Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing: Deep Integration of AI and Process Manufacturing—Perspective

智能过程制造中的数据解析与机器学习——大数据时代的最新进展与展望

尚超 [a], Fengqi You [b,*]

[a] *Department of Automation, Tsinghua University, Beijing 100084, China*
[b] *Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, USA*

ARTICLE INFO

Article history:
Received 6 November 2018
Revised 12 January 2019
Accepted 28 January 2019
Available online 18 October 2019

关键词
大数据
机器学习
智能制造
过程系统工程

摘要

安全、高效、可持续的运行是工业生产过程控制的主要目标。然而，目前的技术严重依赖人为干预，因此在实际应用中体现出明显的局限性。蓬勃发展的大数据时代对流程工业产生了巨大的影响，为实现智能制造提供了前所未有的机遇。这种新的生产方式不仅要求机器能够帮助人类减轻繁重的体力劳动，还要能有效地承担智力劳动，甚至能够实现自主创新。为了实现这一目标，数据分析与机器学习扮演着不可或缺的角色。在本文中，我们回顾了数据分析和机器学习在工业生产过程监控、控制和优化方面的最新进展，着重分析机器学习模型的可解释性和功能性。通过分析实际需求与研究现状之间的差距，为未来的研究方向给出了建议。

© 2019 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. 引言

流程工业在促进全球经济增长、保障社会效益方面扮演着非常重要的角色。例如，世界500强企业中包括许多流程工业公司，如中石化、壳牌和埃克森美孚。随着化学工程、装备制造与信息技术的发展，现代流程工业生产过程的空间规模和功能复杂性迅速增长。这一趋势也给不同层次的最佳和安全操作带来了重大挑战。在底层的控制层，由于不同的装置与过程之间联系紧密，因此多回路、多尺度的耦合现象普遍存在，这直接阻碍了全厂控制运行策略的有效设计。此外，由于流程易受到干扰和故障源的影响，在过程设计阶段很难将这些因素考虑在内，因此异常事件的风险大大增加。在顶层的调度和计划优化中，必须根据外部环境中的各种因素，实时、灵活地做出决策；在全球竞争日益激烈的情况下，为了节约运营成本和提高经济效益，这种决策方式是必不可少的。

为了满足现代流程工业对安全、效率和可持续生产的严格要求，智能制造方面的技术革新迫在眉睫。这些需求同时也为第四次工业革命带来了机遇和挑战。第三次工业革命的兴起源于信息技术的发展，而在过去的30年里，流程工业的繁荣发展很大程度上归功于自动控制技术的广泛应用。正在进行的第四次工业革命中，人们已经普遍认识到，机器不仅要能够代替人类进行繁重的体力劳动——这是前三次工业革命的重点——还应能有效地承担智力劳动，甚至达到能够自主创新的程度。在流程工业中，所有的生产设备与过程都应该是“智能的”，这样它们才能作为一个整体智能地感知环境、发

* Corresponding author.
E-mail address: fengqi.you@cornell.edu (F. You).

2095-8099/© 2019 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文: Engineering 2019, 5(6): 1010–1016
引用本文: Chao Shang, Fengqi You. Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era. *Engineering*, https://doi.org/10.1016/j.eng.2019.01.019

掘新知识，并作出合理决策。此外，机器智能可以分为低级智能和高级智能，低级智能在功能上与人类相近，而高级智能将远远超出人类水平，这是我们未来不断追求的终极目标。

第四次工业革命的一个显著特征是可用数据的爆炸增长，它几乎影响到所有的传统学科，并促使人们对传统技术进行重新审视。作为建模、解释和处理数据并最终实现机器智能的有力武器，数据分析和机器学习在过去的几十年里得到了长足的发展，并且在过程系统工程的技术革新方面发挥了核心作用。数据分析和机器学习最早的应用可以追溯到20世纪80年代末，其标志是神经网络和反向传播算法的迅猛发展[1,2]。后来，包括主成分分析（PCA）、偏最小二乘（PLS）和支持向量机（SVM）在内的统计学习方法受到越来越多的关注，其优点在于其清晰的统计解释、模型训练的简易性和处理小样本问题的良好能力。这些方法主要应用于描述性建模，包括多元统计过程监测（MSPM）和软测量。由于机器学习的能力不断增强，人们能够有效利用数据信息，从而带来建模技术的重大改进[3]。

在当前的大数据时代，数据分析和机器学习在流程工业中得到了越来越广泛的应用。如图1所示，这些方法渗透到流程工业的各个层次，这既包括在过程监控和软测量等底层控制回路中的应用，也包括最优控制和顶层决策等应用[4]。前者的目的是帮助工程人员更好地监测和操作过程，识别过程的关键变化，而不是直接做出决策。相反，最优控制和顶层决策会对工业生产过程造成直接的影响。

在这项工作中，我们拟回顾最近的进展，总结相关文献，并对未来研究方向进行展望。由于篇幅所限，文献综述可能无法做到系统而全面，相反，我们关注两个问题：对于在系统建模中的应用，本文的文献综述关注机器学习模型的可解释性——也就是说，模型背后的物理意义以及与我们对过程理解之间的对应关系。对于在决策中的直接应用，一个重点是数据分析和机器学习的新功能以及扮演的新角色，这指的是机器学习模型旨在描述的关系或现象。通过对目前研究进展中存在的瓶颈进行分析，我们指出了一些未来值得研究的新方向。

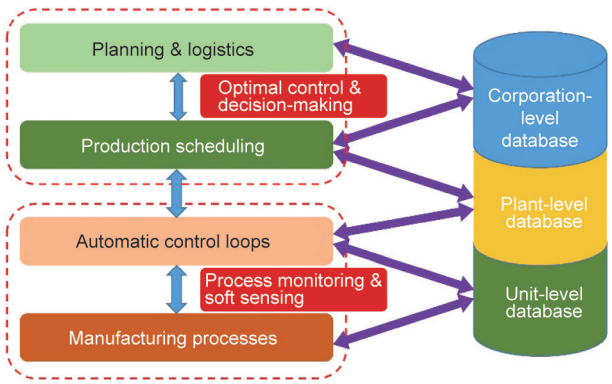

图1. 数据分析和机器学习在流程工业中的分层应用。

本文的提纲如下。在第2节和第3节中，我们将分别使用间接应用（过程监控和软测量）和直接应用（最优控制和顶层决策）来重新讨论典型的数据驱动方法。第4节是对未来研究方向的展望，最后一部分是结语。

2. 间接应用——多元统计过程监测与软测量

2.1. 表示学习——数据分析和机器学习的新思路

建模任务一般可以分为无监督学习和监督学习。在无监督学习中，通过建立描述性模型来描绘输入数据中的隐藏结构；这些主要用于描述过程数据的分布，在此基础上实现过程监控。监督学习主要建立输入与输出之间的函数映射，包括回归和分类，因此预测输出的预测精度是一个关键的因素。在工业生产过程中，快速采样的过程变量，多被用于关键质量变量的软测量建模与预报。近年来，表示学习或特征学习[5]得到了越来越多的关注，其要点在于，需要在构建模型时紧密结合特定领域的知识。这样，模型的可解释性能够得到显著增强，从而进一步提高模型性能。表示学习的一个具体例子是具有分片线性的神经网络在计算机视觉中的广泛应用。由于图形的抽象特征具有局部不变性，即分片线性，因此将特定领域的知识抽象为分片线性单元有助于提高模型的性能[6]。

表示学习为无监督学习和监督学习提供了统一的观点。无监督学习方法可以看作“特征检测器”，用于从输入数据中提取可解释的隐含特征。随后，这些特征可以用作分类器或回归器的输入，从而能够显著提高监督学习的性能。这正是在深度学习技术[7]的思想。换言之，非监督学习和监督学习并不是相互孤立的；相反，前者能够为后者起到积极的推动作用。

总之，一个理想的模型，无论多么复杂，都应该具有清晰的物理解释。接下来一个自然的问题是：怎样的先验知识能够很好地反映过程数据的特征？事实上，这是许多MSPM研究的一个共同的焦点。为了解释清楚这个问题，我们首先回顾有关MSPM的最新进展，然后回顾软测量方法的进展。

2.2. 基于特征学习的 MSPM

由于现代流程工业规模大、耦合性强，微小故障的影响将被显著放大。因此，持续监控操作状态并采取必要的维护措施对于确保安全是至关重要的，尽管这样做需要繁重的手工劳动[8]。自20世纪80年代以来，MSPM已经成为解决这一难题的手段，大量经典的机器学习算法被应用，很好地体现了工业制造业的智能化发展方向。一些综述文章已经对相关工作进行了很好的总结[9,10]。

近年来的研究工作旨在利用有关连续生产过程的特定先验知识，以建立有效的MSPM模型。由于工业过程的过渡过程时间通常较长，因此整个系统通常表现出一定的动态特性。这可以描述为，过程的本质变化具有缓慢变化的特点。因此，变化的快慢程度可以视为一种有意义的特征属性，在此基础上能够有效地描述过程的动态特性[11]。据此，人们提出了基于慢特征分析（SFA）的过程数据建模方法，并实现有效的过程监控和诊断[12,13]。与主成分分析（PCA）、独立成分分析（ICA）、典型变量分析（CVA）等传统监控方法相比，慢特征分析有其独特的性质，能够对工业过程的稳态和暂态行为分别进行描述。因此，通过设计专门针对过程动态异常的监控统计量，可以提供更有意义的信息；这样，操作点的正常切换可以与引起动态异常的实际故障清楚地区分开来。田纳西伊士曼过程的监控结果表明，该种策略可将误报警率降低一个数量级[12]。在缓慢变化准则的驱动下，人们进一步提出了多种改进的监测方法，包括用于自适应监测[14]以及概率监测[15]的递归SFA算法，并且成功应用于压力钻井过程[16]和批次生产过程[17–19]中。

文献[20,21]中提出了另一种动态过程数据分析方法——动态内部主成分分析（DiPCA），其中主时间序列作为潜在变量，基于可预测性被依次提取出来。在这些学者的工作中，使用自回归（AR）模型构建回归模型，并在此基础上定义不同主时间序列的可预测性。在我们看来，DiPCA与SFA相似，都是将潜在变量的动态内容最大化。粗略地说，可预测性可以被看成是慢度的一个特例，因为可以被AR模型很好地描述的时间序列数据往往变化缓慢。对于具有不可忽略动态的过程，上述方法可以提供比传统动态统计模型如动态PCA (DPCA) [22]和动态ICA (DICA) [23]更好的描述性模型。

经典的机器学习模型通常被设计成单模态。对于大型工业生产过程，存在多种操作条件，并且经常在不同模式之间进行切换。因此，在为MSPM设计机器学习模型时，应该将多模态概念化为一种领域特定的特征。多模式过程监控最简单的模型是高斯混合模型（GMM）[24]。遗憾的是，GMM并没有提供关于不同模式之间转换概率的信息，这进一步促进了隐马尔可夫模型（HMM）在多模型监测[25]中的使用。在GMM和HMM中，过程数据在单一模式下的分布都假定为高斯分布，尽管这种假定在实践中有很大的局限性。因此，我们设计了更通用的模型来缓解这一假设[26]。

文献[27]提出了一种不同的过程监控方法，这也符合特征学习的概念。这种方法将一般的过程监控图，如T^2和平方预测误差（SPE），视为低等级类型的特征，随后它们被作为高级过程监控模型（如PCA）的输入。这样，可以实现多个过程监控模型信息的有效融合，从而系统地考虑过程数据的不同特征。由于所提取的特征在统计意义上是高斯分布的，因此使用PCA作为高级过程监控模型是合理的。

2.3. 基于特征学习的软测量

软测量的历史可以追溯到Brosilow和Tong[28]于1978年提出的推理控制策略。作为一种智能感知技术，软测量利用易于测量的过程变量，实现对难以测量但重要的指标（如产品质量和其他环境指标）的在线估计。值得一提的是，关键性能指标（KPI）预测是软测量[29]的又一新兴应用。一些重要的性能指标必须基于成本高昂的实验室化验进行评估，而建立用于质量预报的软测量模型可以提供这些指标的实时估计，进而有助于过程操作人员的决策。简而言之，软测量建模可以视为一个回归问题，因此，人们提出了多种监督学习算法应用于软测量建模，文献[30]全面地回顾总结了相关成果。

文献[11]中首次指出了从表示学习出发实现软测量建模的广阔前景，利用概率SFA (PSFA)来得到缓慢变化的特征，在此基础上建立简单最小二乘回归模型。由于慢特征很好地描述了过程的潜在变化，因此其中一部分往往与质量指标高度相关。与传统的动态PLS方法（DPLS）相比，该方法具有更好的动态预测精度。此外，它还能够综合利用快速采样过程数据和不规则采样质量数据的信息，因而具有半监督学习的特点。后来还开发出了许多拓展形式。参考文献[31]提出了一种贝叶斯学习方法来提取不断变化的动态特征，其中慢特征的缓慢程度被认为是逐渐变化的。在参考文献[32]中，引入了另一层灵活性来处理各种有用的慢特征。随后提出了一

种改良的规范化SFA，用于工业对苯二甲酸加氢精制工艺的质量预报[33]。

值得关注的是，深度学习本身就很好地体现了表示学习的内涵。文献[34]首次将深度学习技术应用于软测量，利用深度神经网络（DNN）构建软测量模型来预测原油蒸馏装置中柴油的切割点温度。DNN的训练过程包括两个步骤：通过无监督学习初始化DNN的权值，进一步通过监督学习根据输入-输出数据调整权值。因此，无监督学习步骤可以视为提取了非线性的潜在特征，这些特征诱导过程变量之间的非线性相关，这有利于进一步建立回归模型。在此基础上，深度学习技术已应用于原油分类[35]和二氧化碳（CO_2）捕捉过程建模[36]，二者均体现了在大规模过程数据建模中使用深度学习技术的优势。深度学习基于无监督学习步骤中的提取特征，因而也可以应用于过程监控和故障诊断[37,38]。

无须赘述，软测量模型也可以建立在其他特征的基础上，比如低维子空间中的隐变量相关性。在这方面最早的方法是主成分回归（PCR），其中特征提取是基于PCA进行的。文献[39]全面总结了低维潜变量模型在软测量中的应用。文献[40]提出利用邻域保持嵌入法学习数据的内在非线性结构，并在此基础上建立软测量预报模型。

3. 直接应用——最优控制和高层决策

3.1. 最优控制中的数据分析和机器学习

在工业生产过程的先进控制中，模型预测控制（MPC）是一种著名而成熟的控制方法，它建立在一个精确已知的数学模型的基础上对过程行为进行描述，进而求解一段时间内最优的控制序列[41]。然而，MPC的基本假设在实际中可能过于理想化的，而且模型失配、不可测干扰、随机噪声等未知因素普遍存在。在这些情况下，一种有前途的方法是将机理模型与数据驱动建模相集成，这在应对不确定性方面显示了巨大潜力[42]。根据功能的不同，它们的应用可以分为两类。

第一类应用是通过拟合过去已有的数据，建立未知因素的预测模型，将不确定性分解为已知的确定性部分和代表预测误差的随机部分。例如，在智能电网的运行和控制中，可以基于天气预报和气候因素等其他信息来源对可再生能源（包括风能和太阳能）发电量进行估计[43]。类似地，若最优控制涉及产品质量等无法在线测量的指标，则可以建立软测量模型提供实时估计，这对于闭环控制是必不可少的。在这两种情况下，机器学习模型如SVM和神经网络都得到了广泛的应用。建立一个好的预测模型有助于显著减轻不确定性的影响，从而获得更好的控制性能。从这个意义上说，预测模型的准确性是一个关键问题。

在MPC中，数据驱动建模的第二类应用涉及以无监督学习方式描述不确定性的分布。在实际应用中，系统不可避免地会受到不确定的扰动，使系统状态偏离标称轨迹。为了解决不确定性的影响，鲁棒MPC（RMPC）[44]和随机MPC（SMPC）[45]最为常用，它们分别使用不同的数学工具来描述不确定性。在RMPC中，不确定集合用来表示不确定性实现的可能区域，而SMPC直接使用了概率分布。最近，在RMPC和SMPC中一个有前景的研究方向是，采用主动学习的观点来合理地对不确定性进行建模。在RMPC中，通常采用传统的基于范数的集合作为不确定性集，然而这缺乏足够的灵活性，不能很好地描述不确定性的分布。因此，用无监督学习方法构造的数据驱动型不确定集合可以很好地解决这一问题。例如，通过主动地从现有数据中学习一个数据密集分布的区域，进而可以将得到的最优控制问题转化为一个易于求解的经典稳健优化（RO）问题[46]。为调整多面体的大小，还提出了一种新的策略，该策略为RMPC的性能给出了概率保证，从而表明最终将得到SMPC的近似解[46]。建立的样本量的理论界大大低于SMPC中的经典结果，因此增强了SMPC的实用性，降低了其保守性。该方法已应用于灌溉控制系统[47]，结果表明，从数据中挖掘有意义的信息可以显著提高系统的安全性和闭环性能。在文献[48]中，采用了一种基于学习的MPC方法来处理自治系统中的重复控制任务。

3.2. 高层决策中的数据分析和机器学习

不确定性环境下的通用优化决策技术可分为随机规划（SP）[49]、鲁棒优化（RO）[50]和分布鲁棒优化（DRO）[51]；这三种技术在能源系统和供应链设计中已经得到了广泛的应用[52,53]。数据驱动决策近年来开始获得越来越多的关注，它集成了基于模型和数据驱动方法来实现不确定性条件下的优化决策。通过有机结合机器学习和数学规划，能够实现更强大、更高效的数据驱动优化框架，这些框架能够有效避免在数据分析和决策[54]之间的不断迭代。

情景规划能为经典的机会约束问题提供近似解，通过采用从过去经验中收集的数据将机会约束转换为大量

确定性约束[55]。确保求解质量的关键是选择足够数量的数据。在这方面已经有了一些理论结果[56]。然而，得到的优化问题包含了大量的约束，这给问题的求解带来了巨大的挑战。由于情景规划具有可分解结构，人们设计了许多分解算法，如L形法[49,57,58]。最近的一个研究热点是采用分布式优化技术[59,60]，通过将原始的大规模问题分解为若干个子问题，然后使用多个处理器来解决这些彼此联系有限的子问题。

在数据驱动RO中，不确定集合通常是直接基于不确定性数据构建的。从机器学习的角度来看，这可以理解为一个无监督学习任务。然而，并不是所有的无监督学习方法都可以用于此目的，主要是因为它需要考虑求解优化问题的复杂性。一方面，无监督学习方法必须足够准确地捕捉不确定性的分布；另一方面，过于复杂的不确定集合会使优化问题难以求解。因此，必须精心设计数据驱动的不确定集，以便在两个相互冲突的目标之间实现理想的平衡。基于这种思想，近年来出现了构建数据驱动不确定集合的非监督学习方法。文献[61]提出了一种基于分段线性核的SVC，它是一种专门面向数据驱动RO的新方法。通过求解二次规划，将大量不确定数据的分布有效地描述为一个紧凑的凸不确定集，从而大大降低了决策的保守性。此外，仅通过调整一个参数就能很容易地选择数据驱动不确定集的数据覆盖率，这为控制保守性和排除异常值提供了一种实用的方法。文献[62]利用PCA和核密度估计建立了数据驱动的不确定集，可以系统地处理数据分布的相关性和不对称性。文献[63,64]研究了数据驱动的不确定集合在多阶段自适应RO (ARO)中的应用，并将其应用于流程工业中的调度优化和计划优化中。结果表明，由于机器学习的强大数据挖掘能力，在多阶段情况下决策的保守性得到了显著降低。特别地，通过主动学习不确定需求的分布区域，在规划应用中可以提升超过20%的净现值。这种方法后来被应用于机组负荷管理[65]、工业蒸汽系统的最优操作[66]和供应链的设计[67]。为了利用多类不确定性数据中的标签信息，文献[68]提出了一种数据驱动的随机RO框架。在文献[69]中，数据驱动的ARO同时考虑了传统的鲁棒性和极大极小后悔准则，从而得到合理的决策。

数据驱动的分布鲁棒优化（DRO）在过去十年中一直是运筹学研究的热门话题。它可以看做某种RO和SP的结合，可以优化一组概率分布上最坏情况下的性能[51]。在数据驱动的DRO中，模糊集起着关键作用，且通常基于数据来确定。在自然界中，不确定性的分布通常是不精确的。为了有效应对分布不确定性，采用一组候选概率分布构建模糊集。描述模糊度最常用的方法是从过去的数据中提取一阶矩和二阶矩信息。文献[70]根据假设检验系统地解决了模糊集大小的设定问题。在流程工业中，DRO首次应用于流程计划优化和调度优化[71]，以及页岩气供应链的优化运行[72]。

4. 展望未来的研究方向

4.1. 过程监测

尽管多种特征提取方法被用于设计过程监控模型，但值得注意的是，提取的特征必须与过程特定的先验知识密切相关。目前，尽管很多过程监测模型本质上是不同特征提取方法的组合，但作为许多监测模型基础的非线性流形和非高斯性假设本质上并不是过程数据所“特有”的。通常，一个好的模型不仅应该能够描述过程属性，而且应该有着清晰的屋里解释，进而很容易被工程技术人员所接受[73]。但是，目前这个问题并没有得到充分的考虑。从这个意义上说，未来的研究应关注诸如缓慢变化、非平稳和因果关系等高阶过程特性[3,74]。此外，特征的设计可以充分结合工程人员的先验知识，如变量的单调性和取值范围信息，因为具有可解释性的模型有助于潜在故障检测后的根源诊断和维护[75]。

在特征的基础上，可以进一步采用迁移学习来综合来自不同操作条件或不同制造设备的数据信息。目前的研究重点是为每种操作条件或每种设备建立单独的模型。尽管这些模型之间存在差异，但也有一些相似之处，因此可以认为这些模型中存在着一些共同的信息。借鉴迁移学习的思想[76]，应该将特征作为描述制造过程基本原理的本质信息来进行发掘，从而提高建模性能以及人们对它的理解。

此外，另一个方向是为过程监控开发对用户友好的可视化技术，以便更好地辅助决策，因为可视化可能有助于更好地理解高维过程数据[77]。

4.2. 软测量

由于工业过程的时变特性，软测量模型的性能很容易随着时间的推移而下降，这就需要大量的人力工作来维护和更新模型。因此，质量预测不仅仅是一个简单的回归问题，而应更多地关注预测模型的自适应更新机制，特别是在经常存在运行条件偏差的情况下[78]。此

外，由于人为因素造成的化验数据数据的不准确也是需要考虑的，如不确定的时间延迟、较大且变化的采样间隔、不同操作人员的采样习惯等。

传统的监督模型通常建立在数据样本独立且同分布的前提下。然而，影响产品质量的过程变量可能要复杂得多。新兴的在线学习理论为没有特定假设的建模任务提供了新的解决思路[79]。例如，在线学习可以系统地处理确定性的、随机的、甚至以对抗规则生成的数据。因此，在大数据时代，使用在线学习技术来解决质量预报问题是一个值得尝试的思路。

同时，随着成像技术的快速发展，越来越多的图像和光谱数据被采集下来，为高精度预测模型的建立提供了有意义的信息。然而，变量的高维性和强相关性也为建模带来了挑战。图像处理和目标识别已经在遥感、自动驾驶等领域发挥了主导作用[80]。虽然这些技术已经应用于流程工业[81,82]，但其发展仍处于起步阶段。因此，利用先进的图像处理技术——尤其是卷积神经网络——以便充分利用流程工业的图像和光谱数据将是一个可行的方向。

4.3. 数据驱动的最优控制

未来的研究工作将会在设计RMPC中的不确定集时结合特定领域的知识。例如，在参考文献[47]中，提出了一个条件不确定集的新概念来描述降雨预报误差分布对预报值的依赖关系。对于其他类型的不确定性，如何设计相关的条件不确定性集的问题值得进一步的个案研究。

作为一种流行的机器学习方法，强化学习（RL）特别适用于在没有模型信息的情况下进行决策[83]。该方法具有数据驱动的特点，能够从本质上适应时变环境。因此，基于RL的控制在处理复杂制造工厂的控制与决策方面具有很大的潜力，而这些工厂的高保真数学模型在实践中很难建立[84,85]。

4.4. 顶层决策

与其他应用相比，顶层决策是最重要的，因为它直接影响一个流程工业企业的经济利润和环境影响。一方面，顶层决策通常是根据决策者的经验在不确定的情况下做出的，有很大改进空间；因此，预计今后将有更多的数据驱动RO和DRO应用于流程工业的决策。另一方面，如何进一步提高数据驱动RO和DRO的求解质量和计算效率，具有一定的研究价值。目前的DRO方法利用矩信息来描述概率分布的模糊性。从理论上讲，不同类型的矩信息可以看作是简单数据分析方法的结果。这种情况为使用先进的无监督学习方法提取高维特征空间中的分布等高级信息提供了思路，并在此基础上进一步考虑模糊性。利用机器学习的建模能力，有望降低概率分布的模糊性，从而得到减轻解的保守性。例如，参考文献[86]中提出的模糊集涉及一系列嵌套集合的概率；然而，关于如何构造这些嵌套集还没有成熟、系统的结果。实际上，可以利用具有不同正则化参数的基于核的机器学习算法来获得嵌套集，从而挖掘大部分数据样本中的信息。

5. 结语

在现代流程工业中，可以收集和存储越来越多的蕴藏有价值信息的数据。通过利用数据，数据分析和机器学习可以帮助感知环境、发现知识，并自动智能地做出决策。本文以数据驱动的监测、预测、控制和优化为研究方向，回顾了该领域的研究现状，分析了有待研究的问题。特别地，我们将数据驱动方法的底层应用（包括监控和软测量）与顶层应用（包括控制和优化）区分开来。对于前者，模型的可解释性是一个主要问题，而对于后者，则特别关注方法的功能性。值得注意的是，尽管大数据在极大程度上改变了流程工业的发展，但大部分数据驱动的方法尚未应用于实践。数据分析和机器学习绝不是解决一切难题的万能钥匙。最重要的是，为了实现成功的应用，有必要充分结合装置和过程的先验知识，这给未来的研究带来了挑战和机遇。

致谢

尚超博士感谢国家自然科学基金的支持（61673236，61433001，61873142）。

Compliance with ethics guidelines

Chao Shang and Fengqi You declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Willis MJ, Di Massimo CD, Montague GA, Tham MT, Morris AJ. Artificial

neural networks in process engineering. IEE Proc Contr TheorAppl 1991;138 (3):256–66.
[2] Willis MJ, Montague GA, Di Massimo C, Tham MT, Morris AJ. Artificial neural networks in process estimation and control. Automatica 1992;28(6):1181–7.
[3] MacGregor J, Cinar A. Monitoring, fault diagnosis, fault-tolerant control and optimization: data driven methods. Comput Chem Eng 2012;47:111–20.
[4] Pillonetto G, Dinuzzo F, Chen T, De Nicolao G, Ljung L. Kernel methods in system identification, machine learning and function estimation: a survey. Automatica 2014;50(3):657–82.
[5] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798–828.
[6] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning; 2010 Jun 21–25; Haifa, Israel; 2010. p. 807–14.
[7] Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Schölkopf B, Platt J, Hofmann T, editors. Advances in neural information processing systems 19: Proceedings of the 2006 Conference. Cambridge: The MIT Press; 2007. p. 153–60.
[8] Shu Y, Ming L, Cheng F, Zhang Z, Zhao J. Abnormal situation management: challenges and opportunities in the big data era. Comput Chem Eng 2016;91:104–13.
[9] Chiang L, Lu B, Castillo I. Big data analytics in chemical engineering. Annu Rev Chem Biomol Eng 2017;8:63–85.
[10] Ge Z, Song Z, Ding SX, Huang B. Data mining and analytics in the process industry: the role of machine learning. IEEE Access 2017;5:20590–616.
[11] Shang C, Huang B, Yang F, Huang D. Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modeling. AIChE J 2015;61(12):4126–39.
[12] Shang C, Yang F, Gao X, Huang X, Suykens JAK, Huang D. Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis. AIChE J 2015;61(11):3666–82.
[13] Shang C, Huang B, Yang F, Huang D. Slow feature analysis for monitoring and diagnosis of control performance. J Process Contr 2016;39:21–34.
[14] Shang C, Yang F, Huang B, Huang D. Recursive slow feature analysis for adaptive monitoring of industrial processes. IEEE Trans Ind Electron 2018;65 (11):8895–905.
[15] Guo F, Shang C, Huang B, Wang K, Yang F, Huang D. Monitoring of operating point and process dynamics via probabilistic slow feature analysis. Chemom Intell Lab Syst 2016;151:115–25.
[16] Gao X, Li H, Wang Y, Chen T, Zuo X, Zhong L. Fault detection in managed pressure drilling using slow feature analysis. IEEE Access 2018;6:34262–71.
[17] Zhang H, Tian X, Deng X. Batch process monitoring based on multiway global preserving kernel slow feature analysis. IEEE Access 2017;5:2696–710.
[18] Zhang S, Zhao C. Slow-feature-analysis-based batch process monitoring with comprehensive interpretation of operation condition deviation and dynamic anomaly. IEEE Trans Ind Electron 2019;66(5):3773–83.
[19] Zhang H, Tian X, Deng X, Cao Y. Batch process fault detection and identification based on discriminant global preserving kernel slow feature analysis. ISA Trans 2018;79:108–26.
[20] Dong Y, Qin SJ. A novel dynamic PCA algorithm for dynamic data modeling and process monitoring. J Process Contr 2018;67:1–11.
[21] Dong Y, Qin SJ. Dynamic latent variable analytics for process operations and control. Comput Chem Eng 2018;114:69–80.
[22] Ku W, Storer RH, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. Chemom Intell Lab Syst 1995;30(1):179–96.
[23] Lee JM, Yoo C, Lee IB. Statistical monitoring of dynamic processes based on dynamic independent component analysis. Chem Eng Sci 2004;59 (14):2995–3006.
[24] Yu J, Qin SJ. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. AIChE J 2008;54(7):1811–29.
[25] Wang F, Tan S, Shi H. Hidden Markov model-based approach for multimode process monitoring. Chemom Intell Lab Syst 2015;148:51–9.
[26] Bai X, Lu G, Hossain MM, Szuhánszki J, Daood SS, Nimmo W, et al. Multi-mode combustion process monitoring on a pulverised fuel combustion test facility based on flame imaging and random weight network techniques. Fuel 2017;202:656–64.
[27] He QP, Wang J. Statistical process monitoring as a big data analytics tool for smart manufacturing. J Process Contr 2018;67:35–43.
[28] Brosilow C, Tong M. Inferential control of processes: part II. the structure and dynamics of inferential control systems. AIChE J 1978;24(3):492–500.
[29] Shardt YAW, Hao H, Ding SX. A new soft-sensor-based process monitoring scheme incorporating infrequent KPI measurements. IEEE Trans Ind Electron 2015;62(6):3843–51.
[30] Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. Comput Chem Eng 2009;33(4):795–814.
[31] Ma Y, Huang B. Bayesian learning for dynamic feature extraction with application in soft sensing. IEEE Trans Ind Electron 2017;64(9):7171–80.
[32] Ma Y, Huang B. Extracting dynamic features with switching models for process data analytics and application in soft sensing. AIChE J 2018;64 (6):2037–51.
[33] Zhong W, Jiang C, Peng X, Li Z, Qian F. Online quality prediction of industrial terephthalic acid hydropurification process using modified regularized slowfeature analysis. Ind Eng Chem Res 2018;57(29):9604–14.
[34] Shang C, Yang F, Huang D, Lyu W. Data-driven soft sensor development based on deep learning technique. J Process Contr 2014;24(3):223–33.
[35] Gao X, Shang C, Jiang Y, Huang D, Chen T. Refinery scheduling with varying crude: a deep belief network classification and multimodel approach. AIChE J 2014;60(7):2525–32.
[36] Li F, Zhang J, Shang C, Huang D, Oko E, Wang M. Modelling of a postcombustion CO2 capture process using deep belief network. Appl Therm Eng 2018;130:997–1003.
[37] Zhang Z, Zhao J. A deep belief network based fault diagnosis model for complex chemical processes. Comput Chem Eng 2017;107:395–407.
[38] Wu H, Zhao J. Deep convolutional neural network model based chemical process fault diagnosis. Comput Chem Eng 2018;115:185–97.
[39] Ge Z. Process data analytics via probabilistic latent variable models: a tutorial review. Ind Eng Chem Res 2018;57(38):12646–112461.
[40] Yuan X, Ge Z, Ye L, Song Z. Supervised neighborhood preserving embedding for feature extraction and its application for soft sensor modeling. J Chemometr 2016;30(8):430–41.
[41] Chu Y, You F. Model-based integration of control and operations: overview, challenges, advances, and opportunities. Comput Chem Eng 2015;83:2–20.
[42] Yan Z, Wang J. Robust model predictive control of nonlinear systems with unmodeled dynamics and bounded uncertainties based on neural networks. IEEE Trans Neural Netw Learn Syst 2014;25(3):457–69.
[43] Appino RR, González Ordiano JÁ, Mikut R, Faulwasser T, Hagenmeyer V. On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages. Appl Energy 2018;210:1207–18.
[44] Saltık MB, Özkan L, Ludlage JHA, Weiland S, Van den Hof PMJ. An outlook on robust model predictive control algorithms: reflections on performance and computational aspects. J Process Contr 2018;61:77–102.
[45] Farina M, Giulioni L, Scattolini R. Stochastic linear model predictive control with chance constraints—a review. J Process Contr 2016;44:53–67.
[46] Shang C, You F. A data-driven robust optimization approach to scenario-based stochastic model predictive control. J Process Contr 2019;75:24–39.
[47] Shang C, Chen WH, Stroock AD, You F. Robust model predictive control of irrigation systems with active uncertainty learning and data analytics. IEEE Trans Contr Syst Technol. Epub 2019 May 31.
[48] Rosolia U, Zhang X, Borrelli F. Data-driven predictive control for autonomous systems. Robot Auton Syst 2018;1:259–86.
[49] Marti K, Kall P. Stochastic programming. Berlin: Springer; 1994.
[50] Ben-Tal A, El Ghaoui L, Nemirovski A. Robust optimization. Princeton: Princeton University Press; 2009.
[51] Delage E, Ye Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res 2010;58 (3):595–612.
[52] Gebreslassie BH, Yao Y, You F. Design under uncertainty of hydrocarbon biorefinery supply chains: multiobjective stochastic programming models, decomposition algorithm, and a comparison between CVaR and downside risk. AIChE J 2012;58(7):2155–79.
[53] Garcia DJ, You F. Supply chain design and optimization: challenges and opportunities. Comput Chem Eng 2015;81:153–70.
[54] Bertsimas D, Thiele A. Robust and data-driven optimization: modern decision making under uncertainty. In: Johnson MP, Norman B, Secomandi N, editors. Models, methods, and applications for innovative decision making. Catonsville: The Institute for Operations Research and the Management Sciences; 2006. p. 95–122.
[55] Calafiore G, Campi MC. Uncertain convex programs: randomized solutions and confidence levels. Math Program 2005;102(1):25–46.
[56] Campi MC, Garatti S. The exact feasibility of randomized solutions of uncertain convex programs. SIAM J Optim 2008;19(3):1211–30.
[57] Birge JR, Louveaux F. Introduction to stochastic programming. 2nd ed. New York: Springer Science & Business Media; 2011.
[58] You F, Grossmann IE. Multicut Benders decomposition algorithm for process supply chain planning under uncertainty. Ann Oper Res 2013;210 (1):191–211.
[59] Carlone L, Srivastava V, Bullo F, Calafiore GC. Distributed random convex programming via constraints consensus. SIAM J Contr Optim 2014;52 (1):629–62.
[60] You K, Tempo R, Xie P. Distributed algorithms for robust convex optimization via the scenario approach. IEEE Trans Automat Contr 2019;64 (3):880–95.
[61] Shang C, Huang X, You F. Data-driven robust optimization based on kernel learning. Comput Chem Eng 2017;106(2):464–79.
[62] Ning C, You F. Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods. Comput Chem Eng 2018;112:190–210.
[63] Ning C, You F. Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty. AIChE J 2017;63(9):3790–817.
[64] Ning C, You F. A data-driven multistage adaptive robust optimization framework for planning and scheduling under uncertainty. AIChE J 2017;63 (10):4343–69.
[65] Ning C, You F. Data-driven adaptive robust unit commitment under wind power uncertainty: a Bayesian nonparametric approach. IEEE Trans Power Syst 2019;34(3):2409–18.
[66] Zhao L, Ning C, You F. Operational optimization of industrial steam systems under uncertainty using data-driven adaptive robust optimization. AIChE J

2019;65(7):e16500.
[67] Zhao S, You F. Resilient supply chain design and operations with decisiondependent uncertainty using a data-driven robust optimization approach. AIChE J 2019;65(3):1006–21.
[68] Ning C, You F. Data-driven stochastic robust optimization: general computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era. Comput Chem Eng 2018;111:115–33.
[69] Ning C, You F. Adaptive robust optimization with minimax regret criterion: multiobjective optimization framework and computational algorithm for planning and scheduling under uncertainty. Comput Chem Eng 2018;108:425–47.
[70] Bertsimas D, Gupta V, Kallus N. Data-driven robust optimization. Math Program 2018;167(2):235–92.
[71] Shang C, You F. Distributionally robust optimization for planning and scheduling under uncertainty. Comput Chem Eng 2018;110:53–68.
[72] Gao J, Ning C, You F. Data-driven distributionally robust optimization for shale gas supply chains under uncertainty. AIChE J 2019;65(3):947–63.
[73] MacGregor JF, Bruwer MJ, Miletic I, Cardin M, Liu Z. Latent variable models and big data in the process industries. In: Proceedings of 9th International Symposium on Advanced Control of Chemical Processes; 2015 Jun 7–10; Whistler, BC, Canada; 2015. p. 521–5.
[74] Shu Y, Zhao J. Data driven causal inference based on a modified transfer entropy. Comput Chem Eng 2013;57:173–80.
[75] Qin SJ. Survey on data-driven industrial process monitoring and diagnosis. Annu Rev Contr 2012;36(2):220–34.
[76] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59.
[77] Wang R, Edgar TF, Baldea M, Nixon M, Wojsznis W, Dunia R. A geometric method for batch data visualization, process monitoring and fault detection. J Process Contr 2018;67:197–205.
[78] Kadlec P, Grbić R, Gabrys B. Review of adaptation mechanisms for data-driven soft sensors. Comput Chem Eng 2011;35(1):1–24.
[79] Morariu O, Morariu C, Borangiu T, Ra˘ileanu S. Manufacturing systems at scale with big data streaming and online machine learning. In: Borangiu T, Trentesaux D, Thomas A, Cardin O, editors. Service orientation in holonic and multi-agent manufacturing. Cham: Springer; 2018. p. 253–64.
[80] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA. Washington, DC: IEEE; 2012. p. 3354–61.
[81] Duchesne C, Liu JJ, MacGregor JF. Multivariate image analysis in the process industries: a review. Chemom Intell Lab Syst 2012;117:116–28.
[82] Chen M, Khare S, Huang B. A unified recursive just-in-time approach with industrial near infrared spectroscopy application. Chemom Intell Lab Syst 2014;135:133–40.
[83] Duan Y, Chen X, Houthooft R, Schulman J, Abbeel P. Benchmarking deep reinforcement learning for continuous control. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA; 2016. p. 1329–38.
[84] Lewis FL, Vrabie D, Vamvoudakis KG. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. IEEE Control Syst 2012;32(6):76–105.
[85] Liu D, Yang X, Wang D, Wei Q. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints. IEEE Trans Cybern 2015;45(7):1372–85.
[86] Wiesemann W, Kuhn D, Sim M. Distributionally robust convex optimization. Oper Res 2014;62(6):1358–76.