移动阅读

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.doi:10.13225/j.cnki.jccs.2018.0710

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.doi:10.13225/j.cnki.jccs.2018.0710

采空区煤自燃预测的随机森林方法

邓　军[1,2],雷昌奎[1,2],曹　凯[3,4],马　砺[1,2],王彩萍[1,2],翟小伟[1,2]

(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.徐州安云矿业科技有限公司,江苏 徐州　221008;4.中国矿业大学 通风防灭火研究所,江苏 徐州　221008)

摘　要: 煤自然发火温度的准确预测是矿井煤自燃防控的关键。为了准确可靠地预测采空区煤自燃温度,在大佛寺煤矿 40106 综放工作面开展了长期的采空区温度和气体观测实验,提出了一种基于随机森林(RF)方法的采空区煤自燃预测模型,并将预测结果与支持向量机(SVM)和 BP 神经网络(BPNN)方法对比。采用粒子群优化算法(PSO)对 RF 和 SVM 超参数进行优化,建立了参数优化的 PSO-RF 和 PSO-SVM 预测模型。结果表明,RF,PSO-RF,SVM 和 PSO-SVM 模型均具有较强的泛化性和鲁棒性;RF 在建模过程中拥有宽广的参数适应范围,当树的数量(n_{tree})超过 100 后,其训练误差趋于稳定,n_{tree}的改变对预测性能没有实质的影响;虽然 PSO 算法可以找到 RF 最优超参数,但默认参数的 RF 模型就能获得满意的预测性能;SVM 预测结果则对超参数十分敏感,PSO 优化可以显著提高其预测精度,其预测性能依赖于超参数的最优选择;BPNN 模型在训练阶段拥有极佳的预测结果,但易出现"过拟合",导致泛化性弱,测试阶段误差较大。通过在其他矿井煤自燃预测中应用,验证了 RF 方法的稳定性和普适性,且无需复杂参数设置和优化就能获得良好的预测性能,可进一步应用于其他能源燃料领域。

关键词: 采空区;煤自燃;随机森林;支持向量机;粒子群优化;温度预测

中图分类号: TD752.2　　**文献标志码:** A　　**文章编号:** 0253-9993(2018)10-2800-09

Random forest method for predicting coal spontaneous combustion in gob

DENG Jun[1,2],LEI Changkui[1,2],CAO Kai[3,4],MA Li[1,2],WANG Caiping[1,2],ZHAI Xiaowei[1,2]

(1.*School of Safety Science and Engineering,Xi'an University of Science and Technology,Xi'an*　710054,*China*;2.*Shanxi Key Laboratory of Prevention and Control of Coal Fire,Xi'an*　710054,*China*;3.*Xuzhou Anyun Mining Technology Co.,Ltd.,Xuzhou*　221008,*China*;4.*Ventilation and Fire Prevention Institute,China University of Mining and Technology,Xuzhou*　221008,*China*)

Abstract: The accurate prediction of coal temperature plays a vital role in preventing and controlling the coal spontaneous combustion in coal mines.To predict the temperature of coal spontaneous combustion in a gob accurately and reliably,a long-term observation test of temperature and gases was implemented in the gob of 40106 fully mechanized top-coal caving face at Dafosi coal mine.A prediction model of coal spontaneous combustion in the gob based on random forest (RF) method was proposed,and the prediction results were compared with the support vector machine (SVM) and BP neural network (BPNN) methods.The particle swarm optimization (PSO) algorithm was employed to optimize the hyper-parameters of RF and SVM for establishing the PSO-RF and PSO-SVM prediction models with optimized parameters.The results indicate that RF,PSO-RF,SVM,and PSO-SVM models all had strong generalization and

收稿日期: 2018-05-28　**修回日期:** 2018-08-20　**责任编辑:** 郭晓炜

基金项目: 国家重点研发计划资助项目(2018YFC0807900);国家自然科学基金资助项目(51504186);陕西省国际科技合作与交流计划资助项目(2016KW-070)

作者简介: 邓　军(1970—),男,四川大竹人,教授,博士生导师。E-mail:dengj518@xust.edu.cn

通讯作者: 雷昌奎(1990—),男,陕西安康人,博士研究生。E-mail:lchangkui@126.com

robustness. RF possessed a wide range of parameters adaptation in the modeling process. When the number of trees (n_{tree}) exceeded 100, the training errors tended to be stable, and the change of n_{tree} had no substantial impact on the prediction performance. Although the PSO algorithm could find the optimal hyper-parameters of RF, the RF model with the default parameters could obtain a satisfactory prediction performance. The prediction results of SVM were very sensitive to its hyper-parameters, PSO optimization could significantly improve its prediction accuracy, and its prediction performance depended on the optimal choice of hyper-parameters. The BPNN model exhibited excellent prediction results in the training stage, but it was prone to "over-fitting", resulting in weak generalization and large errors in the testing stage. Through the application of coal spontaneous combustion prediction in other mines, the stability and universality of the RF method were verified, and good prediction performance could be obtained without complicated parameter settings and optimization, it could be further applied to other energy and fuel fields.

Key words: gob; coal spontaneous combustion; random forest; support vector machine; particle swarm optimization; temperature prediction

煤自燃是影响煤矿安全的主要灾害之一,它不仅会造成资源浪费,设备损坏,甚至会引起瓦斯与煤尘爆炸,诱发二次灾害,造成严重的人员伤亡和财产损失[1-2]。由于采空区内部的隐蔽性和难以接近性,采空区是煤自燃最容易发生的地点之一,尤其是随着综放开采技术的推广与应用,其高产高效的特点使得开采强度和采空区面积不断增大,从而瓦斯涌出量、通风压力和强度也随之增大,这些无疑增加了采空区自燃的危险[3]。因此,采空区煤自燃问题一直是困扰矿井安全生产的亟待解决的难题。及时准确的预测预报是煤自燃防控的前提,目前,气体分析法是矿井最常用的方法之一,它主要是根据煤自燃氧化过程中指标气体的出现和浓度变化来判定和预测煤自燃状态和发展趋势[4-6]。煤自燃是一个复杂的动态物理化学氧化过程,在不同的氧化自热状态,指标气体的浓度会发生相应的变化[7]。因此,可以通过监测煤氧化自热过程中释放的指标气体来预测预报煤自燃。然而,由于煤自燃与气体之间是十分复杂的非线性关系,如何通过一种科学可靠的方法来解释和处理这种非线性关系是解决问题的关键。很多学者在这方面进行了研究与探讨,如王磊等[8]提出了一种将灰色模型和马尔科夫模型相结合对煤自燃进行预测的方法;邬剑明等[9]、周福宝等[10]、靳玉萍[11]应用神经网络方法建立了煤自燃预测模型;高原等[12]、孟倩等[13]、邵良杉等[14]、靳玉萍等[15]、邓军等[16-17]采用支持向量机方法对采空区自然发火进行预测。然而,由于煤矿井下条件的复杂多变性,灰色系统方法难以达到预期的精度;神经网络方法虽然具有优越的非线性处理能力,但其训练过程容易出现"过拟合";支持向量机方法虽然克服了神经网络的不足,但其预测精度对超参数的要求较高。

鉴于此,迫切需要引入一种新的、更有效的方法来解决上述问题。BREIMAN 于 2001 年提出的随机森林方法是一种性能优越的预测工具,BREIMAN 已经证明随机森林具有预测性能高、容错性好和不易"过拟合"的特点[18]。在煤炭领域,随机森林方法主要用于煤炭属性的预测,如发热量[19]、自由膨胀指数[20-21]、可磨性指数[22]等。然而,随机森林方法用于预测采空区煤自燃的研究目前尚鲜见报道。因此,笔者引入随机森林方法进行采空区煤自燃温度预测建模,并提出采用粒子群优化算法对其超参数进行优化,研究超参数对其预测性能的影响,同时与支持向量机和神经网络方法的预测结果进行对比,结合预测结果分析随机森林方法在采空区煤自燃温度预测方面的特点。进一步,将随机森林方法应用于其他矿井,研究其对于不同矿井应用的普适性。

1　方法与实验

1.1　随机森林

随机森林(Random Forest, RF)是一种基于 Bagging 的集成学习方法,可以用于分类和回归问题。它由多个决策树构成,树的构建遵从 CART 策略,并不进行剪枝[23]。RF 结构如图 1 所示,训练集 $S=\{(x_i, y_i), i=1,2,\cdots,N\}$,$(\boldsymbol{X},\boldsymbol{Y}) \in R^M \times R$,输入矩阵 $\boldsymbol{X}$ 由具有 M 个属性的 N 个样本组成,输出 $\boldsymbol{Y}$ 是一个目标向量。

RF 生成步骤如下:

Step. 1,利用 Bootstrap 重抽样方法从原始样本集 S 抽取 n_{tree} 个样本集($S_k, k=1,2,\cdots,n_{tree}$),每次未被抽到的样本组成了 n_{tree} 个袋外数据(out-of-bag, OOB),抽中的样本称为袋内数据(In-bag)。

Step. 2,从 M 个属性中随机选取 m_{try} 个属性作为子集,再从这个子集中选择一个最优属性进行节点分裂,构建 CART 树。

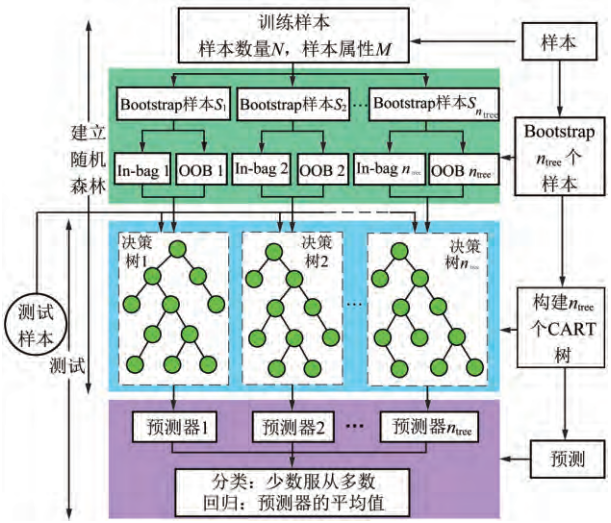

图 1 随机森林算法流程

Fig. 1 Flowchart of the random forest algorithm

Step. 3，每棵树最大限度地生长，不做任何剪裁，m_{try} 值在整个森林的生长过程中保持不变。

Step. 4，在 n_{tree} 轮中共生成 n_{tree} 个 CART 树，由于这 n_{tree} 个决策树在训练集的选择和特征的选择上都是随机的，所以这 n_{tree} 个决策树之间是相互独立的。

Step. 5，将生成的多棵树组成随机森林，用随机森林对新的数据进行预测：对于分类问题，通过少数服从多数的投票方法决定 RF 的预测结果；对于回归问题，将所有回归决策树输出值的平均值作为 RF 的预测值。

对于本文所采用的 RF 回归，其最终输出结果为

$$f(x_t) = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} h_i(x_t) \tag{1}$$

式中，$f(x_t)$ 为 RF 预测结果；$h_i(x_t)$ 为每棵决策树给出的结果。

在 RF 建模时，由于 OOB 数据没有参与决策树构建过程，因此，它可以作为测试集评价预测误差和性能，即

$$MSE_{OOB} = n_{tree}^{-1} \sum_{i=1}^{n_{tree}} (y_i - \hat{y}_i^{OOB})^2 \tag{2}$$

$$R_{RF}^2 = 1 - MSE_{OOB} / \hat{\sigma}_y^2 \tag{3}$$

式中，MSE_{OOB} 为 OOB 数据预测的均方误差；y_i 为真实值，℃；$\hat{y}_i^{OOB}$ 为 RF 对 OOB 数据的预测值，℃；R_{RF}^2 为 OOB 数据预测值的决定系数；$\hat{\sigma}_y^2$ 为 RF 对 OOB 数据预测值的方差。

1.2 现场实验

为了获得真实准确的现场采空区数据，以陕西彬长集团大佛寺煤矿 40106 综放工作面作为现场测试基地进行采空区气体和温度观测实验。40106 工作面走向长度 1 970 m，倾向长度 200 m，采用倾斜长壁式开采，“U”型负压通风。工作面平均煤厚 11.65 m，机采高度为 3.8 m，平均放煤高度为 7.65 m。煤层是低变质程度烟煤，容易发生自燃。

为了全面掌握采空区气体和温度信息，沿工作面倾向方向均匀布置 6 个测点，相邻测点间间距为 40 m。分别在进风和回风巷道两侧铺设束管和测温导线，在距离工作面前方 200～300 m 处通过抽气泵抽取指标气体和测温装置测温。采空区束管监测系统布置如图 2(a) 所示。

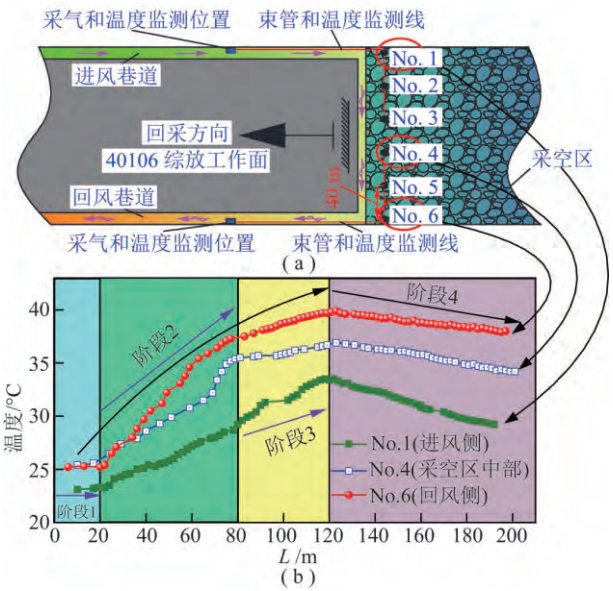

图 2 采空区束管监测系统布置及测点温度

Fig. 2 Layout of the bundle tube monitoring system and temperature measurement in the gob

1.3 模型性能评估指标

采用平均绝对误差(MAE)、平均绝对百分比误差(MAPE)、均方根误差(RMSE)和决定系数(R^2) 4 个指标作为模型预测性能的评估依据，分别定义为

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| \tag{4}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{f_i - y_i}{y_i} \right| \times 100\% \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f_i - y_i)^2} \tag{6}$$

$$R^2 = 1 - \sum_{i=1}^{n} (y_i - f_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7}$$

式中，f_i 为预测值，℃；$\bar{y}$ 为真实值的平均值，℃。

2 随机森林方法预测采空区煤自燃

2.1 数据分析与设置

在工作面回采过程中，由于顶板来压，垮落的

顶板造成部分测点毁坏，最终从进风侧的 No. 1、回风侧的 No. 6 和接近采空区中部的 No. 4 测点获得了完整的数据(图 2(b))。经过为期 33 d 的现场观测，共获得 220 组数据，采空区氧气体积分数降至 4%左右，如图 3 所示。在建模过程中，以测点距工作面的距离(记为 L)，O_2，CO，CO_2 和 CH_4 体积分数作为模型输入变量，煤温(记为 T)作为模型目标输出，其中 160 组样本用于训练建模，剩余 60 组(约总样本数的 1/4)未参与训练建模的样本用于模型性能测试。

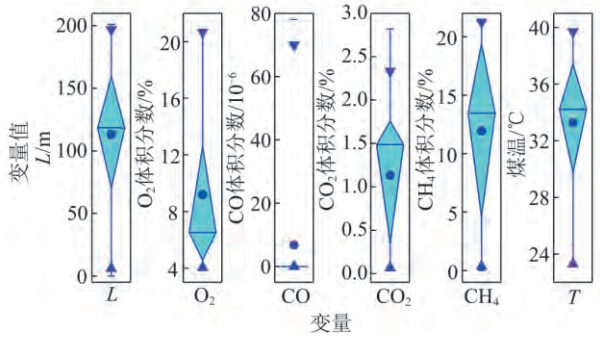

图 3　现场测试数据箱形

Fig. 3　Box plot of the data acquired from in-situ test

为了更清楚的展现采空区温度和气体变化，以 L 和 W(工作面宽度)为 x 轴和 y 轴，采空区监测的温度，O_2，CO，CO_2 和 CH_4 体积分数分别为 z 轴，利用 MATLAB 中的 meshgrid 函数与 griddata 函数配合使用对现场监测数据进行插值，然后用 surfc 函数绘制数据的三维分布和等值线图，如图 4 所示。从图 4 可知，随着工作面的推进，O_2 体积分数持续降低；CO 体积分数从几乎没有增长到中部出现峰值，之后逐渐下降至消失；CO_2 体积分数不断上升，最后趋于稳定，在中部出现较大峰值；CH_4 体积分数在开采过程中一直处于不断上升状态；采空区温度整体呈现先升后降的趋势，并在同一位置，回风侧温度高于进风侧。结合图 2(b)，根据采空区温度变化，可以将其划分为 4 个阶段，各阶段特征和原因分析见表 1。采空区温度和气体呈现出的阶段性特点体现了采空区煤自燃氧化和自热的阶段性，进一步验证了采空区“三带”的存在。

2.2　模型与参数设置

为了对比研究 RF 模型的预测性能，保持与 RF 方法相同训练和测试样本的基础上，利用支持向量机(SVM)和 BP 神经网络(BPNN)方法对煤自燃温度进行训练建模和测试分析。其中，本文所提出的方

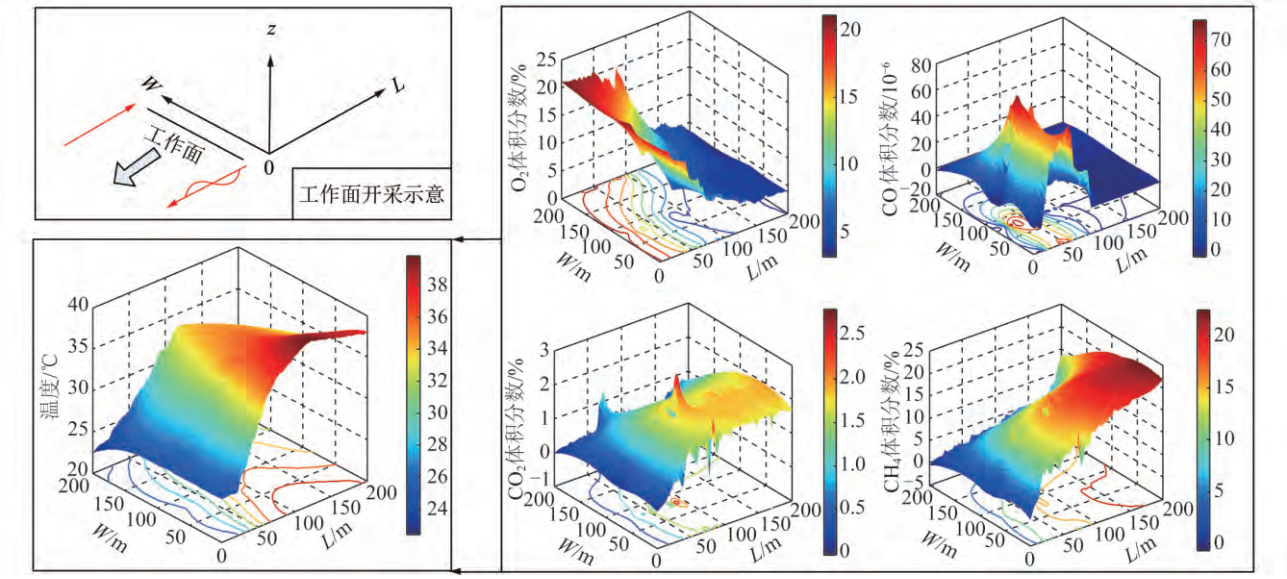

图 4　采空区温度和气体的三维分布

Fig. 4　Three-dimensional distribution of temperature and gases concentration in the gob

表 1　采空区温度分布阶段特征

Table 1　Stage characteristics of temperature distribution in the gob

类别	L/m	采空区温度特征	原因分析
阶段 1	<20	基本趋于稳定	漏风大，顶板冒落程度低，温度难以积聚
阶段 2	20~80	快速上升	采空区漏风和压实程度适当，易于氧化自热
阶段 3	80~120	上升速度减缓，并最终达到最高温度	采空区漏风减少，压实增强
阶段 4	>120	采空区温度开始下降	采空区基本压实，漏风消失

法和模型均基于 Windows8.1 64-bit 系统 MATLAB R2012a 环境下实现。

SVM 是一种基于统计学理论和结构风险化原理的机器学习方法[24]。在支持向量机回归中,输入样本 x 首先通过非线性映射 $\varphi(x)$ 映射到一个高维的特征空间,然后在这个特征空间中构造优化的线性回归函数[25]。在本文中,采用 LIBSVM 工具箱[26]实现 SVM 建模,SVM 核函数选取广泛采用的径向基核函数(RBF)[27]。SVM 回归输出结果为

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x,x_i)+b \tag{8}$$

式中,$f(x)$ 为 SVM 回归函数;α_i 和 α_i^* 为 Lagrange 乘子;$K(x,x_i)$ 为核函数;b 为偏置项。

BPNN 是采用误差反向传播算法的多层前馈网络,是目前使用最普遍的一种神经网络方法[28]。笔者采用标准的 3 层 BPNN,隐层节点数采用经验公式 $2M+1$ 确定,隐层和输出层的传递函数分别为 "tansig"和 "logsig"函数。

由于目前没有统一的标准或者公式对 RF 和 SVM 的超参数进行设置,因此,本文采用粒子群优化算法(PSO)对 RF 和 SVM 的超参数进行优化,建立参数优化的 PSO-RF 和 PSO-SVM 预测模型。

PSO 算法源于对鸟类捕食行为的研究,并应用于求解优化问题。假设在一个 D 维的搜索空间中,由 N_0 个粒子组成的种群 $X=(X_1,X_2,\cdots,X_{N_0})$,其中 $X_i=(x_{i1},x_{i2},\cdots,x_{iD})$ 代表第 i 个粒子在 D 维搜索空间中的位置,亦代表问题的一个潜在解。根据适应度函数可计算出每个粒子位置 X_i 对应的适应度值。第 i 个粒子的速度为 $v_i=(v_{i1},v_{i2},\cdots,v_{iD})$,其个体极值为 $P_i=(P_{i1},P_{i2},\cdots,P_{iD})$,种群的全局极值为 $P_g=(P_{g1},P_{g2},\cdots,P_{gD})$。在每一次迭代过程中,粒子通过个体极值和全局极值更新自身的速度和位置,更新公式为

$$\begin{cases}v_{id}^{k+1}=\omega v_{id}^{k}+c_1r_1(P_{id}^{k}-X_{id}^{k})+c_2r_2(P_{gd}^{k}-X_{id}^{k})\\ X_{id}^{k+1}=X_{id}^{k}+v_{id}^{k+1},(i=1,2,\cdots,N_0;d=1,2,\cdots,D)\end{cases} \tag{9}$$

其中,k 为当前迭代次数;ω 为惯性权重;c_1,c_2 为非负常数,称为加速因子;r_1,r_2 为分布于(0,1)之间的随机数。$c_1=1.5$,$c_2=1.7$,最大迭代次数为 200,种群规模为 20,采用线性递减惯性权重,5 折交叉验证。

根据理论可知,RF 模型超参数主要是 n_{tree} 和 m_{try} 两个参数,笔者采用 Abhishek Jaiantilal 开发的 "randomforest-matlab"工具箱(https://github.com/jrderuiter/randomforest-matlab)建立 RF 模型,其默认参数分别为 $n_{tree}=500$ 和 $m_{try}=M/3$[29]。因此,在优化过程中,超参数范围设置分别为 $n_{tree}\in[1,500]$,$m_{try}\in[1,M]$。RBF 核函数的 SVM 模型主要受惩罚因子 C 和核参数 g 两个参数的影响,为了平衡其泛化性能和预测精度,超参数优化范围分别设置为 $C\in[0.01,100]$,$g\in[0.01,50]$。PSO-RF 和 PSO-SVM 算法描述如下:

Step.1,在搜索范围内初始化粒子位置 X_i 与速度 v_i。

Step.2,将粒子当前位置设置为初始个体极值 P_i,根据适应度函数计算每个粒子的适应度值,其中 PSO-RF 模型的适应度函数为训练样本预测结果的均方根误差,PSO-SVM 模型的适应度函数为训练样本 5 折交叉验证的均方误差;取适应度值最小的粒子对应的个体极值作为最初的全局极值 P_g。

Step.3,更新惯性权重 ω,并根据式(9)更新粒子的速度 v_i 与位置 X_i。

Step.4,计算每次迭代后每个粒子的适应度值。

Step.5,将更新后每个粒子的适应度值与其个体极值的适应度值作比较,如果更优,则更新个体极值,否则保留原值。

Step.6,将更新后的每个粒子的个体极值与全局极值作比较,如果更优,则更新全局极值,否则保留原值。

Step.7,判断是否满足终止条件,若达到最大迭代次数,则终止迭代输出最优参数,否则返回 Step.3。

Step.8,将输出的最优参数赋给 RF/SVM,用于构建最优参数模型。

2.3 结果与讨论

依据上述模型参数设置和算法描述,PSO 对 RF 和 SVM 超参数寻优结果分别为 $n_{tree}=83$,$m_{try}=5$,$C=7.128$,$g=8.224$。各模型训练和测试样本预测结果如图 5 所示,从图 5 可以看出,无论训练样本还是测试样本,RF(默认参数)、PSO-RF 和 PSO-SVM 模型的预测结果基本与零误差线 $y=x$ 重合;SVM 模型(默认参数)的预测结果则都较分散的分布于零误差线 $y=x$ 周围;BPNN 模型训练样本预测结果与零误差线 $y=x$ 吻合很好,但有相当一部分测试样本的预测点偏离零误差线 $y=x$。说明 RF,PSO-RF 和 PSO-SVM 模型预测结果与真实值基本一致,BPNN 模型训练结果很好,但部分测试结果偏差较大,产生了较大的预测误差,然而,SVM 模型训练和测试样本的预测结果均存在较大误差。同时,通过图 6 中模型的绝对百分误差可以看出,RF,PSO-RF 和 PSO-SVM 模型的预测结果均优于 SVM 模型;而 BPNN 模型训练样本获得了最佳的预测精度,但其测试阶段出现了较大的误

差,例如,其训练阶段样本的最大绝对百分误差为 4.41%,但测试阶段样本的最大绝对百分误差高达 15.76%,说明 BPNN 在训练过程中出现了“过拟合”,泛化性差。

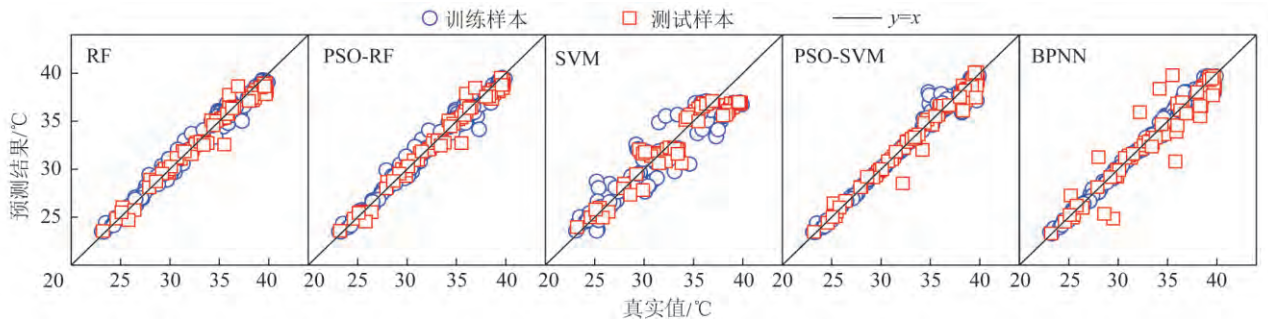

图 5　不同模型训练和测试样本预测结果散点

Fig. 5　Scatter plot of predicted results for different models at the training and testing stages

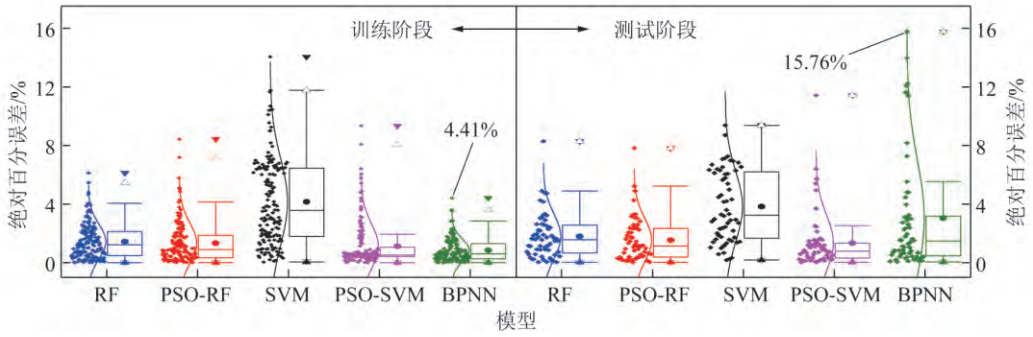

图 6　不同模型训练和测试阶段绝对百分误差箱形

Fig. 6　Box plot of absolute percentage error for different models at the training and testing stages

为了进一步量化对比各个模型的预测性能,表 2 汇总了前文所述的模型性能评估指标。从表 2 可知,不同 m_{try} 值的 RF,PSO-RF,SVM 和 PSO-SVM 模型在测试阶段的 MAE,MAPE,RMSE 和 R^2 与训练阶段的基本没有偏差,均具有较强的泛化性、鲁棒性。然而,BPNN 模型测试样本的预测性能指标 MAE,MAPE 和 RMSE 相对训练样本显著增大,R^2 明显减小,说明 BPNN 模型在训练建模过程中存在“过拟合”现象,导致其泛化性和鲁棒性降低,引起测试样本的误差增大。

表 2　不同模型预测性能指标对比

Table 2　Comparison of predictive performance indicators for different models

模型	m_{try}	性能指标			
		MAE/℃	MAPE/%	RMSE	R^2
RF($n_{tree}=500$)	$m_{try}=1$	0.753^a/0.863^b	2.242^a/2.528^b	0.869^a/1.048^b	0.966 0^a/0.949 8^b
	$m_{try}=2$	0.482^a/0.615^b	1.438^a/1.796^b	0.625^a/0.827^b	0.982 4^a/0.968 7^b
	$m_{try}=3$	0.459^a/0.586^b	1.367^a/1.704^b	0.615^a/0.807^b	0.983 0^a/0.970 2^b
	$m_{try}=4$	0.441^a/0.559^b	1.313^a/1.624^b	0.605^a/0.772^b	0.983 5^a/0.972 7^b
	$m_{try}=5$	0.434^a/0.522^b	1.296^a/1.525^b	0.602^a/0.716^b	0.983 7^a/0.976 5^b
PSO-RF		0.445^a/0.525^b	1.330^a/1.539^b	0.647^a/0.740^b	0.981 2^a/0.974 9^b
SVM		1.410^a/1.327^b	4.158^a/3.825^b	1.733^a/1.606^b	0.864 8^a/0.882 0^b
PSO-SVM		0.393^a/0.461^b	1.126^a/1.343^b	0.682^a/0.797^b	0.979 1^a/0.970 9^b
BPNN		0.286^a/1.008^b	0.841^a/3.035^b	0.403^a/1.626^b	0.992 7^a/0.879 0^b

注: a表示训练样本; b 表示测试样本,下同。

图 7 给出了 RF 建模过程中保持树的数量为默认值不变,即 $n_{tree}=500$,改变参数 m_{try} 值时与其他模

型在测试阶段 MAPE 的对比。随着 m_{try} 的增大，RF 模型的 MAPE 呈现逐渐减小的趋势，当 m_{try} 取值达到 2(默认参数)以后时，MAPE 趋于稳定，说明 RF 模型对其超参数 m_{try} 具有很好的稳定性，m_{try} 的改变对其预测性能影响不大，即使 $m_{try}=1$ 时，也能获得较好的预测结果(MAPE = 2.528%)。然而，SVM 模型预测性能则对其超参数十分敏感，经过 PSO 优化，SVM 的 MAPE 从 3.825%降到了 1.343%，因此，SVM 模型具有突出的泛化性和鲁棒性，但其预测精度依靠超参数的最优选择。

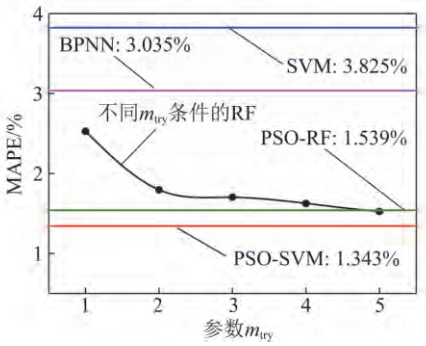

图 7 不同模型测试阶段 MAPE 对比

Fig. 7 Comparison of the MAPE for different models at the testing stage

进一步，根据 RF 模型取不同 m_{try} 值建模过程中对 OOB 数据的预测误差 MSE_{OOB} 和决定系数 R^2_{RF} 变化曲线(图 8)可以看出，煤自燃预测训练过程在 500 棵树(预设值)时停止，当树的数量 n_{tree} 超过 100 时，MSE_{OOB} 和 R^2_{RF} 值基本保持不变，意味着 RF 模型达到了最小误差，也就是在 100 个树之后，OOB 误差保持稳定。因此，当 n_{tree} 达到一定值时，n_{tree} 的改变对 RF 模型预测性能没有实质性的改善，这进一步说明了 RF 模型的稳定性和鲁棒性，对超参数的改变不敏感，具有很宽的参数适应范围。同时，图 8 的结果与 PSO 寻优结果 $n_{tree}=83$ 一致，结合图 7 和表 2，PSO-RF 模型的 MAPE 值(1.539%)基本与 RF 模型最佳预测结果的 MAPE 值(1.525%)重合，从而证实了 PSO 优化 RF 参数的有效性，PSO 方法可以找到 RF 模型最优参数。

3 应用分析

为了研究随机森林方法的普适性和推广应用性，以 XIE 等[30]文中的数据为研究对象，对比分析模型在其他矿井应用中的预测性能。文献[30]中 XIE 等在 Longdong 煤矿 7162 综采工作面通过束管监测系统监测采空区温度和气体信息，采用温升和 CO 体积分数划分煤自燃“三带”。本文整理文献[30]中有效

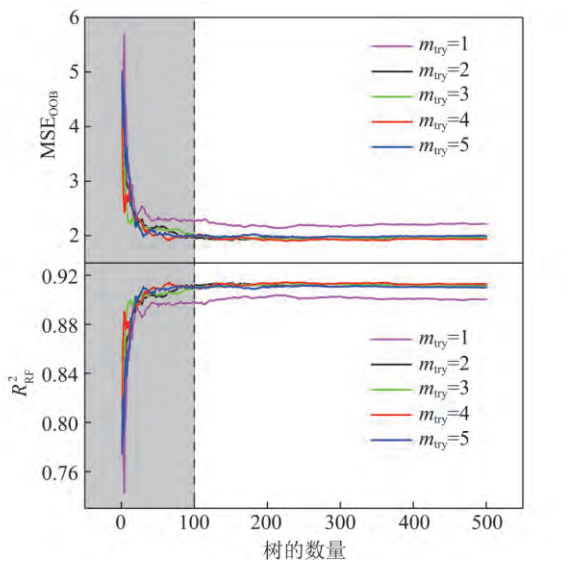

图 8 不同 m_{try} 下的 OOB 数据预测误差和相应的决定系数

Fig. 8 Prediction error and corresponding determination coefficient of OOB data with different m_{try}

数据 60 组，以测点距工作面距离，O_2 和 CO 体积分数为输入变量，采空区温度为目标输出，其中 40 组数据用于训练建模，剩余的 15 组(总样本数的 1/4)用于模型测试。RF 采用超参数 $n_{tree}=500$ 和 $m_{try}=2$ 的设置；PSO 对 SVM 超参数优化范围及 BPNN 模型参数设置保持与前文一致。最终，PSO 优化 SVM 超参数结果为 $C=5.271$，$g=3.711$。

各模型预测结果如图 9 所示，RF，PSO-SVM 和 BPNN 模型在测试和训练阶段的预测结果都紧密分布在零误差线 $y=x$ 周围；SVM 模型训练样本预测结果有一部分偏离零误差线 $y=x$，主要是温度大于 50 ℃的样本误差较大。同时，模型性能指标进一步量化了上述特征，见表 3，经过 PSO 优化，SVM 预测精度显著提高，PSO-SVM 测试样本预测结果的 MAPE 仅为 1.451%，MAE，RMSE 和 R^2 均远远优于 SVM 模型。BPNN 模型则依旧体现出前文所述的特征，在训练建模过程中发生了“过拟合”现象，这与其训练误差最小化的基本原理密切相关。RF 不仅体现出极强的泛化性，而且能实现采空区煤自燃温度的准确预测，测试样本预测结果的 MAPE 仅为 1.842%，说明 RF 方法具有较强的稳定性和普适性，在应用中，无需复杂的参数设置与优化就能获得满意的预测性能。RF 方法优越的非线性处理能力主要由其特殊结构决定的，RF 由一系列非线性处理器(决策树)组合而成，相当于组合很多的非线性关系形成更加复杂的非线性关系处理器。对于处理采空区煤自燃温度与气体之间的非线性关系，RF 是一种简单准确、稳定可靠的方法。

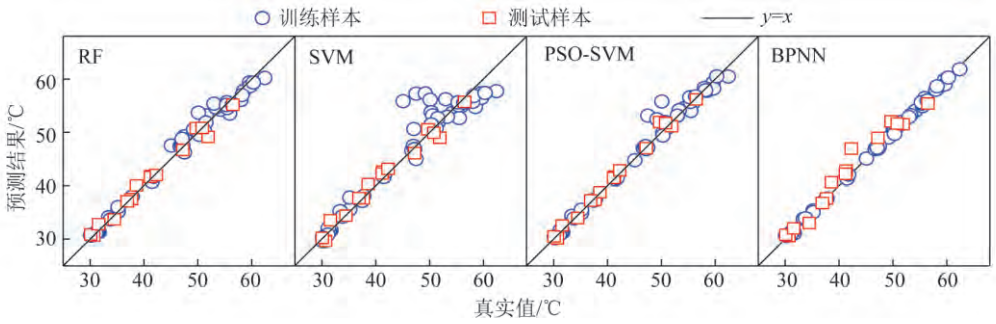

图 9　不同模型在应用时预测结果散点

Fig. 9　Scatter plot of predicted results for different models in the application

表 3　不同模型在应用时预测性能指标对比

Table 3　Comparison of predictive performance indicators for different models in the application

模型	性能指标			
	MAE/℃	MAPE/%	RMSE	R^2
RF	0.905[a]/0.774[b]	1.839[a]/1.842[b]	1.207[a]/1.037[b]	0.985 0[a]/0.983 2[b]
SVM	1.990[a]/1.044[b]	4.070[a]/2.528[b]	3.174[a]/1.265[b]	0.8965[a]/0.974 9[b]
PSO-SVM	0.769[a]/0.599[b]	1.588[a]/1.451[b]	1.431[a]/0.784[b]	0.979 0[a]/0.990 4[b]
BPNN	0.250[a]/1.246[b]	0.556[a]/2.957[b]	0.322[a]/1.696[b]	0.998 9[a]/0.955 0[b]

4　结　　论

(1) 引入 RF 方法预测采空区煤自燃温度，并将预测结果与 SVM 和 BPNN 方法作比较。提出采用 PSO 算法对 RF 和 SVM 的超参数进行优化，建立参数优化的 PSO-RF 和 PSO-SVM 预测模型。

(2) RF 在建模过程中具有宽广的参数适应范围，虽然 PSO 算法能够找到其最优参数，但默认参数的 RF 模型就能获得满意的预测性能。在树的数量 n_{tree} 超过 100 以后，RF 模型训练误差趋于稳定，n_{tree} 的增加对其预测结果没有实质性的改变。

(3) SVM 具有较强的泛化性，但 SVM 预测结果对其超参数十分敏感，PSO 优化能够显著提高其预测精度，SVM 预测性能依赖于超参数的最佳选择; BPNN 拥有最佳的训练结果，但其训练误差最小化的基本原理导致训练过程容易陷入“过拟合”，引起大的预测误差，泛化性弱。

(4) 其他矿井应用分析表明，本文引入的 RF 方法具有普适性，用于采空区煤自燃预测是准确可靠的，预测结果与实际情况相吻合，并且无需复杂的参数设置与优化，可进一步推广应用于其他能源燃料领域。

参考文献(References) :

[1] SONG Z Y，KUENZER C. Coal fires in China over the last decade: A comprehensive review [J]. International Journal of Coal Geology，2014，133: 72-99.

[2] SAFFARI A，SERESHKI F，ATAEI M，et al. Presenting an engineering classification system for coal spontaneous combustion potential [J]. International Journal of Coal Science & Technology，2017，4(2) : 110-128.

[3] DENG J，LEI C K，XIAO Y，et al. Determination and prediction on “three zones” of coal spontaneous combustion in a gob of fully mechanized caving face [J]. Fuel，2018，211: 458-470.

[4] 李林，陈军朝，姜德义，等. 煤自燃全过程高温区域及指标气体时空变化实验研究 [J]. 煤炭学报，2016，41(2) : 444-450.
LI Lin，CHEN Junchao，JIANG Deyi，et al. Experimental study on temporal variation of high temperature region and index gas of coal spontaneous combustion [J]. Journal of China Coal Society，2016，41(2) : 444-450.

[5] HU X C，YANG S Q，ZHOU X H，et al. Coal spontaneous combustion prediction in gob using chaos analysis on gas indicators from upper tunnel [J]. Journal of Natural Gas Science and Engineering，2015，26: 461-469.

[6] 邓军，李贝，李珍宝，等. 预报煤自燃的气体指标优选试验研究 [J]. 煤炭科学技术，2014，42(1) : 55-59.
DENG Jun，LI Bei，LI Zhenbao，et al. Experiment study on gas indexes optimization for coal spontaneous combustion prediction [J]. Coal Science and Technology，2014，42(1) : 55-59.

[7] DENG J，MA X F，ZHANG Y T，et al. Effects of pyrite on the spontaneous combustion of coal [J]. International Journal of Coal Science & Technology，2015，2(4) : 306-311.

[8] 王磊，武术静，李长青. 基于修正灰色马尔科夫模型的煤自燃预测 [J]. 计算机仿真，2014，31(11) : 416-420.
WANG Lei，WU Shujing，LI Changqing. Coal spontaneous combustion prediction based on grey-Markov model [J]. Computer Simulation，2014，31(11) : 416-420.

[9] 邬剑明,王俊峰.基于神经网络的煤层自然发火的非线性预测[J].中国安全科学学报,2004,14(5):15-17.
WU Jianming,WANG Junfeng.A non-linear prediction of coal self-ignition based on neural network [J].China Safety Science Journal,2004,14(5):15-17.

[10] 周福宝,李金海.煤矿火区启封后复燃预测的 BP 神经网络模型[J].采矿与安全工程学报,2010,27(4):494-498,504.
ZHOU Fubao,LI Jinhai.Prediction model for reignition of fire zone after unsealing based on BP neural networks [J].Journal of Mining and Safety Engineering,2010,27(4):494-498,504.

[11] 靳玉萍.基于代数神经网络的煤自燃预测[J].数学的实践与认识,2013,43(18):122-128.
JIN Yuping.Forecasting of coal spontaneous combustion based on algebra neural networks [J].Mathematics in Practice and Theory,2013,43(18):122-128.

[12] 高原,覃木广,李明建.基于支持向量机的采空区遗煤自燃预测分析[J].煤炭科学技术,2010,38(2):50-54.
GAO Yuan,QIN Muguang,LI Mingjian.Analysis on prediction of residual coal spontaneous combustion in goaf based on support vector machine [J].Coal Science and Technology,2010,38(2):50-54.

[13] 孟倩,王永胜,周延.基于粗集-支持向量机的采空区自然发火预测[J].煤炭学报,2010,35(12):2100-2104.
MENG Qian,WANG Yongsheng,ZHOU Yan.Prediction of spontaneous combustion in caving zone based on rough set and support vector machine [J].Journal of China Coal Society,2010,35(12):2100-2104.

[14] 邵良杉,李相辰.不平衡数据下的采空区自然发火预测研究[J].中国安全科学学报,2017,27(6):61-66.
SHAO Liangshan,LI Xiangchen.Prediction of coal spontaneous combustion in caving zone with unbalanced data [J].China Safety Science Journal,2017,27(6):61-66.

[15] 靳玉萍,张兵,高凯.球形支持向量机在煤自燃预测中的应用[J].计算机应用与软件,2013,30(9):57-60.
JIN Yuping,ZHANG Bing,GAO Kai.Application of spheical-SVM in forecasting spontaneous combustion of coal [J].Computer Applications and Software,2013,30(9):57-60.

[16] 邓军,周少柳,马砺,等.基于 PCA-PSOSVM 的煤自燃程度预测研究[J].矿业安全与环保,2016,43(5):27-31.
DENG Jun,ZHOU Shaoliu,MA Li,et al.Research on prediction method of coal spontaneous combustion degree based on PCA-PSOSVM [J].Mining Safety and Environmental Protection,2016,43(5):27-31.

[17] 邓军,雷昌奎,曹凯,等.煤自燃预测的支持向量回归方法[J].西安科技大学学报,2018,38(2):175-180.
DENG Jun,LEI Changkui,CAO Kai,et al.Support vector regression approach for predicting coal spontaneous combustion [J].Journal of Xi'an University of Science and Technology,2018,38(2):175-180.

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

[19] MATIN S S,CHELGANI S C.Estimation of coal gross calorific value based on various analyses by random forest method [J].Fuel,2016,177:274-278.

[20] CHELGANI S C,MATIN S S,MAKAREMI S.Modeling of free swelling index based on variable importance measurements of parent coal properties by random forest method [J].Measurement,2016,94:416-422.

[21] CHELGANI S C,MATIN S S,HOWER J C.Explaining relationships between coke quality index and coal properties by random forest method [J].Fuel,2016,182:754-760.

[22] MATIN S S,HOWER J C,FARAHZADI L,et al.Explaining relationships among various coal analyses with coal grindability index by random forest [J].International Journal of Mineral Processing,2016,155:140-146.

[23] BREIMAN L,FRIEDMAN J H,OLSHEN R A,et al.Classification and regression trees [M].Belmont (California):Wadsworth International Group,1984.

[24] CORTES C,VAPNIK V.Support-vector networks [J].Machine Learning,1995,20:273-297.

[25] 陈同俊,王新,管永伟.基于 SVR 和地震属性的构造煤厚度定量预测[J].煤炭学报,2015,40(5):1103-1108.
CHEN Tongjun,WANG Xin,GUAN Yongwei.Quantitative prediction of tectonic coal seam thickness using support vector regression and seismic attributes [J].Journal of China Coal Society,2015,40(5):1103-1108.

[26] CHANG C C,LIN C J.LIBSVM:A library for support vector machines [J].ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.

[27] 施龙青,谭希鹏,王娟,等.基于 PCA_Fuzzy_PSO_SVC 的底板突水危险性评价[J].煤炭学报,2015,40(1):167-171.
SHI Longqing,TAN Xipeng,WANG Juan,et al.Risk assessment of water inrush based on PCA_Fuzzy_PSO_SVC [J].Journal of China Coal Society,2015,40(1):167-171.

[28] 李慧民,李振雷,何荣军,等.基于粒子群算法和 BP 神经网络的冲击危险性评估[J].采矿与安全工程学报,2014,31(2):203-207,231.
LI Huimin,LI Zhenlei,HE Rongjun,et al.Rock burst risk evaluation based on particle swarm optimization and BP neural network [J].Journal of Mining and Safety Engineering,2014,31(2):203-207,231.

[29] LIAW A,WIENER M.Classification and regression by random forest [J].R News,2002,2:18-22.

[30] XIE Z H,CAI J,ZHANG Y.Division of spontaneous combustion "three-zone" in goaf of fully mechanized coal face with big dip and hard roof [A].2012 international symposium on safety science and engineering in China [C].Procedia Engineering,2012:82-87.