机器学习结合近红外光谱的发酵过程参数预测与质量稳定性控制

于　倩
伊犁川宁生物技术股份有限公司　新疆伊宁　835000

摘　要：本研究聚焦于机器学习与近红外光谱技术的协同应用，旨在解决发酵过程关键参数实时预测及质量稳定性控制难题。通过解析近红外光谱对发酵液中含氢基团振动吸收特性，结合化学计量学预处理手段优化光谱数据质量，构建了涵盖线性模型（偏最小二乘回归）、非线性模型（人工神经网络、循环神经网络）及集成学习框架的综合预测体系。该体系可同步反演葡萄糖消耗速率、微生物密度等多维指标，并通过堆叠策略融合不同算法优势以提升预测精度。基于上述模型设计闭环控制系统，集成光谱监测模块与执行机构，采用比例－积分－微分调节机制实现发酵环境的动态调控。实验表明，此方案能有效识别代谢抑制、污染入侵等异常工况，配合统计过程控制方法量化生产过程能力指数，最终形成覆盖工艺优化、实时监控与持续改进的质量管控闭环。研究突破传统单模态检测局限，为生物制造领域提供了兼具准确性与鲁棒性的智能化解决方案。

关键词：机器学习；近红外光谱；发酵过程；参数预测；质量稳定性控制

现代生物工业对发酵过程精细化管理提出更高要求，传统离线检测手段因时效性差、操作繁琐难以满足连续化生产需求。近红外光谱技术凭借其无损检测、多组分同步分析的特性，成为过程分析技术的重要工具，但其原始光谱易受物理状态干扰且存在特征重叠问题。与此同时，机器学习算法在处理高维非线性数据方面展现独特优势，尤其在复杂系统建模领域具有显著潜力。现有研究多集中于单一模型开发，忽视了发酵过程动态特性与多参数耦合效应的影响，导致模型泛化能力不足。本文针对上述痛点展开系统性探索，重点攻克以下科学问题：①如何利用光谱物理特性建立高效预处理流程；②怎样设计自适应算法体系应对发酵非线性特征；③能否通过智能控制策略实现质量稳定性主动调控。通过理论创新与工程实践相结合，旨在为抗生素、酶制剂等高附加值产品的稳定生产提供技术支撑。

一、近红外光谱技术原理与发酵过程监测特性

1.近红外光谱技术物理基础

近红外光谱波长范围为780–2520 nm，对应含氢基团（C–H、O–H、N–H）的倍频与合频振动吸收。发酵液中葡萄糖、乳酸、乙醇等有机物分子结构包含大量含氢基团，其振动模式与光谱吸收峰位置具有特异性。例如，C–H基团伸缩振动吸收峰位于1100–1350 nm区间，O–H基团弯曲振动吸收峰位于1400–1600 nm区间。通过分析光谱吸收强度与波长分布，可反推发酵液中目标组分浓度。近红外光谱仪采用傅里叶变换或光栅分光技术，将复合光分解为单色光后照射样品，检测器接收透射或反射光信号并转换为电信号，经数据处理系统生成光谱图。该过程无需样品预处理，可直接通过在线探头实现发酵罐内液体或固体的实时检测。

2.发酵过程监测技术优势

与传统色谱法、电化学传感器相比，近红外光谱技术具有显著优势。其一，无损检测特性避免样品消耗与污染风险，支持连续监测需求。其二，多参数同步检测能力可同时获取葡萄糖消耗速率、有机酸积累量、微生物密度（OD值）及产物效价等指标，覆盖发酵全周期。其三，在线探头设计（如固定光栅结构）具备抗震动、耐腐蚀特性，适应工业发酵环境的高温、高压及高湿度条件。模块化测量池支持液体与固体样品切换，适配不同发酵工艺需求。通过旁路循环系统，可实现厌氧发酵过程中挥发性脂肪酸含量的实时检测，优化产气效率[1]。

3.技术局限性分析

近红外光谱技术应用仍面临挑战。光谱数据受样品物理状态（颗粒大小、表面均匀性）与化学状态（水分含量、杂质浓度）影响显著，导致吸收峰位移或强度变化。仪器噪声、光源稳定性及环境温度波动可能引入测量误差。此外，光谱重叠问题导致单一波长吸收峰对应

多种组分，直接解析难度较大。针对上述问题，需通过光谱预处理（如Savitzky–Golay平滑、基线校正）与化学计量学建模（如偏最小二乘回归、主成分分析）提升数据质量，为机器学习算法提供可靠输入。

二、机器学习算法在发酵参数预测中的应用

1.线性模型构建与优化

偏最小二乘回归（PLS）作为经典的线性建模方法，在发酵参数预测领域有着广泛且重要的应用。其核心原理是通过提取光谱数据与目标参数之间的最大协方差方向，从而构建出线性预测模型。在发酵过程中，常常需要同时处理多个变量输入以及多个响应输出的情况，而PLS算法恰好具备这样的能力，能够很好地适用于葡萄糖浓度、pH值等连续参数的预测。然而，发酵过程往往呈现出非线性特性，这使得传统的PLS模型在捕捉复杂关系时存在一定的局限性。为了克服这一问题，核偏最小二乘回归（KPLS）应运而生。KPLS通过引入核函数，将原始的光谱数据映射至高维特征空间，在这个高维空间中，数据之间的关系变得更加线性化，从而增强了模型对复杂关系的捕捉能力。在模型优化过程中，交叉验证是一种常用的方法，通过它来确定主成分数量，避免模型出现过拟合的风险，确保模型具有良好的泛化性能[2]。

2.非线性模型开发与适应性提升

人工神经网络（ANN）是一种模仿生物神经元结构的强大模型，通过构建多层感知机模型来实现对发酵参数的预测。在ANN中，输入层负责接收经过预处理后的光谱数据，这些数据如同生物神经元接收的刺激信号。隐藏层则通过非线性激活函数，如Sigmoid、ReLU等，对输入数据进行特征提取，就像生物神经元对刺激信号进行加工处理。输出层则根据隐藏层提取的特征，预测出目标参数。由于发酵过程具有动态特性，参数会随着时间的推移而发生变化，因此可采用循环神经网络（RNN）及其变体（LSTM、GRU）来处理时序光谱数据，更好地捕捉参数的变化趋势。支持向量机（SVM）则通过寻找最优分类超平面，实现光谱数据与参数类别的映射，在微生物污染等离散事件预测方面表现出色。在模型训练阶段，为了提升模型的泛化能力，需要采用网格搜索与贝叶斯优化算法来调整超参数，如隐藏层节点数、学习率等，使模型能够更好地适应不同的发酵场景。

3.集成学习与模型融合策略

集成学习是一种通过组合多个基学习器的输出来提升预测稳定性的有效方法。随机森林算法通过构建多棵决策树，并采用投票机制来确定最终预测值。在发酵过程中，光谱数据往往会受到各种噪声的干扰，而随机森林算法能够有效地抵抗这种干扰，提高预测的准确性。Adaboost算法则通过动态调整样本权重，聚焦于那些难分类的光谱数据，从而增强模型对异常值的适应性。在模型融合层面，可以采用加权平均或堆叠（Stacking）策略。加权平均策略根据不同模型的重要程度赋予它们不同的权重，然后对预测结果进行加权求和；堆叠策略则更为复杂，它通常将线性模型（如PLS）与非线性模型（如ANN）的预测结果进行综合。

三、基于机器学习的发酵质量稳定性控制体系

1.质量稳定性控制目标与指标体系

发酵质量稳定性控制的核心目标是维持产物的各项质量指标在设定的合理范围内，这些指标包括产物效价、杂质含量以及物理特性（如粒径分布）等。为了实现这一目标，需要建立一套关键控制指标体系。其中，葡萄糖消耗速率标准差能够反映葡萄糖在发酵过程中的消耗波动情况，如果标准差过大，说明葡萄糖的消耗不稳定，可能会影响发酵产物的质量；有机酸积累量波动范围则体现了有机酸在发酵过程中的积累变化，过大的波动可能会对微生物的代谢产生不利影响；微生物密度变异系数则反映了微生物在发酵过程中的生长稳定性，变异系数过大可能意味着发酵过程中存在异常情况。通过实时监测这些指标，可以及时识别发酵过程中的异常状态，如代谢抑制、污染入侵等，并触发相应的调控机制。在构建指标体系时，需要结合历史生产数据和工艺知识，确定各参数的允许波动阈值。例如，在抗生素发酵过程中，如果效价波动超过5%，就需要启动补料策略调整，以确保发酵产物的质量稳定。

2.闭环控制系统设计与实现

闭环控制系统是以机器学习模型为核心构建的一个智能化控制系统。它集成了近红外光谱监测模块、执行机构（如补料泵、温控阀）以及反馈调节算法。在系统运行过程中，近红外光谱监测模块实时采集发酵过程中的光谱数据，这些数据经过预处理后输入到机器学习预测模型中。预测模型根据输入的数据输出参数预测值和控制指令。执行机构根据这些指令调整发酵条件，例如补料泵可以根据指令调整补料速率，温控阀可以根据指令调节发酵温度，从而实现的动态优化。反馈调节算法则采用比例–积分–微分（PID）控制或模型预测控制

（MPC），根据预测误差与历史趋势来调整控制强度。例如，当pH值预测值低于设定值时，系统会自动增加碱液补加速率，以维持微生物的最佳代谢环境，确保发酵过程能够稳定进行。

3.质量稳定性评估与持续改进

质量稳定性评估是通过统计过程控制（SPC）方法来实现的。构建控制图，如X–bar图、R图等，可以直观地监测参数均值与波动范围是否超出控制限。如果参数超出了控制限，就说明发酵过程可能出现了异常情况，需要及时采取措施进行调整。同时，采用过程能力指数（Cp、Cpk）可以量化生产过程满足质量要求的程度，帮助企业了解生产过程的稳定性和能力水平。针对评估过程中发现的稳定性问题，可以通过模型更新与工艺优化来实现持续改进。模型更新采用增量学习策略，定期融入新生产数据，使模型能够适应发酵菌种退化或原料变更等导致的参数分布变化，始终保持良好的预测性能。在工艺优化层面，结合机器学习模型预测结果与实验设计（DOE）方法，调整初始培养基配比或发酵温度等工艺参数，进一步提升发酵质量稳定性，提高产品的质量和产量。

四、技术挑战与未来发展方向

1.当前技术瓶颈分析

当下，机器学习结合近红外光谱用于发酵过程参数预测与质量稳定性控制，仍面临诸多技术瓶颈。光谱数据易受仪器自身噪声、样品异质性干扰，导致数据质量参差不齐，开发更鲁棒的预处理算法迫在眉睫。机器学习模型中，深度神经网络虽预测精度可观，但“黑箱”特性使其缺乏可解释性，难以契合制药行业对工艺透明性的严苛要求。同时，近红外光谱仪及相关在线探头成本高昂，成为中小型发酵企业应用的阻碍。而且，多参数耦合下，单一模型难以全面捕捉发酵动态特性，多模型协同控制策略亟待探索。

2.前沿技术融合趋势

未来，该领域技术发展将围绕多光谱融合、边缘计算与数字孪生展开。拉曼光谱、荧光光谱与近红外光谱技术各有优势，相互融合可构建多维代谢物监测体系，大幅提升对代谢途径的解析能力。边缘计算能将模型部署到发酵现场，实现低延迟实时控制，减轻云端数据传输负担。数字孪生技术通过构建发酵过程虚拟模型，支持离线仿真与工艺优化，有效降低实验成本。例如，结合机器学习模型与代谢通量分析，可预测不同补料策略下的产物产率，为实际生产提供精准指导。

3.产业化应用路径探索

产业化应用面临标准制定、人才培训与生态构建三重挑战。制定近红外光谱在线监测技术标准至关重要，需规范光谱采集、模型验证与设备校准流程，提升行业应用规范性。加强跨学科人才培养，融合发酵工程、光谱分析与机器学习知识，满足智能化生产对复合型人才的需求。构建产学研用协同创新生态，推动技术从实验室走向工业化。例如，通过共建联合实验室，加速机器学习算法在抗生素发酵质量控制中的落地应用，促进产业升级[3]。

结语

本研究创新性地将机器学习与近红外光谱技术深度融合，建立了面向发酵过程的智能预测与质量控制体系。通过构建分层递进的算法架构，有效解决了传统方法在非线性拟合、时序数据处理方面的缺陷，实现了从数据采集到决策执行的全链条自动化。实践验证表明，该系统不仅提升了关键参数预测精度，还通过闭环反馈机制显著增强了工艺抗扰动能力。值得注意的是，研究中提出的增量学习策略与多源信息融合思路，为应对菌种退化、原料波动等实际工况提供了灵活的解决方案。未来工作将着重于开发轻量化边缘计算设备，推动数字孪生技术在工艺仿真中的应用，最终形成可复制推广的行业标杆案例。本成果对于促进我国生物制造业向智能化转型具有重要理论价值与现实意义。

参考文献

[1]张浩，刘振，王玲，等.基于近红外光谱结合机器学习算法检测食用明胶品种溯源的研究[J].河南农业大学学报，2021，55（03）：460-467.

[2]高婧娴，黄扬明，雷春丽，等.机器学习在近红外光谱法判别鲍鱼品种研究中的应用[J].中国农业大学学报，2018，23（09）：166-170.

[3]武小红，蔡培强，武斌，等.基于无监督可能模糊学习矢量量化的近红外光谱生菜品种鉴别研究[J].光谱学与光谱分析，2016，36（03）：711-715.