文章编号: 1671-251X(2023)01-0028-09 DOI: 10.13272/j.issn.1671-251x.2022090032

基于数据填补的煤自燃温度预测模型

翟小伟[1,2,3], 罗金雷[1,2,3], 张羽琛[1,2,3], 宋波波[1,2,3], 郝乐[1,2,3], 周妤婕[1,2,3]

(1. 西安科技大学 安全科学与工程学院, 陕西 西安 710054;
2. 西安科技大学 陕西省煤火灾害防控重点实验室, 陕西 西安 710054;
3. 陕西高校青年创新团队 矿山应急救援创新团队, 陕西 西安 710054)

摘要: 现有煤自燃温度预测模型的建立大多基于较为完整的指标气体样本数据, 但指标气体数据受仪器或人为因素影响, 往往存在数据缺失现象, 导致煤自燃温度预测准确率较低和过拟合等问题。针对上述问题, 提出了将 K 近邻算法(KNN)、随机森林(RF)、决策树(DT)及基于粒子群优化的支持向量回归等填补算法(PSO-SVR)应用于缺失值填补, 缺失数据和填补后的数据通过 RF、SVR 和极限梯度提升树(XGBoost)算法分别进行训练, 并通过 PSO 算法优化参数, 构建了基于数据填补的 RF、XGBoost 和 SVR 煤自燃温度预测模型。利用煤自然发火实验选取 CO, CO_2, CH_4, C_2H_6, O_2 作为指标气体, 并设计整体缺失率为 10%, 20%, 30% 和 CO, CO_2 缺失率为 40%, 50%, 60% 共 6 种随机数据缺失, 采用平均绝对误差百分比(MAPE)作为填补效果评价指标, 采用 MAPE、判断系数 R^2 和均方根误差(RMSE)作为模型性能评价指标, 对 4 种填补算法和 3 种预测模型进行对比。对比分析结果表明: 在 6 种数据缺失情况下, DT 填补算法填补效果优于其他 3 种算法, 在 CO, CO_2 存在较多缺失值时, RF 算法的填补值与实际值的 MAPE 偏大; 在不调参的情况下, XGBoost 模型虽然在训练集效果极好, 但极易过拟合, 而 SVR 模型预测效果极差, 无法满足预测要求; 在 6 种数据缺失情况下, 基于 DT 填补算法的 PSO-SVR、RF 与 PSO-RF 煤自燃温度预测模型的 MAPE 均在 4% 左右, 基于 DT 填补算法的 RF 模型无需优化就能较好地预测出煤自燃温度, 具有良好的稳定性。

关键词: 煤自燃; 温度预测; 指标气体; 数据缺失填补; K 近邻填补算法; 随机森林填补算法; 决策树回归填补算法; 基于粒子群优化的支持向量回归填补算法

中图分类号: TD752 文献标志码: A

Prediction model of coal spontaneous combustion temperature based on data filling

ZHAI Xiaowei[1,2,3], LUO Jinlei[1,2,3], ZHANG Yuchen[1,2,3], SONG Bobo[1,2,3], HAO Le[1,2,3], ZHOU Yujie[1,2,3]

(1. College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, China;
2. Shaanxi Province Key Laboratory of Coal Fire Disaster Prevention and Control, Xi'an University of Science and Technology, Xi'an 710054, China; 3. Mine Emergency Rescue Innovation Team, The Youth Innovation Team of Shaanxi Universities, Xi'an 710054, China)

Abstract: Most of the existing coal spontaneous combustion temperature prediction models are based on relatively complete index gas sample data. However, the index gas data are affected by instruments or human factors. There are often data missing phenomena, resulting in low accuracy and over-fitting of coal spontaneous combustion temperature prediction. In order to solve the above problems, the paper proposes to apply filling algorithms such as K-nearest neighbor algorithm (KNN), random forest algorithm (RF), decision tree algorithm

收稿日期: 2022-09-09; **修回日期**: 2023-01-05; **责任编辑**: 张强。

基金项目: 国家自然科学基金项目(51974236); 陕西省自然科学基础研究计划项目(2021JC-48); 陕西省教育厅青年创新团队建设科研计划项目(21JP078)。

作者简介: 翟小伟(1979—), 男, 陕西富平人, 教授, 博士, 博士研究生导师, 现主要从事矿山重大灾害机理及控制技术研究工作, E-mail: 1150171170@qq.com。

引用格式: 翟小伟, 罗金雷, 张羽琛, 等. 基于数据填补的煤自燃温度预测模型[J]. 工矿自动化, 2023, 49(1): 28-35, 98.
ZHAI Xiaowei, LUO Jinlei, ZHANG Yuchen, et al. Prediction model of coal spontaneous combustion temperature based on data filling[J]. Journal of Mine Automation, 2023, 49(1): 28-35, 98.

扫码移动阅读

(DT) and support vector regression algorithm based on particle swarm optimization (PSO-SVR) to fill in the missing values. The missing data and the filled data are trained by RF, SVR and extreme gradient boosting (XGBoost) algorithm respectively. The parameters are optimized by the PSO algorithm. The RF, XGBoost and SVR coal spontaneous combustion temperature prediction models based on data filling are constructed. CO, CO_2, CH_4, C_2H_6 and O_2 are selected as index gas in coal spontaneous combustion experiment, and six kinds of random data missing are designed. The overall missing rates are designed as 10%, 20% and 30%. The missing rates of CO and CO_2 are designed as 40%, 50% and 60%. The average absolute error percentage (MAPE) is used as the filling effect evaluation index. The MAPE, the judgment coefficient R^2 and the root mean square error (RMSE) are used as the model performance evaluation index. Four filling algorithms and three prediction models are compared. The results of the comparative analysis show the following points. The DT filling algorithm has better filling effect than the other three algorithms in six kinds of missing data cases. When there are more missing values of CO and CO_2, the MAPE between the filling value and the actual values of the RF algorithm is large. The XGBoost model works extremely well in the training set without adjusting the parameters, but it is very prone to overfitting. The prediction effect of SVR model is very poor and the model cannot meet the prediction requirements. In the case of six kinds of data missing, the MAPE of PSO-SVR, RF and PSO-RF coal spontaneous combustion temperature prediction models based on the DT filling algorithm are about 4%. The RF model based on the DT filling algorithm can predict the coal spontaneous combustion temperature without optimization and has good stability.

Key words: coal spontaneous combustion; temperature prediction; index gas; filling in data gaps; K-nearest neighbor filling algorithm; random forest filling algorithm; decision tree regression filling algorithm; support vector regression filling algorithm based on particle swarm optimization

0　引言

煤自燃是煤矿中一种较为常见的灾害，特别是采空区浮煤自燃，不仅会造成大量资源浪费，还使工作人员的生命安全面临严峻挑战，严重影响了煤矿的安全生产[1-3]。因此，准确预测煤自燃程度对于煤矿安全生产具有重大现实意义。

温度是煤自燃最直接的表现，但采空区内部较为隐蔽，难以接近，直接检测采空区的煤体温度较为困难，所以需要采取间接测量技术手段[4]。目前，气体分析法是最常用的煤温间接测量方法之一，可通过测量指标气体浓度，并利用指标气体浓度与煤温存在的非线性关系预测煤自燃温度[5]，很多学者对其进行了深入研究。邓军等[6]利用采空区实测数据建立了基于随机森林（Random Forest, RF）方法的煤自燃预测模型，在不同矿井进行煤自燃温度预测，均具有良好的应用效果。Deng Jun 等[7]采用模拟退火（Simulated Annealing, SA）算法对支持向量机（Support Vector Machine, SVM）的超参数进行优化，构建了 SA−SVM 煤自燃温度预测模型，并通过现场数据对模型进行了验证，结果表明 SA−SVM 模型可较为准确地预测煤自燃温度。周旭等[8]采用极限梯度提升树算法（Extreme Gradient Boosting, XGBoost）建立了煤自燃温度预测模型，并利用粒子群优化算法（Particle Swarm Optimization, PSO）对 XGBoost 模型的随机采样率和最小叶子节点样本权重进行优化，实现了煤自燃温度的准确预测。上述预测模型可较为准确地预测煤自燃温度，模型的建立大多基于较为完整的样本数据，但仪器故障或人为错误难以完全避免，样本数据会存在缺失[9]。指标气体数据的缺失会导致煤自燃温度预测模型存在准确率低、过拟合等问题。因此需对缺失数据进行处理。本文采用 K 最邻近算法（K-Nearest Neighbor, KNN）、决策树（Decision Tree, DT）、RF 和基于粒子群优化的支持向量回归（Particle Swarm Optimization-Support Vector Regression, PSO−SVR）等常用的填补算法对指标气体数据进行填补[10-13]，利用填补后的数据，通过 RF、XGBoost 和 SVR 算法建立了基于数据填补的煤自燃温度预测模型，同时利用 PSO 优化模型参数，并对数据填补后的模型性能进行了对比分析。

1　填补算法与策略

1.1　填补算法

1.1.1　KNN 算法

KNN 算法是最常用的填补算法之一，它简单易用，且模型训练时间短。KNN 算法先根据度量定义寻找与含缺失数据样本最相似的 k 个样本，即 k 个邻居，再计算它们的中位数或众数，并以此作为填补值。

1.1.2 DT 算法

DT 算法训练速度和预测速度较快，能够及时获取预测结果。因此，DT 算法也常被用于数据填补。DT 算法采用分类与回归决策树（Classification and Regression Trees, CART）的回归策略，通过递归二叉分裂划分区域，根据最小均方差的原则寻找分割点，构建回归树对缺失值进行预测填补。

1.1.3 RF 算法

RF 算法预测精度高，不易过拟合，且无需复杂的参数设置和优化。RF 算法基于 Bagging 思想，通过 Bootstrap 抽样建立 DT 基学习器，由于每个随机子集都不同，保证了构建 DT 样本集的多样性。采用随机特征选择方法，DT 上每个节点对应的特征也存在差异，进一步保障了 DT 的多样性，从而极大程度降低了 DT 之间的相关性。对每个 DT 基学习器的回归结果进行平均后得到最终结果。

1.1.4 PSO−SVR 算法

SVR 精度是由核函数决定的，SVR 核函数有线性核函数、高斯径向基核函数（RBF）和多项式核函数等。一般情况下，RBF 核函数在非线性数据上效果显著，因此本文选择 RBF 核函数。

惩罚因子 C、核函数参数 G 是 SVR 的关键参数，本文采用 PSO 对 SVR 关键参数进行优化，形成基于 PSO−SVR 的缺失数据填补方法。设种群粒子的个数为 N，解空间为 D 维，第 i 个粒子的速度为 V_i，其个体极值为 p_{best}，种群的全局极值为 g_{best}。在每一次的迭代中，粒子通过跟踪 p_{best}，g_{best} 来更新自己。粒子速度和位置更新公式为

$$v_{id}^{t+1}=\omega v_{id}^{t}+c_1 r_1(p_{\text{best}id}^{t}-X_{id}^{t})+c_2 r_2(g_{\text{best}id}^{t}-X_{id}^{t}) \tag{1}$$

$$X_{id}^{t+1}=X_{id}^{t}+v_{id}^{t+1} \tag{2}$$

式中：v_{id}^{t+1} 为更新后的粒子速度，t 为当前迭代次数，d 为维度；ω 为惯性权重，取 0.8；v_{id}^{t} 为当前的粒子速度；c_1、c_2 为学习因子，均取 0.5；r_1、r_2 为分布于（0,1）之间的随机数；$p_{\text{best}id}^{t}$ 为当前个体极值；$g_{\text{best}id}^{t}$ 为当前全局极值；X_{id}^{t+1} 为更新后的粒子位置；X_{id}^{t} 为当前粒子位置。

设最大迭代次数为 50，种群规模为 50。根据经验，设置参数优化范围为 $C\in[0.01,100]$，$G\in[0.01,50]$。步骤如下：

（1）初始化粒子位置 X 与速度 V，适应度函数为真实值与填补值的平均绝对误差百分比，初始的个体极值和全局极值分别为粒子的初始位置和适应度值最小的个体极值。

（2）每个粒子根据式（1）和式（2）更新自己的速度和位置。

（3）计算每次迭代后粒子的适应度值。

（4）更新个体极值和全局极值。

（5）判断是否满足终止条件，是否达到最大迭代次数，否则返回步骤（3）。

（6）将输出的最优参数赋给 SVR，用于数据填补。

1.1.5 XGBoost 算法

XGBoost 算法[14]是一种基于梯度提升决策树（Gradient Boosted Decision Tree, GBDT）的优化改进算法，通过多棵 DT 组合来拟合上次回归预测反馈的残差。相较于 GBDT，XGBoost 通过对损失函数二阶泰勒展开以逼近目标函数，求整体最优解，并加入正则项控制模型复杂度，防止过拟合。因而 XGBoost 具有较高的精度和泛化性，目标函数为

$$L=-\frac{1}{2}\sum_{a=1}^{A}\frac{F_a^2}{H_a+\lambda}+\gamma A \tag{3}$$

$$F_a=\sum_{b\in B_a}f_b \tag{4}$$

$$H_a=\sum_{b\in B_a}h_b \tag{5}$$

式中：A 为树中叶节点的数量；F_a 为叶子节点 a 所包含样本的一阶偏导数 f_b 之和；H_a 为叶子节点 a 所包含样本的二阶偏导数 h_b 之和；λ 为固定系数；γ 为复杂度参数；b 为 B_a 样本中的个体，B_a 为叶子节点 a 样本集中的样本。

1.2 填补策略

将当前待填补的不完整特征当作标签，其他的特征和原本的标签组成新的特征矩阵，运用 KNN，DT，RF，PSO−SVR 算法训练填补模型，将模型的输出值作为填补值，具体策略如下：

（1）当多数特征含有缺失值时，要填补一个特征，先将其他特征的缺失值用 0 占位。每完成一次填补，使当前填补好的特征参与到下一个特征的填补中，直至填补所有缺失值。

（2）当少数特征含有大量缺失值时，将第 1 个待填补缺失值的特征作为标签列，其他完整特征和原始标签列作为特征矩阵，但其他含有缺失值的特征不参与计算。填补完第 1 个特征时，使当前填补好的特征参与到下一个特征的填补中，直至填补完所有缺失值。

2 基于数据填补的煤自燃温度预测模型构建

基于数据填补的煤自燃温度预测模型构建流程如图 1 所示。

采用 PSO 优化参数，设 PSO 的惯性权重 ω 为 0.8，

学习因子 c_1、c_2 为 0.5，最大迭代次数为 50，种群规模为 50。RF 主要参数范围设置：树的数量 n_estimators 为[20, 100]，树的深度 max_depth 为[10, 25]；XGBoost 主要参数范围设置：树的数量 n_estimators 为[10, 500]，学习率 learning_rate 为[0, 1]，树的深度 max_depth 为[1, 10]，正则项 reg_lambda 为[1, 25]；SVR 参数设置：$C \in [0.01, 100]$，$G \in [0.01, 50]$。模型构建的具体步骤如下：

（1）将通过实验获取的指标气体浓度作为特征，煤温作为标签，对数据进行标准化。

（2）采用 5 折交叉验证法划分训练集和测试集。

（3）初始化参数并随机生成一组粒子的速度与位置。

（4）通过式（1）和式（2）更新粒子的速度与位置。采用均方根误差（Root Mean Square Error, RMSE）作为适应度函数，计算比较每次迭代后粒子适应度值，更新个体极值和全局极值。

（5）判断是否满足终止条件，是否达到最大迭代次数，否则返回步骤（4）。

（6）将输出的最优参数赋给 RF/XGBoost/SVR 模型，用于煤自燃温度预测。

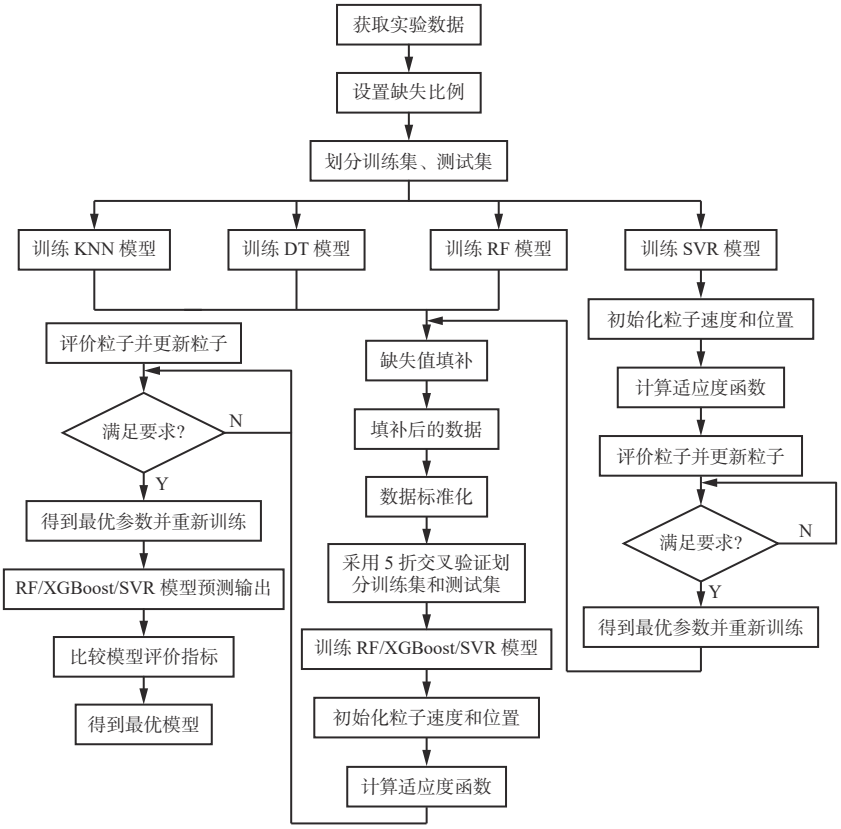


图 1　基于数据填补的煤自燃温度预测模型构建流程

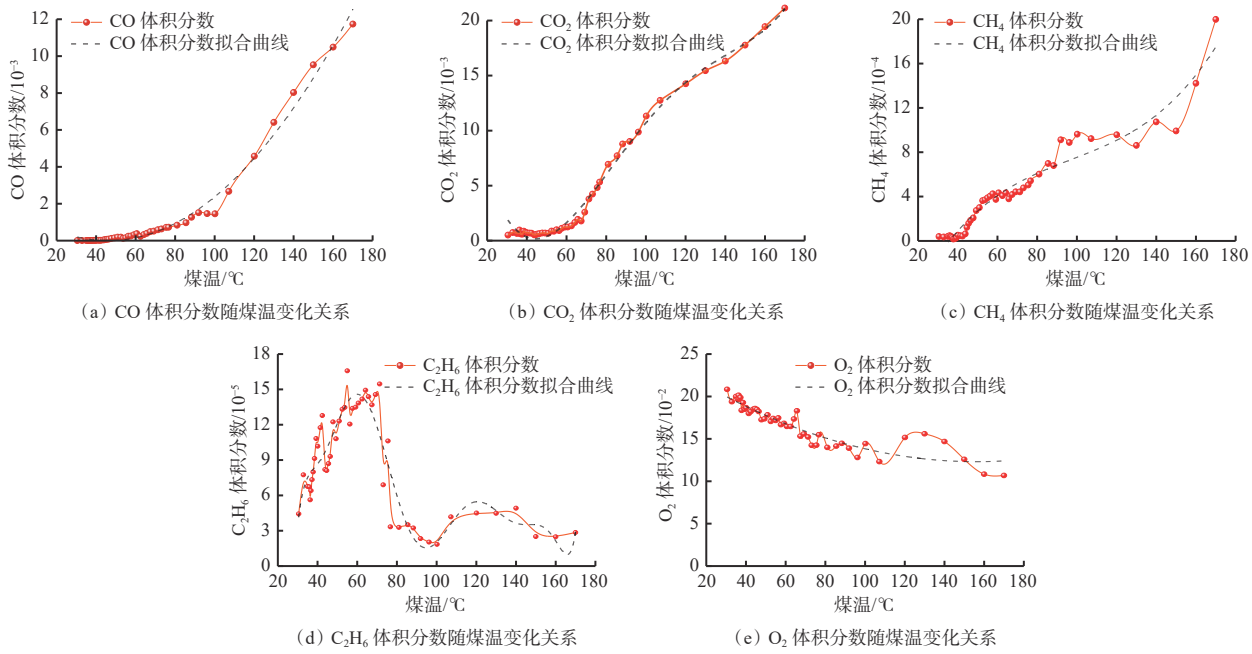
Fig. 1　Process of coal spontaneous combustion temperature prediction model based on data filling

3　指标气体数据获取与缺失值设计

3.1　煤自燃发火实验

从陕西省铜川市柴家沟矿采集新鲜煤样 2 t 左右进行煤自然发火实验。结合常用于预测预判煤自燃危险程度的气体指标[15]，选取 CO, CO_2, CH_4, C_2H_6, O_2 体积分数作为表征煤自燃危险程度的指标。根据实验结果绘制柴家沟矿煤样指标气体随煤温变化关系图，如图 2 所示。

从图 2 可看出：CO 体积分数随着煤温的上升表现为指数形式，且上升过程呈阶段性变化；CO_2 体积分数随着煤温的升高而上升，且上升过程表现出较为明显的阶段性变化；随着煤温的不断上升，CH_4 体积分数也逐渐升高；C_2H_6 体积分数随着煤温的上升呈先上升后下降趋势，煤温超过 70 ℃ 后，C_2H_6 体积分数达到峰值；O_2 体积分数随着煤温持续上升逐渐下降。初期阶段能检测到少量的 CO, CO_2, CH_4 气体，CO, CO_2 体积分数变化幅度较小。当煤温超过临界温度 70 ℃ 后，CO, CO_2 体积分数曲线的斜率明显增大，CH_4 体积分数也明显增大，当煤温持续上升至干裂温度 100 ℃ 后，煤体发生剧烈的氧化反应，CO, CO_2, CH_4 体积分数快速上升。实验结果表明：CO, CO_2, CH_4, C_2H_6, O_2 体积分数的变化与煤温相对应，具有明显的相关性，可作为煤自燃预警指标。



（a）CO 体积分数随煤温变化关系　（b）CO_2 体积分数随煤温变化关系　（c）CH_4 体积分数随煤温变化关系

（d）C_2H_6 体积分数随煤温变化关系　（e）O_2 体积分数随煤温变化关系

图 2　柴家沟矿煤样指标气体随煤温变化关系

Fig. 2　Relationship between index gas of coal sample and coal temperature in Chaijiagou Mine

3.2　缺失值设计

在实际测量中，出现缺失值的情况有以下 2 种：① 由于传感器、气相色谱仪等设备异常或者人为误操作，测量出现较大误差，出现偏差过大的异常值，预处理之后会出现多数特征含有缺失值的情况。② 部分气体传感器、气相色谱仪等设备故障，造成部分气体数据存在缺失，出现少数特征含有大量缺失值的情况。因此，本文设计 2 种缺失值：所有特征都含有缺失值和少数特征含有大量缺失值。

设备异常导致的缺失具有随机性，所以设置数据缺失类型为完全随机缺失，设置整体缺失率为 10%, 20%, 30%。由通过 RF 方法获得的特征重要性（表 1）可知，CO, CO_2 的特征重要性较高，对预测结果有较大影响，为更好地比较填补算法的优劣，选择 CO, CO_2 作为含有较多缺失值的特征，缺失率为 40%, 50%, 60%。

表 1　特征重要性

Table 1　Importance of characteristics

特征	CO	CO_2	CH_4	C_2H_6	O_2
特征重要性	0.259	0.427	0.186	0.028	0.101

4　填补效果与模型对比分析

4.1　模型评价指标

为客观评估模型的性能，本文选择常用的 3 种评估指标来评估模型精度：平均绝对误差百分比（Mean Absolute Percentage Error, MAPE）、判断系数 R^2 和 RMSE。

$$P_{\mathrm{E}}=\frac{100\%}{n}\sum_{c-1}^{n}\left|\frac{\hat{y}_c-y_c}{y_c}\right| \tag{6}$$

$$R^2=1-\frac{\sum_c(y_c-\hat{y}_c)^2}{\sum_c(y_c-\bar{y})^2} \tag{7}$$

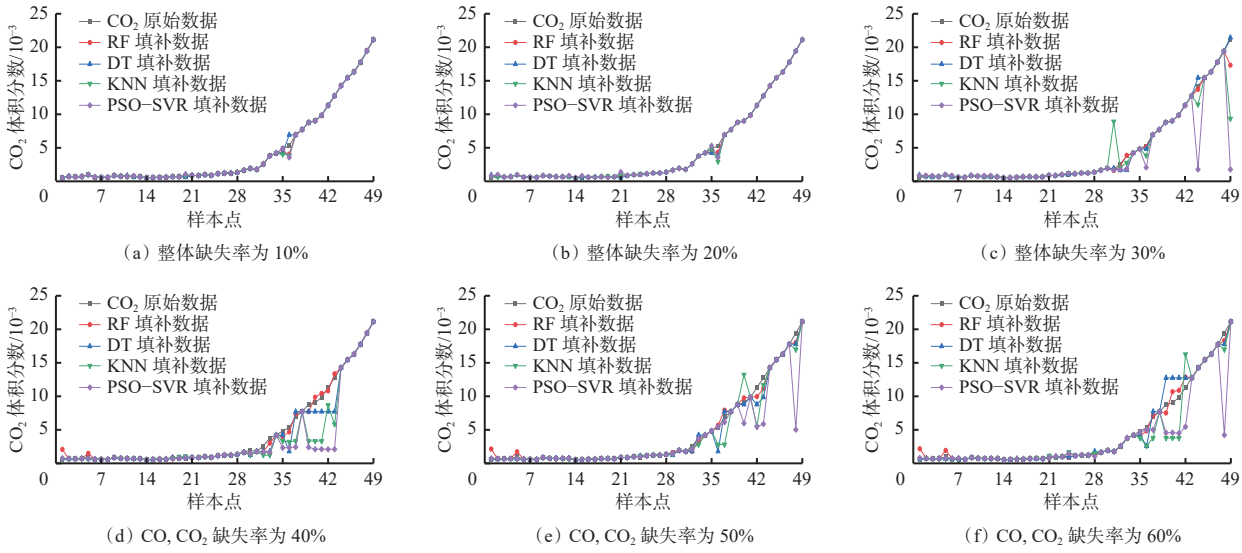
$$M_{\mathrm{E}}=\sqrt{\frac{1}{n}\sum_{c-1}^{n}(\hat{y}_c-y_c)^2} \tag{8}$$

式中：P_{E} 为平均绝对误差百分比；$\hat{y}_c$为第 c 个样本的预测值，c=1, 2, …, n, n 为样本数量；y_c为第 c 个样本的实际值；$\bar{y}$为样本平均值；M_{E} 为均方根误差。

4.2　填补算法效果对比分析

对不同缺失比例的数据集分别采用 KNN、RF、DT 和 PSO−SVR 填补算法进行填补实验，每种算法均重复填补 100 次（填补值取 100 次的均值），部分特征的填补值与原始数据的对比如图 3 和图 4 所示。计算所有特征的填补值和原始数据的 MAPE（取 100 次实验的均值），并相加，结果如图 5 所示。

从图 3 和图 4 可看出：数据缺失率为 10%, 20% 时，4 种填补算法的填补数据与原始数据相比，均无明显差异，随着缺失率的增大，所有算法的填补效果呈下降趋势。在 CO, CO_2 缺失率为 40%, 50%, 60% 时，RF 算法填补的前 7 个样本点数据质量差，特别是 CO_2 数据有较明显的差距，但整体填补效果较好；DT 算法填补数据与原始数据的总体差异较小，但 CO, CO_2 缺失率为 60% 时，差异明显增大；缺失率为 30% 时，KNN 与 PSO−VR 算法的填补数据开始振荡，填补效果显著下降。

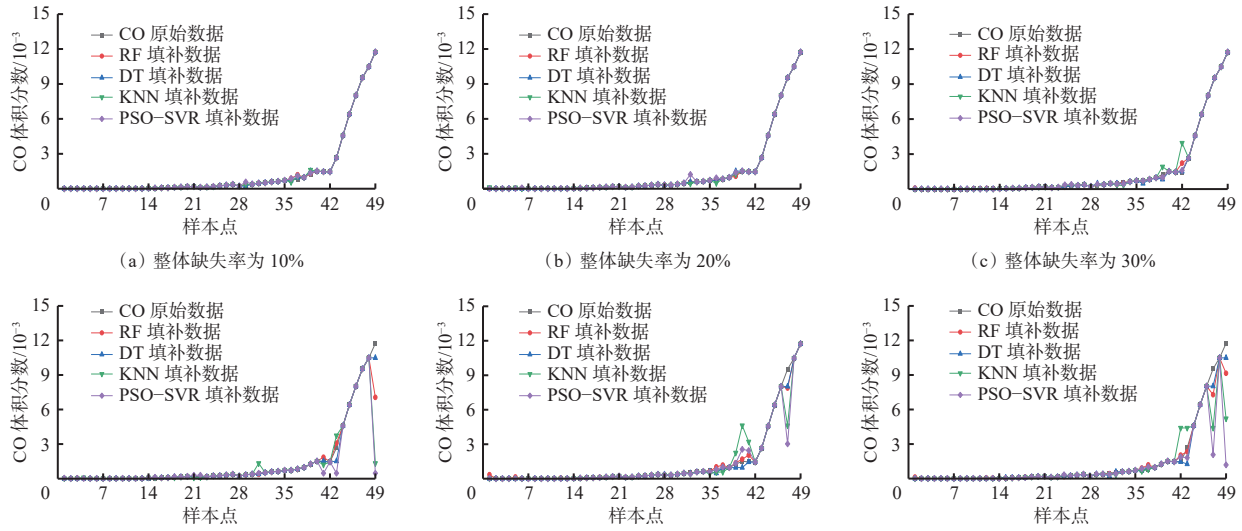


（a）整体缺失率为 10%　（b）整体缺失率为 20%　（c）整体缺失率为 30%

（d）CO, CO_2 缺失率为 40%　（e）CO, CO_2 缺失率为 50%　（f）CO, CO_2 缺失率为 60%

图 3　不同缺失率下 CO_2 体积分数的数据填补效果对比

Fig. 3　Filling effects comparison of CO_2 volume fraction data at different miss rates



（a）整体缺失率为 10%　（b）整体缺失率为 20%　（c）整体缺失率为 30%

（d）CO, CO_2 缺失率为 40%　（e）CO, CO_2 缺失率为 50%　（f）CO, CO_2 缺失率为 60%

图 4　不同缺失率下 CO 体积分数的数据填补效果对比

Fig. 4　Filling effects comparison of CO volume fraction data at different miss rates

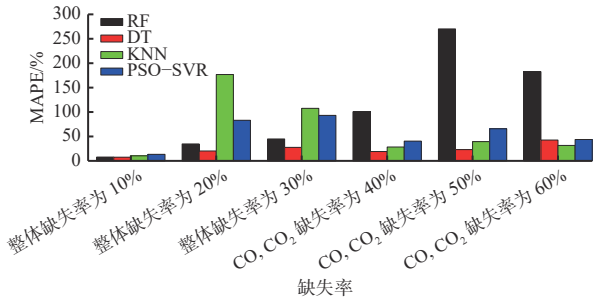


图 5　填补效果对比

Fig. 5　Comparison of filling effect

从图 5 可看出：对于 6 种不同缺失数据，DT 填补算法的 MAPE 最小。CO，CO_2 缺失率为 40%，50% 和 60% 时，RF 的 MAPE 偏大，这是由于煤自然发火初始阶段的指标气体浓度较低，RF 填补数据与原始数据相差较大，导致 MAPE 显著增大。

综上可知，DT 算法填补的效果最优。

4.3　模型精度对比分析

基于完整数据的模型预测精度指标对比见表 2。可看出 XGBoost 模型在训练阶段的效果极好，但测试阶段的 MAPE、RMSE 明显增大，R^2 明显减小，明显过拟合，对于小样本数据，XGBoost 算法极易过拟合；SVR 模型在不调参的情况下效果显著低于其他模型，而 RF 模型优于其他 2 个模型。

不调参的 XGBoost、SVR 模型精度极低，因此仅讨论不同填补算法对 RF 模型在测试集上预测精度的影响，如图 6 所示。

从图 6 可看出，数据缺失对煤自燃温度预测模型精度有较大影响。整体缺失率为 10% 时，数据填

表 2　基于完整数据的模型评价指标对比

Table 2　Comparison of model evaluation index based on complete data

预测模型	模型评价指标		
	训练集/测试集 RMSE/℃	训练集/测试集 MAPE/%	训练集/测试集 R^2
RF	1.856 /4.460	1.539 /4.034	0.997/0.978
XGBoost	0.001/4.544	0.001/4.650	1.000/0.975
SVR	30.782/30.994	24.678/30.190	0.198/0.198

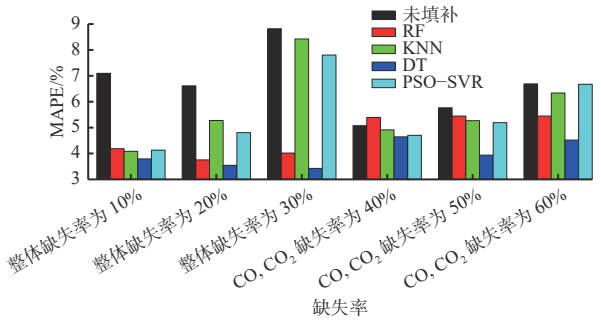


图 6　基于不同填补算法的 RF 预测模型在测试集上的精度对比

Fig. 6　Precision comparison of RF prediction models based on different filling algorithms in test set

补后的 RF 预测模型精度和基于完整数据的 RF 预测模型精度相差均不大，MAPE 在 3.7%～4.2% 之间。整体缺失率为 20%, 30% 时，基于 RF、DT 填补的预测模型的 MAPE 分别为 3.7%, 3.5%, 而基于 KNN、PSO−SVR 填补的预测模型的 MAPE 随着缺失率增加而大幅增加，最大可达 8.3%。CO, CO_2 缺失率为 40%, 50%, 60% 时，基于 RF、DT 填补的预测模型的 MAPE 分别稳定在 5.4% 左右和 4.3% 左右，而基于 KNN、PSO−SVR 填补的预测模型的 MAPE 呈增长趋势，预测精度相对较低。基于 DT 填补的 RF 预测模型在 6 种缺失率下的 MAPE 平均值为 4%，明显低于其他模型。

综上所述，在 6 种缺失情况下，基于 DT 填补算法的预测模型精度总体优于基于其他填补算法的预测模型。

4.4　PSO 优化参数后模型性能对比分析

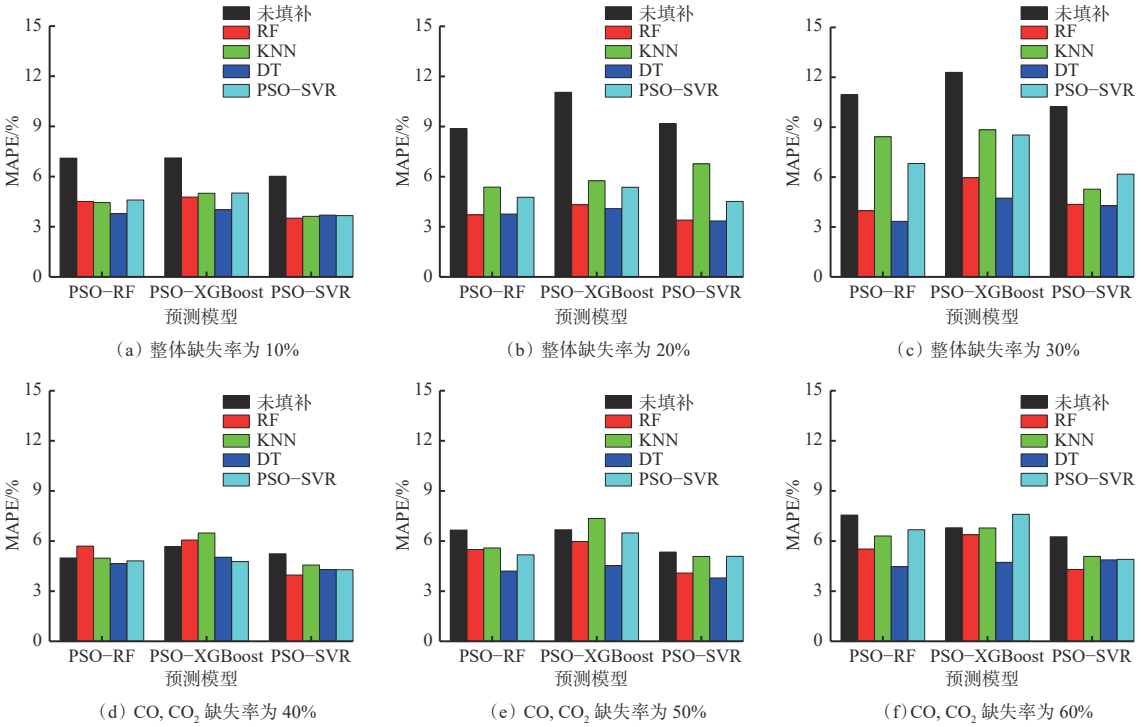
PSO 优化后模型指标见表 3。可看出 PSO 算法调参后，XGBoost 和 SVR 模型精度均有较大的提升，而 RF 模型却与调参前的效果差距不大，说明 RF 模型不进行参数优化也有较好的预测精度；PSO−XGBoost 模型过拟合情况减弱，精度提高，在训练集的效果最好；PSO−SVR 预测模型相较于调参前模型精度显著提高，在训练集和测试集上的预测效果相差最小，无过拟合情况，在测试集的效果最好，泛化性较强。

表 3　基于完整数据的 PSO 优化后的模型指标对比

Table 3　Comparison of PSO optimized model index based on complete data

预测模型	模型评价指标		
	训练集/测试集 RMSE/℃	训练集/测试集 MAPE/%	训练集/测试集 R^2
PSO−RF	1.847/4.211	1.715/4.344	0.997/0.976
PSO−XGBoost	1.235/4.400	0.414/3.912	0.999/0.979
PSO−SVR	2.323/2.427	2.910/3.325	0.995/0.990

PSO 优化后的填补数据在测试集的预测精度对比和模型 MAPE 平均值如图 7 和表 4 所示。



（a）整体缺失率为 10%　（b）整体缺失率为 20%　（c）整体缺失率为 30%

（d）CO, CO_2 缺失率为 40%　（e）CO, CO_2 缺失率为 50%　（f）CO, CO_2 缺失率为 60%

图 7　不同缺失率下基于不同填补算法的预测模型性能对比

Fig. 7　Performance comparison of prediction models based on different filling algoriths under different miss rates

表 4　不同缺失率下预测模型的平均 MAPE

Table 4　Mean MAPE of prediction models under different miss rates　%

预测模型	未填补	填补算法			
		RF	KNN	DT	PSO−SVR
PSO−RF	7.7	4.8	5.6	4.05	5.5
PSO−XGBoost	8.3	5.6	6.7	4.50	6.3
PSO−SVR	7.0	3.9	5.1	4.04	4.8

从图 7 和表 4 可看出：基于 DT 填补算法的预测模型的 MAPE 明显低于基于其他填补算法的预测模型，说明 DT 算法填补后的模型预测准确率更高。基于 DT 算法填补的 PSO−RF 模型和 PSO−SVR 模型在测试集的 MAPE 均值均在 4% 左右，但 RF 与 PSO−RF 模型的 MAPE 相差极小，说明 RF 无需调参，而 SVR 需调参后才能满足预测要求。

综上可知，基于 DT 填补算法的 RF 煤自燃温度预测模型预测性能最优。

5　结论

（1）在 6 种数据缺失情况下，DT 算法填补效果优于其他 3 种算法。在 CO，CO_2 存在较多缺失值时，RF 算法的填补值与实际值的 MAPE 偏大。

（2）基于 4 种填补算法分别建立了 RF、XGBoost 和 SVR 煤自燃温度预测模型。在不调参的情况下，XGBoost 模型虽然在训练集的效果极好，但极易过拟合，而 SVR 模型预测效果极差，均无法满足预测要求。在 6 种数据缺失情况下，基于 DT 填补算法的 RF 预测模型的 MAPE 的平均值为 4%，不进行参数优化也有较好的预测精度，能够满足实际需求。

（3）在 6 种数据缺失情况下，基于 DT 填补算法的 PSO−SVR、RF 与 PSO−RF 预测模型的 MAPE 均在 4% 左右，而基于 DT 填补算法的 RF 模型无需优化就能较好地预测出煤自燃温度，具有良好稳定性。

参考文献（References）：

[1] 邓军，白祖锦，肖旸，等. 煤自燃灾害防治技术现状与挑战[J]. 煤矿安全，2020，51(10)：118-125.
DENG Jun，BAI Zujin，XIAO Yang，et al. Present situation and challenge of coal spontaneous combustion disasters prevention and control technology[J]. Safety in Coal Mines，2020，51(10)：118-125.

[2] 王德明，邵振鲁，朱云飞. 煤矿热动力重大灾害中的几个科学问题[J]. 煤炭学报，2021，46(1)：57-64.
WANG Deming，SHAO Zhenlu，ZHU Yunfei. Several scientific issues on major thermodynamic disasters in coal mines[J]. Journal of China Coal Society，2021，46(1)：57-64.

[3] 王德明. 煤矿热动力灾害及特性[J]. 煤炭学报，2018，43(1)：137-142.
WANG Deming. Thermodynamic disaster in coal mine and its characteristics[J]. Journal of China Coal Society，2018，43(1)：137-142.

[4] 郭庆. 采空区煤自燃预警技术及应用研究[D]. 徐州：中国矿业大学，2021.
GUO Qing. Research on early warning technology and application of coal spontaneous combustion in goaf[D]. Xuzhou：China University of Mining and Technology，2021.

[5] ONIFADE M，GENC B，BADA S. Spontaneous combustion liability between coal seams：a thermogravimetric study[J]. International Journal of Mining Science and Technology，2020，30(5)：691-698.

[6] 邓军，雷昌奎，曹凯，等. 采空区煤自燃预测的随机森林方法[J]. 煤炭学报，2018，43(10)：2800-2808.
DENG Jun，LEI Changkui，CAO Kai，et al. Random forest method for predicting coal spontaneous combustion in gob[J]. Journal of China Coal Society，2018，43(10)：2800-2808.

[7] DENG Jun，CHEN Weile，WANG Caiping，et al. Prediction model for coal spontaneous combustion based on SA-SVM.[J]. ACS Omega，2021，6(17)：11307-11318.

[8] 周旭，朱毅，张九零，等. 基于PSO−XGBoost的煤自燃程度预测研究[J]. 矿业安全与环保，2022，49(6)：79-84.
ZHOU Xu，ZHU Yi，ZHANG Jiuling，et al. Study on prediction model of coal spontaneous combustion based on PSO-XGBoost[J]. Mining Safety & Environmental Protection，2022，49(6)：79-84.

[9] 彭志江. 面向小样本数据的特征分析技术研究[D]. 成都：电子科技大学，2021.
PENG Zhijiang. Feature analysis technology for small sample data[D]. Chengdu：University of Electronic Science and Technology of China，2021.

[10] 郑晓亮. 基于瓦斯含量法的煤与瓦斯突出预测关键技术研究[D]. 淮南：安徽理工大学，2018.
ZHENG Xiaoliang. Research on key technology of coal and gas outburst prediction based on gas content method[D]. Huainan：Anhui University of Science and Technology，2018.

[11] 陈娟，王献雨，罗玲玲，等. 缺失值填补效果：机器学习与统计学习的比较[J]. 统计与决策，2020，36(17)：28-32.
CHENG Juan，WANG Xianyu，LUO Lingling，et al. Comparison of machine learning and statistical learning in the imputation of missing values[J]. Statistics & Decision，2020，36(17)：28-32.

（下转第 98 页）

Mine Automation，2013，39(9)：112-115.

[8] 董立红，王杰，厍向阳. 基于改进Camshift算法的钻杆计数方法[J]. 工矿自动化，2015，41(1)：71-76.
DONG Lihong，WANG Jie，SHE Xiangyang. Drill counting method based on improved Camshift algorithm[J]. Industry and Mine Automation，2015，41(1)：71-76.

[9] 高瑞，郝乐，刘宝，等. 基于改进ResNet网络的井下钻杆计数方法[J]. 工矿自动化，2020，46(10)：32-37.
GAO Rui，HAO Le，LIU Bao，et al. Research on underground drill pipe counting method based on improved ResNet network[J]. Industry and Mine Automation，2020，46(10)：32-37.

[10] 党伟超，姚远，白尚旺，等. 煤矿探水卸杆动作识别研究[J]. 工矿自动化，2020，46(7)：107-112.
DANG Weichao，YAO Yuan，BAI Shangwang，et al. Research on unloading drill-rod action identification in coal mine water exploration[J]. Industry and Mine Automation，2020，46(7)：107-112.

[11] YAN Sijie，XIONG Yuanjun，LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence，New Orleans，2018：5361-5368.

[12] FANG Haoshu，XIE Shuqin，TAI Yuwing，et al. RMPE：regional multi-person pose estimation[C]. Proceedings of the IEEE International Conference on Computer Vision，Venice，2017：2334-2343.

[13] HU Jie，SHEN Li，SUN Gang. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Salt Lake City，2018：7132-7141.

[14] KE Qiuhong，BENNAMOUN M，AN Senjian，et al. A new representation of skeleton sequences for 3D action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Honolulu，2017：3288-3297.

[15] 卫少洁，周永霞. 一种结合Alphapose和LSTM的人体摔倒检测模型[J]. 小型微型计算机系统，2019，40(9)：1886-1890.
WEI Shaojie，ZHOU Yongxia. Human body fall detection model combining Alphapose and LSTM[J]. Journal of Chinese Computer Systems，2019，40(9)：1886-1890.

[16] 杨世强，李卓，王金华，等. 基于新分区策略的ST-GCN人体动作识别[J/OL]. 计算机集成制造系统：1-16[2022-03-29]. http://kns.cnki.net/kcms/detail/11.5946.TP.20211022.1500.014.html.
YANG Shiqiang，LI Zhuo，WANG Jinhua，et al. ST-GCN human action based on new partition strategy[J/OL]. Computer Integrated Manufacturing Systems：1-16[2022-03-29]. http://kns.cnki.net/kcms/detail/11.5946.TP.20211022.1500.014.html.

（上接第 35 页）

[12] 陈利成，陈建宏. 基于数据填补-机器学习的煤与瓦斯突出预测效果研究[J]. 中国安全生产科学技术，2022，18(9)：69-74.
CHEN Licheng，CHEN Jianhong. Study on prediction effect of coal and gas outburst based on data imputation and machine learning[J]. Journal of Safety Science and Technology，2022，18(9)：69-74.

[13] 郑晓亮，来文豪，薛生. MI和SVM算法在煤与瓦斯突出预测中的应用[J]. 中国安全科学学报，2021，31(1)：75-80.
ZHENG Xiaoliang，LAI Wenhao，XUE Sheng. Application of MI and SVM in coal and gas outburst prediction[J]. China Safety Science Journal，2021，31(1)：75-80.

[14] LI Zhuoxuan，SHI Xinli，CAO Jinde，et al. CPSO-XGBoost segmented regression model for asphalt pavement deflection basin area prediction[J]. Science China（Technological Sciences），2022，65(7)：1470-1481.

[15] 任万兴，郭庆，石晶泰，等. 基于标志气体统计学特征的煤自燃预警指标构建[J]. 煤炭学报，2021，46(6)：1747-1758.
REN Wanxing，GUO Qing，SHI Jingtai，et al. Construction of early warning indicators for coal spontaneous combustion based on statistical characteristics of index gases[J]. Journal of China Coal Society，2021，46(6)：1747-1758.