第 14 卷 第 35 期 2014 年 12 月
1671—1815(2014)35-0084-06

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 14 No. 35 Dec. 2014


# 基于 Lasso 方法的碳钢土壤腐蚀率预报研究

鲁 庆 穆志纯*
(北京科技大学 自动化学院,北京 100083)

**摘 要** 提出了一种基于 Lasso 的 step adaptive Lasso with Pls (SALP) 方法,采用 Bayesian Bootstrap 算法重构样本,通过多模型集成对变量进行预选,以消除数据扰动和离群点对于模型性能的影响。应用偏最小二乘权重系数改善 Lasso 方法在处理小样本数据时的参数估计不准确问题。采用碳钢在土壤中的腐蚀数据为研究对象,建立了腐蚀率模型。实验证明: SALP 方法建立的模型可以准确地拟合和预测土壤中的碳钢腐蚀率变化。该方法适用于自然环境中材料腐蚀率的预测及类似研究领域。

**关键词** Lasso 腐蚀率 碳钢 自然环境腐蚀 土壤
**中图法分类号** TP391.9; **文献标志码** A

材料在自然环境中的腐蚀率预报是腐蚀研究的重点和难点之一。由于自然环境的材料腐蚀实验周期漫长、实验站点分布广,而且自然环境的腐蚀影响因素复杂多变,实验数据往往存在高维度、小样本和数据质量较低等问题。目前腐蚀率预报常用的方法包括多元回归[1]、神经网络[2]、灰色系统[3]等。这些方法虽然各具优势,也存在一些问题,例如多元回归方法需要较大的样本容量,很难处理小样本数据;神经网络方法可以较好的拟合实验数据,但在解释腐蚀影响因素方面却存在局限性;灰色系统处理高维数据时需要复杂的数据预处理过程等。

Lasso 方法最早由 Tishirani[4] 提出,能够在高维变量空间中获取稀疏线性模型,目前广泛应用于图像识别[5]、基因分析[6]等领域。它将变量选择和参数估计融合在模型训练过程中,通过 L1 罚函数压缩模型系数,系数被压缩为 0 的变量被视为无关变量,只有显著变量被保留在最终模型中,即在完成模型系数估计的同时也实现了变量选择。

本文将 Lasso 方法引入材料在自然环境中的腐蚀率预报研究中。提出了一种基于 Lasso 的 SALP 方法,以实际项目中的碳钢土壤腐蚀数据检验 SALP 方法的性能,并与采用 Adaptive Lasso、神经网络以及支撑向量回归等方法训练的模型进行比较,实验结果表明,SALP 方法能够较好地处理高维、小样本数据建模问题,性能优于其他方法。建立的模型可以准确的拟合和预测碳钢在土壤的腐蚀率,并具有较好的可解释性。

2014 年 7 月 30 日收到 国家科技基础性工作专项基金项目(2012FY113000)资助

第一作者简介: 鲁 庆(1975—),男,博士研究生。研究方向: 数据挖掘、系统建模。E-mail: luqing_qhd@hotmail.com。

* 通信作者简介: 穆志纯,教授,博士生导师。研究方向: 复杂系统的建模与智能控制、信息自动化。E-mail: mu@ies.ustb.edu.cn。

## 1 Lasso 方法

Lasso 方法是一种基于 L1 罚的正则化方法,以下简述其原理。

对于任意样本容量为 $n$ 的数据集 $\{X,Y\}_1^n$,其中 $X$ 为 $p$ 维预报变量集,$X=(x_{1j},\cdots,x_{nj})^{\mathrm{T}}$; $j=1,\cdots,p$; $Y$ 为响应变量,$Y=(y_1,\cdots,y_n)^{\mathrm{T}}$; 假定线性模型可以合理的描述变量 $X$ 与 $Y$ 的关系,模型的形式如式(1)所示。

$$Y=\sum_{j=1}^{p}x_j\beta_j+\varepsilon \tag{1}$$

式(1)中 $\beta_j$ 为模型中待估计的回归系数,$\varepsilon$ 是随机误差项,代表了不能用变量 $X$ 解释的因素对于 $Y$ 的影响。

对于回归系数 $\beta$ 的估计,常用的方法是最小二乘方法: 令 $\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$,这时模型残差的平方和 $\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2$ 为最小值。

但这种方法要求样本容量 $n$ 和变量集维数 $p$ 至少要满足 $n\geqslant p+1$ 的关系,对于高维数据和小样本数据,最小二乘方法并不适用。

Lasso 方法的基本原理是对模型回归系数的绝对值之和 $\sum_{j=1}^{p}|\beta_j|$ 进行惩罚,系数 $\beta$ 的估值如式(2)所示。

$$\beta_{(\mathrm{Lasso})}=\underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2\}\times\sum_{j=1}^{p}|\beta_j|\leqslant t \tag{2}$$

式(2)中 $t$ 是调控参数。

使用 Lasso 方法估计的模型回归系数 $\beta_j$ 要满足二个条件,一是使残差平方和 $\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$ 最小,二是受 $\sum_{j=1}^{p}|\beta_j|\leqslant t$ 的约束。

通过调整式(2)中的参数 $t$,Lasso 方法使回归系数 $\beta_j$ 收缩。如果 $t$ 值充分小,部分回归系数将收缩为0,这样对应的变量 $x_j$ 就不会在最终模型中出现。Lasso 方法在估计模型回归系数的同时,也完成了变量选择的目标。在实际使用中,参数 $t$ 一般以极小化期望预测误差为标准,通过数据驱动自适应选取。

研究表明[7],与最佳子集等变量选择方法对比,Lasso 方法通过收缩模型的回归系数实现变量选择和系数估值,过程具有相对连续性,虽然所建立模型的估值是有偏的,但通过偏倚的较小增加,降低了预测值的方差,可以提高模型的预测精度。

Lasso 问题的求解算法包括二次规划算法[4]、同伦算法[8]和最小角度回归(least angle regression, Lars)算法[9]等。其中 Lars 算法由 Efron 提出,可以快速有效的求解 Lasso 问题。

Lars 算法的基本思想如图 1[9]所示: 首先在预报变量集 $X$ 中寻找与响应变量 $Y$ 相关度最大的变量 $x_1$; 然后沿 $x_1$ 方向以最大的步幅逼近 $Y$,直到出现另一个变量 $x_2$,它与变量 $x_1$ 同当前残差 $e_k$ 的相关度相等并且大于其他变量与残差的相关度; 在得到第二个变量后,Lars 算法以 $x_1$ 与 $x_2$ 的角分线方向逼近 $Y$,直到出现第三个变量 $x_3$; Lars 将沿着三个变量共同的角分线方向逼近 $Y$。如此继续,直到取得了所有自变量。

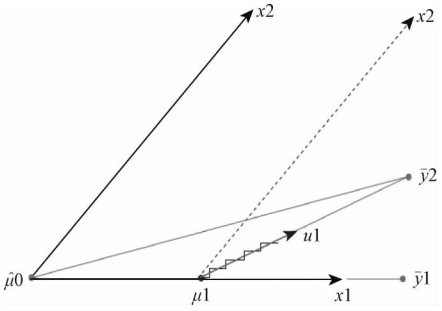


图 1　Lars 算法的解路径示意图

Fig. 1　The Schematic diagram of Lars algorithm solution path

Efron 证明[9],通过改进,Lars 算法可以获取所有的 Lasso 解。

## 2　Step Adaptive Lasso with Pls (SALP) 方法

Khan[10]证明 Lars 算法对于数据的扰动和离群点比较敏感,少量的数据变化会导致变量选择的较大变化。在应用 Lars 算法求解 Lasso 问题时,也需要面对这一问题。文献[10]提出了一种基于稳健相关矩阵的 Lars 变形算法,并使用 Bootstrap 算法稳定变量选择结果。

借鉴 Khan 的思路,现提出 SALP 方法: 将建立模型过程分别为两步,第一步,采用 Bayesian Bootstrap[11]算法重构数据样本。在这组重构样本上,使用 Lasso 方法训练多个模型。通过对多模型变量选择结果进行集成来实现变量预选,以降低无关变量引入模型的可能。第二步,在变量预选的基础上,使用 Lasso 方法进行最终模型的系数估计。Bayesian Bootstrap 算法按照 Dirichlet 分布的概率对样本进行抽样,研究表明,对于未知分布、小样本数据,该算法对比 Bootstrap 算法具有优势[12]。

在第一步的变量预选中,可以认为: 如果某个变量被较多模型选择,则表示该变量入选模型的可信度较高。因此可以通过考察变量被模型选择的次数来确定预选范围。

当变量被模型选择的次数属于或接近正态分布、卡方分布等常见分布时,可以构建变量属于模型的置信区间并进行判断,例如某一变量入选次数处于置信区间之外,则可将该变量排除。

当变量入选模型的次数无法用常见分布拟合时,可以用 Parzen 窗等非参数方法对未知分布进行估计,但更直接的方法是考察其经验分布。通过计算相关的分位数,将某一分位数对应的次数作为门限,据此对变量进行选择。3.2 节的检验将证明,变量预选过程应该以减少无关变量入选模型的几率为目标。如果在预选步骤移出较多的候选变量,将使最终模型的预报偏差变大。

由于 Lasso 方法对式(2)中各系数 $\beta_j$ 采用同样的 L1 罚,Zou[13]证明,Lasso 方法的变量选择过程是不连续的,或者说不具有"先知"(Oracle properties)性质。

所谓先知性质,定义为:

设 $A=\{i\mid \hat{\beta}_i\neq 0\}$,设 $\hat{\beta}_i(G)$ 是通过方法 $G$ 得到的模型系数估计。如果方法 $G$ 具有先知性质,则 $\hat{\beta}_i(G)$ 将满足如下条件:

(1) $\lim_{n\to\infty}P(\{i\mid \hat{\beta}_i(G)\neq 0\}=A)=1$;

(2) $\sqrt{n}(\hat{\beta}_i(G)-\beta_A)\to N(0,\Sigma)$,其中 $\Sigma$ 为真实模型参数的协方差矩阵。

Zou 认为: 这将导致最终模型中包含无关变量及估值的偏差。为改善这一问题,Zou 提出了 Adaptive Lasso 方法,对预报变量采取不同的惩罚权重系

数。Adaptive Lasso 方法的变量系数估计值如式(3)。

$$\beta_{(\text{Adaptive}-\text{Lasso})}=\underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}\omega_j|\beta_j|\} \tag{3}$$

式(3)中权重系数 $\omega_j=|\beta_{\text{ols}}|^{-\gamma},\gamma>0$。$\beta_{\text{ols}}$为最小二乘回归系数,当变量存在多重共线性问题,Zou 建议采用岭回归[14]系数替代 $\beta_{\text{ols}}$。

将 Adaptive Lasso 应用到 SALP 方法中,但对于小样本数据,使用最小二乘回归或岭回归权重系数的模型预报偏差较大,为寻求改善,采用偏最小二乘算法(partial least squares,Pls)系数 $\beta_{\text{pls}}$ 替代 $\beta_{\text{ols}}$。偏最小二乘算法由 S. Wold 等人提出[15],它可以在样本个数少于变量个数的条件下进行回归建模。在处理样本容量小、自变量多、变量间存在严重相关性问题方面有独特的优势。

SALP 方法的基本流程描述如下:

Input: 数据集$\{X,Y\}_1^n$,重构样本数量 $k$;

Output: 模型中待估计的回归系数 $\beta$;

Step1: 按照 Lasso 方法的要求进行数据预处理,令处理后的数据集满足: $\sum_{i=1}^{n}y_i=0,\sum_{i=1}^{n}x_{ij}=0,\sum_{i=1}^{n}x_{ij}^2=1$;

Step2: 使用 Bayesian Bootstrap 算法重构数据集,并生成训练集 $X(l),l=1,\cdots,k$;

Step3: 定义变量 $\text{Vote}(j)=0$,用于记录变量 $x_j$ 被模型选择的次数;

For $l=1$ to $k$

以 $X(l)$ 为训练集,使用 Adaptive Lasso 方法训练模型 $M(l)$;

For $j=1$ to $p$ //统计变量入选模型次数

If 变量 $x_j$ 被模型 $M(l)$ 选中

$\text{Vote}(j)=\text{Vote}(j)+1$

End If

End For

Step4: 根据变量入选模型次数统计结果对预报变量进行预选,生成新的数据集$\{X^*,Y\}_1^n$,$X^*$ 是 $X$ 的子集;

Step5: 在$\{X^*,Y\}_1^n$ 中抽取训练集,使用基于 Pls 权重系数的 Adaptive Lasso 方法训练模型,输出结果。

## 3 实验结果与讨论

### 3.1 数据说明

以碳钢在土壤中的腐蚀数据为研究对象,数据的样本容量为 70 个,包含了 104 个变量。其中变量 $x_1$ 至 $x_{13}$ 为碳钢埋件实验测量记录,变量 $y$ 是碳钢在土壤的腐蚀率记录。$x_{14}$ 至 $x_{104}$ 为检测预报变量与响应变量的非线性关系以及预报变量的交互效应是否存在而设置的虚拟变量。其中变量 $x_{14}$ 至 $x_{26}$ 是原变量的二次项集合,即$\{x_1^2,\cdots,x_i^2,\cdots,x_{13}^2\}$;变量 $x_{27}$ 至 $x_{104}$ 是原变量的交互项集合,即$\{x_1\times x_2,\cdots x_{(i-1)}\times x_i,\cdots,x_{12}\times x_{13}\}$,研究的目标是训练模型来拟合、预报腐蚀率 $y$,并考察其影响因素。

表 1 报告了实验数据的变量名称、变量定义和变量的均值、标准差。

表 1 碳钢的土壤腐蚀实验数据

Table 1 Experimental data of carbon steel corrosion in soil

| 变量名称 | 定义 | 单位 | 均值 | 标准差 |
|---|---|---|---|---|
| $x_1$ | 碳钢埋试时间 | a | 7.37 | 10.20 |
| $x_2$ | 土壤电阻值 | 欧姆 | 141.39 | 199.28 |
| $x_3$ | 含水量 | % | 20.54 | 7.05 |
| $x_4$ | 土壤 PH 值 | — | 7.60 | 1.41 |
| $x_5$ | 有机质含量 | % | 0.67 | 0.53 |
| $x_6$ | 氮含量 | % | 0.04 | 0.02 |
| $x_7$ | $HCO_3^-$ 离子含量 | % | 0.02 | 0.02 |
| $x_8$ | $CL^-$ 离子含量 | % | 0.07 | 0.20 |
| $x_9$ | $SO_4^{2-}$ 离子含量 | % | 0.15 | 0.33 |
| $x_{10}$ | $Ca^{2+}$ 离子含量 | % | 0.05 | 0.09 |
| $x_{11}$ | $Mg^+$ 离子含量 | % | 0.01 | 0.01 |
| $x_{12}$ | $K^+$ 离子含量 | % | 0.00 | 0.01 |
| $x_{13}$ | $Na^+$ 离子含量 | % | 0.06 | 0.14 |
| $x_{14}\sim x_{26}$ | $x_1-x_{13}$ 的二次项 | — | — | — |
| $x_{27}\sim x_{104}$ | $x_1-x_{13}$ 的交互项 | — | — | — |
| $y$ | 腐蚀率 | mm/a | 3.20 | 1.79 |

### 3.2 建立模型

由于碳钢的土壤腐蚀数据采集周期长,过程复杂,其中可能混杂离群点、错误记录等。为消除数据扰动的影响,首先进行 SALP 方法的第一步,对变量进行预选。

按照算法要求,将实验数据中预报变量集 $X$ 的 104 个变量标准化处理,响应变量 $Y$ 对中处理。然后采用 Bayesian Bootstrap 算法对预处理后的数据进行抽样,重构 500 个数据集。对于每个重构的数据集,采用基于 Pls 权重系数的 Adaptive Lasso 方法训练模型。

由于变量入选模型的次数分布不属于正态分布、卡方分布等常见分布,考察其经验分布。次数分布的最小值、10% 分位数、25% 分位数、中位数和最大值分别为 0、11、34、75、432。从分位数可以看出,有 10% 的变量被模型选中的次数少于 11 次,而 25% 的变量被模型选中的次数少于 34 次。

为确定变量预选门限,将被模型选择次数少于 11 次、34 次、75 次的变量分别移除,形成 3 个新的数据集。每个新数据集的预报变量是原预报变量集合的子集,样本容量为 70 个。新数据集记为数据集1、数据集2和数据集3。对于每个新数据集,随机

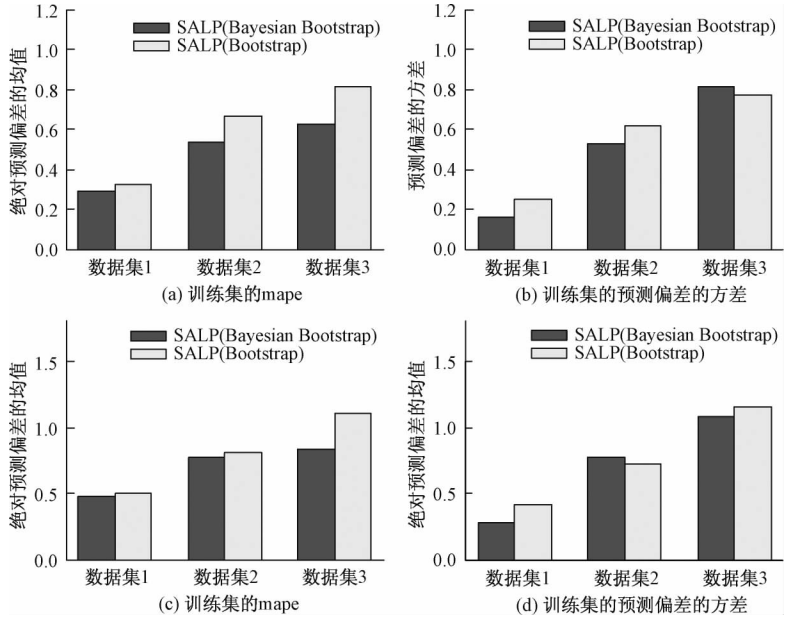


**图 2　SALP( with Bayesian Bootstrap) 和 SALP( with Bootstrap) 方法在不同数据集的性能对比**

Fig. 2　The performance comparison of SALP( with Bayesian Bootstrap) and SALP ( with Bootstrap) methods on different training set

抽取 60 个样本生成训练集,其余 10 个样本作为检验集。

应用基于 Pls 权重系数的 Adaptive Lasso 方法在 3 个新训练集上分别训练模型以确定变量预选门限。作为对比,本文还采用 Bootstrap 算法重构样本并重复了上述过程。实验结果如图 2 所示,对应的数据记录在表 2 中。

本文采用二个指标检测模型的性能,一是绝对预测偏差的均值 *mape*( mean of absolute prediction error) ,定义为 $mape = \frac{1}{m}\sum_i |y_i - y_i^*|$,其中 m 为数据集的样本容量,$y_i - y_i^*$ 是实际值和模型预报值的偏差; 二是预测偏差的方差。

图 2( a) 表示基于不同抽样算法的 SALP 方法在 3 个训练集上的 mape 对比; 图 2( b) 表示了对应的预测偏差的方差对比; 图 2( c) ~( d) 则分别表示了对应指标在测试集上的表现。图 2 显示,采用 Bayesian Bootstrap 算法抽样的模型性能略优于采用 Bootstrap 算法的对应模型。这可能是由于 Bootstrap 的应用在取决于经验分布的选取和样本数的大小[16]。当样本量小不足以提供总体分布信息时,其结果并不可靠。

从图 2 和表 2 的记录可以发现,如果在变量预选过程中移除的变量过多,如数据集 2 和数据集 3,模型的预报性能会变得较差。而适当的变量预选,如数据集 1,却可以明显的提高模型性能。因此可以认为在变量预选过程中,应该以减少无关变量入选模型可能性为目标,如果移除变量过多可能造成最终模型漏掉重要变量或过于稀疏,进而影响其预报性能。

**表 2　不同变量选择门限的模型性能对比**

**Table 2　Model performance Comparison with different variable selection strategy**

| 数据 | | 指标 | 算法 | |
|---|---|---|---|---|
| | | | SALP( with Bayesian Bootstrap) | SALP( with Bootstrap) |
| 数据集 1 | 训练集 | *mape* | 0. 288 | 0. 328 |
| | | 偏差方差 | 0. 160 | 0. 249 |
| | 检验集 | *mape* | 0. 483 | 0. 501 |
| | | 偏差方差 | 0. 288 | 0. 414 |
| 数据集 2 | 训练集 | *mape* | 0. 534 | 0. 655 |
| | | 偏差方差 | 0. 531 | 0. 617 |
| | 检验集 | *mape* | 0. 771 | 0. 810 |
| | | 偏差方差 | 0. 776 | 0. 723 |
| 数据集 3 | 训练集 | *mape* | 0. 629 | 0. 817 |
| | | 偏差方差 | 0. 813 | 0. 776 |
| | 检验集 | *mape* | 0. 934 | 1. 018 |
| | | 偏差方差 | 1. 076 | 1. 151 |

根据实验结果,将变量被模型选择次数的门限定为 11 次。

在完成变量的预选后,进行 SLAP 方法的第二步: 以数据集 1 作为变量预选结果,随机抽取 60 个样本组成训练集,其余 10 个样本作为检验集。采用基于 Pls 权重系数的 Adaptive Lasso 方法训练模型。

### 3.3 模型性能检验与讨论

作为对比，在原数据集上选择与 SALP 方法第二步相同的样本组成训练集和检验集，训练了基于岭回归权重系数的 Adaptive Lasso 模型、支持向量回归(SVR)模型和神经网络模型。这里的数据集包含全部的变量。其中 Adaptive Lasso 方法的调控参数 $\lambda$ 和 $\gamma$ 通过 10 折交叉认证选取；SVR 模型使用 R 软件包 E1071 计算；神经网络模型使用 R 软件包 NNET 实现。后两者的参数主要采用了默认值。

在检验集上测试各模型的结果如图 3 所示。

从图 3 可以看出，SALP 方法所建立的模型可以较好的预测碳钢在自然环境土壤中的腐蚀率，尽管模型中包含较多的变量，但没有出现过拟合情况。在处理与实现对象数据类似的高维小样本数据时，SALP 方法的性能优于未经特殊优化的 Adaptive Lasso、SVR 和神经网络等方法，而且由于最终输出为线性模型，具有较好的可解释性。

在验证算法的有效性后，为了充分利用数据中的信息，采用全部 70 个样本重新训练模型。最终模型形如式(1)，共包含 35 个变量。回归系数 $\beta$ 的估值记录在表 3 中。

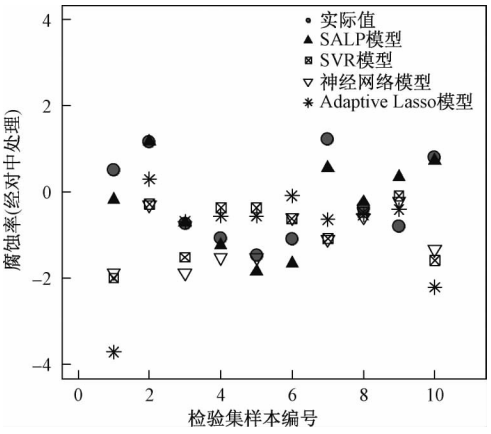


图 3 SALP 模型与其他模型在测试集的预测结果对比

Fig. 3 The prediction comparison of SALP model with other model based on test set

SALP 方法对预报变量进行了标准化处理，因此表 3 的模型系数 $\beta$ 为标准化系数，可以通过系数的绝对值排序来判断预报变量对于响应变量的重要性[17]。对本文的使用数据而言，碳钢在土壤中的埋试时间、土壤中 $Ca^{2+}$ 离子含量和 $Mg^{+}$ 离子含量的交互项、$K^{+}$ 离子含量、$Mg^{+}$ 离子含量的二次项、埋试时间和土壤 pH 值的交互项等是碳钢在土壤中腐蚀率的主要影响因素。

由于模型为线性模型，通过控制其他预报变量，可以考察某个变量对于腐蚀率的净效应。例如根据表 3 中的记录，碳钢在土壤中的埋试时间每增加 1 个标准差，腐蚀率就会下降 309.1 个标准差；土壤含水量和 $CL^{-}$ 离子含量的交互项系数为 214.8，表现出了变量间可能存在较强的相互促进作用；与之相反，土壤氮含量和 $Ca^{2+}$ 离子含量的交互项系数为 −143.08，说明二个变量可能对腐蚀率的作用存在相互削弱的关系等等。当然，任何数学模型都不可能精确地描述现实世界和自然[18]，SALP 模型的可解释性能够为腐蚀因素分析提供线索，但模型产生的结果是否符合现实规律，仍然需要通过实践进行验证。

表 3 碳钢的腐蚀率模型系数估值

Table 3 Coefficients of carbon steel corrosion rate model

| 变量描述 | 系数 $\beta$ | 变量描述 | 系数 $\beta$ |
|---|---|---|---|
| 埋试时间 | −309.31 | 电阻 × $CL^{-}$ | 46.96 |
| 含水量 | −24.47 | 电阻 × $Ca^{2+}$ | −156.17 |
| $HCO_3^{-}$ | 70.24 | 电阻 × $K^{+}$ | 69.16 |
| $K^{+}$ | −266.38 | 含水 × PH | 41.34 |
| (埋试时间)$^2$ | 95.88 | 含水量 × 有机质 | 150.00 |
| (电阻)$^2$ | −29.08 | 含水量 × 氮含量 | −124.67 |
| (pH)$^2$ | −22.40 | 含水量 × $CL^{-}$ | 214.48 |
| ($HCO_3^{-}$)$^2$ | −116.16 | PH × 有机质 | −124.15 |
| ($SO_4^{2-}$)$^2$ | −70.73 | PH × 氮含量 | 135.24 |
| ($Mg^{+}$)$^2$ | −279.11 | 有机质 × 氮含量 | −29.92 |
| 埋试时间 × 电阻 | 24.00 | 有机质 × $HCO_3^{-}$ | 22.80 |
| 埋试时间 × PH | 217.45 | 有机质 × $SO_4^{2-}$ | 198.11 |
| 埋试时间 × $HCO_3^{-}$ | −70.61 | 有机质 × $K^{+}$ | −49.03 |
| 埋试时间 × $Mg^{+}$ | 94.11 | 氮含量 × $Ca^{2+}$ | −143.08 |
| 电阻 × PH | 37.48 | 氮含量 × $Mg^{+}$ | −113.91 |
| 电阻 × 氮含量 | 21.78 | 氮含量 × $K^{+}$ | 153.83 |
| $HCO_3^{-}$ × $K^{+}$ | 108.44 | 氮含量 × $Na^{+}$ | −50.00 |
| $Ca^{2+}$ × $Mg^{+}$ | 299.98 | | |

## 4 结束语

提出了一种基于 Lasso 的 SALP 方法，改善了 Lasso 方法求解过程中对数据扰动敏感的问题和处理小样本数据时的参数估计问题。采用碳钢在自然环境土壤中的腐蚀数据进行方法检验，实验结果表明，SALP 方法建立的模型能够准确的预测碳钢在土壤中的腐蚀率，具有较好的解释性。该方法适用于自然环境中材料腐蚀率的预测及类似研究领域。

### 参 考 文 献

1 邓永先，吕国志，张有宏，等. 对结构腐蚀增长模型的探讨. 科学技术与工程，2007;7(19)：4956—4960
Deng Yongxian，Lü Guozhi，Zhang Youhong，*et al*. Discussion of structure corrosion growth. Science Technology and Engineering，2007;7(19)：4956—4960

2 Sadowski L. Non-destructive investigation of corrosion current density in steel reinforced concrete by artificial neural networks. Archives of Civil and Mechanical Engineering，2013;13(1)：104—111

3 Liang Ping, Du Cuwei, Li Xiaogang. Grey relational space analysis of effect of environmental factors on corrosion resistance of x70 pipeline steel in ying tan soil simulated solution. Corrosion & Protection, 2009; (4): 23—27

4 Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine, 1997, 16(4): 385—395

5 杨　关, 张向东, 冯国灿, 等. 图模型在彩色纹理分类中的应用. 计算机科学, 2011; 38(10): 273—277
Yang Guan, Hang Xiangdong, Feng Guocan, *et al*. Applications of graphical models in color texture classification. Journal of Frontiers of Computer Science & Technology, 2011; 38(10): 273—277

6 Tibshirani R, Saunders M, Rosset S, *et al*. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005; 67(1): 91—108

7 Hastie T, Tibshirani R, Friedman J H. 统计学习基础—数据挖掘、推理与预测. 范　明, 柴玉梅, 译. 北京: 电子工业出版社, 2004: 337—384
Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Tr. by Fan Ming, Cgau Yumei. Beijing: Electronic Industry Press, 2004: 337—384

8 Osborne M R, Presnell B, Turlach B A. On the Lasso and its dual. Journal of Computational and Graphical Statistics, 2000; 9(2): 319—337

9 Efron B, Hastie T, Johnstone I, *et al*. Least angle regression. The Annals of Statistics, 2004; 32(2): 407—499

10 Khan J A, van Aelst S, Zamar R H. Robust linear model selection based on least angle regression. Journal of the American Statistical Association, 2007; 102(480): 1289—1299

11 Rubin D B. The bayesian bootstrap. The Annals of Statistics, 1981; 9(1): 130—134

12 Clyde M A, Lee H K H. Bagging and the bayesian bootstrap. Artificial Intelligence and Statistics. Key West: 2001: 169—174

13 Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 2006; 101(476): 1418—429

14 Hoerl A E, Kennard R W. Ridge regression: applications to nonorthogonal problems. Technometrics, 1970; 12(1): 69—82

15 Wold S, Ruhe A, Wold H. The collinearity problem in linear regression, The partial least square (PLS) approach to generalized inerses. SIAM J Science Statistics Computer, 1984; 5(2): 735—743

16 Roff D A. Introduction to computer-intensive methods of data analysis in biology. New York: Cambridge University Press, 2006: 126—127

17 王海燕, 杨方廷, 刘　鲁. 标准化系数与偏相关系数的比较与应用. 数量经济技术经济研究, 2006; (9): 150—155
Wang Hai-yan, Yang Fang-ting, Loi Lu. Comparison and application of standardized regressive coefficient & partial correlation coefficient. The Journal of Quantitative & Technical Economics, 2006; (9): 150—155

18 吴喜之, 马景义, 吕晓玲, 等. 数据挖掘前沿问题. 北京: 中国统计出版社, 2009: 7—13
Wu Xizhi, Ma Jingyi, Lv Xiaoling, *et al*. The frontier issues of data mining. Beijing: China Statistics Press, 2009: 7—13.

# Corrosion Rate Prediction of Carbon Steel in Soil based on Lasso Method

LU Qing, MU Zhi-chun*

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China)

**[Abstract]** A Lasso-based method was proposed named as Step Adaptive Lasso with Pls (SALP). To eliminate the data perturbation or outlier influence on model training, SALP method uses Bayesian Bootstrap algorithm to reconstruct data sets and establishes models based on them. Then the predictors can be prescreened through integrating the result of those models. In order to deal with small sample size problem of data set, the partial least squares weighting factor is applied. By using SALP method, this study established a corrosion rate model based on the corrosion data of carbon steel in soil. The result shows that, the model established by SALP could precisely describe and predict the corrosion rate of carbon steel in soil. The method can be applied to prediction of material corrosion rate in natural environment or similar research field.

**[Key words]** Lasso　corrosion rate　carbon steel　corrosion of the natural environment　soil