文章编号：1001-8719(2020)00-0000-00

基于 GBDT 和新型 P-GBDT 算法的催化裂化装置汽油收率寻优模型的构建与应用

王 伟[1]，汪 坤[1]，杨 帆[2]，戴超男[2]，金继民[2]，金宝宝[2]
(1. 中国石化武汉分公司，湖北 武汉 430082；2. 联想大数据智能应用实验室，四川 成都 610041)

摘要：催化裂化装置是一个高度非线性和相互强关联的多变量系统，基于数据挖掘技术的分析方法是优化该工艺过程的一类有力工具。笔者利用某石油化工企业集散控制系统(Distributed control system，DCS)和实验室信息管理系统(Laboratory information management system，LIMS)的工业生产实时数据，分别从指标与汽油收率的正负相关性、工业经验以及模型重要性筛选等方面选取了 182 个关键影响参数，利用梯度提升决策树(GBDT)算法构建催化裂化汽油收率的预测模型，预测相应的汽油产率。基于 GBDT 集成学习框架构建了 P-GBDT 模型，引入了特征扰动和特征权重，增大经验可控参数的权重，解决了普通 GBDT 模型对特征缺乏偏好、经验可控参数特征的权重较小的问题。结果显示，由 P-GBDT 算法构建的汽油收率预测模型预测结果的准确率、R^2、均方根误差等指标相比由 GBDT 算法构建的基准模型的预测结果明显更好，对真实收率的拟合效果更为接近，对优化改进实际可控装置操作条件具有更好的指导意义。

关 键 词：P-GBDT 算法；催化裂化；收率寻优模型；人工智能算法
中图分类号：TE65 **文献标识码**：A **doi**：10.3969/j.issn.1001-8719.2020.01.000

Construction and Analysis of Gasoline Yield Prediction Model for Fluid Catalytic Cracking Unit (FCCU) Based on GBDT and P-GBDT Algorithm

WANG Wei[1], WANG Kun[1], YANG Fan[2], DAI Chaonan[2], JIN Jimin[2], JIN Baobao[2]
(1. *SINOPEC Wuhan Company*, *Wuhan* 430082, *China*; 2. *Data Intelligence Application Laboratry*, *Lenovo Group*, *Chengdu* 610041, *China*)

Abstract: Catalytic cracking unit is a highly nonlinear and strongly correlated system. Currently, the data mining techniques are powerful analytical methods for optimizing this process. Based on the industrial production data collected by the laboratory information management system and the distributed control system, 182 key indicators are selected and a gasoline yield prediction model is built based on gradient-growth decision tree (GBDT algorithm). Afterwards, a P-GBDT model is constructed with reference to the GBDT framework, by introducing feature disturbances and feature weights, and increasing the weight of empirically controllable parameters. The results show that P-GBDT has significant higher accuracy with smaller R^2 value and the root mean square error.

Key words: P-GBDT algorithm; fluidic catalytic cracking; product yield optimization model; artificial intelligence algorithms

催化裂化(FCC)是炼油厂重质油轻质化的主要工艺之一，该工艺在高温和催化剂的作用下将蜡油和重油转化成液化气、汽油、柴油等轻质油品，在炼油工业生产中占有重要的地位[1-5]。由于催化裂化

收稿日期：2018-11-12
基金项目：王伟，男，高级工程师，硕士，从事炼油规划与技术改造工作
通讯联系人：杨帆，男，研究员，硕士，从事大数据及工业智能相关研究工作，E-mail：yangfan24@lenovo.com

系统装置工艺复杂，连续程度高，反应机理复杂，其反应过程和产物收率受原料油性质、反应再生催化剂性质以及操作条件相互影响，是一个高度非线性和相互强关联的系统，难以全面地用传统的数学模型来描述[6-8]。近年来，使用基于数据挖掘技术的分析方法成为解决该类问题的新方向[9-14]。

目前，数据挖掘技术被应用于各种行业，包括国内外金融业[15]、互联网[16]、通信[17]、电子商务[18]等，并呈现快速发展的趋势。得益于日益完善的采集装置及数据存储设备，可以采集到各种原料、催化剂、操作条件等实时数据，这些数据可以有效反映催化裂化的反应过程。因此，相比传统的机理分析方法，可以基于数据分析寻找新的优化方式。通过已有数据建立合理的统计学分析模型，对重要指标以及反应过程进行分析，进一步提高原料利用率与对应产品的产率。

近年来，数据挖掘已逐步被用于催化裂化等工艺的优化。李鹏等[19]使用非线性主成分分析法确定了结焦关键性参数，并结合神经网络等预测方法针对结焦趋势构建了结焦诊断模型。Zahedi 等[20]基于误差反向传播神经网络和径向基神经网络建立了催化重整的预测模型，并使用单变量优化方法对温度和压力等工艺参数进行优化。郝关玉等[21]通过多种软测量建模方法分别对相关指标与产品收率进行分析，并使用最小二乘支持向量机(LSSVM)算法建立催化裂化产品收率的预测模型。目前，现有的优化预测模型多由神经网络、支持向量机(SVM)、决策树等算法构建，且用于模型建立的参数多为基于反应原理与生产经验筛选出的经验影响参数。通过对现有方法的研究和改进方法的尝试，发现在模型中仅使用工业经验已知的关键影响参数进行拟合，得到的拟合结果有一定欠缺。其中可直接调控的参数权重较小，不利于模型的改进和实际生产中的寻优。如果在此基础上加入不可控的参数进行训练，则会出现一定的过拟合现象，结论在测试数据以及实际生产中的表现都有所下降。为了改进方法的不足，增大模型中经验可控参数的权重，使模型的拟合泛化能力增强，需要选择建立其它的模型，对催化裂化装置的产品产率进行拟合。

笔者根据某炼油化工公司催化裂化装置提供的实时过程数据，使用 GBDT 算法建立了基于经验可控指标与重要相关参数的汽油收率预测模型。在该算法的基础上，引入特征权重和采样比率，通过人工调整不同特征在模型中的权重大小以及模型每次迭代时所采用的特征数量，从而调整经验可控指标以及其它指标对模型的影响，优化模型，提高模型的拟合能力和泛化能力。

1　数据预处理

1.1　数据整理格式

笔者以从某石油化工企业的集散控制系统(Distributed control system，DCS)及实验室信息管理系统(Laboratory information management system，LIMS)采集到的数据作为研究对象，进行整理后得到初始数据库。通过 LIMS 系统采集到从 2016 年 8 月 4 日至 2018 年 3 月 20 日的数据样本，分析频次为 1 次/周。LIMS 数据包括原料油和再生催化剂的性质相关数据。DCS 系统装置的采集时间段为 2017 年 10 月 21 日至 2018 年 4 月 25 日共 6 个月，记录频次约为 1 次/15 s，该装置主要采集操作变量和系统物料平衡数据。将采集到的数据按照键值对的格式整理，每条数据由时间戳和指标值 2 个字段构成，分别保存为键和值，并将所有数据按时间戳进行升序排序，方便进一步清洗及计算。

1.2　数据清洗与插值

采集到的数据需要通过数据清洗来保证训练数据的正确性和有效性，以提高模型运算的效率。由于一些客观原因，如装置测量波动、数据采集系统偶发问题或者人为因素等，原始数据可能存在异常情况，如部分数据存在异常值、缺失、重复、不完整、噪音等，此外，还可能存在部分冗余数据。对于这些异常数据，需要根据经验与对催化裂化工艺参数的理解进行清洗。数据清洗的原则遵循以下几点：

(1)剔除数据格式错误的数据。

(2)利用莱特准则判定异常数据，并使用时间临近数据的加权平均值替代异常值；对于缺省值，也看作异常值进行处理。

(3)统一相同时间戳的记录，选择其中合法的数值并取均值。

通过分析清洗后数据的特点，发现 2 个采集系统的数据采集频率差异较大，需要对分析指标的监控采集频率进行统一处理。笔者使用 60 min 作为统一间隔，将采集频率小于 60 min 的数据进行平均处理，在时间间隔内采样并取该间隔内的均值；将采集频率大于 60 min 的指标进行插值处理，合理扩大

已有的参数。根据原始数据的特点，一般由 DCS 装置采集的数据需要进行采样处理，而由 LIMS 装置采集到数据需要进行插值处理。

为了使插值得到的数据曲线更平滑，笔者采用 3 种插值方法结合进行插值：(1)直接使用前一次的测量值插值；(2)线性插值；(3)二阶 B 样条插值。

将采用 3 种插值方法分别插值后的均值作为最终插值的结果。对于测量值前后差异较大的数据而言，3 种插值方法都能有效弥补采样时间段内的缺省值，且处理得到结果值差异不大。将它们的均值作为插值结果，得到的数据稳定性更高，更可能反映出未测量到的趋势。

1.3 特征筛选

经过数据清洗与插值的 LIMS 和 DCS 数据包含近 2000 个分析指标，其中包括可控指标和监控所得参数。直接将所有指标应用于产品收率预测，会增大计算的复杂度，影响模型的可解释性、容易出现过拟合，同时可能会降低重要参数的所得权重。因此，筛选收集到的数据指标中可能影响产品产率的关键指标是一个优化模型特征的过程。此外，进行特征筛选可以有效消除冗余特征，降低训练数据的维度和计算的复杂度，同时提高模型的泛化能力。

筛选模型特征一般从系统工艺流程、催化裂化反应过程以及采集数据自身特点等方面来考虑。笔者将采集数据按照 60 min 的粒度进行时间对齐，并构建相应的特征筛选算法如下所示。

算法：特征筛选算法。

输入：来自 LIMS 和 DCS 的全部分析指标构成的特征总集合 A。

输出：筛选后的有效特征集 F。

(1)从特征总集合 A 中删除部分意义重复或无效的特征。

(2)根据行业经验：分析系统工艺流程、催化裂化反应过程筛选特征，选择经过机理验证的对催化裂化过程有重要影响的因素作为特征指标，称为经验可控指标并构成对应集合 S_1。

(3)依据采集指标与产率的相关性构建 filter 进行过滤：计算各指标与产率的 Pearson 相关系数或传递熵，并以此筛选得到部分模型特征。笔者使用 Pearson 相关性系数计算指标与汽油收率的相关性，选择相关性最高的 M 个指标，作为特征集合 S_2。

(4)合并 S_1 与 S_2 所得模型特征，进行特征处理并作为候选特征集 F。

(5)选择可对输入特征进行权重打分的机器学习模型构建 wrapper 用于筛选特征：

选择随机森林算法构建预测模型，使用已有特征集 F 进行训练，筛选出训练后特征权重最小的 m 个特征，构成待剔除特征集 S_3，并移除 Q 中的经验可控特征，更新后的 Q 为 $Q = Q - S_1$。

对于 Q 中的每个特征 q：

①构建新的候选特征集 $F' = F - q$，使用 K-Folder 交叉验证的方式训练随机森林模型，并与使用特征集 F 交叉验证得到的训练结果进行对比。

②选择训练结果更好的特征集作为新的 F。

③若未遍历完 Q 中的全部特征，则重复①②；否则，执行(6)。

(6)最终得到的特征集合 F 为特征筛选结果。

使用该算法筛选特征，结合工业经验，从操作条件、原料、催化剂等方面可以筛选出对汽油收率有重要影响且直接可控的 10 个经验指标，其中包括原料入口温度控制、提升管出口温度、汽提蒸汽(中)流量和回炼油入提升管流量等。计算指标与汽油产率的 Pearson 相关性进行筛选，当所得相关系数值高，表明其一定程度上能真实反映或逼近产品收率的变化，可能是影响催化裂化系统中汽油收率的关键指标。部分经验可控指标可能与汽油产率的相关性并不显著，需要结合所得的 2 个特征集合共同构成候选特征集。

随机森林模型是将多棵决策树集成学习的一种算法，使用该算法构建特征选择模型稳定性较好，在训练中会对特征和样本进行有放回的随机抽样，多次抽样有利于选择到最优的特征。使用交叉验证筛选用于训练的模型特征，剔除了 64 个特征，其中包括粗汽油入塔-吸收解吸塔、稳定汽油入塔-吸收解吸塔和吸收解吸塔上中段回流流量控制等指标。通过构建模型 wrapper 对特征进行筛选，得到除经验可控指标外 172 个模型重要性较高的特征。

通过特征筛选，最终得到 182 个原始特征指标，可以被应用于特征处理与模型训练。除此之外，为了避免随机波动的影响，增加特征数据的稳定性，笔者通过处理原始特征指标，增加对应的滚动均值作为新的特征。使用简单移动平均的方式计算各特征参数每 4 h 的滚动均值，忽略累计量的影响，并将滚动均值作为新的特征与原有特征数据合并。得到的滚动均值曲线可以有效平滑可能的异常值以及随机突变，同时反映实际数据的变化趋势。通过对

筛选后的数据进行特征平滑，扩增后的总特征数为 364，包括原有特征指标与对应的滚动特征均值。

2　现有模型的评估

为评估现有模型，笔者首先采用均值预测，即采用训练集中所有汽油产率的均值作为未来汽油产率的预测值，并以该模型的预测效果作为现有模型评估的基准；其次，采用 GBDT 算法构建汽油产率预测模型，进行汽油产率预测。

2.1　模型评估指标与寻优方法

仅使用单一评估标准来评价汽油产率预测模型的效果会得到不完全的结论，如仅使用平均绝对误差(Mean absolute error，MAE)，可能多个模型之间的值差异不大，不能够得到明显的判别；同时，也无法反应真实产率与预测值的拟合情况。为了更好的评估回归模型的效果，笔者采用 3 种方式共同评价模型的准确度和拟合效果：准确率(Precision)、决定系数(R^2)和均方根误差(Root mean square error，RMSE)。

根据采集数据的特点，采用式(1)来计算准确率，其中，$\hat{y}_i$ 为汽油产率预测值，y_i 为汽油产率实际值，N 为测试样本的个数。

$$\text{Precision}=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|\hat{y}_i-y_i|}{|y_i|}\right)\times 100\% \quad (1)$$

准确率用来衡量预测值与实际值的整体偏差程度，一定程度上而言，准确率越高，模型的预测值与真实数据的偏差越小。

除此之外，由于真实汽油收率的分布近似正态分布，且较为集中，使用历史汽油收率的均值作为预测值的准确率也可以达到 97.32%。在使用准确率作为评价标准的基础上，还需要去掉收率的均值来评价收率的变化程度，同时评估模型对汽油收率的预测效果。针对真实汽油收率的整体特点，使用决定系数 R^2 作为评估标准，可以很好地反应预测模型的拟合效果。R^2 的表达公式如式(2)所示。

$$R^2=1-\frac{\sum_i^N(y_i-\hat{y}_i)^2}{\sum_i^N(y_i-\bar{y}_i)^2} \quad (2)$$

其中，$\bar{y}_i$ 表示为训练集中汽油收率的均值。R^2 反映了汽油收率预测值与实际值的拟合程度，一般而言，R^2 值越高，模型预测值与实际值的拟合程度越好。

相比 R^2 和准确率，均方根误差表示预测值与实际值误差的平方和与预测次数 N 比值的平方根，能更好地反映用来衡量预测值与实际值之间的偏差，由式(3)计算均方根误差的值。

$$\text{RMSE}=\sqrt{\frac{1}{N}(y_i-\hat{y}_i)^2}\times 100\% \quad (3)$$

均方根误差和准确率主要反映预测值与实际值之间的偏离程度，而 R^2 更倾向反映预测值与实际值之间的线性相关性和拟合程度，并与直接使用均值进行预测的效果进行对比。同时使用上述 3 个评估标准作为指标，可以全面地反映原有模型的预测效果以及寻优模型的改进效果。

2.2　GBDT 模型构建

通过 boosting 方式集成的梯度提升决策树(Gradient boosting decision tree，GBDT)算法是一种迭代的决策树算法，通过采用加法模型(即基函数的线性组合)，以及不断减小训练过程产生的残差来完成数据分类或者回归。在 GBDT 算法的基础上，可以对模型进行改进：参考随机森林的思想，对训练特征进行有放回采样，赋予不同的采样权重以保证经验可调的特征对应采样概率更高。为了对比改进方法的模型效果，使用 GBDT 算法构建模型并训练，将基础模型调优得到的结果作为寻优与改进的基础标准。

笔者采用 GBDT 算法的开源系统实现 lightGBM 回归模型，以预测汽油产率为目标建立回归树模型，并将预测结果作为基准值用于与优化之后的模型效果进行对比。为了保证筛选出的所有特征指标都有合理的数据，截取 2017 年 11 月 4 日至 2018 年 3 月 19 日的数据作为整体数据集，选择 2017 年 11 月 4 日至 2018 年 3 月 12 日的特征数据和汽油收率实际值作为训练集，共 3096 条数据；将其余 2018 年 3 月 13 日到 2018 年 3 月 19 日的数据共 168 条作为测试集，用来验证模型的预测效果和拟合程度。

GBDT 算法可以由式(4)来表示。

$$f_m(x)=\sum_{m=1}^{M}T(x;\theta_m) \quad (4)$$

其中，$T(x;\theta_m)$为拟合残差得到的决策树，θ_m 为树的参数，M 为迭代次数。对于 GBDT 算法而言，每一次都在之前建立决策树损失函数的梯度下降方向上建立新的树。即每轮迭代开始时，计算当前损失函数的负梯度的值，并将其作为残差的估计去拟合一个新的回归树。将每轮迭代训练得到的树加权求和，可以得到最终的模型输出。

2.3　GBDT 模型效果与分析

通过经验设置与局部网格搜索的方法调整

GBDT 回归模型的超参数，并对比其交叉验证的结果，根据 2.1 节中的 3 个评估标准，可以得到使用该算法进行学习的相对较好的预测模型。真实测量的汽油产率可能出现过大或过小的异常值，采用莱特准则处理异常值之后，绘制实际产率数据的曲线，并对比 GBDT 模型对催化裂化汽油产率的预测结果，得到工业实际值与预测值的对比曲线如图 1 所示。由图 1 可知，模型的预测值总体趋势与工业实际值吻合较好；当实际的汽油产率短时间内波动较大时，模型的预测值与汽油产率实际值之间的偏差较大。

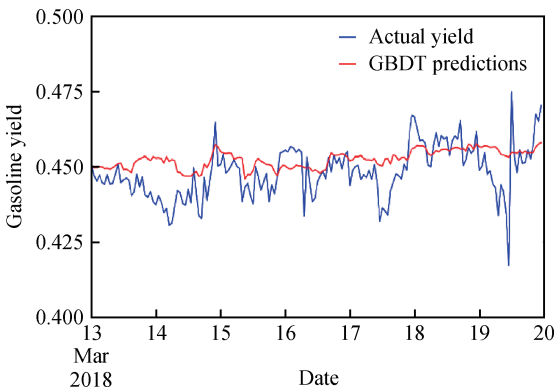

图 1 汽油收率预测值与实际值的对比

Fig. 1 Gasoline yield prediction comparison

对于已得到的较优的模型，其评估结果如表 1 所示。可以看出，预测模型的准确率达到 98.65%，明显高于直接使用均值预测的准确率，验证了该预测模型的可行性和有效性；R^2 为 0.67，表明所选模型特征可以有效解释汽油产率的部分变化趋势，且明显优于直接使用均值进行产率预测的效果；汽油产率预测值与实际值的均方根误差为 0.80%，相对于均值预测，误差明显降低。计算 R^2 时使用训练集的均值作为数据整体均值，更能反映模型对未来数据的预测能力，同时与直接使用均值预测汽油产率的效果进行对比。由准确率和均方根误差可知，基于 GBDT 构建的产率预测模型对汽油产率能够起到较为良好的预测效果，得到的预测值与实际值的整体误差较小。但预测所得的 R^2 较小，且无法准确拟合汽油产率中部分变化较大的波动，说明拟合效果有待提升，对于实际值中变化较大的趋势无法很好的预测，仍然可以进行进一步的改进与优化。

基于 GBDT 回归算法构建的汽油收率预测模型对整体真实汽油收率的预测效果较好，但无法很好地拟合波动较大的趋势；同时经验可控参数特征的权重较小，部分权重较大的特征为依赖监控的其他特征参数，不利于工业控制中的参数寻优。因此，需要尝试新的寻优方法，对汽油产率预测模型进行改进或优化，尝试使经验可控参数的权重更大，同时提升模型的预测效果和泛化能力。

表 1 GBDT 和均值预测结果

Table 1 Results of GBDT model and mean value model

Predictions sets	Precision/%	R^2	RMSE/%
GBDT predictions	98.65	0.67	0.80
Mean value predictions	97.32	0	1.39

在实际的工业生产过程中，不同的装置对催化裂化装置产品收率的影响程度不同，整个生产系统对不同装置的偏好也有所不同。因此，在构建算法模型时，模型也应该具备对不同的特征指标具有不同偏好的特性。针对真实数据的状态，笔者提出了基于 GBDT 的改进算法 P-GBDT 模型。

3 基于 P-GBDT 的汽油收率预测

3.1 P-GBDT 模型构建

基于 GBDT 集成学习框架构建 P-GBDT 算法模型，依据部分特征在实际工艺中的重要性引入特征扰动和特征权重，即采样比率 P 和特征权重 W。引入特征扰动以及依赖权重设置的特征偏向，可保证该类指标在预测模型中的特征重要性更高，并显著增大经验可控指标在模型中训练所得的参数权重。

P-GBDT 的基学习器采用与 GBDT 算法相同的 CART 分类回归树，将训练中损失函数的负梯度作为树的学习目标。该模型的重要超参数包括模型最大迭代次数、采样比率 P 和特征权重 W，需要在训练前进行人为设定。其中，权重 W 为由二维数组构成的矩阵，矩阵最大行数表示模型迭代的最大次数，对应列数表示输入数据的特征个数。权重矩阵 W 的一行表示一次迭代中训练数据的特征权重分布，使不同的特征在迭代采样中根据特征重要性产生对应的采样偏好，同时为模型引入特征扰动。

采样比率 P 表示对特征进行采样的比例。用一维数组构建模型中每次迭代的样本采样比率，该数组的长度表示模型迭代的次数，其中各数值表示该层模型选取的特征数量占全部特征数的比例。每次迭代开始时，同时依据特征的权重大小和采样比率对特征进行采样。

用 T 表示训练数据集及模型的输入，其中

$T=\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_N,y_N)\}$，$x_i\in R^n$，$y_i\in R$；模型迭代最大次数 M，损失函数为 $L(y,f(x))$，P-GBDT 模型的构建过程如 3.3 中所示。

3.2　特征分组及权重设置

通过分析操作装置、反应过程以及真实采集数据，结合业务经验和训练特征的实际物理意义，笔者将筛选出的训练特征初步划分为经验可控特征和普通特征两组特征分组。经验可控特征由经业务经验分析或验证的对产品收率有重要影响的特征组成。当该部分特征指标的参数值发生变化时，仅对产品收率值产生影响，不对其他参与训练的特征指标值或整个生产系统产生显著影响。普通特征由其他经 Pearson 相关性分析和模型筛选得到的特征指标构成，与产品收率存在一定的线性或非线性关系。

依据特征分组对特征权重进行设置，在训练中分别为经验可控特征和普通特征设置一个固定的权重值，且经验特征的权重值大于普通特征的权重值。为经验可控的特征赋予更高的权重，即该组特征在模型的构建过程中重要性更高，可以使其被采样的概率更高，加速模型的收敛；增大经验可控参数的权重，也缓解了 GBDT 模型对训练特征缺乏偏好的问题。

3.3　实验结果与分析

本次实验所使用的数据集与 GBDT 模型相同，其中 2017 年 11 月 04 日至 2018 年 03 月 12 日的数据作为训练集，2018 年 03 月 13 日至 2018 年 03 月 19 日的数据作为验证集。构建 P-GBDT 模型的部分重要训练参数设置如表 2 所示，并使用以下算法步骤构建 P-GBDT 模型：

(1)参数设置：设置模型迭代次数为 M，采样比率为 P，特征权重 W，则在第 m 次迭代中，第 i 个特征被选中的概率 p_i^m 为：

$$p_i^m=\frac{W_i^m}{\sum_1^K W_i^m} \tag{5}$$

其中，K 为特征总数。

(2)模型初始化：给定训练数据，由此训练出第 0 颗树，其表达式如式(6)所示。

$$f_0(x)=c_0=arg\ \min_c\sum_{i=1}^N L(y_i,c) \tag{6}$$

其中，$f_0(x)$ 表示第 0 棵数，$L(\cdot)$ 为损失函数，N 为样本个数，y_i 为第 i 个样本的汽油产率，c 为参数，c_0 为 c 的最优值。

(3)对于当前模型迭代次数 m，$m=1,2,\cdots,M$，每次迭代的步骤如下：

(a)对于训练样本 $i=1,2,\cdots,N$，计算损失函数在当前模型的负梯度值 r_{mi}：

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{7}$$

(b)根据已知的特征权重和特征采样比率，按照式(5)中的概率对特征进行采样，并作为第 m 次迭代的训练特征。

(c)对 r_m 拟合一个回归树，训练得到第 m 棵树，如果该树有 J 个叶子节点，则第 j 个叶节点表示为 R_{mj}，$j=1,2,\cdots,J$。

(d)对 $j=1,2,\cdots,J$，计算线性搜索的最优步长 c_{mj}，即参数 c 的最优值：

$$c_{mj}=arg\ \min_c\sum_{i=1}^N L(y_i,f_{m-1}(x_i)+c) \tag{8}$$

(e)更新模型 $f_m(x)=f_{m-1}(x)+\sum_{j=1}^J c_{mj}I(x\in R_{mj})$，其中 x 为样本。

(4)在 M 次迭代后，得到 P-GBDT 模型：

$$\hat{f}(x)=f_M(x)=f_0(x)+\sum_{m=1}^M\sum_{j=1}^J c_{mj}I(x\in R_{mj}) \tag{9}$$

表 2 中的重要参数主要包括 P-GBDT 模型特征和样本采样所需参数。将每次迭代的样本采样比例统一设置为 0.92，分别设置经验可控特征与其他普通特征的权重均设置为 2.0 和 1.0，通过设置不同的权重从而控制采样概率的差异。使用网格寻优搜索得到其他训练参数的最优取值。对比经过参数调优的 P-GBDT 模型与基础模型的拟合效果，分别绘制测试集的预测结果如图 2 所示。

表 2　重要训练参数及权重设置

Table 2　Typical hyperparameters and weights

Sampling rate	Mass of empirically controllable parameter	Mass of others	Max iterations	Max depth
0.92	2.0	1.0	52	3

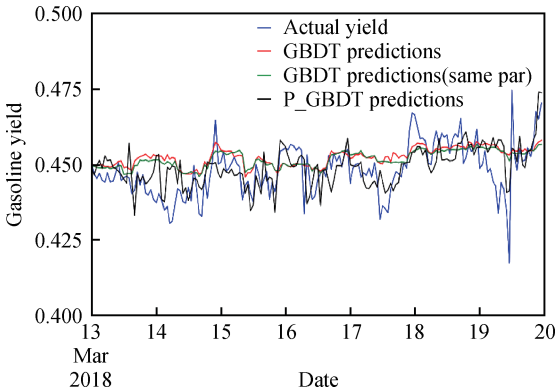

图 2 P-GBDT 模型汽油收率预测对比

Fig. 2 Gasoline yield prediction of GBDT and P-GBDT models

在相同时间段的测试集上，分别使用 P-GBDT 模型和 GBDT 模型对汽油产率进行预测。测试集验证时间较短，汽油收率实际值随着时间的变化相对验证集数据更不稳定，在 3 月 17 日至 19 日之间有明显波折。P-GBDT 模型对汽油收率的预测结果在前期有较小偏离，对 14 日之后的曲线趋势拟合较好，可以真实预测出汽油产率曲线的突变。使用 GBDT 模型的预测结果相对更为平稳，可以拟合出部分变化趋势较小的汽油产率曲线，但无法对汽油收率的突然变化进行预测。对比 2 种模型对汽油收率趋势的拟合，基于 P-GBDT 模型的预测效果显然更优。该模型与其他基础模型在训练集和测试集上的评估标准值如表 3 所示。

表 3 GBDT 和 P-GBDT 回归模型对比结果

Table 3 Results of GBDT and P-GBDT regression models

Predictions set	Precision/%	R^2	RMSE/%
GBDT training	99.67	0.93	0.33
GBDT predictions	98.65	0.67	0.80
P-GBDT training	99.51	0.95	0.29
P-GBDT predictions	98.71	0.71	0.75
GBDT training (with hyperparameters of P-GBDT)	98.60	0.32	1.07
GBDT predictions (with hyperparameters of P-GBDT)	98.54	0.64	0.84
Mean value predictions	97.32	0	1.39

由表 3 可知，在测试集上 P-GBDT 的模型准确率相比 GBDT 提高了 0.06%，可知 P-GBDT 在预测准确性上优于 GBDT。对于 R^2 而言，P-GBDT 模型测试集的 R^2 值在 GBDT 模型的基础上提高了 0.04，其对汽油产率变化趋势的拟合效果明显优于 GBDT 模型。此外，GBDT 和 P-GBDT 在测试集上的均方根误差值分别为 0.80%和 0.75%，误差降低了 0.05 百分点，表明 P-GBDT 模型预测的汽油产率与实际值的偏差更小。结合 2 个模型在测试集上对模型准确率、R^2 和均方根误差的对比结果可知，P-GBDT 模型对汽油收率预测的拟合效果明显优于采用 GBDT 构建的预测模型。

分别对比 2 个模型在训练集上和测试集上的表现，P-GBDT 和 GBDT 模型都存在一定程度上的过拟合。但是 P-GBDT 的过拟合现象明显比 GBDT 更弱，其训练集上的准确率小于 GBDT 在训练集上的准确率，R^2 之间的差值更小。说明 P-GBDT 改进算法不仅可以提升模型预测的性能，对模型过拟合现象也有一定的缓解。

为了对比相同条件下 GBDT 与 P-GBDT 模型的预测效果，利用 P-GBDT 的决策树深度和最大迭代次数训练对应 lightGBM 模型，该模型并非使用网格搜索得到的最优模型，因此，其训练效果比原有 GBDT 参考模型效果更差。使用相同超参数训练 GBDT 及 P-GBDT 回归模型的对比评估标准值如表 3 所示。

从同参数 GBDT 模型的训练集表现可以看出，未使用最佳参数时，GBDT 模型的表现依然优于直接使用平均值对汽油产率的预测。该模型在训练集上的表现很差，在预测集上反而较好，但预测曲线较为平缓，仅能拟合真实汽油产率的部分波动趋势，依然存在较大的误差。训练集与预测集的差异表明，使用与 P-GBDT 相同的超参数训练 GBDT 模型，会出现欠拟合。该模型的表现和 P-GBDT 模型预测结果的对比如图 2 所示。

通过分析并对比 P-GBDT 模型与 GBDT 参考模型的效果，P-GBDT 的各项评估表现其明显优于 GBDT，且能较好地拟合真实汽油产率的波动趋势，缓解了模型构建中的过拟合。证明该 P-GBDT 改进模型是合理可行的，可以提升催化裂化反应中汽油产率的预测性能，并提升经验可控指标在模型中的权重。

4 结 论

基于 LIMS 及 DCS 系统中的工业生产数据，通过分析指标与真实汽油收率的相关性，结合工业经验可控参数以及模型重要性筛选、关联指标的剔除，

选择了 182 个潜在影响催化裂化汽油产率的关键参数作为模型的输入特征，并进行进一步的特征处理。利用 GBDT 算法构建催化裂化汽油收率的预测模型，并将模型对汽油产率的预测效果作为基准值，可以得到以下结论：

(1)基于 GBDT 集成学习框架构建 P-GBDT 模型，加入特征扰动和特征权重，并增大经验可控参数的权重。结果发现，由 P-GBDT 算法构建的汽油收率预测模型预测结果的准确率为 98.71%，R^2 为 0.71，均方根误差为 0.75%，相比由 GBDT 算法构建的基准模型的预测结果明显更好，对真实汽油收率的拟合效果更为接近。

(2)通过对比原 GBDT 模型的预测效果，笔者构建的 P-GBDT 模型能更为精确地预测催化裂化装置中汽油收率，相比于基础模型的拟合效果更优，针对经验可控的重要指标增大其参数权重，解决了 GBDT 模型对特征缺乏偏好，经验可控参数特征的权重较小的问题。增大模型中经验可控指标的权重，对优化改进实际可控装置操作条件具有良好的指导意义。

参 考 文 献

[1] SALVADO F C, TEIXEIRA-DIAS F, WALLEY S M, et al. A review on the strain rate dependency of the dynamic viscoplastic response of FCC metals [J]. Progress in Materials Science, 2017, 88(1): 186-231.

[2] PALOS R, GUTIERREZ A, ARANDES J M, et al. Catalyst used in fluid catalytic cracking (FCC) unit as a support of NiMoP catalyst for light cycle oil hydroprocessing[J]. Fuel, 2018, 216(15): 142-152.

[3] ETIM U J, BAI P, ULLAH R, et al. Vanadium contamination of FCC catalyst: Understanding the destruction and passivation mechanisms [J]. Applied Catalysis A: General, 2018, 555(5): 108-117.

[4] 杨朝合，陈小博，李春义，等. 催化裂化技术面临的挑战与机遇[J]. 中国石油大学学报，2017，41(6)：171-177. (YANG Chaohe, CHEN Xiaobo, LI Chunyi, et al. Challenges and opportunities of fluid catalytic cracking technology[J]. Journal of China University of Petroleum, 2017, 41(6): 171-177.)

[5] 卢春喜，范怡平，刘梦溪，等. 催化裂化反应系统关键装备技术研究进展[J]. 石油学报(石油加工)，2018，34(3)：441-454. (LU Chunxi, FAN Yiping, LIU Mengxi, et al. Advances in key equipment technologies of reaction system in RFCC Unit [J]. Acta Petrolei Sinica (Petroleum Processing Section), 2018, 34(3): 441-454.)

[6] XIONG K, LU C X, WANG Z F, et al. Quantitative correlations of cracking performance with physiochemical properties of FCC catalysts by a novel lump kinetic modelling method[J]. Fuel, 2015, 161(1): 113-119.

[7] AMBLARD B, SINGH R, GBORDZOE E, et al. CFD modeling of the coke combustion in an industrial FCC regenerator[J]. Chemical Engineering Science, 2017, 170(1): 731-742.

[8] 熊凯，卢春喜. 催化裂化(裂解)集总反应动力学模型研究进展[J]. 石油学报(石油加工)，2015，31(2)：293-306. (XIONG Kai, LU Chunxi. Research progresses of lump kinetic model of FCC and catalytic pyrolysis[J]. Acta Petrolei Sinica (Petroleum Processing Section), 2015, 31(2): 293-306.)

[9] ALARADI A A, ROHANI S. Identification and control of a riser-type FCC unit using neural networks [J]. Computers & Chemical Engineering, 2002, 26(3): 401-421.

[10] 周小伟，袁俊，杨伯伦. 应用 BP 神经网络的二次反应清洁汽油辛烷值预测[J]. 西安交通大学学报，2010，44(12)：82-86. (ZHOU Xiaowei, YUAN Jun, YANG Bolun. Prediction of octane number for clean gasoline obtained from secondary reactions based on back-propagation neural network [J]. Journal of Xi'an Jiaotong University, 2010, 44(12): 82-86.)

[11] 苏鑫，吴迎亚，裴华健，等. 大数据技术在过程工业中的应用研究进展[J]. 化工进展，2016，35(6)：1652-1659. (SU Xin, WU Yingya, PEI Huajian, et al. Recent development of the application of big data technology in process industries[J]. Chemical Industry and Engineering Progress, 2016, 35(6): 1652-1659.)

[12] 苏鑫，裴华健，吴迎亚，等. 应用经遗传算法优化的 BP 神经网络预测催化裂化装置焦炭产率[J]. 化工进展，2016，35(2)：389-396. (SU Xin, PEI Huajian, WU Yingya, et al. Predicting coke yield of FCC unit using genetic algorithm optimized BP neural network [J]. Chemical Industry and Engineering Progress, 2016, 35(2): 389-396.)

[13] 陈露. 基于数据挖掘的技术原油评价系统研究[D]. 西安：西安石油大学，2012.

[14] 方伟刚. 数据挖掘技术在催化裂化 MIP 工艺产品分布优化中的应用研究[D]. 上海：华东理工大学，2016.

[15] TIAN X, HAN R, WANG L, et al. Latency critical big data computing in finance[J]. The Journal of Finance and Data Science, 2015, 1(1): 33-41.

[16] GE M, BANGUI H, BUHNOVA B. Big data for internet of things: A survey[J]. Future Generation Computer Systems, 2018, 87(1): 601-614.

[17] RATHORE M M, PAUL A, AHMAD A, et al. Real-time secure communication for Smart city in high-speed big data environment[J]. Future Generation Computer Systems, 2018, 83: 638-652.

[18] QUAN S F. The effect of big data processing on the development of e-commerce in cloud computing environment[J]. Computer Knowledge and Technology, 2013, 19(20): 4762-4770.

[19] 李鹏, 郑晓军, 明梁, 等. 大数据技术在催化裂化装置运行分析中的应用[J]. 化工进展, 2016, 35(3): 665-670. (LI Peng, ZHENG Xiaojun, MING Liang, et al. Application of big data technology in operation analysis of catalytic cracking [J]. Chemical Industry and Engineering Progress, 2016, 35(3): 665-670.)

[20] ZAHEDI G, MOHAMMADZADEH S, MORADI G. Enhancing gasoline production in an industrial catalytic-reforming unit using artificial neural networks [J]. Energy & Fuels, 2008, 22(4): 2671-2677.

[21] HAO Guangyu, HU Jinhua. Application of soft measurement technology in prediction of catalytic cracking product yield[C]//The 13th Annual Meeting of China Petroleum and Chemical Industry Automation, 2014.