文章编号:1671-4598(2022)08-0116-06 DOI:10.16526/j.cnki.11-4762/tp.2022.08.019 中图分类号:TP181,E920.8 文献标识码:A

基于集成学习的装备小样本试验缺失数据插补方法研究

马 亮,郭力强,刘丙杰,杨 静
(海军潜艇学院,山东 青岛 266199)

摘要:针对装备试验数据量有限和装备测试数据易缺失的现状,提出了一种基于集成学习的回归插补方法;以随机森林和XGBoost算法为回归器,通过设定快速填充基准和特征重要性评估策略的方法,改进数据子集重建和训练集与测试集的迭代划分策略,使用Optuna框架实现回归器超参数的自动优化,在某型导弹发射试验上进行实例验证;结果表明,使用集成学习算法的回归插补效果明显优于传统的统计量插补法以及KNN和BP神经网络,在不同缺失比例下的回归确定系数结果均保持在0.95以上,能有效解决装备小样本试验数据缺失的问题,并利用KEEL公测数据集验证了该方法的推广价值和通用性。

关键词:小样本试验;集成学习;随机森林;XGBoost;数据插补

Research About Interpolating Missing Data on Small Sample Trials of Equipment Based on Ensemble Learning

MA Liang, GUO Liqiang, LIU Bingjie, YANG Jing
(Navy Submarine Academy, Qingdao 266199, China)

Abstract: For the current situations that the amount of equipment test data is limited and the equipment test data is prone to missing, a regression interpolation method based on ensemble learning algorithm is proposed. The algorithms for Random Forests and XGBoost are used as the regressor for interpolating the missing data by setting fast filling benchmarks and feature importance assessment strategies, the data subset reconstruction and iterative partitioning strategies for the training and test sets are improved, and the hyperparameters of regressor by using the Optuna framework are automatically optimized. Based on this method, a type of missile launch trial is used for validating. The results show that the regression interpolation effect of the ensembel learning algorithm is significantly better than that of the traditional statistical interpolation method as well as KNN and BP neural networks. And the regression determination coefficient under the different missing proportions is maintained above 0.95, which can effectively solve the missing data problem of small sample tests of equipment. In addition, the promotion and universality of this method are validated by using the KEEL public test dataset.

Keywords: small sample trials of equipment; ensemble learning; random forests; XGBoost; data interpolation

0 引言

大型装备具有高复杂性、高耦合度、非线性等特点,装备性能影响因素多,关系复杂。由于大型装备试验次数少,试验数据量有限,易导致测试数据易缺失,覆盖性不高。为全面了解装备性能影响因素,有必要对缺失数据进行插补。

目前,常用的数据插补方法有统计量插补法和回归预测插补法。随着机器学习和数据挖掘技术的兴起,回归预测插补法引起了广泛的关注。文献[1]针对统计量插补法效果不理想的问题,使用BP神经网络模型对研究流域降水数据进行插补,取得了较好的插补精度。但模型泛化能力弱,对样本数量要求高。文献[2]提出了联合极大似然估计与EM算法相结合的多重插补方法,但是仅在缺失数据比例小于35%时得到理想结果,存在适用范围受限的问题。文献[3]提出了一种基于均值插补法的EM算法,并与基于线性回归和Bootstrap的修正EM算法进行了比较,结论认为基于Bootstrap的修正EM算法更准确。然而,EM算法在原理上需要有庞大的数据集作为支撑,以保证估计值渐近无偏并服从正态分布,否则可能会陷入局部极值。文献[4]提出了基于K近邻的插补方法,可应用于连续型数

收稿日期:2021-11-16; **修回日期**:2022-03-15。
基金项目:国防科技创新特区项目(20-163-05-***)
作者简介:马 亮(1973-),女,吉林镇赉人,博士,教授,主要从事水下发射技术方向的研究。
刘丙杰(1979-),男,山西曲沃人,博士,副教授,主要从事故障诊断与可靠性工程方向的研究。
通讯作者:郭力强(1992-),男,山东青岛人,硕士研究生,主要从事武器装备智能决策技术方向的研究。
引用格式:马 亮,郭力强,刘丙杰,等.基于集成学习的装备小样本试验缺失数据插补方法研究[J].计算机测量与控制,2022,30(8):116-121.

据的插补。随机森林的创始人 Breiman 在文献［5］首次使用模型对缺失数据插补，验证了随机森林对缺失数据插补的可行性。文献［6］分别使用极限学习机、BP 神经网络、支持向量机和 XGBoost 等机器学习模型对潜热通量缺失数据进行插补，结果表明 XGBoost 与极限学习机的插补效果最佳。

国内外对数据插补方法的研究表明，数据插补问题的本质是一个高噪声的多重回归预测问题。由于大型复杂装备试验具有小样本的特点，这就要求回归器在小样本上具有较好的拟合效果和泛化能力。相比于传统的决策树（decision tree）[7]，支持向量机（support vector machine）[8]等对样本数量要求高、容易过拟合的单一模型，集成学习算法会综合多个基学习器的建模结果，以此获取比单个模型更好的预测效果，从理论上适合应用于解决装备小样本试验缺失数据插补问题。

本文以基于 Bagging 算法的随机森林和 Boosting 算法的 XGBoost 为代表，分析集成学习的建模原理，探讨应用集成学习算法解决装备小样本试验数据缺失问题的有效性，并在 KEEL 公测数据集上对该方法进行了推广验证。

1　理论分析

集成学习算法（ensemble learning）是通过在数据上构建多个模型，综合考虑所有模型的建模结果来提高预测性能的算法。由多个模型集成的模型叫做集成学习器，组成集成学习器的每个模型叫做基学习器，并要求其性能至少要超过随机学习器[9]。由于集成学习算法在处理预测和分类问题中的出色性能，近年来成为 KDDcup、Kaggle、天池等大型数据竞赛中的夺冠利器。

1.1　随机森林

随机森林是一种以决策树模型为核心的 Bagging 算法，用于完善单决策树在处理回归和分类问题中精度不高、容易过拟合的缺陷。2001 年 Leo Breiman 将其提出的 Bagging 理论与 CART 决策树，以及随机子空间方法（random subspace method）相结合，提出了一种非参数分类与回归算法—随机森林（random forest）[10]。

随机森林的基本思想可归纳为：首先使用自助重抽样（Bootstrape）的方法从数据量为 N 的原始样本集 D 中有放回地随机抽取生成 m 个与原始样本集 D 同样大小的训练样本集 D_m，在此基础上构建对应的决策树；然后在对决策树的每一个节点进行分支时，从全部 Z 个特征中随机抽取一个特征子集，即只选择部分特征进行分支，并从这些子集中选择一个最优分支来建立决策树；最后将构建的多棵决策树作为集成学习器形成随机森林，通过对每个基学习器的结果进行评估或者多数表决的方法决定集成学习器的结果[11]。

Bagging 算法的核心思想是通过并行构建多个尽可能相互独立的基学习器，借助基学习器之间的“独立性”来降低模型整体的方差，从而获得比单颗决策树更低的泛化误差和更强的泛化能力。泛化误差可理解为模型在未知数据集上预测误差，一般可认为由偏差、方差和噪声构成。其中偏差是预测值与真实值之间的差异，用于衡量模型的精度；方差是模型在不同数据集上输出的结果的方差，用于衡量模型稳定性；噪音是数据收集过程当中不可避免的、与数据真实分布无关的信息。

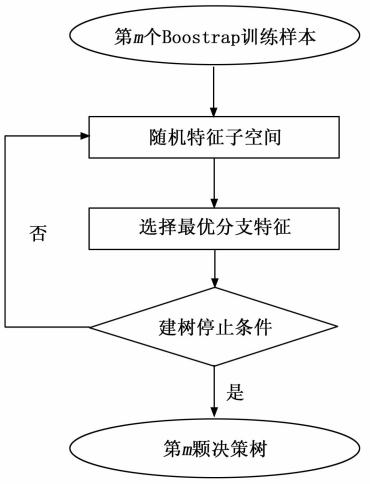

图 1　随机森林构建单颗决策树的流程图

在回归类问题中，假设随机森林模型构建了 n 个相互独立的基学习器，任意决策树上的输出结果为 X_i，全部决策树预测结果的方差可表示为 $Var(X_i)$，随机森林的输出结果即为 $\overline{X}=\sum_{i=1}^{n}X_i/n$，预测结果的方差可表示为 $Var(\overline{X})$。若任意决策树预测结果的方差 $Var(X_i)=\sigma^2$，则有 $Var(\overline{X})=\sigma^2/n$。当 n 为正整数，基学习器之间相互独立时，必然有 $Var(\overline{X})$ 小于 $Var(X_i)$，这是随机森林的泛化能力总是强于单一决策树的根本原因。

然而，由于所有基学习器都是在相同的原始样本集上进行采样训练，很难实现完全相互独立。假设任意决策树之间的相关系数为 ρ，可将随机森林输出结果的方差表示为：

$$Var(\overline{X})=\sigma^2/n+(n-1/n)*\rho*\sigma^2$$

即单颗决策树之间的相关性越弱，随机森林通过降低方差提升的泛化能力越强。因此，为确保 Bagging 算法的有效性，一方面单个基学习器的误差率至少要保证小于 50%，另一方面要尽可能降低基学习器之间的相关性。

由于随机森林引入样本自助重抽样和随机分支策略，通过样本有放回抽样和特征无放回抽样样本的方式，从样本和特征两个维度出发，尽可能使得每颗决策树的训练相互独立。同时可以使用袋外数据（out of bag data）测试模型，有效地提升了效率；在决策树的每个节点只选择部分特征进行分支，从而使树的增长只依赖于该部分的特征，而不是全部特征，在处理高维数据上有着出色性能[12]；最终的输出结果由多颗决策树的输出结果平均或投票决定，因此对数据噪声和异常值有了较好的容忍性。

1.2　XGBoost 算法

相比于与 Bagging 算法，Boosting 的核心思想是按顺序

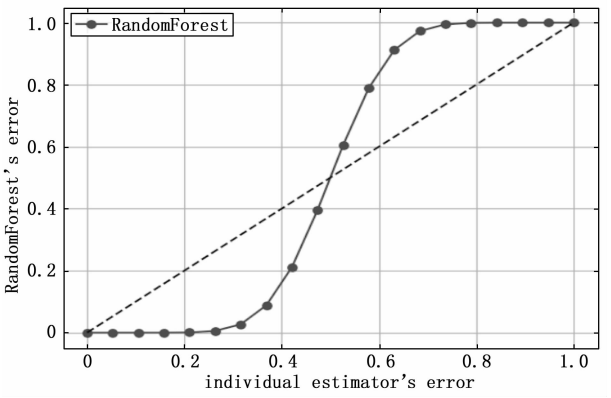

图 2 基学习器误差率变化影响示意图

依次构建相互关联和影响的基学习器，通过某种集成结果输出规则来多次提升弱学习器对样本的预测结果，达到降低模型的整体偏差的目标，构成性能出色的强学习器。假设衡量模型预测结果与真实结果的差异的损失函数为 $L(x,y)$，基学习器为 $f(x)$，算法集成输出结果为 $H(x)$，则 Boosting 算法的建模流程可表示为：依据上一个弱学习器 $f(x)_{t-1}$ 的结果计算损失函数 $L(x,y)$，并使用 $L(x,y)$ 结果自适应影响下一个弱学习器 $f(x)_t$ 的构建。经 T 次迭代后，根据某种集成输出规则计算所有弱学习器的 $f(x)_0 \sim f(x)_T$ 的预测结果得到 $H(x)$。

Boosting 算法的开山代表是 AdaBoost[13]（adaptive boosting）算法，它的基本思想是：首先在抽取的训练样本上建立一棵决策树，根据该决策树预测的结果 $f(x)_t$ 和损失函数值 $L(x,y)$，增加被预测错误的样本 x_i 在数据集中的样本权重 w_i，并让加权后的数据集被用于训练下一棵决策树，即通过调整训练数据的分布来间接影响后续弱学习器的构建。

2001 年 Jerome Friedman 提出一种全新的损失函数计算公式—弗里德曼均方误差[14]：

$$impurity\_decrease = \frac{w_l w_r}{w_l + w_r} * \left(\frac{\sum_l (r_i - y_i)^2}{w_l} - \frac{\sum_r (r_i - y_i)^2}{w_r} \right)^2$$

w 是叶子节点上的样本权重，$r_i = y - H(x_i)$ 是样本 x_i 上的残差。即通过这种拟合残差的方式直接影响后续弱学习器的构建结构，并且采用调和左右叶子节点权重的分支规则加速了 CART 决策树的预测效率，在 AdaBoost 的基础上，形成 GBDT（gradient boosting decision tree）算法。

2015 年 T. Q. Chen 提出 XGBoost[15]（eXtreme gradient boosting）算法，在 GDBT 的基础上改进集成输出结果的评估策略，通过对损失函数进行二阶泰勒展开和添加正则项的方法，有效避免了过拟合问题并加快了模型收敛速度。其输出结果的评估策略可表示为式（1）：

$$\hat{y}_i = \sum_{c=1}^{C} f_c(x_i), f_c \in F_{\text{CART}} \tag{1}$$

$\hat{y}_i$ 表示模型的预测值；C 表示 CART 树的数量；f_c 表示第 c 个子树；x_i 表示第 i 个输入样本；F_{CART} 表示所有 CART 树集合。XGBoost 的目标函数由损失函数和正则项两个部分组成：

$$Obj^{(t)} = \sum_{i=1}^{m} (y_i, \hat{y}_i^{(t-1)}) + f_t(x_i)) + \Omega(f_c) \tag{2}$$

$$\Omega(f_c) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{3}$$

$Obj^{(t)}$ 表示第 t 次迭代的目标函数；$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 次迭代的预测值；$\Omega(f_c)$ 表示第 t 次迭代树模型的正则项；γ 和 λ 表示正则项系数；T 表示该模型的叶子节点个数。对式（2）中的目标函数使用泰勒公式展开得：

$$Obj^{(t)} \cong \sum_{i=1}^{m} \left[f_t(x_i) g_i + \frac{1}{2} (f_t(x_i))^2 h_i \right] + \gamma T + \frac{1}{2}\lambda$$

$$Obj^{(t)} \cong \sum_{j=1}^{T} \left[w_j \sum_{i \in I_j} g_i + \frac{1}{2} w_j^2 \left(\sum_{i \in I_j} h_i + \lambda \right) \right] + \gamma T \tag{4}$$

其中：g_i 表示样本 x_i 的一阶导数；h_i 表示样本 x_i 的二阶导数；w_j 表示第 j 个叶子节点的输出值，I_j 表示第 j 个叶子节点含有样本的子集。简化式（4）中的目标函数，可定义：

$$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \tag{5}$$

观察可得，目标函数是一个凸函数：

$$Obj^{(t)} \cong \sum_{j=1}^{T} \left[w_j G_j + \frac{1}{2} w_j^2 (H_j + \lambda) \right] + \gamma T \tag{6}$$

在式（6）中对 w_j 求导，令一阶导数等于 0，可求得使目标函数达到最小值的 w_j，即：

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

$$Obj_{\min}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{7}$$

式（8）可以用来评价树模型的得分，其数值越小，树模型的得分越高。由此可以得出用于树模型进行分枝的得分公式：

$$G_L = \sum_{i \in I_L} g_i, G_R = \sum_{i \in I_R} g_i, H_L = \sum_{i \in I_L} h_i, H_R = \sum_{i \in I_R} h_i$$

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{8}$$

2 数据插补方法

2.1 统计量插补法

统计量插补法的基本思想是通过计算缺失特征的某一统计量，并用这个值来插补该特征的所有缺失数据。常用的统计量有平均值、中位数值和众数值。该方法思想简单，应用最为广泛，可根据缺失数据类型分别处理：通常情况下，若缺失值为连续数据，则使用对该特征未缺失部分样本平均值或中位数值插补的方法；若缺失值为离散数据，则以该特征未缺失部分样本出现频率最多的数值进行插补。

2.2 回归预测插补法

回归预测插补法的基本思想是通过构建回归模型，从数据中挖掘特征矩阵与标签之间的关联规则，训练能够映射特征矩阵与标签关系的回归器，实现对缺失数据的预测，并将预测结果作为插补结果。由于数据集的标签和特征是

人为标注的，因此可以通过自定义特征矩阵和标签，重新构建数据子集的方式进行回归预测。

2.2.1　单特征值缺失情况

在样本数量为 n，特征数量为 m 的数据集上，假设某一特征 x_i 存在缺失值，将存在缺失值的特征 x_i 定义为数据集的标签 y，其余 $m-1$ 个未存在缺失值的特征和原始标签定义为数据集的特征矩阵 $\boldsymbol{X}$。那么，不含缺失值的样本中既存在特征值也存在标签值，将其作为用于构建回归模型训练集（training set）；包含缺失值的样本中只存在特征值没有标签值，将其作为用于检测回归模型性能的测试集（testing set），其中测试集中缺失的标签值即为回归预测的对象。通过这种方式完成数据子集的重建以及训练集和测试集划分。

2.2.2　多特征值缺失问题

当数据集的多个特征值存在缺失时，可采用如下方法实现数据子集的重建和迭代划分：

Step1：根据缺失数据集的特点，设定快速填充基准和特征重要性评估策略。快速填充基准可使用平均值、中位值等统计量，以快速对缺失值较少的特征进行插补；特征重要性评估策略可根据特征缺失程度大小或特征对标签的贡献度，以此确定对多个缺失特征的插补顺序 U_{index}。一般来说，当某一特征的缺失值越少，损失的信息也就越少，使用统计量快速插补的效果越好；反之特征的缺失值越多，损失的信息也就越多，使用统计量填充的效果越差，应当优先选中进行回归预测。

Step2：根据多个缺失特征的插补排序 $U_{\text{index}}=[U_1,U_2,U_3,\ldots,U_t(t\leqslant m)]$，使用快速填充基准对 U_1 外的 $m-1$ 特征的缺失值进行填充，将多特征值缺失问题转换成单特征值缺失问题，按照单特征值缺失的情况进行数据子集的重建以及训练集和测试集划分。

Step3：选择数据预处理方式和回归结果评估指标，采用 5 折交叉验证的方法优化回归模型超参数后，训练回归器得到 U_1 缺失值的预测结果，并将本轮的数据插补结果返回至缺失数据集相应位置，作为下一次迭代回归的输入。

Step4：按照 U_{index} 排序重复 Step2～3 过程，经 t 次迭代回归预测后，最终得到整个缺失数据集的插补结果，再根据回归结果评估指标对快速填充基准、特征重要性评估策略及回归器进行改进和提升。

3　实验与分析

潜射导弹因其机动范围广、攻击突然性强，装弹数量多、反击威力大等特点备受各大军事强国重视。潜射导弹弹射出筒经过水介质出水至空中再进入预定弹道，水下飞行段和出水段是潜射导弹发射所独有的过程，对发射结果影响巨大。

出筒速度是影响水下弹道姿态和发射精度的关键因素之一[16-17]。受众多客观条件影响，发射深度、海流速度、海浪高度和气幕弹等都会对出筒速度产生影响[18-20]。为确保导弹安全可靠的发射，需根据试验数据的有关信息对导弹出筒速度的影响因素进行分析。但是受水下发射特殊环境的影响，观测数据易出现缺失，因此有必要对试验缺失数据进行有效插补。

3.1　实验设计

实验在 Windows 10 环境下进行，使用 JupyterLab 3.10 IDE 和 Python 3.9 Kernel，调用 Scikit-Learn 库版本为 1.01。CPU 配置为 AMD Ryzen 5-5600H，主频 3.30 GHz，内存 16 GB。

以某型潜射导弹发射试验样本为原始数据集，以发射深度、潜艇航速、海流速度、海面波高作为原始特征，出筒速度为原始标签数据，采用完全随机缺失方式，缺失数据标记为 NaN 值，建立缺失数据集。

初始采用平均值作为快速填充基准，将数据缺失程度作为特征重要性评估策略，使用 Z-Score 标准化的方法对特征数据进行预处理。为反映插补数据对真实数据的整体拟合程度，以回归确定系数 R^2（$R^2=SSR/SST$）作为插补效果评估指标，用于检验和提升回归器插补效果。由于回归器的泛化性能受超参数的影响较大，模型的复杂程度越高，超参数的数量越多。因此，为提高数据插补的效率，使用 Optuna 框架[21] 采用 TPE（tree-structured parzen estimator approach）[22]贝叶斯过程对回归器超参数实现自动优化。数据集的重建划分和回归器训练过程部分代码如下所示：

```
    X_full 实验数据集特征矩阵 y_full 实验数据集标签
    X_missing 缺失数据集特征矩阵 y_full 缺失数据集标签
    X 重建数据集特征矩阵 y 重建数据集标签
from sklearn.impute import SimpleImputer as SI
from sklearn.preprocessing import StandardScaler
import optuna
    按照特征缺失程度进行插补排序
sortindex = np.argsort(X_missing.isnull().sum(axis=0)).values
for i in sortindex:
    构建新特征矩阵
X = X_missing
X = pd.concat([X.iloc[:,X.columns!=i],y_full],axis=1)
    使用统计量对含有缺失值的特征进行快速填充
X = SI(missing_values=np.nan,strategy='mean').fit_transform(X)
    构造新标签
y = X_missing.iloc[:,i]
    新标签真实数据
y_full = X_full.iloc[:,i]
    数据预处理
scaler = StandardScaler()
X = scaler.fit_transform(X)
scaler = StandardScaler()
X = scaler.fit_transform(X)
    划分训练集测试集
Y_train = y_new[y_new.notnull()]
Y_test = y_full[y_new.isnull()]
```

```
X_train = X_new_mean[Y_train.index,:]
X_test = X_new_mean[Y_test.index,:]
  自动优化回归器超参数
best_params,best_score = optimizer_optuna(300,"TPE")
  实例化回归模型
Reg = Regressor(best_params)
  训练回归器
Reg = Reg.fit(X_train,Y_train)
  预测缺失数据
Y_predict = Reg.predict(X_test)
  将预测结果返回至缺失数据集
X_missing.loc[X_missing.iloc[:,i].isnull(),i] = Y_predict
  评估数据插补效果
loss = criterion (X_full,X_missing)
```

3.2 结果分析

初始设置特征矩阵缺失比例为 20%，同时使用统计量法和回归预测法对缺失数据进行插补。为评估集成学习算法的插补性能，分别使用基于 Bagging 思想的 Random Forests 和 Boosting 思想的 XGBoost 两种典型的算法作为回归器，并使用 KNN 和 MLP 算法与之比较。统计量法和回归预测法对缺失数据的插补结果如图 3 所示。

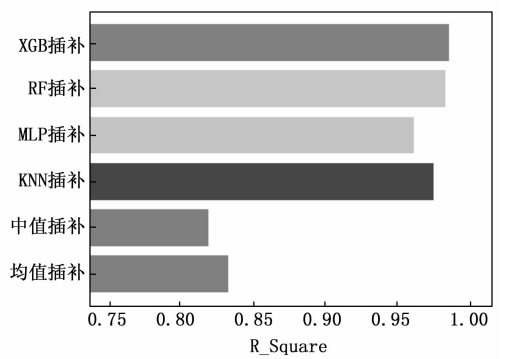

图 3 20%缺失比例下插补效果评估图

在特征矩阵缺失比例为 20%的情况下，使用平均值和中位数值插补得到的 R^2 结果分别为 0.833 和 0.820；使用 KNN、MLP、Random Forest 和 XGBoost 回归器插补得到的 R^2 结果分别为 0.975、0.961、0.983、0.980。可以看出回归插补法的效果明显优于统计量插补法，其中使用集成算法构件的回归器性能最优。

为进一步检验回归预测插补法的适用范围，将特征矩阵缺失比例调整为 10%、15%、20%、25%、30%，分别建立缺失数据集进行插补实验，得到如表 1 和图 4 所示的结果。

表 1 不同缺失比例下的插补效果

缺失比例/%	均值插补	KNN 插补	MLP 插补	RF 插补	XGB 插补
10	0.901	0.972	0.973	0.984	0.983
15	0.872	0.983	0.981	0.993	0.993
20	0.833	0.975	0.961	0.983	0.980
25	0.803	0.965	0.950	0.973	0.976
30	0.756	0.939	0.924	0.954	0.951

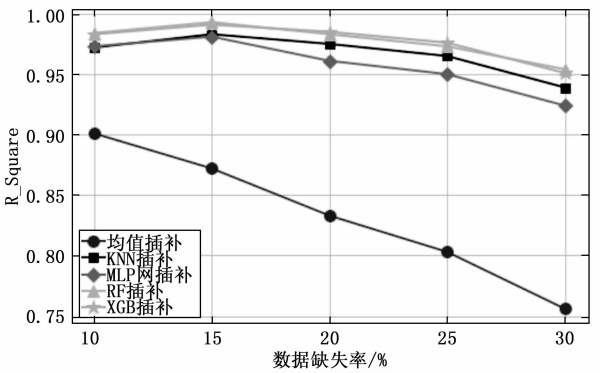

图 4 不同缺失比例下的插补效果

由图 4 可知，回归预测插补法具有较好的适用性，在不同缺失比例下，均能保持较好的数据插补效果。随着数据缺失比例的增大，统计量插补法的效果越差，回归插补法的优势体现的越明显，其中，使用集成算法的回归器在 25%、25%和 30%等高缺失比例下，R^2 结果均能保持在 0.95 以上，验证了该方法的可靠性，可应用于装备小样本试验缺失数据插补。

3.3 方法推广

除了装备小样本试验，插补缺失数据是数据分析和处理工作中必不可少的一环，有必要验证该方法的通用性和推广价值。因此，使用 KEEL 公测数据集“dee”和“wizmir”进行上述数据插补实验。其中“dee”数据集的样本数量为 365，特征数量为 6；“wizmir”数据集的样本数量为 1461，特征数量为 6。得到的插补结果如表 2 所示。

表 2 KEEL 公测数据集下的插补效果

插补方法	dee dateset		wizmir dateset	
	10%	20%	10%	20%
均值插补	0.903	0.813	0.891	0.801
KNN 插补	0.961	0.911	0.941	0.895
MLP 插补	0.937	0.850	0.949	0.772
RF 插补	0.969	0.923	0.958	0.922
XGB 插补	0.966	0.933	0.959	0.931

由表 2 可知，该方法在公测数据集上具有通用性，在 10%和 20%缺失比例下，R^2 结果仍能保持在 0.92 以上。但是回归预测插补法在回归器的构建、超参数优化以及模型性能评估上需要投入大量的时间成本和计算代价。综合衡量，统计量插补法适合在样本数量大、缺失特征数目少、整体数据缺失比例较低（<10%）的数据集；而回归预测插补法更适合用于缺失比例较高、特征相关性显著的数据集。

4 结束语

针对装备小样本试验次数少，测试数据易缺失的特点和常用数据插补方法对数据量要求高、计算复杂的问题，本文提出了一种基于集成学习的数据插补方法。验证结果表明，该方法明显优于统计量插补法，模型的数据拟合能

力和泛化能力较强，在不同缺失比例下均有较好表现，能偶有效解决装备小样本试验数据缺失问题，同时具有一定的推广价值和通用性。为进一步提高缺失数据的插补效果，下一步将结合时序数据的预测处理方法，进一步优化缺失数据插补效果。

参考文献：

[1] 田　琳，王　龙，余　航，等. 基于 BP 神经网络的缺测降水数据插补 [J]. 云南农业大学学报，2012，27 (2)：281－284.

[2] 游晓锋，丁树良，刘红云. 缺失数据的估计方法及应用 [J]. 江西师范大学学报，2011，35 (3)：328－330.

[3] 荆文君，张晓琴，常王华. 一种基于成分数据的修正 EM 算法 [J]. 中北大学学报，2013，34 (5)：485－487，499.

[4] TROYANSKAYA O，CANTOR M，SHERLOCK G，et al. Missing value estimation methods for DNA microarrays [J]. Bioinformatics，2001，17 (6)：520－525.

[5] BREIMAN L. Manual-setting up，using，and understanding random forests V4.0 [EB/OL]. [2003－02－13]. https://www.stat.berkeley.edu/～breiman/Using_random_forests_v4.0.pdf.

[6] 司　超. 基于极限学习机的潜热通量插补研究和应用 [D]. 北京：北京林业大学，2019.

[7] QUINLAN J R. Induction of decision trees [J]. Machine Learning，1986，1 (1)：81－106.

[8] CORTES C，VAPNIK V. Support vector machine [J]. Machine Learning，1995，20 (3)：273－297.

[9] 周志华. 机器学习 [M]. 北京：清华大学出版社，2016.

[10] BREIMAN L. Random forests [J]. Machine Learning，2001，45 (1)：5－32.

[11] 王奕森，夏树涛. 集成学习之随机森林算法综述 [J]. 信息通信技术，2018，12 (1)：49－55.

[12] 陈慧佳. 基于 Random Forest 的缺失数据补全策略研究 [D]. 南昌：南昌大学，2016.

[13] FREUND Y，SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences，1997，55 (1)：119－139.

[14] FRIEDMAN J H. Greedy function approximation：a gradient boosting machine [J]. The Annals of Statistics，2001，29 (5)：1189－1232.

[15] CHEN T Q，GUESTRIN C. Xgboost：A scalable tree boosting system [C] //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，ACM，2016：785－794.

[16] 刘丙杰，罗珩娟，李　发. 潜射导弹出筒速度影响因素相关性分析 [J]. 兵工自动化，2020，39 (7)：63－64，74.

[17] 罗珩娟，刘丙杰，张　庆. 基于数据驱动的潜射导弹出筒速度模型辨识 [J]. 航天控制，2020，38 (6)：55－63，73.

[18] 佟力永，肖　凡，张　涛. 海基远程弹道导弹精度分析样本生成方法研究 [J]. 航天控制，2016，34 (2)：15－18.

[19] 王亚东，袁绪龙，张宇文. 波浪对导弹垂直发射水弹道影响研究 [J]. 兵工学报，2012，33 (5)：630－635.

[20] 尚书聪，孙建中，程　栋，等. 筒口气幕环境的导弹出筒过程受力影响 [J]. 哈尔滨工程大学学报，2012，33 (11)：1423－1434.

[21] AKIBA T，SANO S，YANASE T，et al. Optuna：A next-generation hyperparameter optimization framework [C] //The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining，ACM，2019：2623－2631.

[22] OZAKI Y，TANIGAKI Y，et al. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems [C] //GECCO'20：Proceedings of the 2020 Genetic and Evolutionary Computation Conference，2020：533－541.

(上接第 115 页)

[16] 曾小华，陈虹旭，崔　臣，等. 考虑铁损的永磁同步电机无位置传感器控制算法 [J]. 东北大学学报 (自然科学版)，2021，42 (1)：102－110.

[17] 王　祯，尹项根，陈　玉，等. 基于连续控制集模型预测控制的 MMC 桥臂电流控制策略 [J]. 电力系统自动化，2020，44 (10)：85－91.

[18] 郑春菊，孟　鑫，周　群，等. 三相多驱动系统带移相电流控制的谐波消除方法 [J]. 电力系统保护与控制，2021，49 (12)：114－123.

[19] 刘桓龙，李　顺，谢迟新. 基于液压泵/马达逆向驱动的电机启动电流控制方法 [J]. 西南交通大学学报，2021，56 (4)：720－729.

[20] 陈建福，谭　喆，刘仁亮，等. 基于改进鲁棒重复控制与 QPR 的光伏电流控制策略 [J]. 电力科学与技术学报，2021，36 (3)：100－110.

[21] 李添幸，马瑞卿，白　浩，等. 五相永磁同步电机三次谐波电流控制比较研究 [J]. 西北工业大学学报，2021，39 (4)：865－875.

[22] 齐洪峰，王轶欧，闫一凡. 无速度传感器永磁同步电机预测电流控制策略 [J]. 北京交通大学学报，2020，44 (2)：119－128.

[23] 涂　震，赵　阳，余佳佳，等. 基于双矢量的永磁同步电机模型预测电流控制 [J]. 武汉大学学报 (工学版)，2020，53 (8)：721－727.

[24] 周新秀，周咏平，张　旨，等. 基于参数辨识的内置式永磁同步电机最大转矩电流比电流预测控制 [J]. 光学精密工程，2020，28 (5)：1083－1093.

[25] 孟　柳，章回炫，范　涛. 永磁同步电机静止参数辨识及电流环控制器自动参数整定 [J]. 兵工学报，2021，42 (10)：2114－2122.

[26] 罗　丹，廖志贤，黄国现，等. 单相光伏微网逆变器的建模与电流跟踪数值模拟 [J]. 计算机测量与控制，2019，27 (7)：180－183，189.

[27] 刘宪爽，肖文波，吴华明，等. 基于 LabView 的光纤电流互感器一体化测试系统设计 [J]. 计算机测量与控制，2017，25 (7)：39－42.