文章编号:1671-4598(2020)04-0032-04 DOI:10.16526/j.cnki.11—4762/tp.2020.04.007 中图分类号:TP277 文献标识码:A

基于 PCA—RFR 的传感器故障定位方法

潘 磊,赵忠盖,刘 飞
(江南大学 物联网工程学院,江苏 无锡 214122)

摘要:基于数据驱动的故障诊断方法近些年来得到广泛的研究和应用,但这些方法主要针对于故障检测,对于故障根源的定位尚未得到充分解决;为此,提出一种基于主成分分析(PCA)和随机森林回归(PFR)的因果分析故障定位方法(PCA—PFR),该方法通过将离线故障数据段中的变量作为输入,与之对应的统计量作为输出建立随机森林回归模型,然后通过模型的变量重要性度量来得到过程变量对统计量的因果关系系数,其中值越大的变量被认为越有可能是引起故障发生的故障变量;最后通过一个数值案例和 TE 过程仿真实验,表明该方法的有效性。

关键词:随机森林回归;变量重要性度量;主成分分析;故障定位;因果分析

Fault Location Method Based on Random Forest Regression

Pan Lei, Zhao Zhonggai, Liu Fei
(College of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: Data—driven fault diagnosis methods have been widely studied and applied in recent years, but these methods are mainly focus on the fault detection, and how to determine the fault source has not been fully solved. For this, presents a causal analysis fault location method (PCA—PRF) based on principal component analysis (PCA) and random forest regression (RFR), In this method, the variables in the off—line fault data segment and the corresponding statistics are treated as the input and the output, respectively, and then the causal coefficient from process variables to statistics can get according to the importance measurement of variables based on the random forest regression model, The variable with larger importance measurement is considered to be more likely fault. A numerical case and the Tenessee Eastman Process (TEP) simulation experiment are employed to demonstrate the application of the proposed method, shows the effectiveness of this proposed method.

Keywords: random forest regression; variable importance; principal component analysis; fault isolation; causal analysis

0 引言

随着现代工业过程传感器检测的广度和深度逐渐扩展,基于数据驱动的故障检测和诊断方法也越来越具有吸引力,在保证工况安全运行的同时,还能有效地提高产品质量[1]。主成分分析(principal component analysis, PCA)作为一种常用的工业数据信息提取方法,可提取样本中的特征信息和残差信息,在过程监控中得到了广泛的应用[2]。PCA 方法通过提取统计量对过程进行监控,若待检测样本统计量大于统计量控制限,则认为此时发生了故障[2-3]。在 PCA 方法检测到故障后需要对故障进行定位,典型的定位方法包括贡献图[4]、重构法[5]和重构贡献[6](reconstruction—based contribution, RBC)。这些故障定位方法认为变量贡献值越大越有可能是故障变量,然而由于受到拖尾效应的影响,一些非故障变量的贡献值同样增大并可能会超过故障变量贡献值,从而导致误诊[2]。基于 PCA 的传统故障定位方法的本质是基于数据之间的相关关系而不是因果关系[7],而这就使得其无法从根本上解决拖尾效应的影响,从而影响其在实际应用中的效果。

基于数据驱动的因果分析方法由于其在厘清变量之间因果关系中的作用,近些年来,开始逐渐应用在工业过程的故障定位与故障路径识别中。其中,文献[8]使用互相关函数法对带时滞的变量进行因果关系假设检验,但该方法仅适用于线性系统。针对非线性过程,文献[9]提出的传递熵(transfer entropy)因果分析方法对变量间传递的信息熵进行因果关系假设检验,但该方法易受噪声影响,且计算量大。对此文献[10]提出一种符号化传递熵(symbolic transfer entropy, STE)的因果分析方法,相比于传递熵中的核概率密度估计,该方法对序列进行符号化来得到变量的概率分布,从而大大减小了计算量,但该方法需要优化选择的参数较多,并且样本量较少时会影响估计的概率分布的准确性,从而影响最终预测的结果。由于基于数据驱动的因果分析方法都需要对变量进行逐对分析,当过程变量数为 n,则需要求出 $n\times(n-1)$ 组变量之间的因果关系,然而实际工业过程故障发生后,在控制回路的作用下,收到故障影响的变量可能有限,这就会导致大

收稿日期:2019-08-15; **修回日期**:2019-08-26。
基金项目:国家自然科学基金项目(61573169)。

作者简介:潘 磊(1993-),男,安徽芜湖人,硕士研究生,主要从事工业过程监控与诊断方向的研究。
赵忠盖(1976-),男,湖北荆州人,博士,教授,主要从事间歇过程建模与估计、工业过程监控与诊断方向的研究。
刘 飞(1965-),男,江苏无锡人,博士,教授,主要从事先进控制、工业系统监控与诊断、间歇过程智能装备与系统方向的研究。

量无用的计算，进而影响故障定位的效率。对此文献［7］提出一种利用重构贡献筛选出贡献率大的故障候选变量集，再针对平稳、非平稳变量序列，分别使用格兰杰因果（granger causality，GC）和动态时间规整（dynamic time warping，DTW）对筛选出来的变量集进行因果分析。与之相似，文献［11］提出使用 lasso 重构筛选故障候选变量集，再使用高斯回归进行故障因果分析。虽然这二种方法缩小了故障定位范围降低了计算量，但变量筛选是否可能会遗漏掉故障变量并不可知。文献［12］为降低计算量提出一种计算过程变量对于统计量的动态归一符号化传递熵（symbolic dynamic－based normalized transfer entropy，SDNTE）的故障定位方法，该方法仅需求得 n 组过程变量对于统计量的 STE 因果关系系数，从而显著缩小了故障定位所需花费的时间，但该方法使用的是 STE 方法来度量变量间的因果关系，存在前文所提的优化选择的参数过多、估计概率分布时对样本需求量大等限制。因此，在 SDNTE 方法的框架下，如何使用一种更加简洁并能适应于小样本集的方法对因果关系系数进行度量是一个值得研究的问题。

综上，本文在 SDNTE 方法的思路下提出一种基于 PCA 与随机森林回归算法（random forest regression，RFR）[13] 相结合的 PCA－RFR 故障定位新方法，通过利用 RFR 的变量重要性度量得到过程变量对统计量的因果关系系数，辨识其中值最大的变量作为故障变量。相比于 SDNTE 方法，PCA－RFR 无需优化选择参数，并且可以对小样本集建立良好的模型。最后通过一个数值仿真，并在 TE 过程仿真实验中将本文提出的方法与 RBC、GC[7] 和 SDNTE 方法的定位效果进行了对比，表明该方法定位效果的优越性。

1　基本理论

1.1　基于 PCA 的统计过程监控方法

PCA 统计过程监控方法是将数据投影到二个正交的主元空间和残差空间上，并分别构建相应的检测统计量来进行监控过程运行状况的一种方法。

假设正常工况下的样本集为 $\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$，$n$ 为样本数，m 为变量数。标准化处理后，使其均值为 0 标准差为 1。对协方差矩阵 $\boldsymbol{S}$ 奇异值分解得到：

$$\boldsymbol{S}\simeq\boldsymbol{X}^T\boldsymbol{X}/(n-1)=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T+\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{P}} \tag{1}$$

其中：$\boldsymbol{P}\in\boldsymbol{R}^{m\times l}$，$\widetilde{\boldsymbol{P}}\in\boldsymbol{R}^{m\times(m-l)}$ 分别为主元和残差载荷向量，l 为主元个数，$\boldsymbol{\Lambda}$、$\widetilde{\boldsymbol{\Lambda}}$ 分别为主元和残差特征值组成的对角阵。

任意一个样本可分解为：

$$x=\boldsymbol{P}\boldsymbol{P}^Tx+\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{P}}^Tx=\boldsymbol{C}x+\widetilde{\boldsymbol{C}}x=\hat{x}+\tilde{x} \tag{2}$$

式中，$\boldsymbol{C}$ 和 $\widetilde{\boldsymbol{C}}$ 分别代表主元和残差空间的投影矩阵。

PCA 故障检测统计量指标包括 T^2、SPE 以及 φ 统计量，其中 φ 统计量作为 T^2 和 SPE 统计量的合成指标，使用起来更加方便简单[14]。

SPE 统计量：

$$SPE(x)=\|\tilde{x}\|^2=x^T\widetilde{C}x \tag{3}$$

T^2 统计量：

$$T^2(x)=x^T\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^Tx=x^T\boldsymbol{D}x \tag{4}$$

φ 统计量：

$$\varphi(x)=\frac{SPE(x)}{\delta_\alpha^2}+\frac{T^2(x)}{T_\alpha^2} \tag{5}$$

式中，δ_α^2 和 T_α^2 为 SPE、T^2 的控制限，包括统计量 φ 在内，这 3 种统计量控制限皆可由其抽样分布得到[15]。若统计量超过了相应的控制限，则认为过程发生了异常，从而实现故障的检测。

1.2　RFR 算法

随机森林（random forest，RF）算法是一种由很多学习器组成的集成学习算法，它通过数据的随机重采样（bootstrap）和结点随机分裂技术的应用来降低决策树之间的相关性，进而提高模型的预测性能。RF 常用于分类和回归，基础学习器使用的是分类回归树（classification and regression tree，CART），它是一种结构为二叉树的决策树。

对于 CART 回归树的构建，假设 x 与 y 分别为输入和输出，并且是连续变量。在 x 所在的空间中，每个输入特征空间被递归的划分为二个子区域，并令每个子区域样本对应的 y 的均值作为输出值，使用平方误差最小化准则进行特征选择，构建二叉回归决策树。

RFR 算法由多个 CART 回归树集成而成，具体建模过程如图 1 所示。利用 bootstrap 重采样出 b 组训练样本集 $x_i(i=1,2,\cdots,b)$ 和相应的袋外数据（out－of－bag，OOB）集 $oob_i(i=1,2,\cdots,b)$，由于是等量随机重采样，其中每组训练样本集中会随机抽取到原始样本中约 63% 的样本，原始样本中剩余的 37% 样本即为 OOB 数据。OOB 数据被用来进行模型测试以及变量重要性度量。将每组训练样本使用结点随机分裂技术生成 CART 回归树 $h_i(i=1,2,\cdots,b)$，并将生成的 b 颗决策树组成随机森林回归模型 $f=\{h_1,h_2,\cdots,h_b\}$。当对待检测样本进行预测时，将 b 颗决策树预测值的均值作为最终预测结果。

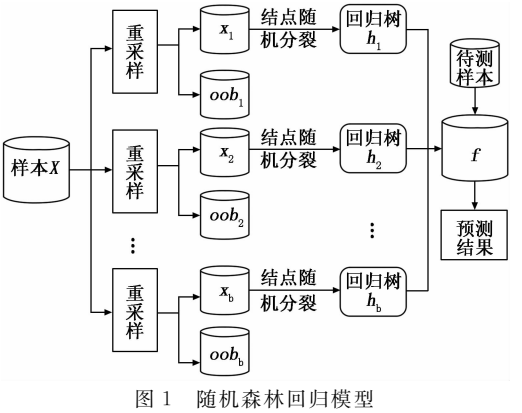

图 1　随机森林回归模型

2 基于 PCA－RFR 的故障定位方法

2.1 变量重要性度量测量因果关系系数

由于 PCA 模型下的统计量是对所有过程变量在相应空间变化信息的度量，从本质上说，所有过程变量对于统计量都存在因果关系。在预测模型下，通过去除变量 x 的作用，判断其对于预测输出变量 y 的影响程度即可得到该变量对于 y 的因果关系系数。

当对训练数据建立好随机森林回归模型后，变量重要性度量作为随机森林回归模型的一个属性，通过对 OOB 数据中输入变量 x 随机置换后（消除变量 x 信息影响）对于输出变量 y 预测精度的降低程度来衡量该输入变量对于输出变量的重要性。本文这里将其借以利用，将其作为输入变量对于输出变量的因果关系大小的度量。

对于变量重要性度量的计算主要分为以下 4 个部分：

1）对已建好的 RFR 模型 $f=\{h_1,h_2,\ldots,h_b\}$，将 $oob_i(i=1,2,\ldots,b)$ 数据带入相应的决策树进行预测，得到均方误差 $MSE_i(i=1,2,..,b)$，均方误差的定义为：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{6}$$

式中，y 和 $\hat{y}$ 分别为真实值和预测值，n 为样本数。

2）将 oob 数据中变量 $x_j(j=1,2,\ldots,m)$ 随机置换，带入步骤 1 得到均方误差 MSE_i^j，其中 m 是变量数。

3）计算决策树 $h_i(j=1,2,\ldots,b)$ 在变量 x_j 随机置换前后的均方误差的差：

$$VI_i^j=MSE_i-MSE_i^j \tag{7}$$

4）变量 x_j 的变量重要性度量值：

$$VI^j=\frac{1}{b}\sum_{i=1}^{b}VI_i^j \tag{8}$$

SDNTE 方法在计算变量间因果关系系数时需要对符号数、窗口大小等参数进行优化选择，同时需求大量样本建立概率统计模型，文献［12］中使用多达 72 000 组样本。相比较而言，变量重要性度量的计算就要简单容易的多，无需优化选择参数，随机森林回归模型一旦建好，即可得到变量重要性度量的数值，同时也能应用在少量样本场合，如对几百个样本便可建立良好的模型。由于随机森林回归通过并行构建决策树，这使得模型的构建也非常快速。

2.2 基于 PCA－RFR 的故障定位流程

PCA 模型下的混合指标 φ 是对样本在主元和残差空间变化程度的度量，当发生故障时，φ 统计量会增加并超过控制限，通过判断过程变量对于统计量指标 φ 的因果关系系数大小就可以进一步辨识出故障变量。

本文首先通过 PCA 模型筛选出发生故障的数据段，再将故障数据段的过程变量作为输入，对应的 φ 统计量作为输出建立 RFR 模型，最后通过模型的变量重要性度量系数值，系数值越大则表明该变量越有可能是引起 φ 统计量变化的变量，因此可将其辨识为故障变量。

基于 PCA－RFR 的因果分析故障定位流程如图 2 所示。具体步骤如下：

1）对工业过程的正常历史样本数据建立 PCA 模型，并得到统计量 φ 的控制限。

2）结合建立好的 PCA 模型和 φ 统计量控制限对离线采集数据进行故障检测，筛选出故障数据段并得到与其对应的 φ 统计量。

3）建立故障数据段的过程变量与 φ 统计量的 RFR 模型。

4）通过模型得到过程变量的变量重要性度量，对于变量重要性度量值最大的变量辨识为故障变量。

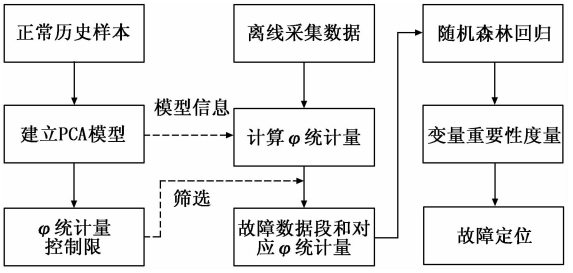

图 2 基于随机森林回归的故障定位方法流程

3 仿真

3.1 数值仿真

参照文献［6］的数值仿真案例，其系统结构如式（9）所示：

$$\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\x_5\\x_6\end{bmatrix}=\begin{bmatrix}-0.2310 & -0.0816 & -0.2662\\-0.3241 & 0.7055 & -0.2158\\-0.217 & -0.3056 & -0.5207\\-0.4089 & -0.3442 & -0.4501\\-0.6408 & 0.3102 & 0.2372\\-0.4655 & -0.433 & 0.5938\end{bmatrix}\begin{bmatrix}t_1\\t_2\\t_3\end{bmatrix}+noise \tag{9}$$

其中：t_1，t_2 和 t_3 为均值为 0，标准差分别为 1，0.8，0.6 的随机变量。噪声的均值为 0，标准差为 0.2。

通过对该仿真模型生成 1000 组正常样本并建立 PCA 模型。再通过对变量 x_3 添加一个均值为 0 标准差为 1 的随机故障，生成 1000 组故障样本，建立该 1000 组故障样本变量数据与其 φ 统计量的随机森林回归模型，得到 6 个输入变量的变量重要性度量如图 3 所示，由图可知，PCA－RFR 算法可以得到故障变量 3 的准确定位。

3.2 TE 过程仿真

Tenessee Eastman（TE）过程是一个由纳西－伊斯曼公司公开的基于实际化工生产过程仿真系统，TE 过程很好地模拟了实际复杂工业过程的主要特征，因此被广泛地应用于控制、优化、过程监控和故障诊断的研究中，其过程流程参见文献[17]。TE 过程共有 5 个操作单元组成：反应器、分离器、循环压缩机、汽提塔和冷凝器。包含 41 个测量变量和 12 个控制变量，其中测量变量又分为 22 个过程测

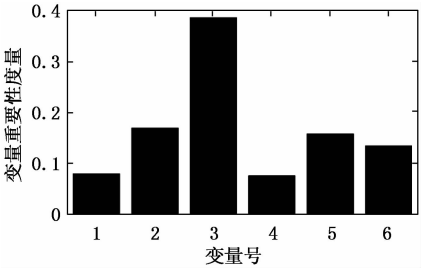

图 3　数值仿真变量重要性度量

量变量和 19 个成分测量变量。TE 过程有 21 种故障类型，包括阶跃、随机、缓慢漂移、阀粘滞等故障类型。

因为基于数据驱动的因果分析方法大多存在变量序列为平稳的限制，因此，本文采用 22 个连续过程测量变量进行研究，选择故障 8－12 共 5 种随机故障进行分析。正常和故障条件下的运行时间都为 960 个时刻，采样时间为 3 分钟，其中故障均从 160 时刻加入。

首先对正常数据建立 PCA 模型，选择故障条件下第 161 到 960 时刻共 800 个样本作为故障样本集，然后将故障样本集的过程变量作为输入，对应的 φ 统计量作为输出建立随机森林回归模型。随机森林回归算法在 Sklearn 机器学习库中决策树数量的默认参数为 10 颗，为了保证模型预测性能，这里将决策树数量参数选择为 100 颗。

故障 11 是反应器冷却水入口温度变化的随机故障，直接与之相关联的变量是变量 x_9（反应器温度）和变量 x_{21}（反应器冷却水出口温度）。基于 GC、RBC、SDNTE 以及 PCA－RFR 方法对故障 11 的故障定位结果分别如图 4 的（a）、（b）、（c）、（d）所示，其中 GC 因果分析方法首选通过筛选出故障候选变量集 x_3，x_4，x_5，x_6，x_8，x_9，x_{14}，x_{17}，再通过变量间的因果关系指向识别出故障变量 x_9，但图中其余的孤立变量，无法判断其中是否存在故障变量。其余三种方法虽然同样都实现了故障变量 x_9 的识别，但 PCA－RFR 对于变量 x_9 的定位效果要明显更加突出。

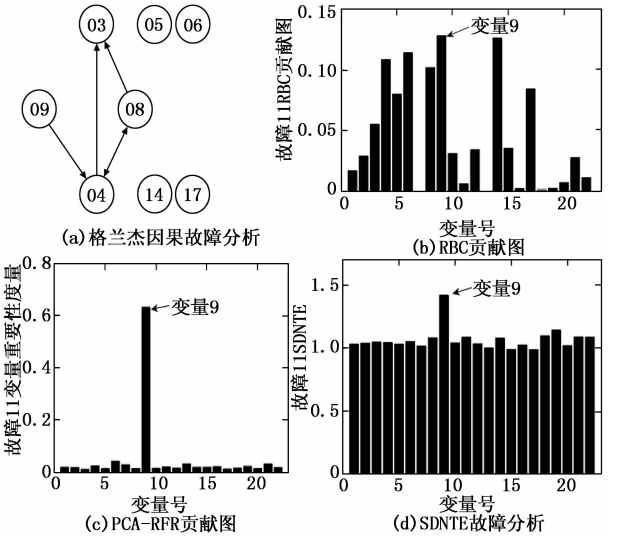

图 4　故障 11 的故障定位

故障 12 是冷凝器冷却水入口温度变化的随机故障，通过影响冷凝器对气体的冷凝效果直接影响到下游的变量 x_{13}（分离器压力变量）。基于 GC、RBC、SDNTE 以及 PCA－RFR 方法对故障 11 的故障定位结果分别如图 5 的（a）、（b）、（c）、（d）所示。其中 GC 方法定位出故障变量 x_{12}，筛选变量时遗漏了故障变量 x_{13}，基于 RBC 方法定位出故障变量 x_{15}，基于 SDNTE 方法定位出故障变量 x_{22}，而基于 PCA－RFR 方法定位出故障 x_{13}。因此，对于故障 12 仅有 PCA－RFR 实现了对故障变量的准确识别。

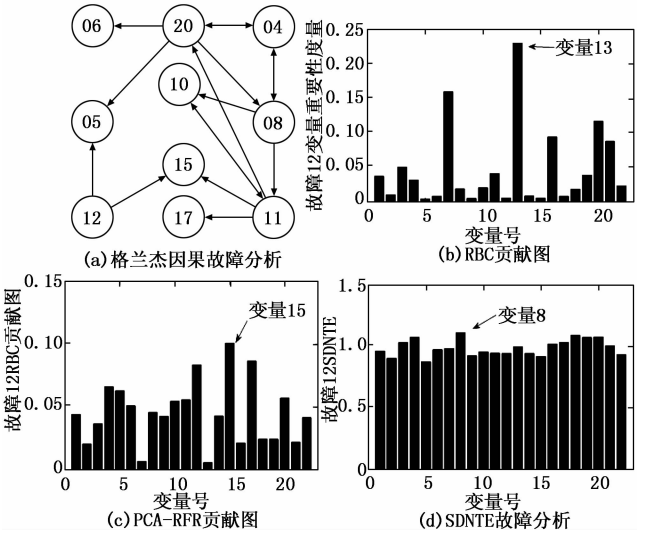

图 5　故障 12 的故障定位

对于 TE 过程的 5 种随机故障，基于 GC、RBC、SDNTE 和 PCA－RFR 方法故障定位结果如表 1 所示。通过对比可以看出仅有 PCA－RFR 实现了对所有故障变量的准确识别。其在这 5 组故障中的定位效果要明显优于其他方法，验证了该方法的有效性和优越性。

表 1　故障定位方法对比

序号	故障变量	PCA－RFR	RBC	SDNTE	GC
8	x_{16}，x_4	x_{16}	x_{20}	x_{11}	x_3 x_4
9	x_9，x_{21}	x_9，x_{21}	x_6	x_{19}	x_4，x_8
10	x_{18}	x_{18}	x_{18}	x_{20}	x_3，x_{18}
11	x_9，x_{21}	x_9	x_9	x_{21}	x_9
12	x_{13}	x_{13}	x_{15}	x_{22}	x_{12}，x_4

4　结束语

提出了一种基于 PCA－RFR 的故障定位新方法，该方法利用变量重要性度量来衡量故障数据中过程变量对于 φ 统计量的因果关系大小，认定其中值越大的变量越有可能是故障变量。通过仿真实验验证了 PCA－RFR 方法在故障定位中的有效性。但该方法目前仅应用于离线数据故障定位中，基于随机森林回归模型的在线故障定位还有待进一步的研究。

（下转第 40 页）