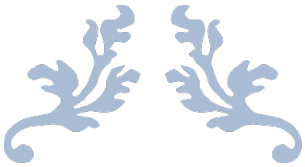
天津大学
Tianjin University



基于改进的综合预测评价模型求解矿石加工质量控制问题

王栋樟，张祎凡，高晨萱

3020244126, 3020234508, 3020001065

Time: 2022.05.10

MAY.2022

摘　要

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。在矿石加工过程中，温度是极其重要的一个参数。如何通过系统温度预测产品质量，或通过目标产品质量、合格率等信息推测系统设定温度是矿石加工产业的重点问题。

针对问题 1 首先研究所给附件数据情况，合理划分数据集，然后设定原矿参数与系统Ⅰ、Ⅱ的设定温度为输入参数，表征产品质量的参数为输出参数，放入所搭建的基于遗传算法优化的 BP 神经网络、多元线性回归模型、Xgboost 预测模型、随机森林预测模型和决策树预测模型五种模型中进行训练，并将各自对应预测评价指标利用基于因子分析法与熵权法的模糊综合评价模型进行评估，Xgboost 预测模型获得最高评分，选取该预测模型进行求解，得到 2022-01-23 在不同设定温度下产品质量结果为$I = [80.15518, 23.90084, 11.51756, 16.11099] / [79.39446, 23.30358, 12.43314, 16.68909]$。

针对问题 2 根据题干要求，将其转化为类似问题 1 所求解的问题，设定原矿参数与表征产品质量的参数为输入参数，系统Ⅰ、Ⅱ的设定温度为输出参数，重复问题 1 的求解过程，最终随机森林预测模型获得最高评分，选取该预测模型进行求解，得到 2022-01-24 在不同产品质量参数值下设定温度结果为$T = [1399.69, 851.3281] / [1153.077, 755.4725]$。

针对问题 3 对于新的附件数据，重新分析，进行数据集的划分，首先设定原矿参数、过程数据及系统Ⅰ、Ⅱ的设定温度为输入参数，表征产品质量的参数为输出参数，再次重复问题 1 的求解过程，最终随机森林预测模型获得最高评分，利用混淆矩阵等手段，对该模型对于产品是否合格的预测进行评估，得出可利用输入参数直接预测合格率的结论。然后将输出参数仅设为合格率，再次利用随机森林预测模型进行问题求解，得到 2022-04-08/09 的合格率结果为$R = 0.258768 / 0.248601$。

针对问题 4 基于问题 3 的预测模型，首先对常见温度区间进行小规模遍历，发现合格率关于温度存在着阶梯性变化趋势，然后将温度区间扩大到全部可能温度，增大步长，进行遍历求解，发现 2022-04-10/11 的指定合格率均不能达到。

最后对于每一问的模型进行准确性和敏感度检验，本文结论正确合理，具有一定应用价值和普适性，同时对模型优缺点进行分析总结，并对存在的缺点提出改进方案设想与推广方向。

关键词：矿石加工质量控制问题，改进的综合预测评价模型，Xgboost 预测模型，随机森林预测算法，混淆矩阵

一、问题重述

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。

在矿石加工过程中，为保证矿石质量，需要严格控制加工过程中的电压、水压、温度等参数。如图 1 所示，矿石加工过程需要经过Ⅰ、Ⅱ两个系统，两个环节不分先后，生产过程中保持除温度以外的电压、水压等其他条件不变。

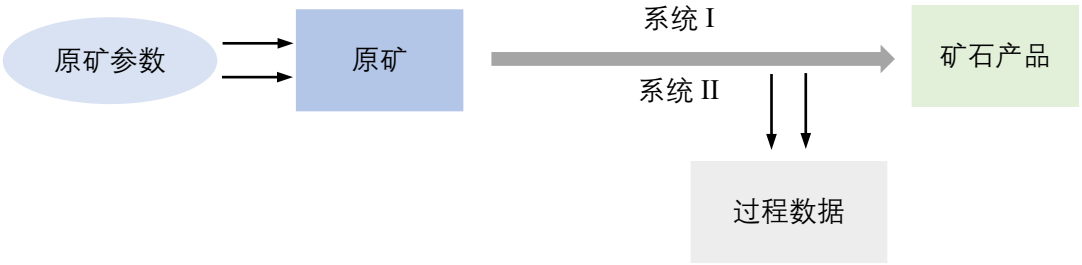


图 1 矿石加工过程

工作人员下达调温指令后，系统温度改变为设定温度（可能会有轻微波动），同时记录系统的实时温度。在每一次传入调温指令的两个小时之内，不会再传入新的调温指令，两个小时以后，矿石加工过程结束，此时可以检测表征矿石产品质量的四个评价参数指标。

根据上述条件，需要研究以下四个问题：

问题 1：给定一定期间内所加工的原矿石的参数，系统设定温度以及表征产品质量的四个参数。要求建立利用系统温度预测产品质量的数学模型，并给出待预测的可能性最大的产品指标。

问题 2：在第一问的基础上，建立由原矿参数和目标产品质量估计系统设定温度的数学模型，并给出待预测的可能性最大的系统设定温度。

问题 3：要求在给定原矿参数和过程参数的情况下，建立由系统设定温度预测矿石合格率的数学模型。并结合给定一定期间内原矿参数、过程参数及系统设定温度，并依据该模型对合格率进行预测，同时建立数学模型对预测结果的准确性进行评价。

问题 4：在问题三的基础上，建立数学模型分析在指定合格率的条件下，如何设定系统温度的方法，解决以下问题：（1）适当的敏感性分析；（2）结果准确性分析；（3）给定原矿参数和过程数据，判断产品的合格率是否达到要求，若可达到，则给出系统设定温度。

二、问题分析

2.1 问题 1 分析

在问题 1 中，需要利用已知的原矿参数与系统Ⅰ、Ⅱ的设定温度共六个参数，预测表征产品质量的四个参数。首先基于假设，将已知数据进行预处理，并对数据集进行划分，利用基于遗传算法的神经网络、多元线性回归方程、Xgboost、随机森林和决策树五种模型进行拟合预测，并采用基于因子分析法与熵权法的模糊

综合评价法，评价五种模型的拟合效果。拟合效果最好，即预测结果与真实值最接近的方法，视为最佳，同时认为利用该方法得出的预测结果是可能性最大的产品指标，并利用该模型预测两种不同的温度设定下的可能性最大的产品指标。

2.2 问题 2 分析

在问题 2 中，需要利用已知的原矿参数和产品目标质量参数共八个参数，预测与之相对应的系统Ⅰ、Ⅱ设定温度的两个参数。与问题 1 类似，保持数据集划分不变，重新划分输入输出参数，采用与问题 1 相同的五种预测模型对训练集进行拟合，并以同样的评价模型评估五种模型的拟合效果，认为拟合效果最好的方法得出的预测结果是可能性最大的系统设定温度，并利用该模型预测不同产品指标对应的最大可能性的系统设定温度。

2.3 问题 3 分析

在问题 3 中，需要利用原矿参数、过程数据和系统温度共十个参数，预测产品合格率。基于新的假设，对新的数据进行预处理，并重新划分数据集，将输出参数设置为表征产品质量的四个参数，再次利用评价模型评估五种预测模型的拟合效果，取拟合效果最好的模型进行后续的预测。同时利用混淆矩阵，评价预测结果与真实结果在产品是否合格方面的偏差。

重新进行数据处理及数据集的划分，将输出参数设置为表征产品质量的合格率这一单一指标，继续利用上述选定的预测模型，进行拟合预测，获得准确性最高的预测结果。

2.4 问题 4 分析

在问题 4 中，需要根据指定合格率，推断所需的系统设定温度。基于问题 3，继续使用原矿参数、过程数据与系统温度为输入参数，合格率为输出参数的预测模型。改变输入的系统温度，获取系统的输出合格率，遍历整个温度范围，获取指定合格率对应的温度范围。建立新的数学模型，提高该过程的效率。最后分析其准确性与敏感度。

三、模型假设

1. 假设系统 I、II 的加工环节不分先后，两系统设定温度间不存在相互影响与因果关系；
2. 假设矿石加工最终结果只受原矿参数、系统设定温度以及过程数据的影响，不考虑其他条件（如电压、水压等）的影响；
3. 假设系统实时温度与调温指令设定的温度相同，忽略轻微波动；
4. 假设产品质量指标的唯一温度影响因素是两小时前的系统实时温度；
5. 假设忽视小概率事件，对于参数未完全给出的数据组舍去；
6. 假设对于合格率的计算是以天为单位进行计算的，即合格率在一天内不会发生变化。

四、符号说明

表 1 符号说明

符号	说明	单位
T_1	系统Ⅰ设定温度	℃
T_2	系统Ⅱ设定温度	℃
M_i	原矿的第i参数	
P_i	第i个过程参数	
I_A	产品评价指标 A	
I_B	产品评价指标 B	
I_C	产品评价指标 C	
I_D	产品评价指标 D	
R	合格率	
MSE	均方误差	
BPGA	基于遗传算法的 BP 神经网络模型	

五、模型建立

纵观全局发现，问题 1 至问题 4 均为预测问题，因此我们建立五个时序预测模型及一个综合评价模型，以实现模型的正确选取及保证严谨性。同时所建立的时序预测模型均可通过更改输入、输出参数得到复用，因此给出拟采用六个模型的详细建立过程，作为问题求解过程的基础。

5.1 基于遗传算法的 BP 神经网络模型

5.1.1 BP 神经网络

BP 神经网络是一种多层前馈神经网络，常用的为输入层-单隐含层-输出层的三层结构，其主要实现步骤如下：

1. 输入的信号特征数据先映射到隐含层（激活函数实现），再映射到输出层（默认采用线性传递函数），得到期望输出值；

2. 将期望输出值和实际测量值做比较，计算误差函数 J，再将误差反向传播，通过梯度下降等算法来调节 BP 网络的权值和阈值；

3. 重复该过程，直到满足设定的目标误差或者最大迭代次数等终止准则，停止训练。

5.1.2 基于遗传算法优化 BP 神经网络

传统 BP 神经网络存在着网络结构、初始连接权值和阈值的选择对网络训练的影响较大，但是又无法准确获得的问题，可以通过引入启发式算法遗传算法来获得更好的全局搜索能力[1]。

主要思想为将参数作为问题的决策变量，模型的精度作为问题的目标函数。遗传算法 GA 优化 BP 神经网络的算法流程图如下：

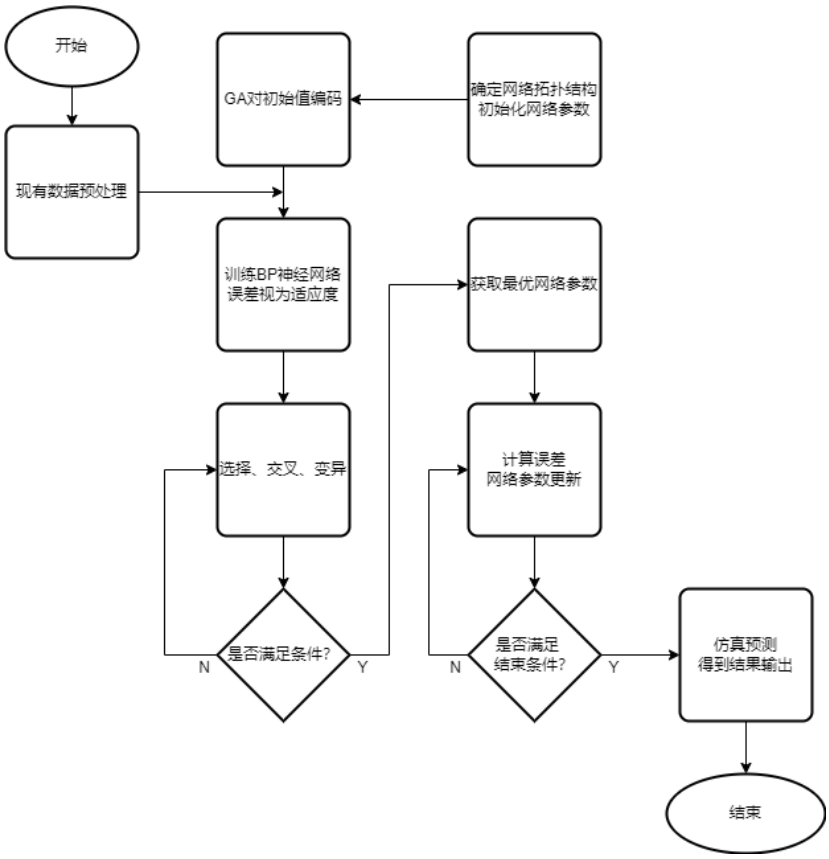


图 2 基于遗传算法的 BP 神经网络模型流程图

5.2 多元线性回归模型

在遇到有些无法用机理分析建立数学模型的时候，通常采取搜集大量数据的办法，基于对数据的统计分析去建立模型，其中统计回归模型用途最为广泛。

回归模型确定的变量之间是相关关系，在大量的观察下，会表现出一定的规律性，可以借助函数关系式来表达，这种函数就称为回归函数或回归方程。

回归模型解题步骤：

1. 根据实验数据画出散点图；
2. 确定经验公式的函数类型；
3. 通过最小二乘法得到正规的方程组；
4. 求解方程组，得到回归方程的表达式。

建立多元正态线性回归模型本质上是利用不同的方程组确定不同的线性参数。

5.3 随机森林预测模型

随机森林算法属于组合算法（集成学习）中作为装袋法的突出算法。装袋法的本质是在将多个模型装入同一个袋子后，让这个袋子作为一个新的模型来实现预测需求。装袋法的主体算法过程为：

1. 在样本集上重采样（有重复的）选出 n 个样本；
2. 在所有属性上，对这 n 个样本建立分类器（ID3、C4.5、CART、SVM、

Logistic 回归等）；

3. 重复以上两步 m 次，即获得了 m 个分类器；
4. 将数据放在 m 个分离器上，最后根据这 m 个分类器的投票结果，决定数据属于哪一类。

随机森林预测模型的实现步骤为[2]：

1. 使用装袋法在行列上进行随机抽样；
2. 构建很多决策树分类器组合而成的森林，单个的决策树分类器使用随机方法构成；
3. 构建从原训练集中通过有放回抽样得到的自助样本为学习集；
4. 随机取出参与构建该决策树的变量；
5. 单个决策树在产生学习集和确定参与变量后，使用 CART 算法计算，无需考虑过度拟合的问题；
6. 各个决策树分类器简单多数选举决定分类结果。

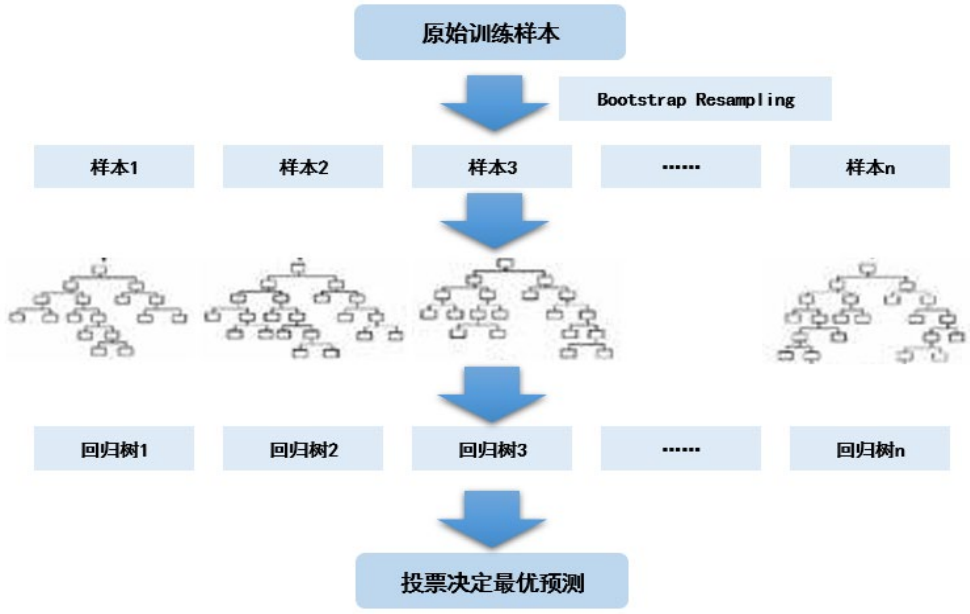


图 3 随机森林模型流程图

5.4 Xgboost 预测模型

Xgboost 预测模型，是一种特殊的梯度提升决策树，其极大程度的提升了算法的效率和准确性[3]。

Xgboost 预测模型算法步骤为：

1. 选择最优切分变量 j 与切分点 s，求解

$$\min\left[\min\sum_{xi\in R1(j,s)}(yi-c1)^2+\min\sum_{xj\in R2(j,s)}(yj-c2)^2\right]$$

遍历变量 j，对固定的切分变量 j 扫描切分点 s 选择使上式最小值的对(j;s)。其中 R_m 是被划分的输入空间，c_m 是空间 R_m 对应的输出值；

2. 用选定的对 (j;s)划分区域并决定相应的输出值

$$R_1(j,s)=\{x\mid x(j)\le s\},R_2(j,s)=\{x\mid x(j)>s\}$$

$$c^m = \frac{1}{Nm} \sum_{xi \in R1(j,s)} yi\,, x \in Rm, m = 1,2$$

3. 继续对两个子区域递归的调用上述步骤，最终将输入空间划分为 M 个区域 $R_1,R_2,\cdots,R_m$，生成决策树

$$f(x) = \sum_{m=2}^{M} c_m I\, x \in Rm$$

4. 当输入空间划分确定时，可以用平方误差来衡量回归树对于训练数据的拟合程度，用平方误差最小的准则不断地递归划分子树，直到平方误差满足需求

$$\mathrm{L}\big(\mathrm{y}, \mathrm{f}(\mathrm{x})\big) = \sum_{xi \in Rm} \big(yi - f(xi)\big)^2$$

Xgboost 引入了集成学习（boosting 方法），并采用并行计算等方式极大的加速了模型计算速度。Boosting 的核心思想就是所有弱分类器的结果相加等于预测值，然后下一个弱分类器去拟合误差函数对预测值的梯度/残差(这个梯度/残差就是预测值与真实值之间的误差)，从而不断地减小残差，直到满足系统的误差要求。

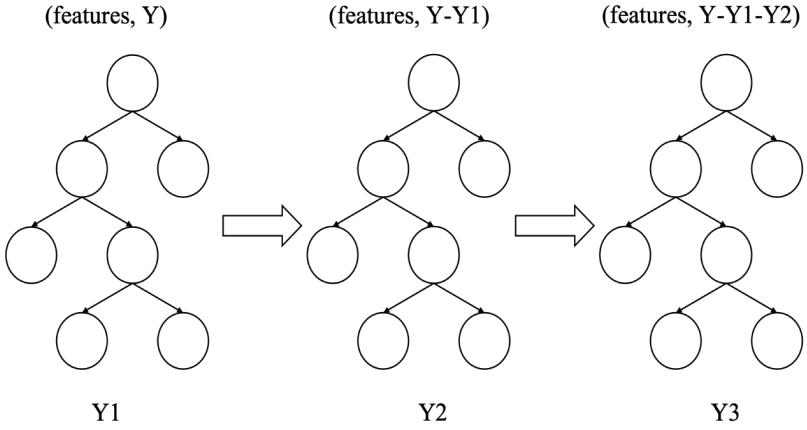


图 4 Xgboost 模型流程图

5.5 决策树预测模型

决策树是一种从无次序、无规则的样本数据集中推理出决策树表示形式的分类规则方法。

决策树学习的算法通常是一个递归地选择最优特征，并根据该特征对训练数据进行分割，使得各个子数据集有一个最好的分类的过程。

在决策树算法中利用信息熵的概念，信息熵的计算公式：

$$H = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

其中 x_i 是变量。$p(x_i)$是变量 x_i 发生的概率

目标是创建一个模型，通过从数据特性中推导出简单的决策规则来预测目标

变量的值。

5.6 基于因子分析法与熵权法的模糊综合评价模型

5.6.1 因子分析法

由于预测模型可以给出多种预测评价因子，而因子分析是指研究从变量群中提取共性因子的统计技术，认为很多变量之间具有相关性，而这种相关性本质上是指多个变量可能存在着一个共同的影响因子，这一个影响因子就可以通过多个函数得到多个变量的近似。那么我们就可以用这一个隐变量去表示多个变量的信息。

具体步骤：

1. 相关性检验，一般采用 KMO 检验法和 Bartlett 球形检验法两种方法来对原始变量进行相关性检验，以确定原有若干变量是否适合于因子分析；
2. 计算样本的相关矩阵 R；
3. 求相关矩阵 R 的特征根和特征向量；
4. 根据系统要求的累积贡献率确定公共因子的个数；
5. 计算因子载荷矩阵 A；
6. 对载荷矩阵进行旋转，以求能更好地解释公共因子；
7. 确定因子模型；
8. 根据上述计算结果，求因子得分，对系统进行分析数学模型的公式化表示

给出因子分析法的公式化表示：

设$X=(X1,X2,\dots Xp)$为观测的随机向量且$E(X)=\mu, Var(X)=\Sigma$,则

$$X=\mu+AF+\varepsilon$$

$$E(F)=0, Var(F)=I_m$$

$$E(\varepsilon)=0, Var(\varepsilon)=D, Cov(F,\varepsilon)=0$$

5.6.2 基于熵权法的模糊评价模型

利用模糊数学理论，结合预测评价指标，针对现有方法的优点和存在的问题，利用评价模型的优劣程度由轻到重逐渐变化的模糊特性，可以获得更科学和更合理的评价结果。在预测结果模糊综合评价中，需确定影响模型优劣的各主要因素，确定评价因子集、评价集、隶属函数，然后通过计算各因素的权重和隶属度，得到综合隶属度，确定预测级别[4]。

熵权理论是一种客观赋权方法。在预测结果模糊评价中，通过对熵的计算确定权重，就是根据各项监测指标值的差异程度，确定各指标的权重。

基于熵权法的模糊评价模型整体过程：

1. 将原始矩阵进行标准化，消除评价指标的量纲所导致的差异性；
2. 计算不同指标的信息熵熵值；
3. 获取指标权值；
4. 利用熵权法对模糊评价法进行优化，定权；
5. 计算评价度 S；
6. 得到各个模型的综合评价结果。

实例计算结果表明，熵权法是一种高效准确的赋权方法，在预测模糊综合评价中有重要应用价值。

六、问题求解

6.1 问题 1

在问题 1 中，需要利用已知M_1, M_2, M_3, M_4与T_1, T_2共六个参数，预测I_A, I_B, I_C, I_D的四个参数。

6.1.1 问题 1 的求解过程

首先，对所给数据进行预处理，经过分析发现，所测量给出的数据的时间步长不一致，根据假设，进行数据的匹配预处理，同时将数据缺失点进行数据的清洗，得到下列数据：

表 2 问题 1 数据预处理结果

	T1	T_2	M_1	M_2	M_3	M_4	I_A	I_B	I_C	I_D
0	1173.63	813.92	49.24	90.38	46.13	28.16	78.15	26.21	12.93	14.59
1	854.55	767.64	49.24	90.38	46.13	28.16	78.39	25.22	12.93	14.28
2	855.34	767.99	49.24	90.38	46.13	28.16	79.22	24.6	12.41	13.7
3	853.57	766.2	49.24	90.38	46.13	28.16	79.52	23.88	11.55	13.56
4	854.81	768.08	49.24	90.38	46.13	28.16	80.04	23.48	11.55	13.47
5	594.59	686.44	49.24	90.38	46.13	28.16	80.51	22.41	11.72	13.38
……	……	……	……	……	……	……	……	……	……	……
232	1404.68	930.64	54.74	93.05	49.03	21.48	80.16	21.78	10.85	17.9
233	1404.85	931.16	54.74	93.05	49.03	21.48	79.79	22.58	11.2	17.05
234	1404.76	931.28	54.74	93.05	49.03	21.48	80.19	21.69	10.68	17.19

利用数据预处理后的 234 条数据，进行数据的预测，将数据集划分为训练集与测试集，其划分比例为 9：1，将输入参数设置为M_1, M_2, M_3, M_4与T_1, T_2的六个参数值，将输出参数设置为I_A, I_B, I_C, I_D的四个参数值，将设置好的数据集放入到基于遗传算法优化的 BP 神经网络、多元线性回归模型、Xgboost 预测模型、随机森林预测模型和决策树预测模型五种模型进行拟合预测。对预测结果进行评定，我们选取了包括均值误差（Bias）、均方误差（MSE）、均方根误差（RMSE）、绝对评价误差（MAE）、百分比误差（MAPE）、对称评价误差（SMAPE）、相关系数（R^2）等这一系列的预测模型的评价指标，利用基于因子分析法与熵权法的模糊综合评价模型的评价建立的预测模型。

因子分析法可得以上 7 组评价指标间具有极高的线性相关性，均大于 0.999，即可以仅利用其中的一组指标来评价模型的优劣，此处我们选取被绝大多数机器学习模型所采用的误差评价指标，即 MSE。

那么分别选取计算各个预测模型对于其四个输出参数分别的 MSE，由于难以通过直观法确定数据的优劣，因此采用熵权法定权，利用模糊分析法进行打分与评价，分别进行排名得到下述结果：

表 3 问题 1 模型评价结果

MSE	BPGA	多元线性回归方程	Xgboost 预测模型	随机森林预测模型	决策树预测模型	权值
指标 A	2.118898	35.75122	0.61946	0.588692	0.530758	0.217627
指标 B	4.087406	66.40706	0.908161	0.917643	1.151825	0.228627
指标 C	3.813388	7.391496	0.225529	0.250585	0.268829	0.314987
指标 D	19.23591	220.5902	4.560269	4.289639	7.759417	0.238758
评分	0.159253	0	0.953918	0.943377	0.789749	
排名	4	5	1	2	3	

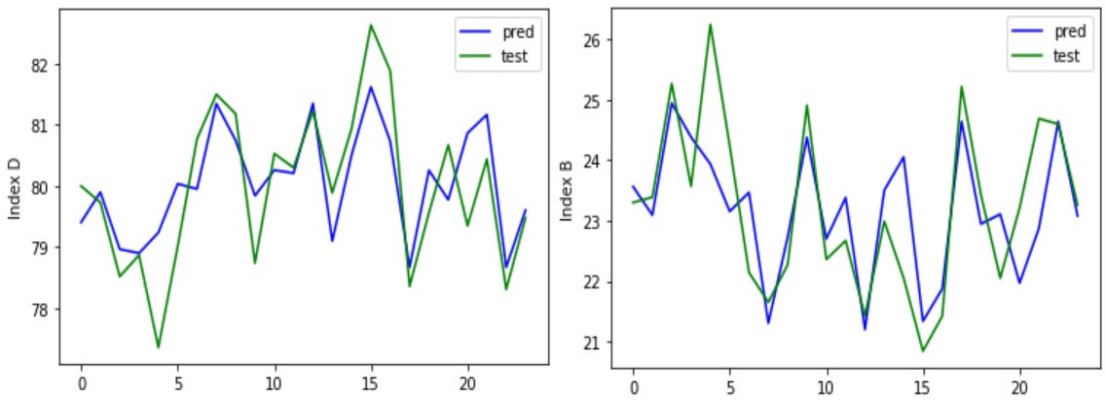
则我们发现对于使用 Xgboost 预测模型得分最高，即相对的 MSE 越小，即预测结果与真实值最接近的方法，视为最佳，同时认为利用该方法得出的预测结果是可能性最大的产品指标，则采用 Xgboost 的时序预测模型进行数据预测，将给出的 2022 年 1 月 23 号的数据进行输入，将预测结果进行输出，得到质量指标的预测结果，得到下表:

表 4 问题 1 模型预测结果

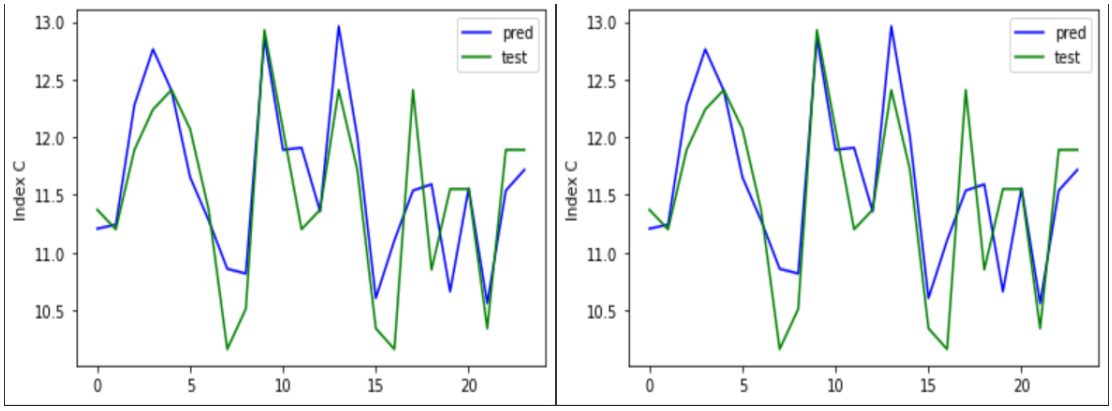
时间	系统 I 设定温度	系统 II 设定温度	指标 A	指标 B	指标 C	指标 D
2022-01-23	1404.89	859.77	80.15518	23.90084	11.51756	16.11099
2022-01-23	1151.75	859.77	79.39446	23.30358	12.43314	16.68909

6.1.2 问题 1 结果分析与检验

在进行预测结束后，我们同样绘制测试集的预测结果与真实结果的对比曲线图。



(a) (b)



(c) (d)

图 5 问题 1 结果对比曲线

(a) I_A 的对比曲线 (b) I_B 的对比曲线 (c) I_C 的对比曲线 (d) I_D 的对比曲线

我们发现真实值曲线与预测值曲线相对比，曲线的趋势一致，同样的曲线的拟合程度很好，偏差并不大，从而得知利用该模型预测两种不同的温度设定下的产品指标可能性最大。

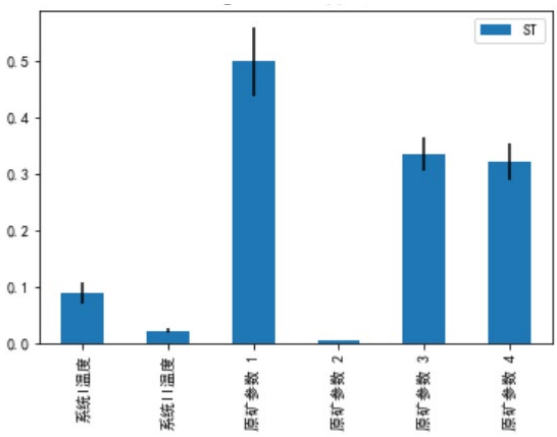


图 6 问题 1 结果敏感度分析

同时利用 Python 的 SALib 对于该模型进行敏感度分析，那么给出全阶指数的可视化敏感度图像，其主要是衡量模型输入对模型输出方差的贡献，我们可以看到模型对于T_1, T_2有着较好的敏感性，也就是可以做到在相同（或者相近）的系统温度下生产出来的产品质量可能有比较大的差别，满足题干的要求。

6.2 问题 2

在问题 2 中，需要利用已知的M_1, M_2, M_3, M_4和I_A, I_B, I_C, I_D的八个参数，估计与之相对应的T_1, T_2的两个参数。

6.2.1 问题 2 的求解过程

问题 2 还是对问题 1 的延续与拓展，即我们当前对T_1, T_2参数未知，但已知M_1, M_2, M_3, M_4参数，则我们可以继续利用问题 1 所预处理后的数据结果，进行数据的预测，将数据集划分为训练集与测试集，其划分比例为 9：1，仅需进行将输

入参数与输出参数进行重新设置，将输入参数设置为M_1, M_2, M_3, M_4与I_A, I_B, I_C, I_D的八个参数值，将输出参数设置为T_1, T_2的两个参数值，将设置好的数据集放入到基于遗传算法优化的 BP 神经网络、多元线性回归模型、Xgboost 预测模型、随机森林预测模型和决策树预测模型五种模型进行拟合预测。对预测结果进行评定，我们选取了包括 MSE 在内的 7 种预测模型的评价指标（与问题 1 一致），利用基于因子分析法与熵权法的模糊综合评价法的预测评价模型。

因子分析法可得以上 7 组指标同样具有极高的线性相关性，均大于 0.999，即可以仅利用其中的一组数据进行评价模型的优劣，此处我们依旧选取 MSE 作为我们预测模型评定的唯一指标。

然后利用与问题 1 类似的步骤，利用熵权法定权，利用模糊分析法进行打分与评价，分别进行排名得到下述结果。

表 5 问题 2 模型评价结果

MSE	BPGA	多元线性回归方程	Xgboost 预测模型	随机森林预测模型	决策树预测模型	权值
温度 1	166839	757933	30225.68	27073.9	17665.05	0.460922
温度 2	25164.05	113032.6	2742.437	2048.58	6098.036	0.539078
评分	0.073714	0	0.66498	0.835996	0.635413	
排名	4	5	2	1	3	

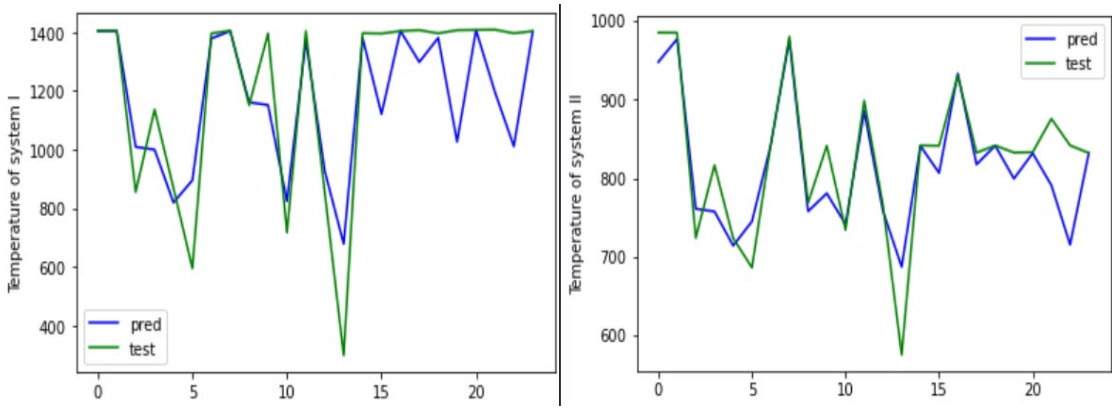
则我们发现对于使用随机森林时序预测模型给分最高，即相对的 MSE 越小，即预测结果与真实值最接近的方法，视为最佳，同时认为利用该方法得出的预测结果是可能性最大的系统温度设定，则采用随机森林时序预测模型进行数据的预测，将给出的 2022 年 1 月 24 号的数据进行输入，将预测结果进行输出，得到系统设定温度的预测结果，得到下表。

表 6 问题 2 模型预测结果

时间	指标 A	指标 B	指标 C	指标 D	系统 I 设定温度	系统 II 设定温度
2022-01-24	79.17	22.72	10.51	17.05	1399.69	851.3281
2022-01-24	80.10	23.34	11.03	13.29	1153.077	755.4725

6.2.2 问题 2 结果分析与检验

在进行预测结束后，我们同样绘制测试集的预测结果与真实结果的对比曲线图。



(a) (b)

图 7 问题 2 结果对比曲线

(a) T_1 的对比曲线 (b) T_2 的对比曲线

我们可以看到真实值曲线与预测值曲线相对比，曲线的趋势一致，同样的曲线的拟合程度很好，偏差并不大，从而得知利用该模型通过不同的产品指标推测所对应的系统设定温度是具有着最大可能性的。

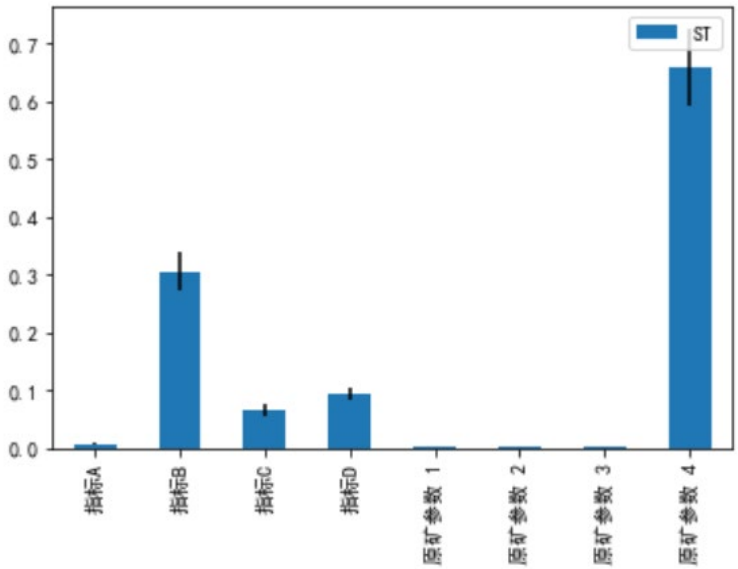


图 8 问题 2 结果敏感度分析

同样给出全阶指数的可视化敏感度图像，们可以看到模型对于I_{B}, I_C, I_D有着较好的敏感性，也就是可以做到同一组产品质量可能有多种调温方法都可以得到，满足题干的要求。

6.3 问题 3

在问题 3 中，需要利用给定的M_1, M_2, M_3, M_4和P_1, P_2, P_3, P_4，通过T_1, T_2预测R。

6.3.1 问题 3 的求解过程

和前面的问题类似，我们依旧是利用模型去预测R，同时我们又知道规定

I_A, I_B, I_C, I_D满足销售条件的产品为合格品，不满足为不合格品，即我们需要R与I_A, I_B, I_C, I_D存在着唯一确定的函数关系，因此我们可以利用模型首先预测I_A, I_B, I_C, I_D，再去利用函数关系映射到R上。

首先，由于新增了P_1, P_2, P_3, P_4以及数据范围的变化，我们需要重新对数据进行预处理。根据我们给出的对于过程参数的假设，我们将所给的数据利用 Python 进行匹配处理，即按每小时一条数据的标准进行处理，同时将数据缺失点进行数据的清洗，得到下列数据：

表 7 问题 3 第 1 次数据预处理结果

	T_1	T_2	M_1	M_2	M_3	M_4	P_1	P_2	P_3	P_4	I_A	I_B	I_C	I_D
1	1273.86	938.16	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	79.95	21.42	10.68	17.63
2	1273.51	937.49	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	80.2	21.2	10.16	16.92
3	1272.84	936.67	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	80.38	20.75	10.16	15.75
4	1404.5	959.9	55.26	108.03	43.29	20.92	1.25	3.09	242.44	164.45	79.75	23.3	11.55	15.84
5	1406.01	960.29	55.26	108.03	43.29	20.92	1.25	3.09	242.44	164.45	79.27	23.48	11.89	14.91
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……
1636	495.47	557.68	54.4	105.14	49.03	20.82	1.25	3.09	303.85	144.41	78.86	25.4	11.37	11.42
1637	494.41	572	54.4	105.14	49.03	20.82	1.25	3.09	313.31	172.28	79.1	25.58	11.37	11.55
1638	495.03	571.61	54.4	105.14	49.03	20.82	1.25	3.09	313.31	172.28	79.32	24.82	11.03	11.55

然后利用新的数据预处理后的 1638 条数据，进行数据的预测，将数据集划分为训练集与测试集，其划分比例为 9：1，将输入参数设置为原矿参数、系统温度以及过程参数的十个数据值，将输出参数设置为产品质量参数的四个数据值，将设置好的数据集放入到基于遗传算法优化的 BP 神经网络、多元线性回归模型、Xgboost 预测模型、随机森林预测模型和决策树预测模型五种模型进行拟合预测。对预测结果进行评定，我们选取了包括 MSE 在内的 7 种预测模型的评价指标（与问题 1 一致），利用基于因子分析法与熵权法的模糊综合评价法的预测评价模型。

因子分析法可得以上 7 组数据，同样具有极高的线性相关性，均大于 0.999，即可以仅利用其中的一组数据进行评价模型的优劣，此处我们依旧选取 MSE 作为我们预测模型评定的唯一指标。

然后利用与问题一类似的步骤，利用熵权法定权，利用模糊分析法进行打分与评价，分别进行排名得到下述结果。

表 8 问题 3 模型评价结果

MSE	BPGA	多元线性回归方程	Xgboost 预测模型	随机森林预测模型	决策树预测模型	权值
指标 A	3.589778	10.32482	0.386302	0.341544	0.612659	0.23903
指标 B	5.890188	18.21536	0.661089	0.766767	1.07771	0.227363
指标 C	5.919012	9.423675	0.325694	0.335157	0.636437	0.273718
指标 D	603.8859	130.0681	3.499776	2.621968	3.328074	0.259889
评分	0.039061	0.004128	0.905889	0.959479	0.605593	
排名	4	5	2	1	3	

我们同样可以发现对于预测产品参数而言，随机森林的模型得分最高，即相对的均方误差越小，即预测结果与真实值最接近的方法，视为最佳，同时认为利用该方法得出的预测结果是可能性最大的产品指标。

然后由于预测结果和真实结果存在着一定的相对偏差，则我们使用 Excel 的数据处理函数去分别判定测试集的真实的产品参数的指标所对应的合格与否和测试集的预测的产品参数的指标所对应的合格与否，然后利用 Python 去构建混淆矩阵模型，去判定这种预测性的分类方式是否存在着一致性，所得的混淆矩阵的热力学直方图如下图所示。

图 9 问题 3 结果混淆矩阵分析

其中，对于表征分类结果优劣性的混淆矩阵而言，其横轴表示的预测值的不同结果，纵轴表示的是真实值的不同结果，则我们可以将这四个区域依次标注为①、②、③、④，其中①号区域标识真阳性（TP），②号区域标识假阴性（FN），③号区域标识假阳性（FP），④号区域标识真阳性（TN），则该随机森林的预测分类模型的调和平均（F1）为 PPV=2*TP/(2*TP+FP+FN)=0.5909。

可以看到其调和平均较高，达到了具有很好平衡的精准率与召回率，同时我们可以去分析观察对角线的数据，可以发现真阳性与真阴性可以达到很好的数据比例，看颜色可以知道该随机森林的预测分类模型的准确度较高，其数据的部分偏差并不会对于判断数据的结果的分类。同时可以观察到这部分真实的合格率为 28.05%，而预测的合格率为 25.61%，两者的拟合偏差仅为 7.9%，两者不存在着较大偏差。

因此，可以做合理推断，仅对R作为唯一的数据输出结果，将I_A, I_B, I_C, I_D进行封装处理，直接利用包括M_1, M_2, M_3, M_4、T_1, T_2以及P_1, P_2, P_3, P_4的 10 个参数作为输入数据，直接获得唯一的R参数指标，是具有着极强标识性的，不会出现明显的数据偏差。

那么我们则对每天看作一个合格率的监测点，利用所给的每小时数据去进行合格率的计算，去进行数据的二次预处理，获得一个新的数据参数列，所得数据如下：

表 9 问题 3 第 2 次预处理数据结果

	T1	T_2	M_1	M_2	M_3	M_4	P_1	P_2	P_3	P_4	R
1	1273.86	938.16	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	0.2272727
2	1273.51	937.49	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	0.2272727

3	1272.84	936.67	55.26	108.03	43.29	20.92	1.25	3.09	226.16	181.23	0.2272727
4	1405.21	960.68	55.28	102.38	46.13	20.1	1.25	3.09	243.62	143.4	0.1666667
5	1404.19	959.9	55.28	102.38	46.13	20.1	1.25	3.09	243.62	143.4	0.1666667
… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …
1636	494.41	572	54.4	105.14	49.03	20.82	1.25	3.09	313.31	172.28	0.0909091
1637	495.03	571.61	54.4	105.14	49.03	20.82	1.25	3.09	313.31	172.28	0.0909091

然后利用新的数据二次预处理后的 1638 条数据，进行数据的预测，将数据集划分为训练集与数据集，其划分比例为 9：1，将输入参数设置为M_1, M_2, M_3, M_4、T_1, T_2以及P_1, P_2, P_3, P_4的 10 个数据值，而将输出数据仅设为R，将设置好的数据集放入到之前监测最优的随机森林时序方法进行二次的数据的拟合与预测。

将给出的 2022 年 4 月 8/9 号的数据进行输入，将预测结果进行输出，得到合格率的预测结果，得到下表：

表 10 问题 3 模型预测结果

时间	系统 I 设定温度	系统 II 设定温度	合格率
2022-04-08	341.40	665.04	0.258768
2022-04-09	1010.32	874.47	0.248601

6.3.2 问题 3 结果分析与检验

在进行预测结束后，我们绘制测试集的预测结果与真实结果的对比曲线图。

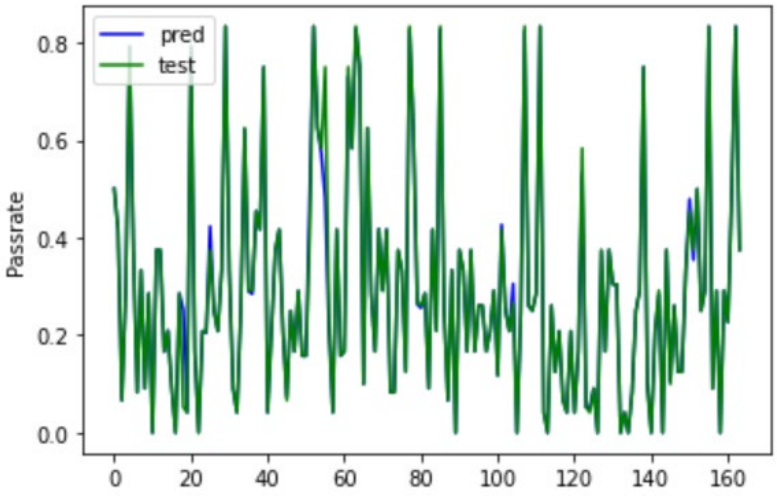


图 10 问题 3 结果对比曲线

我们可以看到这种预测的结果与真实的合理率几乎一致，仅存在部分的偏差，更加佐证了随机森林预测方法的结果是具有着很好的数据拟合性，我们使用合格率作为唯一参数是具有着正确性的。

6.4 问题 4

对于问题 4 而言，根据问题 3 中的结果，建立数学模型分析在指定合格率的条件下，如何设定系统温度的方法。

6.4.1 问题 4 的求解过程

由于R的本质实质上是对于I_A, I_B, I_C, I_D的概括性解读，即利用合格率逆推温度，我们缺失了其中间值，即I_A, I_B, I_C, I_D的准确性指标参数，因此，我们没有选择继续使用与上述三问类似的方式去，将合格率作为输入参数的一员，将设定温度作为输出参数的结果，去进行模型的拟合，评价，最后得到结果。而是选择继续利用问题三所搭建的随机森林时序预测模型（通过M_1, M_2, M_3, M_4、T_1, T_2以及P_1, P_2, P_3, P_4预测R的预测模型），将T_1, T_2继续作为输入参数，进行拟合尝试，以获取R，就是通过不断的改变输入值，以为了获得想要的数据输出。

由于同一组产品质量可能有多种调温方法都可以得到，那么对于类似的合格率可能会由差距较大的温度所获得，因此对于启发式算法是难以适用的，仅可能利用循环去绘制R曲面，寻求预测曲面与所需R平面的交点以得到合适系统设定温度。

首先，在样例数据中出现T_1, T_2数据最多的一段 100℃温度区间中，进行步长为 1℃的单个温度数据的改变，同时得到预测的R，进行曲线绘制。

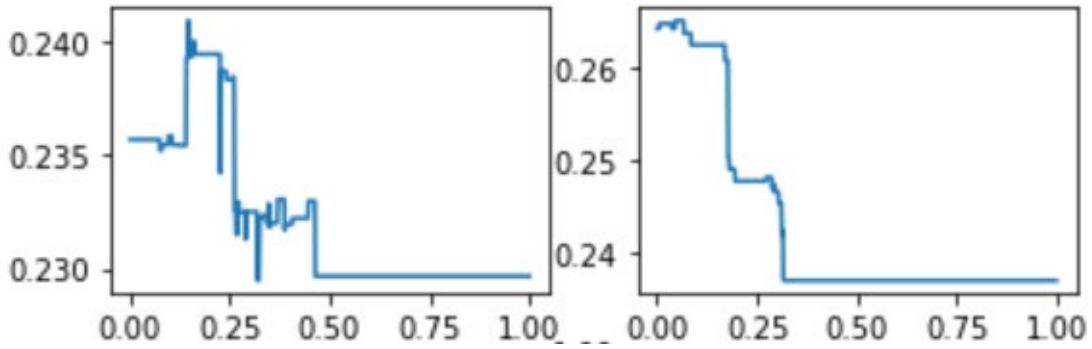
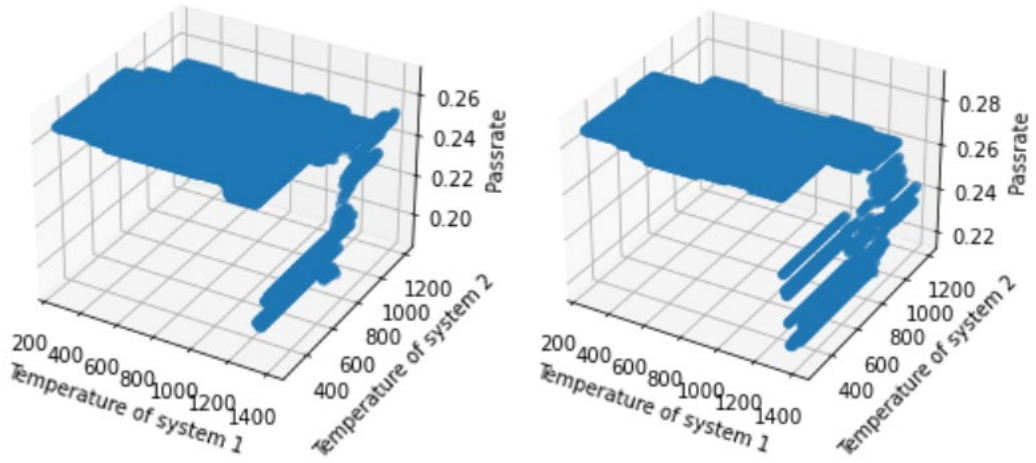


图 11 步长为 1℃的预测的R的变化

那么我们可以通过图像看出，对于单个温度的小幅度改变而言，预测所得的R是不会发生很大改变的，即R在温度改变时是阶段性改变的，在温度改变 15℃以内时，R不会出现超过 0.0001 的变化

因此我们可以加大步长，减少运算时间，同时扩大温度范围，获取更多的数据可能，以达到全方位的探索可能的数据结果，即将系统Ⅰ的设定温度设置在 200℃至 1500℃，将系统Ⅱ的设定温度设置在 300℃至 1300℃（根据所给系统温度的设定范围进行部分扩张所得），步长设置为 10℃，运行模型获取三维图像



(a) (b)

图 12 问题 4 结果二维曲线

(a) 2022-04-10 的二维曲线 (b) 2022-04-11 的二维曲线

那么我们就可以看到在系统的温度可能的范围内一切的合格率的结果，利用希望是合格率与该平面相交，看是否存在交点，补全问题结果：

表 11 问题 4 模型预测结果

时间	合格率	能否达到	系统 I 设定温度	系统 II 设定温度
2022-04-10	80%	否		
2022-04-11	99%	否		

6.4.2 问题 4 结果分析与检验

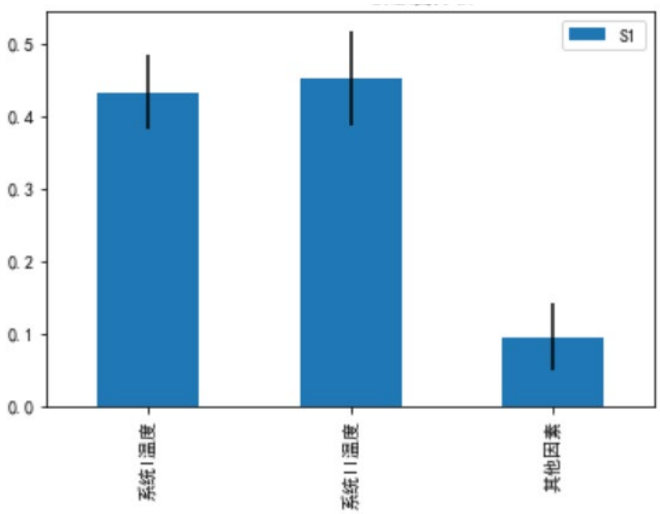


图 13 问题 4 结果敏感度分析

继续利用 Python 的 SALib 对于该模型进行敏感度分析，给出全阶指数的可

视化敏感度图像，我们可以看到模型对于T_1, T_2也有着较好的敏感性，也就是可以做到在相同（或者相近）的系统温度下生产出来的产品质量可能有比较大的差别，满足题干的要求，同时也侧面证明了我们实验模型与过程的合理性。

将随机森林的利用系统温度、原矿参数以及过程数据预测合格率的模型进行预测结果和真实结果进行可视化对比。

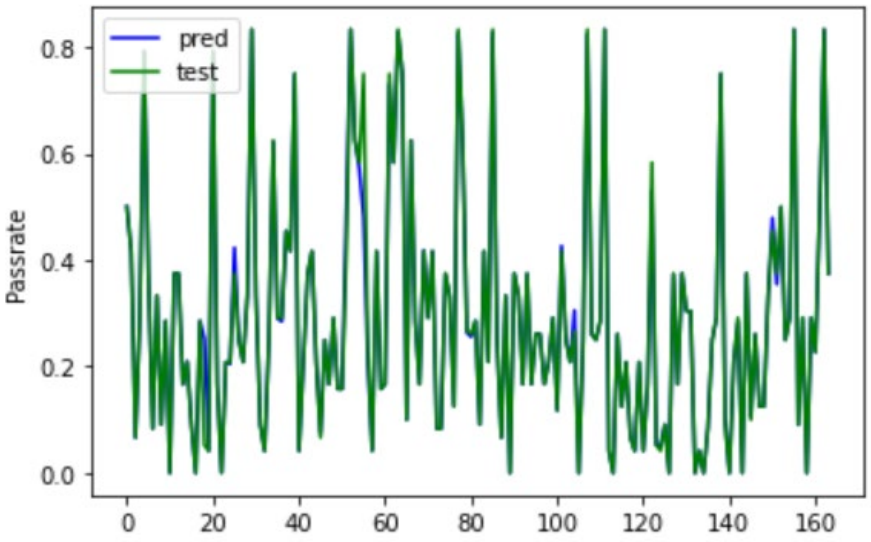


图 14 问题 4 结果对比曲线

我们可以发现这些曲线基本上相拟合，数据并未出现较大的偏差，仅有小幅度的扰动，同时利用随机森林预测模型的内置参数去进行预测的评价，评价分数可以达到 0.99 左右（越接近 1 越好），则随机森林模型对于合格率的预测准确度是可信的，准确性是可以保证的。

七、模型评价、改进与推广

7.1 模型的优点

1. 搭建基于因子分析法与熵权法的模糊综合评价模型，评价多个预测模型的预测的优劣性。因子分析法可以使得对预测模型的评价指标得到充分的利用，对预测模型的评价角度充分全面；熵权法可将同一评价指标的多个模型样本点结合确定权重，考虑了多个样本间的联系,可削弱异常值的影响,使评价结果更准确、合理；模糊综合评价模型则通过精确的数字手段处理模糊的评价对象，能对蕴藏信息呈现模糊性的资料作出比较科学、合理、贴近实际的量化评价。该预测综合评价模型，可以很好地对于搭建的多个预测模型进行全面、准确、贴近现实的评估，充分有力地证明最终所选的预测模型的优越性与合理性；

2. 在面对问题需求时，利用预测综合评价模型进行评估，可以做到按照实际情况具体地选取预测模型方案，使得预测结果更具准确性，避免出现单一模型存在狭隘性及模型合理性难以证明的问题；

3. 对所给数据进行科学化的数据清洗，更好地保证了预测结果的稳定性与准确性；

4. 所建立的模型有效解决了小数据集机器学习难以避免的数据过拟合及离

群点影响过大的问题。

7.2 模型的缺点

1. 预测模型主要基于机器学习进行，对其过程数据难以给出很好的解释，自身行为不确定性较高；

2. 模型仅能实现输入到输出的直接预测，欠缺对于输入参数间内在联系的把握。

7.3 模型的改进

1. 对于机器学习预测模型，给出更多限制与定义，从而实现数据结果的可控化，将确定性算法与启发式算法相结合，以实现预测结果的准确、高效、可解释；

2. 深层次分析输入参数，挖掘其内在联系性，对预测模型进行修正，更好的提高预测结果的准确性；

3. 扩大所用预测模型的范围，使用更多的传统预测模型，或是深度学习时序预测模型，扩大评估范围。

7.4 模型的推广

1. 若给出确定的矿石处理原理，或参数指标的实际物理意义，即可进一步分析联系性，对于特定的过程可以拟合确定的处理曲线，应用于实际工业生产，根据希望的产品指标调控系统设定温度，做到实时反馈与调节，尽可能的提升原矿处理的效率，做到节能与效益的双赢；

2. 若给出数据量更大的数据集，通过对大数据量的全面学习，获得更为准确的预测结果，同时可以根据预测结果，得到矿石处理系统的主要影响因素，并给出针对不同原矿参数的系统温度设定建议；

3. 给出不同原矿处理系统的数据集，通过学习处理，发掘不同系统间的联系，提高所给模型的普适性。

八、参考文献

[1] 刘浩然,赵翠香,李轩,王艳霞,郭长江.一种基于改进遗传算法的神经网络优化算法研究[J].仪器仪表学报,2016,37(07):1573-1580.DOI:10.19650/j.cnki.cjsi.2016.07.017.

[2] 黄卿,谢合亮.机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析[J].数学的实践与认识,2018,48(08):297-307.

[3] 郭澎涛,李茂芬,罗微,林清火,唐群锋,刘志崴.基于多源环境变量和随机森林的橡胶园土壤全氮含量预测[J].农业工程学报,2015,31(05):194-200+202+201.

[4] Zou Z H, Yi Y, Sun J N. Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment[J]. Journal of Environmental sciences, 2006, 18(5): 1020-1023.