钢水“脱氧合金化”配料方案的优化

摘要

本文通过对附件和参考资料数据的分析，主要运用线性回归模型、N 折交叉验证、XGB(XGBoost)优化算法、贝叶斯超参数优化，遗传算法等方法研究了钢水“脱氧合金化”配料方案的优化问题。

针对问题一，结合炼钢过程中的真实情况与附件所提供的数据，由于缺失脱氧合金化后从钢水中排出钢渣的质量数据，无法确定反应后的钢水净重，因此我们提出三种计算方案。在考虑反应前后由于加入合金料与排出钢渣引起的钢水质量的变化下：**方案 a.**）假设反应前钢水中不存在钒(V)元素，建立模型求解；**方案 b.**）利用给出的转炉终点值的磷(P)元素，建立模型求解。在不考虑反应前后加入合金料与排出钢渣引起的钢水质量的变化下：**方案 c.**）历史数据中加入合金料的质量小于钢水净重的 2.7%，带来的误差可忽略不计，建立模型求解。最终根据模型交叉检验结果和真实情况下元素收得率的大小，方案 c 计算出的 C，Mn 元素的历史收得率最为合理，平均值分别为 92%与 89.2%。通过多变量线性回归分析推断 C 元素的收得率主要与加入的钒铁、硅铁等附带有一定脱氧性的合金料和转炉终点时 C 含量相关，Mn 元素的收得率也主要与加入的硅铁等附带有一定脱氧性合金料和转炉终点时 Mn 含量相关。同时在查阅相关资料后，也证实了脱氧剂对元素收得率影响这一因素。

针对问题二，在问题一的基础上，我们构建了线性回归模型对 C、Mn 两种元素的收得率进行了预测，常用的线性回归模型调优方式有在原有模型中添加正则化项构成 Lasso 回归以减少模型过拟合程度等。然而本模型在数据中并未呈现过拟合现象，因此我们换用准确率更高的 XGB 模型[1]，并采用 N 折交叉验证和贝叶斯超参数优化的方式智能调参。最终模型的在 Mn、C 元素历史收得率数据中表现的均方误差(MSE)分别为 0.00019 和 0.0031，较原线性回归模型的预测误差分别降低 88%和 51.5%。

针对问题三，需要建立脱氧合金化成本优化模型。能给出最优的配料方案。模型优化的目标由 **1)合金成本函数**与 **2)连铸正样中元素与目标值差值**两部分组成。通过查阅资料发现，P 和 S 元素的最终成分受投入合金外因素影响，无法进行有效的预测，同时我们通过计算也验证了这一点。所以最终模型的优化目标设定为总成本较历史数据每吨钢水平均成本下降百分比与 C，Mn，Si 三种元素最终含量与目标值差距百分比的加权值。因为国家标准中钢种化学成分范围与目标值差距在 10%左右，借鉴神经网络中 Sigmoid 激活函数对硬阈值函数的改进，不采用硬阈值函数，而设定权重比为 1:10:10:10。这样也将有约束的多目标优化问题转化为单目标优化问题。然后将目标函数代入遗传算法作为适应度函数进行智能优化求解。由于计算成本问题，在模型演示部分我们按照历史记录中的前 3 条样本进行计算，最终平均成本降低 30.91%，连铸正样各元素含量平均在目标值的 96%以上。具体内容请见下文与附件。

针对问题四，结合建模过程中得到的数据与模型以及对数据进行的一些有效探索为炼钢厂在保证钢水质量并最大限度减少合金钢生产成本以及指标控制两个方面,以及数据管理与数据埋点,操作规范等三个方面通过建议信提出可行性建议。

由于代码实现与数据附录较多，我们在本文附录中仅给出了附录文件目录，方便老师在支撑材料中查看和测试

关键字： 回归分析 N 折交叉验证 XGB 预测模型 贝叶斯超参数优化 遗传算法 Python

目录

一、问题重述

炼钢过程中的脱氧合金化是钢铁冶炼中的重要工艺环节。对于不同的钢种在熔炼结束时，需加入不同量、不同种类的合金，以使其所含合金元素达标，最终使得成品钢在某些物理性能上达到特定要求。随着钢铁行业中高附加值钢种产量的不断提高，如何通过历史数据对脱氧合金化环节建立数学模型，在线预测并优化投入合金的种类及数量，在保证钢水质量的同时最大限度地降低合金钢的生产成本，是各大钢铁企业提高竞争力所要解决的重要问题。国外从上世纪九十年代开始研究计算机自动配料，到目前为止，已经形成了具备以合金收得率预测及成本优化算法为主体的自动配料模型，该模型可以实现自动脱氧合金化的功能。国内钢铁企业除部分车间具有引进的脱氧合金化模型外，其他炼钢车间尚未采用这一技术，而是按照不同元素的固定收得率或经验值计算各种合金的加入量，难以实现当前炉次合金配料的自动优化和成本控制。

合金收得率指脱氧合金化时被钢水吸收的合金元素的重量与加入该元素总重量之比。在钢水脱氧合金化过程中，合金收得率受多种因素影响，难以采用显式表达式确定。通过对低合金钢种前期冶炼数据的采集，得到历史真实数据附件。

建立合适的数学模型，分析并解答以下问题。

问题 1：钢水脱氧合金化主要关注 C、Mn、S、P、Si 五种元素的含量，请根据附件 1 计算 C、Mn 两种元素历史收得率，并分析影响其收得率的主要因素。

问题 2：在问题 1 的基础上，构建数学模型，对 C、Mn 两种元素收得率进行预测，并进一步改进模型及算法，尽可能提高这两种元素收得率的预测准确率。

问题 3：不同合金料的价格不同，其选择直接影响钢水脱氧合金化的成本。请根据问题 2 中合金收得率的预测结果及附件 2,建立数学模型，实现钢水脱氧合金化成本优化计算，并给出合金配料方案。

问题 4：请根据你们的研究结果，给炼钢厂领导写一封建议信（一页以内）。

二、模型的假设

1.假设所有合金料中所含元素的收得率与合金料的收得率相同；

2.假设在对每种元素的“连铸正样”数据采集前已经完成对钢水的排渣处理；

3.假设空气中没有气体没有参与钢水合金脱氧合金化过程；

4.假设脱氧合金化反应前后钢水质量不变

5.假设加入合金料对钢水温度没有影响

三、符号说明

为了便于描述问题，我们在此列出文中主要使用一些符号和基本变量，其他一些变量将在文中陆续说明（以硅(Si)元素为例）。

符号	代表含义
$M^{Si}_{合金配料}$	某合金配料中硅元素总质量
$M^{Si}_{加入}$	脱氧合金化时加入的硅元素的总质量
$M_{吸收Si}$	脱氧合金化时被钢水吸收的硅元素的总质量
$p^{Si}_{钒铁}$	硅元素在“钒铁”合金料中所占重量百分比
$M^{Si}_{加入}$	脱氧合金化时加入的硅元素的总质量
$p^{Si}_{硅锰面}$	硅元素在“硅锰面”合金料中所占重量百分比
$M^{Si}_{吸收}$	脱氧合金化时吸收的硅元素的总质量
$p^{Si}_{连铸正样}$	脱氧合金化之后钢水中硅元素的含量
$p^{Si}_{转炉终点}$	脱氧合金化之前钢水中硅元素的含量
$M^{Si}_{加入}$	脱氧合金化时加入的硅元素的总质量
α_{Si}	硅元素的收得率
$Cost$	成本
c_i	某合金料每吨的价格

四、问题分析

4.1 问题一的分析

由于附件一每条炉号数据中包含的子数据较多，为了得到与本体更加直观的数据，使用 Excel 对数据进行清洗和预处理，随后导入 Python 中 pandas 模块运算与分析。

首先，考虑排出钢渣引起的钢水质量的变化，为了计算出 C、Mn 两种元素的历史收得率，首先根据数据给出的信息，可以利用钢水净重与转炉终点时元素所占百分比求得脱氧合金化前各元素在钢水中的质量。利用附件一中脱氧合金化时投入的合金料的质量

与附件二合金料成分数据可以计算出加入某种元素的总重量$M_{加入}$。通过查阅合金料的收得率，结合转炉终点和连铸正样值与钢水反应前的质量，逆推求出脱氧合金化后钢水的重量。再通过计算钢水中 C，Mn 反应前后的质量差得到被吸收的总质量$M_{吸收}$，其与$M_{加入}$的比即为 C，Mn 两种元素的历史收得率。由于附件给出的转炉终点数据只有碳、锰、硫、磷和硅元素，我们首先假设反应前钢水中不存在钒(V)元素，在**方案 a** 中通过钒(V)元素计算。分析**方案 b** 时考虑到主要影响锰元素[2]，硅元素重量变化的合金料是脱氧反应中的重要的还原剂脱氧剂锰硅，最终与大量碳元素和少量硫元素结合在钢水中上浮形成SiO_2与Al_2O_3等钢渣被排出[3]，计算碳、锰、硫和硅无法准确代表脱氧合金化时钢水一共吸收的重量，因此方案 b 使用选用磷元素计算。

然后，在**方案 c** 中我们不考虑反应前后钢水质量的变化，观察附件一中历史炼钢数据发现，每条炉号数据中加入合金料的质量均小于钢水净重的 2.7%，故可以将其忽略不计。通过$M_{钢水} \times (p_{连铸正样} - p_{转炉终点})$计算出元素被吸收的总质量，进而计算碳和锰元素的历史收得率。

通过观察发现计算出的收得率数值范围的实际意义最为合理,然后以转炉终点时钢水的七种的物理特性与投入合金的重量为自变量进行回归分析,并利用传统 t 检验进行假设验证以及五折交叉验证两种方式,来验证模型的有效性.虽然 t 检验中有数个参数 p 值较大,在不拒绝零假设的情况下需要承担风险但是因为 C 和 Mn 交叉验证的 MSE 分别为 0.0060 和 0.0016，因此从机器学习角度考虑,认为模型参数合理。

4.2 问题二的分析

本体要求在问题一计算出的 C、Mn 两种元素历史收得率的基础上构建模型并改进以提高预测准确度。通过问题一中对 C、Mn 两种元素历史收得率的回归分析，我们已经采用最小二乘估计构建了基础的线性回归模型。通过输入变量预测 C、Mn 元素的收得率，观察模型预测的数据，判断模型是否存在为了贴近数据而产生了过拟合现象，如果出现则在模型中引入正则化构建 Lasso 回归或岭回归。

同时我们考虑引入 XGB 预测模型，并使用贝叶斯超参数优化[4]的方式对模型智能调参，搜索对于 XGB 模型更合适的基尼指数记性最优切分特征决策树(CART 决策树)的深度(max_depth)，树的个数(n_estimators)和学习率(learning_rate)三个重要参数。目前主流的超参数调优方法有网格搜索和随机搜索等，但是由于赋值超参数后每次都需要重新对模型进行评价。出于效率的考虑，我们采用贝叶斯超参数优化方法，使模型可以充分利用之前的评价信息，从而减少对超参数搜索的尝试次数。以此构建更高效、更准确和更低复杂度的预测模型来提升对 C、Mn 元素收得率预测的准确度，并对模型进行先关检验，与原线性回归模型的预测准确度做出对比。

4.3 问题三的分析

由之前的问题可知，脱氧合金化过程中投入合金料的数量和比例不同，不同元素的收得率不同，总成本不同。而其主要影响因素就是钢水的初始状况和投入的合金料。而我们要得到就是合金料的自动优化模型，该模型需要在已知钢水的初始情况时，自动计算出最佳配料方案，但是由于我们进行脱氧合金化是为了得到我们需要的合金，而不同合金的钢种具有一定的国家标准，同时材料中也给出了钢厂炼钢时的目标值。因此模型的优化目标需要有两个，第一总成本尽量低，第二，得到的最终合金成分接近目标值。为了求解该有约束优化问题，采用遗传算法构建求解模型，并自定义成本函数作为遗传算法的适应度函数，自定义成本函数由 **1)合金成本函数**与 **2)连铸正样中元素与目标值差值**两部分组成。

在对附件的数据清理之后，发现除个位数的 Q345B 样本外，几乎都是化学成分标准一致的 HRB 系列钢种，同时 Q345B 钢的化学标准基本与 HRB 一致，因此我们这里使用材料中给出的钢种化学成分目标值进行计算,不再添加其他不同元素含量内控范围的钢号.为了减少数量级带来的误差,因此选择成本较历史记录中每吨钢水脱氧合金化的总成本下降的百分比与元素含量与目标值含量差距的百分比作为两个部分组合时候的实际数值。由于连铸正样中化学成分的目标值与国家标准的内控区间的差异基本在上下 10%左右。借鉴神经网络的激活函数中，Sigmoid，Relu 函数对硬阈值函数的改进，我们这里采用权重比为 1:10:10:10 的加权值的形式作为我们的目标函数，同时这样也将有约束的多目标优化问题转化为单目标优化问题。为了测试该模型在模型演示部分，按照历史记录中的前 3 条样本进行计算，结果的平均成本降低 30.91%，大幅度降低了成本。同时 C，Mn，Si 三种元素与目标值的差距的平均值，即误差的平均值分别为 3.94%，2.05%，3.84%。不仅完全在国家标准范围内，同时也很接近目标值，均能达到目标值含量的 96%以上，模型效果较好。

4.4 问题四的分析

在前三问的基础上，集合金属所得率模型求得的结果，为炼钢工厂提供数据预测，如何能够最大限度降低合金钢的生产成本的改进方案。通过结果与数据分析所得的结果，对炼钢工厂的未来发展提出了可行性的意见与建议报告，希望能够帮助工厂提高在钢铁企业中的竞争力。

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据的分析和处理

附件一中给出的炼钢历史数据中，每条炉号数据中包含的子数据较多。1.)对于问题一中 Mn 元素收得率的分析中比较重要的子数据包括转炉终点 Mn、连铸正样 Mn 等。

但导入 Excel 后发现发现转炉终点 Mn 数据在 7A05154~7A06618 炉号数据中缺失，连铸正样 Mn 数据在 7A05154~ 7A06059 炉号数据中缺失，由于缺失后对于计算收得率和分析模型没有贡献，故进行计算的时候全部剔除。同时还存在少量数据出现转炉终点温度，转炉终点 C，转炉终点 S 与转炉终点 P 等回归分析中自变量数据缺失的情况，加入它们会对模型精度造成折扣，故也将其删除。2.)对于问题一中 C 元素收得率的分析中比较重要的子数据包括连铸正样 C 等，但是由于转炉终点 Mn 也作为 C 元素收得率的影响因子之一，需要将缺失转炉终点 Mn 的 C 元素相关数据剔除。最终将问题一计算分析 Mn 元素历史收得率可用的 222 条和计算分析 C 碳元素的历史收得率可用的 116 条数据保存为 csv 文件，方便后期用 pandas 导入做运算。

5.1.2 计算各方案的元素收得率

计算某个元素的收得率需要用到的数据为加入所有合金料中某种元素的总重量$M_{加入}$，钢水脱氧合金化时元素吸收的总质量$M_{吸收}$。通过计算 $\frac{M_{吸收}}{M_{加入}}$ 即可计算出要求的元素的收得率。

其中，元素的$M_{吸收}$值可以通过附件一中加入的不同合金料重量的$M_{合金配料}$、附件二中元素在不同合金料中所占的质量比重得出。如问题一中加入的锰元素的质量为：

$$加入锰元素的质量：\ M_{加入}^{Mn} = M_{硅锰面} \times p_{硅锰面}^{Mn} + \ldots + M_{锰硅合金} \times p_{锰硅合金}^{Mn}$$

在脱氧合金化时共吸收的锰元素的质量为：

$$吸收锰元素的质量：\ M_{吸收}^{Mn} = M_{反应后钢水} \times p_{连铸正样}^{Mn} - M_{反应钢水} \times p_{转炉终点}^{Mn}$$

（一）通过方案 a 求解收得率

考虑反应前后由于加入合金料与排出钢渣引起的钢水质量的变化，假设在脱氧合金化反应前钢水中不存在钒元素，通过附件二可计算出在脱氧合金化过程中加入的钒元素的重量。

$$M_{加入}^{V} = M_{钒氮合金} \times p_{钒氮合金}^{V} + \ldots + M_{钒铁(FeV50-A)} \times p_{钒铁(FeV50-A)}^{V}$$

由此计算出的所有炉号的$M_{加入}^{V}$值，因为脱氧合金化反应前钢水中不存在钒元素，故钢水中增加的钒元素全部来源于加入的合金料，通过查阅资料获得钒铁矿钒元素（V）的平均收得率为 96.5%，

所以，

$$\alpha_V = \frac{M_{反应后钢水} \times p_{连铸正样}^{V}}{M_{钒氮合金} \times p_{钒氮合金}^{V} + \ldots + M_{钒铁(FeV50-A)} \times p_{钒铁(FeV50-A)}^{V}}$$

解得

$$M_{反应后钢水}=\frac{M_{钒氮合金}\times p^{V}_{钒氮合金}+\ \ldots\ +M_{钒铁(FeV50-A)}\times p^{V}_{钒铁(FeV50-A)}}{p^{V}_{连铸正样}}\times\alpha_{V}$$

可计算出所有炉号反应后钢水的质量数据，具体分布见图 1。

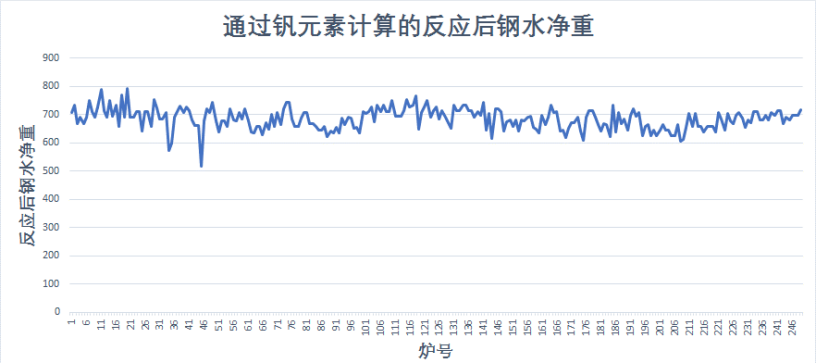


图 1 通过钒元素计算的反应后钢水净重

通过图 1 可以看出，求得的钢水净重普遍小于 800 千克，这显然与真实情况相违背。分析公式可以看出主要原因为 $p^{V}_{连铸正样}$ 的值较大而导致 $M_{反应后钢水}$ 较小，说明原钢水中存在钒元素。因此假设不成立，**方案 a 不合理**。

（二）通过方案 b 求解收得率

考虑反应前后由于加入合金料与排出钢渣引起的钢水质量的变化，通过附件二可计算出在脱氧合金化过程中加入的磷元素的重量。

$$M^{P}_{加入}=M_{钒铁(FeV50-A)}\times p^{P}_{钒铁(FeV50-A)}+\ \ldots\ +M_{硅铁(合格块)}\times p^{P}_{硅铁(合格块)}$$

由此计算出的所有炉号的$M^{P}_{加入}$值，因为脱氧合金化反应前钢水中存在钒元素，故钢水中增加的钒元素为反应后钢水中磷元素所占的百分比与反应前磷元素所占百分比的差，通过查阅资料获得硅锰面（硅锰渣）等合金料中磷元素（P）的平均收得率为 92.5%，

所以，

$$\alpha_{V}=\frac{M_{反应后钢水}\times p^{P}_{连铸正样}-M_{反应前钢水}\times p^{P}_{转炉终点}}{M_{钒铁(FeV50-A)}\times p^{P}_{钒铁(FeV50-A)}+\ \ldots\ +M_{硅铁(合格块)}\times p^{P}_{硅铁(合格块)}}$$

解得

$$M_{反应后钢水}=\frac{M^{P}_{加入}\times\alpha_{V}+M_{反应前钢水}\times p^{P}_{转炉终点}}{p^{P}_{连铸正样}}$$

可计算出所有炉号反应后钢水的质量数据，具体分布见图 2。

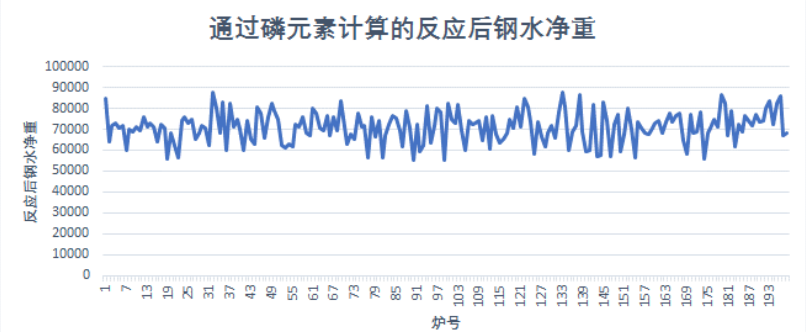


图 1 通过磷元素计算的反应后钢水净重

通过图 2 可以看出，求得的钢水净重普遍在 70000 千克左右，符合真实情况中钢水的重量。所以根据

$$\alpha_{C,Mn} = \frac{M_{反应后钢水} \times p^{C,Mn}_{连铸正样} - M_{反应前钢水} \times p^{C,Mn}_{转炉终点}}{M^{C,Mn}_{加入}}$$

即可求出方案 b 下 C 元素和 Mn 元素的收得率。具体数值分布为图 3，图 4。

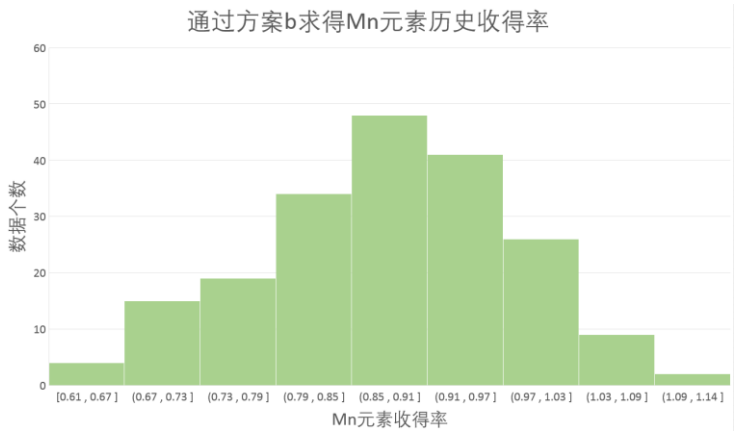


图 3 通过方案 b 求得 Mn 元素历史收得率数值分布

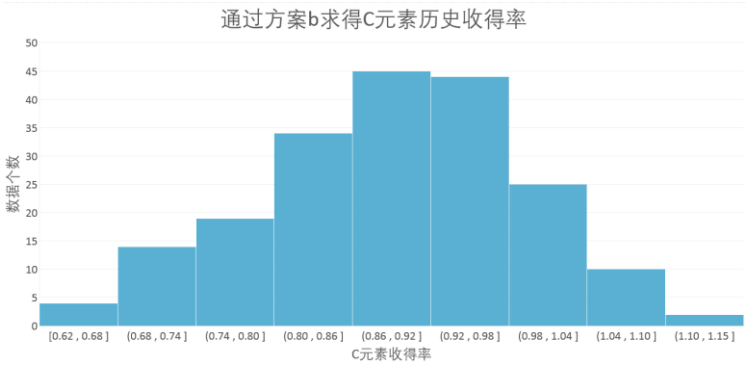


图 4 通过方案 b 求得 C 元素历史收得率数字数值分布

通过图 3，图 4 可以看出，通过方案 b 求得 Mn 元素与 C 元素的历史收得率数值分布十分分散，数据方差较大，因此模型在“模型检验与回归分析”部分中有待进一步检验。

（三）通过方案 c 求解收得率

通过方案 a 与方案 b 都未能得到充分合理的历史收得率数据，在不考虑反应前后加入合金料与排出钢渣引起的钢水质量的变化下，脱氧合金化前的钢水净重与脱氧合金化后的钢水净重相等。所以根据

$$\alpha_{C,Mn} = \frac{M_{钢水净重} \times (p^{C,Mn}_{连铸正样} - p^{C,Mn}_{转炉终点})}{M^{C,Mn}_{加入}}$$

即可求出方案 c 下 C 元素和 Mn 元素的收得率。具体数值分布为图 5，图 6。

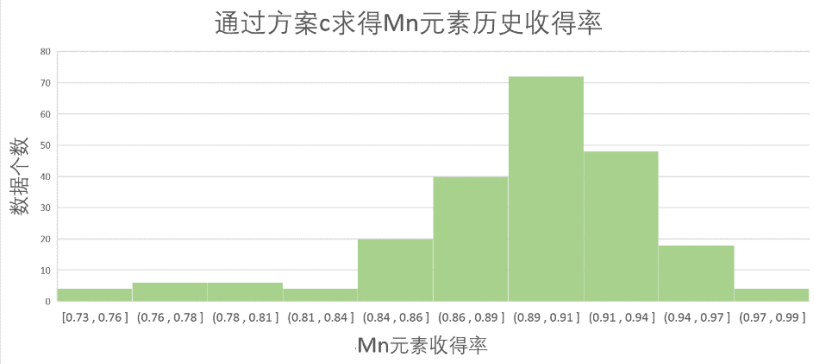


图 5 通过方案 c 求得 Mn 元素历史收得率数字数值分布

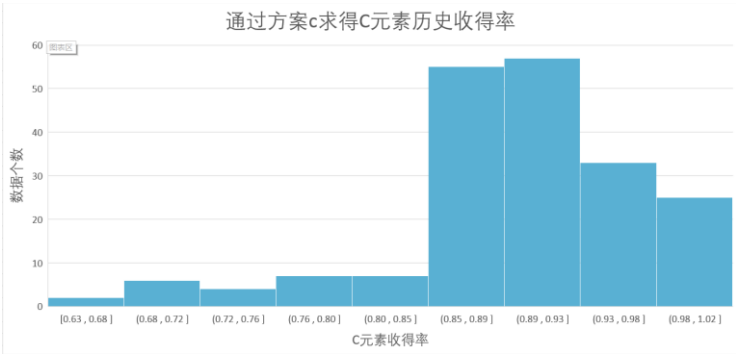


图 6 通过方案 c 求得 C 元素历史收得率数字数值分布

通过图 5，图 6 可以看出，通过方案 c 求得 Mn 元素与 C 元素的历史收得率数值分布较为集中，其中 Mn 元素的历史收得率普遍集中在 86%~92%,C 元素的历史收得率普遍集中在 85%~94%，数据相对较为合理。

5.1.3 建立求解模型，并检验与回归分析

通过方案 b 和方案 c 对 C，Mn 两种元素历史收得率的求解，我们采用多变量线性回归模型分析元素收得率与加入各个合金料的重量、转炉终点时各个元素的含量等因子线性关系，并最终在模型检验通过后得出的影响 C 元素和 Mn 元素收得率的主要因素。所以，我们提出以下模型

$$\alpha = k_0 + k_1 p^{C}_{转炉终点} + k_2 p^{Mn}_{转炉终点} + \cdots + k_{n-i} M_{硅铁(合格块)} + k_{n-i+1} M_{锰硅合金} + \cdots + k_{21} M_{转炉终点温度}$$

其中α为元素的收得率，$k_1 \ldots k_n$为多变量线性回归中各因子的你和系数，k_0为拟合残

差。模型利用最小二乘法[5]对数据进行拟合。其中模型的，

$$X=\begin{bmatrix} k_0 \\ k_1 \\ k_2 \\ \vdots \\ k_{n-i} \\ k_{n-i+1} \\ \vdots \\ k_{21} \end{bmatrix} \qquad \beta=\begin{bmatrix} 1 \\ p^{C}_{转炉终点} \\ p^{Mn}_{转炉终点} \\ \vdots \\ M_{硅铁(合格块)} \\ M_{锰硅合金} \\ \vdots \\ M_{转炉终点温度} \end{bmatrix} \qquad y=\alpha$$

根据最小二乘法的计算原理 $y=X\beta$，模型需要优化 β 使 $(y-X\beta)^T(y-X\beta)$ 靠近最小值，即

$$\frac{\partial(y^Ty-2y^TX\beta+\beta^TX^TX\beta)}{\partial\beta}=0$$

得到：

$$\hat{\beta}=(X^TX)^{-1}X^Ty$$

便可通过 $\hat{y}=X\hat{\beta}$ 得到元素目标收得率 $\hat{y}$ 在回归直线上$x=X$ 时的拟合值。将方案 b 和方案 c 中求解的 C 元素、Mn 元素历史收得率导入模型求解，同时根据模型的 21 个因子的权重系数k_i即可得出他们对 C 元素和 Mn 元素各自收得率的贡献度，即对结果的影响程度。

在对方案 b 数据使用最小二乘估计后，在对 Mn 元素预测值的模型检验时发现模型的决定系数R^2为-0.728，五组交叉验证结果分别为[0.390，0.555，0.120，0.024，-4.683]，与最优指标 1 相差甚远；同时模型均方误差MSE为 2.404，五组交叉验证结果分别为[-1.826，-0.523，-1.895，-1.262，-6.513]，因此方案 b 的数据所拟合的回归方程效果较差，模型不合理。

而在对方案 c 数据使用最小二乘估计后，在对 Mn 元素和 C 元素预测值的模型检验时发现模型的决定系数R^2分别为 0.726 和 0.678，较为靠近最优指标 1；同时模型均方误差MSE分别为 0.0016 和 0.0064，拟合的回归方程效果较好，模型合理。各因子的贡献度大小见表 1。

表 1　各因子对 C，Mn 元素贡献率的大小

参数	对 C 元素收得率的贡献度	对 Mn 元素收得率的贡献度
转炉终点温度	-6.17938e-04	-1.24417e-3
转炉终点 C	-9.14411e-02	-2.05757e-4
转炉终点 Mn	-1.77883e-03	-1.55452e-2
转炉终点 S	4.63004e-03	-4.07091e-3
转炉终点 P	-6.23188e-03	3.83748e-3
转炉终点 Si	-6.50159e-03	-1.42593e-4

钢水净重	3.95624e-02	3.15108e-2
氮化钒铁 FeV55N11-A	2.73873e-03	6.49430e-4
低铝硅铁	7.11237e-17	9.71445e-17
钒氮合金(进口)	3.88954e-02	-3.41295e-3
钒铁(FeV50-B)	1.96583e-02	-1.06332e-2
硅铝钙	-1.24090e-02	-3.19721e-3
硅铝合金 FeAl30Si25	-2.63413e-02	-2.43977640e-03
硅锰面（硅锰渣）	-1.08292e-02	-4.61035370e-02
硅铁(合格块)	-1.04083e-17	-3.46944695e-17
硅铁 FeSi75-B	7.78509e-03	1.50122307e-02
石油焦增碳剂	-6.76518e-02	-8.43659343e-04
锰硅合金 FeMn64Si27(合格块)	-3.66350e-02	-1.72720250e-01
锰硅合金 FeMn68Si18(合格块)	-4.42156e-02	-1.73745229e-01
碳化硅(55%)	-1.57859e-02	-1.79490813e-03
硅钙碳脱氧剂	-5.94544e-03	-2.90007632e-03
偏置(k_0)	0.90266	0.89276

由表 2 可知，C 元素的收得率主要与加入的钒铁、硅铁等附带有一定脱氧性的合金料和转炉终点时 C 含量相关，Mn 元素的收得率也主要与加入的硅铁等附带有一定脱氧性合金料和转炉终点时 Mn 含量相关。同时在查阅相关资料后，也证实了脱氧剂对元素收得率影响这一因素。

5.2 问题二模型的建立与求解

5.2.1 分别建立 C、Mn 元素收得率的线性回归方程并预测

根据问题一中线性回归模型，

$$\alpha = k_0 + k_1 p^{C}_{转炉终点} + k_2 p^{Mn}_{转炉终点} + \cdots + k_{n-i} M_{硅铁(合格块)} + k_{n-i+1} M_{锰硅合金} + \cdots + k_{21} M_{转炉终点温度}$$

利用最小二乘法求解出每个因子的系数k_i，我们可以求得 C，Mn 元素收得率的线性回归方程。将附件一中 21 个自变量数据输入可以得到此线性模型下对 C，Mn 元素收得率预测[6]的结果，部分预测数据见表 2 和表 3，完整数据见附录附件九与附件十一。

表 2　线性回归模型预测 C 元素收得率

炉号	C 元素收得率预测值	炉号	C 元素收得率预测值
7A06878	0.918466	7A06759	0.904953
7A06877	0.907154	7A06758	0.920215
7A06876	1.025897	7A06757	0.971313
7A06875	0.921321	7A06756	0.93894
7A06874	0.935047	7A06755	0.901914
7A06873	0.978644	7A06754	0.986114

7A06872	1.020854	7A06753	0.867822
…	…	…	…
7A06813	0.862048	7A06629	0.977764
7A06812	0.882503	7A06628	0.966798
7A06811	0.938053	7A06627	0.962272
7A06810	0.762517	7A06625	0.95071
7A06809	0.850342	7A06624	0.919376
7A06808	0.877	7A06623	0.987108
7A06807	0.856642	7A06622	0.970652
7A06805	0.904106	7A06620	0.939678

表 3　线性回归模型预测 Mn 元素收得率

炉号	Mn 元素收得率预测值	炉号	Mn 元素收得率预测值
7A06878	0.875469	7A06759	0.880712
7A06877	0.876305	7A06758	0.884685
7A06876	0.96014	7A06757	0.898138
7A06875	0.905179	7A06756	0.865087
7A06874	0.90746	7A06755	0.884832
7A06873	0.891456	7A06754	0.858507
7A06872	0.932257	7A06753	0.916609
…	…	…	…
7A06813	0.893761	7A06629	0.93237
7A06812	0.87457	7A06628	0.944797
7A06811	0.915622	7A06627	0.941651
7A06810	0.819459	7A06625	0.933106
7A06809	0.883007	7A06624	0.948934
7A06808	0.884415	7A06623	0.951398
7A06807	0.871213	7A06622	0.95591
7A06805	0.884558	7A06620	0.941113

根据预测结果，对此模型回归模型进行检验，在这里我们采用 N 折交叉验证的方式，将 C，Mn 历史收得率数据随机的分成 N 份，每次通过随机函数选择 N-1 份作为训练集构建模型，剩下的 1 份做测试集检验模型预测结果。在线性回归模型中常用的检验指标有决定系数R^2和均方误差MSE，决定系数解释了模型对观测值的拟合程度，其值贴近于 1 表示模型效果较好。计算公式为

$$R^2 = 1 - \frac{SSE}{SST} \qquad SSE = \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \qquad SST = \sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2$$

其中，SSR 代表模型的回归平方和，SSE 代表模型残差平方和。均方误差 MSE 表现了预测的数据的变化程度，MSE 的值越小，说明预测模型描述的数据具有更好的精确度。 计算公式为

$$\mathrm{MSE} = \frac{SSE}{n} = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$$

最终通过五折交叉验证的平均值得出，Mn 元素和 C 元素收得率预测模型的决定系数R^2分别为 0.726 和 0.678；均方误差MSE分别为 0.0016 和 0.0064。

5.2.2 构建 XGB 模型，提升数据预测准确率

XGB(eXtreme Gradient Boosting-XGBoost)是一个监督模型，将每棵 CART 树叶子节点对应的分数预测值加到一起作为最终预测值，其中 CART 决策树是采用基尼指数记性最优切分特征的决策树。这种算法与决策树回归和线性回归相比有利于实现高效的优化算法。其模型为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

其中 $\hat{y}_i$ 是预测出的元素收得率，K是树的数量，f表示一棵具体的基尼指数记性最优切分特征决策树。F表示所有可能的基尼指数记性最优切分特征决策树。

模型的总损失函数即目标函数为：

$$\mathrm{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

第一项为要优化的损失函数，即预测的元素收得率和真实元素收得率的差。第二项为避免过拟合加入的正则项，这里的正则项由 K 棵树的正则化项相加而来。每一棵基尼指数记性最优切分特征决策树主要需要确定个两部分，首先是树的结构，这个结构将输入的数据进行计算，映射到某个确定的叶子节点当中；第二部分是各个叶子节点的分数值。可以将叶子节点的分数值作为参数，但树的结构无法作为参数，因此采用加法训练来分步骤优化目标函数。公式如下：

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

...

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

对于 XGB 模型中的正则化项$\Omega(f_k)$，其数学模型为

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

其中 T 为叶子节点数，λ表示 L2 正则化系数，γ 表示节点切分难度。对第 t 棵树的优化目标做如下变形：

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2\right] + \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{i=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma \mathrm{T}$$

得到最终目标函数：

$$Obj^* = \frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma \mathrm{T}$$

5.2.3 使用贝叶斯超参数优化算法，对 XGB 模型智能调参

得到此题 XGB 预测模型的目标优化函数后，在求解时模型时很重要的步骤是如何选择超参数找到我们的最优超参数arg^*，使得

$$arg^* = \mathrm{argmax}\ \mathrm{Obj}^*(arg)$$

由于贝叶斯超参数调优模型主要基于高斯分布模型[7]，其通用算法框架为：

Algorithm 1 Bayesian Optimization

1: **for** $t = 1, 2, \ldots$ **do**
2: Find $\mathbf{x}_t$ by optimizing the acquisition function over the GP: $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}}\, u(\mathbf{x}|\mathcal{D}_{1:t-1})$.
3: Sample the objective function: $y_t = f(\mathbf{x}_t) + \varepsilon_t$.
4: Augment the data $\mathcal{D}_{1:t} = \{\mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t)\}$ and update the GP.
5: **end for**

最终通过贝叶斯超参数优化后，C 元素历史收得率的 XGB 预测模型中基尼指数记性最优切分特征决策树的深度(max_depth)参数为 1，树的个数(n_estimators)为 59，学习率(learning_rate)为 0.10193；Mn 元素历史收得率的 XGB 预测模型中基尼指数记性最优切分特征的决策树的深度(max_depth)参数为 1，树的个数(n_estimators)为 293，学习率(learning_rate)为 0.49963。

5.2.4 用最终的 XGB 模型对数据进行预测，并检验模型预测准确率

将附件一中 21 个自变量数据输入可以得到次线性模型下对 C，Mn 元素收得率预测的结果，部分预测数据见表 4 和表 5，完整数据见附录附件十和附件十二。

表 4　XGB 模型预测 C 元素收得率

炉号	C 元素预测值	炉号	C 元素预测值

7A06878	0.9303199	7A06759	0.9238099
7A06877	0.86456203	7A06758	0.89950955
7A06876	0.9787952	7A06757	0.9303199
7A06875	0.91045046	7A06756	0.92835855
7A06874	0.9456304	7A06755	0.9303199
7A06873	0.9232216	7A06754	0.94361526
7A06872	0.94604385	7A06753	0.89047027
…	…	…	…
7A06813	0.9109757	7A06629	0.92126024
7A06812	0.9284166	7A06628	0.9109177
7A06811	0.9284166	7A06627	0.92835855
7A06810	0.79216635	7A06625	0.9109177
7A06809	0.8446442	7A06624	0.8975482
7A06808	0.8976062	7A06623	0.92835855
7A06807	0.8976062	7A06622	0.9109177
7A06805	0.888567	7A06620	0.92835855

表 5　XGB 模型预测 Mn 元素收得率

炉号	Mn 元素预测值	炉号	Mn 元素预测值
7A06878	0.874677	7A06759	0.882211
7A06877	0.907495	7A06758	0.878263
7A06876	0.962293	7A06757	0.890541
7A06875	0.900517	7A06756	0.88195
7A06874	0.929349	7A06755	0.874553
7A06873	0.908761	7A06754	0.834491
7A06872	0.91769	7A06753	0.910615
…	…	…	…
7A06813	0.896383	7A06629	0.93697155
7A06812	0.901509	7A06628	0.9407592
7A06811	0.917518	7A06627	0.9256418
7A06810	0.836415	7A06625	0.9407592
7A06809	0.882072	7A06624	0.9245715
7A06808	0.878848	7A06623	0.94099295
7A06807	0.879279	7A06622	0.9312757
7A06805	0.879864	7A06620	0.93710107

根据预测结果，与原线性回归模型采用相同的五折交叉检验，得到 C、Mn 元素收得率 XGB

预测模型的均方误差MSE分别为 0.00217 和 0.00019，较原线性模型预测误差分别降低了 51.5%和 88%。

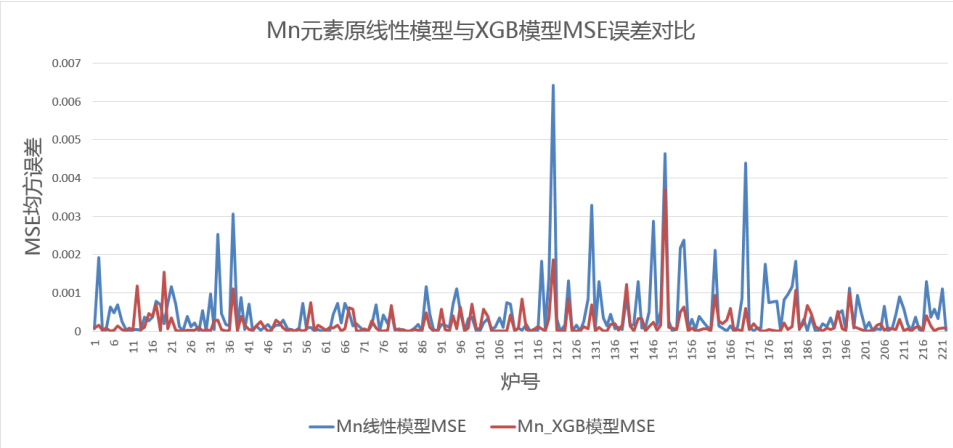


图 7 Mn 元素原线性模型与 XGB 模型 MSE 误差对比

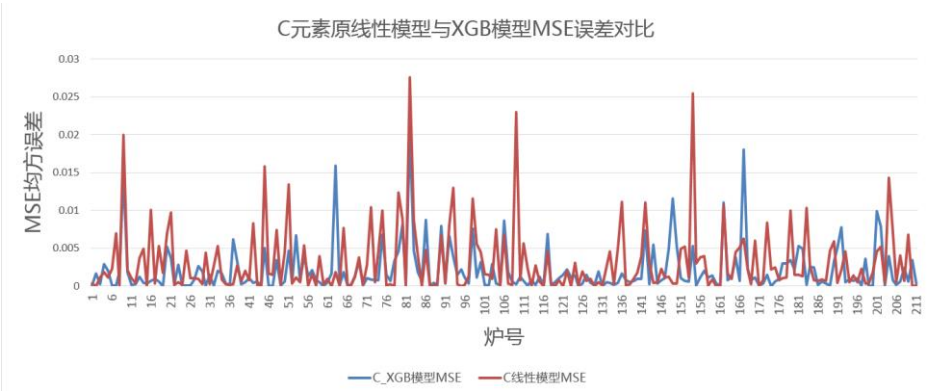


图 8 C 元素原线性模型与 XGB 模型 MSE 误差对比

5.3 问题三模型的建立与求解

根据问题描述，我们需要得到的是脱氧合金化成本优化模型。该模型的作用为，当需要将一炉钢水加入不同合金料进行脱氧合金化时，进行自动配料，使脱氧合金化的成本最低。

这里我们模型的优化目标应该由两个部分组成，第一部分是脱氧合金化总成本，第二个是脱氧合金化后五种元素与钢种中五种目标含量的差距。总成本尽量小，而元素成本差距即需要尽量小，并且还有不能超过国家标准中元素含量内控区间的影响要求。

从数学角度考虑，该问题数属于有约束优化问题[8]，但是由于目标函数中成本计算需要使用我们构建的所得率预测模型来计算，而我们给出的所得率计算模型为 xgboost 模型，所以我们此处很难给出一个有约束优化问题的具体解法。面对这种情况，我们可以使用智能优化算法来解决该问题。这里我们选取遗传算法作为最终方法，将我们的优化目标作为遗传算法中的适应度函数来，将投入不同合金料的重量作为遗传算子来求得自动配料方案。

而遗传算法相较于一般求解方法的好处在于适应范围比较宽广。当后期我们需要对新的钢种进行自动化配料与成本优化时，只需要修改相应的预测模型与目标函数中的钢号的元素目标值即可。极大的提高了程序的通用性。

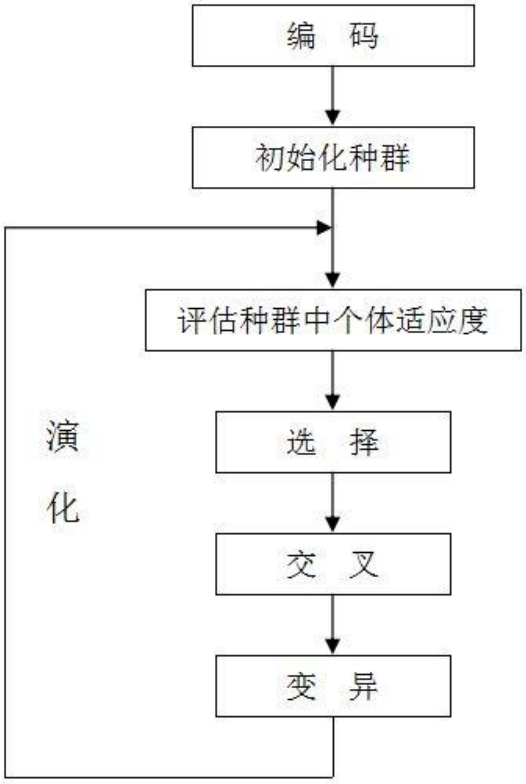
5.3.1 遗传算法简介

对于求解函数最大或最小值问题，一般可以描述为以下数学模型

$$\begin{cases} \max f(X) \\ X \in R \\ R \subset U \end{cases}$$

其中 X 为决策函数，$max\ f(x)$或者$min\ f(x)$为目标函数，$X \in R$，$R \subset U$ 为约束条件。遗传算法是人工智能领域进化学派中一种用于解决最优化问题的搜索启发式算法。该算法借鉴了进化生物学中一些研究而来。包括进化学说中的，DNA，遗传，变异，基因交叉，自然选择等内容。但是遗传算法是否能够求解目标函数的全局最优值，很大程度上由适应度函数确定，不过我们也可以通过模型多次运算结果进行一定的初步判断。

5.3.2 遗传算法具体步骤



(a) 初始化，设定最大计划次数 T，迭代计数器 t=0，随机生成 N 个遗传算子的族群作为初始族群 P(t)

(b) 个体评价，通过适应度函数计算族群中每一个遗传算子的适应度
(c) 选择运算，通过设定的保留个数或者保留比例，选择复数最佳遗传算子保留
(d) 交叉运算，将交叉算子作用于保留得遗传算子，新的遗传算子
(e) 变异运算，将变异运算作用于交叉运算后新的遗传算子.得到的新的遗传算子，即为下一代族群 P(t+1)，t=t+1。
(f) 终止条件判断，如果 t=T，终止计算，将通过适应度函数计算得到的最佳遗传算子作为最优解输出，否则重复 b-f 几个步骤。

5.3.3 适应度函数的确定

(1) 适应度函数第一部分：连铸正样元素含量与钢号成分目标值的差距

我们首先对五种元素的所得率进行计算并按照问题二中高准确率模型的构建方法构建相应模型，并以此计算钢水连铸正样时五种元素的最终含量。但是根据计算结果发现，P，S 的收得率十分不稳定并有转化率超过 100%的现象发生。

同时即使只留下少部分所得率在正常范围内的数据，然后根据第二题的方法构建高准确率预测模型，其预测结果比较差。其中基于 P 构建的预测模型 MSE 为 1.370，绝对误差为 0.955。其中基于 S 构建的预测模型，MSE 为 41.264，绝对误差为 5.185。只有 Si 元素的预测模型较好,其误差中 MSE 为 0.000540，绝对误差为 0.0177。具体计算过程与构建模型详见附件：预测模型_整合。查阅资料发现，P、S 两种元素的所得率不仅仅受投入合金种类与数量影响。其他操作，如氧含量情况，合金料投入时间等等也有很大影响。因此根据数据，模型以及资料推断，数据和模型的较大误差可能是由数据外的未记录因素导致的。所以最终决定目标函数中元素与目标值差距这一部分,我们只选取 C，Mn，Si 这三种元素进行计算。根据所得率计算连铸正样中元素的成分的公式如下：

$$p_{连铸正样}^{C,Mn} = \frac{M_{加入}^{C,Mn} \times \alpha_{C,Mn} + M_{反应前钢水} \times p_{转炉终点}^{C,Mn}}{M_{反应后钢水}}$$

考虑到钢种不同,元素的目标值也不同。经过数据清洗后的数据中,主要为:HRB500D，HRB400D，HRB400B，Q345B，四种钢号的钢种。由于三种 HRB 型号的钢种国家标准一致，而钢种 Q345B 仅有个位数级别的样本且与 HRB 型号的目标值基本重合，因此我们这里采用题目中给予的 HRB400B 的化学成分目标值我们计算中统一的目标值。

(2) 适应度函数中的第二部分：总共加入合金料的成本

由于成本与元素百分比数量级上的差距，因此选择成本较历史记录中每吨钢水脱氧合金化的总成本下降的百分比与元素含量与目标值含量差距的百分比作为两个部分组合时候的实际数值。计算可得历史记录中每吨钢水的脱氧合金化成本为 343.06933462317204 元/吨。计算过程与代码详见附录附件十九和附件二十。根据投入的合金料与题目附件二计算合金料总成本的公式如下(以 Mn 元素为例)：

$$Cost_{加入合金料成本} = \sum_{i=0}^{n} M \times c_i$$

5.3.4 适应度函数中权重的取值

对于每吨钢水的成本部分其计算结果为百分率的浮点小数形式，而元素差距亦同。同时由于连铸正样中化学成分的目标值与国家标准的内控区间的差异基本在上下百分之十左右。同时借鉴神经网络的激活函数中，Sigmoid，Relu 函数对硬阈值函数的改进，我们这里采用权重比为 1 比 10 比 10 比 10 的加权平均和的形式作为我们的目标函数，同时这样也将有约束的多目标优化问题转化为单目标优化问题[9]。目标函数公式为：

$$Cost = Cost_{\text{加入合金料成本}} + 10 \times \left(P^{Mn}_{\text{连铸正样}} - P^{Mn}_{\text{目标值}}\right) + 10 \times \left(P^{C}_{\text{连铸正样}} - P^{C}_{\text{目标值}}\right) + 10 \times \left(P^{Si}_{\text{连铸正样}} - P^{Si}_{\text{目标值}}\right)$$

这样做的好处在于避免了硬阈值函数中位于限定范围外的数据无法进行有效的交叉于变异操作的问题。对遗传算法的迭代收敛有一定的加速过程。

5.3.5 求解结果演示

由于计算成本问题，每次计算不同的样本都需要一定的时间，因此我们这里给出历史数据中前三条样本的最终配料方案，作为模型的演示结果。

由样本一的计算结果可知，合金成本相较于原来下降了 31.22%，降低了接近三分之一的成本。 C，Mn，Si 三种元素与目标值的差距，仅仅是原目标值的：5.94% ，0.37% ，1.31%。相较于原目标值差距十分的小，完全达到了最终要求。

由样本二的计算结果可知，此时合金成本相较于原来下降了 26.90%，降低了四分之一以上的成本。Mn，Si 三种元素与目标值的差距，仅仅是原目标值的：0.25% ，4.94% ，8.69%。相较于原目标值差距十分的小，完全达到了最终要求。

由样本三的计算结果可知，此时合金成本相较于原来下降了 34.61%，降低了三分之一以上的成本。C,Mn,Si 三种元素与目标值的差距,仅仅是原目标值的:5.65% ,0.85% ，1.57%。相较于原目标值差距十分的小，完全达到了最终要求。得出的配料方案具体见表 6。

表 6　成本优化后自动配料配料方案

参数	样本一配料方案	样本二配料方案	样本三配料方案
转炉终点温度	1.64400000e+03	1.65100000e+03	1.68400000e+03
转炉终点 C	6.50000000e-04	4.10000000e-04	3.50000000e-04
转炉终点 Mn	1.10000000e-03	1.10000000e-03	1.10000000e-03
转炉终点 S	3.00000000e-04	3.70000000e-04	2.40000000e-04
转炉终点 P	1.40000000e-04	2.80000000e-04	2.00000000e-04
转炉终点 Si	4.00000000e-03	3.00000000e-03	4.00000000e-03
钢水净重	7.44000000e+04	7.00000000e+04	7.82500000e+04
氮化钒铁 FeV55N11-A	0	0	0
低铝硅铁	0	0	0
钒氮合金(进口)	0	0	0
钒铁(FeV50-B)	0	0	0
硅铝钙	6.25637515e+01	6.25637515e+01	6.25637515e+01
硅铝合金 FeAl30Si25	1.01752049e+02	1.01752049e+02	1.01752049e+02
硅锰面（硅锰渣）	8.71979330e+02	8.71979330e+02	8.71979330e+02
硅铁(合格块)	0	0	0
硅铁 FeSi75-B	0	0	0
石油焦增碳剂	1.12937479e+02	1.12937479e+02	1.12937479e+02

锰硅合金 FeMn64Si27(合格块)	1.10268315e+03	1.10268315e+03	1.10268315e+03
锰硅合金 FeMn68Si18(合格块)	0	0	0
碳化硅(55%)	0	0	0
硅钙碳脱氧剂	1.45328436e+02	1.45328436e+02	1.45328436e+02

5.4 问题四的建议信内容

尊敬的钢厂领导：

您好！通过在该问题求解与建模的过程中一些有效的认识和探索，对于如何保证钢水质量的同时最大限度减少合金钢生产成本以及工厂规范化等方面，提出以下几点建议，希望能够对贵工厂的发展有所帮助。

一、使用智能系统而非人的经验

合金元素收得率是影响合金料使用量最重要的判断因素，但在实际生产中收得率是统计一段时间内合金加料记录得出来的，缺乏实时性，与炉次实际的收得率存在差距。同时根据金属所得率模型表明，不同元素所得率与投入合金的变化关系并不完全线性一致，尤其考虑到五种元素的预估，其应该更接近于非凸函数。因此仅靠人的经验来预估最终结果往往会产生不可避免的误差，同时当投料比例与以往经验差距较大时，误差也可能会变得更大。因此建议贵公司建立氧合金化的自动配料与成本优化系统,摆脱经验计算,从而准确实时的获取不同钢种的合金元素收得率的自动配料方案。

二、转炉终点时各种特性的影响

数据和模型证明，转炉终点时炉中钢水的七种特性对所得率也有一定影响，但是我们缺乏对转炉终点时钢水特性的有关控制，因此最好也建立相关的系统进行智能预测而非进行人力计算或者估计。

三、完善转炉终点含有的许多元素未检测数据

我们在问题一第一小问的第二种计算方案中发现，转炉终点时的钢水除含有五种元素以外还含有钒等其他元素，但是这些数据在转炉终点中并没有，但是这些金属可能会对其他元素所得率产生一些影响，并一定会对钢的最终性能有一定影响。因此建议检测相关内容，并记录有关数据。

四、加强操作规范化建设：

磷元素对钢的性能影响较大，但在数据中即使投料相近，磷元素所得率的值很并不是很稳定，因此根据数据我们无法有效的预测磷元素所得率与最终成分。经查阅资料，我们认为可能是由于脱磷时的操作不统一造成的。因此建议加强对相关操作的规范化建设。

五、优化数据管理

本问题中钢水净重只给出了一个数据，但经查阅资料得知，在脱氧合金化过程中，钢水质量必然会发生一定变化，这一变化如果忽略会产生许多误差，如：多个数据显示 C 元素的收得率超过 100%，同时这对所有元素的最终所得率都会产生较大影响,最终影响自动配料与成本优化系统的准确率。

因此，希望厂方能够注重数据质量，尽量减少数据缺失问题，同时最好对其他可记录的数据提前做好"数据埋点"，存储更多数据，以备不时之需。在新时代，建立智能化工业已成必然，但智能化不仅仅是多些传感器，多些技术人员这么简单，依靠数据与算法，建立有效，高效的现代化的智能系统才是重中之重。

以上是我们的几点建议，希望能够对贵工厂的发展起到一定作用。同时也真诚希望贵公司能尝试我们的自动配料方案模型，如果有疑问请及时与我们联系，我们会进一步根据您提供的情况对模型进行改进。谢谢！

六、模型的评价与推广

问题一，根据附件一给出的数据使用不同的计算方案计算量 C 和 Mn 的历史所得率，并选取了最佳的方案三。然后根据线性回归模型对收得率进行回归分析，并利用交叉验证与基于 t 检验的假设验证进行了计算。得出了不同因素对 C 和 Mn 两种元素所得率的影响，为之后的计算提供了数据支持。

问题二，在问题一的基础上，使用贝叶斯优化与交叉验证对 XGBoot 模型进行智能调参得到准确率更高的模型。大幅度降低了模型误差。

问题三，在问题一二的基础上，借鉴神经网络中 Sigmoid 函数对硬阈值函数的改进，建立自定目标函数，并将其作为遗传算法的适应度函数，建立基于 XGB 拟合模型与遗传算法的脱氧合金化自动化配料与成本优化模型。然后将历史数据中的前三条样本代入数据进行求解，得出合理，有效，低成本的自动配料方案，证明了该模型的高效性。

本文所建立的拟合模型具有较高的准确率，性能较强，且计算成本较低。自动化配料与成本优化模型的不仅计算结果高效合理，并且有较强的普适性，不同于线性规划，只需要替换钢种的目标值与拟合模型就可以进行有效的迁移。具有较广的适用范围。

七、参考文献

[1] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System[P]. cornell university,2016.

[2] 韩培伟. 锰硅铁合金炉外精炼的基础研究[D].北京科技大学,2017.

[3] 文松,柴晓伟,姜杰.工业混合气脱氧剂的研究[J].盐业与化工,2015,44(03):10-14.

[4] 崔佳旭,杨博.贝叶斯优化方法和应用综述[J].软件学报,2018,29(10):3068-3090.

[5] 王媛. 线性回归模型的二阶最小二乘估计[D].北京交通大学,2016.

[6] 冷建飞,高旭,朱嘉平.多元线性回归统计预测模型的应用[J].统计与决策,2016(07):82-85.

[7] J. Snoek, H. Larochelle, and R. Adams.:Practical Bayesian Optimization of Machine Learning Algorithms[P]. NIPS, page 2960-2968. (2012)

[8] 叶秉如,方道南.大型多目标线性规划解法研究[J].水科学进展,1995(04):270-277.

[9] 白鹤松.基于多目标线性规划的决策模型研究[J].哈尔滨理工大学学报,2008,13(06):57-59.

八、附录

附件目录：

注意：附件中的程序为 .ipynb 格式,需要通过 jupyter notebook 打开,其文件在运行的同时可以以 markdown 格式保存和显示图片,文档.因此部分计算结果和可视化结果依旧保存在原文件中.打开后无需运行即可查看上次运行的结果。文件全部用 utf 格式保存,否则程序读取会出错。

附件一：根据 P 计算的收得率.csv ：问题一第一小问求所得率的第二种方案时,使用磷

计算出的所得率结果。
附件二：data_0_c.csv ： 在 excel 中初步清洗后的数据,程序加载后利用清洗规则进一步筛选数据。
附件三：data_0.csv ： 利用程序进一步清洗后的数据,不需要再次清洗即可使用。
附件四：df_data_1_C_plot.csv ： 含有 C 所得率的数据
附件五：df_data_1_Mn_plot.csv ： 含有 Mn 所得率的数据
附件六：df_data_1_P_plot.csv ： 含有 P 所得率的数据
附件七：df_data_1_S_plot.csv ： 含有 S 所得率的数据
附件八：df_data_1_Si_plot.csv ： 含有 Si 所得率的数据
附件九：C_Predict.csv：回归分析时，线性回归模型预测的 C 所得率结果
附件十：C_Predict_XGB.csv：问题二中，XGB 模型预测的 C 所得率结果
附件十一：Mn_Predict.csv：回归分析时，线性回归模型预测的 Mn 所得率结果
附件十二：Mn_Predict_XGB.csv：问题二中，XGB 模型预测的 Mn 的所得率结果
附件十三：money_0.csv ： 原题中附件二中的数据

附件十四：第一问_方案二.ipynb ： 问题一第一小问求所得率的第二种方案的代码
附件十五：所得率计算_整合.ipynb ： 问题一求所得率的第三种方案
附件十六：随机森林_回归分析.ipynb ： 基于随机森林计算特征重要程度的代码
附件十七：回归分析_New.ipynb:第二问的大部分代码的整合,包括数据清洗,C 和 MN 的基于五折交叉验证的线性回归,基于留一法和留 P 法(p=2)的线性回归,全部数据的回归分析,以及 P 值检验(基于 statsmodels 的 t 检验)等 使用贝叶斯优化对 XGBoost 进行超参数优化并构建新的 XGB 模型并持久化以备第三问使用。
附件十八：预测模型_整合.ipynb ： 对基于问题一第一小问方案三计算所得率结果与第二问构建高准确率预测模型的方法构建模型并持久化的代码整合，以及 P,S,Si 三种元素的有关计算。
附件十九：第三问_计算目标函数需要的函数.ipynb ：第三问中制定遗传算法适应度函数所需要的部分函数，包括模型加载，通过所得率计算最终成分，## 通过投入金属计算成本，计算历史数据中每吨钢水所需投入合金元素之平均成本，初步整合后的适应度函数
附件二十：遗传算法 with deap.ipynb ： 使用 deap 类库实现的遗传算法,在程序中保存不同样本的最终计算结果。