控制与决策
Control and Decision

**基于混合虚拟样本生成的铈镨/钕组分含量预测**

陆荣秀, 赖路璐, 杨辉, 朱建勇

引用本文:
陆荣秀,赖路璐,杨辉,朱建勇. 基于混合虚拟样本生成的铈镨/钕组分含量预测[J]. 控制与决策, 2023, 38(4): 1129−1136.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1078

---

**您可能感兴趣的其他文章**

**Articles you may be interested in**

基于复合生成对抗网络的对抗样本生成算法研究

Research on generative adversarial example algorithm based on multiple GANs

控制与决策. 2023, 38(2): 528−536 https://doi.org/10.13195/j.kzyjc.2021.0028

特征响应权重自适应的IoU网络跟踪算法改进

Improvement of IoU network tracking with adaptive weighted characteristic responses

控制与决策. 2022, 37(7): 1752−1762 https://doi.org/10.13195/j.kzyjc.2021.0148

小样本条件下基于属性权重Shapley值分配的粗糙集决策模型

Rough set decision−making model based on shapley value assignment of attribute weight under the condition of small sample

控制与决策. 2022, 37(10): 2677−2684 https://doi.org/10.13195/j.kzyjc.2020.1709

基于分类特征约束变分伪样本生成器的类增量学习

Class incremental learning based on variational pseudo−sample generator with classification feature constraints

控制与决策. 2021, 36(10): 2475−2482 https://doi.org/10.13195/j.kzyjc.2020.0228

基于WGRA−FCM样本相似性度量的转炉炼钢终点碳温软测量方法

End point carbon temperature measurement method based on WGRA−FCM for sample similarity measurement

控制与决策. 2021, 36(9): 2170−2178 https://doi.org/10.13195/j.kzyjc.2020.0128

# 基于混合虚拟样本生成的铈镨/钕组分含量预测

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,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. 铈镨/钕萃取现场数据验证和对比实验分析表明,所提出的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

**引用格式:** 陆荣秀,赖路璐,杨辉,等. 基于混合虚拟样本生成的铈镨/钕组分含量预测 [J]. 控制与决策, 2023, 38(4): 1129-1136.

# 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 the virtual sample are generated using the midpoint interpolation method. Then, the input data of the 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 the 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 the SCN is established. Through the field data verification and comparative experimental analysis of CePr/Nd extraction, the results show that the proposed method can effectively solve the problem of the small sample, which 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、产品产量和质量在世界的经济发展中占据重要地位[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中组分含量的检测具有重大意义[2]. 文献[3-4]分别采用加权最小二乘支持向量机(WLSSVM)和改进的即

**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)；国家自然科学基金项目(61863014, 61733005, 61963015).

**责任编委:** 孙宗耀.

†通讯作者. E-mail: ecjtu_rxlu@163.com.

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3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本文方法的合理性、有效性、适用性.

## 1 虚拟样本定义

文献[7]在图像识别领域定义了虚拟样本,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$\{T_x, f(T_y)\}$. 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$\{x_{\rm small}, y_{\rm small}\}$通过样本属性间的转换关系$\{T_x, f(T_x)\}$和$\{T_y, f(T_y)\}$,合理生成新样本集,即分别如下式所示:

从输入属性角度进行扩增:

$$\begin{cases} \{x_{\rm small}\} \xrightarrow{T_x} \{x^{\rm vs}_{\rm small}\},\ x^{\rm vs}_{\min} \leqslant x^{\rm vs}_{\rm small} \leqslant x^{\rm vs}_{\max};\\ \left.\begin{matrix} x^{\rm vs}_{\rm small}\\ y_{\rm small}\end{matrix}\right\} \xrightarrow{f(T_x)} \{y^{\rm vs}_{\rm small}\},\ y^{\rm vs}_{\min} \leqslant y^{\rm vs}_{\rm small} \leqslant y^{\rm vs}_{\max}. \end{cases} \tag{1}$$

从输出属性角度进行扩增:

$$\begin{cases} \{y_{\rm small}\} \xrightarrow{T_y} \{y^{\rm vs}_{\rm small}\},\ y^{\rm vs}_{\min} \leqslant y^{\rm vs}_{\rm small} \leqslant y^{\rm vs}_{\max};\\ \left.\begin{matrix} x_{\rm small}\\ y^{\rm vs}_{\rm small}\end{matrix}\right\} \xrightarrow{f(T_y)} \{x^{\rm vs}_{\rm small}\},\ x^{\rm vs}_{\min} \leqslant x^{\rm vs}_{\rm small} \leqslant x^{\rm vs}_{\max}. \end{cases} \tag{2}$$

$\{x_{\rm small}, y_{\rm small}\}$是原始数据小样本,$x^{\rm vs}_{\min}$、$x^{\rm vs}_{\max}$和$y^{\rm vs}_{\min}$、$y^{\rm vs}_{\max}$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$\hat{H}$得到. 超平面$H$与$\hat{H}$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.

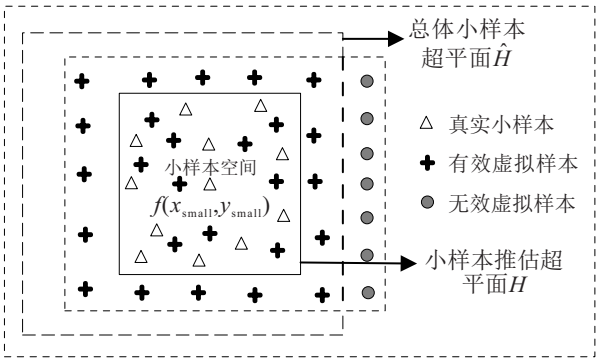


**图1 小样本与虚拟样本关系**

## 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$Y \in R^d \to X = \{x_1, x_2, \ldots, x_m\} \in R^m$,假设SCN模型此时已有$L-1$个隐含层节点,即

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

其中:$w_l$、$\beta_l$和$b_l$分别是网络的输入权值、输出权值和阈值,$\varphi_l(\cdot)$是激活函数.

若SCN模型输出的残差$e_{L-1}$不满足下式:

$$|e_{L-1}| = |Y - Y_{L-1}| =$$

$$||[e_{L-1,1}, e_{L-1,2}, \dots, e_{L-1,n}]|| < \varepsilon, \tag{4}$$

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

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

其中$\varepsilon$表示模型的容忍误差,且$\varepsilon > 0$.

由输入变量$X$得到的第$L$个隐含层节点的输出矩阵$H_L$表示为

$$H_L = [\phi_L(w_L^{\mathrm{T}} X_1 + b_L), \dots, \phi_L(w_L^{\mathrm{T}} X_m + b_L)], \tag{6}$$

其中激活函数$\phi(x) = \dfrac{1}{1+\mathrm{e}^{-x}}$.

若SCN隐含层的输出权值表示为$\beta = [\beta_1, \beta_2, \dots, \beta_L]$,则隐含层输出与网络输出之间的线性关系可表示为

$$Y = H_L \cdot \beta. \tag{7}$$

为了尽可能地简化计算,采用欧氏距离相似度,在SCN的隐含层输出进行线性中点插值,并选取下式的最优欧氏距离$d$作为插值的终点行:

$$d = \min\{d_q\},\ q = 1, 2, \dots, r, \tag{8}$$

其中$d_q$是不同行间的欧氏距离.

若隐含层输出矩阵$H_L$的第1行$H_{h11}$为插值起始行,隐含层矩阵中与起始行最接近的那一行是隐含层输出矩阵需要插值的第2行$H_{h21}$,则采用线性中点插值后隐含层输出为

$$H' = \begin{bmatrix} H_{h11} & H_{h12} & \cdots & H_{h1L} \\ \dfrac{H_{h11}+H_{h21}}{2} & \dfrac{H_{h12}+H_{h22}}{2} & \cdots & \dfrac{H_{h1L}+H_{h2L}}{2} \\ H_{h21} & H_{h22} & \cdots & H_{h2L} \end{bmatrix}, \tag{9}$$

其中$H_{hij}$是隐含层输出矩阵$H_L$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

$$H'_L = \begin{bmatrix} H'_{h11} & H'_{h12} & \dots & H'_{h1L} \\ H'_{h21} & H'_{h22} & \dots & H'_{h2L} \\ \vdots & \vdots & \ddots & \vdots \\ H'_{hN1} & H'_{hN1} & \dots & H'_{hNL} \end{bmatrix}. \tag{10}$$

那么,插值后虚拟样本输出为

$$Y' = \frac{Y_1 + Y_2}{2}. \tag{11}$$

合并多组插值后,虚拟样本的输出$Y'$为

$$Y' = [Y_1', Y_2', \dots, Y_L']^{\mathrm{T}}. \tag{12}$$

虚拟样本的输入$X'$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,即

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

其中:$w^{\dagger}$是SCN输入权值矩阵的广义逆;$\phi_L^{-1}(\cdot)$是激活函数Sigmoid的逆,表达式为

$$\phi_L^{-1}(\cdot) = \ln\left(\frac{x}{1-x}\right). \tag{14}$$

若输入权值矩阵$w_L$是可平方阵,则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

$$(w_L)^{\dagger} = ((w_L)^{\mathrm{T}} w_L)^{-1} (w_L)^{\mathrm{T}}. \tag{15}$$

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

$$S_{v1} = (X', Y') = (X'_{v1}, Y'_{v1}),\ v_1 = 1, 2, \dots, N_{v1}. \tag{16}$$

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.

### 2.2　GA优化MD-MTD的虚拟样本生成技术

大趋势扩散技术(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,而且在未考虑到样本数据集的真实分布的情况下扩增虚拟样本,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.

step 1: 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:

$$\mathrm{CL} = \mathrm{median}(x_m). \tag{17}$$

step 2: 计算中心点CL的左偏度$S_{\mathrm{KL}}$、右偏度$S_{\mathrm{KR}}$和方差$S^2_{x_m}$,即

$$S_{\mathrm{KL}} = \frac{N_L}{N_L + N_R + 1},\ S_{\mathrm{KR}} = \frac{N_R}{N_L + N_R + 1}, \tag{18}$$

$$S^2_{x_m} = \frac{\sum_{i=1}^{m} (x_m - \bar{x}_m)^2}{m-1}, \tag{19}$$

其中$N_L$、$N_R$是小于和大于中心点的样本数.

step 3: 确定$x_m$可扩展的上界RB和下界LB,即

$$\mathrm{LB} = \begin{cases} \mathrm{CL} - \dfrac{N_L}{N_L + N_R}\sqrt{\dfrac{-2S^2_{x_m} \times \ln(10^{-20})}{N_L}}, \\ \qquad \mathrm{LB} < \min; \\ \min,\ \mathrm{LB} > \min. \end{cases} \tag{20}$$

$$\mathrm{RB}=\begin{cases}\mathrm{CL}+\dfrac{N_R}{N_L+N_R}\sqrt{\dfrac{-2S_{x_m}^2\times\ln(10^{-20})}{N_R}},\\ \qquad \mathrm{RB}<\max;\\ \max,\ \mathrm{RB}>\max.\end{cases}\tag{21}$$

step 4: 在不同区域生成虚拟样本:

在[LB, min]区域内有

$$x_m=\mathrm{LB}+s(\min-\mathrm{LB}).\tag{22}$$

在[min, max]区域内有

$$x_m=\begin{cases}\mathrm{LB}+\sqrt{s(\mathrm{RB}-\mathrm{LB})(\mathrm{CL}-\mathrm{LB})},\\ \qquad 0<s<\dfrac{\mathrm{RB}-\mathrm{CL}}{\mathrm{RB}-\mathrm{LB}};\\ \mathrm{RB}-\sqrt{(1-s)(\mathrm{RB}+\mathrm{LB})(\mathrm{RB}-\mathrm{CL})},\\ \qquad \dfrac{\mathrm{RB}-\mathrm{CL}}{\mathrm{RB}-\mathrm{LB}}\leqslant s<1.\end{cases}\tag{23}$$

在[max, RB]区域内有

$$x_m=\max+s(\mathrm{RB}-\max).\tag{24}$$

其中$s$是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,建立满足下式精度要求的SCN模型:

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

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$Y'_{v2}$. 其中:$Y_{v2}$和$\hat{Y}_{v2}$分别是样本的实际输出值和预测值;$\lambda$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

$$\begin{aligned}&\min\ g(x)=\left|\frac{y_{\mathrm{test}}-\hat{y}_{\mathrm{test}}}{y_{\mathrm{test}}}\right|\times 100.\\ &\text{s.t. }\mathrm{LB}_t\leqslant x_t\leqslant\mathrm{RB}_t,\ t=1,2,\ldots,m_t;\\ &-0.05\leqslant\left|\frac{y_{\mathrm{test}}-\hat{y}_{\mathrm{test}}}{y_{\mathrm{sst}}}\right|\leqslant 0.05.\end{aligned}\tag{26}$$

其中:$g(x)$是虚拟样本优化问题的目标函数;$x_t=\{x_1,x_2,\ldots,x_{m_t}\}$是$m_t$个决策变量;$\mathrm{LB}_t$、$\mathrm{RB}_t$分别是$x_t$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)、(3)可以得到,虚拟样本生成的合理性原则如下:

$$\text{if}\{x_{\mathrm{small}}\}\xrightarrow{T_x}\{x^{\mathrm{vs}}_{\mathrm{small}}\},\ x^{\mathrm{vs}}_{\min}\leqslant x^{\mathrm{vs}}_{\mathrm{small}}\leqslant x^{\mathrm{vs}}_{\max}:\begin{cases}\text{reasonable},\ y^{\mathrm{vs}}_{\min}\leqslant y^{\mathrm{vs}}_{\mathrm{small}}\leqslant y^{\mathrm{vs}}_{\max};\\ \text{unreasonable},\ \text{otherwise}.\end{cases}\tag{27}$$

$$\text{if}\{y_{\mathrm{small}}\}\xrightarrow{T_y}\{y^{\mathrm{vs}}_{\mathrm{small}}\},\ y^{\mathrm{vs}}_{\min}\leqslant y^{\mathrm{vs}}_{\mathrm{small}}\leqslant y^{\mathrm{vs}}_{\max}:\begin{cases}\text{reasonable},\ x^{\mathrm{vs}}_{\min}\leqslant x^{\mathrm{vs}}_{\mathrm{small}}\leqslant x^{\mathrm{vs}}_{\max};\\ \text{unreasonable},\ \text{otherwise}.\end{cases}\tag{28}$$

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,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下原则进行:

$$\begin{cases}y_m=Y'_{v1}\text{ or }y_m=Y'_{v3}\text{ or }Y'_{v1}=Y'_{v3},\ \text{delete};\\ \text{otherwise},\ \text{reserve}.\end{cases}\tag{29}$$

式(29)表示:如果生成的虚拟样本与原始数据样本相同,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$S'_{v1}$、$S'_{v3}$,将其与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.

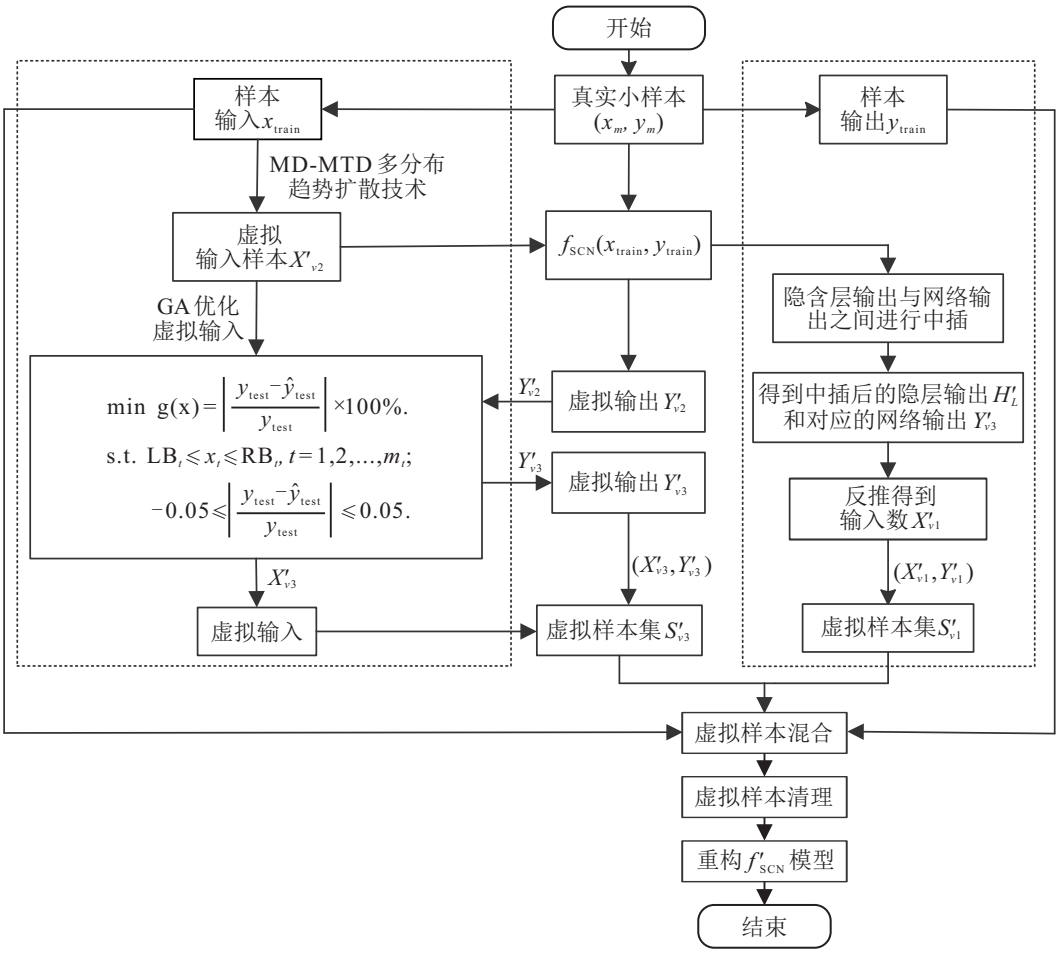


**图 2　虚拟样本生成工作流程**

## 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.

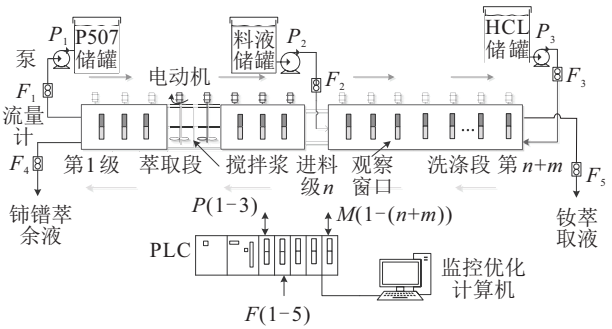


**图 3　稀土萃取分离工艺流程**

图3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$T_{\rm tr}=\{X_{\rm tr}, Y_{\rm tr}\}=\{x_{\rm Htr}, x_{\rm Str}, x_{\rm Itr}, y_{\rm tr}\}\in R^{50\times 4}$. 剩余15组原始样本为测试样本,表示为$T_{\rm te}=\{X_{\rm te}, Y_{\rm te}\}=\{x_{\rm Hte}, x_{\rm Ste}, x_{\rm Ite}, y_{\rm te}\}\in R^{15\times 4}$. 下面设计6组对比实验:

**实验1**　基于SCN的真实小样本组分含量模型;

**实验2**　基于SCN隐含层插值的组分含量模型;

**实验3**　基于MD-MTD虚拟样本生成的组分含量模型;

**实验4**　基于GA优化MD-MTD虚拟样本生成的组分含量模型;

**实验5**　基于合成少数过采样技术(SMOTE)算法虚拟样本生成的组分含量预测模型;

**实验6**　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,使用最大相对误差(MRE)和均方根误差(RMSE)作为评价指标,即

$$\mathrm{MRE}=\max\left(\left|\frac{Y-\hat{Y}}{\hat{Y}}\right|\times 100\,\%\right),\tag{31}$$

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

其中:$\hat{Y}$、$Y$分别是测试集样本实际值和预测值,$n$是测试样本数量.

### 4.2 仿真结果

#### 4.2.1 合理性分析

为了验证本文所提方法生成的虚拟样本的合理性,将实验2、实验4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,与实验1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$r=[0.9,0.99,0.999,0.999\,9,0.999\,99,0.999\,999]$. GA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)、(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,考虑到实验2、实验4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)、(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,列出实验1、实验2、实验4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1 | | 实验2 | | 实验4 | |
|---|---|---|---|---|---|---|
| | 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2和实验4,当虚拟样本数量分别从0增加到50时,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2、实验4和实验6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6的81个虚拟样本.对比实验5,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6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 6组实验对比结果比较**

| 实验 | 虚拟样本数量 | MRE/% | RMSE |
|---|---|---|---|
| 1 | 0 | 4.65 | 1.75 |
| 2 | 50 | 3.00 | 0.77 |
| 3 | 50 | 4.20 | 1.23 |
| 4 | 50 | 2.89 | 0.52 |
| 5 | 50 | 3.95 | 1.19 |
| 6 | 81 | 2.84 | 0.38 |

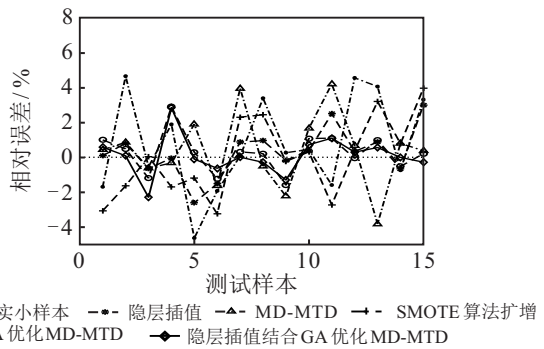


**图4 6组实验的相对误差变化**

分析表3和图4:对比实验2、实验5与实验1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,即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;对比实验2、实验4与实验5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;对比实验2、实验5与实验6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,即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7: 使用与实验1相同的训练集和测试集,分别采用SCN、支持向量机(SVM)、最小二乘支技向量机(LSSVM)、加权最小二乘支持向量机(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8: 将实验1中原有的61个真实样本数据与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的参数设置与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.

分析表5可知: 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;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.

**表4　数据扩增前后各模型参数设置**

| 模型参数 | SVM | | LSSVM | | WLSSVM | |
|---|---|---|---|---|---|---|
| | 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 | – | – | – | – |
| $\gamma$ | – | – | 330.142 6 | 353.755 7 | 199.457 | 342.769 2 |
| $\sigma$ | – | – | 0.437 1 | 0.517 9 | 0.009 8 | 0.013 6 |
| $W_H$ | – | – | – | – | 0.418 1 | 0.402 1 |
| $W_S$ | – | – | – | – | 0.396 5 | 0.368 8 |
| $W_I$ | – | – | – | – | 0.185 4 | 0.229 1 |
| $m_H$ | – | – | – | – | 2.277 6 | 2.333 2 |
| $m_S$ | – | – | – | – | 2.160 0 | 2.140 1 |
| $m_I$ | – | – | – | – | 1.009 8 | 1.329 4 |

**表5　数据扩增前后各模型的测试能值**

| 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 |

## 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)提出了基于SCN隐含层插值生成虚拟样本方法; 2)在MD-MTD方法基础上,提出了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符合期望空间,确保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3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地填充真实小样本间的信息间隔,提高数据样本的完备性并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]. 北京: 中国地质大学(北京), 2011.
(Shang Y. Research on the international competitiveness of China's rare earth industry[D]. Beijing: 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(2): 201-205.
(Tian H, Guo Z H, Li L Y. Soft-sensing in rare earth extraction[J]. Journal of the Chinese Society of Rare Earths, 2015, 33(2): 201-205.)

[3] 朱建勇, 张旭乾, 杨辉, 等. 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(2): 780-788.

(Zhu J Y, Zhang X Q, Yang H, et al. Soft-sensing of Pr/Nd component content under different single illumination conditions[J]. CIESC Journal, 2019, 70(2): 780-788.)

[4] 陆荣秀, 饶运春, 杨辉, 等. 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(8): 1846-1854.
(Lu R X, Rao Y C, Yang H, et al. Prediction of Pr/Nd component content based on improved just-in-time learning algorithm[J]. Control Theory & Applications, 2020, 37(8): 1846-1854.)

[5] 朱宝. 虚拟样本生成技术及建模应用研究[D]. 北京: 北京化工大学, 2017.
(Zhu B. Research on virtual sample generation technologies and their modeling application[D]. Beijing: Beijing University of Chemical Technology, 2017.)

[6] Chang C J, Li D C, Huang Y H, et al. 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]. Massachusetts: Massachusetts Institute of Technology Cambridge, 1992: 69-94.

[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, 38(3): 16-19.
(Yu X, Yang J, Xie Z Q. Research on virtual sample generation technology[J]. Computer Science, 2011, 38(3): 16-19.)

[9] 易令, 吕忠元, 丁进良, 等. 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): 2153-2160.
(Yi L, Lyu Z Y, Ding J L, et al. Data pretreatment approach for crude oil hydrogen properties prediction[J]. Control and Decision, 2018, 33(12): 2153-2160.)

[10] Li D C, Wu C S, Tsai T I, et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge[J]. Computers & 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(9): 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(9): 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, et al. 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: 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, 30(9): 835-838.
(Chen W, Ding Q L. Study on an XML approximately duplicated data cleaning method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2004, 30(9): 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, 24(4): 569-573.
(Jia W J, Chai T Y. Soft-sensor of element component content based on multiple models for the rare earth cascade extraction process[J]. Control Theory & Applications, 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, 8(6): 727-734.
(Wang C X, Zhang T, Ma C S. Improved SMOTE algorithm for imbalanced datasets[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(6): 727-734.)

## 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、控制与优化的研究, 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.

(责任编辑: 闫 妍)