控制与决策
Control and Decision

面向工业软测量应用的定制化生成对抗数据填补模型

姚邹静， 赵春晖， 李元龙， 付川， 乔红麟

引用本文:
姚邹静, 赵春晖, 李元龙, 等. 面向工业软测量应用的定制化生成对抗数据填补模型[J]. 控制与决策, 2021, 36(12): 2929−2936.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0974

面向工业软测量应用的定制化生成对抗数据填补模型

姚邹静[1], 赵春晖[1†], 李元龙[2], 付 川[2], 乔红麟[2]

(1. 浙江大学 控制科学与工程学院，杭州 310027；2. 阿里巴巴集团，杭州 310024)

摘 要: 在工业领域,数据缺失十分普遍,对解决下游任务(如软测量、异常检测)造成阻碍,这些任务大多依赖完整而高质量的数据集构造模型. 现有缺失数据填补方法很少考虑数据填补后的具体下游任务(软测量). 如何根据下游任务针对性地进行数据填补是当前研究中的挑战之一. 为此,提出一种加入临时软测量模块的对抗生成数据填补模型(SSIGAN). 与生成对抗数据填补模型(GAIN)相比,SSIGAN模型显式地考虑了软测量损失对数据填补模型的影响,通过临时软测量模块指导对质量相关变量的修复,实现数据填补的“定制化”,用于更精准的工业软测量建模. 通过某工业炼钢过程中的终点成分软测量实验,验证了所提出方法对软测量质量相关变量缺失数据填补效果以及最终软测量效果的提升.

关键词: 工业过程；缺失数据；数据填补；生成对抗网络；软测量；转炉炼钢

中图分类号: TP183 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0974 **开放科学(资源服务)标识码(OSID):**

引用格式: 姚邹静, 赵春晖, 李元龙, 等. 面向工业软测量应用的定制化生成对抗数据填补模型 [J]. 控制与决策, 2021, 36(12): 2929-2936.

Customized generative adversarial data imputation model for industrial soft sensing

YAO Zou-jing[1], *ZHAO Chun-hui*[1†], *LI Yuan-long*[2], *FU Chuan*[2], *QIAO Hong-lin*[2]

(1. College of Control Science and Engineering，Zhejiang University，Hangzhou 310027，China；2. Alibaba Group，Hangzhou 310024，China)

Abstract: Missing data is quite common in the industrial field, resulting in problems in downstream applications such as soft sensing and anomaly detection, as most data driven methods used in these applications rely on complete and high-quality dataset to build a high-quality model. Current data imputation methods hardly take its following applications like soft sensing into consideration. A considerable challenge is how to refine missing data repair according to its downstream application. In this paper, we propose an imputation generative adversarial network with soft sensor (SSIGAN) which considers the loss of soft sensors as data. Compared with the imputation generative adversarial network (GAIN), the proposed SSIGAN model introduces the influence of data imputation on the soft sensor. The temporary soft sensor model gives guidance for better repair of quality-related variables. Thus, “customized” data imputation can be achieved for building a more accurate industrial soft sensor. An experiment of soft sensing of end-point composition in a steel-making process is conducted and verifies the improvement of data imputation of quality-related variables and that of the soft sensor with the proposed data imputation model.

Keywords: industrial process；missing data；data imputation；generative adversarial network；soft sensor；converter steelmaking

0 引 言

工业领域中,数据缺失是大多数软测量模型建立时需要解决的首要问题[1]. 过程工业中数据缺失有多种原因, 最常见的原因为硬件传感器故障,包括传感器维护或拆卸, 其他原因包括传感器和数据库之间的数据传输错误,数据库访问错误等. 数据缺失会降低样本信息质量,而高质量的建模数据是高质量软测量模型的基础. 软测量模型一般基于各种数据驱动方法[2-5],若数据大量缺失则无法建立准确模型. 如果数据缺失问题未能得到妥善解决,不但会增大后

收稿日期: 2020-07-16；**修回日期:** 2020-11-10.

基金项目: 浙江省工业化与信息化融合联合基金项目(U1709211)；浙江省重点研发基金项目(2019C03100).

责任编委: 邹长亮.

†通讯作者. E-mail: chhzhao@zju.edu.cn.

续软测量建模难度,也会增加最终软测量结果误差,不利于工业生产,扩大经济损失. 数据缺失可分为3种类型[6]: 1)数据缺失情况与变量本身及其他变量均无关,则称数据完全随机缺失(missing completely at random, MCAR); 2)数据缺失情况仅与变量本身有关,则称数据随机缺失(missing cat random, MAR); 3)数据缺失情况与其他变量有关,则数据为非随机缺失(missing not at random, MNAR). 本文针对工业现场中最常见的数据完全随机缺失情况进行研究.

在实际操作中,解决数据完全随机缺失问题的通用方法为完全删除含缺失元素的数据样本. 此法操作简便,且后续计算均基于已知的观察值进行. 然而,保留的数据与完整数据存在偏差,部分关键信息的丢失会降低最终软测量模型精度. 若数据中变量个数众多,即使单个变量数据缺失比例小,在完全缺失数据删除操作下,大部分数据都会被删除,保留数据的数据分布可能与删除前的情况有很大偏差[7].

另一大类方法为数据填补,其目的是尽量构造完整数据集,并减少由于数据填补造成的估计偏差. 传统的数据填补方法可分为单一填补法和多重填补法(multiple imputation, MI)[8-9]. 单一填补法结构简单效率高,其中均值填补最常用,此时所有插补值集中在均值点上,会严重扭曲数据分布,在均值上形成尖峰. 多重填补法可以改善单一填补法中数据偏差较大且未考虑数据随机性的问题. Donald于1978年提出的多重填补法[10],以贝叶斯理论为基础,引入抽样过程,多次迭代填补数据集,可采用回归、Logistic回归、判别分析及马尔科夫蒙特卡罗等多种不同的方法进行数据填补[11]. 多重填补法将数据单次填补的不确定性纳入考量,可更好地保持变量间关系,从而增加估计的有效性.

最新的机器学习类数据填补方法可分为判别方法和生成方法. 判别方法包括MICE多重插补[12-14]、MissForest[15-17]和矩阵补全[18-20];生成方法主要包括基于深度学习的算法,如变分自动编码器(variational autoencoder, VAE)[21-23]、去噪自动编码器(denoising autoencoder, DAE)[24-26]和生成对抗网络(generative adversarial network, GAN)[27-31]. MissForest[15]训练时不需要完整的数据,对其他缺失元素经过简单填补后,对某选定元素进行缺失填补,逐步填补后续缺失元素,因此会因为简单填补造成的数据分布扭曲影响数据填补质量. 基于VAE[21-22]的数据填补方法对基础分布进行了假设,无法拟合数据分布比较复杂时的情况. 而基于DAE的方法大多在训练过程中需要完整数据[24],部分衍生方法不需要完整数据,但会使用均值替代缺失值[25]. 基于GAN的方法常用作图像填补[28],同样需要完整数据来训练模型. Yoon等[29]提出一种GAIN模型用于通用数据填补,该模型既能通过神经网络形成复杂的数据分布,又不需要完整数据集进行模型训练,但文中未结合工业软测量应用进行讨论. 在许多情况下,数据缺失是问题固有结构的一部分,特别是数据完全随机缺失时,基本无法获得完整数据集.

以上缺失数据填补方法均未能结合数据填补后的具体应用目的进行调整,数据填补与数据填补后的下游任务是割裂的两部分. 如何考虑填补效果对后续应用的影响,从而对应用中的关键参数针对性地提高填补精度,实现定制化数据填补,成为当前研究中的挑战. 本文提出一种加入软测量模块的对抗生成数据填补模型(imputation generative adversarial network with soft sensor, SSIGAN),主要贡献在于显式地考虑了软测量损失对GAIN数据填补模型的影响,以提升对质量相关变量的修复效果,实现数据填补的“定制化”,用于更精准的工业软测量建模.

1 定制化SSIGAN模型

本节对文中提出的定制化SSIGAN模型进行详细阐述. 该模型考虑软测量这一具体应用目的来进行数据填补.

1.1 GAN模型结构

SSIGAN具备GAN的基本骨架,包含一个生成器(G)和一个判别器(D). 在基本的GAN中,G捕捉真实数据样本的潜在分布, 并生成新的数据样本,D判别输入是真实数据还是生成的样本. GAN的核心思想为G和D相互对抗迭代优化,寻找二者间的纳什均衡[32]. 基本的GAN模型结构如图1所示.

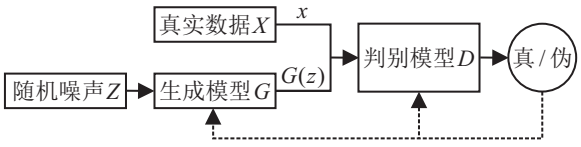


图1 基本的GAN模型结构

1.2 SSIGAN模型结构

本文提出的SSIGAN模型结构如图2所示. 除G与D, SSIGAN的结构中还包含提示矩阵生成器(hint generator),以及一个用于指导数据填补的临时软测量模型(temporary soft sensor). GAIN则不含临时软测量模型模块.

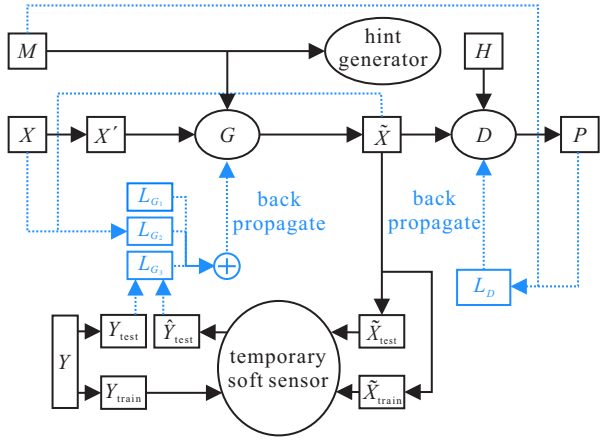


图 2　SSIGAN 模型结构

SSIGAN 模型显式地考虑软测量损失对数据填补模型的影响,将软测量损失引入 G 的损失函数中. 因此,SSIGAN 中的 G 不仅需生成尽可能真实的数据,使生成的填补数据能欺骗 D,同时需要降低数据填补波动对后续软测量的影响,间接提升质量相关变量的修复精度.

图 2 中,黑色实箭头表示数据流. 输入数据为 $X \in \boldsymbol{R}^{I\times J}, M \in \boldsymbol{R}^{I\times J}, Y \in \boldsymbol{R}^{I\times 1}$. X 为含缺失量的数据矩阵; M 为显示数据缺失情况的掩码矩阵, $M(i,j)$ 对应 $X(i,j)$ 分量的缺失情况,若该处数据缺失则为 0,不缺失则为 1; Y 为软测量标签矩阵, $Y(i1)$ 对应 X 中第 i 条样本对应所需软测量元素的真实值,或称软测量标签; I 为每次迭代时输入的样本条数; J 为用于软测量的预测变量个数. 将 X 和 M 输入 G, G 为全连接层,输出得数据填补后的矩阵 $\tilde{X}$;将 M 输入提示矩阵生成器,生成提示矩阵 $H \in \boldsymbol{R}^{I\times J}$;将 $\tilde{X}$ 和软测量标签矩阵 Y 划分为训练集和测试集后,训练临时软测量模型,输出测试集的软测量损失,以此指导数据填补过程; $\tilde{X}$ 和 H 被输入 D, D 也为全连接层,输出得判别矩阵 $P \in \boldsymbol{R}^{I\times J}$. G、D 的全连接层均为 3 层,分别是输入层、隐含层和输出层,神经元个数分别为 $2J$、J、J.

本节会对 SSIGAN 的主要组件及其对应的计算公式以及 SSIGAN 的训练算法进行阐述.

1.2.1　SSIGAN 的主要组件

SSIGAN 主要组件包括 G、D、提示矩阵生成器、临时软测量模型. 本小节将详细介绍这些组件的输入输出以及对应计算公式.

$$X' = M \odot X + (1-M) \odot Z, \tag{1}$$

$$\hat{X} = G(X', M), \tag{2}$$

$$\tilde{X} = M \odot X + (1-M) \odot \hat{X}. \tag{3}$$

G: G 有两部分输入,一部分是 X 加入了随机噪声 Z 后的 X',对应式 (1);一部分是掩码矩阵 M, G 输出估计矩阵 $\hat{X}$,对应式 (2),需还原未缺失数据得到矩阵 $\tilde{X}$,对应式 (3).

对于单条样本 x,对应有 x'、m 输入 G. x' 部分为未缺失分量,即 $m \odot x$,部分为代替缺失分量的随机噪声分量,即 $(1-m) \odot z$, m 为对应的掩码向量. G 生成 $\hat{x}$ 之后,需将未缺失分量还原,变为 $\tilde{x}$.

SSIGAN 中的 Z 与标准 GAN 中引入的噪声变量类似[32]. G 生成数据的过程可以视为根据神经网络拟合的某种数据分布 $P(X)$ 生成样本. 在此框架中,目标分布 $P(X|X')$ 本质上是 $\|1-M\|$ 维,因此传递到生成器中的噪声为 $(1-M) \odot Z$ 而不是 Z,使得噪声的尺寸与待填补数据目标分布的尺寸相匹配.

D: D 的输入为 G 的输出 $\tilde{X}$,及提示矩阵 H. 提示矩阵 H 在下文中有阐述. D 输出 P, $P(i,j)$ 对应 $\tilde{X}(i,j)$ 为真实数据而非生成数据的概率,该概率最小为 0,最大为 1,对应下式:

$$P = D(\tilde{X}, H). \tag{4}$$

提示矩阵生成器: 提示矩阵生成器的输入为掩码矩阵 M, SSIGAN 通过它生成的提示矩阵 H 控制传入 D 中有关掩码矩阵 M 的信息量. 若 H 不包含有关 M 的“足够”信息,则无法保证 G 最后能够学习到基于真实数据的数据分布[28]. 设矩阵 $B(IJ)$, $B(i,j)$ 从 $\{0,1\}$ 集合中随机采样产生, H 的计算如下所示:

$$H = B \odot M + 0.5 \odot (1-B). \tag{5}$$

对于单条样本 h,易知当 $b(j) = 1$ 时, $h(j) = m(j)$,传输 $m(j)$ 的信息;当 $b(j) = 0$ 时, $h(j) = 0.5$,未传输 $m(j)$ 的信息. M 的部分信息由此通过 H 传入 D 中.

临时软测量模型: 将 G 的输出 $\tilde{X}$ 与软测量标签矩阵 Y 按比例 4 : 1 划分为训练集 $\tilde{X}_{\text{train}}$、Y_{train} 和测试集 $\tilde{X}_{\text{test}}$、Y_{test}. 使用训练集建立软测量模型,本文中采用 XGBOOST 方法[33] 建立该模型,将 $\tilde{X}_{\text{test}}$ 输入该模型,可得 $\tilde{X}_{\text{test}}$ 对应的软测量值 $\hat{Y}_{\text{test}}$,如下所示:

$$(Y_{\text{test}}, \hat{Y}_{\text{test}}) = \text{temporary Soft sensor}(\tilde{X}, Y). \tag{6}$$

1.2.2　SSIGAN 的损失函数

由于 D 的目标是尽可能分辨数据是否真实,一般设定 D 的损失函数为一个交叉熵损失函数,如下式所示:

$$L_D(M,P) = -\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}[M(i,j)\log(P(i,j)) +$$

$$(1-M(i,j))\log(1-P(i,j))], \tag{7}$$

但由于SSIGAN中有提示矩阵H,某条输出$\tilde{x}$对应的b中某分量$h(j)=m(j)$时,D无需G输出的$\tilde{x}$的信息,即可判断其分量$\tilde{x}(j)$是否为生成数据,因此该部分的损失无法帮助训练D和G. 因此,只将$b(j)=0$时的交叉熵损失计入D的损失函数中,如下所示:

$$L_D(M,P,B)=$$
$$-\frac{1}{Bk}\sum_{i=1}^{I}\sum_{j=1;B(i,j)=0}^{J}[M(i,j)\log(P(i,j))+$$
$$(1-M(i,j))\log(1-P(i,j))]. \tag{8}$$

其中Bk为B中$B(i,j)=0$元素的总个数.

由此,D的优化目标变为最小化$L_D(M,P,B)$.

G的目标是尽量使生成的填补数据接近能欺骗D,并生成尽量真实的数据,同时要考虑数据填补对软测量模型的影响,因此G的损失函数包括3部分.

第1部分为填补后数据进入D造成的损失,与D的损失计算类似,对于某条输出$\tilde{x}$只将$b(j)=0$时的损失计入,损失计算公式为

$$L_{G_1}(M,P,B)=$$
$$-\frac{1}{Bk}\sum_{i=1}^{I}\sum_{j=1;B(i,j)=0}^{J}(1-M(i,j))\log(1-P(i,j)). \tag{9}$$

第2部分为未缺失数据的重构误差,确保生成数据的分布与真实数据分布相似. 由于实际数据缺失场景中,仅有未缺失值的真实值,无法获取缺失数据的真实值,使用G输出的未还原未缺失数据的$\hat{X}$,计算对未缺失数据的重构误差,如下式所示:

$$L_{G_2}(X,\hat{X},M)=$$
$$-\frac{1}{Mk}\sum_{i=1}^{I}\sum_{j=1}^{J}(M(i,j)X(i,j)-M(i,j)\hat{X}(i,j))^2. \tag{10}$$

Mk为M中$M(i,j)=0$元素的总个数,即未缺失元素个数.

第3部分需体现数据填补对软测量模型的影响. 使用$\tilde{X}$和相应软测量标签Y,可训练临时软测量模型:按照4:1划分训练数据与测试数据,得到训练集$\tilde{X}_{\text{train}},Y_{\text{train}}$,以及测试集$\tilde{X}_{\text{test}},Y_{\text{test}}$. 使用XGBOOST方法训练软测量模型,使用决定系数R^2作为模型评价指标,测试集上的软测量结果为$\tilde{Y}_{\text{test}}$,则损失计算公式为

$$L_{G_3}(Y_{\text{test}},\hat{Y}_{\text{test}})=1-R^2=$$
$$\frac{\sum_{i=1}^{I}(\hat{Y}_{\text{test}}(i)-Y_{\text{test}}(i))^2}{\sum_{i=1}^{I}(\bar{Y}_{\text{test}}-Y_{\text{test}}(i))^2}. \tag{11}$$

$\bar{Y}_{\text{test}}$为Y_{test}的均值.

由此,G的优化目标变为最小化L_G,损失计算公式为

$$L_G(M,P,B,X,\hat{X},Y_{\text{test}},\hat{Y}_{\text{test}})=$$
$$L_{G_1}(M,P,B)+\alpha L_{G_2}(X,\hat{X},M)+$$
$$\beta L_{G_3}(Y_{\text{test}},\hat{Y}_{\text{test}}). \tag{12}$$

α和β为超参数,分别是L_{G_2}和L_{G_3}在L_G中的权重系数.

1.3 面向软测量应用的SSIGAN模型建立步骤

首先准备训练集和测试集. 总样本条数为Q,A为训练集中样本条数,C为测试集中样本条数,$Q=A+C$. J为用于软测量的预测变量个数.

图3为面向软测量应用的SSIGAN模型建立步骤,具体如下.

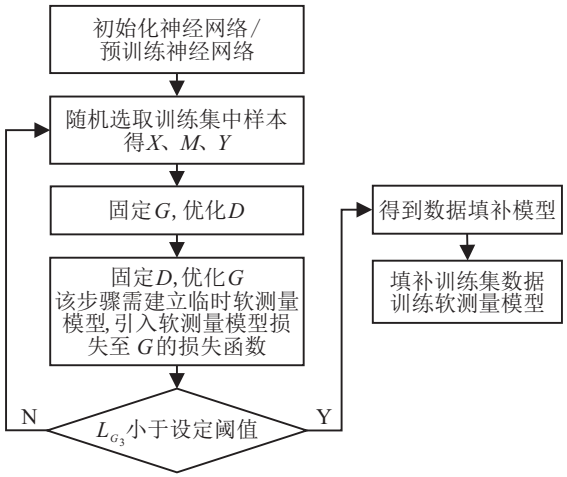


图3 面向软测量应用的SSIGAN模型建立步骤

step 1: 初始化G和D的神经网络参数.

step 2: 从训练集中随机采取I条样本,得到所需含缺失量的数据矩阵$X\in\boldsymbol{R}^{I\times J}$,及其对应的掩码矩阵$M\in\boldsymbol{R}^{I\times J}$、软测量标签矩阵$Y\in\boldsymbol{R}^{I\times 1}$.

step 3: 固定G,使用Adam算法优化D. 优化前需用现有的G得到数据填补后的矩阵$\tilde{X}\in\boldsymbol{R}^{I\times J}$,并通过提示矩阵生成器得提示矩阵$H\in\boldsymbol{R}^{I\times J}$,用待优化的$D$得到判别矩阵$P\in\boldsymbol{R}^{I\times J}$,用于优化的损失函数$L_D$计算对应1.2.2中的式(8).

step 4: 固定D,使用Adam算法优化G. 优化前需用现有的D得到判别矩阵$P\in\boldsymbol{R}^{I\times J}$,用待优化的$G$得到数据填补后的矩阵$\tilde{X}\in\boldsymbol{R}^{I\times J}$,且需计算$\tilde{X}\in\boldsymbol{R}^{I\times J}$与对应的$Y\in\boldsymbol{R}^{I\times 1}$建立临时软测量模型后

的测试损失. 用于优化的损失函数损失函数 L_G 计算对应 1.2.2 中的式(12); 然后返回 step 2 顺序进行, 直至 L_{G_3} 小于设定阈值 ε, 若小于设定阈值 ε, 则 SSIGAN 数据填补模型训练完毕.

step 5: 训练软测量模型. 面向软测量应用时, 先将训练集中的含缺失量的数据矩阵 $X \in \boldsymbol{R}^{A\times J}$ 输入生成器中完成数据填补, 得 $\tilde{X} \in \boldsymbol{R}^{A\times J}$, 与对应的 $Y \in \boldsymbol{R}^{A\times 1}$ 一起训练软测量模型, 由于 SSIGAN 损失函数中引入的软测量损失为使用 XGBOOST[33] 训练软测量模型所得, 此处也对应选用 XGBOOST 方法.

通过上述步骤, 可得数据填补模型以及软测量模型. 本文提出的面向软测量应用的 SSIGAN 数据填补模型中的软测量方法不局限于 XGBOOST 方法. 若需改变软测量方法, 则对应改变 SSIGAN 中的损失函数以及 step 5 中的软测量方法. 为提升模型收敛速度, 可采用 GAIN 模型进行神经网络参数的预训练代替 step 1.

2　案例研究

本文将面向软测量应用的 SSIGAN 数据填补模型应用于转炉炼钢的终点元素含量预测中铝 (Al) 成分的预测, 并基于真实炼钢过程的数据集进行了实验, 以验证其有效性. 转炉炼钢以铁水、废钢、铁合金为主要原料, 不借助外加能源, 靠铁液本身的物理热和铁液组分间化学反应产生热量, 在转炉中完成炼钢, 钢水需要达到要求的温度与成分, 终点元素含量预测在转炉炼钢中尤为重要[34]. 转炉炼钢生产过程中, 常有记录员遗漏等原因导致的数据缺失问题. 因此, 研究该场景下缺失数据填补后对终点元素进行软测量, 对实际生产很有帮助. 表 1 介绍了数据集的情况. 数据集经归一化处理, 采用 min-max 标准化将每个变量转化至 [0, 1] 范围内. 随机选取 80 % 的数据作为训练集, 剩下 20 % 的数据作为测试集.

表 1　数据集介绍

属性	说明
样本个数	2 875
过程变量个数	72
软测量目标变量	铝 (Al)
过程变量的构成	12 个加合金前各元素含量 14 个不同合金的添加量 2 个钢水的温度和质量 11 个操作变量 33 个其他变量

2.1　不同数据缺失率下的模型表现

基于原始数据, 分别设定数据缺失率为 0.1、0.2、0.3、0.4、0.5、0.6, 且数据为完全随机缺失, 使用不同缺失率数据集进行实验得到相应的软测量结果. 分别使用 4 种方法填补缺失数据作对比实验, 即本文 SSIGAN 方法、GAIN[28] 方法、MissForest[15] 方法以及均值填补方法. 数据填补后使用 XGBOOST[33] 建立软测量模型. 使用缺失率为 0 的数据集建立软测量模型, 其模型效果作为实验对比基准.

实验时, 选取常用的模型误差评价指标决定系数 R^2 和均方根误差 RMSE 作为软测量效果评价指标, 计算公式如下所示

$$R^2 = \frac{\sum_{i=1}^{C}(\hat{\mathrm{Tar}}\mathrm{get}(i) - \bar{\mathrm{Tar}}\mathrm{get}(i))^2}{\sum_{i=1}^{C}(\mathrm{Target}(i) - \bar{\mathrm{Tar}}\mathrm{get}(i))^2}. \tag{13}$$

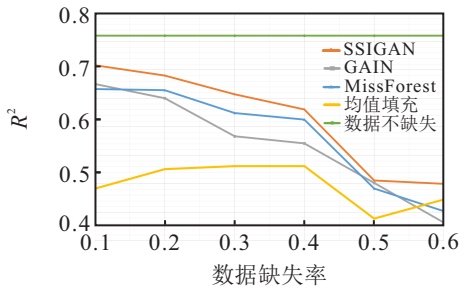
$$\mathrm{RMSE} = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(\hat{\mathrm{Tar}}\mathrm{get}(i) - \mathrm{Target}(i))^2}. \tag{14}$$

其中: Target 为目标参数的真实值, $\bar{\mathrm{Tar}}\mathrm{get}$ 为目标参数真实值的均值, $\hat{\mathrm{Tar}}\mathrm{get}$ 为对目标参数的估计结果.

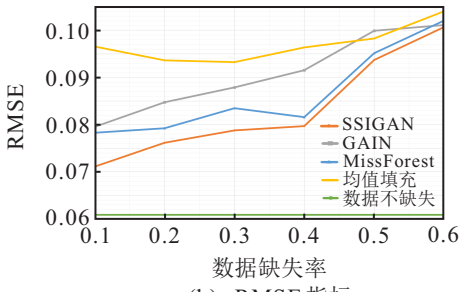
模型更新时的参数设定说明如下. 设定每次迭代时输入的样本条数 I 为 900, 超参数 α 和 β 分别为 10、100, 均为多次实验所得经验值. 阈值 ε 使用均值填补后的 R^2 代入 $(1 - R^2 \times 1.2)$ 计算得到, 若超过 8 000 次迭代更新仍未达到终止条件, 则停止模型迭代. 该阈值 ε 以均值填补后软测量效果为参考, 使得模型在优化至相对均值填补效果更好时停止更新. SSIGAN 的数据填补效果相对均值填补效果的好坏由此时与 R^2 相乘的系数决定, 该系数需大于 1, 本文中取 1.2, 为经验设定值. 在 GAIN 中, 只有超参数 α, 令其等于 10.

图 4 展示了不同数据缺失率下使用各模型进行数据填补后建立的软测量模型效果. 其中, 数据填补模型对应折线越接近数据不缺失时对应直线, R^2 越接近 1, RMSE 越接近 0, 说明该数据填补方法对最终软测量效果提升越明显.

由图 4 可以看出, 随着数据缺失率上升, 各模型对应的软测量效果均呈下降趋势, 而 SSIGAN 模型的表现始终优于其他几种模型. 以 R^2 作为评价指标, 在 6 种不同缺失率的数据集中, 对比 GAIN、MissForest 以及均值填补后的软测量模型, SSIGAN 数据填补后的软测量模型平均分别有 8.46 %、5.31 % 和 23.16 % 的提升. 以 RMSE 作为评价指标, 则平均分别有 8.43 %、3.97 % 和 14.28 % 的提升. 由上述分析, SSIGAN 数据填补模型对最终的软测量模型效果的提升高于其他方法.



(a) R^2 指标



(b) RMSE 指标

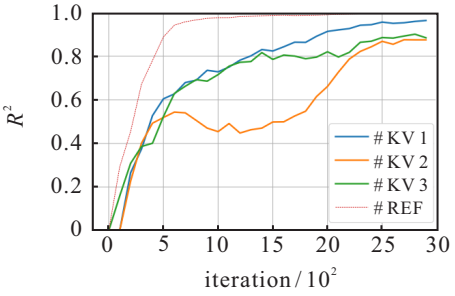
图 4 不同数据缺失率下的软测量模型效果

2.2 模型对质量相关变量的修复

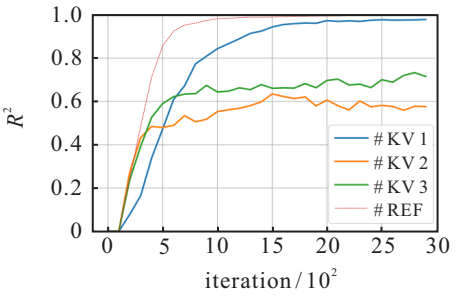
对于软测量这一下游任务,过程变量可被划分为软测量的质量相关变量和非相关变量,从机理分析可得3个关键的质量相关变量,分别为钢种、实际值-硅锰合金、硅铁. 实验中还人为增加了一个与软测量不相关的随机变量作为参考变量. 本小节将研究数据填补环节中模型在关键变量上的表现,探究基于SSIGAN数据填补的软测量模型优于其他方法的内在原因.

以GAIN模型作为对比模型,使用0.2缺失率下的数据集进行实验. SSIGAN和GAIN模型都可以实现对缺失数据的填补和对未缺失数据的估计. 生成器的输出 $\tilde{X}$ 在没有执行还原未缺失数据步骤前,是估计矩阵 $\hat{X}$, $\hat{X}$ 不仅对缺失元素进行填补,也对未缺失元素进行估计. 将决定系数 R^2 和均方根误差RMSE作为评价指标,可得两种模型在全部数据集中对质量相关变量的缺失数据填补效果和未缺失数据估计效果. R^2、RMSE的计算分别对应式(13)和(14). 为方便图表展示,若计算得 R^2 小于0,则将其置为零. SSIGAN和GAIN模型训练时参数设置与2.1中的设置对应一致.

表2中展示了SSIGAN与GAIN模型训练结束后在全部数据集中对软测量质量相关变量的缺失元素填补以及非缺失元素估计的效果对比情况. 括号内为SSIGAN对比GAIN模型所得效果的提升百分比. 图5展示了SSIGAN模型与GAIN模型在迭代更新时全部数据集中对软测量质量相关变量非缺失元素估计效果的变化情况,使用了 R^2 作为评价指标,对应计算公式(13). 图5中编号#KV1~#KV3对应3个关键的质量相关变量,分别为钢种、实际值-硅锰合金、硅铁;编号#REF对应参考变量.



(a) SSIGAN



(b) GAIN

图 5 SSIGAN与GAIN对质量相关变量未缺失元素估计效果的变化

由表2可以看出,SSIGAN实现数据填补时对软测量质量相关变量的缺失数据填补效果以及未缺失元素估计效果都优于GAIN模型. 以 R^2 为指标,质量相关变量缺失元素的填补质量至少比GAIN提升4.23 %,以RMSE为指标,则至少提升1.73 %. 虽然在“钢种”未缺失元素的估计效果上SSIGAN表现均略

表 2 质量相关变量上SSIGAN与GAIN的表现

评价指标	R^2				RMSE			
效果	缺失元素填补		未缺失元素估计		缺失元素填补		未缺失元素估计	
模型	SSIGAN	GAIN	SSIGAN	GAIN	SSIGAN	GAIN	SSIGAN	GAIN
参考变量	0.000 00 (+0.00 / %)	0.000 00	0.995 67 (−0.14 / %)	**0.997 06**	0.307 30 (−2.99 / %)	**0.298 39**	0.018 87 (−21.5 / %)	**0.015 53**
钢种	**0.013 34** (+infty / %)	0.000 01	0.957 47 (−2.20 / %)	**0.979 03**	**0.234 32** (+1.73 / %)	0.238 44	0.047 99 (−42.4 / %)	**0.033 70**
硅锰合金	**0.171 58** (+66.7 / %)	0.102 93	**0.865 30** (+47.9 / %)	0.585 14	**0.163 44** (+3.90 / %)	0.170 08	**0.058 59** (+43.0 / %)	0.102 83
硅铁	**0.398 90** (+4.23 / %)	0.382 71	**0.844 59** (+18.2 / %)	0.714 36	**0.139 78** (+5.54 / %)	0.147 98	**0.075 80** (+26.2 / %)	0.102 76

逊于GAIN,但在其缺失元素的填补效果上表现仍占优. 对参考变量进行缺失元素填补和未缺失元素估计时,SSIGAN表现均略逊于GAIN,但由于参考变量在设定时为随机变量,与软测量这一下游任务无关,对它的缺失元素填补或未缺失元素估计的效果好坏对后续软测量没有影响.

由图5可以看出,随着模型迭代更新,SSIGAN对质量相关变量未缺失元素的估计效果逐渐在总体上优于GAIN,并在某些质量相关变量中大幅超过GAIN. 比如对#KV2和#KV3,即硅锰合金和硅铁,未缺失元素估计效果明显优于GAIN. 生成器损失函数中的$L_{G_2}(X,\hat{X},M)$ 即对应未缺失元素估计误差,可知随着SSIGAN模型的优化迭代,L_{G_2}中由质量相关变量造成的损失比GAIN中对应损失小,且该部分损失也随着模型迭代训练呈下降趋势. 结合表2中的数据,可知SSIGAN对质量相关变量的缺失数据填补效果均优于GAIN,数据填补效果优势明显. 图4中对应的SSIGAN最终软测量模型效果比GAIN也有很大提升. SSIGAN中生成器损失函数L_G中对比GAIN多构造的软测量损失L_{G_3},可以指导模型在具体软测量质量相关变量未知的情况下,自动向着有利于提升质量相关变量修复效果的方向优化,根据下游软测量任务定制化数据填补,从而提升软测量模型的最终效果.

3 结　论

本文提出了一种面向工业软测量的SSIGAN通用数据填补模型,并通过对比实验验证了该模型的有效性. 所提出SSIGAN模型将原本割裂的数据填补和软测量结合在一起,首次提出并实现了对工业软测量定制的数据填补. 模型中的生成器产生的完整数据集被用来建立临时软测量模型,并向生成器的损失函数中加入该临时模型的测试损失,在未知具体软测量质量相关变量时,指导生成器增强对这些变量的数据修复,使模型向着有利于下游软测量任务的方向迭代更新,使最终的工业软测量模型更加精准. 之后的工作还可考虑面向其他具体应用目的定制化数据填补,比如过程监测等.

参考文献(References)

[1] Shang C, Yang F, Huang D X, et al. Data-driven soft sensor development based on deep learning technique[J]. Journal of Process Control, 2014, 24(3): 223-233.

[2] Yuan X F, Huang B, Wang Y L, et al. Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE[J]. IEEE Transactions on Industrial Informatics, 2018, 14(7): 3235-3243.

[3] Qin Y, Zhao C H, Huang B. A new soft-sensor algorithm with concurrent consideration of slowness and quality interpretation for dynamic chemical process[J]. Chemical Engineering Science, 2019, 199: 28-39.

[4] Zhao C H. A quality-relevant sequential phase partition approach for regression modeling and quality prediction analysis in manufacturing processes[J]. IEEE Transactions on Automation Science and Engineering, 2014, 11(4): 983-991.

[5] Wang J, Zhao C H. Mode-cloud data analytics based transfer learning for soft sensor of manufacturing industry with incremental learning ability[J]. Control Engineering Practice, 2020, 98: 104392.

[6] Little R J A, Donald B Rubin. Statistical analysis with missing data[M]. Brisbane: JohnWiley & Sons, 1987: 1-23.

[7] Afifi A A, Elashoff R M. Missing observations in multivariate statistics I. review of the literature[J]. Journal of the American Statistical Association, 1966, 61(315): 595-604.

[8] Ehrlinger L, Grubinger T, Varga B, et al. Treating missing data in industrial data analytics[C]. The 13th International Conference on Digital Information Management (ICDIM). Berlin, 2018: 148-155.

[9] 曹阳, 谢万军, 张罗漫. 多重填补的方法及其统计推断原理[J]. 中国医院统计, 2003, 10(2): 77-81.
(Cao Y, Xie W J, Zhang L M. Methods of multiple imputation and related inference theory[J]. Chinese Journal of Hospital Statistics, 2003, 10(2): 77-81.)

[10] Donald B Rubin. Multiple imputation for nonresponsein surveys[M]. NewYork: JohnWiley & Son, 1987: 75-107.

[11] 袁中英. 多元线性回归模型中缺失数据填补方法的效果比较[D]. 长沙: 中南大学, 2008.
(Yuan Z Y. Comparison of data imputation methods on linearŋregression model[D]. Changsha: Central South University, 2008.)

[12] van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations inR[J]. Journal of Statistical Software, 2011, 45(3): 1-67.

[13] Hegde H, Shimpi N, Panny A, et al. MICE vs PPCA: Missing data imputation in healthcare[J]. Informatics in Medicine Unlocked, 2019, 17: 100275.

[14] 张莹. 基于多重插补的代谢综合征自我预测模型研究[D]. 杭州: 浙江大学, 2019.
(Zhang Y. Study on a self-rediction model for metabolic syndrome based on interpolated features[D]. Hangzhou: Zhejiang University, 2019.)

[15] Stekhoven D J, Bühlmann P. MissForest: Non-parametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012, 28(1): 112-118.

[16] 张晓琴, 程誉莹. 基于随机森林模型的成分数据缺失值填补法[J]. 应用概率统计, 2017, 33(1): 102-110.
(Zhang X Q, Cheng Y Y. Imputation of missing values for compositional data based on random forest[J]. Chinese

Journal of Applied Probability and Statistics, 2017, 33(1): 102-110.)

[17] Chai Z, Zhao C H. Multiclass oblique random forests with dual-incremental learning capacity[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5192-5203.

[18] Cai Jian-Feng, Emmanuel J Candès, Shen Zuowei. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal of Optimization, 2010, 20(4): 1956-1982.

[19] Zhao Kang, Chong Peng, Qiang Cheng. Top-N recommender system via matrix completion[C]. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, 2016: 179-184.

[20] Liu X F, Wang X, Zou L, et al. Spatial imputation for air pollutants data sets via low rank matrix completion algorithm[J]. Environment International, 2020, 139: 105713.

[21] McCoy J T, Kroon S, Auret L. Variational autoencoders for missing data imputation with application to a simulated milling circuit[J]. IFAC-PapersOnLine, 2018, 51(21): 141-146.

[22] Boquet G, Morell A, Serrano J, et al. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection[J]. Transportation Research Part C: Emerging Technologies, 2020, 115: 102622.

[23] Feng L J, Zhao C H, Sun Y X. Dual attention-based encoder-decoder: A customized sequence-to-sequence learning for soft sensor development[J]. IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2020.3015929.

[24] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]. Proceedings of the 25th International Conference on Machine Learning. New York: ACM Press, 2008: 1096-1103.

[25] Abiri N, Linse B, Edén P, et al. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems[J]. Neurocomputing, 2019, 365: 137-146.

[26] 杜婧涵. 基于深度学习的机场噪声监测数据补全研究[D]. 南京: 南京航空航天大学, 2019.
(Du J H. Research on airport noise monitoring data recovery based on deep learning[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2019.)

[27] Shang Chao, Palmer Aaron, Sun Jiangwen, et al. VIGAN: Missing view imputation with generative aversarial networks[C]. Proceedings of IEEE International Conference on Big Data. Boston, 2017: 766-775.

[28] Li Yijun, Liu Sifei, Yang Jimei, et al. Generative face completion[J/OL]. 2017, arXiv:1704.05838v1.

[29] Yoon Jinsung, Jordon James, Schaar Mihaela. GAIN: Missing data imputationusing generative adversarial Nets[C]. Proceedings of the 35th International Conference on Machine Learning. Sweden, 2018: 5689-5698.

[30] 王守相, 陈海文, 潘志新, 等. 采用改进生成式对抗网络的电力系统量测缺失数据重建方法[J]. 中国电机工程学报, 2019, 39(1): 56-64.
(Wang S X, Chen H W, Pan Z X, et al. A reconstruction method for missing data in power system measurement using an improved generative adversarial network[J]. Proceedings of the CSEE, 2019, 39(1): 56-64.)

[31] Chaiand Z, Zhao C. A fine-grained adversarial network method for cross-domain industrial fault diagnosis[J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(3): 1432-1442.

[32] Goodfellow I, Pouget-Abadi J, Mirza M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014: 2672-2680.

[33] Chen Tianqi, Carlos Guestrin. XGBoost: A scalable tree boosting system[J/OL]. 2016, arXiv: 1603.02754.

[34] Jalkanen H, Holappa L. Converter steelmaking[C]. Treatise on Process Metallurgy. Amsterdam: Elsevier, 2014: 223-270.

作者简介

姚邹静(1997−), 女, 博士生, 从事工业软测量的研究, E-mail: yzjing@zju.edu.cn;

赵春晖(1979−), 女, 教授, 博士生导师, 从事工业大数据分析与应用(包括状态监测、故障诊断、软测量)等研究, E-mail: chhzhao@zju.edu.cn;

李元龙(1987−), 男, 研究员, 博士, 从事智慧工厂的研究, E-mail: xunyuan.lyl@alibaba-inc.com;

付川(1986−), 男, 工程师, 从事云计算、大数据应用的研究, E-mail: fuchuan.fc@alibaba-inc.com;

乔红麟(1985−), 男, 研究员, 从事智慧工厂的研究, E-mail: kenny.qhl@alibaba-inc.com.

(责任编辑：孙艺红)