过程工业大数据建模研究展望

刘 强 [1, 2] 秦泗钊 [2, 3]

摘 要 人们对大数据的认识已从 “3Vs” (Volume — 大容量; Variety — 多样性; Velocity — 处理实时性)、“4Vs” (“3Vs” 与 Value — 价值)、到现今的 “5Vs” (“4Vs” 与 Veracity — 真实性). 在此背景下, 首先分析过程工业大数据的 “5Vs” 特性; 接下来, 综述现有数据建模方法, 并结合过程工业大数据特有性质 (包括: 多层面不规则采样性、多时空时间序列性、不真实数据混杂性) 论述现有数据建模方法应用于工业大数据建模时的局限; 最后, 探讨过程工业大数据建模有待研究的问题, 包括: 1) 多层面不规则采样数据的潜结构建模; 2) 用于事件发现、决策和因果分析的多时空时间序列数据建模; 3) 含有不真实数据的鲁棒建模; 4) 支持实时建模的大容量数据计算架构与方法.

关键词 过程工业大数据, 多层面数据潜结构建模, 多时空时间序列数据建模, 大数据计算架构

引用格式 刘强, 秦泗钊. 过程工业大数据建模研究展望. 自动化学报, 2016, **42**(2): 161−171

DOI 10.16383/j.aas.2016.c150510

Perspectives on Big Data Modeling of Process Industries

LIU Qiang[1, 2] QIN S. Joe[2, 3]

Abstract The understanding of big data goes through three stages, i.e., “3Vs” (Volume, variety and velocity), “4Vs” (“3Vs” and value), and “5Vs” (“4Vs” and veracity). In the era of big data of process industries, the “5Vs” characteristics of industrial big data are analyzed. After that, the existing methods on data modeling are reviewed while the corresponding limitations are analyzed under industrial big data circumstances with specific characteristics, i.e., multi-layer irregularly sampling, multiple temporal and spatial time series, and non-veracity with outlier. Finally, the perspectives on industrial big data modeling are discussed, including: i) latent structure modeling of multi-layer irregularly sampled big data; ii) multiple temporal and spatial time-series data modeling for event discovery, decision-making, and causality analysis; iii) robust modeling of data with non-veracity samples; and iv) data-friendly system architecture and method towards big data real-time modeling.

Key words Process industrial big data, multi-layer data latent structure modeling, multiple temporal and spatial time-series data modeling, big data computing framework

Citation Liu Qiang, Qin S. Joe. Perspectives on big data modeling of process industries. *Acta Automatica Sinica*, 2016, **42**(2): 161−171

1 过程工业大数据的发展

1.1 大数据发展及其价值

随着传感器技术、计算机技术、通信技术、物联网、数据存储等技术的发展, 互联网、过程工业等行业产生并存储了大容量数据, 且随时间指数级增长[1−2], 预计到 2027 年每年将产生 1 YB (10^{24} Bytes)[3]. 这些数据不仅容量大, 各行业对其的认识已由 “3Vs” (Volume — 大容量, Variety — 多样性, Velocity — 处理实时性)、“4Vs” (3Vs 与 Value — 价值)、到现今的 “5Vs” (4Vs 与 Veracity — 真实). 麦肯锡咨询公司将 “大数据” 定义为超出传统数据库软件工具抓取、存储、管理和分析能力的数据群[4]. “大数据” 一词 1970 年首次出现在有关大气与海洋环境的文章中[5], 21 世纪以后由计算机、工程和数学科学等引领相关研究, 但两者之间有本质

收稿日期 2015-08-13 录用日期 2015-10-23
Manuscript received August 13, 2015; accepted October 23, 2015
国家自然科学基金 (61304107, 61490704, 61573022, 61290323, 61203102), 中国博士后科学基金 (2013M541242), 博士后国际交流计划派出项目 (20130020), 中央高校基本科研业务费 (N130408002, N130108001) 资助
Supported by National Natural Science Foundation of China (61304107, 61490704, 61573022, 61290323, 61203102), the China Postdoctoral Science Foundation (2013M541242), the International Postdoctoral Exchange Fellowship Program (20130020) and the Fundamental Research Funds for the Central Universities (N130408002, N130108001)
本文责任编委 姜斌
Recommended by Associate Editor JIANG Bin
1. 东北大学流程工业综合自动化国家重点实验室 沈阳 110819 中国 2. 美国南加州大学化工系 洛杉矶 90089 美国 3. 香港中文大学 (深圳) 深圳 518172 中国
1. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China 2. Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, CA 90089, USA 3. School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China

上的不同.

大数据中蕴含大价值, 根据麦肯锡公司的研究, 大数据创造价值有 5 种方式: 1) 通过提高信息透明化与可用度开启价值; 2) 采集更细节信息来揭示变化与提升性能; 3) 个性化产品与服务; 4) 通过复杂的分析改进决策; 5) 改进下一代产品和服务的研发. 大数据价值促使人们重新审视统计学、计算机科学、工程学等传统的数据分析工具, 通过数据建模对大容量数据进行分析来获取知识, 例如: 1) 互联网领域, 2008 年谷歌公司研发谷歌趋势系统, 利用 50 000 000 次流感关键词 (如温度计、流感症状、胸闷) 的互联网搜索实时数据, 以及美国疾控中心流感传播历史数据, 通过 15 000 000 模型发现 45 个特征来建立流感预测模型, 近实时地预测 2009 年美国流感 H1N1 的爆发[6]; 2) 经济活动领域, Preis 等利用谷歌趋势系统对 98 个经济词汇的搜索数据建模, 发现搜索量增加发生在金融大亏损之前, 分析出网络搜索行为和经济指数间的联系[7]; 3) 社会安全领域, PredPol 公司与警方以及加州大学合作, 利用洛杉矶地区 80 年的 130 万个犯罪纪录数据, 采用地震预测算法预测犯罪位置和时间段, 相应加大巡逻密度, 使 2011 年 11 月至 2012 年 4 月间盗窃罪和暴力犯罪分别下降了 33 % 和 21 %.

1.2 过程工业数据发展及其大数据特点

相比较互联网大数据近年从无到有的迅猛发展, 过程工业的数据基础更好, 上世纪 70 年代就可由计算机集散控制系统 (Decentralized control system, DCS) 采集用于过程控制与设备状态监控的设备及传感器数据. 随着信息化的发展, 过程工业不仅在时间上不断存储积累这些过程运行数据, 还在空间上扩展采集设备、人之间及内部传输的数据, 从而获得时间与空间两个维度上不同尺度的大容量数据, 以及分散于各生产部门的多源不同类型的文本、图像、声音等数据. 据麦肯锡咨询公司大数据报告统计, 过程工业的数据存储量高于其他行业 — 2010 年的数据存储量接近 2 EB[8], 而且增速是其他大数据领域的两倍[9].

过程工业大数据蕴含大价值, 麦肯锡全球研究院发布的《Big data: the next frontier for innovation, competition, and productivity》中已指出过程工业可以从大数据分析和应用中提高生产力、降低消耗. 以工业大数据为价值源, 到 2020 年的总体价值将近 1.3 万亿美元[9]. 过程工业大数据价值产生方式主要是通过集成设计与运行时的生产数据、采购的原料数据及销售过程中积累的点击流和用户行为数据等, 更好地决策来改进过程运行、提高生产效率、提高产品质量、减少缺陷产品、满足用户需求. 一方面, 可以减少 20 %～50 % 的产品开发时间[10]; 另一方面, 基于大数据主动预测[8], 实现快速分析及执行、降低错误决策的后果[11]. 目前, 已有一些初步的应用, 比如: 1) Vestas 风力发电机制造公司对天气数据与其涡轮仪表多时空大数据交叉分析, 改进发电机的布局并进行起、停、改变迎风角等运行决策, 提高发电机布局的效率、增加电力输出和延长寿命[12]; 2) 通用公司在亚特兰大建立能源监测和诊断中心, 采集全球上千台燃气轮机数据、振动、温度等 10 年的数据, 通过比较历史数据和实时数据, 监测燃气轮机异常运行趋势, 预测燃气轮机故障, 提前检修与维护, 年节约 0.75 亿美元[8].

过程工业大数据的 “5Vs” 特性体现在:

1) 大容量 (Volume) 体现在容量大的相对性. 据 Robert Hilliard 文章 “It′s time for a new definition of big data”, 过程工业产生 GB 级别数据, 比如 10 万个传感器、每个传感器每秒产生 8 Bytes 数据, 每小时产生近 3 GB 的数据. 相比较社会网络大数据而言, 容量并不大, 但由于采样率高, 采样时间段长, 信息密度大, 仍可称大数据[13].

2) 多样性 (Variety) 体现在多层面、多类别, 以及不规则采样. 过程工业分层次运行, 采集的时间序列数据既有高维且快速率动态采样的过程数据, 又有不规则采样的指标数据. 数据的多层面特性, 数据存储形式包括图像、文本, 以及时间序列数据不规则采样性, 体现了过程工业特有的多样性数据特征.

3) 速度 (Velocity) 体现在实时建模的需求. 过程运行工况及质量指标实时控制与优化, 要求实时数据处理, 对建模实时性、模型在线更新提出要求.

4) 价值 (Value) 体现在通过数据建模实现价值. 时间尺度的时间序列建模, 空间尺度的潜结构建模都属于实现数据价值的建模范畴.

5) 真实性 (Veracity) 体现在鲁棒建模的需求. 实际工业过程, 受测量仪表或变送器等故障以及异常干扰的影响, 导致测量数据中混杂不真实数据, 具有离群点、缺失点等异常样本. 比如, 赤铁矿磨矿过程由于矿石的 “磁团聚” 特性导致磨矿粒度实际测量值出现大偏差, 导致建模样本中出现离群点.

尽管过程工业数据丰富, 但根据 Garterner 公司的分析, 由于缺少有效的分析工具以及高效的计算技术来提取有用信息, 工业大数据还未充分利用[14]. 目前主要是将数据压缩、短时间段的数据存档, 仅在特殊运行状况下进行数据恢复与分析, 而不是像亚马逊、谷歌等将历史数据视为资产用于常规的决策过程中. 过程工业大数据面临的挑战在于挖掘历史大数据中蕴含的知识. 但知识不是直接呈现在数据里, 而是呈现于用于揭示数据的模型. 为此, 本文集中考虑通过过程工业大数据的建模实现大数

据价值.

2 数据建模与大数据建模研究动态

2.1 数据建模方法

过程工业物料变化频繁、控制复杂、多级运行, 难以采用传统机理建模方法. 以潜结构建模为代表的数据建模方法因其具有降维、便于可视化的优点得到重视, 相关研究可追溯到二十多年前, 已在化工、薄膜制造、半导体、钢铁等行业中成功应用, 同时成为学术界的研究热点[15].

潜结构建模方法强调数据间的潜结构, 有别于传统的输入输出模型和因果模型, 现有方法利用过程正常运行数据建立潜结构模型, 在此基础上定义故障检测指标及其控制限以进行故障检测与诊断. 可以在产品质量控制与用户反馈发生前实现故障的检测与故障原因诊断, 并可诊断出会导致严重故障与不安全事故的异常工况. 相关建模方法分为三类: 1) 潜结构建模的基本方法, 包括无监督的主元分析 (Principal component analysis, PCA)、独立元分析 (Independent component analysis, ICA)[16−17], 以及有监督的偏最小二乘 (Partial least squares, PLS) 等; 2) 多模态及非线性过程数据潜结构建模方法; 3) 过程数据强自相关互相关的动态过程潜结构建模方法[18].

2.1.1 潜结构建模基本方法

潜结构建模策略中, 单一层面的无监督潜结构模型包括主元分析和独立元分析模型. 其中, 主元分析方法以方差最大为目标对高维数据进行潜结构分解, 提取描述过程主要变化的潜变量; 独立元分析方法在假设潜在变量非高斯分布的前提下, 将观测数据分解为统计上独立元的线性组合, 提取测量值不服从高斯分布的部分. 二者的主要区别是 PCA 以提取方差最大的潜变量为目标; ICA 以提取非高斯潜变量为目标, 更适于过程数据不满足高斯分布的情况[16−17].

PCA 和 ICA 属于单一层次无监督建模方法, 只能描述过程数据间的相关关系, 无法建立过程数据与产品质量间的关系. 针对该问题, 将数据划分为过程数据空间与指标数据空间, 采用 PLS 这类多层面有监督建模方法描述过程数据与产品质量数据间的关系, 通过过程数据矩阵与指标数据矩阵相互交换分解信息, 找到输入空间到输出空间的预测能力意义上的最优特征方向, 提取与指标有关的潜变量.

2.1.2 多模态与非线性过程潜结构建模方法

实际过程运行时, 由于运行任务与设定值的变化、外界环境的改变、设备重组等, 会导致正常工况发生改变, 具有多模态特性. 对于这类经常发生正常工况切换的过程, 学者提出多模态建模方法, 有多模型建模、局部学习. 比如, Singhal 等提出基于 PCA 相似因子的模式匹配方法[19]; Yoo 等为批次式反应器设计了一种多模态建模方法[20]; Kano 等提出外部分析与 ICA 相结合的建模方法, 将过程变量分为外部变量和主变量, 在此基础上, 将主变量分解为去除外部变量影响的部分和与外部变量有关的部分, 使之适应过程操作条件的改变[21].

对于变量间具有强非线性关系的过程, 文献 [22] 提出 KPLS (Kernel PLS) 方法, 通过核运算将非线性过程数据投影到高维的特征空间, 再通过建立质量数据与高维特征空间的线性 PLS 模型来描述原始过程数据与质量数据间的非线性关系; 文献 [23] 发展了 KPLS 方法, 提出 CKPLS (Concurrent KPLS) 方法, 将过程与质量空间划分为共有子空间、质量特有子空间和过程特有子空间, 对过程变化与质量变化具有更强的解释性; 上述 KPLS 与 CKPLS 方法可通过核运算来描述变量间的非线性关系, 相比较线性 PLS 方法而言建模精度更高, 比如将其用于过程监控与故障诊断时会降低线性 PLS 方法的漏报率与误报率, 但同时因核运算随样本数增加而指数级增大的特性会增加建模的计算负荷.

2.1.3 动态过程潜结构建模方法

当过程数据是强自相关或动态互相关的动态时间序列时, 传统的静态潜结构建模方法无法描述过程的动态性. 过去二十多年, 多位学者提出了动态潜结构建模方法. 比如, Ku 等提出动态主元分析建模方法[18], 利用时间窗构造变量的增广矩阵并进行奇异值分解来建立变量间的自相关和互相关关系. 但该方法的局限性是自相关与互相关混杂导致模型参数较多, 而且难以解释潜在的动态关系.

基于子空间建模的动态潜结构建模方法可避免上述问题, 比如, Negiz 等使用规范变量状态空间模型来描述动态过程, 其等价于向量自回归滑动平均时间序列模型. 使用规范变量分析 (Canonical variate analysis, CVA) 算法的随机实现来处理自相关、互相关、共线性的大规模过程建模[24]. 其所构建的状态变量为过去测量值的线性组合, 用来解释数据的未来变化. 文献 [25] 对 CVA 和 PLS 两种方法进行比较研究, 结果表明 CVA 可对故障提供更快速的检测, 但由于协方差阵中较小特征值的影响, PLS 通常可获得比 CVA 方法数值更稳定的结果. 文献 [26] 采用子空间辨识方法建立过程的子空间模型. 主元分析子空间辨识方法利用变量的一致性空间法建立状态空间模型[27]. Li 等建立了动态主元分析方法与子空间辨识方法间的关系[28]. 在有过程噪声和测量噪声的条件下, 提出一致动态主元分析算法, 即间接

动态主元分析建模方法. 最近, Ding 等将子空间辨识与基于模型的故障检测技术相结合, 提出基于观测器的故障检测方法[29]. 另一种称为线性高斯状态空间的动态概率模型是 Wen 等提出的[30], 用于动态过程监控, 使用最大期望算法进行卡尔曼滤波的结构选择. 为了同时提取变量间的动态自相关和互相关信息, 文献 [31] 提出含义更明确的动态潜变量模型.

然而, 上述模型均未考虑与质量动态相关的过程建模. 在对有监督的 PLS 建模方法进行动态扩展时, 因为 PLS 模型由内模型和外模型组成, 所以动态 PLS 建模有更多可能性. Kaspar 等提出偏最小二乘改进算法, 但只将动态项引入内模型, 而外模型沿用静态 PLS 模型, 形成内外不统一的动态偏最小二乘模型[32]. 针对该问题, 文献 [33] 提出内部与外部模型统一的内在动态 PLS 建模方法. Li 等提出改进的动态偏最小二乘算法, 采用动态全潜结构模型将过程数据空间分为四个子空间, 描述质量相关的变化[34]. 但动态全潜结构模型只能描述过程数据中可以预测的输出变化. 针对上述问题, Liu 等提出动态并发潜结构建模方法, 建立多层面并发潜结构建模与诊断框架分离出不可预测的输出变化[35].

2.2 大数据建模

2.2.1 互联网公司大数据建模

目前, 大数据建模的相关研究主要由谷歌、亚马逊等互联网公司引领, 与传统数据建模方法相比, 互联网公司大数据建模需处理的数据规模更大、数据的使用范围更广. 研究重点是以分布式存储和分布式处理为基础, 构建适合实时数据分析处理的存储结构和计算引擎, 包括大数据平台架构与大数据计算框架两部分.

1) 大数据平台架构方面: Dean 等根据谷歌文件系统和映射化简 (MapReduce) 思想创建的 Hadoop 是目前最流行的大数据处理平台[36]. 具有高可靠性、高扩展性、高效性、高可伸缩性等优点. 改进 Hadoop 性能及定制大数据处理应用为目前的研究热点, 包括: 对 Hadoop 平台性能的改进[37]、Hadoop 上构建数据仓库、Hadoop 和数据库连接[38]、高效查询处理[39]、索引构建和使用[40]、数据挖掘[41]、推荐系统等. Hadoop 相关的大数据处理工具还包括: MapR、Cloudera、Hortonworks、BigInsights 和 ASTERIX.

2) 大数据计算框架方面: 主流的计算框架分为流处理模式、批处理模式和混合模式三种. a) 流处理模式将数据视为流, 当新的数据到来立刻处理并返回结果, 典型框架包括: 实时计算框架 Storm 和 S4 等; b) MapReduce 是最具代表性的批处理框架, 致力于通过大规模廉价服务器集群实现大数据的并行处理. 近年研究包括应用领域扩展、实时性提升和易用性改进等, 如文献 [42] 提出在 MapReduce 上的增量式数据挖掘方法来缩短数据挖掘的时间; HOP[43] 在 MapReduce 处理过程中引入管道的概念, 使数据在各个任务间以管道的方式交互, 增加了任务的并发性, 提高了数据处理的实时性, 还有一些研究通过改进 MapReduce 模型迭代计算的效率来提高其实时性, 如 HaLoop[44]、iMapReduce[45]、iHaLoop[46] 和 PrIter[47]; c) 针对批处理模式实时性不强的不足, 有研究尝试将流处理和批处理模式融合, 主要思路是利用 MapReduce 模型实现流处理, 如文献 [48] 探讨了将 MapReduce 模型应用到流处理单遍分析应用时在架构上应当进行怎样的调整; 在此分析基础上, 文献 [49] 介绍了利用 MapReduce 实现的适用于单遍分析的可扩展平台; 流 MapReduce[50] 结合事件流处理的特点, 对 MapReduce 中的 Mapper 和 Reducer 进行重新定义, 增加了持续的、低延迟的数据处理能力.

随着待处理数据容量的指数级增加, 传统数据建模方法无法求解或求解过慢. 针对该问题, 学者研究高速、可实现的大数据建模算法[51]. 比如, 文献 [52] 提出并行处理的迭代聚类分析算法; 文献 [53] 提出基于 Hadoop 的 K 均值聚类分析算法; 文献 [54] 提出基于 MapReduce 的层次聚类分析算法; 文献 [55] 提出快速计算大数据均值、方差、协方差的分布式求解算法.

2.2.2 过程工业大数据建模

互联网公司的大数据技术强调非结构化数据的存储与管理, 以及基于 “大数据” 的查询、统计与决策的简单应用, 通常很简单的统计分析就可以揭示部分信息. 过程工业大数据建模是面向过程工业决策、优化、故障诊断、控制等应用, 解决相对复杂的数据建模问题, 通常需要更深入的数据建模方法. 目前, 相关研究还处于起步阶段, 主要关注 “大数据” 集成于现有控制优化系统带来的实时数据处理、复杂数据分析的难度, 特别是利用过程历史运行大数据来建模设计阶段未考虑的运行变化.

2.3 过程工业大数据的新特点对建模带来挑战

2.3.1 过程工业大数据的新特点

随着信息化发展和 DCS 的广泛采用, 现代过程工业向大规模、动态性、集成化发展, 多单元、多产品生产、动态运行, 采集到如图 1 所示的多层面多

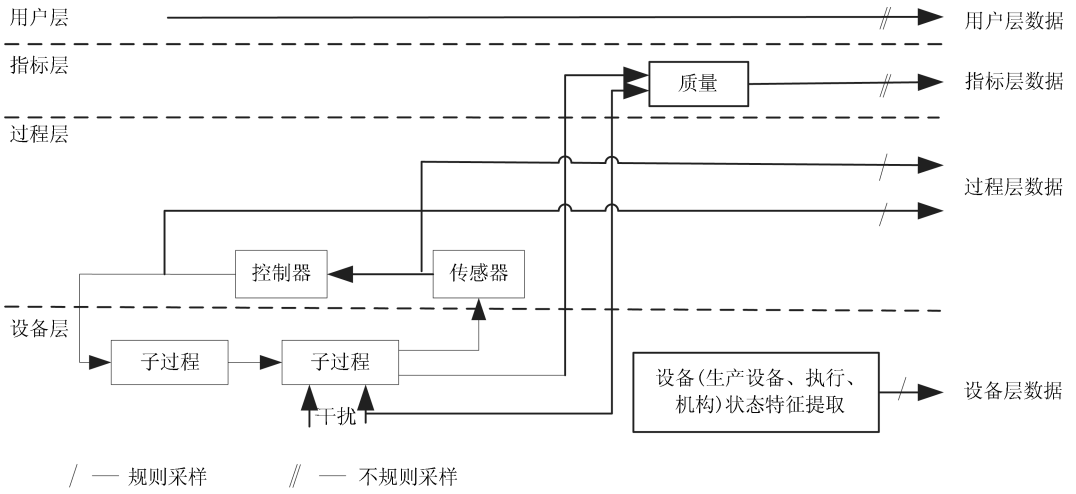


图 1　过程工业多层面、不规则采样时间序列数据

Fig. 1　Multi-layer irregularly sampling time-series data of process industries

时空时间序列的过程工业大数据[56]. 其中, 底层为设备层, 采集毫秒级的设备运行状况数据; 上一层为过程层, 采集规则采样的过程控制数据; 再上一层为指标层, 采集多种形式不规则采样的产品质量数据; 最上一层为用户层, 采集来自售后服务与社会网络媒体的用户反馈数据. 比如, 钢铁生产冷轧连续退火过程包括如下大数据: 1) 设备层辊子振动、轴承温度数据等; 2) 过程层以采样率 0.01 s 采集传动辊速度、电流、带钢张力数据等; 3) 指标层人工不规则采样带钢宽度、厚度数据, 视频监控带钢表面粗糙度图像数据等; 4) 用户层反馈用户对产品质量的定量与定性评价数据.

过程工业大数据数据采样率高、信息密度大; 相比较互联网大数据强调数据的非结构化, 过程工业大数据强调数据的多层面不规则采样性、多时空时间序列性、以及不真实数据混杂性. 具体分析如下.

1) 多层面不规则采样性: 过程工业既有高维动态的过程数据, 又有不规则采样的指标数据, 比如磨矿粒度指标、竖炉焙烧过程磁选管回收率往往难以在线测量, 只能通过人工化验获得, 具有大延迟和不规则采样的特点.

2) 多时空时间序列性: 随着过程工业采样变量规模的增大、数据采样率的增大、历史数据采样时间段的增大, 采集存储到更大容量的多时空大数据. 过程数据与指标数据不仅空间上具有相关性, 由于储能环境的存在、动态运行操作、成批次运行使得过程变量具有强自相关与互相关关系, 时间尺度上数据也呈稀疏分布, 且具有时间序列相关关系.

3) 不真实数据混杂性: 数据采集、传输、存储过程中的异常以及传感器自身漂移, 过程层的传感器设备故障、指标层人为误读数等会造成采集的高维动态数据中混杂离群点、缺失点等不真实数据.

2.3.2　过程工业大数据特点对传统数据建模方法带来的挑战

1) 针对多层面不规则采样性, 现有数据建模方法多集中在对规则采样数据的建模与分析, 无法对不规则采样的数据进行建模与分析.

2) 针对多时空时间序列性, 高维动态的过程数据是带有强相关的时间序列, 需要动态数据潜结构分析方法. 但是, 目前数据建模方法主要是提取空间潜结构的静态数据统计分析, 未考虑时间动态性, 不能从历史动态大数据中提取运行工况特征信息.

3) 针对不真实数据混杂性, 现有数据建模方法需要无污染的数据或对预处理后的数据建模. 否则, 模型参数受少数异常点影响大, 模型失配时有发生[57].

2.3.3　计算机学科大数据最新进展带来的机遇

计算机学科的机器学习、数据挖掘领域大数据研究的最新进展为解决上述问题带来机遇:

1) 相对于过程工业现有数据建模方法强调处理规则采样数据, 而将不规则及间接采样数据闲置不用, 计算机行业已可从高度非结构化数据中提取有价值的信息[56];

2) 相对于过程工业仅发表了少量时间序列趋势分析的研究论文, 数据挖掘领域已开发了完整的趋势分析技术来处理时间序列建模问题[56];

3) 相对于过程工业现有数据建模方法需要无污染的数据来建模, 数据挖掘与机器学习领域认为存在不真实数据是不可避免的, 但可利用大容量数据提取对不真实数据鲁棒的模型[56].

过程工业大数据建模与传统过程工业数据建模相比, 有如下异同:

1) 二者都是面向工业过程决策、优化、控制、故

障诊断的实际应用需求;

2) 二者针对的数据特点不同, 传统过程工业数据建模针对的是小变量规模、短时间段的规则采样数据; 而过程工业大数据建模针对的是更大范围时空尺度的不规则采样时间序列数据, 且其中混杂了不真实数据信息;

3) 二者解决途径不同, 传统过程工业数据建模方法集中于自动化学科, 主要采用多元统计建模[58]、系统辨识方法[59−60] 来建模规则采样的数据, 方法可扩展性差, 在变量或样本数过大时往往难以使用, 且建模前需要对建模数据进行离群点预处理等; 大数据建模方法多来源于计算机学科, 强调从大数据中挖掘知识, 可利用统计机器学习、数据挖掘算法建模不规则采样的多时空时间序列大数据.

为此, 需要结合计算机学科统计机器学习领域的最新进展将两大类方法有效融合, 研究过程工业大数据建模的新问题.

3 过程工业大数据建模最新进展与研究展望

3.1 多层面潜结构建模

过程工业大数据建模的主要方向是利用大数据建模技术扩展用于建模的数据类别. 传统方法主要是以 PCA 算法为代表的单一层次潜结构建模. 然而, 实际的工业过程大多分层次运行 (分为用户层、指标层、过程层和设备层), PCA 建模无法描述层间的潜结构关系, 从而无法分析过程变化对产品质量的影响. 指标不规则采样数据具有大采样迟延, 通常与高维过程数据具有动态强关联和潜结构. 通过建立二者间的动态潜结构模型, 就能够预测指标状况, 挖掘数据中潜在的运行结构信息, 用于决策、优化、控制与故障诊断.

Wold 提出将数据划分为过程数据空间与指标数据空间两层结构, 采用 PLS 模型将过程数据空间进一步划分为主元子空间和残差子空间来描述过程与指标的多层面潜结构建模方案. 然而, PLS 分解的主要问题在于: PLS 主元子空间中包含和输出变量无关或正交的信息[61], 主元子空间中可能包含与输出无关的变化[62]. 另一个问题是 PLS 残差子空间并不是最小化方差的子空间. 由于 PLS 目标是最大化过程数据和指标数据之间的协方差, 不是按主元方差降序排列潜变量, 排在后面的潜变量可能比之前的潜变量的方差更大. 近期, 文献 [63] 进一步分解残差子空间, 提出了全潜映射结构方法来解决这些问题. 但全潜映射结构方法只能描述过程数据中可以预测的输出变化. 为此, 文献 [64] 提出 CPLS (Concurrent PLS) 方法建立多层面并发潜结构建模框架, 以分离层间共有变化、过程层内特有变化和指标层内特有变化, 并采用该方法为过程监控提供了一个完整的输出变量监控和简洁的输入数据空间监控. 在该理论框架内, 对于分层子空间内的特有变化, 文献 [65] 定义各子空间内的多块和子块贡献, 提出了多级并发潜结构建模方法. 此外, 近年统计机器学习领域深度学习理论建模数据内在层次关系[66], 采用多层的方式用较少的参数来表示复杂的函数模型, 获得数据更抽象的特征表达, 提取数据中蕴含的层次关系.

上述多层面潜结构建模的现有成果, 解决的是规则采样的过程数据和指标数据的建模问题. 然而, 实际过程的指标数据具有大延迟和不规则采样. 指标数据大延迟与不规则采样情况下的多层面数据潜结构建模是新的研究方向. 可以通过对层间数据的预处理与再采样建立层间的潜结构模型, 相关研究内容是多层次间共有变化的潜结构建模和层内特有变化的潜结构建模[56]. 1) 层间共有变化建模: 依据慢尺度指标数据的采样时刻, 利用动态时间窗口对高速率过程数据动态采样, 构建动态关联的层间数据对, 在此基础上利用层间潜结构投影算法建立层间共有的潜结构模型, 其中, 动态时间窗口长度选择以及层间潜结构投影目标为有待研究的问题; 2) 层内特有变化建模: 利用原始过程动态数据和层间共有潜结构模型, 将层间共有变化剔除产生过程快尺度的特有变化数据, 再利用动态多维时间序列建模理论为基础建立过程层内快尺度潜结构模型和指标层慢尺度潜结构模型, 其中, 确定最大限度获取动态潜结构信息的建模目标为有待研究的问题.

上述多层面潜结构建模实现过程中, 如果建模变量数或样本数过大, 可导致算法无法实现或计算时间过长. 从该角度的研究方向是结合计算机学科大数据建模的并行 (或称分布式)、在线 (或称迭代) 计算框架与方法, 研究多层面潜结构建模的分布式在线实现方法[67], 比如研究加速文献 [33] 内在动态 PLS 的实现方法.

3.2 多时空时间序列数据建模

过程工业运行可采集并存储高维动态的多时空时间序列数据, 高维与动态是多时空时间序列数据的两大特征. 目前, 针对高维数据建模的主要解决方案是潜结构建模, 以上述多层面潜结构建模为研究热点. 针对动态数据建模的研究热点是时间序列建模, 目前最常用的是拟合平稳序列的模型, 可细分为自回归模型、滑动平均模型和自回归滑动平均模型.

对于过程工业采集的多时空时间序列数据, 现有的数据建模方法使用的是定义好的变量和选定的短时间段数据样本, 并未有效利用多源传感信息与历史运行数据. 为有效利用过程工业运行历史大数

据, 提取时间尺度更大、空间范围更广的知识, 实现分类、聚类、异常检测等. 比如, 提取历史运行模式特征来进行类似运行模式的快速检索. 有待研究的新建模问题包括: 1) 针对实时运行工况, 利用对历史数据的高效检索实现过程运行特征提取与相似度建模, 并实现快速匹配识别, 解决时间大范围多尺度的大数据建模问题; 2) 为实现多单元同类设备部件或系统特征的建模, 解决空间大范围多尺度的大数据建模问题.

尽管建模数据具有多时空来源, 一般而言多数过程数据仍是规则采样的时间序列数据. 为此, 可利用计算机领域广泛研究的时间序列数据挖掘与表示方法[68] 实现过程工业大数据压缩、快速匹配, 多时空时间序列数据建模的任务包括数据表示与索引、相似度建模、时间序列数据检索与匹配[68−69], 其核心任务是分段, 即获取支持索引、聚类和分类的时间序列简化表示. 这方面有待研究的关键问题包括:

1) 时间序列表示与索引方面: 为描述多变量时间序列数据特征, 采用多变量潜结构建模理论提取潜变量; 在此基础上, 建立时间序列索引机制来实现时间序列大数据压缩, 比如采用文献 [69] 中提出采用自适应分段总近似技术实现时间序列索引.

2) 时间序列相似度建模方面: 利用特定运行工况时段数据以及更长时间段的历史故障数据集, 实现运行模态的时间序列相似度建模, 比如文献 [69] 中介绍动态时间扭曲技术实现时间序列相似度建模.

3) 基于相似度的历史运行模态高效检索与匹配: 结合历史运行知识建立历史运行模态的分类标签, 为实现运行模态的实时 (或近实时) 识别, 进行历史运行模态大数据的高效检索与快速匹配识别, 文献 [70] 中介绍基于在线分段线性表示的序列匹配技术来实现时间序列匹配.

时间大范围的大数据建模局限于单一或局部变量, 而空间大范围多时间尺度的大数据建模则面临上千倍的大数据检索容量, 这一方面的研究主要以动态时间规整 (Dynamic time warping, DTW) 替代传统的欧式距离搜索算法[71]. DTW 已证明为时间序列的最好度量, 但其计算复杂度过大. 虽然 Jegou 等的成果具有大容量数据特征搜索的能力[71], 但该类工作是通过近似搜索实现的. 文献 [71] 归纳了大数据多时空高维时间序列建模的精确搜索问题, 传统的搜索优化方法包括: 1) 在计算候选时间序列距离时, 使用 DTW 下限距离 LBkeogh, 而不是计算样本向量间的完整距离; 2) 在计算距离 (如欧式距离, LBkeogh 距离) 过程中, 采用早放弃原则, 即在子序列而非完整序列间的距离超过已知最短距离时就终止后续计算; 3) 利用计算机的多核计算对传统方法线性加速. 除上述方法外, 文献 [71] 中还提出时间序列大数据建模的新方案, 主要包括: 1) 建模时简化归一化和在线归一化, 而不是处理归一化后的数据; 2) 使用早放弃原则时, 重新排序, 重点搜索距均值更远的时间序列.

为实现时间序列数据的高效检索与快速匹配识别, 新研究方向还包括结合流 MapReduce 与大数据框架, 在线分段线性表示的分布式实现. 此外, 在时间序列建模的基础上, 为描述时间序列发生的因果关系, 可采用 Granger 因果分析提取时间序列内部因果关系. 通过建立因果与预测的直接关系及统计假设测试来确定一个时间序列是否可用于预测另一时间序列, 比如用于识别全厂异常的根本原因[72].

3.3 含有不真实数据的鲁棒建模

针对离群点、缺失点等因数据真实性会影响建模精度与可用性的问题, 主要有建模前的数据预处理以及鲁棒建模两种处理方式.

数据预处理是在建模前获得用于建模的数据, 再采用常规数据建模方法. 文献 [73] 分别探讨了针对离群点和缺失点的数据预处理方法. 主要是替代法和删除异常样本点两种处理方式. 然而, 过程工业数据预处理所需要的过程知识在建模前往往难以准确获得; 再者, 数据预处理算法多为批处理算法, 当测量数据累积时, 算法的计算负荷逐渐加大.

鲁棒建模可以有效克服上述缺点, 已有一些初步的研究成果. 主要有样本加权方法和鲁棒学习方法. 比如, 针对显著离群点问题, 文献 [74] 使用样本加权建立鲁棒的支持向量机数据建模方法; 针对更普遍存在的非显著离群点的数据建模问题, 文献 [75] 改进现有的鲁棒学习算法, 基于鲁棒 3σ 原则更新鲁棒代价函数端点, 抑制离群点的不良影响, 提高模型的鲁棒性. 统计机器学习领域的低阶矩阵近似为鲁棒建模提供了新的发展方向, 低阶矩阵近似是将数据表示成一个低阶矩阵与稀疏矩阵之和, 将不真实数据项纳入稀疏矩阵, 而以低阶矩阵建模数据潜结构. 比如, Mackey 等提出分而治之的框架[76], Candés 等采用该框架来解决鲁棒 PCA 问题[77], 其通过使用零范数或 1-范数替代传统 PCA 方法的 2-范数使其对离群点和缺失点不敏感.

然而, 过程工业数据虽强相关但不低秩, 与矩阵压缩理论中的假设不符, 使用传统低阶矩阵近似方法建模, 找到的低秩数据还包含了特征值小的变化, 影响了主方向的选择和稀疏阵的分离, 造成建模结果不准确. 改进的思路是以分离出表征潜元的低秩部分、表征建模误差的残差部分以及表征离群样本的稀疏部分为目标, 修改优化目标项, 以实现鲁棒的数据建模. 对于缺失点问题, 可以将问题建模为具有低秩约束的矩阵补充问题, 再采用矩阵分解技术与

凸优化技术求解. 此外, 解决数据间的不一致、甚至对立的问题, 减小数据缺失对建模的影响, 可采用贝叶斯模型通过概率建模数据之间的相关性.

3.4 支持实时建模的大容量数据计算架构与方法

目前, 过程数据采集分析系统普遍以 DCS 为基础构建, 采用的实时数据库难以与高级运行数据库集成. 这类系统架构不能支持多源历史、实时大容量数据的快速建模. 传统数据建模方法对于大容量数据, 难以在现有计算平台实现. 甚至最常用的数据建模方法, 比如 PCA 算法, 在数据量过大时难以求解.

现有的以数据为中心的大数据计算架构用于过程工业大数据建模时, 存在以下问题:

1) MapReduce 是面向批处理的并行计算模型, 导致其数据处理的实时性不强, 难以满足过程工业实时建模的要求.

2) 现有的大数据分析框架是通用的体系结构, 虽擅长简单查询, 但深度分析能力不强, 难以满足过程工业多层面不规则采样、多时空时间序列数据复杂建模的需求.

为此, 过程工业大数据建模难以直接应用现有大数据计算架构来实现. 并行 (或称分布式)、在线是大数据建模的计算途径. 并行是基于分而治之思想, 在线是动态、数据流的学习策略. 有待研究的问题还包括:

1) 建立计算资源合并的软硬件体系结构, 通过同时编排在多个计算平台上的任务提高资源利用率.

2) 建立数据模式提取的计算架构. 为实现大数据时间序列模式的处理, 包括对时间序列模式的特征提取、挖掘、检索、聚类、分类等. 当运行中标定出数据模式后 (或者通过标定/提取部分特征), 建立快速从历史数据中检索出最一致或者最不一致的时段的计算架构, 用于分析理解该时段故障, 并为后续预测、数据模态识别提供支持. 此外, 通过对类似数据模态的聚类, 为该类模态建模提供基础数据.

3) 建立任务并行和数据并行的分布式数据建模方法. 由于传统软件可伸缩性差、分布式处理系统 Hadoop 深入分析能力不强, 将开源统计分析软件 (如 R、Weka 等) 与 MapReduce 技术 (Hadoop 软件) 相集成, 实现数据计算的并行处理, 增强 Hadoop 的深入分析能力.

计算方法方面, 近年理论计算科学领域的随机算法采用随机投影和随机选取技术把问题变为中等规模问题, 再利用已有数据建模方法解决大数据建模问题[78]. 其中, 随机投影技术将高维数据投影到随机选取的低维子空间来以高概率保持投影前后向量之间的欧氏距离, 随机选取技术则是随机选取部分数据样本代表整个数据样本再进行数据建模.

4 结论

过程工业大数据中蕴含的大价值推动学术界和工业界进行相关研究, 过程工业应结合自身大数据特点, 以及二十世纪九十年代以来在数据驱动建模方面的研究优势, 利用计算机行业已开发的 “大数据” 管理平台, 推进各级部门间数据的共享, 面向工业过程决策、优化、控制、故障诊断进行大数据建模理论方法研究与应用实践.

References

1 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 97−107

2 Syed A R, Gillela K, Venugopal C. The future revolution on big data. *International Journal of Advanced Research in Computer and Communication Engineering*, 2013, **2**(6): 2446−2451

3 Condliffe J. The problem with big data is that nobody understands it [Online], available: http://gizmodo.com/59062-04/the-problem-with-big-data-is-that- nobody-understands-it, April 30, 2012.

4 Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A H. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute Report [Online], available: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, June, 2011.

5 Halevi G, Moed H. The Evolution of big data as a research and scientific topic: overview of the literature. *Special Issue on Big Data, Research Trends*, 2012, (30): 1−37

6 Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*, 2009, **457**(7232): 1012−1014

7 Preis T, Moat H S, Stanley H E. Quantifying trading behavior in financial markets using Google trends. *Scientific Reports*, 2013, **3**: 1684

8 GE intelligent platform. Industrial big data cloud promotes innovation, competition and growth using big data. *Automation Panorama*, 2012, (12): 40−42
(GE 智能平台. 工业大数据云利用大数据集推动创新、竞争和增长. 自动化博览, 2012, (12): 40−42)

9 Zhong Lu-Yin. Industrial data growth rate will be two times the other big data fields. *People's Posts and Telecommunications News*.
(钟路音. 工业数据增速是其他大数据领域的两倍. 人民邮电报.)
[Online], available: http://www.cnii.com.cn/wlkb/rmydb/content/2013−08/27/content_1210645.htm, August 27, 2013.

10 Industrial Big Data. Know the future-automate processes. Software for data analysis and accurate forecasting [Online], available: http://differentia.co/qlikview/docs/Blue-Yonder-White-Paper-Industrial-Big-Data.pdf, October 23, 2015.

11 Obitko M, Jirkovský V, Bezdíček J. Big data challenges in industrial automation. *Industrial Applications of Holonic and Multi-Agent Systems, Lecture Notes in Computer Science*. Berlin Heidelberg: Springer, 2013, **8062**: 305−316

12 Schroeck M, Shockley R, Smart J, Romero-Mora-les D, Tufano P. Analytics: the real-world use of big data [Online], available: http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf, 2013.

13 Hillard R. It′s time for a new definition of big data [Online], available: http://mike2.openmethodology.org/blogs/information-development/2012/03/18/its-time-for-a-new-definition-of-big-data, March 18, 2012.

14 Yan J. Big data, bigger opportunities [Online], available: http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf, April 9, 2013.

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

16 Kano M, Tanaka S, Hasebe S, Hashimoto I, Ohno H. Monitoring independent components for fault detection. *AIChE Journal*, 2003, **49**(4): 969−976

17 Lee J M, Qin S J, Lee I B. Fault detection and diagnosis based on modified independent component analysis. *AIChE Journal*, 2006, **52**(10): 3501−3514

18 Ku W F, Storer R H, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**(1): 179−196

19 Singhal A, Seborg D E. Evaluation of a pattern matching method for the Tennessee Eastman challenge process. *Journal of Process Control*, 2006, **16**(6): 601−613

20 Yoo C K, Villez K, Lee I B, Rosén C, Vanrolleghem P A. Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnology and Bioengineering*, 2007, **96**(4): 687−701

21 Kano M, Hasebe S, Hashimoto I, Ohno H. Evolution of multivariate statistical process control: application of independent component analysis and external analysis. *Computers & Chemical Engineering*, 2004, **28**(6−7): 1157−1166

22 Rosipal R. Kernel partial least squares for nonlinear regression and discrimination. *Neural Network World*, 2003, **13**(3): 291−300

23 Sheng N, Liu Q, Qin S J, Chai T Y. Comprehensive monitoring of nonlinear processes based on concurrent kernel projection to latent structures. *IEEE Transactions on Automation Science and Engineering*, 2015, (99): 1−9, DOI: 10.1109/TASE.2015.2477272

24 Negiz A, Çinar A. Statistical monitoring of multivariable dynamic processes with state-space models. *AIChE Journal*, 1997, **43**(8): 2002−2020

25 Simoglou A, Martin E B, Morris A J. Statistical performance monitoring of dynamic multivariate processes using state space modelling. *Computers & Chemical Engineering*, 2002, **26**(6): 909−920

26 Qin S J. An overview of subspace identification. *Computers & Chemical Engineering*, 2006, **30**(10−12): 1502−1513

27 Wang J, Qin S J. A new subspace identification approach based on principal component analysis. *Journal of Process Control*, 2002, **12**(8): 841−855

28 Li W H, Qin S J. Consistent dynamic PCA based on errors-in-variables subspace identification. *Journal of Process Control*, 2001, **11**(6): 661−678

29 Ding S X, Zhang P, Naik A, Ding E L, Huang B. Subspace method aided data-driven design of fault detection and isolation systems. *Journal of Process Control*, 2009, **19**(9): 1496−1510

30 Wen Q J, Ge Z Q, Song Z H. Data-based linear Gaussian state-space model for dynamic process monitoring. *AIChE Journal*, 2012, **58**(12): 3763−3776

31 Li G, Qin S J, Zhou D H. A new method of dynamic latent variable modeling for process monitoring. *IEEE Transactions on Industrial Electronics*, 2014, **61**(11): 6438−6445

32 Kaspar M H, Ray W H. Dynamic PLS modelling for process control. *Chemical Engineering Science*, 1993, **48**(20): 3447−3461

33 Dong Y N, Qin S J. Dynamic-inner partial least squares for dynamic data modeling. In: Proceedings of the 9th International Symposium on Advanced Control of Chemical Processes (ADCHEM). Whistler, British Columbia, Canada: IFAC, 2015. 117−122

34 Li G, Liu B S, Qin S J, Zhou D H. Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: the dynamic T-PLS approach. *IEEE Transactions on Neural Networks*, 2011, **22**(12): 2262−2271

35 Liu Q, Qin S J, Chai T Y. Quality-relevant monitoring and diagnosis with dynamic comcurrent projection to latent structures. In: Proceedings of the 19th IFAC World Congress. Cape Town, South Africa: IFAC, 2014. 2740−2745

36 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2004. 137−149

37 Dittrich J, Quiané-Ruiz J A, Jndal A, Kargin Y, Setty V, Schad J. Hadoop++: making a yellow elephant run like a cheetah (without it even noticing). *Proceedings of the VLDB Endowment*, 2010, **3**(1−2): 515−529

38 Su X Y, Swart G. Oracle in-database hadoop: when mapreduce meets RDBMS. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2012. 779−790

39 Silva Y A, Reed J M. Exploiting MapReduce-based similarity joins. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2012. 693−696

40 Gudmundsson G P, Amsaleg L, Jonsson B P. Distributed high-dimensional index creation using Hadoop, HDFS and C++. In: Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing. Annecy, France: IEEE, 2012. 1−6

41 Yang L, Shi Z Z, Xu L D, Liang F, Kirsh I. DH-TRIE frequent pattern mining on Hadoop using JPA. In: Proceedings of the 2011 IEEE International Conference on Granular Computing. Kaohsiung: IEEE, 2011. 875−878

42 Böse J H, Andrzejak A, Högqvist M. Beyond online aggregation: parallel and incremental data mining with online MapReduce. In: Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud. New York, USA: ACM, 2010. Article No. 3

43 Condie T, Conway N, Alvaro P, Hellerstein J M, Elmeleegy K, Sears R. MapReduce online. In: Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2010. 313−328

44 Bu Y Y, Howe B, Balazinska M, Ernst M D. HaLoop: efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, 2010, **3**(1−2): 285−296

45 Zhang Y F, Gao Q X, Gao L X, Wang C R. iMapReduce: a distributed computing framework for iterative computation. *Journal of Grid Computing*, 2012, **10**(1): 47−68

46 Elnikety E, Elsayed T, Ramadan H E. iHadoop: asynchronous iterations for MapReduce. In: Proceedings of the 2011 IEEE 3rd International Conference on Cloud Computing Technology and Science. Athens: IEEE, 2011. 81−90

47 Zhang Y F, Gao Q X, Gao L X, Wang C R. PrIter: a distributed framework for prioritizing iterative computations. *IEEE Transactions on Parallel and Distributed Systems*, 2013, **24**(9): 1884−1893

48 Mazur E, Li B D, Diao Y L, Shenoy P J. Towards scalable one-pass analytics using MapReduce. In: Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Ph. D Forum. Shanghai, China: IEEE, 2011. 1102−1111

49 Li B D, Mazur E, Diao Y L, McGreor A, Shenoy P. A platform for scalable one-pass analytics using MapReduce. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2011. 985−996

50 Brito A, Martin A, Knauth T, Creutz S, Becker D, Weigert S, Fetzer C. Scalable and low-latency data processing with stream MapReduce. In: Proceedings of the 2011 IEEE 3rd International Conference on Cloud Computing Technology and Science. Athens: IEEE, 2011. 48−58

51 Sato-Ilic M. Preface to Part III Adaptive big data analytics. *Procedia Computer Science*, 2012, **12**: 211

52 Yan W Z, Brahmakshatriya U, Xue Y, Gilder M, Wise B. p-PIC: parallel power iteration clustering for big data. *Journal of Parallel and Distributed Computing*, 2013, **73**(3): 352−359

53 Zhao W Z, Ma H F, He Q. Parallel K-means clustering based on MapReduce. *Cloud Computing*, 2009, **5931**: 674−679

54 Gao H, Jiang J, She L, Fu Y. A new agglomerative hierarchical clustering algorithm implementation based on the map reduce framework. *International Journal of Digital Content Technology and Its Applications*, 2010, **4**(3): 95−100

55 Ordonez C, Pitchaimalai S K. Fast UDFs to compute sufficient statistics on large data sets exploiting caching and sampling. *Data & Knowledge Engineering*, 2010, **69**(4): 383−398

56 Qin S J. Process data analytics in the era of big data. *AIChE Journal*, 2014, **60**(9): 3092−3100

57 Alma Ö G. Comparison of robust regression methods in linear regression. *International Journal of Contemporary Mathematical Sciences*, 2011, **6**(9): 409−421

58 Zhou Xiao-Jian. Enhancing ε-support vector regression with gradient information. *Acta Automatica Sinica*, 2014, **40**(12): 2908−2915
(周晓剑. 考虑梯度信息的 ε-支持向量回归机. 自动化学报, 2014, **40**(12): 2908−2915)

59 Cao Peng-Fei, Luo Xiong-Lin. Wiener structure based modeling and identifying of soft sensor systems. *Acta Automatica Sinica*, 2014, **40**(10): 2179−2192
(曹鹏飞, 罗雄麟. 基于 Wiener 结构的软测量模型及辨识算法. 自动化学报, 2014, **40**(10): 2179−2192)

60 Qian Fu-Cai, Huang Jiao-Ru, Qin Xin-Qiang. Research on algorithm for system identification based on robust optimization. *Acta Automatica Sinica*, 2014, **40**(5): 988−993
(钱富才, 黄姣茹, 秦新强. 基于鲁棒优化的系统辨识算法研究. 自动化学报, 2014, **40**(5): 988−993)

61 Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 2002, **16**(3): 119−128

62 Li G, Qin S J, Zhou D H. Output relevant fault reconstruction and fault subspace extraction in total projection to latent structures models. *Industrial & Engineering Chemistry Research*, 2010, **49**(19): 9175−9183

63 Zhou D H, Li G, Qin S J. Total projection to latent structures for process monitoring. *AIChE Journal*, 2010, **56**(1): 168−178

64 Qin S J, Zheng Y Y. Quality-relevant and process-relevant fault monitoring with concurrent projection to latent structures. *AIChE Journal*, 2013, **59**(2): 496−504

65 Liu Q, Qin S J, Chai T Y. Multiblock concurrent PLS for decentralized monitoring of continuous annealing processes. *IEEE Transactions on Industrial Electronics*, 2014, **61**(11): 6429−6437

66 Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2009, **2**(1): 1−127

67 Qin S J. Process monitoring in the era of big data. In: Proceeding of the 9th International Symposium on Advanced Control of Chemical Processes (ADCHEM). Plenary Talk, Whictler, British Columbia, Canada: IFAC, 2015.

68 Fu T C. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 2011, **24**(1): 164−181

69 Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 2003, **7**(4): 349−371

70 Rakthanmanon T, Campana B, Mueen A, Batista G, Westover B, Westover B, Zhu Q, Zakaria J, Keogh E. Addressing big data time series: mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data*, 2013, **7**(3): Article No. 10

71 Jegou H, Douze M, Schmid C, Perez P. Aggregating local descriptors into a compact image representation. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE, 2010. 3304−3311

72 Yuan T, Qin S J. Root cause diagnosis of plant-wide oscillations using Granger causality. *Journal of Process Control*, 2014, **24**(2): 450−459

73 Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 2009, **33**(4): 795−814

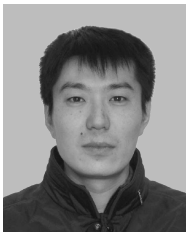
74 Suykens J A K, de Brabanter J, Lukas L, Vandewalle J. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 2002, **48**(1−4): 85−105

75 Zhang Shu-Ning, Wang Fu-Li, He Da-Kuo, Jia Run-Da. Modeling method of online robust least-squares-support-vector regression. *Control Theory & Applications*, 2011, **28**(11): 1601−1606
(张淑宁, 王福利, 何大阔, 贾润达. 在线鲁棒最小二乘支持向量机回归建模. 控制理论与应用, 2011, **28**(11): 1601−1606)

76 Mackey L W, Talwalkar A, Jordan M I. Divide-and-conquer matrix factorization. *Advances in Neural Information Processing Systems*, 2011, **24**: 1134−1142

77 Candés E J, Li X D, Ma Y, Wright J. Robust principal component analysis. *Journal of the ACM*, 2011, **58**(3): Article No. 11

78 Mahoney M W. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 2011, **3**(2): 123−224



刘 强 东北大学流程工业综合自动化国家重点实验室讲师, 美国南加州大学化工系博士后. 主要研究方向为基于数据的复杂工业过程建模与故障诊断.
E-mail: liuq@mail.neu.edu.cn
(**LIU Qiang** Lecturer at the State Key Laboratory of Synthetical Automation for Process Industries (Northeastern University), China, and Postdoctor at the Department of Chemical Engineering, University of Southern California, USA. His research interest covers statistical process monitoring, fault diagnosis of complex industrial processes.)

秦泗钊 香港中文大学 (深圳) 教授, IEEE 会士、IFAC 会士. 主要研究方向为统计过程监控、故障诊断、模型预测控制、系统辨识、建筑能源优化与控制性能监控. 本文通信作者.
E-mail: joeqin@cuhk.edu.cn
(**QIN S. Joe** Professor at the Chinese University of Hong Kong, Shenzhen, China. He is a Fellow of the International Federation of Automatic Control and a Fellow of IEEE. His research interest covers statistical process monitoring, fault diagnosis, model predictive control, system identification, building energy optimization, and control performance monitoring. Corresponding author of this paper.)