顶底差压	0.012459
还原器流化氢气流量	0.021391	2#催化汽油进装置流量	0.012238
精制汽油出装置硫含量	0.020547	燃料气进装置压力	0.012092

R-101 下部床层压降	0.019506	C-201 下部进料管温度	0.011762
反应器质量空速	0.018731	原料性质辛烷值 RON	0.011319
由 PDI2104 计算出	0.018077	反吹氢气温度	0.011203
加热炉主火嘴瓦斯入口压力	0.017812	待生吸附剂性质-S wt%	0.01095
D101 原料缓冲罐压力	0.017221	8.0MPa 氢气至循环氢压缩机入口	0.01013
加氢裂化轻石油进装置流量	0.016543	加热炉进口温度	0.009496
冷氮气过滤器 ME-114 差压	0.016427	EH-101 加热元件温度	0.009221

该 28 个变量的相关系数矩阵如下图所示 4-6 所示：

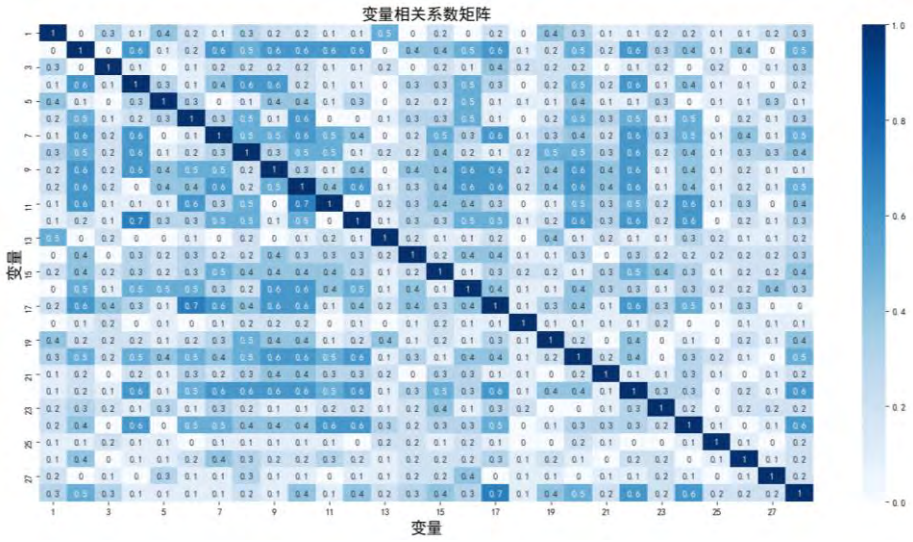


图 4-6 筛选出 28 个变量的相关系数矩阵

由图中看出，大部分变量之间的相关系数都较小，颜色较浅的的色块占绝大多数，证明筛选变量之间的独立性较强。

综上所述，通过基于 EM 算法的 GMM 模型对 354 个操作变量聚类，并且将每类中的变量及性质变量利用信息增益理论计算其对于辛烷值损失的信息增益，并通过信息增益排序及斯皮尔曼相关系数的筛选，最终筛选出 28 个具有代表性独立性的变量。

5.问题三分析与求解

5.1 问题三分析

通过上述问题二对变量的分析与筛选，主要通过 GMM 聚类及信息增益排序筛选出了共计 28 个具有代表性、独立性的变量（包括原料辛烷值、产品硫含量及待生吸附剂性质 S 等），并对变量加以展示说明其筛选的优越性。

在问题三中，需要针对 325 个样本数据，利用上述筛选的 28 个变量预测每个样本的辛烷值损失。由于样本数据具有代表性，且各样本均是不同时间下采集的计算结果，根据本文的样本数据及筛选变量的特点，结合实际汽油精制化学工艺流程并查阅相关文献，建立了基于慢特征分析的即时学习（SFA-JITL），该模型在预测预测流程有如下步骤：

（1）在筛选出主要变量的基础上，对筛选出的主要变量进行慢特征分析，共筛选出 7 维慢特征；

（2）搭建即时学习框架模型，对 325 个样本的顺序输入利用即时学习框架中的局部建模方法不断对新增样本预测辛烷值损失，并在此过程中利用局部慢特征权重分析对慢特征权重进行调整，已达到良好的预测效果。

（3）利用已有搭建模型并进行预测，采用预测指标，即预测值与真实值之间的差距对模型的有效性进行检验。

综上，问题三预测模型建模过程的主要技术路线如下图 5-1 所示：

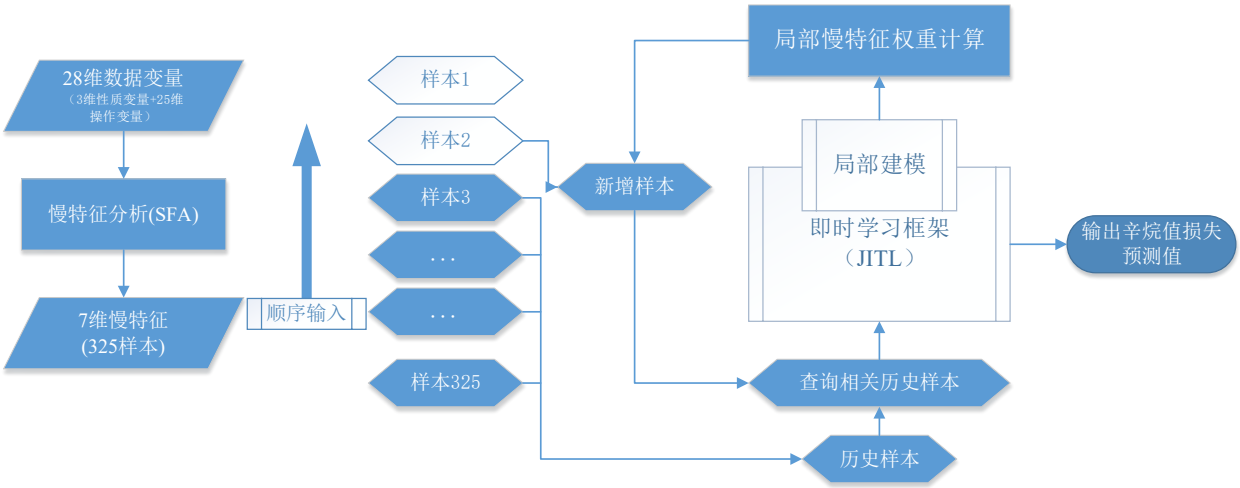


图 5-1 辛烷值损失预测模型技术路线

5.2 SFA-JITL 预测模型原理及框架

5.2.1 慢特征分析原理

慢特征分析（Slow Feature Analysis, SFA）是由 Hinton[5]于 1989 年首次给出基本概念，由 Laurenz Wiskott 于 2002 年首次提出算法框架的一种无监督学习算法[6]。在大数据中，缓慢的变化通常代表系统的固有特征，而快速的变化常常暗示数据噪声，SFA 能够通过抽象出系统缓慢的变化，将工业过程中最具本质的特征信息提取出来，非常适合用于流程工业这类场合。近年来，慢特征分析算法在工业领域取得较好进展，比如，黄健[7]等针对工业过程中的故障检测问题，采用慢特征分析算法提取过程的本质动态特征，并在数值系统和 Tennese Eastman 过程进行仿真验证，证实了基于慢特征分析算法的模型效

果优于主成分分析法；王鹤莹[8]等针对初馏塔这一类工业过程的故障问题，采用慢特征分析技术，从流程数据中提取特征信息用于实时监测。

SFA 的基本原理概括如下：

给定 L 维输入序列 $\delta=\left[\delta_1,\delta_2,\cdots,\delta_n\right]$，通过求解如下优化问题，使输出序列（即慢特征）$s=\left[s_1,s_2,\cdots,s_n\right]$ 变化缓慢：

$$\begin{aligned}&\arg\min_s \sum_{i=2}^{n}\left\|s_i^{'}\right\|_2^2\\&s.t.\quad SS^T=I_l\end{aligned}\tag{5-1}$$

其中：$s_i^{'}=s_i-s_{i-1}$，I_l 代表一个单位对角矩阵。

为了进一步简化上述优化问题，假设慢特征（记作 SF ）与输入数据具有一定的线性关系，并将该关系式表示为

$$S=\wp^T\tag{5-2}$$

其中，$\wp$是一个由特征向量组成的映射矩阵，由此，上式等价于：

$$\sum_{i=2}^{n}\left\|s_i^{'}\right\|_2^2=\operatorname{trace}\left(\dot{S}\dot{S}^T\right)=\operatorname{trace}\left(\wp^T\left(\dot{\delta}\dot{\delta}^T\right)\wp\mathcal{D}\right)\tag{5-3}$$

其中, $\operatorname{trace}(\bullet)$ 是跟踪运算符；　$\dot{S}=\left[s_2^{'},\cdots,s_n^{'}\right]$；　$\dot{\delta}=\left[\dot{\delta}_2,\cdots,\dot{\delta}_n\right]$

此外，必须保证：

$$\dot{\delta}\dot{\delta}^T=I_L\tag{5-4}$$

使用 SFA 算法进行计算时，第 1 步是进行白化处理以消除各个变量之间的相关关系。使用奇异值分解实现该操作：

$$SS^T=\wp^TSS^T\wp=\wp^T\wp=I_l\tag{5-5}$$

因此，矩阵 $\wp$ 是正交矩阵。至此，可以得到，SFA 算法的目标是寻找正交矩阵$\wp$，使目标函数最小化。综上，原本复杂的全局优化问题简化为一个简单的广义特征分解问题：

$$\begin{aligned}&\arg\min\operatorname{trace}\left(\wp^T\left(\dot{S}\dot{S}^T\right)\wp\right)\\&s.t.\quad \wp\,\wp^T=I_L\end{aligned}\tag{5-6}$$

其中，根据特征向量的定义，它可以由特征值矩阵（ Λ ）线性表示为：

$$\left(\dot{\delta}\dot{\delta}^T\right)\wp=\Lambda\wp\tag{5-7}$$

综上，慢特征分析的直观原理图可用图 5-2 表示：

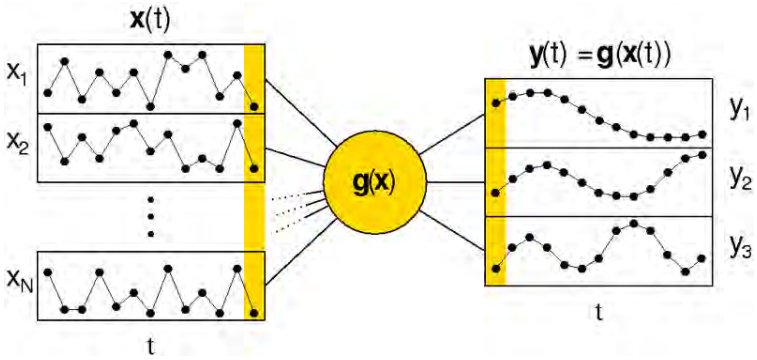


图 5-2 慢特征原理效果图

5.2.2 即时学习（JITL）

即时学习是一个最近开发的非线性过程建模策略，它建立在数据库和局部建模技术的基础之上。传统的全局建模方法是参数化模型，其模型与训练数据一起离线训练并用于在线预测，与之不同的是，即时学习建模策略会自主选择最相关的历史样本来构建在线局部线性模型，能够实现这一特性的理论基础是复杂的非线性模型可以通过一系列简单的线性模型来近似。因此，即时学习建模策略可以很好地描述过程特性的变化。通常，在即时学习模型框架中执行 4 个步骤：1）每当查询到新样本时，设计一个相似标准从历史数据集中选择最相关的样本用于局部模型训练；2）使用所选择的相关样本围绕查询样本构建和训练局部模型；3）使用训练模型为查询样本预测输出变量 4）预测完成后立即丢弃局部模型，并延迟新的局部模型直到下一个查询样本到来。图 1 给出了传统全局建模和即时学习建模的示意图。

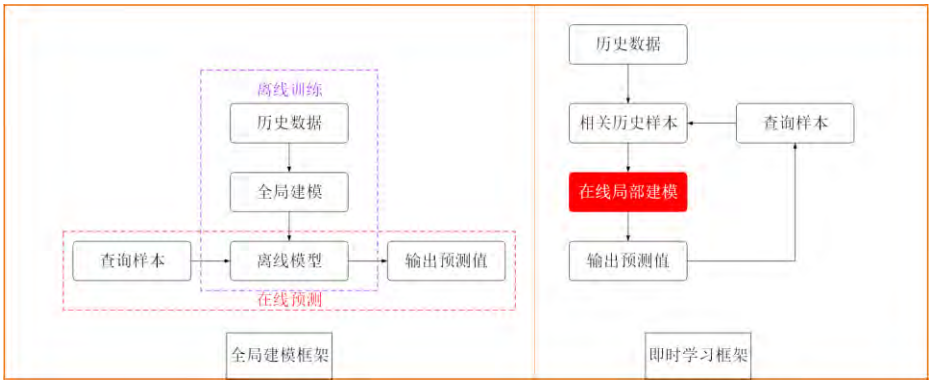


图 5-3 传统全局建模和即时学习框架对比示意图

即时学习建模策略中最关键的一步是从历史数据集中选择最相关的样本用于局部模型训练，通常采用测量查询样本和历史样本之间的某种“距离”的方法进行选择。“距离”用于指示应为哪些历史样本分配优先级以构造当前查询样本的局部模型。一般而言，较大的距离通常表示历史样本与查询样本的相似性较低，反之亦然。在计算所有历史样本的权重之后，可以按降序对它们进行排序，并选择具有最大权重的样本作为相关的局部模型样本。一旦为查询样本选择了相关样本，就可以将在线局部模型构建为 $y_i = f\left(x_i\right)$。

此外，局部建模算法有很多选择，在本次汽油催化裂化的模型建模过程中我们采用的在线局部模型为线性回归模型。即时学习框架可以将复杂的预测非线性建模转换为局部建模中的线性模型，通过对不断加入的新的查询样本进行分析预测，从而给出整个模型样本的辛烷损失预测值。

5.2.3 局部加权慢特征分析

尽管 5.2.1 中提到的常规的 SFA 算法在工业工程类相关预测中很有效，但是仍然存在一些需要解决的问题：

（1）首先，常规的 SFA 算法为线性算法，不能详尽全面地描述特征数据信息之间的非线性关系。而在本文研究的问题中，由于工艺过程的复杂性以及设备的多样性，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，因此，常规的 SFA 算法不能满足该类工业要求。

（2）其次，根据样本的分布差异，所选样本对查询向量的重要性不同。 但是，传统的即时学习预测的框架将所有选定的样本均等地对待，并为其分配相同的权重，这可能无法有效地从数据中提取非线性信息。

为了解决上述问题，接下来将线性 SFA 拓展为非线性形式，由此提出了一种基于局部加慢行特征分析算法。

（1）定义我们已经挑选出的m个特征变量所对应的输入矩阵为δ_{selected}，定义输出矩阵为Y_{selected}，定义一个新的变量δ_{new}，同时定义一个相似度对角矩阵变量Ψ：

$$\Psi = \text{diag}\left\{\Psi_1, \Psi_2, \cdots, \Psi_m\right\} \tag{5-8}$$

其中，$\Psi_i (i = 1,2,...,m)$ 代表δ_{selected}和δ_{new}之间的相似度，它是根据相应的潜在分布间的不相似性定义的，$std(\bullet)$是标准差：

$$\Psi_i = \exp\left(-\frac{SKL\left(\delta^i_{\text{selected}}, \delta_{\text{new}}\right)}{\mu \cdot \text{std}\left(\text{SKL}\left(\delta^i_{\text{selected}}, \delta_{\text{new}}\right)\right)} \right) \tag{5-9}$$

μ是一个可变常数，当μ趋近于正无穷时，已选定的特征变量的权重均接近 1，此时为常规的线性 SFA 方法；通过训练，不断改变μ的取值，我们可以得到不同的权重矩阵$\wp$，进而确定最优权重。训练方法如下：

$$\delta_{\wp} = \frac{\sum_{i=1}^{m} \Psi_i \delta^i_{\text{selected}}}{\sum_{i=1}^{m} \Psi_i} \tag{5-10}$$

$$Y_{\wp 0} = \frac{\sum_{i=1}^{m} \Psi_i Y^i_{\text{selected}}}{\sum_{i=1}^{m} \Psi_i} \tag{5-11}$$

（2）构建局部加权慢特征分析算法的前提是所有的变量都已进行加权平均处理，因此，我们首先对用于训练参数的样本变量做加权平均处理：

$$\overline{\delta_{\text{selected}}} = \delta_{\text{selected}} - I_{m*1}\delta_{\delta} \tag{5-12}$$

$$\overline{\delta_{\text{new}}} = \delta_{\text{new}} - \delta_{\wp} \tag{5-13}$$

$$\overline{Y_{\text{selected}}} = Y_{\text{selected}} - I_{m*1}Y_{\wp} \tag{5-14}$$

（3）对所有训练样本的相关变量进行加权处理后，通过奇异值分解原理（SVD），对其协方差矩阵S进行特征值分解，得到U 和D：

$$S = \frac{1}{m-1}\left(\Psi_1 \cdot \overline{\delta_{\text{selected}}}\right)^T \Psi_1 \cdot \overline{\delta_{\text{selected}}} \tag{5-15}$$

$$S = UDU^T \tag{5-16}$$

（4）进行白处理，令$Q = D^{-\frac{1}{2}}U^T$，定义$Z = \Psi_1 \cdot \delta_{\text{selected}}$，那么协方差矩阵$S$等价于：

$$S = \frac{1}{m-2}Q^T\left(\Psi_1 \cdot \overline{\delta_{\text{selected}}}\right)^T \Psi_1 \cdot \overline{\delta_{\text{selected}}}\,\overline{Q} = \frac{1}{m}\dot{Z}^T\dot{Z} = P^T\Lambda P \tag{5-17}$$

（5）此时，使得目标函数最小的正交矩阵P已得到，由此可以计算目标矩阵$\wp$，即：

$$\wp = PQ = PD^{-\frac{1}{2}}U^T \tag{5-18}$$

（6）最终，可以计算局部加权慢特征SF：

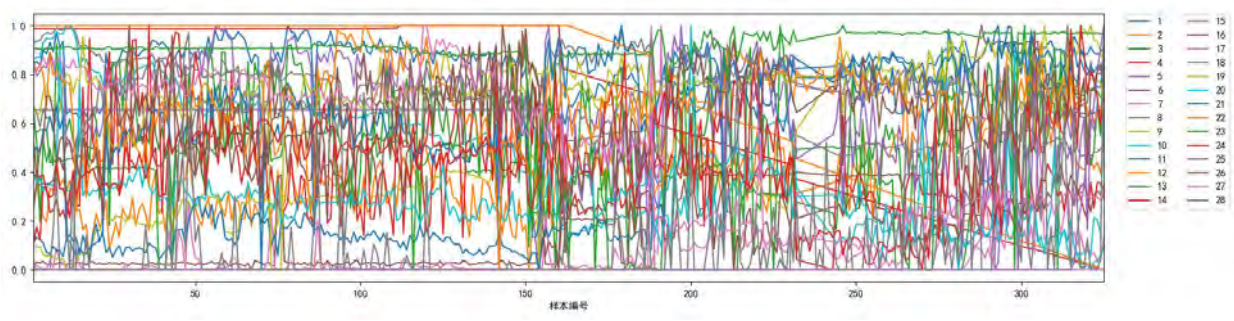
$$SF = \overline{\delta}_{\text{selected}} \cdot \wp \tag{5-19}$$

由此，可以将原始数据映射为多个按照缓慢程度排列的慢特征，迅速变化的数据常含有大量噪声，而在多个样本中变化幅度较小的特征常被视为本质特征。

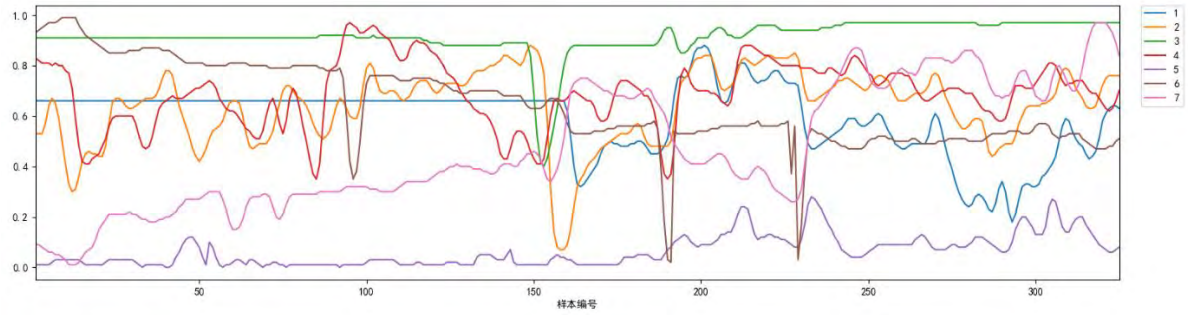
5.3 SFA-JITL 辛烷值损失预测模型

本文基于慢特征分析的即时学习预测模型框架基于 Python 程序编写，分别为慢特征选取部分，局部加权慢特征分析以及即时学习框架搭建。

首先，在慢特征分析选取阶段，首先将筛选出的 28 维变量进行归一化，并且按照样本的实际采样计算时间进行排序，应用相应的慢特征分析公式，可得如下图的处理结果：



(a)28 维变量归一化分布



(b) 7 维慢特征

图 5-4　慢特征分析特征变化图

由图 5-3 所示，相比于之前的特征的杂乱无章、震荡性以及特征冗余，经过慢特征变化后形成的 7 维慢特征具有明显的分层，各特征分管不同的区域属性，在归一化表示上，数值在高低数值上的分布较为均匀，并且在时序样本属性上的表示更为平滑，能够更好地为之后的即时学习框架提供良好的数据基础。

根据 5.2.2 及 5.2.3 的内容，搭建的即时学习框架并利用局部加权特征对筛选的慢特征进行训练权重，在即时学习框架中，每一个样本是按顺序输入的，并不是一次性输入所有样本，因此这种框架的预测反而更好地适用于实际工程使用，通过不断新加入的样本对预测模型进行不断的改进，并且对慢特征等权重也会不断地根据新加入的样本不断调整。综上，即时学习框架会根据其模型框架如图 5-4 所示：

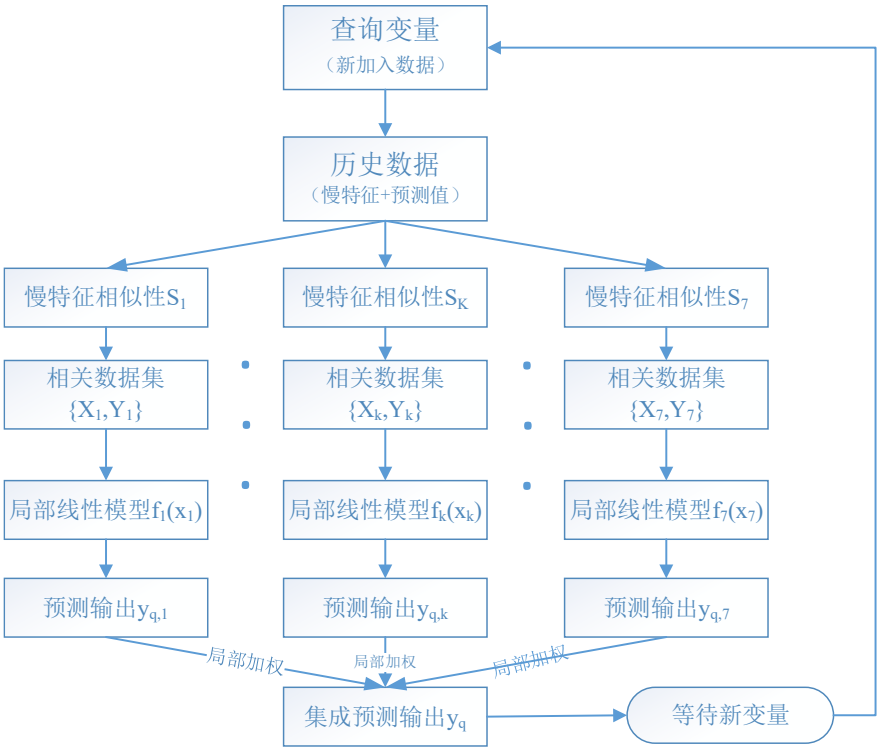


图 5-5 SFA-JITL 辛烷值模型预测框架

根据上述 SFA-JITL 辛烷值损失预测模型，对 325 个样本的辛烷值进行预测，其结果与真实值的差距如图 5-5 所示：

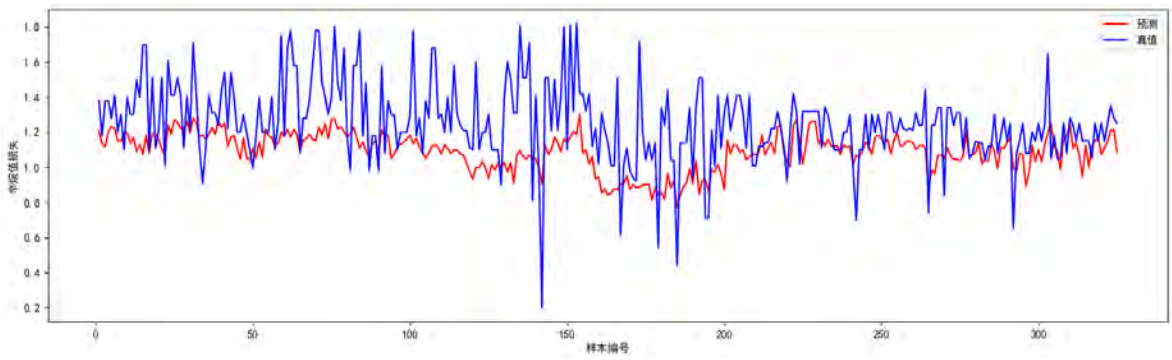


图 5-6 辛烷值损失预测结果

本文共选取了三个指标来评价辛烷值预测结果并说明模型的有效性，模型主要指标有 RMSE(Root Mean Square Error,均方根误差)，MAE(Mean Absolute Error ，平均绝对误差), MAPE(Mean Absolute Percentage Error 平均绝对百分比误差），各指标的主要计算公式如下所示：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{5-20}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{5-21}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{5-22}$$

根据上述的预测指标结果，我们将该模型与线性回归模型、卡尔曼滤波等基本方法进行比较，模型的主要指标对比如下表所示

表 5-1 不同模型的预测指标结果

预测模型	评价指标		
	RMSE	MAE	MAPE/%
线性回归	0.979	0.649	88.359
卡尔曼滤波	0.534	0.391	25.542
SFA-JITL	0.261	0.199	16.124

由上述结果表明，SFA-JITL 辛烷值损失预测模型具有良好的优越性及有效性，能够对辛烷值损失预测进行良好的建模。另外，通过与真实值的比较以及与其他基础模型预测结果的比较，说明了模型的有效性，并且其本身模型特性：例如不断地更新查询历史样本，通过众多变量提取出慢特征等，也十分符合实际工艺流程的使用，其本身模型框架也在化工工艺流程中开始得到重视并使用。

6. 问题四分析与求解

6.1 问题四分析

在本题中，在保证产品硫含量不大于$5\mu g/g$的前提下，需要挑选预测结果中降幅大于30%的样本，并阐述其所对应的建模主要变量经过优化后的操作条件。

经过基于慢特征分析的即时学习预测模型预测后，我们得到了325个数据样本的辛烷值损失预测值。显而易见地，各数据样本优化效果有所不同。然而，来源于某石化企业的这325个数据样本，其汽油产品辛烷值损失平均为1.37个单位，而同类装置的最小损失值只有0.6个单位，还有56.21%的优化空间。因此，我们挑选出降幅大于30%的样本，并详细分析所对应的主要变量经过优化后的操作条件，采取求均值的方法，获得最大可能适用于所有样本的操作条件，进而为石化企业的长期、高效运营提供科学稳健的理论支撑。

6.2 挑选结果分析

首先，我们定义降幅Δ为：

$$\Delta = \frac{\left|R_1^i - R_0^i\right|}{R_0^i} \times 100\% \tag{6-1}$$

其中，R_0^i代表第i个数据样本原来的辛烷值损失，R_1^i代表第i个数据样本的辛烷值预测损失。

在基于慢特征分析的即时学习预测模型得到的预测结果中，共有56个降幅大于30%的样本，分别是第13,15,16,21,31,43,59,61,62,64,69,70,71,76,79,83,84,86,101,107,108,114,121,131,132, 133,135,137,138,149,151,153,161,166, 173,180,181,182,192,193,200,204,205,208,225, 264,266,267,268,269,272,274,275,296,303, 314。将这56个数据样本重新编号为1至56，比如，原来的挑出的第一个样本（原编号13）为以下分析中的样本1，原来的挑出的第二个样本（原编号15）为以下分布的样本2，依次类推。最高降幅48.3797%，其中有6个样本降幅大于40%，15个样本降幅介于35%至40%之间，如图1所示。该56个样本的辛烷值损失降幅情况如图2所示。具体来看，辛烷值、待生吸附性、硫含量和辛烷值损失的操作条件如图3所示，所有建模主变量，即慢特征的操作条件见上传附件。

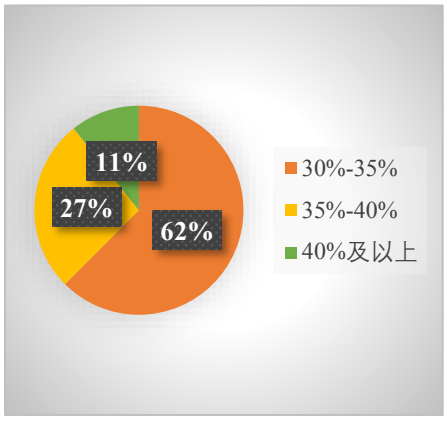


图 6-1 325样本辛烷值损失降幅占比图

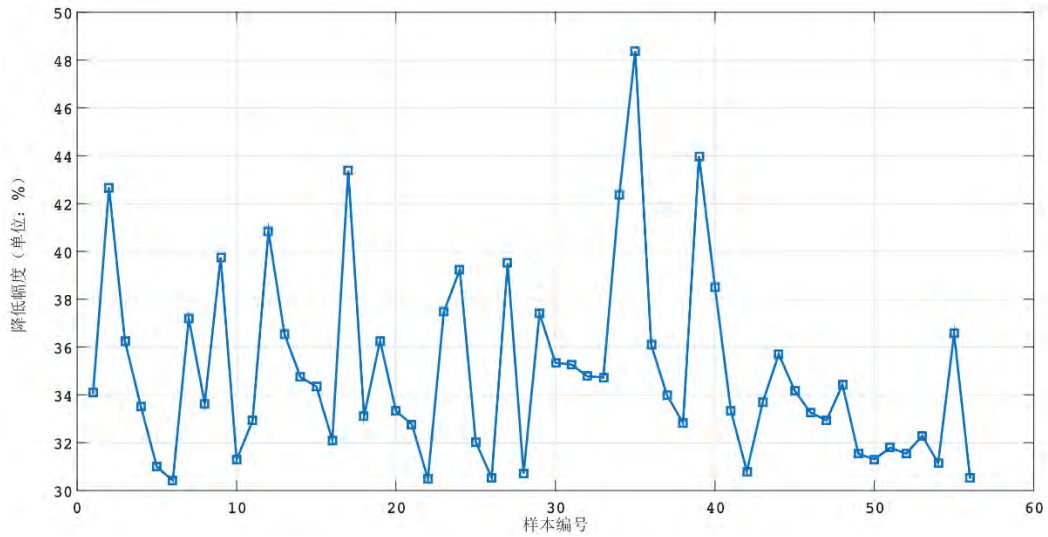
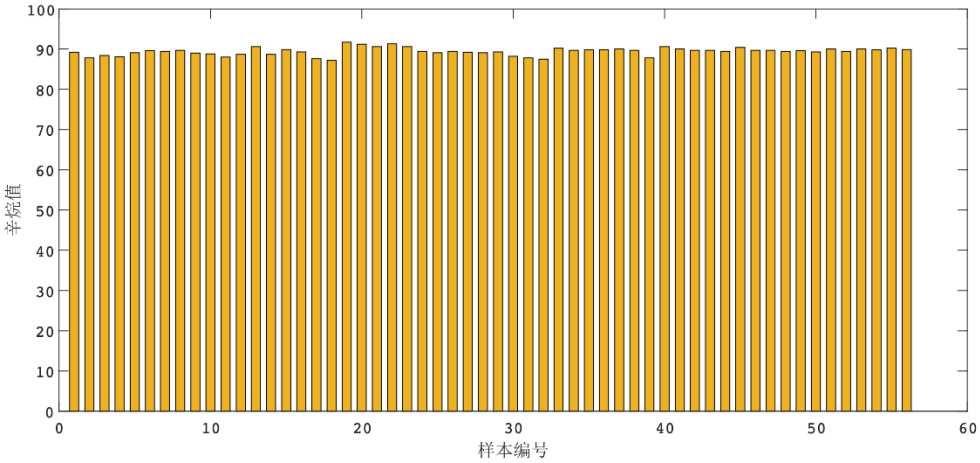
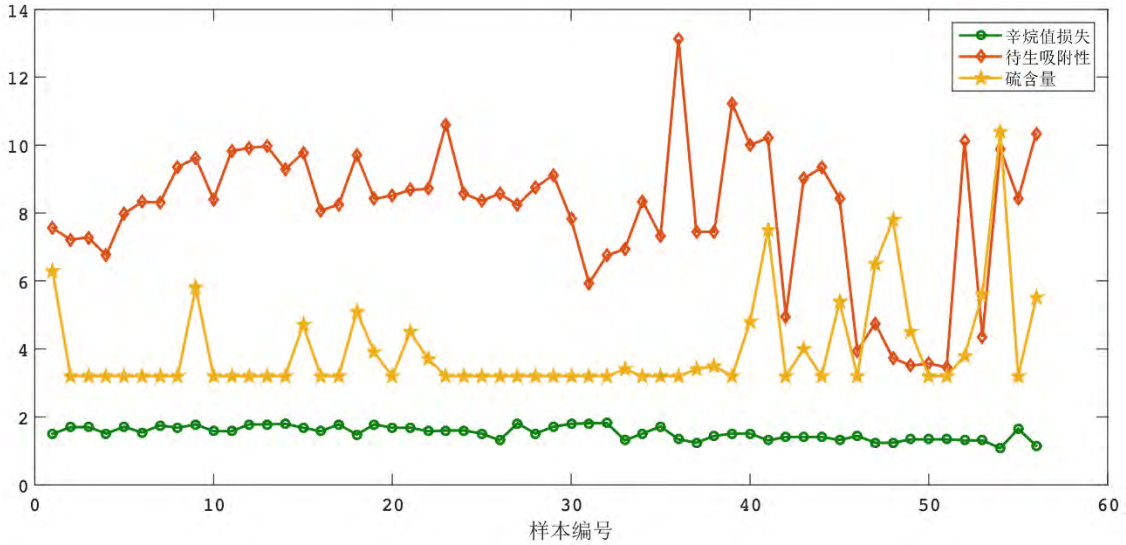


图 6-2 辛烷值损失降幅图



（a）辛烷值



(b)待生吸附性、硫含量和辛烷值损失

图 6-3 几个主要变量操作条件图

6.3 最终操作条件

由 6.2 可以看出，每个数据样本的 28 个特征值都不尽相同。为了给石化企业提供稳健的操作标准，我们需要寻找一个普适性的操作条件。因此，本文将以上 56 个数据样本的 28 个本质特征分别求取平均值，并将其做为优化后的最终操作条件，结果如表 1 所示。

表 6-1 主变量优化后的操作条件

变量	E-101D 壳程出口管温度	还原器温度	精制汽油出装置温度
取值	119.3667892	283.6317378	35.11150791
变量	K-103A 进气温度	D121 去稳定塔流量	还原器流化氢气流量
取值	-3697.838594	10.82371756	696.1775144
变量	精制汽油出装置硫含量	R-101 下部床层压降	反应器质量空速
取值	1.331756534	60.83005655	4.638538
变量	由 PDI2104 计算出	加热炉主火嘴瓦斯入口压力	D101 原料缓冲罐压力
取值	29.53255	0.077542	-23.9725
变量	加氢裂化轻石脑油进装置流量	冷氮气过滤器 ME-114 差压	过滤器 ME-101 出口温度
取值	4498.37	0.371231	426.8711
变量	A-202A/B 出口总管温度	D-101 脱水包液位	再生器顶底差压
取值	53.41813	-4.96842	37.8867
变量	2#催化汽油进装置流量	燃料气进装置压力	C-201 下部进料管温度
取值	51.66669	0.411233	124.5893
变量	原料性质-辛烷值 RON	反吹氢气温度	待生吸附性质
取值	89.41786	239.1042	8.080266
变量	8.0MPa 氢气至循环氢压缩机入口	加热炉进口温度	EH-101 加热元件温度
取值	2178.883	364.5616	427.3214
变量	产品性质-硫含量 μg/g		
取值	427.3214		

7. 问题五分析与求解

7.1 问题五分析

由题目五的要求可知，各主要操作变量每次允许调整幅度值均为某个固定值 Δ，即每次调整各主要操作变量只能在上一步的基础上增加或降低相应的固定值 Δ。本题主要根据问题四中提供的优化方案，对 133 号样本的各主要操作变量进行逐步的迭代，对模型进行逐步的优化，最终将主要操作变量迭代至问题四中得到的优化方案上。考虑到该工业过程为连续的操作过程，可以假设在每一步调整各主要操作变量后，最终产品的辛烷值和硫含量都会及时地发生变化，即每组主要操作变量都对应着该条件下相应产品辛烷值和硫含量值，这样便可以得到每一步迭代过程中在当前步操作变量的设置下产品的辛烷值、辛烷值损失和硫含量值数据，于是便可以得到在整个迭代优化过程中对应的成品汽油的辛烷值和硫含量的变化轨迹曲线，以下给出了具体求解过程。

7.2 问题五求解

对 133 号样本的各操作变量逐步进行调整，直到调整至问题四中得到的最终的所有操作变量最优值附近。整体的调整方案如下：

计算当前第 i 次迭代步下每个主要操作变量 $Z_n^{(i)}$ 与最优值 Z_n^o 之间的差值，根据插值的具体情况，我们可以得到在第 $i+1$ 次迭代下主要操作变量 $Z_n^{(i+1)}$ 为：

$$Z_n^{(i+1)}=\begin{Bmatrix} Z_n^{(i)}+|\Delta_n|,\ if\ Z_n^{(i)}-Z_n^{(i)}\le -|\Delta_n| \\ Z_n^{(i)},\ if\ -|\Delta_n|<Z_n^{(i)}-Z_n^{(i)}<|\Delta_n| \\ Z_n^{(i)}-|\Delta_n|,\ if\ Z_n^{(i)}-Z_n^{(i)}\ge |\Delta_n| \end{Bmatrix} \tag{7-1}$$

同时我们设定每次迭代，所有的主要操作变量都要进行调整更新，直到所有的主要操作变量都满足迭代终止条件 $-|\Delta_n|<Z_n^{(i)}-Z_n^{(i)}<|\Delta_n|$，调整优化过程中，若某个变量已经满足迭代终止条件，则不再进行调整。假设第 n 个主要操作变量的最大迭代步数为 $step_n$，这样可以计算出整个调整优化过程的最大迭代步数 $step$ 为：

$$step=\max_n\{step_n\}=\max_n\left\{\left\lfloor\frac{\left|Z_n^{(0)}-Z_n^{o}\right|}{|\Delta_n|}\right\rfloor\right\} \tag{7-2}$$

对于 133 号样本，计算出所需的最大迭代步数 $step=3910$，所有主要操作变量需要迭代调整优化需要的步数整理为下表 5-1 所示。

表 7-1：所有主要操作变量需要迭代调整优化需要的步数

变量编号	迭代步数	变量编号	迭代步数	变量编号	迭代步数
120	1	189	1	47	0
76	21	64	0	30	0
20	0	354	48	168	3
265	3910	346	0	10	0
73	0	199	0	350	0
4	0	325	0	57	4
22	1	173	16	302	25
233	0	167	0		
342	0	81	0		

由表 5-1 可知，完成整个优化过程需要 3910 步，这是由于该变量对应的$|\Delta_n|=1$相对较小所导致，这里我们为方便起见，将其$|\Delta_n|$设为 80，以方便后续的展示，这样该变量的迭代步数变为了 49，于是后续整个优化过程现在只需要 49 步完成。分别如图 5-1 和 5-2 所示，展示优化过程中，汽油产品辛烷值和硫含量的变化轨迹。图 5-1 展示了产品汽油辛烷值与产品辛烷值损失的变化轨迹，从图中可知辛烷值损失在优化迭代过程中基本保持下降，图中标出了某些主要操作变量调整完成时的位置，如第 57 号操作变量（加热炉进口温度）在第 4 步迭代完成后就不再进行调整了，其他位置处的类似。

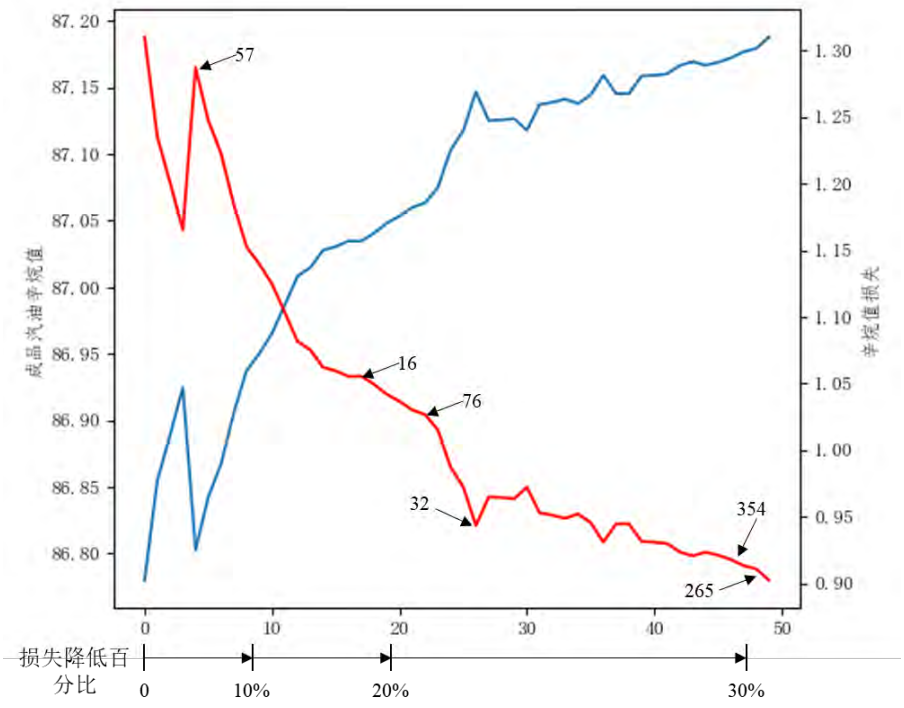


图 7-1 产品汽油辛烷值与辛烷值变化轨迹

分析图 5-2 可知，虽然在迭代优化过程中产品汽油硫含量有一定程度的升高，但是并没有超出$5\mu g/g$，这说明在整个优化过程中，根据每一步优化后的各操作变量设定条件

下，所制得的产品汽油的硫含量始终能够满足汽油产品标准的要求，且能够在保证满足对产品硫含量标准要求下，使得优化后方案下的产品辛烷值损失有较大程度的下降，也就能够为工业流程提供一定的参考，创使得催化裂化得到的汽油在满足燃烧标准的条件下，为工厂创造更高的经济效益。

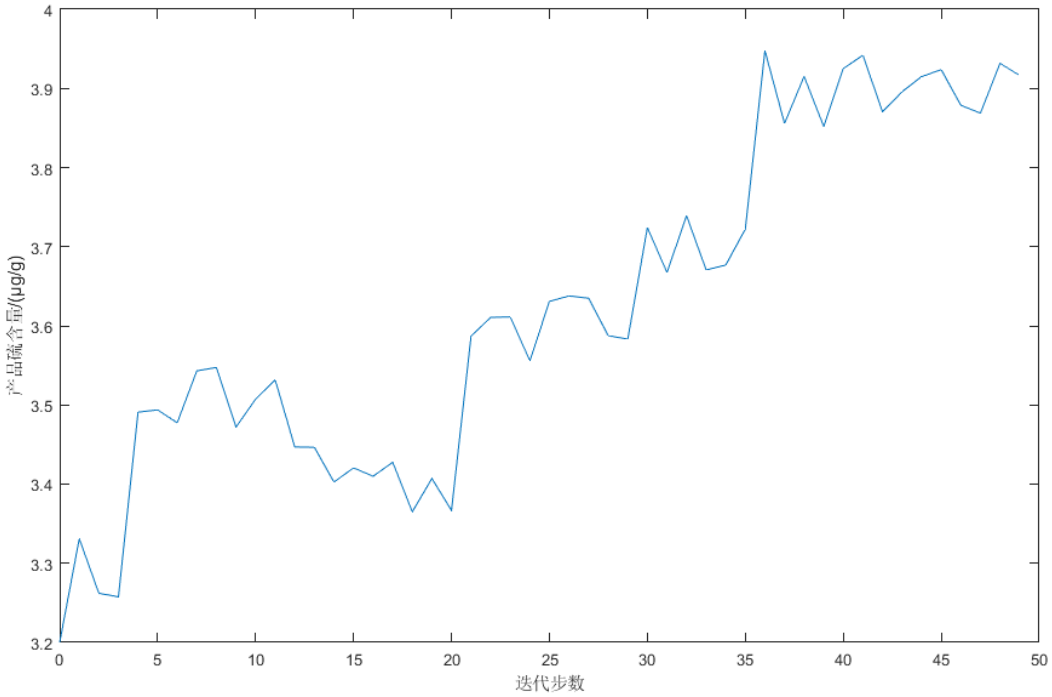


图 7-2 产品硫含量的变化轨迹

8. 结论与模型评价

8.1 结论

（1）利用多种数学原理和计算软件，完成了数据处理

对于问题 1，预处理提供了 285 号和 313 号样本共时长为 2 个小时的数据采集记录（包括非操作变量和操作变量），同时还提供了所有操作变量的相关信息。① 由于非操作变量属于相关产品和原材料的固有属性且采集频率较低，只提供了一条数据记录，因此不再对提供的非操作变量进行处理。②对于提供的非操作变量数据中存在的各种不良数据或错误数据，先后采用了相应的处理手段：对于由于设备、外界条件导致的变量值缺失为空值的异常数据作了取前后一段时间内平均的补全处理；对于变量值超出可调控最小或最大范围的数据，根据该操作变量的可调控范围，直接将其取为相应的最小值和最大值，最后剔除掉所有操作变量数据组中的最大值和最小值以实现更加良好的平均；对于变量值超出 3 σ 区间范围的数据，考虑利用 3 σ 准则将每个位点采集到的跳跃值数据剔除出去。经处理发现该原始数据中不存在任何的数据缺失，但两个样本中均存在同样的 4 个操作变量的所有数据值都超出范围值的情况，可能为单位不一致所导致。删除不符合操作变量可调控范围的最大最小值要求的数据并继续剔除最大值和最小值后，两个样本各剔除 0 个和 65 个超出 3 σ 区间的值。

（2）建立基于 EM 算法的 GMM 聚类对 354 个操作变量进行聚类，将其分为 30 类，，并通过 t‐sne 降维可视化技术对聚类结果进行可视化；另外，对 12 维性质变量及 30 类中每一聚类簇中根据信息增益理论计算其关于辛烷值损失的信息增益值，在信息增益计算结果排序的基础上结合斯皮尔曼相关系数，对 42 维变量进行筛选，剔除相关性较大及信息增益值较低的变量，最终筛选出包括 E-101D 壳程出口管温度、还原器温度、精制汽油出装置温度、K-103A 进气温度、D121 去稳定塔流量在内的 28 维具有代表性和独立性的变量（包括 3 维性质变量及 25 维操作变量）。

（3）建立了基于慢特征分析的即时学习（SFA-JITL）辛烷值损失预测框架。该模型首先对 28 个主要变量进行了慢特征分析筛选出 7 维慢特征，其次搭建即时学习框架对样本序列依次输入完成每个样本值辛烷值损失的预测，并在加入更新样本的过程中利用局部加权慢特征方法对慢特征权重不断调整。此外，SFA-JITL 辛烷值预测模型结果与真实值对比并于其他模型相比较验证了模型的有效性，RMSE,MAE,MAPE 值分别为 0.261，0.199，16.124，优于线性回归模型及卡尔曼滤波模型 。

（4）经过基于慢特征分析的即时学习预测模型预测后，得到了 325 个数据样本的辛烷值损失预测值。显而易见地，各数据样本优化效果有所不同。然而来源于某石化企业的这 325 个数据样本，其汽油产品辛烷值损失平均为 1.37 个单位，而同类装置的最小损失值只有 0.6 个单位，还有 56.21%的优化空间。因此，在保证产品硫含量不大于 $5\mu g/g$ 的前提下，我们挑选出降幅大于 30%的样本，并详细分析所对应的主要变量经过优化后的操作条件，采取求均值的方法，获得最大可能适用于所有样本的操作条件，进而为石化企业的长期、高效运营提供科学稳健的理论支撑

（5）在题目要求条件下，对各主要操作变量进行一步一步地迭代，在每一步迭代中，经过模型的分析预测，得到了在当前步操作变量的设置下产品的辛烷值、辛烷值损失和硫含量值数据，画出了整个迭代优化过程中对应的成品汽油的辛烷值和硫含量的变化轨迹曲线。最终共迭代过程共有 49 步，数据显示在优化过程中，汽油产品辛烷值损失在优化迭代过程中基本保持下降，而产品汽油硫含量有一定程度的升高，但是并没有超出 5 μ g/g，这说明在整个优化过程中，根据每一步优化后的各操作变量设定条件下，所

制得的产品汽油的硫含量始终能够满足汽油产品标准的要求，且能够在保证满足对产品硫含量标准要求下，使得优化后方案下的产品辛烷值损失有较大程度的下降，也就能够为工业流程提供一定的参考，创使得催化裂化得到的汽油在满足燃烧标准的条件下，为工厂创造更高的经济效益。

8.2 模型评价

模型优点：

(1)设计了较为合理的方法对原始数据进行预处理，同时还考虑到数据采集设备自身存在的白噪声误差，使用 3σ准则对数据进行了处理，使原始数据更为精确。

(2)在对基于慢特征分析的即时学习（SFA-JITL）辛烷值损失预测过程中，考虑到实际工艺流程，能够对不断加入的样本进行预测更新，通过数学模型与试验相结合的方法，使得验证的结果更加具有说服力。

模型缺点：

(1)本文中辛烷值损失预测模型及优化方案是根据 325 个样本进行分析处理得到的，在后续的研究中可以根据新加入的样本进行分析，探讨个变量和之间的内在联系。

(2)在变量特征提取筛选的过程中，通过不同变量簇之间的聚类及计算信息增益值可得到相关结果，根据 AIC\BIC 准则及题目的推荐变量数筛选，后续需根据更加科学的筛选指标体系构建适用于汽油精制的变量筛选体系。

参考文献

[1] 江火生.催化裂化装置反应温度对产品分布的影响[J].广东石油化工学院学报,2020,30(04):23-26.

[2] 胥红玉.浅谈汽油辛烷值的影响因素[J].品牌与标准化,2020(05):49-50+52.

[3] 寿建祥,何阳.催化重整技术对炼化一体化总流程的支撑作用[J].炼油技术与工程,2020,50(02):49-54.

[4] 张美霞,李丽,杨秀,孙改平,蔡雅慧.基于高斯混合模型聚类和多维尺度分析的负荷分类方法[J/OL].电网技术:1-15[2020-09-19].https://doi.org/10.13335/j.1000-3673.pst.2019.1929.

[5]Hinton, G. E. Connectionist learning procedures[J]. Artificial Intelligence, 1989, 40(1): 185-234.

[6] Laurenz Wiskott,Terrence J. Sejnowski. Slow Feature Analysis: Unsupervised Learning of Invariances[J]. Neural Computation,2002,14(4).

[7]黄健,杨旭.基于在线加权慢特征分析的故障检测算法[J/OL].上海交通大学学报:1-9[2020-09-19].https://doi.org/10.16183/j.cnki.jsjtu.2020.99.012.

[8]王鹤莹,陈夕松,迟慧,梅彬,段佳.基于慢特征分析的初馏塔故障预警研究[J].石油化工应用,2020,39(04):82-85+106.