基于流形降维和梯度提升树的大气腐蚀速率预测模型

梁喜旺，付冬梅，杨焘

（北京科技大学 自动化学院，北京 100083）

摘要：目的 为了挖掘大气腐蚀速率与金属化学成分和暴露时间两个影响因素的定量关系，针对数据集特点，提出一种局部保持投影（Locality Preserving Projection）和梯度提升树（Gradient Boosting Decision Trees）结合的大气腐蚀速率预测模型（LPP-GBDT）。**方法** 采用 LPP 算法对金属化学成分进行降维处理，得到金属化学成分低维特征，然后引入时间因素，并利用 GBDT 进行建立预测模型。以青岛海洋大气环境下积累的 16 年内的腐蚀速率数据进行模型性能验证，**结果** LPP-GBDT 模型测试集平均绝对误差为 1.73 μm/a，平均绝对百分误差为 6.30%。正交化 LPP-GBDT 模型测试集平均绝对误差为 1.21 μm/a，平均绝对百分误差为 4.42%。**结论** 与多个典型预测模型相比，LPP-GBDT 模型基于暴露时间和化学成分因素实现了大气腐蚀速率较为准确的预测，对特定环境下金属选材具有一定的参考价值。

关键词：金属化学成分；腐蚀速率；流形方法；梯度提升决策树

DOI：10.7643/ issn.1672-9242.2018.06.008

中图分类号：TJ07; TG172 **文献标识码：**A

文章编号：1672-9242(2018)06-0041-07

Predicting Method of Atmospheric Corrosion Rate Based on Manifold Dimension Reduction and Gradient Boosting Decision Trees

LIANG Xi-wang, *FU Dong-mei*, *YANG Tao*

(School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China)

ABSTRACT: Objective To analyze quantitatively the relationship between atmospheric corrosion rate and the two factors including outdoor exposure time and chemical components of steel materials, an atmospheric corrosion rate predicting model was proposed in combination with locality preserving projection (LPP) and gradient boosting decision trees (GBDT). **Methods** First, LPP was applied to have dimension reduction process on chemical components of steels to get low-dimensional features. Then GBDT were used to build a predicting model. Corrosion rate data of marine atmospheric environment within 16 years in Qingdao were used to validate the proposed model. **Results** Testing MAE and MAPE of LPP-GBDT model were 1.73 μm/a and 6.30% respectively. Testing MAE and MAPE of LPP-GBDT with orthogonalization were 1.21 μm/a and 4.42% respectively. **Conclusion** Compared with other common predicting methods, the proposed model has preferable prediction accuracy and offers some reference to steels selection in specific environment.

KEY WORDS: chemical components of steels; corrosion rate; manifold methods; GBDT

收稿日期：2018-03-26；**修订日期：**2018-03-26

基金项目：国家重点研发计划（2017YFB0702104）；博士后科学基金（2017M620615）；中央高校基本科研业务费（FRF-TP-16-082A1）

作者简介：梁喜旺（1993—），男，河北人，硕士研究生，主要研究方向为大气腐蚀数据挖掘。

通讯作者：付冬梅（1963—），女，辽宁人，博士，教授，主要研究方向为数据挖掘与共享、智能算法与控制、图形分析与处理。

大气环境下的金属腐蚀作为一种常见现象，会造成严重的经济损失、安全隐患、资源浪费[1-2]，研究和掌握大气腐蚀规律具有重要的工程意义。大气腐蚀受到大气环境、金属化学成分含量和暴露时间等多方面因素影响，不同于基于腐蚀速率与环境因素关系的研究，预测新环境下特定材料的腐蚀行为，文中分析了特定大气环境下金属化学成分含量和暴露时间因素对大气腐蚀速率的影响，建立了腐蚀速率预测模型。

文中数据集具有高维、非线性且小样本的特点，化学成分对腐蚀速率的影响非常复杂，多达 14 种的化学元素影响程度各不相同，部分元素之间还存在相互作用的现象。由于样本种类有限，含有某些元素如铌、镭的金属比较少，这些特征变化不大，出现大量 0 值，带来了特征突变、数据冗余等问题，为建模预测带来困难。针对这些问题，文中首先对化学成分数据进行降维处理，得到更为约简、预测能力更强的特征。一般认为，腐蚀现象的发生是有一定条件的，各个化学成分之间存在一定形态的约束关系，这种约束关系决定了金属材料自身的耐腐蚀性。常用的主成分分析 PCA[3]是基于数据欧式距离全局结构的线性降维方法，可能会破坏数据间的非线性约束关系。流形方法[4]在保持数据全局或局部约束关系的同时，寻找一个映射子空间，使得降维后数据更加接近原始数据的非线性本质，比较具有代表性的有 ISOMAP，LLE，LE 等。等度规映射 ISOMAP 是多维尺度分析的拓展，尽量保持全局流形上两点距离不变；局部线性嵌入 LLE 在样本点和它的邻域点之间构造一个重构权向量，在低维空间中保持权值不变；拉普拉斯特征映射 LE 构造样本点之间的关联矩阵，并在重构低维嵌入时，保持高维空间中距离近的点在低维空间距离也近。上述流形方法虽然能实现高维数据的约简，却不能得到高维空间到低维空间的显式映射，降维处理只限于训练样本，难以应用到测试样本，此问题能通过引入线性化过程得以解决[5]。局部保持投影 LPP 是[6] LE 算法的线性化算法，依据流形思想，保持局部信息，并得到高维数据到低维嵌入的线性映射。文中采用 LPP 算法对金属化学成分进行降维处理，此外，为了较好地重构低维嵌入，提高局部保持能力，对 LPP 算法进行正交化改进。

LPP 降维后的低维特征未与大气腐蚀速率建立联系，需要利用一定的建模方法实现腐蚀速率的预测。腐蚀速率预测领域常用的方法有灰色预测模型[7]、人工神经元网络[8]和 CART 回归树[9]等。典型的灰色 GM(1,1)模型适合单一时序预测，难以引入金属化学成分的影响；神经元网络虽然能实现基于多个因素的预测，但需要大量样本和复杂的网络结构，且易于过拟合；CART 回归树从单个特征入手，遍历所有特征，寻找最优划分特征和最优划分点，并在子空间重复划分，比较适合文中数据。单个回归模型结构简单，预测精度较低，容易出现过拟合现象，并对噪声敏感[10]。针对这些问题，文中采用梯度提升决策树算法。GBDT 是近年来最有效的机器学习方法之一，是一种基于 CART 树的集成模型，最早由 Friedman 提出[11]，具有较好的健壮性和泛化能力，能有效提升预测准确性。同时，GBDT 模型的可解释性比较好，能够分析影响腐蚀的关键因素。

文中主要利用 LPP 算法挖掘了高维、非线性且小样本数据的本质特征，并结合 GBDT 模型实现了大气腐蚀速率的预测，同时与几种典型预测模型进行对比研究。

1 LPP 算法及正交化改进

LPP 作为流形学习的重要分支，是一种典型的基于近邻图的降维方法，是拉普拉斯特征映射 LE 算法的线性化算法。为了方便表示，设原始数据集为 $X=\{x_1,x_2,\cdots,x_n\}$，$x_i\in\mathbb{R}^D$，低维嵌入为 $Y=\{y_1,y_2,\cdots,y_n\}^T$，$y_i\in\mathbb{R}^d$，满足 $Y=\boldsymbol{A}^TX$，其中，$\boldsymbol{A}=\left[a_1,a_2,\cdots,a_d\right]\in\mathbb{R}^{D\times d}$，为线性映射矩阵。LPP 的目标是在寻找最优映射的同时，保持原始数据中的局部几何结构，通过 k 近邻法构建近邻图 $G=\{X,W\}$，若 x_i 和 x_j 互为近邻点，则通过热核函数为两点赋予连接权值，定义如式(1)所示。

$$W_{ij}=\exp\left(\frac{-\left\|x_i-x_j\right\|^2}{t}\right) \tag{1}$$

式中：$\|\cdot\|$ 代表 L2 范数；t 为热核参数。

LPP 优化目标函数[12]：

$$\min_{\boldsymbol{A}}\sum_{i,j=1}^{n}\left\|\boldsymbol{A}^{\mathrm{T}}x_i-\boldsymbol{A}^{\mathrm{T}}x_j\right\|^2\boldsymbol{W}_{ij}\quad\text{s.t.}\sum_{i=1}^{n}D_{ii}\left\|\boldsymbol{A}^{\mathrm{T}}x_i\right\|^2=\boldsymbol{I} \tag{2}$$

式中：$\boldsymbol{I}$ 为单位矩阵；$\boldsymbol{D}$ 为对角线矩阵，为 $\boldsymbol{W}$ 矩阵的行求和或列求和，即 $D_{ii}=\sum_j W_{ji}$。设 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ 为拉普拉斯矩阵。为了得到唯一解，需要满足约束条件 $\sum_{i=1}^{n}D_{ii}\left\|\boldsymbol{A}^{\mathrm{T}}x_i\right\|^2=\boldsymbol{I}$。

由式(1)热核函数定义可知，原始高维空间距离较近的点之间具有较大的连接权值，因此，映射到低维空间中的点只有保持较近的距离才能使得目标函数达到最小。采用该方法计算的连接权值 W_{ij} 保证了高维空间中处于近邻的数据点在低维空间中距离也很近。

显然，可以将式(1)改写成：

$$\min_{\boldsymbol{A}}\sum_{i,j=1}^{n}\left\|\boldsymbol{A}^{\mathrm{T}}x_i-\boldsymbol{A}^{\mathrm{T}}x_j\right\|^2W_{ij}/\sum_{i=1}^{n}D_{ii}\left\|\boldsymbol{A}^{\mathrm{T}}x_i\right\|^2 \tag{3}$$

先考虑分子项

$$\sum_{i,j=1}^{n}\left\|\boldsymbol{A}^{\mathrm{T}}x_i-\boldsymbol{A}^{\mathrm{T}}x_j\right\|^2W_{ij}=$$

$$\mathrm{tr}\left[\boldsymbol{A}^{\mathrm{T}}\left(\sum_{i,j=1}^{n}(x_i-x_j)(x_i-x_j)^{\mathrm{T}}W_{ij}\right)\boldsymbol{A}\right] \tag{4}$$

式中：tr(·)代表矩阵迹操作。令 e_i 表示单位向量，第 i 个元素为 1，其余为 0，因此有：

$$\sum_{i,j=1}^{n}(x_i-x_j)(x_i-x_j)^{\mathrm{T}}W_{ij}=$$

$$\boldsymbol{X}\left[\sum_{i,j=1}^{n}(e_i-e_j)(e_i-e_j)^{\mathrm{T}}W_{ij}\right]\boldsymbol{X}^{\mathrm{T}} \tag{5}$$

展开括号内项，并重新合并项可得：

$$\sum_{i,j=1}^{n}(e_i-e_j)(e_i-e_j)^{\mathrm{T}}W_{ij}=2\boldsymbol{D}-2\boldsymbol{W} \tag{6}$$

因此可得：

$$\sum_{i,j=1}^{n}\left\|\boldsymbol{A}^{\mathrm{T}}x_i-\boldsymbol{A}^{\mathrm{T}}x_j\right\|^2W_{ij}=2\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}) \tag{7}$$

同理可得：

$$\sum_{i=1}^{n}D_{ii}\left\|\boldsymbol{A}^{\mathrm{T}}x_i\right\|^2=2\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}) \tag{8}$$

因此可将 LPP 优化问题(1)转化成式(9)所示的矩阵迹之比形式。

$$\min_{\boldsymbol{A}}\frac{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})} \tag{9}$$

通常来说，矩阵迹之比优化问题是非凸的，同时不存在闭式解，一般转化为更为简单的比值之迹形式[12]，如式(10)所示：

$$\min_{\boldsymbol{A}}\mathrm{tr}((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}) \tag{10}$$

上式能够通过以下广义特征值问题求解：

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\boldsymbol{A}}a=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\boldsymbol{A}}a \tag{11}$$

A 由式(11)的前 d 个最小特征值对应的特征向量组成。

LPP 兼顾了局部最小映射和保持全局信息，但 LPP 得到的映射 $\boldsymbol{A}$ 是非正交的，由式(3)和欧式距离定义，低维空间中 y_i 和 y_j 的距离可以表示为式(12)。可见，非正交的 $\boldsymbol{A}$ 在数据重构的过程中必然造成原始欧式空间结构不能完全被恢复。

$$\mathrm{dis}(y_i,y_j)=\sqrt{(x_i-x_j)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}(x_i-x_j)} \tag{12}$$

通过正交化投影矩阵 $\boldsymbol{A}$，使得 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{I}$，那么原始数据空间结构能被完全保持，局部信息损失降低。此外，降维后数据正交，特征区分度更高，有利于建模预测。文中采用一种基于 QR 分解的正交化 LPP 方法[6]。由式(10)可得出一个结论：若 $\hat{\boldsymbol{A}}$ 为它的一个最优解，则 $\hat{\boldsymbol{A}}\boldsymbol{V}$ 也是它的一个最优解，$\boldsymbol{V}$ 是任意可逆矩阵，因为：

$$\mathrm{tr}((\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{A}}\boldsymbol{V})^{-1}\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{A}}\boldsymbol{V})=$$

$$\mathrm{tr}((\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{A}}) \tag{13}$$

QR 分解是一种应用广泛的矩阵分解方式，将矩阵分解为正交矩阵 $\boldsymbol{Q}$ 和上三角矩阵 $\boldsymbol{R}$ 的乘积形式，对式(10)最优解 $\hat{\boldsymbol{A}}$ 进行 QR 分解：$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{A}}\boldsymbol{R}$。可得 $\tilde{\boldsymbol{A}}=\hat{\boldsymbol{A}}\boldsymbol{R}^{-1}$，由上述结论可知，$\tilde{\boldsymbol{A}}$ 也是优化问题(10)的最优解，并满足正交约束条件：$\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\mathrm{T}}=\boldsymbol{I}$。

文中采用的正交化 LPP 算法首先求解原始 LPP 算法投影矩阵 $\hat{\boldsymbol{A}}$，然后对 $\hat{\boldsymbol{A}}$ 进行 QR 分解，得到正交矩阵 $\tilde{\boldsymbol{A}}$，最终得到低维嵌入 $\boldsymbol{Y}=\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{X}$。LPP 算法为非监督学习，低维数据集没有与腐蚀速率建立联系，需要借助回归模型实现腐蚀速率预测。

2　GBDT 模型

梯度提升决策树（GBDT）是一种提升算法，其原理是将大量简单 CART 树在提升过程中进行集成，以提高树模型的预测能力。由于基于决策树算法，GBDT 具有较好的模型可解释性[13]，为分析腐蚀影响因素的重要性提供了一种方法。

2.1　GBDT 基本算法

假设输入训练样本集为：$\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$。为了寻找回归树的最优组合，在每次迭代过程中顺序添加新的回归树来减少预测误差，新加入的回归树建立在之前所有树的负梯度之上。

估计函数 $f(x)$预测 y 的损失函数 $L(f)$定义为：

$$L(f)=\sum_{i=1}^{N}L(y_i,f(x_i)) \tag{14}$$

在回归问题中，一般为平方误差损失：

$$L(f)=\sum_{i=1}^{N}\frac{1}{2}(y_i-f(x_i))^2 \tag{15}$$

在梯度提升框架 M 次迭代中，全局函数估计 $\hat{f}(x)$ 可以由加法模型表示：

$$f_M(x)=\sum_{i=0}^{M}f_i(x) \tag{16}$$

其中，$f_0(x)$为初始值，定义为：

$$f_0(x)=\arg\min_{c}\sum_{i=1}^{N}L(yi,c) \tag{17}$$

在迭代次数 $m=1,2,3,\cdots,M$ 中，对样本 $i=1,2,3,\cdots,N$ 计算负梯度：

$$g_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{18}$$

利用 (x_i,g_{mi})，$i=1,2,3,\cdots,N$，拟合一棵 CART 回归树，得到第 m 棵树，其对应的叶子节点区域为 R_{mj}，$j=1,2,\cdots,J$，为回归树 m 的叶子节点个数。对叶子区域 $j=1,2,\cdots,J$，计算最佳拟合值，并更新强学习器：

$$c_{mj}=\arg\min_{c}\sum_{x\in R_{mj}}L(y_i,f_{m-1}(x_i)+c) \tag{19}$$

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{20}$$

迭代结束后，得到全局强学习器：$f(x)=f_M(x)$。为了防止过拟合，需要对 GBDT 进行正则化处理，在式(20)加入正则化项 η，即学习率（learning rate），则有：

$$f_m(x)=f_{m-1}(x)+\eta\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{21}$$

η 的取值范围为：$0<\eta\leqslant 1$。对于同样的训练集学习效果，较小的 η 需要更多的迭代次数，即回归树的总棵数；较大的 η 容易出现过拟合，通常同时调节迭代次数和学习率来决定模型的预测性能。

2.2 GBDT 的模型可解释性

在腐蚀速率预测中，模型的可解释性十分重要，GBDT 模型通过计算特征重要性来分析影响腐蚀的关键因素，Friedman 在 GBM 论文中[12]提出的方法：

设特征总数为 D，特征 d $(d=1,2,\cdots,D)$ 的全局重要性通过特征在单个树中的平均值来衡量：

$$I_d^2=\frac{1}{M}\sum_{m=1}^{M}I_d^2(T_m) \tag{22}$$

式中：M 是树的数量；T_m 为第 m 棵树。特征 d 在单棵树中的重要性为：

$$I_d^2(T)=\sum_{j=1}^{J-1}\hat{i}_j^2\mathbb{1}(v_j\sim d) \tag{23}$$

式中：J 为树的叶子节点数量；v_j 是和节点 j 相关联的特征；$\hat{i}_j^2$ 是节点 j 分裂后平方损失的减少值；$\mathbb{1}(v_j\sim d)$ 为示性函数，当 v_j 与特征 d 相关联时，示性函数值为 1，否则为 0。

3 实验结果与分析

3.1 数据集准备和分析

文中采用数据来源于中国腐蚀与防护网黑色金属大气腐蚀数据库青岛腐蚀站点数据，包含了暴露时间、碳、硅、锰、硫、磷等共 14 种化学元素含量参数和实验金属的腐蚀速率，共 16 种实验金属，80 个样本，部分腐蚀速率数据见表 1。对于每一个站点而言，每年的平均环境因素变化不大，为了便于分析，可忽略环境因素影响，分析特定站点下的金属合金元素含量和暴露时间对腐蚀的影响。

3.2 预测性能评估方法

文中采用平均绝对误差 MAE 和平均绝对百分误差 MAPE 来评估模型的预测效果。平均绝对误差 MAE 计算预测值和实际值之间偏差绝对值的平均，计算公式为：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{24}$$

为了评估预测误差相对于实际值的大小，还采用了平均绝对百分比误差 MAPE，计算公式为：

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{y_i}\times 100\% \tag{25}$$

式中：N 为样本数量；y 为实际值；$\hat{y}$ 为模型预测值。

表 1 青岛腐蚀站点部分数据

种类	时间/a	各元素质量分数/%														腐蚀速率/($\mu m\cdot a^{-1}$)
		碳	硅	锰	硫	磷	铬	钼	铝	钒	钛	铜	铌	镍	镉	
D36	1	0.14	0.39	1.4	0.018	0.022	0.05	0	0.025	0.03	0.02	0.05	0.03	0	0	54.5
D36	2	0.14	0.39	1.4	0.018	0.022	0.05	0	0.025	0.03	0.02	0.05	0.03	0	0	37.4
D36	4	0.14	0.39	1.4	0.018	0.022	0.05	0	0.025	0.03	0.02	0.05	0.03	0	0	28.5
D36	8	0.14	0.39	1.4	0.018	0.022	0.05	0	0.025	0.03	0.02	0.05	0.03	0	0	21.8
D36	16	0.14	0.39	1.4	0.018	0.022	0.05	0	0.025	0.03	0.02	0.05	0.03	0	0	18.5
3C	1	0.14	0.4	0.9	0.027	0.035	0	0	0	0	0	0.08	0	0	0	58.9
3C	2	0.14	0.4	0.9	0.027	0.035	0	0	0	0	0	0.08	0	0	0	40.7
3C	4	0.14	0.4	0.9	0.027	0.035	0	0	0	0	0	0.08	0	0	0	30
3C	8	0.14	0.4	0.9	0.027	0.035	0	0	0	0	0	0.08	0	0	0	23.9
3C	16	0.14	0.4	0.9	0.027	0.035	0	0	0	0	0	0.08	0	0	0	19.8

3.3 模型建立过程

LPP-GBDT 预测模型分为两部分，第一步利用 LPP 对金属化学成分数据进行降维处理，第二步利用低维数据训练 GBDT 模型，实现腐蚀速率的预测。该模型需要调节的参数共有 5 个，分别为 LPP 算法的目标维数 d、近邻点个数 k、热核函数参数 t、GBDT 算法的迭代次数（回归树数量）M、学习率 η。考虑

到时间开销和计算机性能，分两步优化参数，采用留一法交叉验证，以 GBDT 预测的平均绝对误差作为评价标准。

以确定 LPP 参数为例，首先将 GBDT 模型参数固定为 M=100，η=0.1，优化 LPP 算法 3 个参数。参数区间设置为：d 为区间[2,13]内的整数；k 为区间[2,27]内的整数；t 为区间[0.05,2]内的浮点数，步长为 0.05。同时搜索了效果对比方法 PCA 的参数，参数优化结果见表 2。

表 2　参数优化结果

参数	PCA	LPP	正交化 LPP
d	4	4	4
k	*	12	15
t	*	1.20	1.20

LPP 算法降维处理可以视为一个特征重构的过程，LPP 降维结果如式(26)—(29)所示，其中，$Featurei$（i=1,2,⋯,4）表示构造的低维特征：

$$\begin{aligned}Feature1 = &-0.152w_C - 0.080w_{Si} - 0.028w_{Mn} + \\ &0.227w_S - 0.241w_P - 0.026w_{Cr} + 0.055w_{Mo} - \\ &0.003w_{Al} + 0.158w_V - 0.375w_{Ti} + 0.027w_{Cu} - \\ &0.194w_{Nb} + 0.005w_{Ni} - 0.808w_{Re}\end{aligned} \tag{26}$$

$$\begin{aligned}Feature2 = &0.015w_C - 0.057w_{Si} - 0.0561w_{Mn} - \\ &0.614w_S - 0.542w_P + 0.058w_{Cr} + 0.001w_{Mo} + \\ &0.010w_{Al} - 0.465w_V - 0.217w_{Ti} + 0.169w_{Cu} + \\ &0.147w_{Nb} + 0.071w_{Ni} - 0.028w_{Re}\end{aligned} \tag{27}$$

$$\begin{aligned}Feature3 = &-0.045w_C - 0.095w_{Si} - 0.070w_{Mn} + \\ &0.600w_S - 0.514w_P - 0.093w_{Cr} + 0.066w_{Mo} + \\ &0.001w_{Al} - 0.118w_V - 0.300w_{Ti} - 0.032w_{Cu} - \\ &0.081w_{Nb} - 0.037w_{Ni} + 0.484w_{Re}\end{aligned} \tag{28}$$

$$\begin{aligned}Feature4 = &-0.170w_C - 0.040w_{Si} - 0.1300w_{Mn} - \\ &0.382w_S - 0.008w_P + 0.080w_{Cr} + 0.094w_{Mo} - \\ &0.001w_{Al} + 0.634w_V - 0.415w_{Ti} + 0.083w_{Cu} - \\ &0.304w_{Nb} + 0.118w_{Ni} - 0.323w_{Re}\end{aligned} \tag{29}$$

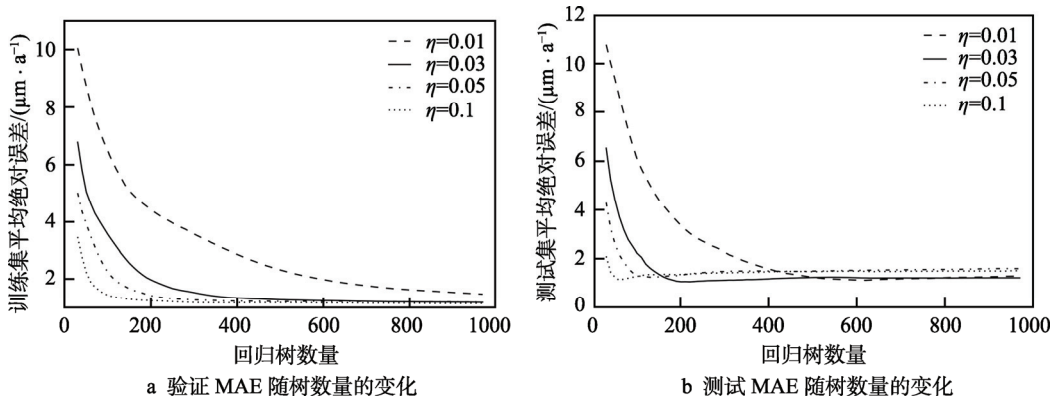
从降维结果可以看出，金属化学成分数据集通过不同的降维方法降至 4 维具有较好的预测能力，说明此数据集的本征维数极有可能为 4 维，需要更多后续研究加以验证。

在获得 LPP 参数后，优化 GBDT 参数，区间设置为：M 为区间[30,1000]内的整数，以 10 为步长；学习率 η 分别取 0.01，0.03，0.05，0.1，结果如图 1 所示。可以看出，训练集误差随回归树数量的增加而降低并趋于不变，降低速度随 η 的增加而变大。当 η 比较大时，测试集很快出现过拟合现象；若 η 太小，则需要较多的基学习器个数（M）。结合训练、测试误差及模型复杂度综合考虑，确定 GBDT 的参数为：M=600，R=0.03。

3.4　预测模型性能检验

为了验证文中建立模型的预测性能和泛化能力，随机选取 4 种金属的共 20 个样本作为测试集，其余 60 个样本作为预测模型训练样本，采用留一交叉验证训练模型参数。基于原始数据，不同模型预测结果见表 3。其中，SVR 支持向量机非线性回归，核函数为 RBF，惩罚系数 C=10，松弛变量 ξ=0.1。ANN 为多层感知器模型，设置 3 层网络，迭代次数为 600。CART 回归树取 50 次实验结果平均值。

通过单个模型仿真结果看出，实验采用的单个模型的预测效果普遍较低，训练误差和测试误差都比较大。几种模型相比而言，CART 回归树的预测误差较低，比较适合本文数据集建模。通过梯度提升算法的引入，建立多棵回归树，GBDT 极大地提升了单棵 CART 回归树的预测效果，预测误差降低近一半。



a 验证 MAE 随树数量的变化　　b 测试 MAE 随树数量的变化

图 1　GBDT 参数优化结果

为了进一步提高 GBDT 的预测性能，采用 LPP 和正交化 LPP 算法对原始数据进行降维处理，为了验证 LPP 方法降维的有效性，采用 PCA 算法作为参考，仿真结果见表 4。

表 3　基于原始数据模型预测性能对比

模型类型	模型名称	训练样本		测试样本	
		MAE/(μm·a^{-1})	*MAP*/%	*MAE*/(μm·a^{-1})	*MAPE*/%
单个模型	多元线性回归	10.51	32.96	8.84	32.27
	SVR	9.51	29.83	9.60	35.03
	ANN	5.31	16.68	10.42	38.04
	CART	3.96	12.44	5.87	21.43
集成模型	GBDT	1.62	5.09	3.47	12.69

表 4　基于降维数据模型预测性能对比

模型名称	训练集		测试集	
	MAE/(μm·a^{-1})	*MAPE*/%	*MAE*/(μm·a^{-1})	*MAPE*/%
GBDT	1.62	5.09	3.47	12.69
PCA-GBDT	1.94	6.09	6.24	22.77
LPP-GBDT	1.54	4.84	1.72	6.30
正交化 LPP-GBDT	1.23	3.87	1.21	4.42

相比于基于原始数据建立的 GBDT 模型，PCA-GBDT 的训练误差变化不大，但测试误差几乎提高了 1 倍，模型的泛化能力大大降低。可见，PCA 的线性降维过程破坏了金属化学成分之间的复杂非线性关系。采用 LPP 算法降维 GBDT 模型的训练、测试误差都降低，拟合和泛化能力明显提升，预测性能明显改善。其中正交化 LPP-GBDT 取得了最低的测试误差，比原始数据 GBDT 提高近 8%，验证了 LPP 方法构造的简约化特征具有更高的回归预测能力，同时也验证了正交化处理在提高局部能力和增加数据区分度方面的优势。

3.5　腐蚀速率影响因素重要性分析

GBDT 是解释性比较好的模型，对原始数据集建模预测时，通过 2.2 中所述方法对模型进行分析，各影响因素重要性结果见表 5，特征重要性合计为 1，平均值为 0.0667。

表 5　影响因素重要性排序

时间	硫	碳	磷	铜	钼	钒	锰	钛	硅	铬	铌	镭	铝	镍
0.431	0.173	0.097	0.073	0.07	0.052	0.043	0.03	0.015	0.006	0.004	0.004	0.002	0	0

可以看出，暴露时间是影响大气腐蚀速率的主导因素，在金属化学成分中，硫、碳、磷、铜、钼、钒、锰的影响比较大，其中硫、磷、铜的重要性符合相关文献的描述[14]，硫、碳降低金属的耐腐蚀性，磷、铜、锰增强金属的耐腐蚀性。青岛站点为典型的海洋大气环境，大气中海盐粒子较多，钼有利于防止氯离子的存在所产生的点蚀倾向，钒具有耐酸、耐盐的特性，因此钼和钒具有较高的特征重要性。硅通常被认为具有增强耐腐蚀性的作用，能促进耐腐蚀的稀土元素的富集[15]，但实验结果却没有印证这一结论，原因可能是本文样本中含稀土元素的金属极少，或硅在湿热的大气环境下的作用更为明显[14]。此外，由于样本金属材料的种类限制，一些合金元素对腐蚀速率的影响不是很明显，需要扩充样本种类作进一步研究。

4　结论

1）针对高维、非线性和小样本数据集，通过与其他典型方法的比较，GBDT 取得了较好的预测效果，并分析了众多因素对于腐蚀速率的影响程度，为特定环境下金属材料的合金元素的调整提供一定的参考。

2）LPP 及其正交化改进方法能有效处理高维非线性数据，线性重构简约化特征。实验结果表明，LPP 算法的引入进一步提升了 GBDT 的预测性能。

3）文中建立的 LPP-GBDT 模型不仅适用于青岛腐蚀站点腐蚀数据，还可推广到其他大气环境下的腐蚀速率预测。

参考文献：

[1] LI X, ZHANG D, LIU Z, et al. Materials Science: Share Corrosion Data[J]. Nature, 2015, 527(7579): 441.

[2] 高蒙, 孙志华, 刘明, 等. 7B04 铝合金在 NaCl 沉积与 SO_2 环境下的大气腐蚀行为[J]. 环境技术, 2016, 34(5): 9-13.

[3] JOLLIFFE I T, CADIMA J. Principal Component Analysis: A Review and Recent Developments[J]. Philosophical Transactions, 2016, 374(2065): 20150202.

[4] IZENMAN A J. Introduction to Manifold Learning[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2012, 4(5): 439-446.

[5] WANG R, NIE F, HONG R, et al. Fast and Orthogonal Locality Preserving Projections for Dimensionality Reduction[J]. IEEE Transactions on Image Processing, 2017, PP(99): 1.

[6] HE X, NIYOGI P. Locality Preserving Projections[J]. Advances in Neural Information Processing Systems, 2004, 16(1): 186-197.

[7] 黄海军, 李婵, 王俊. 典型大气腐蚀介质的灰色预测模型分析[J]. 装备环境工程, 2012, 9(1): 13-16.

[8] 邓志安, 李姝仪, 李晓坤, 等. 基于模糊神经网络的海洋管线腐蚀速率预测新方法[J]. 中国腐蚀与防护学报,

2015, 35(6): 571-576.

[9] BRIAN R. Tree: Classification and Regression Trees[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2016(1): 14-23.

[10] CHOU J S, NGO N T, CHONG W K. The Use of Artificial Intelligence Combiners for Modeling Steel Pitting Risk and Corrosion Rate[J]. Engineering Applications of Artificial Intelligence, 2016, 65: 471-483.

[11] FRIEDMAN J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[12] ZHANG L, QIAO L, CHEN S. Graph-optimized Locality Preserving Projections[J]. Pattern Recognition, 2010, 43(6): 1993-2002.

[13] YANG S, WU J, DU Y, et al. Ensemble Learning for Short-term Traffic Prediction Based on Gradient Boosting Machine[J]. Journal of Sensors, 2017(4): 1-15.

[14] 梁彩凤, 侯文泰. 钢的大气腐蚀预测[J]. 中国腐蚀与防护学报, 2006, 26(3): 129-135.

[15] 陶鹏, 孙金全, 董彩常, 等. 海洋大气环境中含稀土耐候钢暴露 1 年的耐蚀性能研究[J]. 装备环境工程, 2017, 14(5): 21-24.