DOI:10.19651/j.cnki.emt.2209216

基于改进麻雀算法优化深度极限学习机的缺失数据预测*

张文帅 王占刚
(北京信息科技大学信息与通信工程学院 北京 100020)

摘 要: 数据缺失降低了数据的可利用性,因此如何预测缺失数据变得尤为重要。针对缺失数据问题,提出一种改进麻雀搜索算法优化深度极限学习机的预测算法。首先,将 Singer 混沌映射、柯西-高斯变异策略和余弦权重因子与麻雀搜索算法结合;其次利用改进后的麻雀搜索算法优化深度极限学习机中的各极限学习机中自动编码器的输入权重与偏置,进行缺失数据预测。实验表明,在小数据量,低缺失率下时,改进麻雀搜索算法优化深度极限学习机相较于麻雀搜索算法优化深度极限学习机、粒子群优化深度极限学习机、深度极限学习机,其稳定性强,预测精度最高;在均方根误差、平均绝对误差等评价指标上改进麻雀搜索算法优化深度极限学习机优于对比算法。

关键词: 缺失数据预测;深度极限学习机;麻雀搜索算法;混沌映射;变异策略
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 520.20

Missing data prediction based on improved sparrow algorithm optimized deep extreme learning machine

Zhang Wenshuai Wang Zhangang
(School of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing 100020, China)

Abstract: Missing data reduces data availability. Prediction of missing data becomes very important. A prediction algorithm named ISSA-DELM(Improved Sparrow Search Algorithm optimized Deep Extreme Learning) was proposed to solve the problem of missing data. First of all, singer chaotic map, Cauchy-Gaussian mutation strategy and cosine weight factor combined with sparrow search algorithm. Secondly, the input weights and biases of the autoencoders in each extreme learning machine in the deep extreme learning machine are optimized by ISSA. Then ISSA-DELM is applied to predict missing data. The experimental results show that, ISSA-DELM has strong stability and the highest prediction accuracy compared with SSA-DELM、Particle Swarm Optimization DELM(PSO-DELM)、DELM in the case of small data volume and low miss rate. The evaluation indexes, such as RMSE, MAE and the coefficient of determination are better than the compared algorithms.

Keywords: missing data prediction; deep extreme learning machine; sparrow search algorithm; chaotic mapping; mutation strategy

0 引 言

云计算、人工智能、大数据等技术的快速发展,为数据分析提供了强大的技术支撑,然而在数据挖掘分析中,数据缺失的问题无法避免,导致数据分析等一系列工作受到影响。在小数据集背景下,数据缺失问题将会带来更加严重的后果,这在生态环境领域尤为突出。在生态环境领域数据缺失的原因有:数据采集困难、人为因素、设备故障等。因此,如何在生态环境领域对缺失数据进行高效预测成为数据处理阶段亟待解决的问题。

目前,国内外学者们针对数据缺失问题提出许多预测方法,并将其应用到各个领域中。一类是基于统计学方法,如邓子畏等提出一种基于随机期望最大化算法的缺失数据预测算法用于混凝土泵车数据预测[1];Bashir 等[2]提出一

收稿日期:2022-03-09
* 基金项目:国家重点研发计划课题资助(2018YFC1800203)、北京市科技创新服务能力建设-基本科研业务费(市级)(科研类)(PXM2019_014224_000026)资助

种基于矢量自回归模型，将期望最小化算法与预测误差最小化方法结合在一起进行数据猜测，上述的方法往往未考虑数据对象本身的类别，导致预测结果的准确性较差。Gondara 等[3]提出了一种基于超完全深度去噪自动编码器的多重填补模型进行缺失数据填充，该方法需要缺失比例不高，比较多完整数据进行模型训练。另一种是基于机器学习方法，如马创等[4]使用基于遗传算法与支持向量机进行了水质缺失数据预测，但是所提算法对参数选择核函数较为敏感；卢森骧等[5]利用加权随机森林算法填补了单轴信号的缺陷并通过模糊推理系统完成三轴信号反演融合，最终预测出较为准确的缺陷尺寸，然而该方法有点过于复杂；还有学者将去噪自编码器与生成对抗网络结合用来处理缺失率较高的工业物联网数据，但该模型无法高精度进行填充[6]。

相较于当前缺失数据预测方法，深度极限学习机(deep extreme learning，DELM)训练速度快，泛化性好，可以很好地提取非线性数据特征。Li 等[7]利用深度极限学习机预测氮氧化合物的含量，证明了深度极限学习机的有效性；吐尔逊·买买提等[8]将 DELM 应用到柴油机尾气排放预测，相较于支持向量机和 BP 神经网络，DELM 的鲁棒性和适应性更强。但是深度极限学习机存在随机输入权重和偏置影响 DELM 的训练效果，且易陷入局部最优的问题。

针对以上分析，本文提出了改进麻雀搜索(improved sparrow search algorithm，ISSA)优化深度极限学习机预测算法—ISSA-DELM。通过改进麻雀搜素算法使个体地分布更均匀，提升寻优精度，增强全局寻优能力，利用 ISSA 优化 DELM 的输入权重和偏置，解决 DELM 易陷入局部最值的问题，增强 DELM 预测的稳定性和精确度，将完整数据作为整体输入 ISSA-DELM，对缺失数据进行预测，以期为生态环境领域的缺失数据预测提供一种精度高、稳定性强的新方法。

1　缺失数据预测算法

1.1　改进麻雀搜索算法(ISSA)

麻雀搜索算法(sparrow search algorithm，SSA)[9-10]虽然寻优能力不错，但仍然存在易陷入局部最优等问题。通过改进算法种群的初始化、权重因子与自适应扰动三个方面，从而增强种群的多样性、提升寻优精度，提高全局寻优能力。

1)Singer 混沌映射初始化

在原始 SSA 中，个体的位置通过随机方式初始化产生，导致个体位置分布不均匀，会降低求解精度。而基于混沌理论的混沌序列具有伪随机性和边界性等特点[11]。因此，本文采用 Singer 混沌映射对 SSA 进行初始化：

$$x_{k+1}=\mu(7.86\cdot x_k-23.31\cdot x_k^2+28.75\cdot x_k^3-13.302\,875\cdot x_k^4) \tag{1}$$

其中，x_{k+1} 为引入 Singer 混沌映射的值，$\mu\in(0.9,1.08)$，本文中 μ 取 1。

2)余弦权重因子

SSA 能否找到最优解在很大程度上取决于搜索发现者的能力。个体在搜索范围内的位置是随机分布的。当发现者附近没有相邻的麻雀时，会进行随机搜索。这种方式不仅使收敛速度变慢，也降低收敛精度。为了进一步提高寻优性能，在发现者的位置更新时加入余弦权重因子[12]，其更新公式变为式(2)：

$$X_{i,m}^{t+1}=\begin{cases}X_{i,m}^{t}\cdot\exp\left(-\dfrac{i}{\omega\alpha\cdot iter_{\max}}\right), & R_2<ST\\ X_{i,m}^{t}+Q\cdot\boldsymbol{L}, & R_2<ST\end{cases} \tag{2}$$

其中，t 为当前迭代次数；$iter_{\max}$ 为最大迭代次数；$\omega=\cos^2\dfrac{\pi\cdot t}{2\cdot iter_{\max}}$ 是随迭代次数增加而自适应减少的余弦权重因子。前期时，ω 的值偏大，算法搜索范围较大，有利于跳出局部极值。随着迭代次数的增加，ω 逐渐变小，此时进行精细搜索，这是前期搜索经验对后期搜索的帮助和支持的作用，从而可以提高算法的寻优精度与稳定性。

3)柯西-高斯变异策略

原始的 SSA 位置更新依赖于每次迭代更新后的结果，下一代个体主要通过选择最优适应度进行更新，由于未对个体进行主动的扰动更新，导致算法易陷入局部最优，种群多样性低和算法收敛过早。为了解决这一问题，本文引入了柯西-高斯变异策略[13]，经过柯西-高斯变异扰动后，个体位置的新解可以用以下公式表示。

$$N_{best}^{t}=X_{best}^{t}[1+\lambda_1 cauchy(0,\sigma^2)+\lambda_2 Gauss(0,\sigma^2)] \tag{3}$$

$$\lambda_1=1-\frac{t}{iter_{\max}}\quad \lambda_2=\frac{t}{iter_{\max}} \tag{4}$$

$$\sigma=\begin{cases}1 & f(X_{best}^{t})<f(X_{im}^{t})\\ \exp\left(\dfrac{f(X_{best}^{t})-f(X_{im}^{t})}{f(X_{best}^{t})}\right) & otherwise\end{cases} \tag{5}$$

其中，N_{best}^{t} 为经过柯西-高斯变异策略扰动后的最新位置，σ^2 表示柯西-高斯变异策略的标准差。$cauchy(0,\sigma^2)$ 是一个满足柯西分布的随机变量，$Gauss(0,\sigma^2)$ 是一个满足高斯分布的随机变量。如式(4)所示，在初始阶段 λ_1 较大，使得算法可以在较大的变异步长范围内求最优解。λ_2 的变异步长较小，便于算法在最优解附近求解。随着迭代次数的增加，λ_1 逐渐减小，而 λ_2 不断增大。

1.2　深度极限学习机(DELM)

2004 年，Huang 等[14]提出了一种极限学习机(extreme learning machine，ELM)，ELM 是一种单隐层前馈神经网络，能够应用在监督学习和非监督学习[15]。ELM 的输出为：

$$f(x_i)=\sum_{i=1}^{l}\boldsymbol{\beta}_i g(\boldsymbol{w}_i\cdot x_j,b_i);j=1,2,\cdots,N \tag{6}$$

其中，$\boldsymbol{w}_i=(w_{i1},w_{i2},\cdots,w_{im})^{\mathrm{T}}$ 为输入层和隐含层之

间的权值；$\boldsymbol{\beta}_i=(\beta_{i1},\beta_{i2},\cdots,\beta_{in})^{\mathrm{T}}$ 为隐含层与输出层之间的权值，$\boldsymbol{b}_i$ 是第 i 个隐含层的偏置，$g(\cdot)$ 为隐层激活函数。ELM 矩阵表达式为式(8)。

$$\boldsymbol{H\beta}=\boldsymbol{T} \tag{7}$$

为得到 ELM 的训练结果，通过最小二乘解的 $\hat{\boldsymbol{\beta}}$，使：

$$\|\boldsymbol{H}\hat{\boldsymbol{\beta}}-\boldsymbol{T}\|=\min_{\beta}\|\boldsymbol{H\beta}-\boldsymbol{T}\| \tag{8}$$

其中，$\boldsymbol{H}$ 为 ELM 的隐层输出矩阵；$\boldsymbol{T}$ 为样本期望输出矩阵。输出权值可以确定为式(10)。

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{9}$$

其中，$\boldsymbol{H}^{\dagger}$ 为输出矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆矩阵。

由于 ELM 为单隐含层结构，当输入的数据维度过时，ELM 获取数据特征变得困难。ELM 随机产生权重和阈值，可能导致部分神经元无法发挥作用，降低了对数据特征的学习能力。因此提出深度极限学习机(DELM)，DELM 有着较高的泛化能力，可以解决 ELM 对数据特征的学习能力低的问题。DELM[16-17]是由多个极限学习机自编码器堆叠而成的深度网络。ELM 自编码器就在极限学习机的基础上，增加了一个正则化项。

$$\boldsymbol{\beta}=\boldsymbol{H}^{\mathrm{T}}\left(\frac{1}{C}+\boldsymbol{HH}^{\mathrm{T}}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \tag{10}$$

其中，$\boldsymbol{\beta}=[\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\cdots\boldsymbol{\beta}_n]$，$C$ 为正则项系数。将 $\boldsymbol{\beta}^{\mathrm{T}}$ 作为原网络结构的输入层与隐含层的权值矩阵，可以完成对 ELM-AE 的单层训练。深度极限学习机自编码器(DELM)利用多个 ELM-AE 堆叠进行计算，构建了含多层隐含层的网络结构。

1.3　改进 SSA-DELM 算法

由于各个 ELM-AE 的随机输入权重和偏置会影响 DELM 的训练效果，故引入了改进的麻雀搜索算法来优化 DELM 的参数，以深度极限学习机(deep extreme learning machine，DELM)为基础，利用改进的麻雀搜索算法(ISSA)迭代寻优 DELM 的预训练过程中的各 ELM-AE 输入权重与偏置，以此来提升 DELM 的精确度。ISSA-DELM 算法结合了 ISSA 算法的全局搜索能力和局部开发能力及 DELM 的非线性特征随机映射能力，能够避免模型陷入局部最值。

ISSA-DELM 的具体步骤为：

1) 确定算法的输入与输出，并将数据分为完备数据集和缺失数据集进行归一化操作；

2) 初始化 ISSA 与 DELM 相关参数，如种群数量 N、最大迭代次数 M、发现者比例 Pnum、侦察者比例 Snum、预警值 等，并利用式(1)Singer 混沌映射初始化麻雀种群；

3) 对计算好的种群适应度值进行排序选出当前最优值和最差值；

4) 按照式(2)更新发现者的位置、加入者的位置以及意识到危险的麻雀的位置；

5) 选择柯西-高斯变异策略对当前最优解进行自适应扰动，按照式(3)进行自适应扰动，产生新解；

6) 对当前最优值和新解进行比较，从而确定是否进行位置更新；

7) 进行迭代操作直到达到最大迭代次数，得到全局最优值和最佳适应度值；

8) 得到全局最优值和最佳适应度值作为 DELM 算法的输入，利用训练集不断训练 DELM 网络，直到总误差小于期望误差；

9)对缺失数据集进行预测，输出预测结果。

1.4　参数设置

ISSA-DELM 算法的激活函数设置为 sigmoid；极限学习机自编码器中的正则化系数(C)为 2；麻雀搜索算法中的种群规模(N)为 10；最大迭代次数 M＝100、发现者数(Pnum)、跟随者数(Snum)均为 20%。

在深度学习中，构建的算法容易受隐含层层数及每层神经元个数的影响。通过设置 DELM 的层数分别为 2 层、3 层、4 层、5 层，从而设定合适的隐含层层数。神经元的多少会影响算法的预测结果的精度。通过查阅文献以及经验选择 5 个范围：(5,10)，(5,20)，(5,30)，(5,40)，(5,50)来确定多层隐含层中最合适的神经元个数。

表 1 中的各项指标的数值都经过 20 次的实验选取了各项的平均值，其主要目的减小算法的偶然性。由表 1 可知，当隐含层数为 4 层时，其各项指标最优，隐含层数为 2 层时，指标较差，3 层和 5 层的指标差异较小，当隐含层数为 4 层，神经元数范围为(5，40)时，模型拟合度最高；DELM 的隐含层层数及神经元数并不是越多越好，故通过实验确定深度极限学习机(DELM)的隐含层层数为 4 层，神经元个数范围为(5,40)。

表 1　不同隐含层数和神经元个数范围测试

隐含层	指标	5～10	5～20	5～30	5～40	5～50
2 层	RMSE	2.369	2.341	2.337	2.332	2.352
	MAE	1.794	1.711	1.718	1.798	1.713
	R^2	0.935	0.935	0.936	0.936	0.935
3 层	RMSE	2.144	2.119	2.126	2.131	2.138
	MAE	1.752	1.733	1.7	1.708	1.732
	R^2	0.946	0.945	0.946	0.947	0.945
4 层	RMSE	2.138	1.983	1.976	1.957	1.960
	MAE	1.678	1.623	1.616	1.590	1.594
	R^2	0.952	0.951	0.951	0.958	0.954
5 层	RMSE	2.107	2.344	2.376	2.275	2.285
	MAE	1.698	1.947	1.923	1.854	1.802
	R^2	0.948	0.940	0.945	0.945	0.947

2　实验结果分析

为证明 ISSA-DELM 在生态环境领域对缺失数据预测

的优越性，并为生态环境领域处理数据缺失提供一种新的方法。本文选取了生态领域中数据缺失现象较为常见的空气数据和土壤重金属数据作为实验数据集。实验将在华北地区国控站点的空气污染指数数据集和华北某地区的土壤重金属污染数据集进行，分别与 SSA-DELM、PSO-DELM、DELM 预测模型进行对比。不同算法独立运行 20 次，将 20 次预测结果的平均值作为最终评价结果。实验计算得到各模型预测结果与真实值的误差。为进一步验证各模型性能，实验分析得到了各算法预测结果的标准误差、平均绝对误差及 R^2。

2.1　土壤重金属含量数据集

土壤重金属含量预测缺失数据的准确率及预测的 R^2 如表 2 所示。由表 2 可知，在不同缺失率下，ISSA-DELM 相较与 SSA-DELM、PSO-DELM、DELM，其预测准确率高于其他预测模型。其中在 10% 的数据缺失率时，ISSA-DELM 的预测准确率与 SSA-DELM、PSO-DELM、DELM 相比分别提升了 4.54%，5.24%，8.76%。ISSA-DELM 相较于其他算法，其 R^2 最高，说明 ISSA-DELM 算法的拟合效果最佳。

表 2　预测结果对比

指标	方法	缺失率				
		5%	10%	15%	20%	25%
准确率	ISSA-DELM	0.943	0.944	0.932	0.92	0.909
	SSA-DELM	0.902	0.903	0.895	0.89	0.884
	PSO-DELM	0.895	0.897	0.889	0.884	0.879
	DELM	0.861	0.868	0.863	0.858	0.849
R^2	ISSA-DELM	0.957	0.968	0.964	0.938	0.936
	SSA-DELM	0.913	0.926	0.917	0.893	0.905
	PSO-DELM	0.852	0.878	0.877	0.857	0.875
	DELM	0.769	0.837	0.848	0.808	0.839

值得注意的是，在一些土壤采样点周围存在直接污染企业的排放，会导致该企业周围的采样点会比其他采样点高出许多，导致数据具有突变特征。在实验中发现，本文所提的 ISSA-DELM 模型对突变特征数据进行缺失数据预测依然能得到不错的效果。由此可证明 ISSA-DELM 可应用在有突变特征的数据中。

此外，还计算出不同预测方法的 RMSE 与 MAE 并进行对比，为了更加直观地比较每个方法在不同缺失率下的效果，图 1 和 2 所示分别为 RMSE 和 MAE 的对比，显然，在不同缺失率下，ISSA-DELM 的 RMSE 和 MAE 均为最小，其中在 10% 的缺失率时，ISSA-DELM 的指标最优，由此可证明本文所提方法的有效性。

2.2　空气质量指数数据集

在空气质量指数数据集中，主要对数据中的 PM2.5 的缺失值进行预测。预测结果的对比如表 3 所示。由表 3 可知，在不同缺失率下，ISSA-DELM 的预测准确率和 R^2 均为最优，其

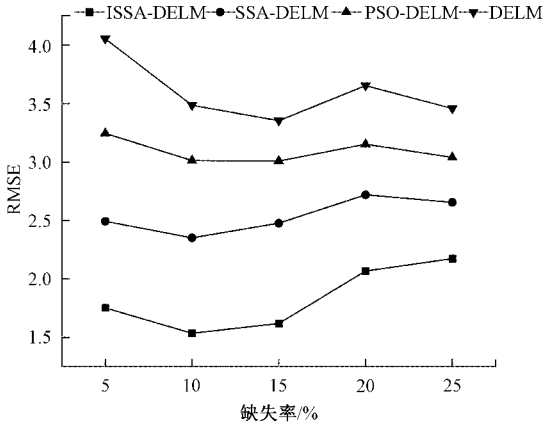

图 1　土壤数据 RMSE 比较

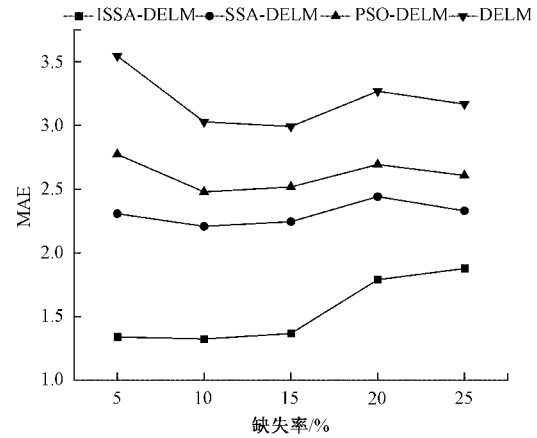

图 2　土壤数据 MAE 比较

中随着缺失率的增加，准确率不断下降，R^2 不断减小，由此可知，在缺失率较小时，ISSA-DELM 有不错的预测效果。

表 3　预测结果对比

指标	方法	缺失率				
		5%	10%	15%	20%	25%
准确率	ISSA-DELM	0.925	0.919	0.904	0.902	0.89
	SSA-DELM	0.892	0.887	0.852	0.858	0.851
	PSO-DELM	0.873	0.876	0.85	0.851	0.845
	DELM	0.846	0.843	0.833	0.834	0.828
R^2	ISSA-DELM	0.967	0.963	0.949	0.915	0.863
	SSA-DELM	0.942	0.896	0.739	0.766	0.718
	PSO-DELM	0.903	0.928	0.765	0.727	0.693
	DELM	0.883	0.805	0.759	0.712	0.681

图 3 和 4 分别为不同缺失率下，ISSA-DELM、SSA-DELM、PSO-DELM、DELM 模型预测结果的 RMSE 和 MAE 经过对比，由图 3 和 4 可知，ISSA-DELM 的 RMSE 和 MAE 均为最小，随着缺失率的不断增加，其 RMSE 和 MAE 的值也不断增加。整体而言，相较于其他预测方法，ISSA-DELM 对缺失数据的预测效果最好。

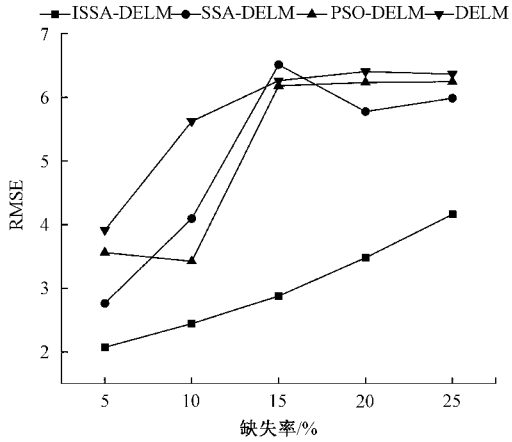

图 3 空气数据 RMSE 比较

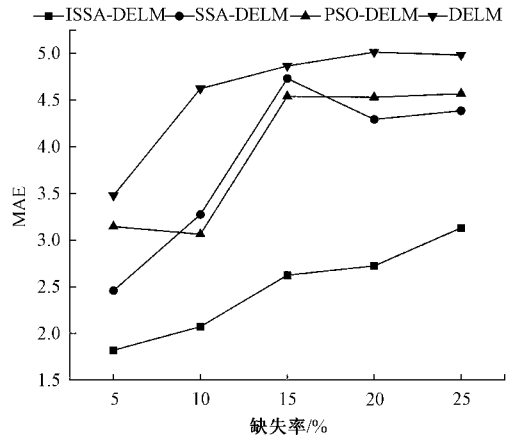

图 4 空气数据集 MAE 比较

3 结 论

本文通过 ISSA 优化 DELM,提出了新的 ISSA-DELM 预测算法。实验结果表明该算法可以增强 DELM 预测的稳定性和精确度,解决现有预测模型易陷入局部最值的问题,有着不错的非线性特征随机映射能力,ISSA-DELM 具有更强的泛化能力,能够加快收敛。这说明 ISSA-DELM 对缺失数据的预测具有一定的工程应用价值,是一种有效的缺失数据预测方法,可以为生态环境领域缺失数据预测提供了新思路。

参考文献

[1] 邓子畏, 唐朝晖, 朱红求,等. 基于改进 EM 算法的混凝土泵车数据治理[J]. 中南大学学报(自然科学版), 2021, 52(2):443-449.

[2] BASHIR F, WEI H L. Handling missing data in multivariate time series using a vector autoregressive model-imputation algorithm [J]. Neurocomputing, 2018, 276: 23-30.

[3] GONDARA L, WANG K. Mida: Multiple imputation using denoising autoencoders [C]. Pacific-Asia conference on knowledge discovery and data mining, Springer, Cham, 2018: 260-272.

[4] 马创,王尧,李林峰. 基于遗传算法与支持向量机的水质预测模型[J]. 重庆大学学报,2021,44(7):108-114.

[5] 卢森骧,神祥凯,张俊楠,等. 基于三轴融合的漏磁内检测数据缺陷反演方法研究[J]. 仪器仪表学报,2021,42(12):245-253.

[6] WANG H, YUAN Z, CHEN Y, et al. An industrial missing values processing method based on generating model[J]. Computer Networks, 2019, 158: 61-68.

[7] LI Y, LI F. NOx prediction method based on deep extreme learning machine[C]. 2018 3rd International Conference on Computational Intelligence and Applications(ICCIA), 2018.

[8] 吐尔逊·买买提,赵梦佳,宁成博,等. 基于深度极限学习机的柴油机尾气排放预测[J]. 科学技术与工程, 2021,21(36), DOI:10.3969/j.issn.1671-1815.2021.36.046.

[9] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[10] 戈一航,杨光永,徐天奇,等. 基于 SSA 优化 PID 在移动机器人路径跟踪中的研究[J]. 国外电子测量技术, 2021,40(9):64-69.

[11] YU Y, GAO S, CHENG S, et al. CBSO: A memetic brain storm optimization with chaotic local search[J]. Memetic Computing, 2018, 10(4): 353-367.

[12] 王言文,邱启荣,王宝坤. 基于 Lasso 和 SVR 的向量夹角余弦变权重组合预测模型[J]. 统计与决策, 2020, 36(18):22-26.

[13] 杜晓昕,张剑飞,郭媛,等. 基于柯西-高斯动态消减变异的果蝇优化算法研究[J]. 计算机工程与科学,2016, 38(6):1171-1176.

[14] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: Theory and applications [J]. Neurocomputing, 2006, 70(1-3): 489-501.

[15] 关心. 基于花朵授粉优化极限学习机的高炉铁水硅含量预测[J]. 电子测量技术,2020,43(4):77-80.

[16] DING S, ZHANG N, ZHANG J, et al. Unsupervised extreme learning machine with representational features[J]. International of Machine Learning and Cybernetics, 2017, 8(2): 587-595.

[17] 王立宪,马宏忠,戴锋. 基于机电联合的 GIL 局部放电趋势预测研究[J]. 电子测量与仪器学报, 2021, 35(10):98-106.

作者简介

张文帅,硕士研究生,主要研究方向为深度学习、数据挖掘分析。

E-mail:18370840069@163.com

王占刚,博士,副教授,主要研究方向为时空模型与可视化、数据挖掘分析。

E-mail:wangzg@bistu.edu.cn