基于混合虚拟样本生成的铈镨/钕组分含量预测方法

陆荣秀[1,2†]，赖路璐[1,2]，杨辉[1,2]，朱建勇[1,2]

(1. 华东交通大学 电气与自动化工程学院, 南昌 330013;
2. 江西省先进控制与优化重点实验室,南昌 330013

摘 要: 针对稀土萃取过程进行质量监控时, 存在采集样本重复率高、有效数据少的小样本问题, 提出一种基于混合虚拟样本生成的稀土萃取过程组分含量预测方法. 首先, 以萃取现场的小样本为基础, 采用中点插值法生成虚拟样本输出数据, 再根据随机配置网络（SCN）中隐含层与输出层、输入层与隐含层间的映射关系, 生成虚拟样本输入数据; 鉴于这些虚拟样本仅能在邻近点产生的局限, 采用结合遗传算法（GA）的多分布趋势扩散技术（MD-MTD）生成优化的虚拟样本集进行补充. 依据数据合理性原则, 将虚拟样本与真实小样本进行融合, 建立基于 SCN 的组分含量预测模型. 通过铈镨/钕萃取现场数据验证和对比实验分析, 结果表明本文方法能有效解决小样本问题, 适用于稀土萃取过程组分含量监控.

关键词: 稀土萃取；组分含量预测；随机配置网络；插值；趋势扩散技术；虚拟样本

中图分类号: TP393 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1078

Prediction Method of CePr/Nd Component Content Based on Hybrid Virtual Sample

LU Rong-xiu[1,2†], *LAI Lu-lu*[1,2], *YANG Hui*[1,2], *ZHU Jian-yong*[1,2]

(1. College of Electrical and Automation, East China Jiaotong University, Nanchang 330013, China; 2. Key Laboratory of Advanced Control and Optimization of Jiangxi Province, Nanchang 330013, China)

Abstract: Aiming at the small sample problem of high sample collection repetition rate and low effective data during quality monitoring of rare earth extraction process, a method for predicting component content of rare earth extraction process based on mixed virtual sample generation is proposed. First of all, based on the small sample extracted from the field, the output data of virtual sample are generated by midpoint interpolation method. Then, the input data of virtual sample are generated according to the mapping relationship between the hidden layer of the stochastic configuration network (SCN) and the output layer, the input layer and the hidden layer. In view of the limitation that the virtual sample can only be generated at neighboring points, a multi-distribution trend diffusion technology (MD-MTD) combined with genetic algorithm (GA) is used to generate an optimized virtual sample set to supplement. According to the principle of data rationality, the virtual sample and the real small sample are merged, and a component content prediction model based on SCN is established. Through the field data verification and comparative experimental analysis of CePr/Nd extraction, the results show that the method in this paper can effectively solve the problem of small sample and the paper is suitable for component content monitoring in the rare-earth extraction process.

Keywords: rare earth extraction; component content; stochastic configuration network; interpolation; megatrend diffusion technique; virtual sample

0 引 言

我国稀土分离企业的生产规模、产品产量和质量对世界的经济发展占据重要地位[1]. 为了保证稀土萃取过程两端出口的产品质量, 根据萃取过程精度控制和工艺控制要求, 常在洗涤段和萃取段设置监测级, 若检测出该级萃取槽体中的组分含量异常, 则需调节洗涤剂和萃取剂流量等工艺参数.

近几年来, 软测量方法的研究对稀土萃取过程

收稿日期: 2021-06-21；**录用日期:** 2021-12-30.

基金项目: 国家重点研发计划项目 (2020YFB1713700), 国家自然科学基金资助项目 (61863014, 61733005, 61963015).

†通讯作者. E-mail: ecjtu_rxlu@163.com.

中组分含量的检测具有重大意义[2]. 文献 [3] 和文献 [4] 分别采用加权最小二乘支持向量机（WLSSVM）和改进的即时学习算法建立软测量模型, 实现稀土萃取过程组分含量的快速预测. 但是, 传统的神经网络往往需要依靠大量的数据作支撑, 而复杂的稀土萃取工业生产现场受环境、检测时间、经济成本的影响, 使得采集到的样本稀缺, 即造成建模数据具有样本容量小、信息间隔大、样本多样性差等小样本问题[5]. 基于此, 通过合理方法扩增数据, 填补真实小样本间的信息间隔, 是具有小样本特征的数据建立有效预测模型的前提.

目前, 针对小样本建模困难的问题, 学术界主要有两方面的研究: 一方面是灰色理论[6], 该方法不适用于规律性不强、分布不均匀的样本; 另一方面是虚拟样本生成方法[7]. 虚拟生成方法按照生成思想有以下三类[8]: 基于扰动思想[9], 基于研究领域具有的先验知识[10] 和基于研究领域的分布函数[11], 但上述方法均具有片面性, 未考虑不同生成思想构造的虚拟样本之间存在互补特性. 文献 [12] 采用改进的大趋势扩散技术结合隐含层插值的方法生成虚拟样本, 实现不同类型虚拟样本的互补, 但该方法未考虑输入、输出属性之间的相关性, 且对于极端分布密集、中间分布稀疏的样本, 虚拟样本集的生成具有较大随机性和不确定性.

对于同样具有小样本特征的稀土萃取过程数据, 本文提出了一种随机配置网络（SCN）隐含层插值结合 GA 优化 MD-MTD 的虚拟样本生成方法. 首先, 从输出属性角度考虑, 采用 SCN 隐含层最优欧氏距离插值方法, 生成一种虚拟样本; 然后, 从输入属性角度考虑, 采用 GA 优化 MD-MTD 方法生成另一种虚拟样本, 最后, 考虑两种不同类型虚拟样本的互补性, 将其与真实小样本融合后进行重复性数据清理并通过稀土萃取过程数据, 验证了本文方法的合理性、有效性、适用性.

1 虚拟样本定义

Poggio 在图像识别领域首先定义虚拟样本[7], 随后, 文献 [13] 对虚拟样本给出定义: 随机生成的训练样本集 $\{x, y\}$, 基于先验知识 K, 可以得到新样本集 $\{Tx, f(T)y\}$. 本文从数据属性角度考虑数据扩增, 对虚拟样本重新定义: 原始数据小样本 $\{x_{small,}, y_{small}\}$, 通过样本属性之间的转换关系 $\{T_x, f(T_x)\}$ 和 $\{T_y, f(T_y)\}$, 合理生成新样本集, 即式 (1)、式 (2).

$\{x_{small,}, y_{small}\}$ 是原始数据小样本, x_{min}^{vs}、x_{max}^{vs} 和 $y_{\min}^{vs}$、$y_{\max}^{vs}$ 分别是虚拟输入和虚拟输出的最小值、最大值. 转换关系 $\{T_x, f(T_x)\}$ 和 $\{T_y, f(T_y)\}$ 可以采用 SCN、ELM（极限学习机）等机器学习方法建立真实小样本集推估超平面 H 和总体超平面 $\hat{H}$ 来得到[19]. 超平面 $H, \hat{H}$ 之间的距离反映了真实小样本与期望样本之间关系, 真实小样本与期望样本空间的关系如图 1 所示.

从输入属性角度进行扩增:

$$\begin{cases} \{x_{small}\} \xrightarrow{T_x} \{x_{small}^{vs}\}, x_{min}^{vs} \le x_{small}^{vs} \le x_{max}^{vs} \\ \left.\begin{matrix} x_{small}^{vs} \\ y_{small} \end{matrix}\right\} \xrightarrow{f(T_x)} \{y_{small}^{vs}\}, y_{min}^{vs} \le y_{small}^{vs} \le y_{max}^{vs} \end{cases} \tag{1}$$

从输出属性角度进行扩增:

$$\begin{cases} \{y_{small}\} \xrightarrow{T_y} \{y_{small}^{vs}\}, y_{min}^{vs} \le y_{small}^{vs} \le y_{max}^{vs} \\ \left.\begin{matrix} x_{small} \\ y_{small}^{vs} \end{matrix}\right\} \xrightarrow{f(T_y)} \{x_{small}^{vs}\}, x_{min}^{vs} \le x_{small}^{vs} \le x_{max}^{vs} \end{cases} \tag{2}$$

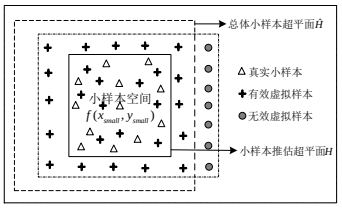


图 1 小样本与虚拟样本关系

2 虚拟样本生成方法

2.1 基于 SCN 隐含层插值的虚拟样本生成方法

对于回归问题, 当训练样本不足时, 可以利用神经网络生成虚拟样本, 从而提高神经网络的学习功能[14]. 填充小样本空间信息间隔最常用的方法是插值法[15], 而随机配置网络的隐含层节点可变, 使得网络具有灵活性[16], 因此, 本文从输出属性角度考虑, 提出基于 SCN 隐含层插值的虚拟样本生成方法, 达到填补真实小样本空间的目的.

SCN 是一个由输入层、隐含层、输出层组成的 3 层前向反馈网络. 给定目标函数 $Y \in R^d \to X = \{x_1, x_2, ..., x_m\} \in R^m$, 假设 SCN 模型此时已有 $L-1$ 个隐含层节点, 即:

$$Y_{L-1} = \sum_{l=1}^{L-1} \beta_l \varphi_l(w_l^T X_m + b_l)\ (L = 1, 2, \cdots, Y_0 = 0) \tag{3}$$

式 (3) 中, w_l, β_l 和 b_l 分别表示网络的输入权值、输出权值和阈值; $\varphi_l(\cdot)$ 是激活函数.

当 SCN 模型输出的残差 e_{L-1} 不满足式 (4) 时, 则需增加新的隐含层节点, 即新的模型有 L 个隐含

层节点, 模型输出为式 (5).

$$|e_{L-1}| = |Y - Y_{L-1}| = |[e_{L-1,1}, e_{L-1,2}, \cdots, e_{L-1,n}]| < \varepsilon \quad (4)$$

式 (4) 中, ε 表示模型的容忍误差, 且 $\varepsilon > 0$.

$$Y_L = Y_{L-1} + \beta_L \phi_L = \sum_{l=1}^{L} \beta_l \phi_l (w_l^T X_m + b_l) \quad (5)$$

根据输入变量 X 得到的第 L 个隐含层节点的输出矩阵 H_L 表示为:

$$H_L = [\phi_L(w_L^T X_1 + b_L), \cdots, \phi_L(w_L^T X_m + b_L)] \quad (6)$$

式 (6) 中, 激活函数 $\phi(x) = \frac{1}{1+e^{-x}}$.

若 SCN 隐含层的输出权值表示为 $\beta = [\beta_1, \beta_2, \cdots, \beta_L]$, 则隐含层输出与网络输出之间的线性关系可以由式 (7) 表示:

$$Y = H_L \cdot \beta \quad (7)$$

为了尽可能地简化计算, 采用欧式距离相似度, 在 SCN 的隐含层输出进行线性中点插值, 并选取式 (8) 的最优欧式距离 d 作为插值的终点行.

$$d = \min\{d_q\} \qquad (q = 1, 2, ..., r) \quad (8)$$

式 (8) 中, d_q 表示不同行间的欧式距离.

若隐含层输出矩阵 H_L 的第一行 H_{h11} 为插值起始行, 与起始行最优欧式距离是隐含层输出矩阵的第二行 H_{h21} , 则采用线性中点插值后隐含层输出为:

$$H' = \begin{bmatrix} H_{h11} & H_{h12} & \cdots & H_{h1L} \\ \frac{H_{h11}+H_{h21}}{2} & \frac{H_{h12}+H_{h22}}{2} & \cdots & \frac{H_{h1L}+H_{h2L}}{2} \\ H_{h21} & H_{h22} & \cdots & H_{h2L} \end{bmatrix} \quad (9)$$

H_{hij} 表示隐含层输出矩阵 H_L 第 i 行的第 j 列元素. 进行多组插值后, 隐含层输出矩阵表示为:

$$H'_L = \begin{bmatrix} H'_{h11} & H'_{h12} & \cdots & H'_{h1L} \\ H'_{h21} & H'_{h22} & \cdots & H'_{h2L} \\ \vdots & \vdots & \ddots & \vdots \\ H'_{hN1} & H'_{hN1} & \cdots & H'_{hNL} \end{bmatrix} \quad (10)$$

那么, 插值后虚拟样本输出为:

$$Y' = \frac{Y_1 + Y_2}{2} \quad (11)$$

合并多组插值后, 虚拟样本的输出 Y' 为:

$$Y' = [Y_1', Y_2', ..., Y_L']^T \quad (12)$$

虚拟样本的输入 X' 则可根据 SCN 的输入层与隐含层之间的非线性映射关系反推得到:

$$X' = (w_L)^{\dagger}(\phi_L{}^{-1}(H'_L) - b_L) \quad (13)$$

式 (13) 中, $w^{\dagger}$ 表示 SCN 输入权值矩阵的广义逆, $\phi_L^{-1}(\cdot)$ 表示激活函数 Sigmoid 的逆, 表达式为:

$$\phi_L^{-1}(\cdot) = \ln(\frac{x}{1-x}) \quad (14)$$

若输入权值矩阵 w_L 是可平方阵, 输入权值矩阵可由矩阵的逆计算得到 $(w_L)^{-1}$, 反之, 依据广义逆的唯一存在性, 输入权值矩阵的广义逆 $(w_L)^{\dagger}$ 为:

$$(w_L)^{\dagger} = ((w_L)^T w_L)^{-1} (w_L)^T \quad (15)$$

综上所述, 经过 N_{v1} 次隐含层输出线性中点插值后可得到 N_{v1} 个虚拟样本:

$$S_{v1} = (X', Y') = (X'_{v1}, Y'_{v1}) \ (v1 = 1, 2, \cdots, N_{v1}) \quad (16)$$

由上述推导可知, 该方法相当于在邻近点之间做插值, 得到的虚拟样本具有特定意义, 使得虚拟样本的生成具有一定的局限性. 故从输入属性角度考虑, 采用 MD-MTD 方法随机生成不均匀虚拟样本, 以弥补 SCN 隐含层插值方法的不足.

2.1.1 GA 优化 MD-MTD 的虚拟样本生成技术

MTD[16] 是在非对称域范围扩展的一种特殊方法, 但是该方法过于简单, 而且未考虑到样本数据集的真实分布情况来扩增虚拟样本, 未能从根本上解决样本分布的不平衡问题[17], MD-MTD 则具有更好的普适性[5].

以真实样本数据集 X 中的某一属性 x_m 为例, 采用 MD-MTD 方法进行数据扩增的方法介绍如下.

(1) 真实小样本输入集 X'_{v2} 的区域扩增:

Step1 计算 x_m 的数据中心点 CL, x_m 的最大值 Max 和最小值 Min , 其中, CL 的表达式如下:

$$CL = \text{median}(x_m) \quad (17)$$

Step2 计算中心点 CL 的左偏度 S_{KL} 、右偏度 S_{KR} 和方差 $S_{x_m}^2$:

$$S_{KL} = \frac{N_L}{N_L + N_R + 1}, S_{KR} = \frac{N_R}{N_L + N_R + 1} \quad (18)$$

$$S_{x_m}^2 = \frac{\sum_{i=1}^{m} (x_m - \bar{x}_m)^2}{m-1} \quad (19)$$

N_L, N_R 表示小于和大于中心点的样本数.

Step3 确定 x_m 可扩展的上界 RB 和下界 LB:

$$LB = \begin{cases} CL - \frac{N_L}{N_L+N_R}\sqrt{\frac{-2S_{x_m}^2 \times \ln(10^{-20})}{N_L}}, & LB < \text{Min} \\ \text{Min}, & LB > \text{Min} \end{cases} \quad (20)$$

$$RB = \begin{cases} CL + \frac{N_R}{N_L+N_R}\sqrt{\frac{-2S_{x_m}^2 \times \ln(10^{-20})}{N_R}}, & RB < \text{Max} \\ \text{Max}, & RB > \text{Max} \end{cases} \quad (21)$$

Step4 在不同区域生成虚拟样本:

在 [LB,Min] 区域内:

$$x_m = LB + s(\text{Min} - LB) \tag{22}$$

在 [Min,Max] 区域内:

$$x_m = \begin{cases} LB + \sqrt{s(RB-LB)(CL-LB)}, \\ \quad 0 < s < \frac{RB-CL}{RB-LB} \\ RB - \sqrt{(1-s)(RB-LB)(RB-CL)}, \\ \quad \frac{RB-CL}{RB-LB} \le s < 1 \end{cases} \tag{23}$$

在 [Max,RB] 区域内:

$$x_m = \text{Max} + s(RB - \text{Max}) \tag{24}$$

其中, s 表示服从正态分布的随机数.

(2) 真实小样本输出集 Y'_{v2} 的区域扩增:

为了保证对应的虚拟样本输出的可靠性, 建立满足如式 (25) 精度要求的 SCN 模型, 将虚拟输入 X'_{v2} 作为测试输入, 通过已经搭建好的 SCN 模型, 生成与之对应的虚拟输出 Y'_{v2}.

$$\left| \frac{Y_{v2} - \hat{Y}_{v2}}{Y_{v2}} \right| \times 100\% \le \lambda \tag{25}$$

Y_{v2} 和 $\hat{Y}_{v2}$ 分别表示样本的实际输出值和预测值, λ 是预测精度的阈值条件, 一般取 5%.

基于此, 采用 MD-MTD 方法即可生成虚拟样本集 (X'_{v2}, Y'_{v2}).

由图 1 可知, 生成的虚拟样本可能不在期望空间内, 就需要将虚拟样本构造过程转换成虚拟样本寻优过程[17], 因此, 本文采用具有优良鲁棒性和全局搜索能力的 GA 算法, 寻找相对 "最优" 的虚拟样本集, 使得预测值与实际值之间的误差最小化, 即 GA 优化 MD-MTD 方法, 具体的数学描述为:

$$\begin{aligned} &\min:\ g(x) = \left| \frac{y_{\text{test}} - \hat{y}_{\text{test}}}{y_{\text{test}}} \right| \times 100 \\ &\text{s.t.:}\quad LB_{\text{t}} \le x_{\text{t}} \le RB_{\text{t}},\ \text{t} = 1, 2, \ldots, m_{\text{t}} \\ &\qquad -0.05 \le \left| \frac{y_{\text{test}} - \hat{y}_{\text{test}}}{y_{\text{sst}}} \right| \le 0.05 \end{aligned} \tag{26}$$

$g(x)$ 是虚拟样本优化问题的目标函数; $x_t = \{x_1, x_2, ..., x_{m_t}\}$ 表示 m_t 个决策变量;LB_t, RB_t 分别表示 x_t 的下界和上界. 求解虚拟样本集的优化问题, 即可转换为求解式 (26) 的非线性不等式约束的优化问题. 通过 GA 优化 MD-MTD 方法生成的虚拟样本集为 $S_{v3} = (X'_{v3}, Y'_{v3})$.

2.1 节中基于 SCN 隐含层插值生成虚拟样本的方法, 是在真实小样本空间特定生成虚拟样本, 不适用于分布不均匀的复杂工业过程; 而本节方法可以随机生成不均匀虚拟样本, 这两种类型的虚拟样本在一定程度上形成互补. 基于此, 本文将两种方法生成的不同类型的虚拟样本进行融合, 以增强虚拟样本生成的稳定性.

3 基于混合虚拟样本模型的建立与评估

3.1 虚拟样本数据合理性判定

由图 1 中真实小样本与虚拟样本之间的关系可知, 符合期望空间的虚拟样本视为有效, 即虚拟样本生成的合理性. 根据式 (2)、式 (3) 可以得到, 虚拟样本生成的合理性原则如下:

$$\text{if}\{x_{small}\} \xrightarrow{T_x} \{x^{vs}_{small}\},\ x^{vs}_{\min} \le x^{vs}_{small} \le x^{vs}_{\max}:$$
$$\begin{cases} \text{reasonable}, & y^{vs}_{\min} \le y^{vs}_{small} \le y^{vs}_{\max} \\ \text{unreasonable}, & \text{others} \end{cases} \tag{27}$$

$$\text{if}\{y_{small}\} \xrightarrow{T_y} \{y^{vs}_{small}\},\ y^{vs}_{\min} \le y^{vs}_{small} \le y^{vs}_{\max}:$$
$$\begin{cases} \text{reasonable}, & x^{vs}_{\min} \le x^{vs}_{small} \le x^{vs}_{\max} \\ \text{unreasonable}, & \text{others} \end{cases} \tag{28}$$

由于数据间存在重复的数据记录问题, 使得融合的数据样本含有 "脏" 数据, 为了提取高质量的有效数据, 就需要对这类数据进行清理[18], 虚拟样本融合前的清理按照如式 (29) 原则进行:

$$\begin{cases} y_m = Y_{v1}' \, or \, y_m = Y_{v3}' \ or \ Y_{v1}' = Y_{v3}', & \text{delete} \\ \text{others}, & \text{reserve} \end{cases} \tag{29}$$

式 (29) 表示: 如果生成的虚拟样本和原始数据样本相同, 则删除虚拟样本, 否则保留原始样本.

综上, 得到相应的虚拟样本集 S'_{v1} 、S'_{v3}, 将其和真实小样本的训练集进行数据融合, 即:

$$\begin{aligned} S_v &= \{S'_{v1}; S'_{v3}; T'_{tr}\} \\ &= \{(X''_{v1}, Y''_{v1}); (X''_{v3}, Y''_{v3}); (X'_{tr}, Y'_{tr})\} \end{aligned} \tag{30}$$

3.2 混合虚拟样本模型的建立

根据 2.1 节、2.2 节所提虚拟样本生成方法, 并结合 3.1 节数据处理原则得到最终的混合虚拟样本建立基于混合虚拟样本生成的组分含量 SCN 模型, 其原理如图 2 所示.

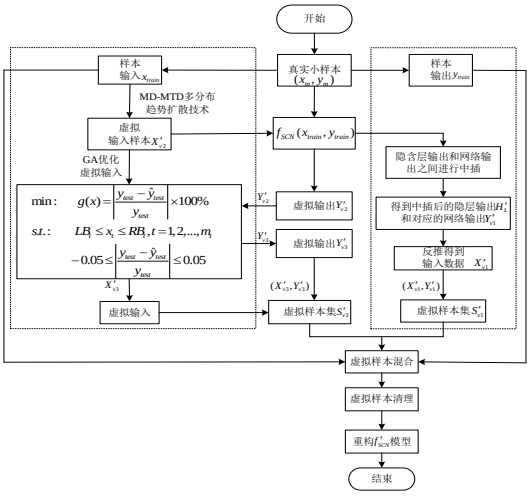


图 2 虚拟样本生成工作流程图

4 实验验证与分析

4.1 稀土萃取过程描述

稀土萃取分离的生产流程如图 3 所示.

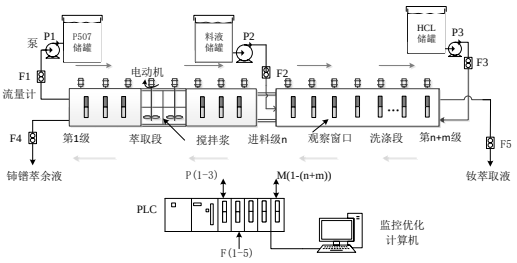


图 3 稀土萃取分离工艺流程

图 6 中, 待分离的稀土料液中含有难萃组分铈（Ce）、镨（Pr）和易萃组分钕（Nd）, 每一萃取级均要经过搅拌、澄清, 在洗涤段出口得到富含 Nd 离子的紫红色萃取液; 萃取段出口得到富含 Pr 离子的淡绿色萃余液[19]. 整个萃取分离过程耗时较长, 从生产现场采集的溶液数据样本重复性高, 使得有效建模的样本数据十分有限, 需采取虚拟样本生成的方式提高组分含量模型的精度.

以江西某稀土公司的铈镨/钕萃取工业生产线为研究对象, 在不同工况的萃取槽体中采集到 65 组样本溶液, 通过离线实验室化验分析检测得到该级组分含量值, 同时将溶液用于图像采集, 提取溶液图像的 H、S、I 特征分量，并以此为辅助变量, Nd 组分含量作为主导变量. 为了测试训练模型的有效性, 随机抽取 50 组原始样本作为训练样本, 表示为:$T_{tr}=\{X_{tr},Y_{tr}\}=\{x_{Htr},x_{Str},x_{Itr},y_{tr}\}\in R^{50\times4}$, 剩余 15 组原始样本为测试样本, 表示为:$T_{te}=\{X_{te},Y_{te}\}=\{x_{Hte},x_{Ste},x_{Ite},y_{te}\}\in R^{15\times4}$. 下面设计 6 组对比实验: (a) 基于 SCN 的真实小样本组分含量模型; (b) 基于 SCN 隐含层插值的组分含量模型; (c) 基于 MD-MTD 虚拟样本生成的组分含量模型; (d) 基于 GA 优化 MD-MTD 虚拟样本生成的组分含量模型; (e) 基于 SMOTE 算法虚拟样本生成的组分含量预测模型; (f) 基于 SCN 隐含层插值和 GA 优化 MD-MTD 的混合虚拟样本生成的组分含量模型. 为了评估组分含量模型的性能, 使用式 (31)、式 (32) 的最大相对误差（MRE）和均方根误差（RMSE）作为评价指标.

$$\mathrm{MRE}=\max\left(\left|\frac{Y-\hat{Y}}{\hat{Y}}\right|\times100\%\right) \tag{31}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{te=1}^{n}\left(Y-\hat{Y}\right)^2} \tag{32}$$

式中, $\hat{Y}$、Y 分别表示测试集样本实际值和预测值. n 表示测试样本数量.

4.2 仿真结果

4.2.1 合理性分析

为了本文所提方法生成的虚拟样本的合理性, 将实验 b、实验 d 依次增加 10 个虚拟样本, 试验 5 次, 每个实验得到 5 种不同的训练数据集, 采用测试集对模型进行测试, 和实验 a 形成对比实验. 参照文献 [16], SCN 的初始参数均设置为可容忍误差 $\varepsilon=0.001$，最大隐含层节点数 $L=50$；最大搜索次数 $T_\max=100$；随机权重范围为:$\beta_L=[0.5,1,3,5,7,9,15,25,50,100,150,200]$, 不等式约束系数 $r=[0.9,0.99,0.999,0.9999,0.99999,0.999999]$. GA 的参数为: 种群数 N=20，遗传代数 F=500, 编码长度 $l=6$, 交叉概率 $p_1=0.7$, 变异概率 $p_2=0.1$.

以每次扩增的 10 个虚拟样本为一组单位, 根据式 (27)、式 (28) 给出的虚拟样本合理性原则, 以虚拟样本生成的合格率判断虚拟样本的合理性.

根据 3.1 节, 考虑到 b 方法、d 方法生成的虚拟样本存在重复数据的情况, 进行数据清理前后, 不同方法生成的虚拟样本数量及合格率如表 1 所示.

表 1 不同方法的虚拟样本数量及合格率

方法	原始样本	虚拟样本	清理后	合格率/%
SCN 隐层插值	50	50	41	100
GA-优化 MD-MTD	50	50	40	100
虚拟样本融合	50	100	81	100

由表 1 可知, 基于 SCN 隐层插值方法和基于 GA 优化 MD-MTD 方法生成的虚拟输入、输出均 100% 符合式 (1)、式 (2) 虚拟样本设定的上下限, 从而验证了所提方法生成虚拟样本的合理性.

4.2.2 有效性分析

为了验证所提单一虚拟样本生成方法的有效性及重新构建组分含量预测模型的有效性, 实验 a、实验 b、实验 d 对应的各个组分含量 SCN 模型的各项性能值如表 2 所示.

表 2 SCN 模型的各项性能值

虚拟样本数	实验 a		实验 b		实验 d	
	MRE(%)	RMSE	MRE(%)	RMSE	MRE(%)	RMSE
0	4.65	1.75	4.65	1.75	4.65	1.75
10	-	-	4.47	1.23	4.61	1.49
20	-	-	4.07	1.21	4.10	1.10
30	-	-	3.55	1.10	3.70	0.91
40	-	-	3.42	0.98	3.12	0.83
50	-	-	3.00	0.77	2.89	0.52

分析表 2, 对于实验 b 和实验 d, 当虚拟样本数量分别从 0 增加 50, 预测模型的 MRE 和 RMSE 值在不断降低, 即采用了虚拟样本生成技术的组分含量模型比未采用该技术的模型性能高, 且随着虚拟样本数的增加, 模型性能不断提高.

为了凸显混合虚拟样本生成方法的优越性, 将实验 b、实验 d 和实验 f 生成的虚拟样本融合真实小样本, 根据 3.1 节的规则, 对重复数据进行清理后得到实验 f 的 81 个虚拟样本. 对比实验 e, 采用实验法确认 SMOTE 算法的近邻参数 K 取 10. 基于本文设计的六组实验, 应用于稀土萃取生产现场数据, 实验对比结果和相对误差分别如表 3 和图 4 所示.

表 3 六组实验对比结果比较

实验	虚拟样本数量	MRE(%)	RMSE
a	0	4.65	1.75
b	50	3.00	0.77
c	50	4.20	1.23
d	50	2.89	0.52
e	50	3.95	1.19
f	81	2.84	0.38

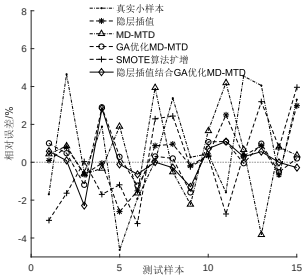


图 4 六组实验的相对误差变化

分析表 3 和图 4: (1) 对比实验 b 实验 e 与实验 a 测试结果可知, 采用虚拟样本生成后的组分含量模型性能指标值均优于基于真实小样本的 SCN 模型测试指标: MRE=4.65%, RMSE=1.75, 说明采用虚拟样本生成的方法对提高稀土萃取过程组分含量的模型预测精度是有效的; (2) 对比实验 b、实验 d 与实验 e 可知, 采用本文所提数据扩增方法比 SMOTE 算法[20] 生成虚拟样本的模型测试精度更高, 对于稀土组分含量预测更具优势; (3) 对比实验 b 实验 e 与实验 f 的测试结果可知, 采用混合虚拟样本生成方法的模型测试性能指标最佳: MRE=2.84%, RMSE=0.38; 说明混合虚拟样本生成方法综合考虑了真实小样本的输入、输出特性, 可以提高虚拟样本生成方法的有效性能, 具有更好的泛化能力.

4.2.3 适用性分析

为了说明基于 SCN 扩增生成的虚拟样本对其他算法也适用, 增加两组对比实验:

实验一: 使用和实验 a 相同的训练集和测试集, 分别采用 SCN、SVM、LSSVM、WLSSVM 方法建立组分含量模型, 得到虚拟样本生成前各模型的性能指标;

实验二: 以实验 a 中原有的 61 个真实样本数据和 81 个虚拟样本数据混合, 随机选择 105 组数据作为训练样本, 剩下 37 组数据作为测试样本, 分别采用 SCN、SVM、LSSVM、WLSSVM 方法进行建模, 对比模型测试的性能指标. 数据扩增前后, SCN 的参数设置和 4.2.1 保持一致, 其他各模型参数设置如表 4 所示; 不同建模方法的模型测试性能结果如表 5 所示.

表 4 数据扩增前后各模型参数设置

模型	SVM		LSSVM		WLSSVM	
参数	扩增前	扩增后	扩增前	扩增后	扩增前	扩增后
c	4	4	-	-	-	-
g	0.5	1	-	-	-	-
γ	-	-	330.1426	353.7557	199.457	342.7692
σ	-	-	0.4371	0.5179	0.0098	0.0136
W_H	-	-	-	-	0.4181	0.4021
W_S	-	-	-	-	0.3965	0.3688
W_I	-	-	-	-	0.1854	0.2291
m_H	-	-	-	-	2.2776	2.3332
m_S	-	-	-	-	2.1600	2.1401
m_I	-	-	-	-	1.0098	1.3294

表 5 数据扩增前后各模型的测试性能

模型	无虚拟样本		虚拟样本为 81	
	MRE(%)	RMSE	MRE(%)	RMSE
SCN	4.65	1.75	2.84	0.38
SVM	10.67	2.94	3.79	1.14
LSSVM	6.52	2.77	3.13	0.86
WLSSVM	5.35	2.22	3.43	0.88

分析表 5 可知: (1) 虚拟样本生成前, 组分含量 SCN 模型的性能比基于 SVM、LSSVM 和 WLSSVM 的组分含量模型性能更优. (2) 当生成 81 个虚拟样本后, 各组分含量模型的测试性能均提高, 说明将混合虚拟样本应用于 SVM、LSSVM、WLSSVM 等方法建模, 也可以实现对 CePr/Nd 组分含量的有效预测.

5 结论

本文针对稀土萃取过程质量检测具有小样本问题, 提出了 SCN 隐含层插值结合 GA 优化 MD-MTD 的混合虚拟样本生成方法, 实现稀土萃取过程组分含量的预测. 本文主要工作包括: (1) 提出基于 SCN 隐含层插值生成虚拟样本方法; (2) 在 MD-MTD 方法基础上, 提出 GA 算法优化 MD-MTD 方法, 将虚拟样本构造过程转换成虚拟样本寻优过程, 使生成的虚拟样本尽可能的符合期望空间, 确保了虚拟样本集的有效扩增; (3) 将生成的混合虚拟样本分别从合理性、有效性、适用性三个方面进行分析, 并通过稀土萃取生产现场采集的过程数据, 验证了本文所提方法能够有效的填充真实小样本间的信息间隔, 提高数据样本的完备性和保证稀土萃取组分含量预测模型的可靠性.

参考文献 (References)

[1] 尚宇. 中国稀土产业国际竞争力研究 [D]. 中国地质大学（北京）,2011.

(Shang Y. Research on the international competitiveness of China's rare earth industry [D]. China university of geosciences (Beijing), 2011.)

[2] 田海, 郭智恒, 李兰云. 稀土萃取分离过程软测量方法的研究 [J]. 中国稀土学报, 2015, 33(002): 201-205.
(Tian H, Guo Z H, Li L Y. Research on soft measurement method of rare earth extraction and separation process[J]. Journal of Rare Earths in China, 2015, 33(2): 201-205.)

[3] 朱建勇, 张旭乾, 杨辉, 陆荣秀. 单光照条件变化的镨/钕元素组分含量软测量 [J]. 化工学报, 2019, 70(02): 780-788.
(Zhu J Y, Zhang X Q, Yang H, Lu R X. Soft measurement of praseodymium/neodymium element content under single light conditions[J]. Journal of Chemical Industry and Engineering(China), 2019, 70(02): 780-788.)

[4] 陆荣秀, 饶运春, 杨辉, 朱建勇. 基于改进即时学习算法的镨/钕元素组分含量预测 [J]. 控制理论与应用, 2020, 37(08): 1846-1854.
(Lu R X, Rao Y C, Yang H, Zhu J Y. Prediction of praseodymium/neodymium component content based on improved instant learning algorithm[J]. Control Theory and Applications, 2020, 37(08): 1846-1854.)

[5] 朱宝. 虚拟样本生成技术及建模应用研究 [D]. 2017.
(Zhu B. Research on Virtual Sample Generation Technology and Modeling Application [D]. 2017.)

[6] Chen, Chien-Chih, Huang. A novel gray forecasting model based on the box plot for small manufacturing data sets[J]. Applied mathematics and computation, 2015, 265: 400-408.

[7] Poggio T, Vetter T. Recognition and structure from One 2D model view: observations on prototypes, object classes and symmetries[R]. Technical Report A. I. Memo 1347, Massachusetts Institute of Technology Cambridge, MA, USA, 1992.

[8] 于旭, 杨静, 谢志强. 虚拟样本生成技术研究 [J]. 计算机科学, 2011, 038(003): 16-19.
(Yu X, Yang J, Xie Z Q. Research on Virtual Sample Generation Technology[J]. Computer Science, 2011, 038(003): 16-19.)

[9] 易令, 吕忠元, 丁进良. 面向原油总氢物性预测的数据扩增预处理方法 [J]. 控制与决策, 2018, 33(12): 44-51.
(Yi L, Lv Z Y, Ding J L, Liu C X. Data amplification preprocessing method for crude oil total hydrogen property prediction[J]. Control and Decision, 2018, 33(12): 2153-2160.)

[10] Li D C, Wu C S, Tsai T I. Using mega trend diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge[J].Computers and Operations Research, 2007, 34(4): 966-982.

[11] Zhou Z H, Jiang Y. NeC4.5: Neural Ensemble Based C4.5[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(6): 770-773.

[12] 乔俊飞, 郭子豪, 汤健. 基于改进大趋势扩散和隐含层插值的虚拟样本生成方法及应用 [J]. 化工学报, 2020, 71(12): 5681-5695.
(Qiao J F, Guo Z H, Tang J. Virtual sample generation method based on improved megatrend diffusion and hidden layer interpolation with its application[J]. CIESC Journal, 2020, 71(12): 5681-5695.)

[13] 汤健, 乔俊飞, 柴天佑. 基于虚拟样本生成技术的多组分机械信号建模 [J]. 自动化学报,2018,44(09):1569-1589.
(Tang J, Qiao J F, Chai T, Y. Multi-component mechanical signal modeling based on virtual sample generation technology[J]. Acta Automatica Sinica, 2018, 44(09): 1569-1589.)

[14] Cho S, Jang M, Chang S. Virtual Sample Generation Using a Population of Networks[J]. Neural Processing Letters, 1997, 5(2): 83-89.

[15] He Y L, Wang P J, Zhang M Q. A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry[J]. Energy, 2018, 147: 418-427.

[16] Wang D H, Li M. Stochastic Configuration Networks: Fundamentals and Algorithms[J]. IEEE transactions on cybernetics, 2017, 47(10): 3466-3479.

[17] Chen Z S , Zhu B , He Y L , et al. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets[J]. Engineering Applications of Artificial Intelligence, 2017, 59(3): 236-243.

[18] 陈伟, 丁秋林. 一种 XML 相似重复数据的清理方法研究 [J]. 北京航空航天大学学报, 2004(09): 835-838.
(Chen W, Ding Q L. A method of cleaning up similar duplicate data in XML[J].Journal of Beijing University of Aeronautics and Astronautics, 2004(09): 835-838.)

[19] 贾文君, 柴天佑. 稀土串级萃取分离过程元素组分含量的多模型软测量 [J]. 控制理论与应用, 2007(04): 569-573.
(Jia W J, Chai T Y. Multi-model soft measurement of element component contents in cascade extraction and separation of rare earth elements [J]. Control theory and application, 2007, 24(4): 569-573.)

[20] 王超学, 张涛, 马春森. 面向不平衡数据集的改进型 SMOTE 算法 [J]. 计算机科学与探索, 2014(06):91-98.
(Wang C X, Zhang T, Ma C S. Improved SMOTE Algorithm for Imbalanced Datasets. Journal of Computer Science and Technology, 2014(06):91-98.)

作者简介

陆荣秀(1976−), 女, 副教授, 博士, 从事复杂工业过程建模与控制、智能检测等研究, E-mail: ecjtu_rxlu@163.com;

赖路璐(1996−), 女, 硕士, 从事复杂系统建模、控制与优化等研究, E-mail:1743316178@qq.com;

杨辉(1965−), 男, 教授, 博士生导师, 从事复杂系统建模、控制与优化、大数据分析等研究, E-mail: yhshuo@263.net;

朱建勇(1977−), 男, 副教授, 博士, 从事复杂工业过程控制与优化, 大数据分析等研究, E-mail: zhujyemail@163.com.