基于虚拟样本生成技术的多组分机械信号建模

汤 健[1,2,3] 乔俊飞[1,2] 柴天佑[3] 刘 卓[3] 吴志伟[3]

摘 要 采用具有多组分、非平稳、非线性等特性的机械振动/振声信号构建数据驱动软测量模型, 是目前工业界测量高能耗旋转机械设备内部难以检测过程参数的常用手段. 针对机械信号产生机理的复杂性导致模型解释性弱, 以及工业过程连续不间断运行和机械设备旋转封闭的特殊性导致获取完备训练样本的经济性差和周期性长等问题, 本文提出一种基于虚拟样本生成 (Virtual sample generation, VSG) 技术的多组分机械信号建模方法. 首先, 将机械信号自适应分解为具有不同时间尺度的平稳子信号并变换为多尺度谱数据; 接着, 采用适合于小样本高维数据建模的改进选择性集成核偏最小二乘 (Selective ensemble kernel partial least squares, SENKPLS) 算法构建面向真实训练样本的基于可行性的规划 (Feasibility-based programming, FBP) 模型, 提出一种综合先验知识和 FBP 模型等手段面向高维谱数据的 VSG 技术, 用以弥补真实训练样本的短缺问题; 然后, 基于互信息 (Mutual information, MI) 对由真实和虚拟训练样本组成的混合建模数据进行自适应特征选择; 最后, 基于约简的混合训练样本采用 SENKPLS 构建软测量模型. 以近红外谱数据和磨矿过程实验球磨机的筒体振动/振声信号验证所提 VSG 技术和面向多组分机械信号建模方法的合理性和有效性.

关键词 多组分机械信号, 高维谱数据, 难以检测过程参数, 数据驱动建模, 虚拟样本生成

引用格式 汤健, 乔俊飞, 柴天佑, 刘卓, 吴志伟. 基于虚拟样本生成技术的多组分机械信号建模. 自动化学报, 2018, **44**(9): 1569−1589

DOI 10.16383/j.aas.2017.c170204

Modeling Multiple Components Mechanical Signals by Means of Virtual Sample Generation Technique

TANG Jian[1,2,3] QIAO Jun-Fei[1,2] CHAI Tian-You[3] LIU Zhuo[3] WU Zhi-Wei[3]

Abstract Mechanical vibration & acoustic signals with characteristics of multiple components, nonstationarity and nonlinearity are always used to construct the data-driven soft sensor model of industrial processes. It is one of the main approaches to measure the difficulty-to-measure process parameters inside those high energy consumption mechanical devices. Duo to the complexity of the production mechanism of these mechanical signals, most of these soft sensor models are difficult to be explained. Moreover, the characteristics of the industrial process′ continuous running and the mechanical equipment′ operation modes lead to the difficulty of high economic cost and long period waiting to obtain sufficient training samples. To solve these problems, a new multi-component mechanical signal modeling method based on virtual sample generation (VSG) technology is proposed. Firstly, the mechanical signals are processed into a set of sub-signals with different time scales by using adaptive multi-component signal decomposition technique; then these sub-signals are transferred to high dimensional multi-scale spectral data. Secondly, an improved selective ensemble kernel partial least squares (SENKPLS) algorithm that suits to model small sample high dimensional data is used to construct a feasibility-based programming (FBP) model with the true training samples; then prior knowledge, FBP models and information entropy are integrated to produce virtual training samples. Thirdly, mutual information (MI) method is used to select the spectral features of the new mixing training samples based on the true and virtual ones. Finally, a soft sensor model is built by using these reduced mixing spectral data. Near-infra spectra data and mechanical vibration and acoustic singals of a laboratory-scale ball mill in grinding process validate the reasonability and effectiveness of the proposed VSG techniques and multi-component mechanical signals-based modeling approach.

Key words Multi-component mechanical signal, high dimentional spectra data, difficulty-to-measure process parameters, data-driven modeling, virtual sample generation (VSG)

Citation Tang Jian, Qiao Jun-Fei, Chai Tian-You, Liu Zhuo, Wu Zhi-Wei. Modeling multiple components mechanical signals by means of virtual sample generation technique. *Acta Automatica Sinica*, 2018, **44**(9): 1569−1589

收稿日期 2017-04-16 录用日期 2017-06-22
Manuscript received April 16, 2017; accepted June 22, 2017
国家自然科学基金 (61573364, 61703089), 流程工业综合自化国家重点实验室开放课题基金资助项目 (PAL-N201504), 矿冶过程自动控制技术国家重点实验室矿冶过程自动控制技术北京市重点实验室 (BGRIMM-KZSKL-2017-07) 资助
Supported by National Natural Science Foundation of China (61573364, 61703089), State Key Laboratory of Synthetical Automation for Process Industries (PAL-N201504), and State Key Laboratory of Process Automation in Mining & Metallurgy Beijing Key Laboratory of Process Automation in Mining & Metallurgy (BGRIMM-KZSKL-2017-07)
本文责任编委 侯忠生
Recommended by Associate Editor HOU Zhong-Sheng
1. 北京工业大学信息学部 北京 100124 2. 计算智能与智能系统北

工业过程的优化运行控制[1] 需要准确检测与生产过程的质量、产量、能耗等指标密切相关的难以检测的过程参数[2−3], 如磨矿过程广泛使用的大型机械设备球磨机内部的料球比、磨矿浓度、充填率等[4−5]. 这些过程参数的实时准确检测一直是工业界亟待解决的难题[6]. 由于流程工业过程的综合复杂特性, 以及机械设备连续旋转和封闭运行的特点, 其内部过程参数难以通过直接检测方式和建立机理模型计算得到[7]. 虽然运行专家可以依据多源信息和多年工作经验对所熟悉的机械设备内部的过程参数进行较为准确的估计, 但专家经验的差异性和精力的有限性难以保证工业过程长期运行在优化状态. 基于这些设备工作中产生的机械振动和振声信号构建数据驱动软测量模型, 是目前该领域重点关注的热点和难点问题[8]. 下文分别从多组分机械信号建模、建模样本非完备和本文研究动机等 3 方面予以描述.

0.1 多组分机械信号建模

机械振动和振声信号通常具有较强的多组分、非线性和非平稳等特性. 基于这些信号进行难以检测过程参数软测量的首要难点问题是: 如何从机械信号中提取模型的输入特征. 这通常包括信号处理和维数约简两个子问题. 通常, 机械信号中蕴含的有价值信息被隐含在宽带随机噪声中[9]. 以磨矿过程的关键设备球磨机为例, 磨机负荷参数与磨机筒体振动和振声信号的功率谱密度 (Power spectral density, PSD) 密切相关[10]. 研究表明, 快速傅里叶变换 (Fast Fourier transform, FFT) 并不适合于具有非平稳特性的振动信号的处理[11]. 离散小波变换 (Discrete wavelet transform, DWT)、连续小波变换 (Continuous wavelet transform, CWT)、小波包变换等时频分析方法已经被广泛应用于基于机械振动信号的故障诊断[12−15], 但这些方法不能自适应分解这些多组分机械信号, 如面对任何一个具体实际应用问题需要为 CWT 选择合适的母小波. 经验模态分解 (Empirical mode decomposition, EMD) 技术突破上述局限, 通过自适应分解获取具有不同时间尺度和物理含义的内禀模态函数 (Intrinsic mode function, IMF)[16], 广泛用于处理多组分机械信号[17−19]. 文献 [20−21] 采用 EMD 对磨矿过程的球磨机筒体振动信号进行处理, 分解得到系列具有不同物理含义和不同时间尺度的子信号. 研究表明, EMD 算法存在频谱分辨率低、虚假组分易造成模态混叠、低能量成分不可分等问题. 集成 EMD (Ensemble EMD, EEMD) 技术克服了 EMD 的模态混合问题[22], 但仍存在蕴含有价值信息的子信号数量有限及其时频特征难以选择等问题.

因机械振动和振声的频域特征明显, 将其进行时频变换是通常采用的处理手段之一. 文献 [23] 提出: 将机械信号直接变换至频域的数据称为单尺度频谱, 而将经多组分信号技术分解得到的不同时间尺度子信号变换至频域的数据称为多尺度频谱. 显然, 当频率分辨率较高 (如 1 Hz) 时, 单/多尺度频谱的维数均高达数千维并且谱特征间具有较强的共线性. 因此, 维数约简成为基于机械信号构建软测量模型需要面对的又一问题[24]. 常用的基于遗传算法 (Genetic algorithm, GA) 的频谱数据特征选择算法存在运行时间长和效率低等问题[25]. 针对机械振动频谱, 文献 [26] 提出基于互信息 (Mutual information, MI) 和潜结构模型的 IMF 及其多类谱特征的自适应选择方法. 研究表明, MI 能有效描述输入和输出数据间的映射关系, 并且更易于理解[27]; 但存在寻优耗时等问题.

我们认为, 构建基于机械频谱特征的难以检测过程参数软测量模型应尽可能地模拟运行专家选择多维度有价值信息进行过程参数认知的机制. 通常, 机械振动/振声信号的多源多尺度频谱是由具有不同物理含义和时间尺度的子信号经时频变换获得. 基于这些谱数据构建选择性集成 (Selective ensemble, SEN) 软测量模型的过程在本质上是选择性的信息融合过程[6]. 文献 [21, 23] 基于这样的思路, 构建了基于分支定界 (Branch and bound, BB) 优化算法和自适应加权融合 (Adaptive weighted fusion, AWF) 算法的 SEN 核潜结构映射或核偏最小二乘 (Kernel partial least squares, KPLS) 算法的软测量模型. 从集成学习理论的视角出发, 这些方法均属于 “操纵输入特征” 的集成构造策略进行模型构建, 所优化选择的是 “有价值的多源信息”. 工业实际中, 运行专家识别难以检测过程参数不仅需要选择有价值的多源信息进行融合, 还需要利用自身积累的历史经验, 即也要选择有价值样本进行认知. 基于遗传算法的选择性集成 (GA-based selective ensemble, GASEN)[28]. 采用 “操纵训练样本” 策略构造集成、采用误差反向传播神经网络 (Back propagation neural networks, BPNN) 构建候选子模型、采用 GA 优选集成子模型和简单加权平均组合集成子模型; 针对 BPNN 训练时间长、容易过拟合和 GASEN 难以采用高维小样本数据直接建模等缺点, 文献 [26] 提出了基于 “操纵训练样本” 集成构造策略的改进选择性集成核偏最小二

京市重点实验室 北京 100124 3. 流程工业综合自动化国家重点实验室 沈阳 110004
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124 3. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004

乘 (Selective ensemble kernel partial least squares, SENKPLS) 算法; 文献 [29] 提出了采用双层 GA 优化改进 SENKPLS 的建模参数; 文献 [30] 提出基于稀疏非线性特征的 KPLS. 显然, 我们需要合适的软测量策略模拟运行专家融合多维度信息的认知机制.

0.2 建模样本非完备

在实际工业应用过程中, 难以检测过程参数建模经常遇到的更为棘手的难点是: 如何获得能够覆盖多种工况和充足完备的建模样本. 研究表明, 充足完备的建模样本对于构建有效的学习模型非常重要. 通常, 流程工业中难以检测过程参数的建模样本仅能在实验设计阶段或工业过程停产重新运行的起始阶段获得; 否则, 需要以牺牲企业的经济效益或较长周期的时间等待为代价. 如何基于短缺、非完备样本构建鲁棒的面向工业应用的数据驱动模型, 一直以来都是个开放性的难题.

为提高模型的泛化性能, 图像识别领域首次提出了基于先验知识从给定小规模真实训练样本产生虚拟训练样本的方法[31−33], 即虚拟样本产生 (Virtual sample generation, VSG) 技术. 目前的已有研究包括: 利用 BPNN 和巨型趋势分散技术[34]、利用运行专家知识[35]、利用噪声构造[36]、利用原始样本分布函数[37] 等. 对面向高维小样本数据的分类问题, 文献 [38] 提出了基于分组的 VSG 用于脱氧核糖核酸 (Deoxyribonucleic acid, DNA) 微阵列数据建模. 上述研究中的 VSG 多面向分类问题[37, 39]. 本文主要关注如何利用 VSG 辅助构建基于多源多尺度谱数据的软测量模型, 即面向回归问题的 VSG.

针对回归建模问题, 文献 [40] 提出基于多层感知器网络的 VSG 技术, 即虚拟样本输入通过选择真实样本输入附近的点产生, 虚拟样本输出通过平均多层感知器网络的不同输出数据获得; 文献 [41] 提出用分散神经网络 (Decentralized neural networks, DNN) 产生虚拟样本和建模小数据集, 表明 DNN 比 BPNN 具有更强的预测性能; 文献 [42−44] 提出基于遗传算法 (GA)、粒子群优化 (Particle swarm optimization, PSO) 算法以及蒙特卡洛与 PSO 相结合的 VSG; 文献 [45] 提出一种的产生通用结构数据的采样方法; 但上述这些方法多采用传统的单模型产生虚拟样本. 针对具有复杂分布的建模数据或高维小样本数据, 传统的单模型难以有效解决模式识别或回归建模等问题. 文献 [46] 提出的 VSG 难以直接面对高维谱数据建模.

针对真实数据与虚拟数据混合利用问题, 文献 [47] 提出将基于原始数据的特定小数据和虚拟的人工数据相结合, 然后对机器学习模型进行更新的策略; 文献 [48] 提出平行学习的概念, 并将其作为机器学习研究方向的一个新型理论框架, 在该框架中提出通过混合人工数据和原始数据进行基于计算实验的预测学习和集成学习. 可见, 将 VSG 结合具体的工业背景进行研究具有重要的理论和现实意义.

0.3 本文研究动机

综上可知, 面对基于多组分机械振动/振声信号的流程工业难以检测过程参数的软测量问题, 有以下问题需要解决: 1) 如何将蕴含着丰富难以检测过程参数信息的多组分机械振动/振声信号自适应地分解为具有不同物理含义的不同组成成分, 为构建寓意明确的软测量模型和探究旋转机械设备内部的工作机理与振动产生机理奠定基础; 2) 如何基于确定性的先验知识和非充足完备的真实训练样本提出适合于高维谱数据的 VSG 技术; 3) 针对基于多组分机械信号, 在利用 VSG 构建虚拟样本输出时需要解决: 如何依据不同时间尺度子信号的谱特征所构建的集成子模型对软测量模型预测输出的贡献率, 对集成子模型的虚拟样本输出进行加权组合以获得统一的虚拟样本输出; 4) 如何选择有价值子信号及其高维频域特征, 以及如何选择性融合多源多尺度谱特征和多工况样本, 以便更有效地模拟工业现场运行专家的认知机制. 因此, 基于多组分机械信号的软测量建模可以归结为一类针对多源多尺度高维谱数据的小样本建模问题.

针对上述需要解决的问题, 综合之前的研究成果, 本文首先提出了一种面向多源多尺度高维谱数据的 VSG 技术用以解决建模样本的短缺非完备问题; 再以混合训练样本结合 MI 自适应特征选择技术获取多源输入特征, 并基于改进的 SENKPLS 算法从操纵训练样本集成构造视角构建软测量模型; 最后采用近红外谱 (Near infrared spectrum, NIR) 数据和磨矿过程实验球磨机筒体振动与振声信号构建的软测量模型验证本文所提出的 VSG 技术和面向多组分机械信号建模方法的合理性和有效性.

1 相关知识

此处主要对本文所采用算法和技术的相关基础知识进行简短描述.

1.1 面向多组分机械信号的建模

1.1.1 多组分机械信号自适应分解

满足特定假设条件的多组分、非线性、非平稳机械信号采用 EMD 可以分解为若干个不同时间尺度的 IMF 和残差之和. 这些理论上均有其物理含义的 IMF 子信号按照频率由高到低排列. 多组分机械

信号与 IMF 信号间的关系可表示为

$$\boldsymbol{x}^{\mathrm{t}} = \sum_{j=1}^{J_{\mathrm{EMD}}} \boldsymbol{x}^{\mathrm{t}}_{\mathrm{EMD}_j} + r^{\mathrm{t}}_{J_{\mathrm{EMD}}} \tag{1}$$

其中, $r^{\mathrm{t}}_{J_{\mathrm{EMD}}}$ 表示分解后的残差.

EMD 在处理机械振动和振声等多组分信号较传统 FFT 和小波变换具有明显优势, 但也存在虚假人工成分导致的模态混叠、分解端点效应、子信号非严格正交、有效子信号数量有限等问题. 基于白噪声统计属性的 EEMD 可以有效克服 EMD 的模态混叠问题, 其基本思路是加入影响整个时频空间的白噪声 A_{noise}, 重复进行整个 EMD 分解过程 M 次后进行平均. EEMD 和 EMD 之间的关系表示为

$$\begin{cases} \boldsymbol{x}^{\mathrm{t}}_{\mathrm{EEMD}_j} = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} (\boldsymbol{x}^{\mathrm{t}}_{\mathrm{EMD}_j})_m \\ \boldsymbol{r}^{\mathrm{t}}_{\mathrm{EEMD}} = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} (\boldsymbol{r}^{\mathrm{t}}_{\mathrm{EMD}_J})_m \end{cases} \tag{2}$$

其中, $\boldsymbol{x}^{\mathrm{t}}_{\mathrm{EMD}_j}$ 表示第 mth 个 EMD 分解的第 jth 个 IMF 子信号, $\boldsymbol{r}^{\mathrm{t}}_{\mathrm{EEMD}_J}$ 表示分解后的残差.

显然, EEMD 的计算消耗与 EMD 相比成倍增长. 针对 EMD 的缺点, 已有研究包括小波包 EMD (Wavelet packet decompsition (WPT)-EMD)、在线 EMD 及各种基于预测模型的端点延拓算法等. 研究表明: 不同算法具有各自的优缺点, 需要结合应用背景确定; 有价值子信号的数量也是有限的, 需要进行优选.

1.1.2 高维谱数据的维数约简

特征提取和特征选择技术均可有效处理高维谱数据的维数约简问题. 特征提取采用线性或是非线性的方式确定合适的低维潜在特征替代原始高维特征, 优点是考虑了全部特征的多数变化信息, 缺点是所提取特征较难解释, 并且所丢弃的残差因含有全部输入变量的信息而可能具有更高的建模贡献率. 常用的基于主成分分析 (Principal component analysis, PCA) 的特征提取以低维潜在特征表征原始高维数据, 但未考虑输入和输出间的相关性[49]. 基于潜结构映射或偏最小二乘 (Projection to latent structure or partial least squares, PLS) 的特征提取可克服这一缺陷, 可逐层提取同时表征输入输出间变化的潜在变量[50]. 文献 [51] 提出了基于 PLS 的通用特征提取框架. 面向多尺度频谱, 文献 [52] 提出了基于球域准则和 PLS 的多尺度频谱特征选择方法. 工业过程的非线性本质使得为输入数据扩展非线性项而进行非线性特性提取的核方法广泛应用[53−57], 如核 PCA (Kernel PCA, KPCA)、核独立主元分析 (Kernel independent principal component analysis, KICA)、核费舍尔判别分析 (Kernel Fisher discriminant analysis, KFDA) 和 KPLS[58−59] 等. 特征选择技术依据某种准则优选重要特征, 其优点是所选择的特征易于解释, 缺点是未被选择的特征可能会降低软测量模型的泛化性能[60]. 研究表明, 相对于其他方法, 基于 MI 的特征选择更易于理解和更有效. 文献 [61] 提出基于 MI 的潜在特征度量方法, 对特征提取与特征选择技术进行了有效组合.

因此, 结合不同的应用需求, 研究不同的维数约简算法是面对实际应用问题的有效策略之一.

1.1.3 基于改进 SENKPLS 的高维谱数据建模

PLS/KPLS 算法除用于提取与输入输出均相关的潜在特征外, 也可直接构建潜结构模型, 类似于人脑逐层进行特征抽取进行认知的机制. 面对多源机械振动/振声信号特征子集, 基于 “操纵输入特征” 的集成构造策略, 文献 [24] 从选择性信息融合的视角构建 SENKPLS 模型.

此处重点描述基于 “操纵训练样本” 集成策略的改进 SENKPLS 算法; 与 GASEN 对比, 该方法可以直接构建基于小样本高维谱数据的软测量模型. 另, 除非特别说明, 本文中的输出均为单变量输出.

集成构造的过程是采用 Bootstrap 算法基于有放回的抽样原则在原始训练样本 $\{(\boldsymbol{z}, y)_l\}_{l=1}^{k}$ 中产生 J 个训练子样本 $\{\{(\boldsymbol{z}^j, y^j)_l\}_{l=1}^{k}\}_{j=1}^{J}$, 其中 J 也是候选子模型数量和 GA 优化种群数量. 下文以第 jth 个训练子样本 $\{(\boldsymbol{z}^j)_l\}_{l=1}^{k}$ 为例对候选子模型的构建过程进行描述.

首先将训练子样本映射到高维空间:

$$K^j = \Phi^{\mathrm{T}}((\boldsymbol{z}^j)_l)\Phi((\boldsymbol{z}^j)_m), \quad l, m = 1, 2, \cdots, k \tag{3}$$

其中, K^j 采用下式中心化:

$$\tilde{K}^j = \left(I - \frac{1}{k}\boldsymbol{l}_k\boldsymbol{l}_k^{\mathrm{T}}\right) K^j \left(I - \frac{1}{k}\boldsymbol{l}_k\boldsymbol{l}_k^{\mathrm{T}}\right) \tag{4}$$

其中, I 是单位阵, $\boldsymbol{l}_k$ 是值为 1, 长度为 k 的向量.

此处, 为所有候选子模型选择相同的核参数 K_{para}和核潜变量 (Kernel latent variable, KLV) 数量 h_{KLV}, 并将全部候选子模型的集合标记为 $\{f^j(\cdot)\}_{j=1}^{J}$. 验证样本 $\{(\boldsymbol{z}_{\mathrm{valid}})_l\}_{l=1}^{k^{\mathrm{valid}}}$ 基于第 j_{th} 个候选子模型的预测输出为

$$\hat{\boldsymbol{y}}^j_{\mathrm{valid}} = f^j(\{(\boldsymbol{z}_{\mathrm{valid}})_l\}_{l=1}^{k^{\mathrm{valid}}}) = \tilde{K}^j_{\mathrm{valid}} U^j \left((T^j)^{\mathrm{T}} \tilde{K}^j U^j\right)^{-1} (T^j)^{\mathrm{T}} \boldsymbol{y}^j \tag{5}$$

$$\tilde{K}^j_{\mathrm{valid}} = \left(K^j_{\mathrm{valid}} I - \frac{1}{k}\boldsymbol{l}_{k^{\mathrm{valid}}}\boldsymbol{l}_k^{\mathrm{T}} K^j\right)\left(I - \frac{1}{k}\boldsymbol{l}_k\boldsymbol{l}_k^{\mathrm{T}}\right) \tag{6}$$

$$K^j_{\mathrm{valid}} = K((\boldsymbol{z}_{\mathrm{valid}})_l, (\boldsymbol{z}^j)_m) \tag{7}$$

其中, k^{valid} 是验证样本的数量. 验证样本的预测误差采用下式计算:

$$\boldsymbol{e}^{j}_{\text{valid}} = \hat{\boldsymbol{y}}^{j}_{\text{valid}} - \boldsymbol{y}_{\text{valid}} \tag{8}$$

第 jth 个和第 sth 个候选子模型的相关系数为

$$c^{\text{valid}}_{js} = \frac{\sum\limits_{l=1}^{k^{\text{valid}}} \boldsymbol{e}^{j}_{\text{valid}}(j,k^{\text{valid}}) \cdot \boldsymbol{e}^{s}_{\text{valid}}(s,k^{\text{valid}})}{k^{\text{valid}}} \tag{9}$$

构建如下的相关系数矩阵:

$$C^{\text{valid}} = \begin{bmatrix} c^{\text{valid}}_{11} & c^{\text{valid}}_{12} & \cdots & c^{\text{valid}}_{1J} \\ c^{\text{valid}}_{21} & c^{\text{valid}}_{22} & \cdots & c^{\text{valid}}_{2J} \\ \vdots & \vdots & c^{\text{valid}}_{js} & \vdots \\ c^{\text{valid}}_{J1} & c^{\text{valid}}_{J2} & \cdots & c^{\text{valid}}_{JJ} \end{bmatrix}_{J\times J} \tag{10}$$

基于 C^{valid} 采用 GA 优化工具箱 (GAOT) 优化候选子模型的随机向量 $\boldsymbol{w}_J = [w_1, \cdots, w_j, \cdots, w_J]$, 并将优化结果记为 $\boldsymbol{w}^* = [w^*_1, \cdots, w^*_j, \cdots, w^*_J]$. 采用如下准则选择集成子模型:

$$\xi_j = \begin{cases} 1, & w^*_j \geq \lambda \\ 0, & w^*_j < \lambda \end{cases} \tag{11}$$

其中, λ 是集成子模型的选择阈值.

将 $\xi_j = 1$ 的候选子模型选择为集成子模型, 并将其数量记为 J^*, 即 SEN 模型的集成尺寸. 将第 j^*th 集成子模型记为 $f^{j^*}(\cdot)$, 其输出为

$$\hat{\boldsymbol{y}}^{j^*}_{\text{valid}} = f^{j^*}(\{(\boldsymbol{z}_{\text{valid}})_l\}^{k^{\text{valid}}}_{l=1}) \tag{12}$$

将全部集成子模型记为 $\{f^{j^*}(\cdot)\}^{J^*}_{j^*=1}$. 考虑到不同集成子模型的贡献率, 采用自适应加权融合 (AWF) 算法获取集成子模型的权重, 等同于求解如下优化问题:

$$\begin{aligned} \min \quad & \sigma^2 = \sum_{j^*=1}^{J^*} (W^{\text{AWF}}_{j^*})^2 \sigma^2_{j^*} \\ \text{s.t.} \quad & \sum_{j^*=1}^{J^*} W^{\text{AWF}}_{j^*} = 1, \ \ 0 \leq W^{\text{AWF}}_{j^*} \leq 1 \end{aligned} \tag{13}$$

其中, σ 是 SEN 模型预测值 $\hat{\boldsymbol{y}}$ 的方差, σ_{j^*} 是集成子模型预测值 $\hat{\boldsymbol{y}}^{j}_{\text{valid}}$ 的方差.

集成子模型的权重采用下式计算:

$$W^{\text{AWF}}_{j^*} = \frac{1}{(\sigma_{j^*})^2 \sum\limits_{j^*=1}^{J^*} \frac{1}{(\sigma_{j^*})^2}} \tag{14}$$

基于上述建模算法, 单个测试样本的输出为

$$\hat{y}_{\text{test}} = \sum_{j^*=1}^{J^*} W^{\text{AWF}}_{j^*} \hat{y}^{j^*}_{\text{test}} \tag{15}$$

其中, $\hat{y}^{j^*}_{\text{test}}$ 是第 j^*th 个集成子模型的预测输出.

1.2　面向建模样本非完备的 VSG 技术

1.2.1　小样本数据集概述

在多数工业过程中, 数据采集与存储系统所带来的多是模型输入维数和低价值训练样本的增加, 这使得输入特征高维、训练样本不完备等问题仍然较为突出[62]. 研究表明, 数量充足、覆盖工况完备的建模样本对构建有效的软测量模型非常重要. 目前, 关于小样本数据的定义具有较大的相对性和主观性[62].

为了确定获得必要的预测性能而需要的最小训练样本的数量, 研究人员提出了概率近似正确、训练样本和输入特征比率等指标[63−64]. 在模式识别领域, 通常认为训练样本数量与输入特征之比应该足够大, 其相互关系可表示为

$$\alpha = \frac{n_{\text{sample}}}{p_{\text{feature}}} \tag{16}$$

其中, n_{sample} 和 p_{feature} 分别表示训练样本和输入特征的数量. 通常, α 的取值为 2, 5 或 10.

文献 [65] 面向分类问题, 研究了分类误差、训练样本数量、输入特征维数和分类算法复杂性间的相互关系. 针对一些典型的分类器, 文献 [66] 描述需要充足完备训练样本的内在原因, 并着重研究在 $\alpha \leq 1$ 时的分类器性能, 即研究 n_{sample} 小于 p_{feature} 时线性分类器泛化性能. 此处, 记维数约简后的特征为 $p_{\text{feature\_redu}}$, 并定义如下指标:

$$\alpha_{\text{redu}} = \frac{n_{\text{sample}}}{p_{\text{feature\_redu}}} \tag{17}$$

若经维数约简后的 α_{redu} 值仍然难以满足构建具有鲁棒预测性能的学习模型的要求, 必须采用其他方法解决训练样本的短缺问题.

1.2.2　虚拟样本的定义

虚拟样本的定义源于图像识别领域: 基于先验知识从物体的 3D 视角出发, 将通过数学变换产生的新图像称为虚拟样本. 文献 [67] 对虚拟样本给出了如下较为通用的定义:

定义 1. 将 $\{\boldsymbol{x}_l, y_l\}$ 记为真实样本, 其中, $\boldsymbol{x}, y \in \mathbf{R}^n$, k 是真实样本数量; 基于先验知识 $Know$, 采用变换 $\{T, f(T)\}$ 产生新样本 $\{T\boldsymbol{x}, f(T)y\}$, 即:

$$\left.\begin{matrix} \{\boldsymbol{x}_l, y_l\} \\ Know \end{matrix}\right\} \xrightarrow{\{T, f(T)\}} \{T\boldsymbol{x}, f(T)y\} \tag{18}$$

这个新样本 $\{T\boldsymbol{x}, f(T)y\}$ 被称为虚拟样本. 变换 $\{T, f(T)\}$ 所采用方法依据应用背景不同而具有差异性.

基于上述定义, 本文给出如下推论:

推论. 给定真实样本数据集 $\{\boldsymbol{x}_l, y_l\}_{l=1}^{k}$, 通过适当的变换可产生虚拟样本数据集 $\{\boldsymbol{x}_{l'}^{\rm VS}, y_{l'}^{\rm VS}\}_{l'=1}^{k^{\rm VS}}$. 该过程可用如下公式表示:

$$\left\{\begin{array}{l}\{\boldsymbol{x}_l\}_{l=1}^{k} \xrightarrow{T} \{\boldsymbol{x}_{l'}^{\rm VS}\}_{l''=1}^{k^{\rm VS}}, \quad \boldsymbol{x}_{\rm low}^{\rm VS} \le \boldsymbol{x}_{l'}^{\rm VS} \le \boldsymbol{x}_{\rm high}^{\rm VS} \\ \left.\begin{array}{l}\{\boldsymbol{x}_{l'}^{\rm VS}\}_{l''=1}^{k^{\rm VS}} \\ \{y_l\}_{l=1}^{k}\end{array}\right\} \xrightarrow{f(T)} \{y_{l'}^{\rm VS}\}_{l''=1}^{k^{\rm VS}}, \quad y_{\rm low}^{\rm VS} \le y_{l'}^{\rm VS} \le y_{\rm high}^{\rm VS}\end{array}\right. \tag{19}$$

其中, $\boldsymbol{x}_{\rm low}^{\rm VS}$ 与 $\boldsymbol{x}_{\rm high}^{\rm VS}$ 和 $y_{\rm low}^{\rm VS}$ 与 $y_{\rm high}^{\rm VS}$ 分别是虚拟样本输入和输出值的下限与上限.

1.2.3 VSG 的关注问题

通常, 若要产生合理的虚拟样本, 需要关注至少如下 4 个问题:

1) 确定虚拟样本产生策略. 方法包括: 利用先验知识、扰动原始样本、对输入数据添加噪声等[67];

2) 确定虚拟样本输入. 方法包括: 先验知识[31]、在真实样本输入点的超域内随机选取[40]、函数化虚拟群体[68]、基于真实样本输入间隔的信息分散技术[41]、间隔核密度估计[37]、Mega 趋势分散 (MTD) 函数[34]、基于模糊数据集的成员函数[69]、组虚拟样本产生[38]、基于模糊理论的产生趋势分散 (GTD)[70]、基于高斯分布[67]、基于 GA 和 BPNN[42];

3) 确定虚拟样本输出. 方法包括: 平均神经网络的输出[40]、基于实际样本输出间隔的信息分散技术[41] 和基于 GA 和 BPNN 的方法[42] 等;

4) 确定虚拟样本数量. 目前, 最优化虚拟样本数量的确定主要基于实验数据确定. 如何在理论上指导或确定优化的虚拟样本数量还是个开放性难题.

另外, 对于分类模型和回归模型, VSG 所关注的问题是具有差异性的, 原因在于: 分类问题的类标是固定的, 变换 $f(T)$ 也许是不必要的; 回归问题的每个输入均需要唯一对应一个输出, 这增大了 VSG 的难度. 笔者认为, 这也是当前关于回归问题的 VSG 研究要远少于分类问题的原因之一.

综上, 结合具体的应用背景进行 VSG 回归问题的实用化研究将更具有挑战性和现实意义, 也是目前研究中的难点之一.

2 基于 VSG 的多组分机械信号建模策略

本文提出的基于 VSG 的多组分机械信号建模策略主要由: 多尺度谱数据获取、虚拟样本产生、谱特征自适应选择和软测量模型构建共 4 个模块组成, 如图 1 所示.

图 1 中, $\{\boldsymbol{x}_{\rm V}^{\rm t}\}_{l=1}^{k}$ 和 $\{\boldsymbol{x}_{\rm A}^{\rm t}\}_{l=1}^{k}$ 分别表示包含 k 个建模样本的机械设备振动和振声时域信号的集合, 其中 $\boldsymbol{x}_{\rm V}^{\rm t}=\{(x_{\rm V}^{\rm t})_n\}_{n=1}^{N}$ 和 $\boldsymbol{x}_{\rm A}^{\rm t}=\{(x_{\rm A}^{\rm t})_n\}_{n=1}^{N}$ 是包含 N 个数据点的相应信号; $f^{\rm DCOM}(\cdot)$ 表示多组分机械信号的自适应分解函数; $f^{\rm Trans}(\cdot)$ 表示频域特征转换函数; $Z^{\rm True}$ 和 $\boldsymbol{y}^{\rm True}$ 分别表示真实训练样本的输入和输出; $S^{\rm True}=\{Z^{\rm True}, \boldsymbol{y}^{\rm True}\}$ 表示真实训练样本的输入输出集; $\{f_{j_{\rm IMF}}^{\rm FBP}(\cdot)\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}$ 表示采用 $S^{\rm True}$ 构建的全部 IMF 的 FBP 模型集; $\{f_{j_{\rm IMF}}^{\rm VSG}(\cdot)\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}$ 表示用于面向全部 IMF 的 VSG 函数集, 其输入为

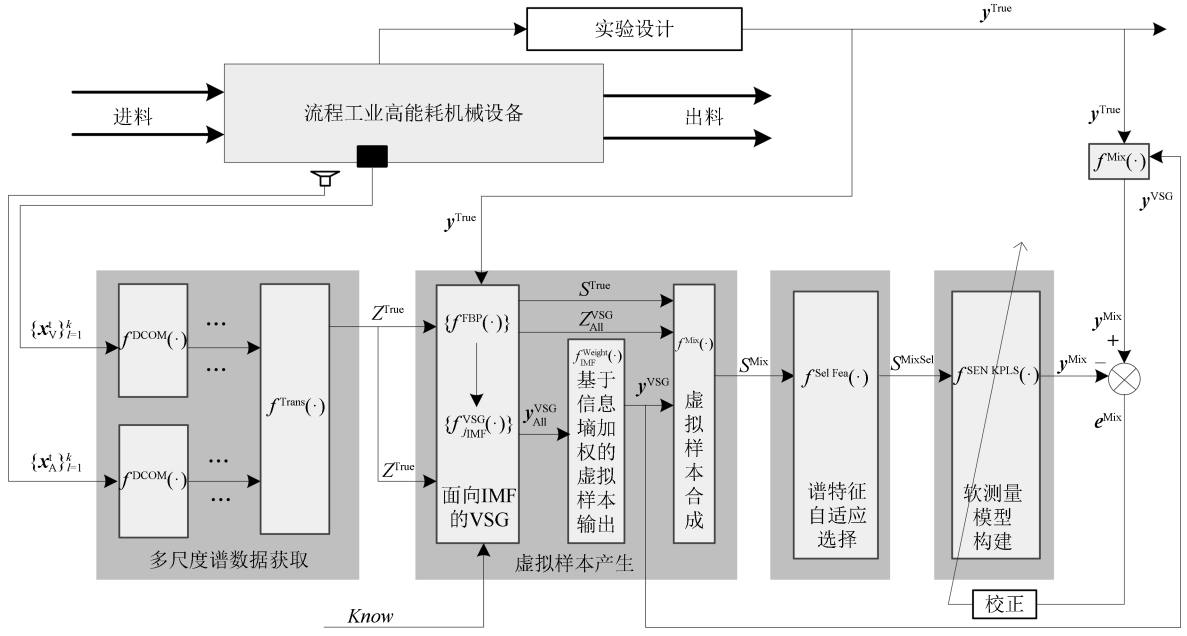


图 1 基于 VSG 的多组分机械信号建模策略

Fig. 1 Multi-component mechanical signal modeling strategy based on VSG

S^{True} 和先验知识 $Know$, 输出为全部 IMF 的虚拟样本集 $S_{\mathrm{All}}^{\mathrm{VSG}}=\{Z_{\mathrm{All}}^{\mathrm{VSG}},\boldsymbol{y}_{\mathrm{All}}^{\mathrm{VSG}}\}$; $f_{\mathrm{IMF}}^{\mathrm{Weight}}(\cdot)$ 表示基于信息熵加权不同 IMF 虚拟样本输出的函数; $f^{\mathrm{Mix}}(\cdot)$ 表示用于获得混合样本集 S^{Mix} 的函数; $f^{\mathrm{SelFea}}(\cdot)$ 表示混合样本集的特征选择函数; S^{MixSel} 表示经特征约简后的混合样本集; $f^{\mathrm{SENKPLS}}(\cdot)$ 表示基于约简混合样本集构建的软测量模型; $\boldsymbol{y}_{\mathrm{All}}^{\mathrm{VSG}}$ 表示全部 IMF 的虚拟样本输出值的集合; $\boldsymbol{y}^{\mathrm{VSG}}$ 表示加权融合后的虚拟样本输出; $\boldsymbol{y}^{\mathrm{Mix}}$ 和 $\hat{\boldsymbol{y}}^{\mathrm{Mix}}$ 表示软测量模型的真值和预测值.

图 1 中不同模块的功能描述如下:

1) 多尺度谱数据获取模块: 其输入为真实的时域机械振动/振声信号, 输出为真实的频域多尺度训练样本; 主要功能是将包含若干数据点的多组分时域信号经自适应分解和时频域转换得到多源多尺度高维谱数据.

2) 虚拟样本产生模块: 是本文所提方法的核心模块, 其输入为真实的训练样本和先验知识, 输出为混合训练样本; 主要功能包括: 面向 IMF 的 VSG, 基于信息熵加权 IMF 虚拟样本输出, 以及虚拟样本合成.

3) 谱特征自适应选择: 其输入为混合训练样本, 输出为经特征选择的约简混合训练样本; 主要功能是自适应地选择有价值的多源多尺度子信号及其谱特征.

4) 软测量模型构建: 其输入为约简混合样本, 输出为难以检测过程参数的预测值; 主要功能是构建基于 "操纵训练样本" 策略的适合于高维谱数据的 SENKPLS 模型.

因此, 上述不同模块分别实现了多组分信号的自适应分解、基于先验知识和真实训练数据的虚拟样本产生、多组分机械信号不同 IMF 集成子模型虚拟样本输出加权融合、多源多尺度特征的自适应选择以及仿运行专家综合多源特征和多工况样本认知机制的建模. 其中, 虚拟样本产生模块是本文所提方法的核心.

3　基于 VSG 的多组分机械信号建模实现

3.1　多尺度谱数据获取模块

以包含 N 个数据点的单个建模样本为例, EEMD 可将机械振动和振声信号分解为

$$\boldsymbol{x}_{\mathrm{V}}^{\mathrm{t}}\xrightarrow{f^{\mathrm{DCOM}(\cdot)}}\sum_{j_{\mathrm{V}}=1}^{J_{\mathrm{V}}^{\mathrm{all}}}(\bar{\boldsymbol{c}}_{\mathrm{V}}^{\mathrm{t}})_{j_{\mathrm{V}}}+(\bar{r}_{\mathrm{V}})_{J_{\mathrm{V}}^{\mathrm{all}}}\tag{20}$$

$$\boldsymbol{x}_{\mathrm{A}}^{\mathrm{t}}\xrightarrow{f^{\mathrm{DCOM}(\cdot)}}\sum_{j_{\mathrm{A}}=1}^{J_{\mathrm{A}}^{\mathrm{all}}}(\bar{\boldsymbol{c}}_{\mathrm{A}}^{\mathrm{t}})_{j_{\mathrm{A}}}+(\bar{r}_{\mathrm{A}})_{J_{\mathrm{A}}^{\mathrm{all}}}\tag{21}$$

其中, $(\bar{\boldsymbol{c}}_{\mathrm{V}}^{\mathrm{t}})_{j_{\mathrm{V}}}$ 和 $(\bar{\boldsymbol{c}}_{\mathrm{A}}^{\mathrm{t}})_{j_{\mathrm{A}}}$ 表示第 j_{V}th 和 j_{A}th 个 IMF 子信号, $(\bar{r}_{\mathrm{V}})_{J_{\mathrm{V}}^{\mathrm{all}}}$ 和 $(\bar{r}_{\mathrm{A}})_{J_{\mathrm{A}}^{\mathrm{all}}}$ 表示残差信号.

用于构建软测量模型的 IMF 的最大数量依据空载时机械设备振动分解结果和先验知识确定. 本文将构建软测量模型的机械振动和振声 IMF 的数量标记为 J_{A} 和 J_{V}, 并将 IMF 重新编号, 如下式所示:

$$\begin{aligned}\bar{\boldsymbol{c}}_{\mathrm{IMF}}^{\mathrm{t}}=&[(\bar{\boldsymbol{c}}_{\mathrm{V}}^{\mathrm{t}})_1,\cdots,(\bar{\boldsymbol{c}}_{\mathrm{V}}^{\mathrm{t}})_{j_{\mathrm{V}}},\cdots,(\bar{\boldsymbol{c}}_{\mathrm{V}}^{\mathrm{t}})_{J_{\mathrm{V}}},(\bar{\boldsymbol{c}}_{\mathrm{A}}^{\mathrm{t}})_1,\cdots,\\&(\bar{\boldsymbol{c}}_{\mathrm{A}}^{\mathrm{t}})_{j_{\mathrm{A}}},\cdots,(\bar{\boldsymbol{c}}_{\mathrm{A}}^{\mathrm{t}})_{J_{\mathrm{A}}}]=\\&[\bar{\boldsymbol{c}}_1^{\mathrm{t}},\cdots,\bar{\boldsymbol{c}}_{j_{\mathrm{IMF}}}^{\mathrm{t}},\cdots,\bar{\boldsymbol{c}}_{J_{\mathrm{IMF}}}^{\mathrm{t}}]\end{aligned}\tag{22}$$

其中, $J_{\mathrm{IMF}}=J_{\mathrm{A}}+J_{\mathrm{V}}$, 表示全部 IMF 的数量.

从机械振动和振声 IMF 可提取至少三类三类特征: 基于 Hilbert 变换 (Hilbert transform, HT) 的多尺度边际谱 (Multi-scale HT, MSHT)、HT 变换的多尺度瞬时幅值和频率的均值及方差 (Multi-scale amplitude and frequency, MSAF) 和基于 FFT 的多尺度功率谱密度 (Multi-scale PSD, MSPSD). 这三类特征均有各自特性, 且均已成功应用在不同领域中[71]. 这些谱特征可视为来自不同视角的多源信息, 其转换过程为

$$\bar{\boldsymbol{c}}_{j_{\mathrm{IMF}}}^{\mathrm{t}}\xrightarrow{f^{\mathrm{Trans}(\cdot)}}\boldsymbol{z}_{j_{\mathrm{IMF}}}^{\mathrm{True}}\tag{23}$$

在实际过程中, 可依据工业实际选择其中的一类或几类特征或全部特征. 本文中以 MSHT 特征为例进行描述. 此处, 将基于不同 IMF 的谱特征统一表示为下式:

$$\boldsymbol{z}^{\mathrm{True}}=[\boldsymbol{z}_{1_{\mathrm{IMF}}}^{\mathrm{True}},\cdots,\boldsymbol{z}_{j_{\mathrm{IMF}}}^{\mathrm{True}},\cdots,\boldsymbol{z}_{J_{\mathrm{IMF}}}^{\mathrm{True}}]\tag{24}$$

其中, $\boldsymbol{z}_{j_{\mathrm{IMF}}}^{\mathrm{True}}$ 表示第 j_{IMF} 个 IMF 的 MSHT 特征.

3.2　虚拟样本产生 (VSG) 模块

3.2.1　VSG 模块的结构与功能

该模块由面向 IMF 的 VSG、基于信息熵加权的虚拟样本输出和虚拟样本合成共 3 个子模块组成, 其相互之间的输入输出关系如图 2 所示.

由图 2 可知, "面向 IMF 的 VSG" 子模块包含 J_{IMF} 个面向高维谱数据的相对独立的二级子模块, 其功能是基于真实的多尺度频谱和先验知识得到多个基于不同 IMF 的虚拟样本输入和输出; "基于信息熵加权的虚拟样本输出" 子模块将多个不同 IMF 的虚拟样本输出加权以获得统一虚拟样本输出; "虚拟样本合成" 子模块是组合真实和虚拟训练样本得到混合虚拟样本.

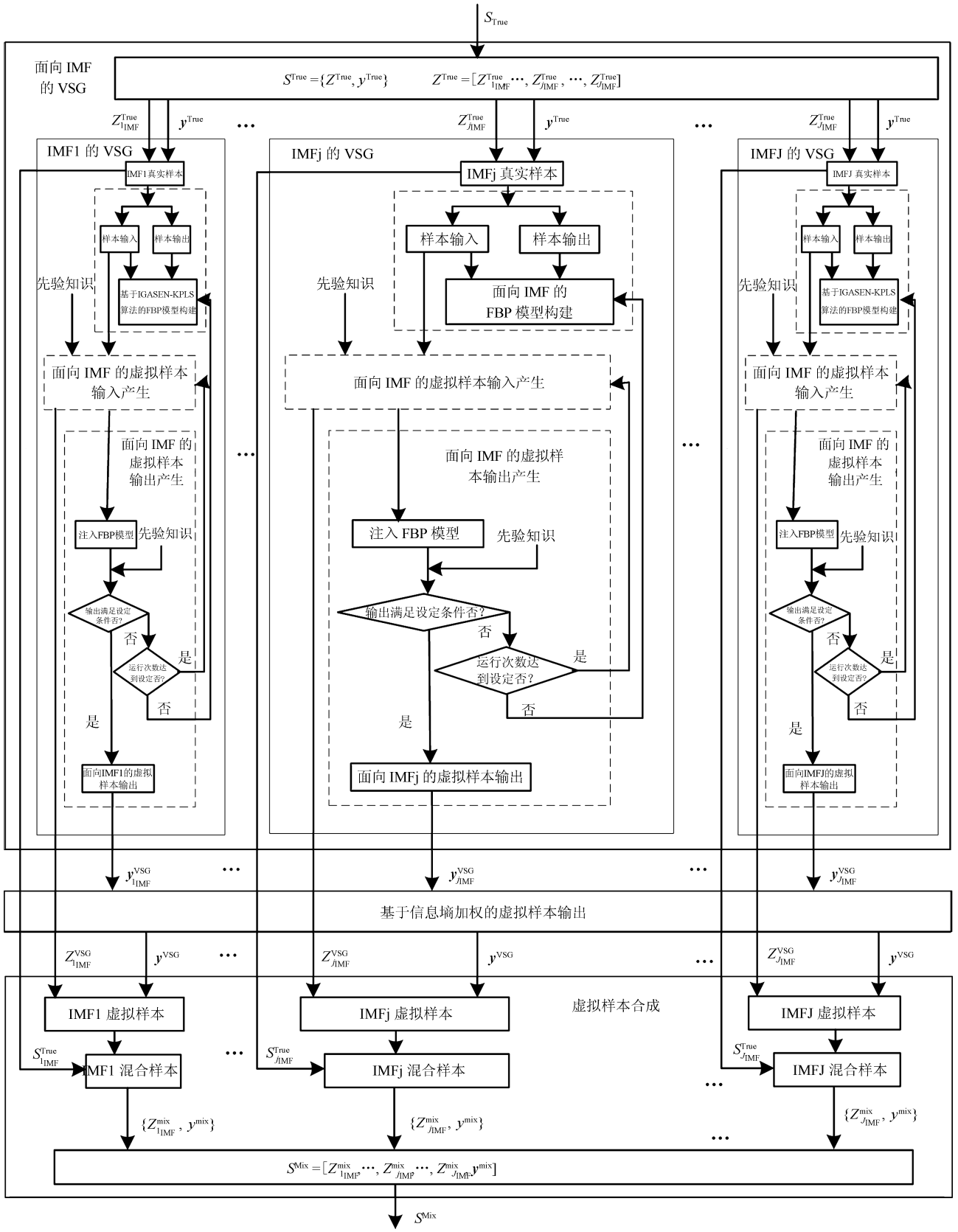


图 2 虚拟样本产生 (VSG) 模块的结构

Fig. 2 Structure of the virtual sample generation (VSG) module

3.2.2 VSG 模块的算法实现

1) 面向 IMF 的 VSG 子模块

将多源多尺度的真实训练样本表示为以不同 IMF 的频谱为输入和待建模过程参数为输出的训练样本子集. 面对实际工业过程, 所具备的先验知识 $Know$ 是: 构建过程参数软测量模型的真实训练样本均具有实际物理含义, 即针对采用实验设计方案产生训练样本的先验知识是已知的.

以第 $S_{j_{\rm IMF}}^{\rm True}=\{(\boldsymbol{z}_{j_{\rm IMF}})_l, y_l\}_{l=1}^{k}$ 个 IMF 训练样本子集 $S_{\rm True}^{j_{\rm IMF}}=\{(\boldsymbol{z}_{j_{\rm IMF}})_l, y_l\}_{l=1}^{k}$ 为例, 对 VSG 过程进行描述. 此处假定两个相邻真实训练样本输入值之间的间隔被等分为 $N_{\rm VSG}$ 个部分, 可知虚拟训练样本输入的数量是 $(N_{\rm VSG}-1)$ 个. 这些虚拟输入可用下式计算:

$$(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}=(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{\rm low}+\frac{\left((\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{\rm high}-(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{\rm low}\right)l'_{\rm VSG}}{N_{\rm VSG}} \tag{25}$$

其中, $1\le l'_{\rm VSG}<N_{\rm VSG}$, $N_{\rm VSG}\ge 2$; $(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}$ 是针对第 $j_{\rm IMF}$th 个 IMF 的第 $l'_{\rm VSG}$th 个虚拟样本输入.

此处, 将基于两个相邻真实训练样本输入之间间产生的虚拟样本输入记为 $\{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{N_{\rm VSG}-1}$. 再假定针对 k 个真实训练样本共存在 $k_{\rm VSG}$ 个间隔利用, 则可产生的虚拟样本输入的总数量为

$$k'_{j_{\rm IMF}}=k_{\rm VSG}(N_{\rm VSG}-1) \tag{26}$$

由上可知, 基于 $\boldsymbol{z}_{j_{\rm IMF}}^{\rm True}$ 产生的虚拟样本输入可表示为 $\{\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG}\}_{l'_{\rm VSG}=1}^{k'_{j_{\rm IMF}}}$. 虚拟样本的第 $l'_{\rm VSG}$th 输入可以采用下式计算虚拟样本输出:

$$(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}=f_{j_{\rm IMF}}^{\rm FBP}\left((\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\right) \tag{27}$$

其中, $f_{j_{\rm IMF}}^{\rm FBP}(\cdot)$ 是基于第 2.1.3 节所描述 SENKPLS 算法利用小样本高维谱数据 $S_{\rm True}^{j_{\rm IMF}}$ 构建的 FBP 模型. 在计算得到 $(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}$ 后, 采用下式判断虚拟输出是否满足如下限定条件:

$$y_{\rm low}^{\rm VSG}\le(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\le y_{\rm high}^{\rm VSG} \tag{28}$$

若不满足式 (28), 则重复式 (27) 和 (28) 表示的过程; 若重复 $N_{\rm times}$ 后, 上式仍然无法满足, 重新构建 FBP 模型 $f_{j_{\rm IMF}}^{\rm FBP}(\cdot)$, 直至满足上述条件; 若满足式 (28), 则获得一个完整的虚拟样本输入输出对, 如下式所示:

$$(S_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}=\{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}},(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\} \tag{29}$$

相应的, 基于 $\boldsymbol{z}_{j_{\rm IMF}}^{\rm True}$ 和 $f_{j_{\rm IMF}}^{\rm FBP}(\cdot)$ 产生的全部虚拟样本输入输出对可表示为 $S_{j_{\rm IMF}}^{\rm VSG}=\{(S_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}$.

上述过程可采用函数 $f_{j_{\rm IMF}}^{\rm VSG}(\cdot)$ 表示为

$$f_{j_{\rm IMF}}^{\rm VSG}(\cdot)=\left\{\begin{array}{l}\{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm True})_l\}_{l=1}^{k}\\ \{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\\ \{(y^{\rm True})_l\}_{l=1}^{k}\end{array}\right\}\xrightarrow{K_{now}}\{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\xrightarrow{f_{j_{\rm IMF}}^{\rm FBP}(\cdot)}\{(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'} \tag{30}$$

因此, 基于真实训练样本子集 $S_{\rm True}^{j_{\rm IMF}}$、先验知识 $Know$ 和 $f_{j_{\rm IMF}}^{\rm FBP}(\cdot)$ 产生虚拟样本的过程可用下式表示:

$$\left.\begin{array}{l}S_{\rm True}^{j_{\rm IMF}}\\ Know\\ f_{j_{\rm IMF}}^{\rm FBP}(\cdot)\end{array}\right\}\xrightarrow{f_{j_{\rm IMF}}^{\rm VSG}(\cdot)}S_{j_{\rm IMF}}^{\rm VSG}=\{\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG},\boldsymbol{y}_{j_{\rm IMF}}^{\rm VSG}\} \tag{31}$$

针对全部 $J_{\rm IMF}$ 个真实训练样本子集, 将所构建的全部 VSG 函数记为 $\{f_{j_{\rm IMF}}^{\rm VSG}(\cdot)\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}$. 全部 IMF 的虚拟样本集合可表示为

$$S_{\rm All}^{\rm VSG}=\{Z_{\rm All}^{\rm VSG},y_{\rm All}^{\rm VSG}\}=\{S_{j_{\rm IMF}}^{\rm VSG}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}=\{\{(\boldsymbol{z}_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}},(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}} \tag{32}$$

2) 基于信息熵加权的虚拟样本输出子模块

由式 (32) 可知, 针对不同的多源多尺度真实训练样本子集产生了 $J_{\rm IMF}$ 个虚拟样本输出值, 原因在于构建了 $J_{\rm IMF}$ 个 FBP 模型. 显然, 需要加权这些不同 IMF 的虚拟样本输出以获得统一的输出值. 第 $l'_{\rm VSG}$th 个虚拟样本的输出为

$$(y^{\rm VSG})_{l'_{\rm VSG}}=f_{\rm IMF}^{\rm Weight}(\cdot)=\sum_{j_{\rm IMF}=1}^{J_{\rm IMF}}w_{j_{\rm IMF}}^{\rm VSG}(y_{j_{\rm IMF}}^{\rm VSG})_{l'_{\rm VSG}} \tag{33}$$

其中, $w_{j_{\rm IMF}}^{\rm VSG}$ 是加权系数. 利用 FBP 模型的预测值和真值, 基于信息熵的加权策略按下式计算:

$$w_{j_{\rm IMF}}^{\rm VSG}=\frac{1}{J_{\rm IMF}-1}\left(1-\frac{1-E_{j_{\rm IMF}}}{\sum\limits_{j_{\rm IMF}=1}^{J_{\rm IMF}}(1-E_{j_{\rm IMF}})}\right) \tag{34}$$

其中

$$E_{j_{\rm IMF}}=\frac{1}{\ln k}\sum_{l=1}^{k}\frac{(e_{j_{\rm IMF}})_l}{\left(\sum\limits_{l=1}^{k}(e_{j_{\rm IMF}})_l\right)}\ln\frac{(e_{j_{\rm IMF}})_l}{\sum\limits_{l=1}^{k}(e_{j_{\rm IMF}})_l} \tag{35}$$

$$(e_{j_{\rm IMF}})_l=\begin{cases}\dfrac{(\hat{y}^{\rm True}_{j_{\rm IMF}})_l-y^{\rm True}_l}{y^{\rm True}_l}, & 0\le\left|\left|\dfrac{(\hat{y}^{\rm True}_{j_{\rm IMF}})_l-y^{\rm True}_l}{y^{\rm True}_l}\right|\right|<1\\ 1 & \left|\left|\dfrac{(\hat{y}^{\rm True}_{j_{\rm IMF}})_l-y^{\rm True}_l}{y^{\rm True}_l}\right|\right|\ge 1\end{cases}\tag{36}$$

其中, $(\hat{y}^{\rm True}_{j_{\rm IMF}})_l$ 表示第个真实训练样本基于 lth 模型的预测值. 经上述过程, 全部的 IMF 虚拟样本输出可表示为 $\{(y^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}$, 其产生过程可用函数 $f^{\rm VSG}(\cdot)$ 表示为

$$f^{\rm VSG}(\cdot)=\begin{cases}\{\{(\boldsymbol{z}^{\rm True}_{j_{\rm IMF}})_l\}_{l=1}^{k}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}} \xrightarrow{k_{\rm now}} \{\{(\boldsymbol{z}^{\rm VSG}_{j_{\rm IMF}})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}\\ \{\{(\boldsymbol{z}^{\rm VSG}_{j_{\rm IMF}})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}\\ \{(y^{\rm True})_l\}_{l=1}^{k}\\ \{\{(y^{\rm VSG}_{j_{\rm IMF}})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}\\ \{(y^{\rm True},\hat{y}^{\rm True}_{j_{\rm IMF}})_l\}_{l=1}^{k}\end{cases}\begin{cases}\xrightarrow{f^{\rm FBP}_{j_{\rm IMF}}(\cdot)}\{\{(y^{\rm VSG}_{j_{\rm IMF}})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}}^{k'}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}}\\ \\ \xrightarrow{f^{\rm Weight}_{\rm IMF}(\cdot)}\{(y^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}\end{cases}\tag{37}$$

3) 虚拟样本合成子模块

经上述过程产生的虚拟样本集可表示为

$$S^{\rm VSG}=\{(\{\boldsymbol{z}^{\rm VSG}_{j_{\rm IMF}}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}})_{l'_{\rm VSG}},(y^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}\tag{38}$$

组合真实和虚拟训练样本可获得构建软测量模型的混合样本 $S^{\rm Mix}$. 为便于理解, 将其重新描述, 其改写过程可表示为

$$\begin{aligned}S^{\rm Mix}=f^{\rm Mix}(\cdot)=\{S^{\rm True};S^{\rm VSG}\}=&\\ &\{\{(\{\boldsymbol{z}^{\rm True}_{j_{\rm IMF}}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}})_l,(y^{\rm True})_l\}_{l=1}^{k};\\ &\{(\{\boldsymbol{z}^{\rm VSG}_{j_{\rm IMF}}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}})_{l'_{\rm VSG}},(y^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}\}=\\ &\{\{\{(\{\boldsymbol{z}^{\rm True}_{j_{\rm IMF}}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}})_l\}_{l=1}^{k};\\ &\{(\{\boldsymbol{z}^{\rm VSG}_{j_{\rm IMF}}\}_{j_{\rm IMF}=1}^{J_{\rm IMF}})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}\},\\ &\{\{(y^{\rm True})_l\}_{l=1}^{k};\{(y^{\rm VSG})_{l'_{\rm VSG}}\}_{l'_{\rm VSG}=1}^{k'}\}\}=\\ &\{\{\boldsymbol{z}^{\rm mix}_{l^{mix}}\}_{l^{mix}=1}^{k+k'},\{y^{\rm mix}_{l^{mix}}\}_{l^{mix}=1}^{k+k'}\}=\\ &\{Z^{\rm mix},\boldsymbol{y}^{\rm mix}\}\end{aligned}\tag{39}$$

由上式可知, 混合样本的数量为 $(k+k')$.

3.2.3 VSG 模块的准确性分析与适应性讨论

VSG 模块是保证本文所提多组分机械信号建模方法的准确性并具有良好预测性能的关键环节. 此处针对产生虚拟样本过程中的关键环节进行准确性分析和适用性讨论.

1) 面向 IMF 构建的 FBP 模型: 此处采用的建模方法是适合于小样本高维数据的 SENKPLS 算法. KPLS 算法是 PLS 的核版本, 能够有效地处理输入变量间的共线性问题, 其用于构建内层模型的潜在变量远远小于原始输入特征数量, 进而能够保证基于 IMF 的高维谱数据构建的 FBP 模型的泛化性能. 另外, 基于 "操纵训练样本" 集成构造策略的 SEN 算法选择有价值训练样本构造软测量模型, 其与 KPLS 算法的结合, 进一步增强了 FBP 模型的泛化性能.

2) 面向 IMF 的虚拟样本输入: 基于多源多尺度 IMF 的真实训练样本子集是基于机械振动/振声信号经自适应分解和时频变换获得; 虽然这些高维谱变量难以得到合理的具体解释, 但原始的机械振动/振声样本均有明确的物理含义; 在产生虚拟样本输入值时, 两个真实训练样本间所划分间隔的大小也是依据先验知识确定的; 并且式 (25) 保证了其合理的取值范围. 这些约束保证了虚拟样本输入值的准确性.

3) 面向 IMF 的虚拟样本输出产生的准确性: 这些不同的 IMF 基于各自的 FBP 模型所产生的虚拟输出值的范围由式 (28) 及其相应的策略予以保证, 由图 2 所给出的 "面向 IMF 的 VSG" 中的算法流程给出了更为清晰的描述, 进而保证了虚拟样本输出值的准确性.

4) 基于信息熵加权的虚拟样本输出: 这些多源多尺度 IMF 均是由原始的机械振动/振声信号分解得到的, 显然这些 IMF 的虚拟样本输入应该对应统一且唯一的虚拟样本输出值, 需要将不同 IMF 的虚拟样本输出值进行加权, 式 (33) ∼ (36) 采用信息熵加权, 由 FBP 模型的预测误差得到不同 IMF 虚拟样本输出值的权系数, 从而保证了最终的虚拟样本输出值的准确性.

5) VSG 模块的适用性讨论: 在 VSG 过程中, FBP 模型的准确性比较重要, 这就要求真实的训练样本虽然稀少, 但尽量保证较宽的工况覆盖范围; VSG 模块的所描述的算法适用于已经将原始多组分信号分解为多个不同的子信号, 并且这些子信号的输入变量间具有较强的共线性的情况; 同时, "面向 IMF 的 VSG 子模块" 中所描述的算法适用于具

有高维共线性特性的真实训练数据产生虚拟样本.

因此, 本文此处所提出 VSG 技术能够保证所产生虚拟样本的准确性. 另外, 本文所提 VSG 技术虽然是面向多组分机械信号建模这类问题提出的, 具有较强的针对性; 同时, "面向 IMF 的 VSG 子模块"所描述的 VSG 技术针对高维谱数据也具有较好的普适性.

3.3　谱特征自适应选择模块

为便于特征选择, 此处将不同 IMF 的输入变量特征进行合并和重新编号, 如下式所示:

$$Z^{\rm mix}=[Z^{\rm mix}_{1_{\rm IMF}},\cdots,Z^{\rm mix}_{j_{\rm IMF}},\cdots,Z^{\rm mix}_{J_{\rm IMF}}]=[(\boldsymbol{z}^{\rm mix}_1),\cdots,(\boldsymbol{z}^{\rm mix})_p,\cdots,(\boldsymbol{z}^{\rm mix}_P)] \tag{40}$$

采用密度估计方法计算第 pth 谱变量 $(\boldsymbol{z}^{\rm mix})_p$ 和难以检测过程参数 $\boldsymbol{y}^{\rm mix}$ 之间的 MI 值:

$$Muin\left(\boldsymbol{y}^{\rm mix};(\boldsymbol{z}^{\rm mix})_p\right)=\int\int p((\boldsymbol{z}^{\rm mix})_p)\times\log\frac{p(\boldsymbol{y}^{\rm mix},(\boldsymbol{z}^{\rm mix})_p)}{p((\boldsymbol{z}^{\rm mix})_p)p(\boldsymbol{y}^{\rm mix})}{\rm d}((\boldsymbol{z}^{\rm mix})_p){\rm d}\boldsymbol{y}^{\rm mix}=H(\boldsymbol{y}^{\rm mix})-H(\boldsymbol{y}^{\rm mix}|(\boldsymbol{z}^{\rm mix})_p) \tag{41}$$

其中, $p((\boldsymbol{z}^{\rm mix})_p)$ 和 $p(\boldsymbol{y}^{\rm mix})$ 是 $(\boldsymbol{z}^{\rm mix})_p$ 和 $\boldsymbol{y}^{\rm mix}$ 的边缘概率密度, $p((\boldsymbol{z}^{\rm mix})_p,\boldsymbol{y}^{\rm mix})$ 是联合概率密度, $H(\boldsymbol{y}^{\rm mix}|(\boldsymbol{z}^{\rm mix})_p)$ 是条件熵, $H((\boldsymbol{z}^{\rm mix})_p)$ 是信息熵.

基于下式对谱特征进行选择:

$$\zeta_{(\boldsymbol{z}^{\rm mix})_p}=\begin{cases}1, & Muin\left(\boldsymbol{y}^{\rm mix};(\boldsymbol{z}^{\rm mix})_p\right)\geq\theta^{\rm Muin}\\ 0, & Muin\left(\boldsymbol{y}^{\rm mix};(\boldsymbol{z}^{\rm mix})_p\right)<\theta^{\rm Muin}\end{cases} \tag{42}$$

其中, $\theta^{\rm Muin}$ 是谱特征的选择阈值, 且 $\theta^{\rm Min}_{\rm Muin}\leq\theta^{\rm Muin}\leq\theta^{\rm Max}_{\rm Muin}$. $\theta^{\rm Min}_{\rm Muin}$ 和 $\theta^{\rm Max}_{\rm Muin}$ 是 $(\boldsymbol{z}^{\rm mix})_p$ 的输入特征与 $\boldsymbol{y}^{\rm mix}$ 之间互信息的最小值和最大值; 采用下式自适应计算 MI 阈值的递增步长:

$$\theta^{\rm step}=\frac{\theta^{\rm Max}_{\rm Muin}-\theta^{\rm Min}_{\rm Muin}}{10} \tag{43}$$

此处简化基于 MI 的特征选择过程是考虑到后续基于 SENKPLS 的软测量模型可以消谱特征间的共线性. 采用如下过程进行有价值 IMF 及其特征的自适应选择: 首先, 将 $\zeta_{(\boldsymbol{z}^{\rm mix})_p}=1$ 的谱特征进行串行组合并构建潜结构模型; 接着, 以步长 $\theta_{\rm step}$ 增加 MI 阈值并重复上述过程; 最终, 具有最小预测误差的阈值被选定为最终阈值, 以此阈值基于式 (42) 完成有价值子信号及其谱特征的选择.

将约简后的多源多尺度谱特征记为 $Z^{\rm mix}_{\rm sel}$, 相应的混合样本可标记为

$$S^{\rm MixSel}=f^{\rm SelFea}(\cdot)=\{Z^{\rm mix}_{\rm sel},\boldsymbol{y}^{\rm mix}\} \tag{44}$$

3.4　软测量模型构建模块

采用第 1.1.3 节的 SENKPLS 算法, 基于约简后的混合样本 $S^{\rm MixSel}$ 构建过程参数软测量模型, 其输出可表示为

$$\hat{\boldsymbol{y}}^{\rm Mix}=f^{\rm SENKPLS}(Z^{\rm mix}_{\rm sel}) \tag{45}$$

4　应用验证

此处的应用验证分为两部分: 首先, 采用近红外谱 (NIR) 数据对本文所提方法的关键部分 "面向 IMF 的 VSG" 技术进行验证; 接着, 采用磨矿过程实验球磨机的筒体振动/振声信号验证本文所提出的基于 VSG 的多组分机械信号建模方法.

4.1　基于近红外 (NIR) 谱的 VSG 技术验证

4.1.1　数据描述及预处理

近红外 (Near infrared, NIR) 谱数据用于估计橙汁的含糖水平. 该数据集可在 http://www.ucl.ac.be 下载, 其包含的训练和测试数据集的大小分别是 150 × 700 和 68 × 700. 本文采用间隔取样原始训练数据方式获取 1/10 (15 个样本) 作为真实的训练样本, 其输入特征变量和待预测的输出如图 3 所示.

采用 PLS 提取真实训练样本的潜在特征 (LV), 其中前 5 个 LV 的贡献率如表 1 所示.

表 1　基于 PLS 提取的潜在特征的贡献率

Table 1　Contribution of the latent features extracted based on PLS

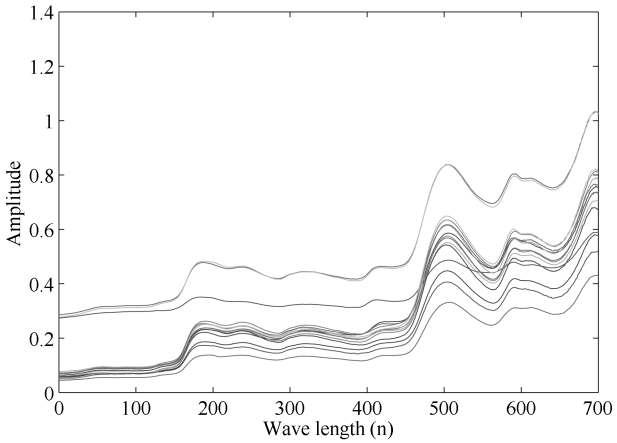
LV#	输入单 LV	输入累计	输出单 LV	输出累计
1	88.85	88.85	23.78	23.78
2	10.96	99.8	19	42.78
3	0.17	99.97	20.8	63.57
4	0.01	99.98	21.77	85.35
5	0.01	99.99	8.63	93.98

由表 1 可知: 在 NIR 谱数据的 700 个输入变量中提取的第 1 个潜在特征可以表征输入数据 88.85 % 的变化, 却只能提取输出数据的 23.78 %的变化; 相应的, 前 5 个 LV 的累计贡献率可以达到 99.99 %和 93.98 %. 这些结果表明了 NIR 数据具有较强的共线性, 也表明了基于潜在变量的建模方法比较适合该类高维数据.

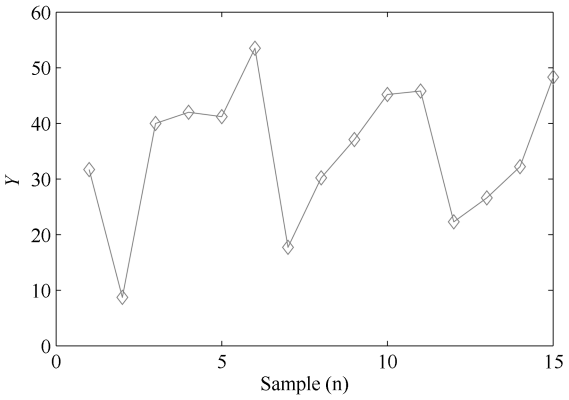
4.1.2　谱数据的 VSG 结果

采用第 1.1.3 节所描述的 SENKPLS 算法构建真实训练样本的 FBP 模型. 其中, 设定候选子模型

的数量 (即 GA 种群数量) 为 20, 集成子模型的选择阈值为 0.05, KPLS 采用径向基函核函数 (Radial basis function, RBF); 核半径和核潜在变量 (KLV) 数量采用网格寻优法, 它们与均均方根误差 (Root mean square error, RMSE) 的关系如图 4 和图 5 示.



(a) 本文采用的 NIR 真实训练样本的输入
(a) Inputs of NIR true training samples using in this study



(b) 本文采用的 NIR 真实训练样本的输出
(b) Outputs of NIR true training samples using in this study

图 3 NIR 真实训练样本输入和输出

Fig. 3 Inputs and outputs of NIR training samples

依据图 4 和图 5 可知, KLV 取 10, 核半径取 600. 为克服 GA 算法的随机性, 采用上述建模参数运行 20 次, 测试数据 RMSE 的均值、最大值、最小值和方差分别为 7.1880, 9.0742, 5.6867 和 0.9779 基于上述 FBP 模型, 采用第 3.2.2 节 "1) 面向 IMF 的 VSG 子模块" 部分描述的 VSG 算法生成虚拟样本. 此处, 选择在两个相邻的真实训练样本之间产生虚拟样本; 这样, 在 $N_{\rm VSG}=2,3,\cdots,10$ 时, 对应的虚拟样本的数量分别是 $14,28,\cdots,126$. $N_{\rm VSG}=8$ 时所生成的虚拟样本输入和输出如图 6 所示.

对比图 6 和图 3 可知, 虚拟样本的输入和输出在趋势上与真实训练样本是一致的.

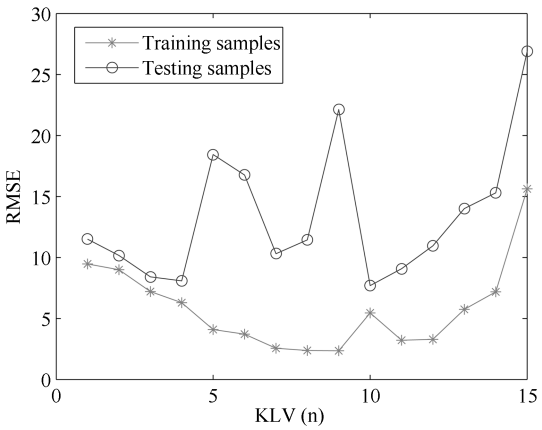


图 4 KLVs 数量与 RMSE 的关系

Fig. 4 Relationships between KLVs′ number and RMSE

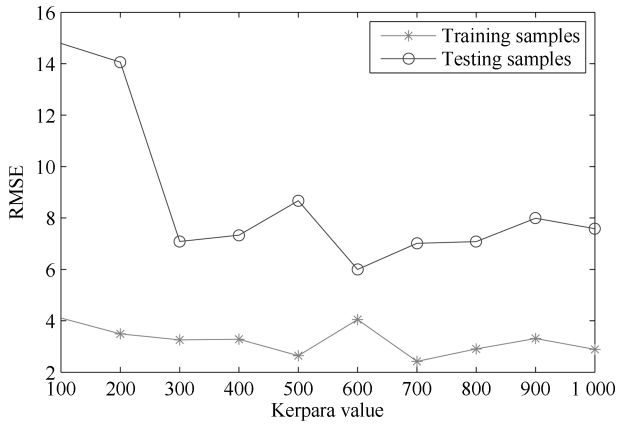


图 5 核半径与 RMSE 的关系

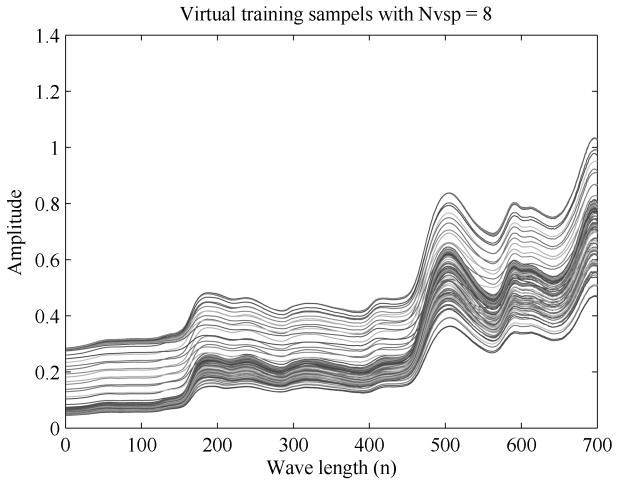
Fig. 5 Relationships between kernel radiu and RMSE

4.1.3 基于 VSG 的模型性能比较结果

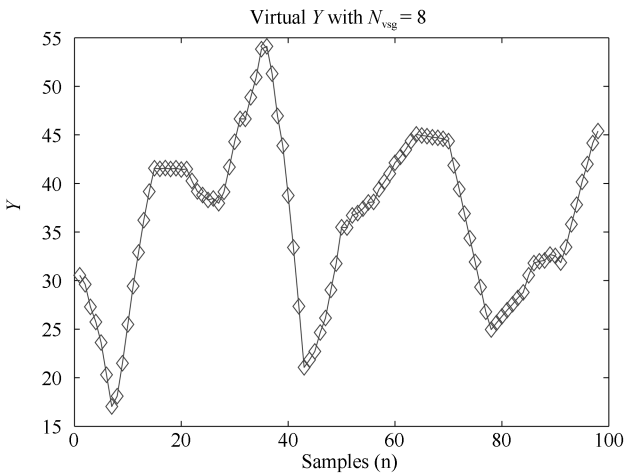
将虚拟与真实样本相混合, 采用第 1.1.3 节的方法构建 $N_{\rm VSG}=2,3,\cdots,10$ 时的混合样本软测量模型. 本文此处主要基于 NIR 谱数据验证 VSG 技术的合理性, 故未对谱特征进行选择. 建模参数的选择过程和所采用的测试样本均与 FBP 模型相同. 采用不同数量的混合样本建立的软测量模型的测试数据统计结果如表 2 所示.

如表 2 所示, 在 $N_{\rm VSG}=9$ (即虚拟训练样本数量为 112) 时, 基于混合样本所构建的软测量模型具有最佳的平均、最大和最小 RMSE, 分别为 6.0723, 6.8627 和 5.8029, 并且方差也只有 FBP 模型的 1/4. 可见, 本文所提 VSG 技术同时提高了软测量模型的预测精度和预测稳定性. 基于其他数量的混合样本所构建的模型的预测性能也多强于 FBP 模型, 模型预测的稳定性均得到提高. 因此, 选择合适的虚拟样本数量对构建有效的软测量模型非常重要.

表 2 只是从测试数据预测性能的视角给出统计结果, 未考虑训练数据. 图 7 给出采用不同值 (即不同数量的虚拟样本) 建模时的训练误差的统计曲线.



(a) 本文产生的 $N_{\mathrm{VSG}}=8$ 的 NIR 虚拟训练样本的输入
(a) Inputs of NIR virtual training samples with $N_{\mathrm{VSG}}=8$ in this study



(b) 本文产生的 $N_{\mathrm{VSG}}=8$ 的 NIR 虚拟训练样本的输出
(b) Inputs of NIR virtual training samples with $N_{\mathrm{VSG}}=8$ in this study

图 6　NIR 虚拟训练样本的输入和输出

Fig. 6　Inputs and outputs of NIR virtual training samples

表 2　基于混合样本建立的 NIR 模型统计结果

Table 2　Statistical results of NIR model based on mixed samples

真实样本数量	虚拟样本数量	核半径值	KLV 数量	均值 (Mean)	最大值 (Max)	最小值 (Min)	方差 (Var)
15	0	600	10	7.188	9.0742	5.6867	0.9779
15	14	65	12	6.9473	9.3558	5.9747	0.8814
15	28	50	15	7.7808	11.2846	6.2231	1.2904
15	42	0.8	13	8.9316	10.3599	7.7212	0.7781
15	56	10	14	7.7027	12.2875	6.0686	1.6912
15	70	60	13	8.3782	11.6499	7.2438	1.0895
15	84	60	10	7.0026	7.6549	6.5843	0.3273
15	98	65	12	7.3832	8.4788	6.8142	0.4051
15	112	75	9	**6.0723**	**6.8627**	**5.8029**	**0.2375**
15	126	19	12	8.1114	10.0179	6.5984	0.8225

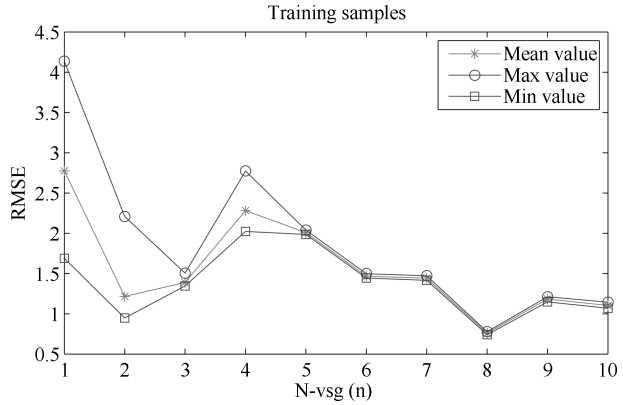


图 7　基于不同数量的虚拟样本构建模型的训练误差

Fig. 7　Training errors of the constructed model based on virtual samples with different numbers

由图 7 可知: FBP 模型 ($N_{\mathrm{VSG}}=1$) 具有最大的预测性能波动范围; $N_{\mathrm{VSG}}=5$ 以后, 混合训练数据构建的软测量模型的预测稳定性较好, 同时训练误差也进一步降低.

综合表 2 和图 7 可知, 本文所提 VSG 技术适合于对小样本高维谱数据, 将其结合具体工业过程进行研究将更具有实际意义.

4.2　基于球磨机机械信号的建模方法验证

4.2.1　磨矿过程及磨机实验描述

1) 磨矿工艺与磨机负荷参数磨矿过程是整个选矿流程中最为重要的作业环节. 国内选矿行业广泛应用两段式闭式磨矿回路 (Grinding circuit, GC) 工艺. 其中一段 GA (GC I) 的工艺流程如图 8 所示.

如图 8 所示, 原矿通过振动给料机给到运输皮带, 然后输送到湿式预选机, 通过磁力选择有用矿石, 抛尾矿; 然后混合来自旋流器的沉砂以及周期性添加的钢球, 通过给矿器进入球磨机; 球磨机依靠筒体旋转带动钢球对矿石进行冲击破碎, 形成矿浆; 矿

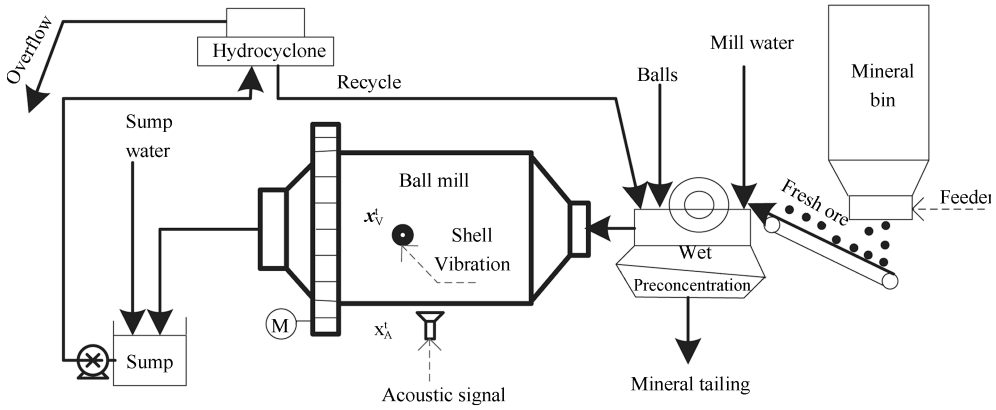


图 8 某选矿厂一段磨矿回路 (GC I) 工艺流程

Fig. 8 Flow chart of the grinding circuit I (GC I) of some mineral grinding process

浆依靠自身的流动性排出磨机后进入泵池, 与泵池内的新加水混合后被泵入旋流器; 旋流器将矿浆分为粒度较细的溢流和较粗的沉砂, 其中, 后者进入球磨机再磨构成球磨的闭路循环, 前者进入磁选机进行选别的沉砂将进入二段磨矿回路 (GC II).

磨矿过程的目标是以最大化的经济效率解离和获得有价值矿物. 通常情况下, 磨矿过程的目标是通过设定固定的产品粒度获得最大化的磨矿生产率 (Grinding production rate, GPR); 最大化 GPR 通常是通过最优化磨矿回路负荷决定的, 而后者通常由 GC I 的球磨机负荷确定. 球磨机是磨矿过程使用的关键重型旋转机械设备, 其内部料球比 (Material to ball volume ratio, MBVR)、磨矿浓度 (Pulp density, PD)、充填率 (Charge volume ratio, CVR) 等磨机负荷参数的异常变化 (如 MBVR 过大、PD 和 CVR 过高等) 均会导致磨机过负荷. 这些磨机负荷参数与磨机研磨产生的机械振动和振声信号的映射机理复杂. 在时域内, 有价值信息被隐含在随机噪声中, 难以提取; 但其频域特征较为明显.

2) 实验球磨机筒体振动/振声数据采集与描述

实验在 XMQL420 × 450 球磨机上进行, 其滚筒外径和长度均为 460 mm. 该磨机由功率为 2.12 kW 的三相电机驱动, 最大钢球装载量为 80 kg, 设计磨粉能力为 10 kg/小时, 转速为 57 转/分钟. 磨机中部开口, 用于添加钢球、物料和水负荷. 实验中采用的物料为铜矿石, 直径均小于 6 mm, 密度为 4.2 t/m^3. 采用直径 30、20 和 15 mm 的钢球作为研磨介质, 配比为 3 : 4 : 3. 磨机筒体振动和振声信号的采样频率为 51 200 Hz 和 8 000 Hz.

考虑到实际磨矿过程的建模和磨机负荷的控制等研究中常把 24 或 48 小时内的钢球负荷当做常量处理[6], 本文进行了球负荷保持不变的 4 组磨机实验: 第一组: 料负荷为 10 kg, 水负荷从 5 kg 增加到 40 kg; 第二组: 料负荷为 20 kg, 水负荷从 2 kg 增加到 20 kg; 第三组: 水负荷为 2 kg, 料负荷从 10 kg 增加到 20 kg; 第四组: 水负荷为 10 kg, 料负荷从 22 kg 增加到 50 kg

4.2.2 多尺度 IMF 获取结果

文献 [23] 给出了零负荷时的筒体振动原始信号和经 EMD 分解为具有不同时间尺度的 IMF 子信号的时频曲线, 结果表明: 不同尺度 IMF 频谱代表单尺度频谱的不同部分. 特别指出的是, 第 13 个 IMF 子信号为 4 周期正弦曲线, 与所处理数据包含的磨机 4 个旋转周期相一致; 通过对频谱的幅值进行比较可知, 第 13 个 IMF 的幅值至少是其他的 100 倍, 表明磨机自身旋转引起的振动很大; 第 1 个 IMF 到第 15 个 IMF 的时频特性表明, 子信号的时间尺度是逐渐增大. 因此, 这些多尺度频谱蕴含着极为不同的磨机负荷参数信息.

本文中, 选择 $A_{\text{noise}} = 0.1$ 和 $M = 10$ 将磨机旋转 4 个周期的信号进行 EEMD 分解, 并将筒体振动和振声信号的 IMF 标记为振动 IMF (Vibration VIMF) 和振声 IMF (Acoustic IMF). 对全部 IMF 后进行 HT 变换, 计算得到 IMF 边际谱. 部分筒体振动和振声多尺度子信号的谱数据 (VIMF1 ∼ VIMF10) 如图 9 和图 10 所示.

由图 9 和图 10 可知, 这些磨机筒体振动和振声信号的谱数据的时间尺度是逐渐增大的, 表明了这些机械信号组成成分的复杂性和可分解特性.

4.2.3 多尺度 IMF 的 VSG 结果

采用不同的多源多尺度谱数据构造 FBP 模型的过程与第 4.1.2 节相同, 此处不在赘述.

依据第 4.2.1 节所描述的实验过程, 可知 4 种工况下真实训练样本的分布如表 3 所示.

以表 4 中真实训练样本 1 和 2 为例进行说明: 两个样本中的料负荷固定为 10 kg, 而水负荷从 5 kg 增大到 15 kg, 因此可以构造出水负荷在

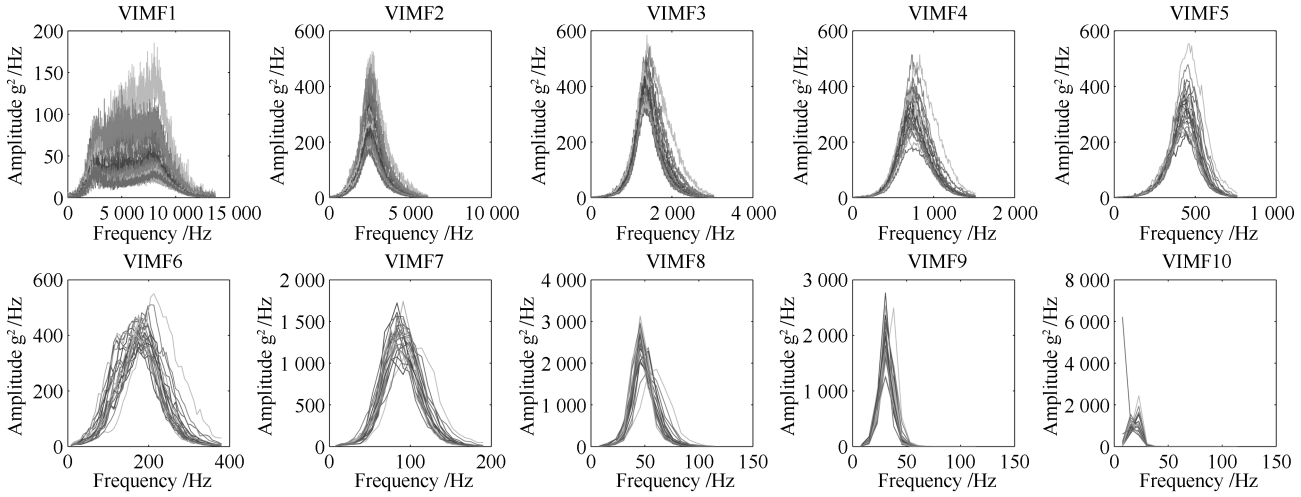


图 9　磨机筒体振动的 VIMF1～10 子信号的真实谱数据

Fig. 9　True spectra data of VIMF1～10 sub-signals from mill shell vibration signal

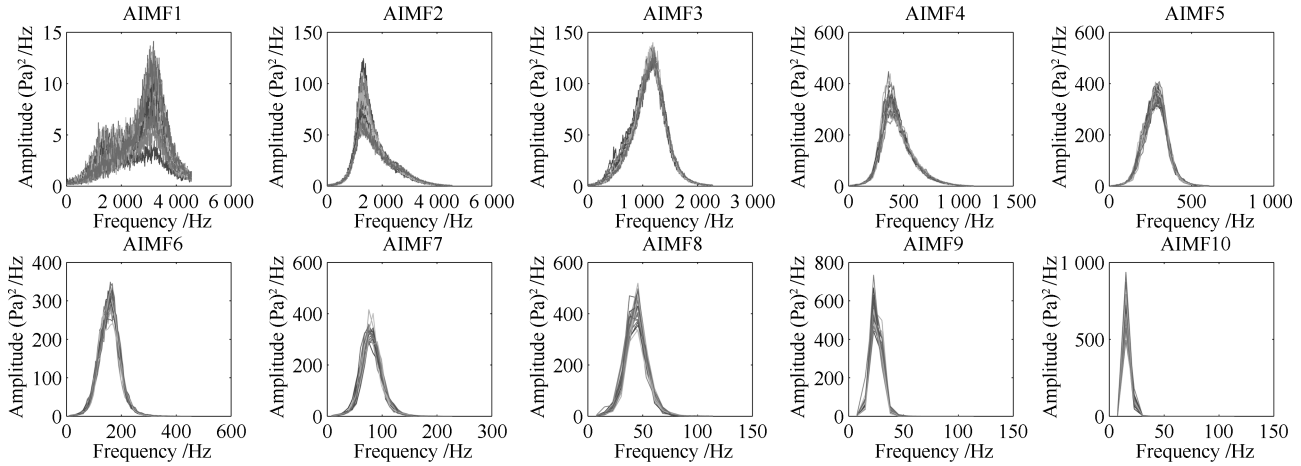


图 10　磨机振声 AIMF1～10 子信号的真实谱数据

Fig. 10　True spectra data of VIMF1～10 sub-signals from mill acoustic signal

表 3　用于产生虚拟样本的真实训练样本分布

Table 3　Distribution of the true training samples for generating virtual samples

样本序号	1	2	3	4	5	6	7	8	9	10	11	12	13
固定负荷 (kg)	料 10	料 10	料 10	水 2	水 2	水 2	料 20	料 20	料 20	水 10	水 10	水 10	水 10
变化负荷 (kg)	水 5	水 15	水 20	料 10	料 1 6	料 20	水 7.5	水 12.5	水 20	料 24	料 28	料 35	料 45

5～15 kg 间变化的虚拟样本输入. 以此类推, 依据表 1 可知能够产生虚拟样本输入的真实训练样本对包括: No. 1 和 No. 2, No. 2 和 No. 3, No. 4 和 No. 5, No. 5 和 No. 6, No. 7 和 No. 8, No. 8 和 No. 9, No. 10 和 No. 11, No. 11 和 No. 12, No. 12 和 No. 13. 当 $N_{\mathrm{VSG}} = 2, 3, \cdots, 10$ 时, 对应的虚拟样本的数量分别是 $9, 18, \cdots, 81$.

依据第 3.2 节所述方法, 当 $N_{\mathrm{VSG}} = 7$ 时针对 CVR 所产生的部分虚拟样本 (VIMF1～10, AIMF1～10) 的输入如图 11 和图 12 所示.

由图 11 和 12 可知, 所提方法可以有效地产生虚拟样本. 同时, 图 9～12 也表明了不同尺度的谱数据需要进行维数约简, 并选择更有价值的 IMF 构建软测量模型.

4.2.4　多尺度 IMF 的混合样本特征选择结果

以 CVR 为例, 采用第 3.3 节所述方法对混合样本进行谱特征选择, 统计结果如表 4 所示.

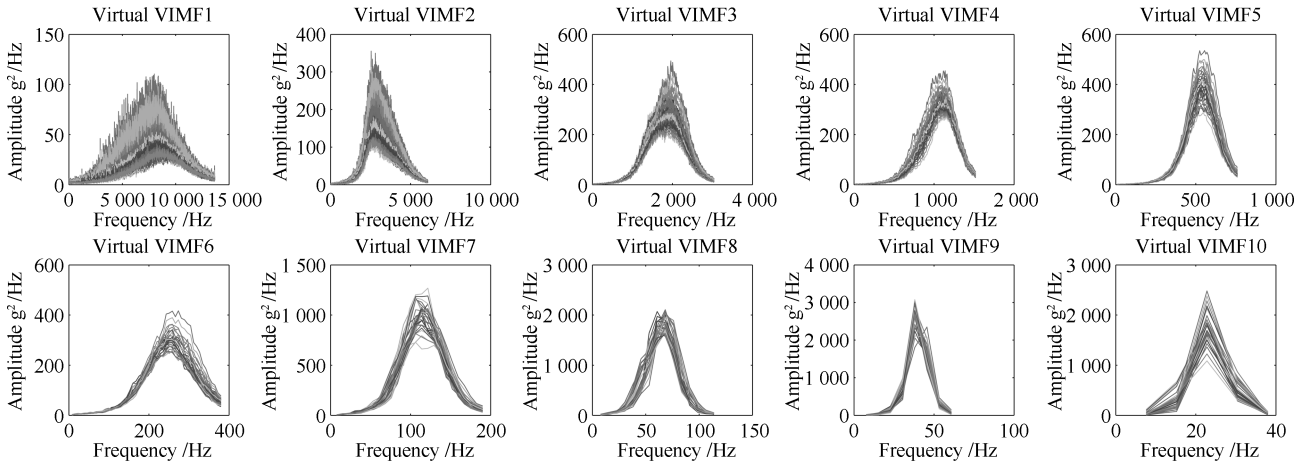


图 11 磨机筒体振动 VIMF1～10 的虚拟谱数据

Fig. 11 Virtual spectra data of VIMF1～10 sub-signals from mill shell vibration signal

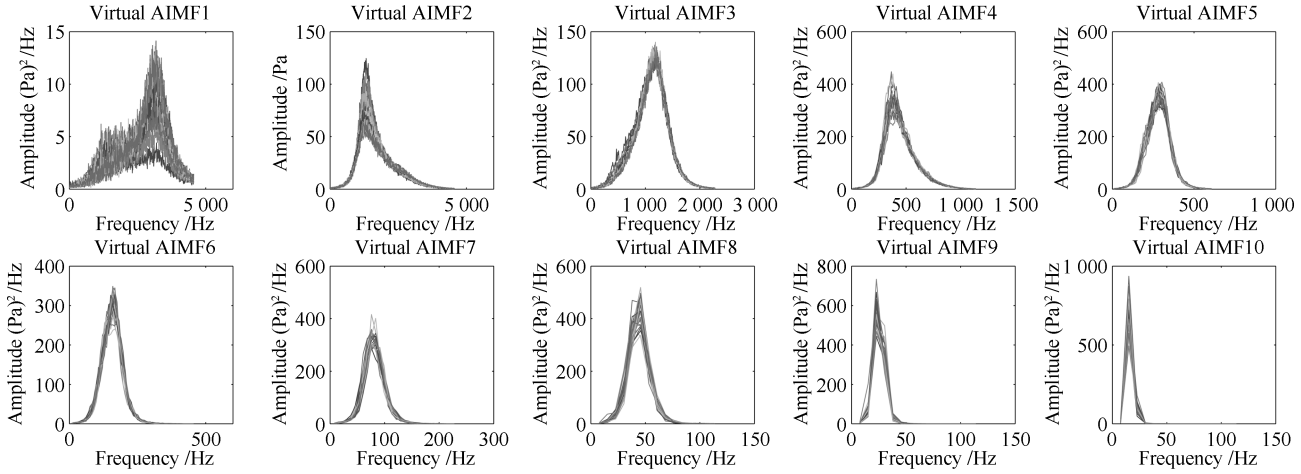


图 12 磨机振声 AIMF1～10 的虚拟谱数据

Fig. 12 Virtual spectra data of AIMF1～10 sub-signals from mill acoustic signal

表 4 面向 CVR 的谱特征选择的统计结果

Table 4 Statistical results of spectra feature selection for CVR

真实样本数量	虚拟样本数量	振动特征数量	振声特征数量	特征数量总和	MI 阈值
13	9	1 344	190	1 534	0.8
13	18	473	71	514	0.9
13	27	3 388	1 878	5 266	0.2
13	36	895	175	1 070	0.8
13	45	3 396	1 870	5 266	0.2
13	54	3 387	1 869	5 256	0.2
13	63	3 403	1 880	5 283	0.1
13	72	3 384	1 865	5 249	0.2
13	81	3 403	1 879	5 282	0.1

表 4 表明: 阈值与谱特征数量间的相关性较大; 从 AIMF 选择的特征数量远小于 VIMF, 可能的原因除筒体振动信号的采样频率远高于声音信号外, 其机理方面的原因还有待于深入的进一步分析.

4.2.5 软测量模型的性能比较结果

软测量模型的性能通常采用测试样本的 RMSE 进行评估. 当不具备足够的大量测试样本时, 训练数据也用于评估软测量模型性能. 留一交叉验证、K-折交叉验证、Bootstrap 及其改进等性能评估方法

得到了广泛应用[72−73]. 针对高维小样本数据, 0.632 Bootstrap 和留一交叉验证评估方法可以得到较佳性能[74].

本文采用 0.632 Bootstrap 评估方法. 假设进行了 R 次的 Bootstrap, 采用 k_r^* 表示从训练样本抽取的样本, $f_r^*(\cdot)$ 示 k_r^* 训练的软测量模型; 定义 0.632 Bootstrap 的均方根相对预测误差 (Root mean square relative error prediction, RMSREP) 如下:

$$RMSREP_{0.632} = 0.632RMSREP_{\rm BCV} + (1-0.632)RMSREP_{\rm app} \tag{46}$$

$$RMSREP_{\rm BCV} = \sqrt{\frac{1}{k^{\rm mix}}\sum_{l^{\rm mix}=1}^{k^{\rm mix}}\frac{1}{R_{-l^{\rm mix}}}\sum_{r:l^{\rm mix}\notin k_r^*}\left(\frac{f_r^*(\boldsymbol{z}_{l^{\rm mix}}^{\rm mix})-y_{l^{\rm mix}}^{\rm mix}}{y_{l^{\rm mix}}^{\rm mix}}\right)^2} \tag{47}$$

$$RMSREP_{\rm app} = \sqrt{\frac{1}{k^{\rm mix}}\sum_{l^{\rm mix}=1}^{k^{\rm mix}}\left(\frac{f_{k^{\rm mix}}(\boldsymbol{z}_{l^{\rm mix}}^{\rm mix})-y_{l^{\rm mix}}^{\rm mix}}{y_{l^{\rm mix}}^{\rm mix}}\right)^2} \tag{48}$$

其中, $r=1,\cdots,R$; R_{-l} 是不包含第 lth 个训练样本所抽取的样本数量; $f_{k^{\rm mix}}(\cdot)$ 表示由全部混合样本训练得到的软测量模型.

采用上述评估指标, 针对文献中采用的不同方法及本文所提方法基于不同数量混合样本建立的软测量模型的统计结果如表 5 所示. 其中: 采用对原始机械振动/振声信号进行 FFT 变换后的单尺度频谱构建基于 PLS/KPLS 单模型, 并作为比较的基本方法; 文献 [24] 采用的方法表示基于单尺度频谱获得特征子集进行选择性信息融合的 SEN 模型; 文献 [20−21] 是基于 EMD 的多尺度频谱进行选择性信息融合的 SEN 模型.

表 5 表明:

1) 本文所提方法的平均预测性能随虚拟样本数量的增加而增加. 本文方法在虚拟样本的数量为 81 时, 其平均预测误差为 0.1290, 与文献 [24] 的最佳的平均预测误差 0.1265 接近. 另外, 文献 [24] 并未对多组分信号进行自适应分解, 在提高软测量模型的可解释性和深入理解磨机研磨机理等方面弱于本文所提方法. 本文方法在平均预测性能上也强于基于单尺度单模型的 PLS/KPLS 方法和基于多尺度频谱选择性信息融合的 SENKPLS 方法.

2) 在所有的软测量模型中, 本文所提方法具有最佳的预测稳定性, 这在工业实际中具有较高的应用价值. 文献 [24] 虽然具有预测误差性能的最小值, 但该方法同时也具有预测误差的最大方差 (0.0677), 是本文所提建模方法的至少 2 倍. 显然, 文献 [24] 的预测性能的稳定性较差. 本文所提方法的测试误差的方差与的关系如图 13 所示.

表 5　基于不同数量混合样本构建的软测量模型的统计结果

Table 5　Statistical results of soft sensor models based on mix samples with different number

	真实样本数量	虚拟样本数量	RMSREP 均值 (Mean)	RMSREP 最小值 (Min)	RMSREP 最大值 (Max)	RMSREP 方差 (Var)
PLS	26	0	0.3445	0.1492	0.5803	0.0977
KPLS	26	0	0.1839	0.0704	0.4598	0.0947
文献 [24]	26	0	0.1265	0.0381	0.4263	0.0677
文献 [20]	26	0	0.3424	0.1967	0.4773	0.0778
文献 [21]	26	0	0.2184	0.0968	0.4418	0.0858
本文	26	0	0.1708	0.0771	0.2829	0.0694
方法	26	9	0.1651	0.118	0.2591	0.0382
	26	18	0.149	0.0966	0.2135	0.0288
	26	27	0.1449	0.0916	0.2159	0.0337
	26	36	0.1345	0.0994	0.1775	0.0172
	26	45	0.1397	0.0909	0.2011	0.0286
	26	54	0.1439	0.0981	0.1914	0.0266
	26	63	0.1321	0.0987	0.1849	0.0208
	26	72	0.1366	0.1069	0.1828	0.0203
	26	81	**0.129**	**0.0988**	**0.1749**	**0.0199**

注: 表 5 中的 26 个真实样本中, 仅是表 3 中所示的 13 个用于产生虚拟样本.

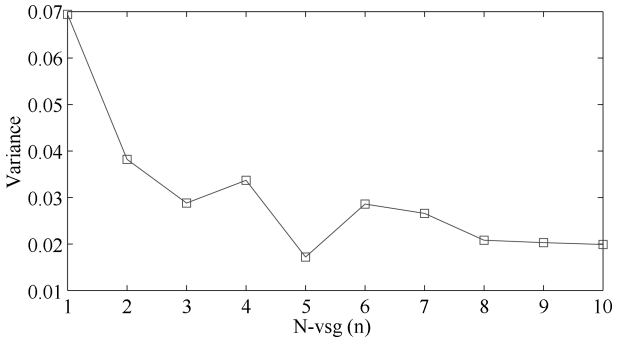


图 13 基于不同 N_{VSG} 值构建的软测量模型测试误差的方差

Fig. 13 Variance of the testing errors based on soft sensor models using different N_{VSG} values

由图 13 可知, 本文所提方法的预测稳定性随虚拟样本数量的增加而提高.

3) 本文所提方法的软测量模型预测误差的均值和最大值随着虚拟样本数量的增加而降低, 如当采用 81 个虚拟样本时, 预测误差的均值和最大值与无虚拟样本时进行比较, 分别从 0.1708 和 0.2829 降低到了 0.1290 和 0.1749. 本文所提方法的测试误差与的关系如图 14 所示.

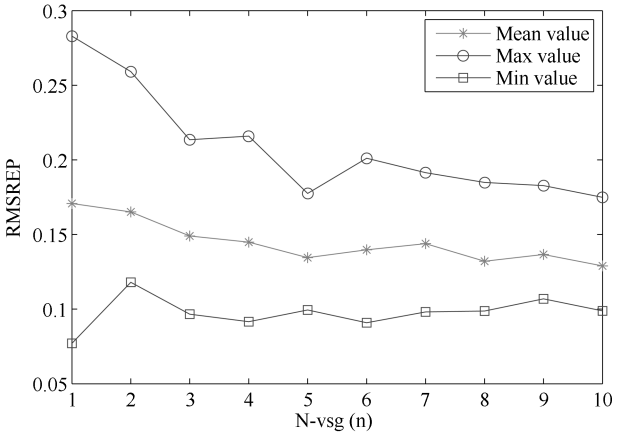


图 14 基于不同 N_{VSG} 值构建的软测量模型的测试误差

Fig. 14 Testing errors based on soft sensor models using different N_{VSG} values

显然, 可依据工业实际确定较适合的虚拟样本的数量. 如何从理论上指导选择合适的虚拟样本数量的研究还有待于深入进行.

综合以上分析结果可知, 本文所提基于 VSG 的机械多组分信号建模方法可以有效提高磨机负荷参数软测量模型的可解释性和预测性能的稳定性. 关于多源多尺度谱特征的具体物理含义, 需要在下一步研究中结合数值仿真模型和更多实验以及运行专家知识逐步获得合理的阐释.

5 结论

本文针对流程工业中与产品质量、产量以及效率等经济指标密切相关的关键机械设备内部参数难以准确直接检测, 只能依据这些设备产生的具有多组分、非平稳、非线性等特性的振动/振声信号进行间接测量, 但建模样本却较为稀缺难以充足完备获取等问题, 提出一种基于虚拟样本生成技术 (VSG) 的多组分机械信号建模方法. 采用近红外谱数据和磨矿过程实验球磨机筒体振动和振声信号构建的软测量模型验证了所提 VSG 技术和面向多组分机械信号建模方法的合理性和有效性.

该文的主要贡献是: 首次提出基于 VSG 结合多组分信号自适应分解的多组分机械信号软测量建模策略, 使得所提方法能够贴合工业实际, 其中所包含的面向小样本高维谱数据的 VSG 技术具有较好的普适性; 综合利用了多组分信号自适应分解、基于互信息的特征选择、基于遗传算法的核潜在模型选择性集成建模等技术, 可以有效地结合特征选择和样本选择策略构建软测量模型, 能够有效模拟工业现场运行专家的认知机制; 对构建物理阐释明确的软测量模型和有效结合复杂工业过程机械设备的虚拟人工系统进行难以检测过程参数软测量具有重要的借鉴意义.

References

1 Chai Tian-You. Industrial process control systems: research status and development direction. *Scientia Sinica Informationis*, 2016, **46**(8): 1003−1015
(柴天佑. 工业过程控制系统研究现状与发展方向. 中国科学: 信息科学, 2016, **46**(8): 1003−1015)

2 Sun Bei, Zhang Bin, Yang Chun-Hua, Gui Wei-Hua. Discussion on modeling and optimal control of nonferrous metallurgical purification process. *Acta Automatica Sinica*, 2017, **43**(6): 880−892
(孙备, 张斌, 阳春华, 桂卫华. 有色冶金净化过程建模与优化控制问题探讨. 自动化学报, 2017, **43**(6): 880−892)

3 Song He-Da, Zhou Ping, Wang Hong, Chai Tian-You. Nonlinear subspace modeling of multivariate molten iron quality in blast furnace ironmaking and its application. *Acta Automatica Sinica*, 2016, **42**(11): 1664−1679
(宋贺达, 周平, 王宏, 柴天佑. 高炉炼铁过程多元铁水质量非线性子空间建模及应用. 自动化学报, 2016, **42**(11): 1664−1679)

4 Zhou P, Chai T Y, Wang H. Intelligent optimal-setting control for grinding circuits of mineral processing. *IEEE Transactions on Automation Science and Engineering*, 2009, **6**(4): 730−743

5 Chai Tian-You. Operational optimization and feedback control for complex industrial processes. *Acta Automatica Sinica*, 2013, **39**(11): 1744−1757
(柴天佑. 复杂工业过程运行优化与反馈控制. 自动化学报, 2013, **39**(11): 1744−1757)

6 Tang Jian, Tian Fu-Qing, Jia Mei-Ying, Li Dong. *Load Soft Sensor of Rotating Mechanical Device based on Frequency Spectral Data-Driven*. Beijing: National Defense Industrial

Press, 2015. 1−63
(汤健, 田福庆, 贾美英, 李东. 基于频谱数据驱动的旋转机械设备负荷软测量. 北京: 国防工业出版社, 2015. 1−63)

7 Tang J, Chai T Y, Liu Z, Yu W. Selective ensemble modeling based on nonlinear frequency spectral feature extraction for predicting load parameter in ball mills. *Chinese Journal of Chemical Engineering*, 2015, **23**(12): 2020−2028

8 Tang J, Qiao J F, Wu Z W, Chai T Y, Zhang J, Yu W. Vibration and acoustic frequency spectra for industrial process modeling using selective fusion multi-condition samples and multi-source features. *Mechanical Systems and Signal Processing*, 2018, **99**: 142−168

9 Zeng Y, Forssberg E. Monitoring grinding parameters by vibration signal measurement-a primary application. *Minerals Engineering*, 1994, 7(4): 495−501

10 Tang J, Zhao L J, Zhou J W, Yue H, Chai T Y. Experimental analysis of wet mill load based on vibration signals of laboratory-scale ball mill shell. *Minerals Engineering*, 2010, **23** (9): 720−730

11 Lei Y G, He Z J, Zi Y Y. Application of the EEMD method to rotor fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 2009, **23**(4): 1327−1338

12 Singh G K, AlKazzaz S A S. Isolation and identification of dry bearing faults in induction machine using wavelet transform. *Tribology International*, 2009, **42**(6): 849−861

13 Cusido J, Romeral L, Ortega J A, Rosero J A, Garcia Espinosa A G. Fault detection in induction machines using power spectral density in wavelet decomposition. *IEEE Transactions on Industrial Electronics*, 2008, **55**(2): 633−643

14 Riera-Guasp M, Antonino-Daviu J A, Pineda-Sanchez M, Puche-Panadero R, Perez-Cruz J. A general approach for the transient detection of slip-dependent fault components based on the discrete wavelet transform. *IEEE Transactions on Industrial Electronics*, 2008, **55**(12): 4167−4180

15 Kankar P K, Sharma S C, Harsha S P. Rolling element bearing fault diagnosis using autocorrelation and continuous wavelet transform. *Journal of Vibration and Control*, 2011, **17**(14): 2081−2094

16 Huang N E, Shen Z, Long S R, Wu M C, Shih H H, Zheng Q, Yen N C, Tung C C, Liu H H. The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1998, **454**(1971): 903−995

17 Faiz J, Ghorbanian V, Ebrahimi B M. EMD-Based analysis of industrial induction motors with broken rotor bars for identification of operating point at different supply modes. *IEEE Transactions on Industrial Informatics*, 2014, **10**(2): 957−966

18 Shukla S, Mishra S, Singh B. Power quality event classification under noisy conditions using EMD-Based De-Noising techniques. *IEEE Transactions on Industrial Informatics*, 2014, **10**(2): 1044−1054

19 Li R Y, He D. Rotational machine health monitoring and fault detection using EMD-based acoustic emission feature quantification. *IEEE Transactions on Instrumentation and Measurement*, 2012, **61**(4): 990−1001

20 Zhao L, Tang J, Zheng W. Ensemble modeling of mill load based on empirical mode decomposition and partial least squares. *Journal of Theoretical and Applied Information Technology*, 2012, **45**(1): 179−191

21 Tang Jian, Chai Tian-You, Cong Qiu-Mei, Yuan Ming-Zhe, Zhao Li-Jie, Liu Zhuo, Yu Wen. Soft sensor approach for modeling mill load parameters based on EMD and selective ensemble learning algorithm. *Acta Automatica Sinica*, 2014, **40**(9): 1853−1866
(汤健, 柴天佑, 丛秋梅, 苑明哲, 赵立杰, 刘卓, 余文. 基于 EMD 和选择性集成学习算法的磨机负荷参数软测量. 自动化学报, 2014, **40**(9): 1853−1866)

22 Wu Z H, Huang N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 2009, **1**(1): 1−41

23 Tang Jian, Chai Tian-You, Cong Qiu-Mei, Liu Zhuo, Yu Wen. Modeling mill load parameters based on selective fusion of multi-scale shell vibration frequency spectrum. *Control Theory Applications*, 2015, **32**(12): 1582−1591
(汤健, 柴天佑, 丛秋梅, 刘卓, 余文. 选择性融合多尺度筒体振动频谱的磨机负荷参数建模. 控制理论与应用, 2015, **32**(12): 1582−1591)

24 Tang J, Chai T, Yu W, Zhao L J. Modeling load parameters of ball mill in grinding process based on selective ensemble multisensor information. *IEEE Transactions on Automation Science and Engineering*, 2013, **10**(3): 726−740

25 Tang Jian, Chai Tian-You, Zhao Li-Jie, Yue Heng, Zheng Xiu-Ping. Soft sensing mill load in grinding process by time/frequency information fusion. *Control Theory and Applications*, 2012, **29**(5): 564−570
(汤健, 柴天佑, 赵立杰, 岳恒, 郑秀萍. 融合时频信息的磨矿过程磨机负荷软测量. 控制理论与应用, 2012, **29**(5): 564−570)

26 Tang Jian, Chai Tian-You, Liu Zhuo. A Soft Measuring Method for Mill Load Parameter, China, Patent 201510303525, June 2015
(汤健, 柴天佑, 刘卓. 一种磨机负荷参数软测量方法, 国家发明专利, 201510303525, 2015 年 6 月)

27 Liu H W, Sun J G, Liu L, Zhang H J. Feature selection with dynamic mutual information. *Pattern Recognition*, 2009, **42**(7): 1330−1339

28 Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 2002, **137**(1−2): 239−263

29 Tang J, Zhang J, Wu Z W, Liu Z, Chai T Y, Yu W. Modeling collinear data using double-layer GA-based selective ensemble kernel partial least squares algorithm. *Neurocomputing*, 2017, **219**: 248−262

30 Zhang X M, Kano M, Li Y. Locally weighted kernel partial least squares regression based on sparse nonlinear features for virtual sensing of nonlinear time-varying processes. *Computers and Chemical Engineering*, 2017, **104**: 164−171

31 Poggio T, Vetter T. Recognition and Structure from One 2D Model View: Observations on Prototypes, Object Classes and Symmetries, Technical Report A. I. Memo 1347, Massachusetts Institute of Technology Cambridge, MA, USA, 1992.

32 Li L J, Peng Y L, Qiu G Y, Sun Z G, Liu S G. A survey of virtual sample generation technology for face recognition. *Artificial Intelligence Review*, 2017, **1**: 1−20

33 Du Y, Wang Y. Generating virtual training samples for sparse representation of face images and face recognition. *Journal of Modern Optics*, 2016, **63**(6): 536−544

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

35 Abu-Mostafa Y S. Hints. *Neural Computation*, 1995, **7**(4): 639−671

36 An G Z. The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation*, 1996, **8**(3): 643−674

37 Li D C, Lin Y S. Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, 2006, **175**(1): 413−434

38 Li D C, Fang Y H, Lai Y Y, Hu S C. Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation. *Information Sciences*, 2009, **179**(16): 2740−2753

39 Chang C J, Li D C, Chen C C, Chen C S. A forecasting model for small non-equigap data sets considering data weights and occurrence possibilities. *Computers and Industrial Engineering*, 2014, **67**(1): 139−145

40 Cho S, Jang M, Chang S. Virtual sample generation using a population of networks. *Neural Processing Letters*, 1997, **5**(2): 21−27

41 Huang C F, Moraga C. A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, 2004, **35**(2): 137−161

42 Li D C, Wen I H. A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*, 2014, **143**: 222−230

43 Chen Z S, Zhu B, He Y L, Yu L A. A PSO based virtual sample generation method for small sample sets: applications to regression datasets. *Engineering Applications of Artificial Intelligence*, 2017, **59**: 236−243

44 Gong H F, Chen Z S, Zhu Q X, He Y L. A monte carlo and PSO based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: an empirical study of petrochemical industries. *Applied Energy*, 2017, **197**: 405−415

45 Coqueret G. Approximate norta simulations for virtual sample generation. *Expert Systems with Applications*, 2017, **73**: 69−81

46 Tang Jian, Sun Chun-Lai, Mao Ke-Feng. A Virtual Sample Generation Method China Patent 201510303525, August 2015
(汤健, 孙春来, 毛克峰. 一种虚拟样本生成方法, 国家发明专利, 201510496474, 2015 年 8 月)

47 Wang F Y. A big-data perspective on AI: Newton, Merton, and analytics intelligence. *IEEE Intelligent Systems*, 2012, **27**(5): 24−34

48 Li Li, Lin Yi-Lun, Cao Dong-Pu, Zheng Nan-Ning, Wang Fei-Yue. Parallel learning — a new framework for machine learning. *Acta Automatica Sinica*, 2017, **43**(1): 1−8
(李力, 林懿伦, 曹东璞, 郑南宁, 王飞跃. 平行学习 — 机器学习的一个新型理论框架. 自动化学报, 2017, **43**(1): 1−8)

49 Tang J, Chai T Y, Zhao L J, Yue H. Soft sensor for parameters of mill load based on multi-spectral segments PLS sub-models and on-line adaptive weighted fusion algorithm. *Neurocomputing*, 2012, **78**(1): 38−47

50 Rosipal R, Trejo L J. Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, 2002, **2**: 97−123

51 Dhanjal C, Gunn S R, Shawetaylor J. Efficient sparse kernel feature extraction based on partial least squares. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(8): 1347−1361

52 Tang J, Yu W, Chai T, Liu Z, Zhou X J. Selective ensemble modeling load parameters of ball mill based on multi-scale frequency spectral features and sphere criterion. *Mechanical Systems and Signal Processing*, 2016, **66−67**: 485−504

53 Joe Qin S. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 2012, **36**(2): 220−234

54 Yin S, Ding S X, Haghani A, Hao H Y, Zhang P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, 2012, **22**(9): 1567−1581

55 Ge Z Q, Song Z H, Gao F R. Review of recent research on data-based process monitoring. *Industrial and Engineering Chemistry Research*, 2013, **52**(10): 3543−3562

56 Yin S, Li X W, Gao H J, Kaynak O. Data-based techniques focused on modern industry: an overview. *IEEE Transactions on Industrial Electronics*, 2014, **62**(1): 657−667

57 Cong Y, Wang S, Fan B J, Yang Y S, Yu H B. UDSFS: unsupervised deep sparse feature selection. *Neurocomputing*, 2016, **196**: 150−158

58 Liu Z, Chai T Y, Yu W, Tang J. Multi-frequency signal modeling using empirical mode decomposition and PCA with application to mill load estimation, *Neurocomputing*, 2014, 169: 392−402

59 Motai, Y. Kernel association for classification and prediction: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(2): 208−223

60 de Lázaro J M B, Moreno A P, Santiago O L, da Silva Neto A J. Optimizing kernel methods to reduce dimensionality in fault diagnosis of industrial systems. *Computers and Industrial Engineering*, 2015, **87**: 140−149

61 Tang J, Liu Z, Zhang J, Chai T Y, Yu W. Kernel latent features adaptive extraction and selection method for multicomponent non-stationary signal of industrial mechanical device. *Neurocomputing*, 2016, **216**: 296−309

62 Li D C, Liu C W. Extending attribute information for small data set classfication. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **24**(3): 452−464

63 Shawe-Taylor J, Anthony M, Biggs N L. Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 1993, **42**(1): 65−73

64 Muto Y, Hamamoto Y. Improvement of the Parzen classifier in small training sample size. *Intelligent Data Analysis*, 2001, **5**(6): 477−490

65 Raudys S J, Jain A K. Small sample size effects in statistical pattern recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, **13**(3): 252−264

66 Duin R P W. Small sample size generalization. In: Proceedings of the 9th Scandinavian Conference on Image Analysis. Uppsala, Sweden: Mendeley, 1995. 957−964

67 Yang J, Yu X, Xie Z Q, Zhang J P. A novel virtual sample generation method based on Gaussian distribution. *Knowledge-Based Systems*, 2011, **24**(6): 740−748

68 Li D C, Chen L S, Lin Y S. Using Functional Virtual Population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, 2003, **41**(17): 4011−4024

69 Li D C, Wu C S, Tsai T I, Chang F M. Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge. *Computers and Operations Research*, 2006, **33**(6): 1857−1869

70 Lin Y S, Li D C. The Generalized-Trend-Diffusion modeling algorithm for small data sets in the early stages of manufacturing systems. *European Journal of Operational Research*, 2010, **207**(1): 121−130

71 Lei Y G, Lin J, He Z J, Zuo M J. A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 2013, **35**(1−2): 108−126

72 Efron B, Tibshirani R. Improvements on Cross-Validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 1997, **92**(438): 548−560

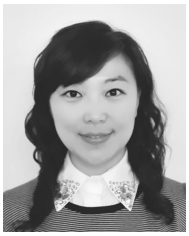
73 Krzanowski W J, Hand D J. Assessing error rate estimators: the leave-one-out method reconsiderd. *Australian and New Zealand Journal of Statistics*, 2010, **39**(1): 35−46

74 Mevik B H, Cederkvist H R. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 2004, **18**(9): 422−429

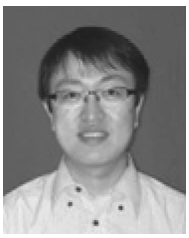
汤 健 北京工业大学教授. 主要研究方向为复杂工业过程智能控制与建模, 数据驱动软测量. 本文通信作者.
E-mail: freeflytang@bjut.edu.cn
(**TANG Jian** Professor at Beijing University of Technology. His research interest covers intelligent control and modeling for complex industrial processes, and data driven-based soft sensor. Corresponding author of this paper.)

乔俊飞 北京工业大学教授. 主要研究方向为智能控制, 神经网络分析与设计.
E-mail: junfeq@bjut.edu.cn
(**QIAO Jun-Fei** Professor at Beijing University of Technology. His research interest covers intelligent control, analysis and design of neural networks.)

柴天佑 中国工程院院士, 东北大学教授. IEEE Fellow, IFAC Fellow, 欧亚科学院院士. 主要研究方向为自适应控制, 智能解耦控制, 流程工业综合自动化理论、方法与技术.
E-mail: tychai@mail.neu.edu.cn
(**CHAI Tian-You** Academician of Chinese Academy of Engineering, professor at Northeastern University, and academician of the International Eurasian Academy of Sciences. His research interest covers adaptive control, intelligent decoupling control, and integrated automation thoery, method and technology of industrial process.)



刘 卓 博士, 东北大学流程工业综合自动化国家重点实验室讲师. 主要研究方向为复杂工业过程建模.
E-mail: liuzhuo@ise.neu.edu.cn
(**LIU Zhuo** Ph. D., lecturer at the State Key Laboratory of Synthetical Automation for Process Industries of Northeastern University. Her research interest covers modeling for complex industries.)



吴志伟 博士, 东北大学讲师. 主要研究方向为复杂工业过程的智能优化控制.
E-mail: wuzhiwei2006@163.com
(**WU Zhi-Wei** Ph. D., lecturer at the State Key Laboratory of Synthetical Automation for Process Industries of Northeastern University. His research interest covers intelligent optimal control for complex industries.)