流程工业的智能制造

摘 要

本文针对高炉炼铁过程中铁水含硅量动态预测与硫含量优化问题，采用数据挖掘技术建立了相应的数学模型，并根据预测与优化结果对高炉炼铁过程提出了指导性建议。

针对铁水含硅量[Si]动态预测问题，首先根据冶炼周期选取多组风量[FL]和喷煤量[PML]作为数据特征，然后通过**基于 *Score Matching* 算法的独立成分分析（*ICA*）方法**将低维样本空间**非线性映射**到高维特征样本空间，得到相互独立的特征向量，将之与前一时刻的[Si]含量组合构成**独立特征矩阵**；将独立特征矩阵进行归一化处理，以其 70%作为训练集，30%作为测试集传入**最小二乘法支持向量机（*LS-SVM*）模型**；通过网格遍历法与交叉检验法，求出 *RBF* **核函数**的两个最优参数：惩罚因子 *C* 和核函数半径 *g*。以 *RBF* 核函数构建的支持向量机对训练集进行训练从而对特征空间进行划分得到最优超平面，以实现对铁水[Si]含量的一步动态预测。在一步动态预测的基础上把预测结果并入特征矩阵归一化后重新训练得到二步动态预测模型。

对于上述模型进行预测成功率检验，当置信区间为±15%时，一步预测与两步预测均能保证 90%以上的正确率；当置信区间变为±8%时，一步预测的准确率为 85.5%，两步预测的准确率为 74.7%；对于炉温的升降趋势预测，一步预测准确率为 82.4%，两步预测准确率为 73.9%。最后用**非线性拟合优度 R_{new}** 对拟合程度进行检验，一步、两步预测模型的拟合优度均在 0.8 以上，模型求解正确。

针对含硫量[S]的优化问题，首先对其与前三组（一个周期）[Si]、[PML]、[PL]数据进行**相关性分析**，得出其影响因素为前一组输出的[Si]和前一时间段输入的[PML]；然后以空间欧氏距离为指标，进行**数据匹配**并通过**样条插值**的方法补全空缺数据绘制其三维关系图，定性分析[S]所受两个因素的影响情况；最后通过遍历定量求出[S]含量的最低值与使其达到最低值的[Si]、[PML]含量。分析结果符合物料守恒规律及炉温与含硫量[S]成反比的关系。

最后对以上问题的模型进行评价和推广，根据分析结论对处理相关实际问题提出了一些指导性建议，具有一定的实际参考价值。

关键词： **基于 *Score Matching* 算法的独立成分分析（*ICA*）方法　非线性映射　最小二乘法支持向量机（*LS-SVM*）模型　相关性分析　样条插值**

目录

1、问题重述

1.1 问题背景

当前我国正面临供需关系失衡造成的产能过剩问题。我国钢铁、煤炭、水泥、石油、石化、铁矿石等几大行业，利润下降幅度巨大，产能过剩严重。截至 2015 年 12 月初，几大行业的生产价格指数（PPI）已连续 40 多个月呈负增长状态，作为国民经济支柱性产业的钢铁冶金等流程工业首当其冲。

革新技术、优化管理是制造业“破局”的方法。2015 年 3 月 5 日，李克强在两会上作《政府工作报告》时提出“中国制造 2025”时指出: 中国制造业与世界先进水平相比，在资源利用效率、产业结构水平、质量效益等方面差距明显。2015 年 11 月，习近平总书记主持召开中央财经领导小组第十一次会议，强调要着力加强供给侧结构性改革，提高供给体系质量和效率，推动我国社会生产力水平实现整体跃升。而这一切都需要从技术手段上予以实现。

具体在高炉冶炼优质铁水的过程中，为了实现产品优质、耗能低等优化目标，我们需要升级智能控制的关键技术，将传统的反馈控制转变为更加及时、快捷的预测控制，通过预测性地动态调整生产过程控制，得到最优的生产效果。考虑到炼铁的一系列最终指标都与控制性中间指标——炉温密切相关，而铁水含硅量[Si]又与炉温呈良好的线性相关；因此，分析并预测铁水含硅量[Si]，以此调控当前高炉各项操作参数也许是条可行的求解方法。

1.2 问题描述

（一）以由[Si]-[S]-FL-PML 依序号排列的 1000 个高炉生产数据为样本，分析[Si]与其它数据之间的关系，建立[Si]预测动态数学模型（包括一步预测模型与两步预测模型），全面论述数学建模的思路。

（二）选取适当的验证样本，验证数学模型的预测成功率（包括数值预测成功率和炉温升降方向预测成功率），并讨论其动态预测控制的可行性。

（三）以铁水含硫量[S]为铁水质量的指标（含硫量越低，铁水质量越好），建立质量指标[S]的优化数学模型，并讨论模型计算得出的预测控制预期效果。

（四）讨论建立复杂流程工业智能控制大数据建模的心得体会。

2、模型假设与符号系统

2.1 模型假设

针对本问题，建立如下假设：

（1） 参数准确性假设：问题一中，按照附件 1 所取参数值均准确无误；

（2） 参数相关性假设：问题一中，超过 5 个炉次之前的数据对铁水含硅量[Si]没有影响。

（3） 拟合平滑性假设：问题三中，选取距离目标向量最近的特征组合所对应铁水含硫量[S]为目标向量相应的[S]。

2.2 符号系统

符号	意义
X	由样本数据每五个 *PML* 与 *FL* 一组得到的原始特征矩阵
y	由样本数据得到的铁水含硅量[Si]向量
A	*ICA* 的系数矩阵
V	*ICA* 的白化矩阵
E	协方差矩阵特征向量的正交矩阵
W	*ICA* 分离后的独立特征矩阵
D	协方差矩阵特征值的对角矩阵
S	假定非高斯分布的源信号矩阵
F	协方差矩阵相应特征向量组成的矩阵
x	观测信号向量
$\tilde{x}$	白化信号向量
ψ	得分函数
$\tilde{J}$	样本的目标函数
C	是惩罚系数
gamma	核函数半径
$\sigma_{j=n}$	第 n 行数据的标准差
K	Kernel Function，即核函数
Z	决策函数
$Rnew$	拟合度指标

·注：未列出的符号以及重复符号以出现处为准。

3、问题一的建模与求解

3.1 问题一的分析

由题目信息可知，前后两炉铁水含硅量是具有相关性的；而炉温，即铁水含硅量[Si]与喷煤、鼓风量有关。由于一个炼铁周期是 6 至 8 个小时，而出铁是 2 小时一次，所以出铁 3 至 4 次才能完成一个周期。

可以初步判断，从炼铁周期开始至结束，喷煤与鼓风量必然都对最后炉温产生影响。因此，为了避免遗漏信息，我们选择提取每一个铁水含硅量[Si]的前 5 组喷煤、鼓风量作为特征，采用 ICA 特征提取技术，得到一个特征相互独立的矩阵；最后，利用最小二乘支持向量机方法，得到核函数的最优参数，取 70%和 30%分别进行训练和预测，得到[Si]的预测非线性模型。

问题一的建模流程图如下：

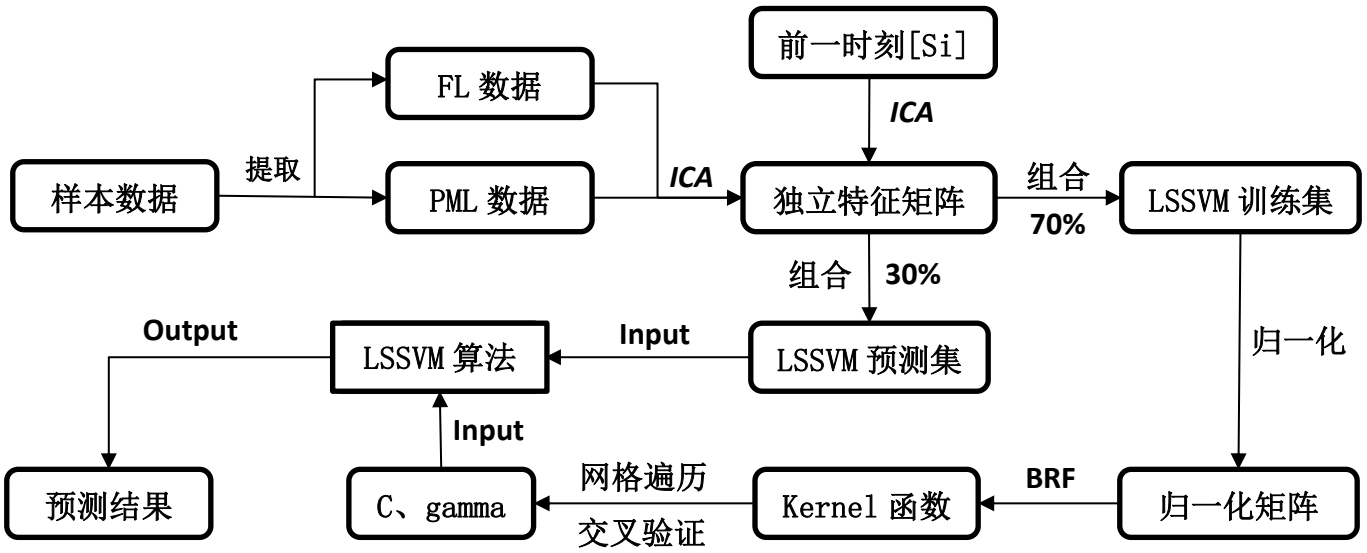

图 3.1 问题一的建模流程图

3.2 建模准备——基于独立成分分析（*ICA*）的特征独立方法

ICA 原先是一种从混合线性信号中恢复出一些基本源信号的方法。

设有 i 个观测信号$x_1, x_2, \dots, x_i$，分别为 n 个非高斯分布独立的源信号$s_1, s_2, \dots, s_n$的线性组合，即$X = A \cdot S$。其中，A 为系数矩阵，$X = [x_1 \ \ x_2 \dots x_i]^T$，$S = [s_1 \ \ s_2 \dots s_n]^T$。*ICA* 的目的，在于找出一个分离矩阵 W，使得源信号估计值 $\hat{S} = W \cdot X = W \cdot A \cdot S \approx S$，从而分离出源信号。[1]

本题中，使用 *ICA* 可以将影响铁水含硅量[Si]的特征分离开来，分散为独立的新特征，便于进行进一步归一化处理与机器学习。使用 *ICA* 方法时，需要做一些必不可少的预处理，而后运用算法求出最终的独立特征矩阵与训练集。

（一）*Centering* 中心化

中心化就是对每一个信号向量 *x* 都减去它的均值向量，使其成为一个零均值的变量。中心化的好处在于，它能够简化 *ICA* 方法；当我们用中心化的 *X* 得到系数矩阵 *A* 以后，可以直接把 *s* 均值向量的矩阵加到中心化的 *S* 中去。[2]

（二）*Whitening* 白化

白化的步骤是，首先将观测信号 *X* 进行白化处理；设白化后的信号为$\tilde{X}$，则：

$$\mathrm{E}\{\tilde{X}\cdot\tilde{X}^T\}=E\{(V\cdot X)\cdot(V\cdot X)^T\}=I \tag{1}$$

$$\mathrm{V}=D^{\frac{1}{2}}F^T \tag{2}$$

式中，*V* 代表白化矩阵，*I* 是单位矩阵，*E* 是协方差矩阵；

$D=diag(\lambda_1,\lambda_2,\dots\lambda_n)$，$\lambda$代表协方差矩阵$E\{XX^T\}$的前 n 个特征值。F 为相应特征向量组成的矩阵。

该式为特征值分解[1]方法的简化式，其具体操作方法如下：[2]

(1) 将观测信号向量 *x* 线性变换为白化信号向量$\tilde{x}$，使其协方差矩阵与单位矩阵相等。

(2) 对协方差矩阵$E\{X\cdot X^T\}$做特征值分解（***EVD***）：

$$E\{X\cdot X^T\}=\boldsymbol{EDE^T} \tag{3}$$

其中 ***E*** 为该协方差矩阵特征向量的正交矩阵，***D*** 为其特征值的对角矩阵

(3) 白化后的信号向量$\tilde{x}$由下式给出：

$$\tilde{x}=\boldsymbol{ED^{-\frac{1}{2}}E^T}x \tag{4}$$

经过白化处理后，同时也得到矩阵$\tilde{X}$；$\tilde{X}$中各信号分量相互正交，消除了原矩阵 X 中各信号分量的相关性。

(4) 系数矩阵 *A* 也变换为新矩阵$\tilde{A}$：

$$\tilde{x}=\boldsymbol{ED^{-\frac{1}{2}}E^T}As=\tilde{A}s \tag{5}$$

不难判断，$\tilde{A}$也是一个正交的矩阵：

$$E\{\tilde{x}\tilde{x}\}=\tilde{A}E\{ss^T\}\tilde{A}^T=AA^T=I \tag{6}$$

原系数矩阵 *A* 有n^2个特征，而正交的新系数矩阵$\tilde{A}$只包含$\frac{n(\mathrm{n}-1)}{2}$个特征；故而，白化过程实际上也是一个特征降维的过程。

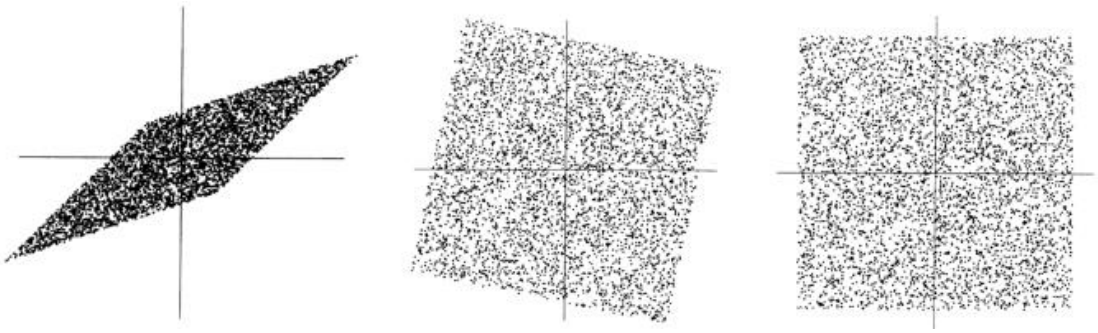

图 3.2　白化原理图：白化过程压缩特征量

（三）Score_Matching 得分匹配算法

一般而言，由数据的概率分布，我们只能得到一个趋于增加的常量。因此，计算模型可行性和最优值的方法难以实现。与其应用传统的蒙特卡罗法，倒不如避开麻烦，着重于研究密度的得分函数（the score function of density）[3]

$$\psi(\eta; W, V) = \nabla_\eta \, log \, p(\eta; W, V) \tag{7}$$

另外，定义 $$\psi_x(.) = \nabla_\eta \, log \, p_x(.) \tag{8}$$

输入数据与得分函数相匹配，就能够使模型得到最优化。在这里，我们通过把平方距离最小化，得到

$$J(W, V) = \frac{1}{2}\int_\eta ||\psi(\eta; W, V) - \psi_x(\eta)||^2 d\eta \tag{9}$$

上述(9)式可以通过强行计算密度的非参数，得出结论；但是，只要对公式求导，我们能够得到

$$\tilde{J}(W, V) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{\partial}{\partial \eta_i}\psi_i(x(t); W, V) + \frac{1}{2}{\psi_i}^2(x(t); W, V)\right] + C \tag{10}$$

这里，$\tilde{J}$表示的是该样本的目标函数，但当仅在$T \to \infty$且非简并最优值存在时有效。该函数在统计学意义上是连续的，其中常数 C 与参数的取值没有关系。

当目标函数最小时，我们得到最终的独立特征矩阵 *W* (999×400)。

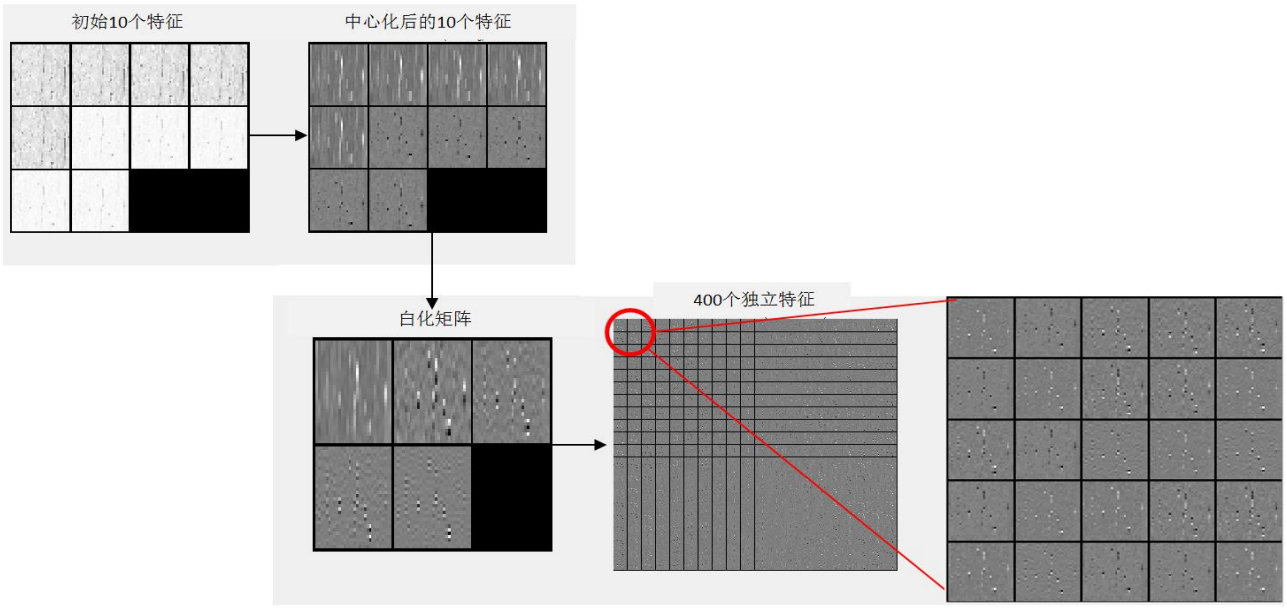

图 3.3 中心化与白化得到独立特征矩阵过程

（四）组成训练集

不妨令$W = [w_1, w_2, \ldots, w_{400}]$，*w* 代表矩阵 *W* 中的列向量；则训练样本经过 *ICA* 动态提取特征后，获得统计独立特征信号与铁水含硅量一起组成 ***LSSVM*** **训练集**$\{\boldsymbol{W}, \boldsymbol{y}\}$，其中y即为由 999 个铁水含硅量数据组成的列向量。

3.3 问题一模型建立

（一）支持向量机(Support Vector Machine，SVM)

支持向量机于 1995 年被提出，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，是一种与学习算法有关的监督学习模型，可以分析数据，识别模式，常用于分类和回归分析。[4]

对于神经网络方法，可以简单概括为：系统随机产生一个超平面并移动它，直到数据中不同类的点刚好正好位于该超平面的不同侧面。（由图 3.4 所示）这种处理机制决定了用于分类的超平面将会非常靠近训练集中的点。但对于支持向量机而言，它的目标不仅只是找到满足分类要求的超平面，而且要使训练集中的点距离超平面距离尽量得远。它的基本思想由图 3.5 所示。[1]

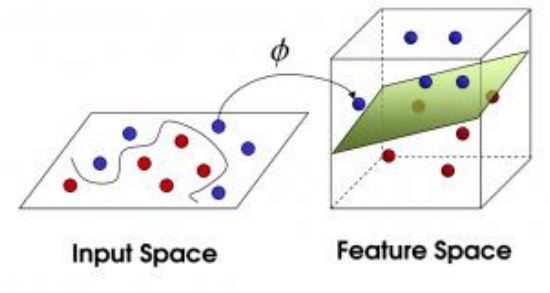

图 3.4　超平面分类原理[6]

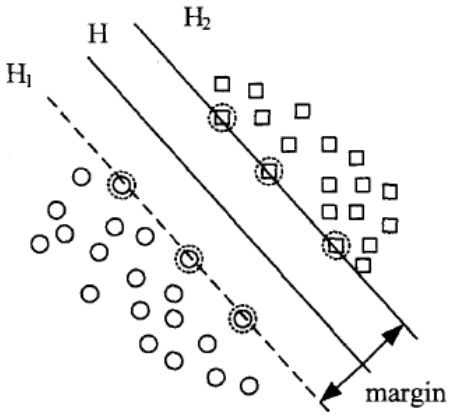

图 3.5　支持向量机分类原理[1]

（二）最小二乘支持向量机(least squares support vector machine,LSSVM)

支持向量机方法在优化求解时，计算复杂度随训练样本个数的增大而增大；且求解的相应问题越复杂，计算速度越慢。因此，对于大规模的寻优问题需要对标准的支持向量机进行改进。最小二乘支持向量机是解决此类问题的有效方案，它在优化问题中引入了平方项，改进优化目标的损失函数为最小二乘线性系统，将不等式约束条件变为等式约束条件，将经典的二次规划寻优问题转化为求解方程组，极大降低了 *SVM* 的运算复杂度。[5]

3.3.1 特征归一化

将由 *ICA* 得到的训练集矩阵 *SW* (999×401)进行**归一化处理：**

$$\begin{bmatrix} x_{[1,1]} & \cdots & x_{[1,401]} \\ \vdots & \ddots & \vdots \\ x_{[999,1]} & \cdots & x_{[999,401]} \end{bmatrix}$$

每列数据减去列均值后除以标准差

$$\begin{bmatrix} \frac{x_{[1,1]}-\overline{x}_1}{\sigma_{j=1}} & \cdots & \frac{x_{[1,j]}-\overline{x}_j}{\sigma_{j=j}} & \cdots & \frac{x_{[1,401]}-\overline{x}_{401}}{\sigma_{j=401}} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \frac{x_{[i,1]}-\overline{x}_1}{\sigma_{j=1}} & \cdots & \frac{x_{[i,j]}-\overline{x}_j}{\sigma_{j=j}} & \cdots & \frac{x_{[i,401]}-\overline{x}_{401}}{\sigma_{j=401}} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \frac{x_{[999,1]}-\overline{x}_1}{\sigma_{j=1}} & \cdots & \frac{x_{[999,j]}-\overline{x}_j}{\sigma_{j=j}} & \cdots & \frac{x_{[999,401]}-\overline{x}_{401}}{\sigma_{j=401}} \end{bmatrix} \quad (17)$$

其中，i 代表行数，j 代表列数， $\sigma_{j=n}$代表第 n 行数据的标准差，$\overline{x}_j = \frac{\sum_{i=1}^{999} x_{[i,j]}}{999}$

3.3.2 *RBF Kernel* 与最优参数

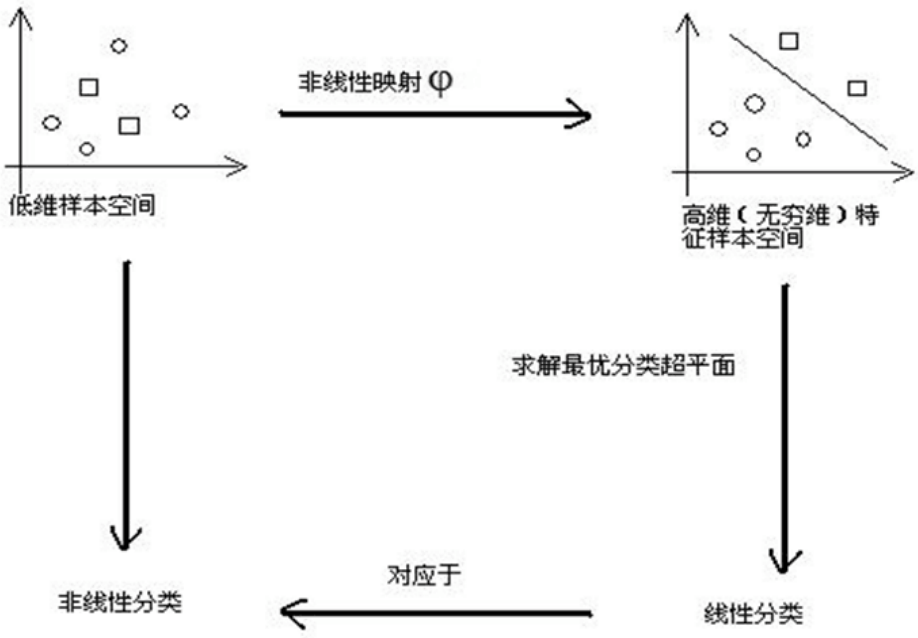

图 3.6 线性分割求解最优分类超平面示意图

对特征空间划分的最优超平面是 SVM 的目标。关于超平面理解可见图四，在低维样本空间里，B 不是线性可分的，即无法被一条直线分为两类。但从二维空间“映射”至更高的三维空间后，B 能够被平面（二维中即为线性）分割开来。

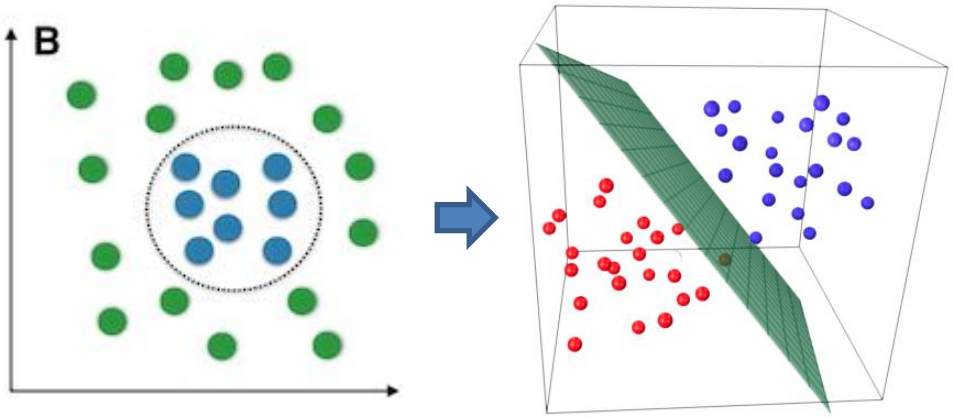

图 3.7 升维后非线性分割变为超平面分割

问题的关键在于，如何构造映射函数，使得非线性关系化为线性关系。

（一）*RBF Kernel* 基于半径的核函数

我们将映射所使用的函数称为内核函数（*Kernel Function*）。

内核函数形式多样，有线性、多项式、指数、三角函数等，在这里我们使用 *RBF Kernel* 求解最优参数。*RBF*（*Radial Based Function*）有两个重要参数：C 和 gamma。其中，C 是惩罚系数，即对误差的宽容度。C 的值小，则决策平面光滑，欠拟合；C 的值大，则训练样本分类较为准确。Gamma 是 RBF 函数自带的参数（核函数半径），它的值决定单个训练样本的影响力，即被映射至特征空间后的分布。gamma 值越大，支持向量越少，上图的点就离切面越远；gamma 值小，支持向量越多，上图的点就离切面越近 [7] 。

RBF Kernel 的公式如下：

$$K(x, x') = \exp(-g\|x - x'\|^2) \tag{18}$$

（二）交叉验证法与网格遍历法

给定 C 和 gamma 的范围后，应用网格遍历法，即尝试各种可能的(C, g)对值，然后进行交叉验证，找出使交叉验证精确度最高的(C, g)对。

交叉验证(Cross Validation)法是一种用于验证分类器的性能的统计方法。它的基本思想是，在某种意义下将原始数据进行分组，一部分做为训练集，另一部分做为验证集。首先，用训练集对分类器进行训练，再利用验证集来测试训练得到的模型，以此来做为评价分类器的性能指标。

我们采用 K-fold Cross Validation(K-CV)作为交叉验证算法，将原始数据均分成 K 组，将每个子集数据分别做一次验证集，其余的 K-1 组子集数据作为训练集，这样就得到 K 个模型。而后，将 K 个模型最终的验证集的分类准确率的平均数作为此 K-CV 下分类器的性能指标。

3.3.3 基于最小二乘向量机方法的问题转化

对于分类问题，由于已知铁水含硅量[Si]与喷煤、鼓风量为高度耦合非线性关系；假定训练样本集为训练样本集数据$(x_i, y_i), i = 1,2, \dots, l, x_i \in R^d, y \in R$，*LSSVM* 将优化问题转变为：

$$\begin{cases} \min\limits_{w,b} \dfrac{1}{2}\|w\|^2 + \dfrac{1}{2}C\sum\limits_{i=1}^{l}{\xi_i}^2 \\ s.t. \quad y_i[\Phi(x_i)\cdot w + b] = 1 - \xi_i \quad i = 1,2,\dots,l \end{cases} \tag{11}$$

用拉格朗日法求解，*LSSVM* 将问题转化为下列二次规划问题：

$$min\ J_{LSSVM} = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{l}{\xi_i}^2 - \sum_{i=1}^{l}a_i\{y_i[\Phi(x_i)\cdot w + b] + \xi_i - 1\} \tag{12}$$

其中a_i为x_i对应的拉格朗日乘子。

根据极值条件：

$$\frac{\partial J}{\partial w} = 0,\ \frac{\partial J}{\partial b} = 0,\ \frac{\partial J}{\partial \xi} = 0,\ \frac{\partial J}{\partial a} = 0 \tag{13}$$

可得方程组：

$$\begin{cases} w = \sum_{i=1}^{l} a_i y_i \Phi(x_i) \\ \sum_{i=1}^{l} a_i y_i = 0 \\ a_i = C\xi_i \\ y_i[\Phi(x_i)\cdot w + b] + \xi_i - 1 = 0 \end{cases} \tag{14}$$

定义核函数$K(x_i, x_j) = \Phi(x_i)\cdot\Phi(x_j)$，进一步将问题变为求线性方程组：

$$\begin{pmatrix} 0 & y_1 & \cdots & y_l \\ y_1 & K(x_1,x_1)+\dfrac{1}{\gamma} & \cdots & K(x_1,x_l) \\ \vdots & \vdots & \ddots & \vdots \\ y_l & K(x_1,x_l) & \cdots & K(x_1,x_l)+\dfrac{1}{\gamma} \end{pmatrix} \begin{pmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_l \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{15}$$

方程组求解得到决策函数：

$$Z = sign\{\sum_{i=1}^{l} a_i y_i k(x_i, x) + b\} \tag{16}$$

这里，Z 即代表用 *LSSVM* 方法最终求得的预测值。（其直接预测一步，两步预测需要在一步预测基础上进行，所用模型一致）

3.4 问题一模型求解

3.4.1 求解过程

利用 Matlab 软件实现对全部基于 *ICA* 与 *LSSVM* 的模型进行编程求解，按照流程图思路，依次将输出的数据输入下一环节。

一、*ICA*—*SVM* 方法的一步预测：
（1）回归的均方误差为 *0.00884679*
（2）回归平方相关系数为 *0.285954*
（3）最优参数值 *gamma = 0.5743，C = 1*

二、*ICA*—*SVM* 方法的两步预测：
（1）回归的均方误差为 *0.0117852*
（2）回归平方相关系数为 *0.0870566 (regression)*
（3）最优交叉检验 *MSE* 值为 *0.00975863*
（4）最优参数值 *C= 1，gamma = 0.25*

利用 Matlab 作图，预检验最优参数的取值范围:

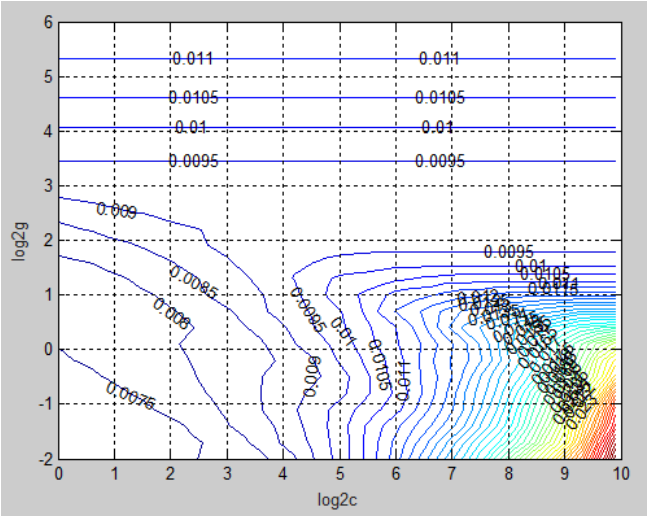

图 3.8 一步模型 利用梯度下降寻找最优参数（C，gamma）

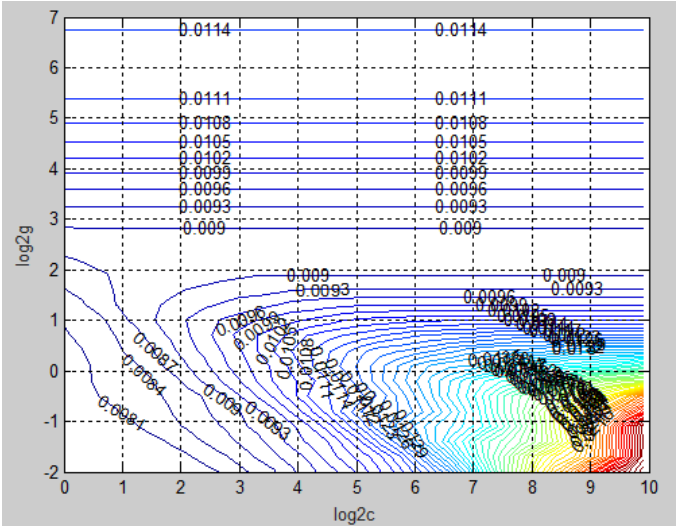

图 3.9 两步模型 利用梯度下降寻找最优参数（C，gamma）

另外，由于 *ICA* 过程基本上都是矩阵变换，具体过程的数据难以给出，最后将以 Matlab 代码的形式展示。

3.4.2 求解结果

（1） 一步模型输出的预测结果

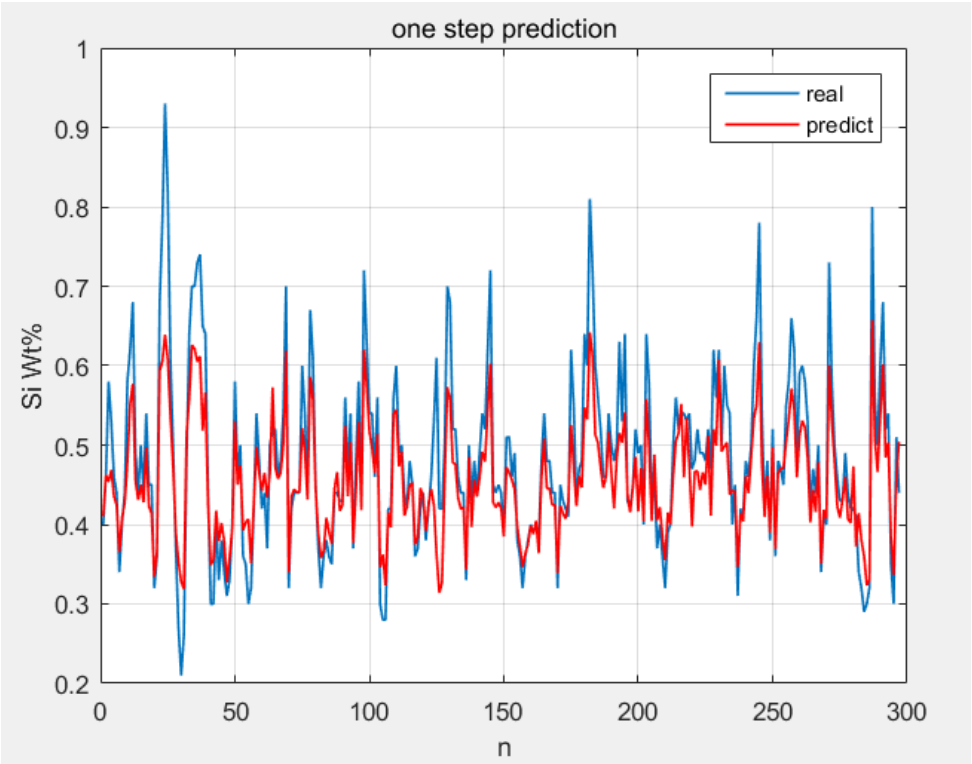

图 3.10 一步模型输出的结果

（2） 两步模型输出的预测结果

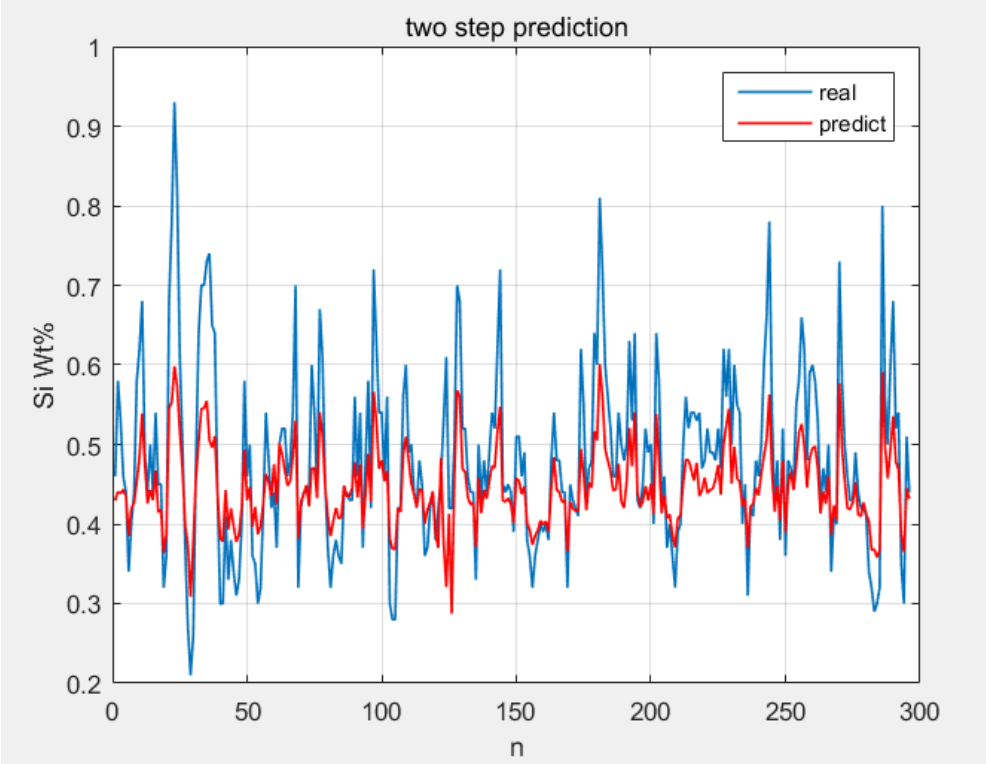

图 3.11 两步模型输出的结果

4、问题二的分析与求解

4.1 问题二的分析

根据问题一的模型，将用于预测的 300 组数据（第 700-第 999 组）作为测试样本进行模型的验证检验；因此直接从问题一中调用数据即可。

考虑到该模型是非线性模型，用于线性模型检验的指标不适用，故采用百分率测试的方法，按误差置信的百分率梯度从 5%至 25%做多组成功率检测。对于炉温升降方向的检测，则用算法分别判断观测值与拟合值在前后炉次中的差值；若都差值都为正或负，认为预测正确；若差值一正一负，认为预测失败。

对于可行性的分析，可以从数值预测成功率和炉温升降方向预测成功率两个角度进行讨论；此外，可以引用非线性拟合优度 *Rnew* 值，判断预测模型的拟合程度，作为讨论可行性的第三个角度。

其中$Rnew = 1 - (\frac{Q}{\sum y^2})^{\frac{1}{2}}$，Q 为$y$的残差平方和。

4.2 问题二的求解

(1) 数值预测成功率
由 MATLAB 编程得到的结果如图：

```
Rnew =

    0.8825

successful rate of prediction within the range of error 25%: 9.966330e-01
successful rate of prediction within the range of error 20%: 9.898990e-01
successful rate of prediction within the range of error 15%: 9.797980e-01
successful rate of prediction within the range of error 10%: 9.124579e-01
successful rate of prediction within the range of error 8%: 8.552189e-01
successful rate of prediction within the range of error 5%: 7.037037e-01
successful rate of prediction for trendency of Si Content: 8.243243e-01
```

图 4.1 第一步数值预测成功率

```
Rnew =

    0.8434

successful rate of prediction within the range of error 25%: 9.932432e-01
successful rate of prediction within the range of error 20%: 9.729730e-01
successful rate of prediction within the range of error 15%: 9.358108e-01
successful rate of prediction within the range of error 10%: 8.310811e-01
successful rate of prediction within the range of error 8%: 7.466216e-01
successful rate of prediction within the range of error 5%: 5.641892e-01
successful rate of prediction for trendency of Si Content: 7.389831e-01
```

图 4.2 第二步数值预测成功率

结果整理为表格如下：

	25%	20%	15%	10%	8%	5%
第一步	99.663%	98.990%	97.980%	91.246%	85.522%	70.370%
第二步	99.324%	97.297%	93.581%	83.108%	74.662%	56.419%

表 4.1 数值预测成功率表

注：图中第二行第二列数据“99.663%”表明，在第一步数值预测中，误差范围在25%以内的预测成功率为 99.663%。其他数据依此类推。

(2) 炉温升降方向预测成功率

```
successful rate of prediction for trendency of Si Content: 8.243243e-01
```

图 4.3 第一步炉温升降方向预测成功率

```
successful rate of prediction for trendency of Si Content: 7.389831e-01
```

图 4.4 第二步炉温升降方向预测成功率

	炉温升降方向预测成功率
第一步	82.432%
第二步	73.898%

表 4.2 炉温预测方向成功率表

(3) 小结

由以上分析求解可知，数值预测和炉温升降方向预测成功率结果如下表：

	25%	20%	15%	10%	8%	5%	炉温升降方向预测
第一步	99.663%	98.990%	97.980%	91.246%	85.522%	70.370%	82.432%
第二步	99.324%	97.297%	93.581%	83.108%	74.662%	56.419%	73.898%

表 4.3 预测成功率表

4.3 动态预测控制

根据已经得到的第一步、第二步数值预测与炉温升降方向预测可知，在进行实时动态预测时，有 97.980%的几率可以让下一步的预测仅有 15%的误差，有 93.581%的几率可以让之后第二步的预测仅有 15%的误差；在炉温方向预测中，第一步成功率有 82.432%，第二步成功率为 73.898%，有相当的参考价值。

除此之外，我们引入专用于非线性回归分析的拟合优度 *Rnew*。

同线性回归分析中的拟合优度R^2类似，*Rnew* 可以用作判断回归拟合对实际观测值的拟合程度，取值范围为[0,1]，越接近 1 则拟合程度越高。该预测模型的 Rnew 值大于 0.8，可以认为拟合程度是较高的。

综上所述，从本模型的预测成功率看，其在实际应用中具有一定可行性。

5、问题三的建模与求解

5.1 问题三的分析

问题三要求建立质量指标[S]的优化模型，并讨论按照优化结果进行[Si]预测的预期效果。对该问题，首先要分析[S]的相关影响因素，然后定性分析这些因素对其含量的影响以及影响程度的变化情况，定量计算使得[S]含量达到最优值时的相关因素取值。

5.2 问题三模型建立

首先对当前的[S]于前三个时间段输出的[Si]、[PML]、[PL]进行相关性分析，得出[S]只与前一时间段输出的[Si]和前一时间段输出的[PML]有关，而与其他数据特征没有相关性。之后，对这两个因素和因变量[S]分析，以空间欧氏距离为指标，通过样条插值的方法补全空缺数据绘制其三维关系图，定性分析[S]所受两个因素的影响情况；最后定量计算[S]含量的最低值与使其达到最低值的[Si]、[PML]含量。

问题三流程图如下所示：

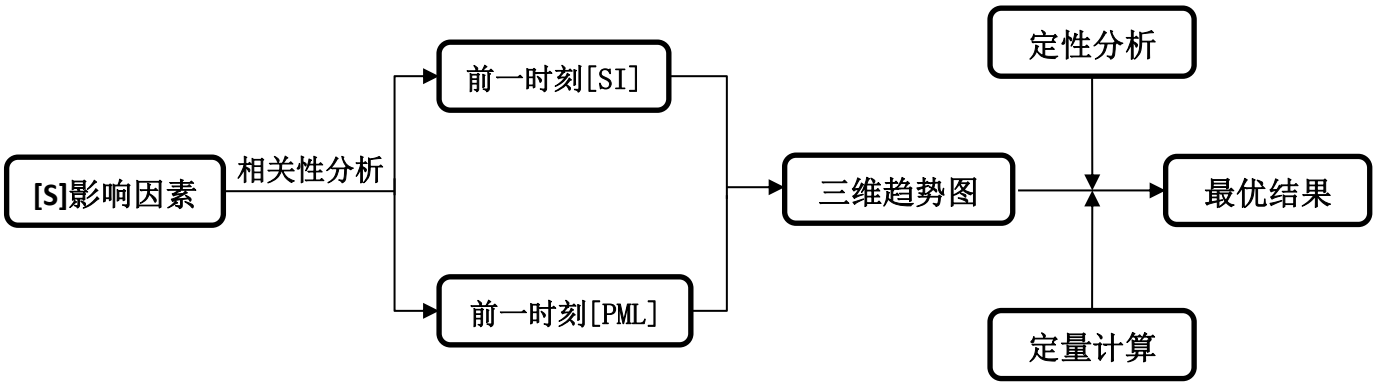

图 5.1 问题三建模流程图

5.2.1 相关性分析

将[S]含量与之前三次输出所对应的[Si]含量、之前三次的喷煤量[PML]、之前三次的风量[FL]做相关性分析，结果如下图所示：

相关性

		S	Si_pre1
S	Pearson 相关性	1	-.159**
	显著性（双侧）		.000
	N	997	997
Si_pre1	Pearson 相关性	-.159**	1
	显著性（双侧）	.000	
	N	997	1000

**. 在 .01 水平（双侧）上显著相关。

相关性

控制变量			S	Si_pre2
Si_pre1	S	相关性	1.000	-.011
		显著性（双侧）	.	.730
		df	0	994
	Si_pre2	相关性	-.011	1.000
		显著性（双侧）	.730	.
		df	994	0

相关性

控制变量			S	Si_pre3
Si_pre1 & Si_pre2	S	相关性	1.000	.011
		显著性（双侧）	.	.731
		df	0	993
	Si_pre3	相关性	.011	1.000
		显著性（双侧）	.731	.
		df	993	0

相关性

		S	PML_pre1
S	Pearson 相关性	1	-.067*
	显著性（双侧）		.035
	N	997	997
PML_pre1	Pearson 相关性	-.067*	1
	显著性（双侧）	.035	
	N	997	1000

*. 在 0.05 水平（双侧）上显著相关。

相关性

		S	PML_pre2
S	Pearson 相关性	1	-.044
	显著性（双侧）		.163
	N	997	997
PML_pre2	Pearson 相关性	-.044	1
	显著性（双侧）	.163	
	N	997	997

相关性

		S	PML_pre3
S	Pearson 相关性	1	-.046
	显著性（双侧）		.148
	N	997	997
PML_pre3	Pearson 相关性	-.046	1
	显著性（双侧）	.148	
	N	997	997

相关性

		S	FL_pre1
S	Pearson 相关性	1	-.026
	显著性（双侧）		.412
	N	997	997
FL_pre1	Pearson 相关性	-.026	1
	显著性（双侧）	.412	
	N	997	997

相关性

		S	FL_pre2
S	Pearson 相关性	1	-.009
	显著性（双侧）		.776
	N	997	997
FL_pre2	Pearson 相关性	-.009	1
	显著性（双侧）	.776	
	N	997	997

相关性

		S	FL_pre3
S	Pearson 相关性	1	-.001
	显著性（双侧）		.968
	N	997	997
FL_pre3	Pearson 相关性	-.001	1
	显著性（双侧）	.968	
	N	997	997

图 5.2 相关性分析

可以看出，[S]含量与前一时刻的[Si]含量和前一时刻的喷煤量[PML]的 *Pearson* 相关性都达到了±0.05 以上；同时，显著性都在 0.05 以下，可以认为是具有显著相关性的。而其它时刻的[Si]含量、喷煤量[PML]以及风量[FL]与[S]的相关性分析的指标没有达到要求，可以认为没有相关性。所以，问题三转化为以前一时刻的[Si]含量和前一时刻的喷煤量[PML]作为自变量，以当前时刻的[S]含量作为因变量的优化问题。

5.2.2 数据匹配与插值拟合

首先，我们将[Si]含量与[PML]含量扩展成矩阵 $W_{111\times34}$ 。其中，矩阵的行向量代表[Si]含量的离散变化取值，从最小值 0.2 以增量 0.01 增至最大值 1.3；列向量代表[PML]含量的离散变化取值，从最小值 0.1 以增量 0.5 增至最大值 17。

然后，以矩阵中的坐标对应的[Si]含量和[PML]含量为目标向量 x，分别求其与原数据 y 之间的欧氏距离：

$$d(x, y) = \left[\sum_{k=1}^{p} \left|x_k - y_k\right|^2\right]^{\frac{1}{2}}$$

找到与之距离最小的原数据，以该数据所对应的[S]值作为目标向量的[S]值，这里可能不止一个与之对应的原数据，对于超过一个的情况，我们选取这些原数据的平均值作为目标向量的[S]值。

得到全部的含硫量[S]以后，对其进一步扩充，将喷煤量[PML]之间的间隔设置为 0.01，采用样条插值的方法进行更平滑的拟合。其中三次样条插值的定义可表述为：

如果函数 $f(x)$ 在节点 $x_0, x_1, \ldots, x_n$ 处的函数值为 $f(x_j) = y_j, j = 0, 1, \ldots, n$

并且关于这个节点集的三次样条插值函数 $s(x)$ 满足插值条件

$$S(x_j) = y_j, j = 0, 1, \ldots, n$$

则称这个三次样条函数 $s(x)$ 为三次样条插值函数。

5.2.3 最优点寻找

采用遍历的思想进行最优点搜寻，可找到取到最小硫含量[S]的硅含量[Si]与喷煤量[PML]的组合。

5.3 问题三的模型求解

5.3.1 插值拟合

采用基于欧式距离的数据匹配方法以及三次样条插值的数值算法，利用 MATLAB 编程可以得到结果如下图所示：

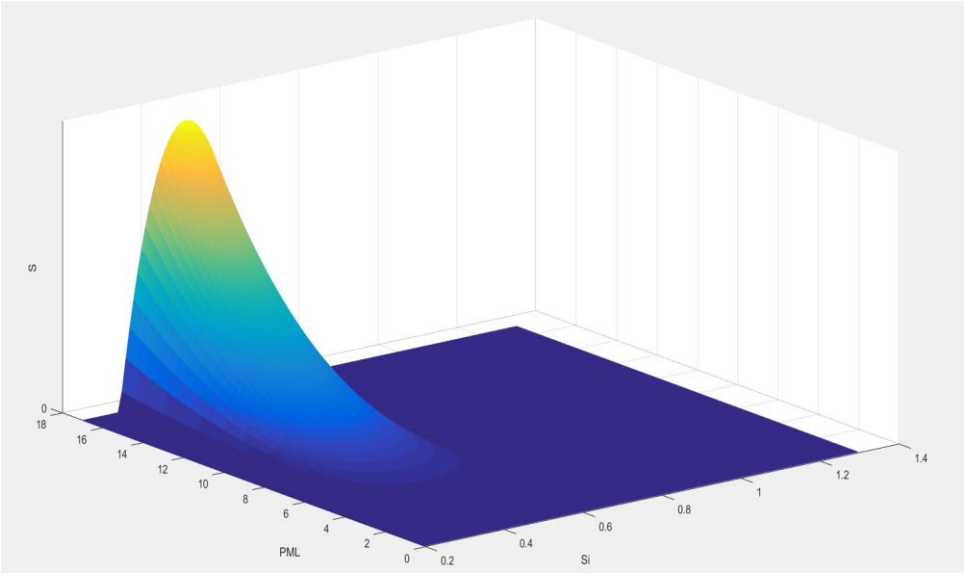

图 5.3 硫含量[S]的三维趋势图

从图中可以看出，在喷煤量比较小的时候硫含量[S]几乎接近 0，随着喷煤量[PML]的增加而逐步增加，这一点符合物料守恒关系[8]。

固定喷煤量，可以发现含硫量[S]随着硅含量[Si]的增加先增加后减小，由题目可知，硅含量[Si]代表着炉温，即当炉温比较低时，含硫量[S]比较高，含硅量[Si]比较小；炉温比较高时，含硫量[S]比较少，含硅量[Si]比较大。这一点符合高炉炼铁的化学反应原理，从而可以知道我们的模型求解正确。

5.3.2 遍历寻求最优解

对目标向量构成的矩阵进行遍历搜索，可得到含硫量[S]的最小值

$$S_{\min} = 0.007$$

此最小值对应的含硅量[Si]与喷煤量[PML]的组合如下表所示：

Si	PML	Si	PML	Si	PML
1.25	5.1	0.73	6.1	0.21	7.1
0.48	5.6	1.07	6.1	0.55	7.1
0.82	5.6	0.3	6.6	0.89	7.1
1.16	5.6	0.64	6.6	1.23	7.1
0.39	6.1	0.98	6.6	0.46	7.6

表 5.1 最优硫含量[S]对应组合

6、模型的优缺点和推广

6.1 问题一模型的优缺点

优点：

(1) 利用最小二乘向量机方法对数值较为准确地做出了预测。当置信区间为±15%时，一步预测与两步预测均能保证 90%以上的正确率；当置信区间变为±8%时，一步预测的准确率为 85.5%，两步预测的准确率为 74.7%。
(2) 能够提供多种置信区间的数据，使得其更好的贴近于不同生产实际的需要。

有待提高之处：

(1) 在个别[Si]的极值处，模型无法进行准确的预测。
(2) 对于炉温的升降趋势预测相较于数值预测不够准确，一步预测准确率为 82.4%，两步预测准确率仅为 73.9%。

改进方向：

(1) 修改参数，使模型在预测炉温的时更加准确。

推广方向：

(1) 其他工业大数据、医疗大数据等方向，比如：某些药品的制备等。

6.2 问题三模型的优缺点

优点：

(1) 利用相关性分析得出[S]与风量无关的结论，方便了模型的求解计算。

有待提高之处：

(1) 拟合平滑性假设（问题三中，选取距离目标向量最近的特征组合所对应铁水含硫量[S]为目标向量相应的[S]）可能带来较大误差，有待消除。

改进方向：

(1) 消除拟合平滑性假设。

推广方向：

(1) 环境治理中水质的改善与优化、城市交通中的拥堵情状况的优化等。

7、复杂流程工业智能控制大数据建模的心得体会

首先，我们要感谢这次赛题让我们有机会利用数据挖掘算法来建模预测复杂流程工业智能控制系统。三天的快速学习不仅让我们获得了关于数据挖掘以及高炉炼铁控制方面的知识，更为我们迈向“大数据世界”的大门做好了铺垫。

做问题一的时候，我们刚开始广泛尝试不同的算法，譬如线性拟合、时间序列、神经网络等，搜索了许多高炉炼铁[Si]含量预测的论文。现在回过头看，其实问题一的数学建模目的就是实现铁水含硅量[Si]的动态预测，以达到对炉温的实时动态控制。动态预测[Si]的前提是找到隐含的[Si]时间序列变化规律，否则就没有可以用于预测[Si]的依据。按照问题一的描述，铁水含硅量[Si]随时间（即炉次）的变化会受前一炉[Si]的影响，同时也可能受到其它几组参数的影响。

为了准确预测[Si]，我们的模型中势必要将其它几组参数考虑进来，这时就涉及到如何选取参数、选取多少参数数据的问题。然而，[Si]与 PML、FL 都不呈线性关系，这一点不仅可以从它们之间的相关性预分析中得知，许多以高炉炼铁预测[Si]为主题的文献也都已明确指出。所以，我们无法直接用 Pearson 相关性、时间序列等线性方法分析出对[Si]有影响的参数。

因此，我们需要找到一种能够抓取尽量多的信息，而后自动对参数进行降维的方法，将对[Si]有显著影响的参数化为某种特征值与特征向量，最后导入预测模型。这样的操作能够避开手动地分离参数，省去不必要的麻烦。

独立成分分析（ICA）恰巧具备这样的能力，它能将由独立分离的源特征线性组合出的特征重新剥离开来，将低维样本空间非线性映射到高维特征样本空间，构成相互独立的特征向量，以便于进行进一步分析潜在规律。

我们也知道，机器学习方法能够从大量数据中总结出潜在规律，尤其适合于问题一中这种要研究的关系并不明晰的情形；最小二乘向量机方法就是一种机器学习方法，它的核心是将一部分数据用于训练，另一部分用于预测，然后给出结论。它的另一个特点在于将求解二次规划问题转化为更容易求解的解方程组问题，简化运算，因此我们才决定采用 LSSVM 的方法。

此外，我们还引入了诸如 Rnew 拟合优度指标来对模型进行检验与进一步的分析；虽然因为提出至今时间不长，它是否适用还值得进一步考量，甚至大

部分做非线性回归分析的文献都没有采用 Rnew 作为评价指标。但是，论文进行到这个阶段，也基本将要完成，我们实际上也已经对各种数据挖掘算法适用场景、预测精度、计算成本有了更深入的了解，所以就决定尝试些新方法、新思路。当然，我们也深刻认识到自身知识的局限，必须要查阅文献、多做预处理，花费大量精力去钻研学习，才能理解不同算法之间的差别与思路。

大数据是一个随着机器计算能力的提高而逐步发展起来的一门科学，它有着广泛的应用场景，将之与传统科学类比为数值解与解析解毫不为过，红极一时的 AlphaGo 的诞生就与大数据紧密相连。在人工智能的发展越发蓬勃的今天，放眼大数据的未来，我们更应该脚踏实地地学好真本事。

8、参考文献

[1] 郑俊华.基于支持向量机的高炉炉温预报的研究.浙江大学硕士学位论文.NO.2006AA04Z184.17→21.2007

[2] Michele Scarpiniti. Different ICA Approaches, Neural Networks. http://ispac.ing.uniroma1.it/scarpiniti/index.htm. 2017.5.29

[3] Urs Koster and Aapo Hyvarinen,A two-layer like model estimated by Score Matching. University of Helsinki and Helsinki Institute for Information Technology.

[4] https://baike.so.com/doc/6186041-6399292.html. 2017.5.30

[5] 郑俊华.基于支持向量机的高炉炉温预报的研究.浙江大学硕士学位论文.NO.2006AA04Z184.27→28.2007

[6] Iddo.Support Vector Machines explained well. http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well/. 2017.5.30

[7] http://scikit-learn.org/stable/modules/svm.html#kernel-functions

[8] https://www.zybang.com/question/c4ff7dd5e02d34a8813779449eb0ff13.html

[9] 宁长龙，《降低铁水硅含量》，本钢技术，2013 年第 5 期：P7-9。

[10] 李静，《基于数据挖掘的高炉铁水温度建模与预报》，内蒙古科技大学，硕士学位论文，2013 年 6 月。

[11] 吴金花，《高炉冶炼过程分析及其铁水硅含量预测模型研究》，燕山大学，硕士学位论文，2016 年 5 月。

[12] 徐雅娜，许桂清，周建常，《神经网络在高炉铁水含硫量预报中的应用》，基础自动化，2001 年 12 月，第 8 卷第 6 期。

[13] 张军红，谢安国，沈峰满，《基于神经网络对铁水含硫量的优化和分析》，材料与冶金学报，2006 年 6 月，第 5 卷第 2 期

9、附录

源程序见上传附件。目录如下：

ICA	---	独立成分分析算法
One_Step_Prediction	---	一步预测模型
Two_Step_Prediction	---	两步预测模型
optimize_S	---	[S]含量优化模型