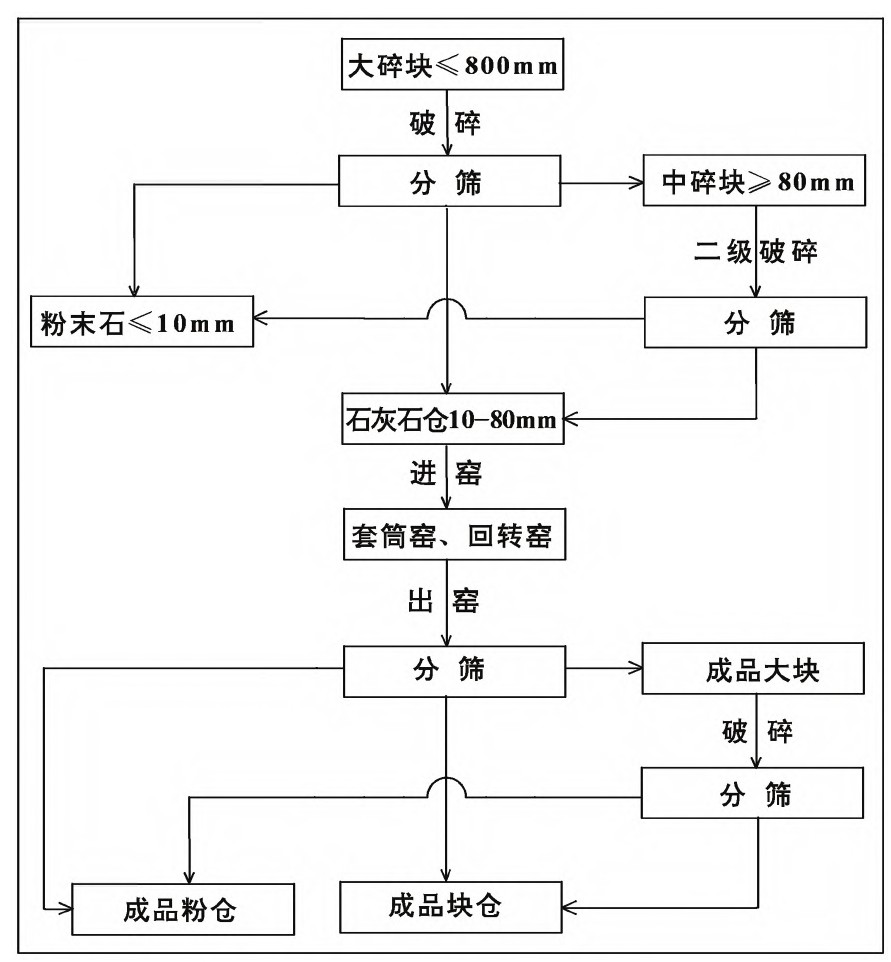
矿石加工质量控制问题

**摘要**

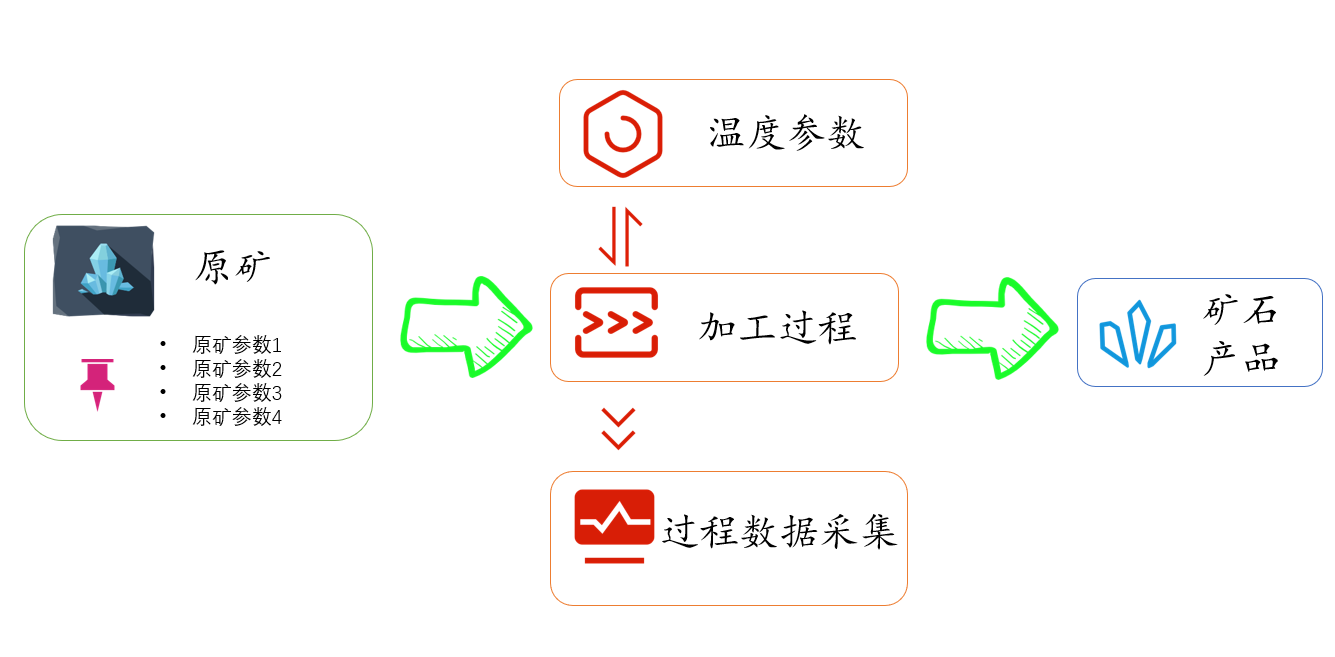
**关键词：支持向量机、CatBoost算法、xgboost算法，投票法，对抗学习。**

# 1 问题背景与重述

2020年9月22日，国家主席习近平在第75届联合国大会一般性辩论上宣布，中国将提高国家自主贡献力度，采取更加有力的政策和措施，二氧化碳排放力争于2030年前达到峰值，努力争取2060年前实现碳中和。提高矿石加工质量，可以直接或间接地节约不可再生的矿物资源以及加工所需的能源，从而推动节能减排，助力“双碳”目标的实现。矿石加工是一个复杂的过程，在加工过程中，电压、水压、温度作为影响矿石加工的重要因素，直接影响着矿石产品的质量。矿工加工实际流程如下：



为了方便建模，假设矿石加工过程需要经过系统I和系统II两个环节，两个环节不分先后，其他条件（电压、水压等）保持不变。简化工艺流程如下：



针对该生产车间2022-01-13至2022-04-07的生产加工数据，进行数学建模，完成下列问题：

(1)问题一要求对附件1中生产加工数据，研究原矿参数和系统设定温度对产品质量指标之间的关系，建立数学模型，给出利用系统温度预测产品质量的方法并预测对应温度的产品质量指标。

(2) 问题二要求进一步探究原矿参数，系统设定温度，产品质量指标之间的关系，在问题一的基础上，分析问题一的逆映射，探究产品目标质量所对应的系统温度并预测对应产品质量指标下可能的系统温度。

(3) 问题三要求利用附件2的数据，通过增加过程数据，利用过程数据改进问题一模型，给出指定系统设定温度，预测矿石产品合格率的方法并给出合格率预测结果。

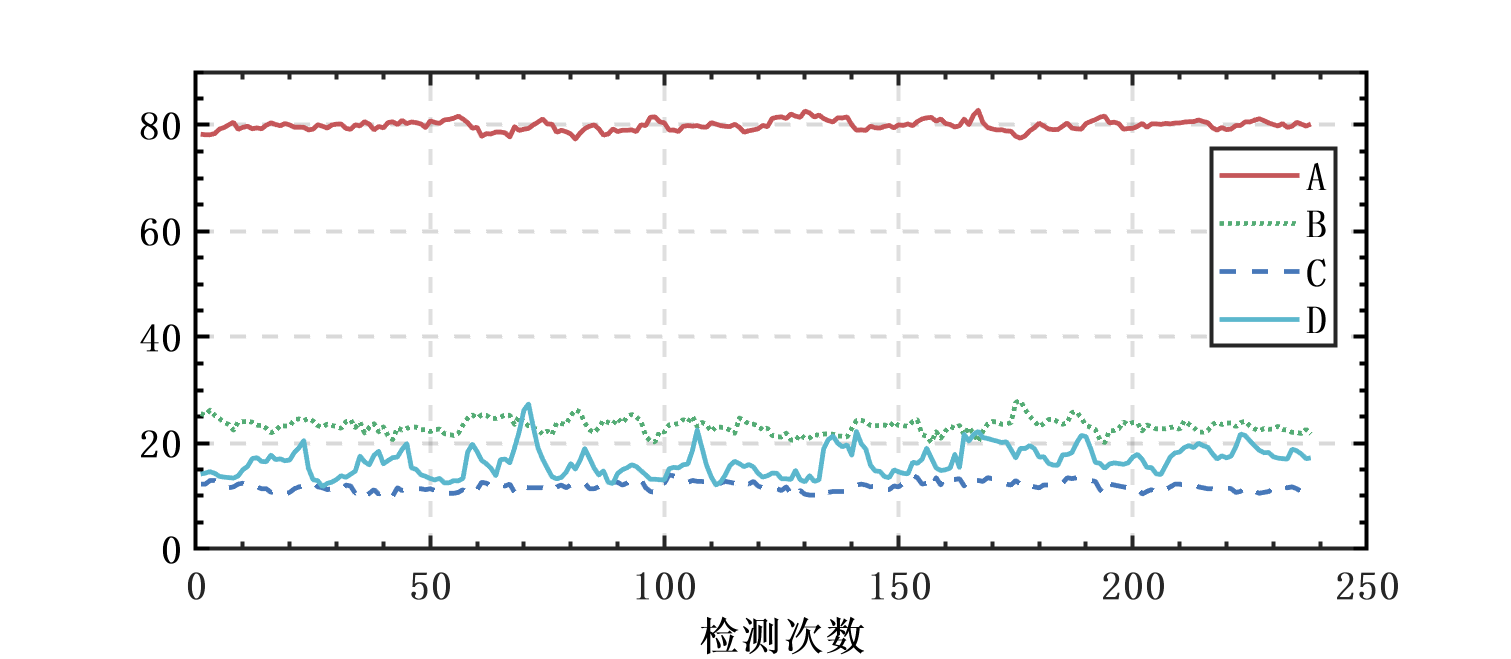
(4) 问题四要求在问题三基础上，进一步分析灵敏度和准确性。判断能否达到2022-04-10和2022-04-11产品的合格率要求，如果可以达到，给出系统设定温度。

# 2 问题分析

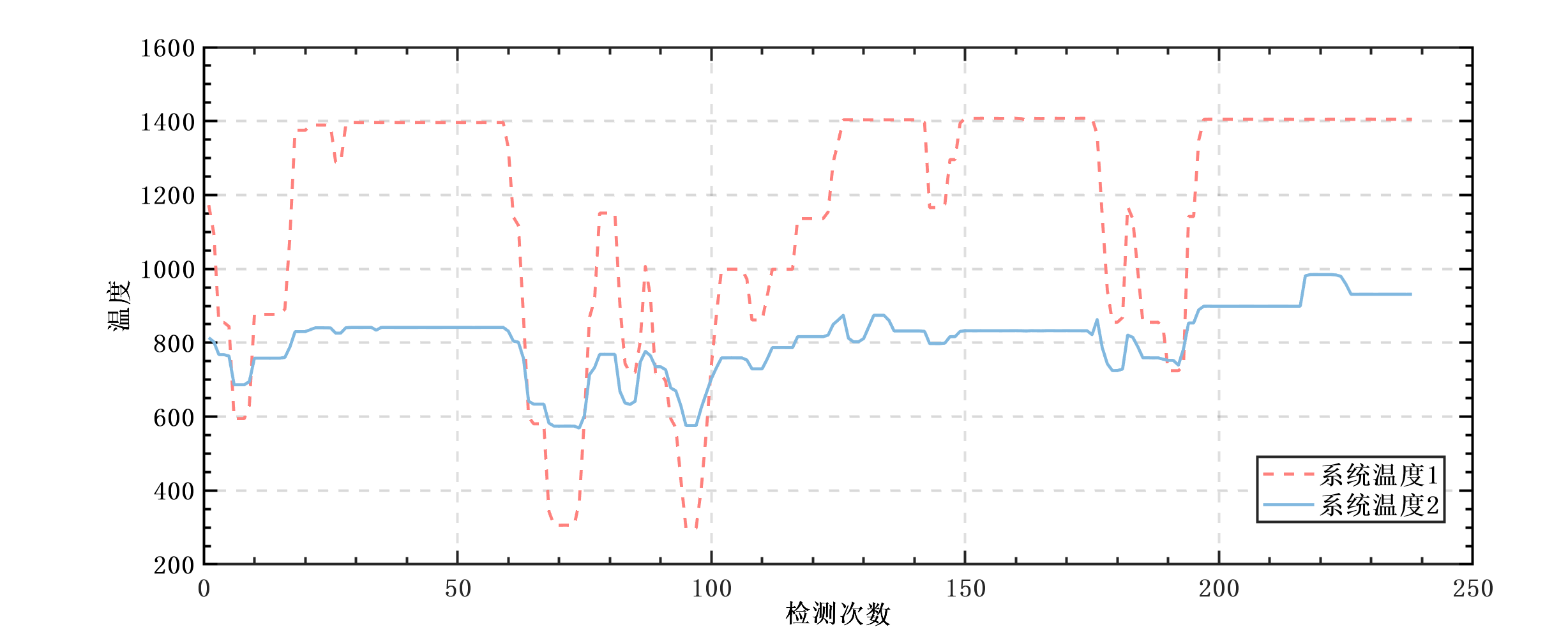
## 2.1 题目整体分析

首先将数据清洗后做整合，整理出系统温度、产品指标以及原矿参数对应小时的数据样本。处理过程中，我们对一小时内的系统温度取均值作为对应时间的数值，而原矿参数设置为当天内24小时原矿参数相同。

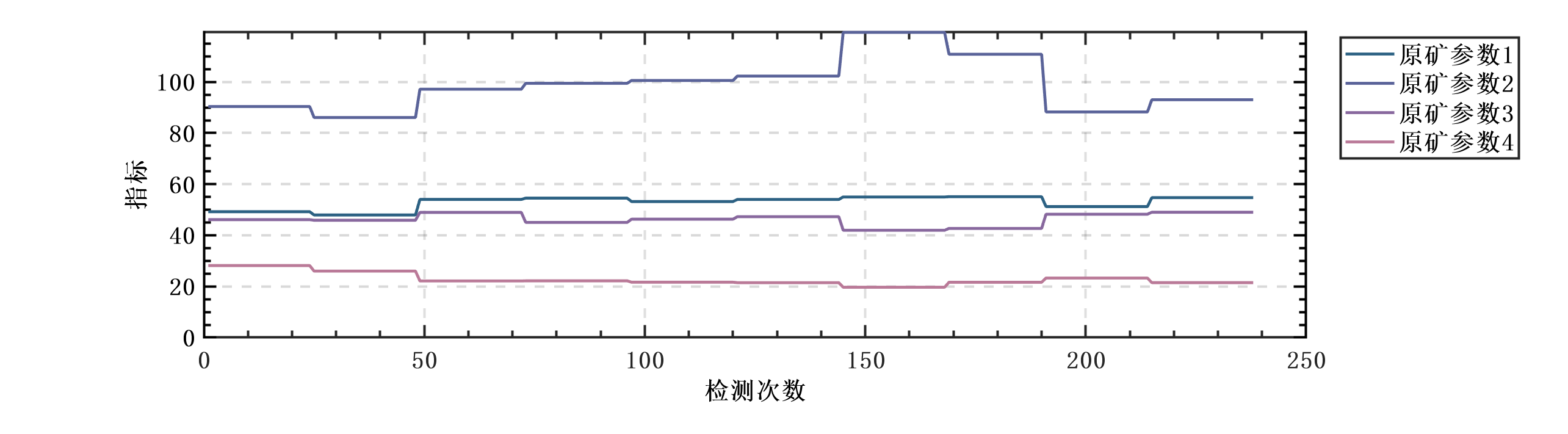
对于该问题, 我们首先需要观察数据的分布情况，探究数据之间的相关性，挖掘其数据本身的特征。为此我们选取其变量数据进行绘图。各趋势图如下图所示：



由XX图可以看出参数A,B,C之间波动范围小，D波动范围大，可能更容易受到影响。

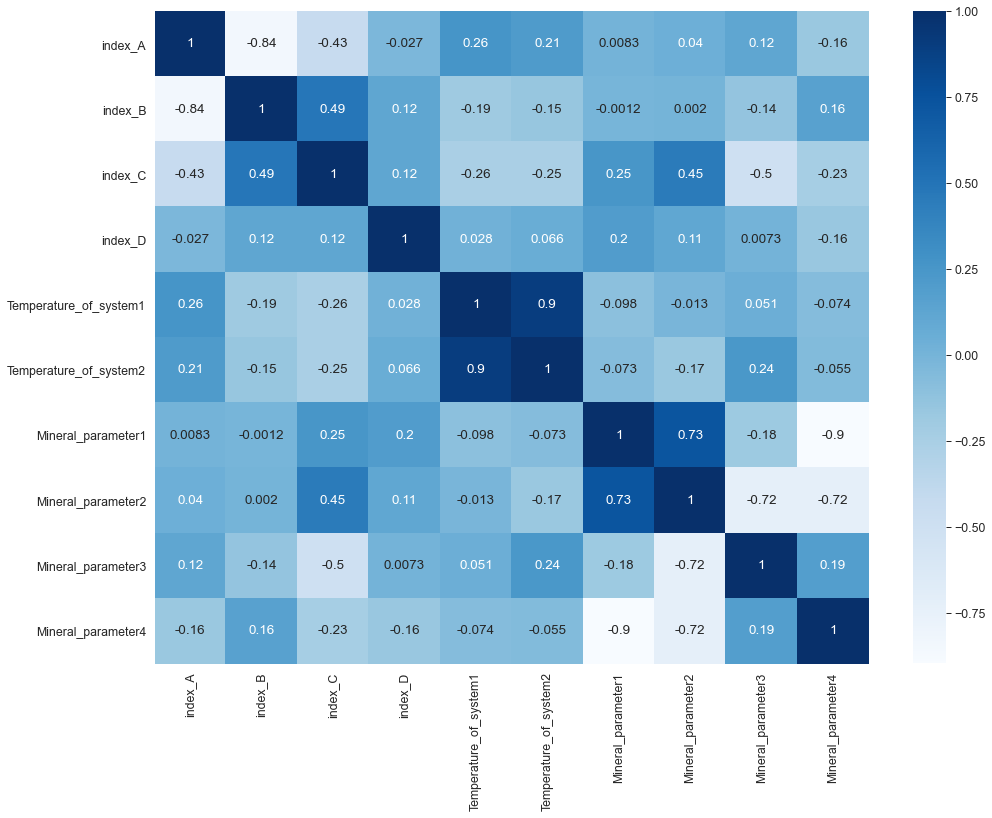


由xx图可以看出系统温度检测时波动幅度较大，尤其是温度1，范围较广，可能是主要影响因素之一。



由xx图可以看出，原矿参数1变化幅度较大，原矿参数2，3呈现负相关，原矿参数4较为平缓。后面因重点探究原矿参数1，2，3对其结果的影响。

此外，我们绘制了其变量相关性统计图：



从图片中可以看出，变量之间相关性不是很明显，其中指标A,B负相关，系统温度1与系统温度2相关性明显，原矿参数之间互相影响。为了进一步探究，绘制俩俩变量之间散点图和核密度估计图。



上面可以看出其变量之间的分布情况和关联度。右上角为俩俩变量之间的散点图，左下角为核密度估计，对角线为变量本身特征。由xx图可以发现其变量之间存在多维度影响，难以直接用线性关系进行拟合。

通过审题我们发现，问题一要对生产加工数据进行分析找出系统温度和产品质量指标的映射关系，而映射关系可以归结为一类模型，每个模型在不同催化剂情况下有特定的参数。问题二需要找出不同催化剂组合及温度作为影响乙醇转化率以及C4烯烃选择性大小的影响因素的权重，第三问则要利用第二问的结果，搜索出使得C4烯烃收率尽可能高的催化剂组合与温度。

## 2.2 问题一的分析

附件1给出了2022-01-13至2022-01-22的生产加工数据，要求建立数学模型，给出利用系统温度预测产品质量的方法。在给定的2022-01-23原矿参数和系统设定温度下，给出产品质量预测结果。

根据上面对数据样本的分析，针对样本高维特征和样本数量不多的特点，在问题一中，决定采用支持向量机进行回归预测，可以有效挖掘高维度特征之间的关系，针对问题一产品质量指标A,B,C,D构建多输出支持向量机模型。由于样本数据偏小的原因，为了增强鲁棒性，引入投票法进行平均，通过搭配不同核函数实现差异化学习，投票法最终的预测结果是多个回归模型预测结果的平均值。

## 2.3 问题二的分析

利用附件1的数据，假设原矿参数和产品目标质量已知，建立数学模型，估计产品目标质量所对应的系统温度。在给定的2022-01-24原矿参数和目标产品质量下，给出系统设定温度。

通过问题一的分析我们了解到矿石加工过程中，不同变量和产品质量指标之间的关系，在问题一基础上，探究其逆映射，针对支持向量机非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数。针对问题一多项式核poly，高斯核函数rbf对非线性内核存在误差，引入随机森林发展而来的集成化学习，在问题一模型上进行改进。通过对比各类集成学习，最后确定引入xgboost和catboost平衡算法的性能。进一步完善模型，根据模型可以很轻易求出问题一二的结果。

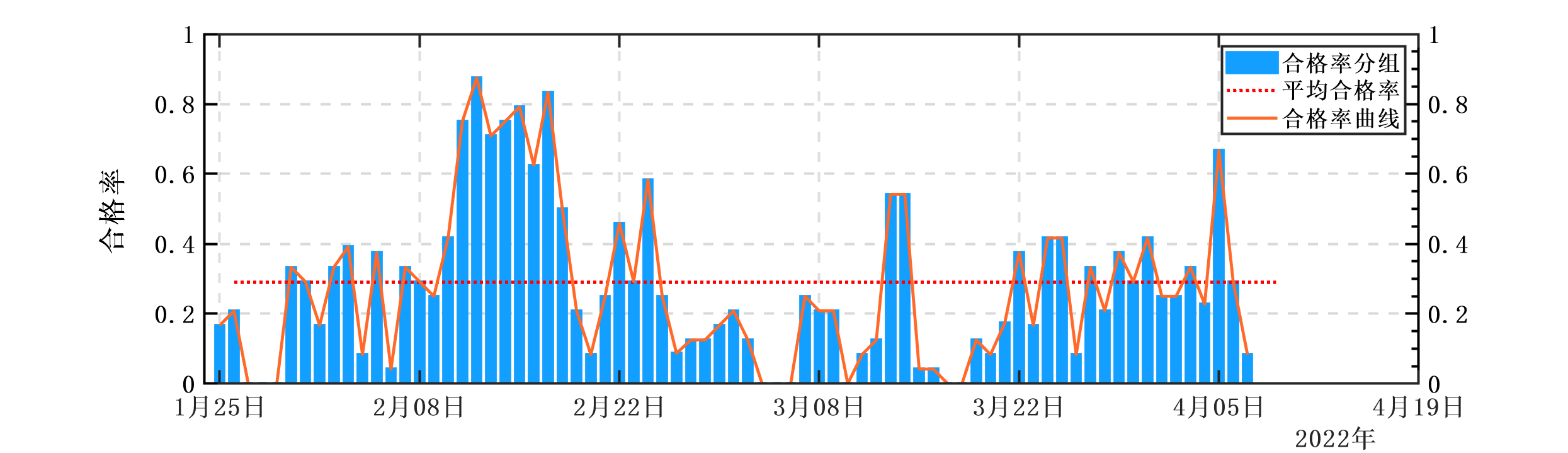
## 2.4 问题三的分析

附件2给出了该生产车间2022-01-25至2022-04-07的生产加工数据及过程数据，满足下表销售条件的产品视为合格产品，合格率=合格产品数/产品总数。请建立数学模型，给出指定系统设定温度，预测矿石产品合格率的方法。给定2022-04-08和2022-04-09原矿参数、过程数据和系统设定温度，给出合格率预测结果。

表格 1合格标准

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标 | 指标A | 指标B | 指标C | 指标D |
| 销售条件 | 77.78 - 80.33 | <24.15 | <17.15 | <15.62 |

对附件2的数据做同样的数据分析，拥有相同的样本特征，利用给出合格标准，统计样本合格率，结果如下：

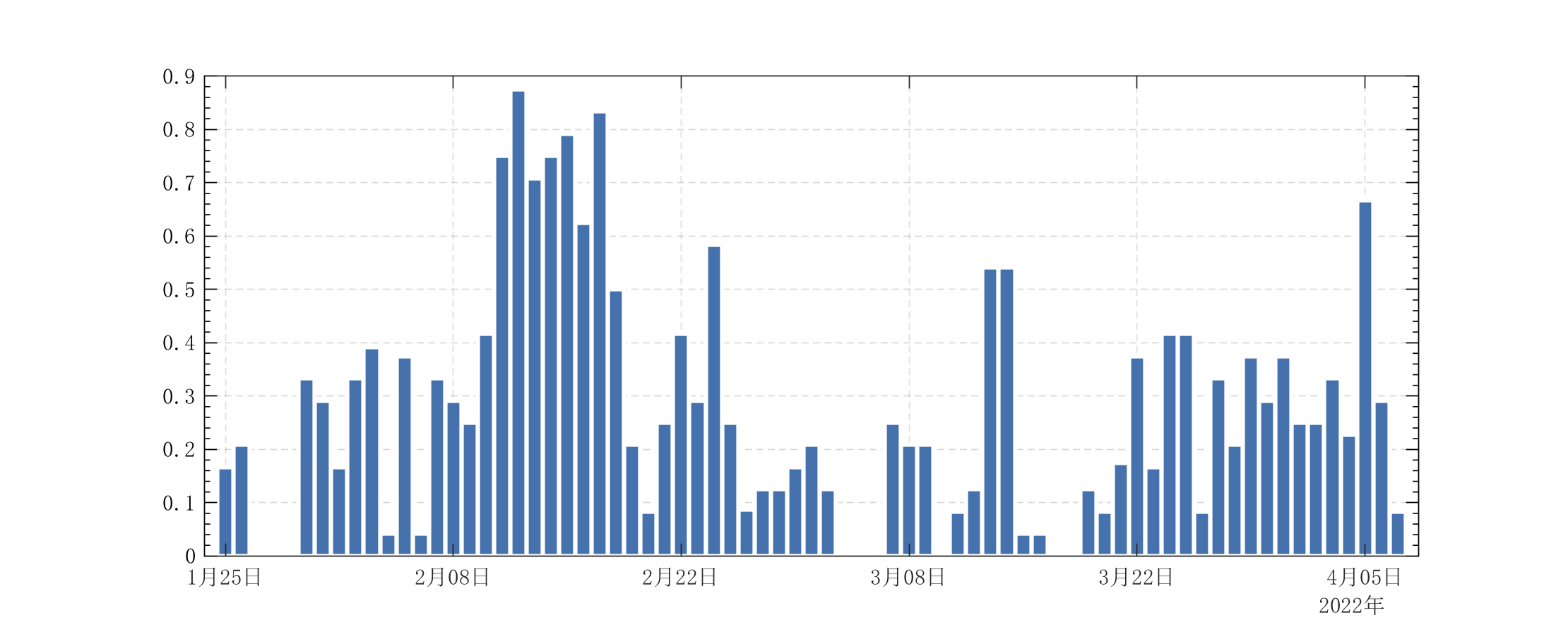


由xx图可知，其整体合格率偏低，如果直接利用模型进行预测得产品质量指标A,B,C,D按标准进行归类合格，其预测存在一定误差，在误差范围内容易影响结果。设计对抗学习，利用问题二改进的模型进行回归预测，生成产品质量指标A,B,C,D，与原产品质量指标A,B,C,D进行混合，构建分类模型。利用对抗思想，评估生成的样本数据（也称fake data即造假数据），与真实数据到底有多接近这样一个评判标准（也就是这里的合格标准），通过对比，合理利用误差，得到在误差允许范围内的分类模型（这里选用随机森林），利用改进的预测模型和对抗分类模型，给出合格率预测结果。

## 2.5 问题四的分析

利用附件2的数据，建立数学模型分析在指定合格率的条件下，设定系统温度的方法，并对模型的准确性和敏感性进行分析。判断能否达到2022-04-10和2022-04-11产品的合格率要求，如果可以达到，给出系统设定温度。

在前面模型的基础上，预测合格率是否能达到。由于整体合格率偏低，利用如下图所示：



由于合格率在0.8以上样本数较少，在上面模型的基础上求其逆映射存在困难，这里采取优化算法来寻找结果。构建非线性映射函数，将合格率作为因变量进行研究。因为自变量温度的取值有无限种取值可能，所以对于此全局寻优问题，我们采用启发式算法—粒子群优化算法（PSO）求解。如果无解则代表不达到给定要求。

# 3 模型假设及符号说明

## 3.1 模型假设

我们已经结合实际情况对本问题进行了系统且具体的分析, 根据上述分析我们做出如下合理的假设，以期利用恰当的数学模型对该问题进行详细的解答.

* 假设系统温度与调温指令设定的温度相同；
* 假设每次温度调节后的两个小时不会输入新的调温指令，系统温度变化幅度不大，可以用两小时的平均值来表示这两小时内任一时刻的温度；

## 3.2 符号说明

为便于描述，我们对通篇所使用的符号及其意义做如下说明。

表格 1 符号意义表

| **符号** | **含义** |
| --- | --- |
|  | 数据的期望 |
|  | 数据的方差 |
|  | 数据的协方差 |
|  | 数据的相关性 |
|  | 事件发生的概率 |
| *Index* | 指标 |
| *T* | 系统温度 |
|  | 原矿参数 |
| *Pp* | 过程数据 |
| C | 惩罚因子 |
| W | 核函数 |
|  | 平均绝对误差 |
|  | 拟合度指数 |

# 4 模型的建立与求解

# 5 模型的评价与推广

## 5.1 模型的优点

### 5.1.1问题一二模型的优点：

1. 非线性映射是SVM方法的理论基础,SVM利用内积核函数代替向高维空间的非线性映射，样本量不是海量数据的时候，准确率高，泛化能力强。

2. 投票法是一种遵循少数服从多数原则的集成学习模型，通过多个模型的集成降低方差，从而提高模型的鲁棒性和泛化能力。

3. XGBoost的决策树是Level-wise增长。Level-wise可以同时分裂同一层的叶子，容易进行多线程优化，过拟合风险较小

4. Catboost采用的策略在降低过拟合的同时保证所有数据集都可用于学习。

### 5.1.2问题三模型的优点：

1. 引入对抗学习GAN网络的思想，以半监督方式训练分类器，使得在误差范围能更准确的预测结果。

2. 对抗学习使得数据增强，具有更广的适用性。

3. 随机森林利用构建决策树，可以处理非线性特征且考虑了变量之间的相互作用。

### 5.1.3问题四模型的优点：

1. 粒子群算法方便简单且易于实现，对比其他搜索算法，其速度更快，方便易操作。

## 5.2 模型的缺点

### 5.2.1问题一二模型的缺点：

1. 模型较为复杂，通过投票法使得模型更加稳定的同时也产生了大量冗余，造成模型训练速度慢。

2. 其过拟合效果联合，可数据样本小时，如问题一样本数据远小于问题三，使得部分模型欠拟合问题，只能减少测试集合，扩大训练集。

3. 由于xgboost和catboost与支持向量机特性不同，在鲁棒性增强的同时，准确性有所下降。

### 5.2.2问题三模型的缺点：

1.对于特征的处理需要大量的内存和时间；

2.对抗训练学习了假的数据，存在误差影响，如果前面模型误差太大会导致误差进一步放大，影响结果。

### 5.2.3问题四模型的缺点：

1有时粒子群在俯冲过程中会错失全局最优解；

2应用PSO算法处理高度复杂问题时，算法可能过早收敛；

3 PSO算法是一种概率算法，搜索过程带有随机性。

## 5.3 模型的推广或改进

1. 模型二对模型一进行了改进，弥补了支持向量机的曲线，也融合了xgboost和catboost的优点，在牺牲一部分准确性的情况下，使得模型更加稳定，对此可以更容易的利用在其他问题上。

2. 针对问题三中对抗学习，可以引入对抗网络和变分自编码器进行模拟，可以增强其模型稳定性和准确率，但也需要更多的样本数据，问题三中数据太少，无法构建神经网络进行预测。

3.在问题四中我们采用了PSO算法，在启发式算法中，还可以升级为人工鱼群法，加入觅食行为，聚群行为，追尾行为，随机行为等控制，提高搜索算法的精确度。

# 6 参考文献

# 7 附录

## 7.1 附录1 工具和软件

本文程序用到了MATLAB R2020a，SPSS，Python，EXCEL（包括一些开源的库和包，用于求解问题）

本文绘图使用了MATLAB，javascript，drawio等软件和语言，用来美化和绘制图片。

## 7.2 附录2代码

### 7.2.1 关联度代码（MATLAB）

|  |
| --- |
| 1. clear;clc; 2. load data 3. x=data; 4. for i=1:6 5. x(:,i)=x(:,i)/x(1,i); 6. end 7. x=x'; 8. n=size(x,1); 9. ck=x(1,:);m1=size(ck,1); 10. bj=x(2:n,:);m2=size(bj,1); 11. for i=1:m1 12. for j=1:m2 13. t(j,:)=bj(j,:)-ck(i,:); 14. end 15. jc1=min(min(abs(t')));jc2=max(max(abs(t'))); 16. rho=0.5; 17. ksi=(jc1+rho\*jc2)./(abs(t)+rho\*jc2); 18. rt=sum(ksi')/size(ksi,2); 19. r(i,:)=rt; 20. end 21. r 22. [rs,rind]=sort(r,'descend') %对关联度进行排序 |