文章编号：1001-8719(2019)04-0807-11

基于人工智能算法的催化裂化装置汽油收率预测模型的构建与分析

杨 帆[1]，周 敏[2]，戴超男[1]，曹 军[3]

(1. 联想数据智能应用实验室，四川 成都 610041；2. 四川理工学院 过程装备与控制工程四川省高校重点实验室，四川 自贡 643000；3. 华东理工大学 机械与动力工程学院，上海 200237)

摘要：基于某石化企业的 LIMS(Laboratory information management system)及 DCS(Distributed control system)系统中的工业生产数据，结合工业经验中已知的影响催化裂化产品收率的重要因素，通过分析监控指标与实际汽油收率的相关性，筛选出与汽油收率的正负相关性较高的分析指标。在此基础上，基于梯度提升决策树 GBDT 算法构建了催化裂化汽油收率的预测模型，并预测了相应的汽油产率。结果表明：由 GBDT 算法构建的汽油收率预测模型预测结果的准确率为 98.9%，R^2 系数为 0.236，平均绝对误差为 0.531%；模型预测结果与实际汽油产率相比，误差率小于 1%，表明构建的模型能精确预测催化裂化装置中汽油等产品收率，有助于在实际生产中优化催化裂化装置的操作条件，从而进一步提升催化裂化装置的经济性能。

关 键 词：人工智能；催化裂化；预测模型；GBDT 算法

中图分类号：TE65 **文献标识码**：A **doi**：10.3969/j.issn.1001-8719.2019.04.024

Construction and Analysis of Gasoline Yield Prediction Model for FCC Unit Based on Artificial Intelligence Algorithm

YANG Fan[1], ZHOU Min[2], DAI Chaonan[1], CAO Jun[3]

(1. *Data Intelligence Application Lab, Lenovo Group, Chengdu* 610041, *China*;
2. *Sichuan Provincial Key Lab of Process Equipment and Control, Sichuan University of Science and Engineering, Zigong* 643000, *China*;
3. *School of Mechanical and Power Engineering, East China University of Science and Technology, Shanghai* 200237, *China*)

Abstract: Industrial data were collected from a petrochemical company's DCS (Distributed control system) and LIMS (Laboratory information management system) systems. Together with the essential factors affecting the product yield of catalytic cracking from industrial experience, indicators with high correlations were filtered out by analyzing their correlations with actual gasoline yield. Subsequently, a prediction model based on gradient-growth decision tree (GBDT algorithm) was constructed to predict the gasoline yield on the catalytic cracking unit. The results show that the accuracy of the GBDT gasoline yield prediction model is 98.9%, the R^2 value is 0.236, and the mean absolute error is 0.531%. Compared to actual gasoline yield, the error of prediction result is less than 1%, indicating that the presented model could predict the product yield such as gasoline on the catalytic cracking unit accurately. This will optimize the operation conditions of FCC, and further enhance the economic performance of FCC unit.

收稿日期：2018-06-11

基金项目：上海市自然科学基金项目(18ZR1409000)和过程装备与控制工程四川省高校重点实验室开放基金项目(GK201818)资助

第一作者：杨帆，男，硕士，从事大数据及工业智能相关研究工作

通讯联系人：曹军，男，助理研究员，博士，从事石油化工及工业大数据相关研究工作，Tel：021-64253810，E-mail：caojun@ecust.edu.cn

Key words: artificial intelligence; fluidic catalytic cracking; prediction model; GBDT algorithm

催化裂化是重质油在酸性催化剂存在下，在500 ℃左右、$1\times10^5\sim3\times10^5$ Pa的压力下发生以裂化反应为主的一系列化学反应生成轻质油、气体和焦炭的过程。目前，我国催化裂化装置生产的柴油和汽油约占成品柴油和汽油总量的30%和70%左右，已经成为重油加工的最重要方法之一[1-5]。催化裂化的工艺过程和产品收率优化的建模分析一直是石油加工领域研究的热点和难点。目前常用的催化裂化过程建模方法有机理建模法[6-8]和统计建模法[9-11]。由于催化裂化是一个高度非线性和相互强关联的系统，其中原料油性质、反应再生催化剂性质，以及反应操作工况条件等因素都会影响到反应过程和产物收率，使用传统的机理模型很难全面地去描述，大数据技术则是解决这一问题的有力工具。

目前，大数据技术正处于其应用的高速发展期，并已经在电子商务[12]、电力[13]、航空[14]以及医疗[15]等领域取得了巨大的成功。随着石化行业生产过程的自动化控制水平日益提高和工艺流程控制系统的不断完善，各种原料数据、催化剂性能数据以及操作工况参数等都能从装置的数据库平台中实时采集。这些数据记录了反应过程的特征、性能和变化，反映了反应过程的本质，大大改进了原来数据收集的不完整。已经积累的海量过程数据为数据挖掘技术在石化领域的应用提供了良好的基础条件。将数据挖掘技术应用于石化反应过程，建立完善的统计学分析模型，可缩短新工艺的开发研究周期、优化工程设计方案、优化装置操作和实现装置的在线优化，多角度全方位地对反应过程及其影响机制进行分析，从而可进一步提高原料利用率和所需产品的产率，具有传统机理分析优化方法无法比拟的优势[16]。这一优势在催化裂化工艺操作工况的优化和产品收率的预测方面体现的更为明显。

目前，已有研究者将神经网络、支持向量机等人工智能算法应用于优化催化裂化工艺。Zahedi等[17]使用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，并采用单变量优化方法优化了温度和压力等工艺参数，使汽油收率从80%增加到82%。李鹏等[18]在中国石化开发的炼油技术分析与远程诊断平台上，运用大数据数据处理技术和积累的海量的催化运行数据进行数据挖掘与分析，对催化裂化装置报警、结焦等问题进行深入探索研究与分析，解决了催化裂化装置报警问题、结焦问题和收率问题，从而进一步提升了催化裂化装置运行水平。陈露[19]通过整理大量原油评价数据，建立了原油性质和催化裂化反应产物分布数据之间的模型，并采用化学计量学校正了该模型，结果表明所建立的原油评价模型具有较好的适用性。孔金生等[20]对催化裂化数据进行了预处理并建立了粗汽油干点的神经网络模型，结果证明该模型具有可靠性。方伟刚[21]以中国石化九江分公司催化裂化装置提供的实时过程数据为基础，进行了产品收率优化的研究，建立了合适的原料油性质聚类模型和产品收率神经网络模型，并使用优化算法对操作条件进行了优化。

笔者以某炼化公司催化裂化装置提供的实时过程数据为基础，建立了合适的原料油性质聚类模型和产品收率预测模型，然后使用优化算法对操作条件进行优化。计算分析的结果有助于进一步提升催化裂化装置的汽油收率，进而增加企业经济效益，并为工业操作提供可靠的技术支持。

1 某炼化公司催化裂化装置实时过程数据预处理

笔者使用的数据均采集自某石化企业的LIMS(Laboratory information management system)及DCS(Distributed control system)系统。通过LIMS系统可采集到原料油和再生催化剂性质的相关数据，其分析频次为1次/周。为了采集到足够多的样本，LIMS数据采集时间段从2016年8月4日至2018年3月20日共近2年。通过DCS系统可采集到操作变量和系统物料平衡数据，每隔15 s记录1次，装置数据采集时间段从2017年10月21日至2018年4月25日共6个月。将DCS和LIMS的数据按“时间戳[分割符]指标值”的格式整理，每条数据由时间戳和指标值两个字段构成。为了方便进一步清洗以及计算，将所有数据按时间戳升序进行排序。

1.1 数据清洗

由于一些客观的装置条件以及人为因素，例如数据采集系统出现问题、数据存储/传输过程中发生

错误等，采集到的数据可能存在部分数据缺失、重复、不完整、噪音、异常等情况；除此之外，原始数据中还存在部分冗余数据。为了保证训练数据的正确性和有效性，同时提高模型运算的效率，需要对数据进行清洗。对于不同类型的数据异常，相应的清洗方法如下：

①数据格式错误。每条数据必须满足给定的格式，其中第一个字段是 timestamp 类型，第二个字段是 float 类型。可以直接删除格式错误的数据。

②数据值异常。计算每个指标的均值 μ 和标准差 σ，使用莱特准则，将($\mu\pm3\sigma$)范围外的数据定义为异常值，并使用时间临近的数据做加权平均作为替代。

③数据重复。同一指标的数据中可能存在多条相同的时间戳记录，需要选择其中合法的数值并取均值。

④数据缺失。缺少某些应有时刻的数据，可以将缺省数据点看作异常值，使用时间临近的数据做加权平均。

1.2　数据时间对齐

针对不同的分析指标，其监控采集频率可能不同。为了便于数据分析与计算，需要将这些数据在时间上对齐：即对所有分析指标使用统一的时间间隔。通过分析数据的特点，笔者将 60 min 作为参考时间间隔。对于采集时间间隔小于 60 min 的数据，需要做采样处理，通过设置 60 min 大小的时间窗口，取该时间窗口里的数值的均值。对于采集时间间隔大于 60 min 的参数，需要做插值处理。根据不同系统的采集频率，一般 DCS 系统采集到的数据需要进行采样处理，而由 LIMS 系统收集到的数据则需要进行插值处理。笔者采用 3 种插值方法：

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。

以监测点指标数据“混合原料密度(20 ℃)”在 2016 年 11 月至 2017 年 3 月 30 日之间的数据为例，该数据的原始分布如图 1 所示。

按照以上思路分别使用 3 种插值方法计算，结果数据如图 2、3 及 4 所示。其中：图 2 为直接使用前一次的测量值插值所得数据；图 3 为线性插值处理结果；图 4 为二阶 B 样条插值处理结果。3 种插值方法都可以弥补时间间隔内的缺省值，且处理得到的结果相似。

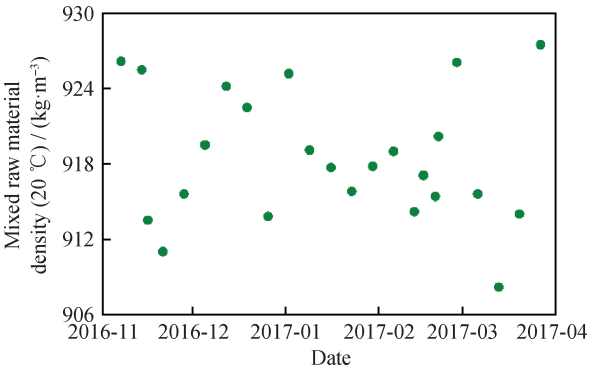

图 1　某炼化公司 2016 年 11 月至 2017 年 4 月之间混合原料密度(20 ℃)未插值处理的数据

Fig. 1　Mixed raw material density data (20 ℃) without interpolation between November 2016 and April 2017 from a refining and chemical company

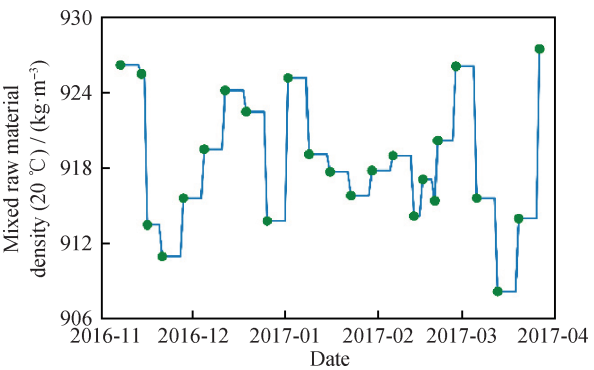

图 2　对图 1 数据直接使用前一次的测量值后处理结果

Fig. 2　Results of processing before nearest value interpolation for data of Fig. 1

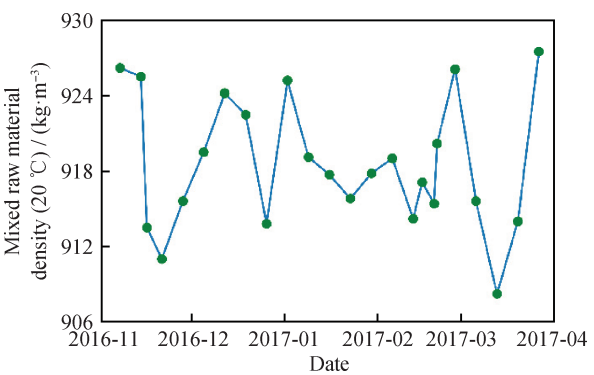

图 3　对图 1 数据线性差值处理结果

Fig. 3　Results of processing linear interpolation for data of Fig. 1

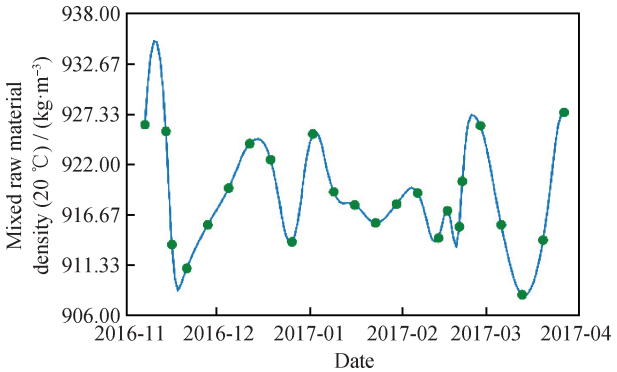

图 4 对图 1 数据二阶 B 样条插值处理结果

Fig. 4 Results of processing quadratic B-spline interpolation for data of Fig. 1

2 针对已处理的某炼化公司数据构建预测模型

2.1 预测模型的选用

利用机器学习模型预测产品收率的方法已经被一些文献提及或使用，其中较大比例采用神经网络构建模型[20]。神经网络的优点在于拟合能力非常强，理论上能逼近任意非线性映射，且自学习与自适应性强。另一方面，由于在神经网络模型中有较多超参数需要确认，往往需要较长时间的反复调参，才能取得好的效果；同时，其可解释性较差，不利于研究输入向量各分量之间以及它们与输出的相关性。

根据以上分析，笔者决定采用另一种拟合能力较强的模型：由多棵决策树构成的集成学习模型。树模型对于真实分布的拟合效果较好，具有一定的可解释性，且可以用于特征筛选。多决策树模型的典型代表是使用 bagging 方式集成的随机森林模型和通过 boosting 方式集成的梯度提升决策树(Gradient boosting decision tree，GBDT)模型。通过对比这两种模型在实际数据集上的应用，发现 GBDT 在产品收率上的表现更好。因此，本研究选择使用 GBDT 构建预测模型。

GBDT 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)，不断减小训练过程产生的残差来完成数据分类或者回归。在训练过程中，每轮迭代开始时，计算损失函数的负梯度在当前模型的值，将其作为残差的估计去拟合一个回归树；每次迭代都会生成一颗新的决策树，将每轮训练得到的树加权求和，可以得到输出的最终模型。GBDT 的主要特点，即是通过在每轮训练中让损失函数尽可能快地减小，以便尽快地收敛达到局部最优解或者全局最优解。

2.2 模型特征的选取

筛选 DCS 与 LIMS 系统中与汽油产率正负相关性较强的指标是一个优化模型特征选择的过程。它们的测量值变化与汽油收率线性相关系数较高，一定程度上可以更好地反映或逼近真实收率的变化趋势，可能是影响汽油收率的关键指标。在模型的构建中，考虑将这些潜在的关键指标作为特征，可以有效降低训练数据集的维度，同时提高模型的学习性能。

采集到的 LIMS 和 DCS 数据中包括近 2000 个分析指标，其中大部分不适用于产品收率预测，因此需要对已有的分析指标进行筛选。首先，使用大数据分析的方法，筛选出与产品收率相关性较高的指标。使用 Pearson 系数作为考察相关性的依据，将采集到的指标与汽油收率按 60 min 的时间粒度，依照 2.2 节中介绍的方法进行时间对齐后，计算 Pearson 相关系数。其中，DCS 数据中与汽油收率正相关性较高的 27 个指标如表 1 所示，负相关性较高的 22 个指标如表 2 所示。

对于 LIMS 系统中的数据，由于无法明确 LIMS 数据中各个指标在监测间隔里的变化过程，基于现有数据无法比较 3 种不同插值方法的优劣。考虑到仅采用一种插值方法得到的数据可能有一定偏差，笔者选择同时使用 3 种插值方法分别处理数据。LIMS 系统中计算得到的正相关性较强的 21 个指标如表 3 所示，负相关性较强的 25 个指标如表 4 所示。

除此之外，由于催化裂化反应的特性，可以从工业经验角度考虑，筛选出影响产品收率的关键指标作为模型特征。根据催化裂化产品的生成过程，并结合工业经验，可以筛选出部分线性相关性不高，却对产品收率有重要影响的经验指标，包括提升管反应器出口温度、原料中饱和烃与胶质含量、汽提蒸汽流量、催化剂活性指数等[20]。将这些重要指标作为参照指标，计算其与产品收率的相关性作为参照相关性，用于从以上正负相关性较强的指标中筛选出相关性大于或接近参照相关性的指标。作为参考的重要因素的相关性如表 5、6 所示。

经过筛选的指标的因变量需要进行人工去除。整理筛选出的正负相关性较强指标，并结合工业经

表 1　DCS 数据中与汽油收率正相关性较高的指标

Table 1　Factors with high positive correlations for gasoline yield in DCS data

Index name	Correlation coefficient
Flow rate of catalytic gasoline to S-Zorb catalyst/($m^3 \cdot h^{-1}$)	0.69
Crude gasoline mass flow rate control/($t \cdot h^{-1}$)	0.36
Crude gasoline mass flow rate into the tower/($t \cdot h^{-1}$)	0.35
Prefractionation flow control of stable gasoline mass flow rate into the tower/($t \cdot h^{-1}$)	0.32
Mass flow rate of central circulating water/($t \cdot h^{-1}$)	0.29
Interface control/%	0.28
Oil mass concentration in water/($mg \cdot L^{-1}$)	0.28
Stripping section density/($kg \cdot m^{-3}$)	0.28
Liquid level control/%	0.26
Circulating water mass flow rate/($t \cdot h^{-1}$)	0.25
Mass flow rate of vacuum wax oil/($t \cdot h^{-1}$)	0.24
Mass flow rate of reuse med-pressure steam/($t \cdot h^{-1}$)	0.24
Burning oil mass flow rate of furnace/($t \cdot h^{-1}$)	0.24
Cumulative mass flow rate control of light gasoline/($t \cdot h^{-1}$)	0.24
Oxidation air volume flow rate of tower/($m^3 \cdot h^{-1}$)	0.24
Gas volume accumulation in old CO furnace/m^3	0.24
Coking liquid hydrocarbon mass flow rate into tower/($kg \cdot h^{-1}$)	0.24
Reforming cumulant mass flow of medium pressure steam from new CO furnace/t	0.24
Mass flow rate of circulating water of CO plant pipeline/($t \cdot h^{-1}$)	0.24
Compensation mass flow of steam temperature and pressure in catalytic line/($t \cdot h^{-1}$)	0.24
Cumulative steam mass flow control/t	0.24
Cumulant mass flow of atmospheric (coking) wax oil into riser tube/t	0.24
Cumulant mass flow of heavy oil flow control/t	0.24
Nitrogen volume flow rate/($m^3 \cdot h^{-1}$)	0.24
Softened water mass flow rate/($t \cdot h^{-1}$)	0.24
Cumulant of mid-pressure steam mass flow into pipe network from old CO furnace/t	0.24
Demercaptan fresh water mass flow rate/($t \cdot h^{-1}$)	0.24

表 2　DCS 数据中与汽油收率负相关性较高的指标

Table 2　Factors with high negative correlations for gasoline yield in DCS data

Index name	Correlation coefficient
Catalytic conversion-steam mass flow rate/($t \cdot h^{-1}$)	−0.28
Circulating water volume flow rate in demercaptan zone/($m^3 \cdot h^{-1}$)	−0.27
Flash vapor volume flow rate/($m^3 \cdot h^{-1}$)	−0.26
Oxidizing wind pressure of tower/MPa	−0.26
Temperature of steam to deaerator/℃	−0.26
Mass flow rate of circulating water into device/($t \cdot h^{-1}$)	−0.26
Upper density of settler/($kg \cdot m^{-3}$)	−0.25
Boundary indication of return tank/%	−0.25
Temperature of tank/℃	−0.24
De-temperature water mass flow rate of furnace/($t \cdot h^{-1}$)	−0.24
Steam temperature and pressure compensation mass flow rate of catalytic line/($t \cdot h^{-1}$)	−0.24
Transient gas volume flow rate of old CO furnace/($m^3 \cdot h^{-1}$)	−0.24
Re-use mid-pressure steam mass flow rate/($t \cdot h^{-1}$)	−0.24
Material level of regenerate degassing tank/t	−0.23
Feed temperature control of tower/℃	−0.22
Feed temperature of tower/℃	−0.22
Density of middle riser tube/($kg \cdot m^{-3}$)	−0.22
Butterfly valve manual control/%	−0.22
Unqualified gasoline inlet mass flow rate/($t \cdot h^{-1}$)	−0.21
Pressure difference of lubricating oil filter/MPa	−0.21
Total steam temperature/℃	−0.21
Density of semi-regenerative oblique tube/($kg \cdot m^{-3}$)	−0.20

表 3 LIMS 数据中与汽油收率正相关性较高的指标

Table 3 Factors with high positive correlations for gasoline yield in LIMS data

Index name	Specific surface area of equilibrium catalyst/($m^2 \cdot g^{-1}$)	Abundant gas mass fraction/%	Temperature of FCCU-10% stable gasoline/℃	Equilibrium catalyst-micro reactivity index
Prevalue interpolation	0.53	0.21	0.53	0.29
Linear interpolation	0.53	0.50	0.43	0.32
Spline interpolation	0.53	0.43	0	0.31
Average value	0.53	0.38	0.31	0.31

Index name	Slurry-solid mass concentration/($g \cdot L^{-1}$)	Safety gas-hydrogen mass fraction/%	Initial boiling point of FCCU stable gasoline/℃	Light diesel oil-sulfur mass fraction/($\mu g \cdot g^{-1}$)
Prevalue interpolation	0.53	0.21	0.42	0.13
Linear interpolation	0.53	0.30	0.41	0.32
Spline interpolation	0.53	0.30	−0.03	0.30
Average value	0.53	0.27	0.26	0.25

Index name	Turbidity of circulating water of low temperature heat source/($kg \cdot m^{-3}$)	Density (20 ℃) of light diesel oil/($kg \cdot m^{-3}$)	Raw hydrocarbon-*trans*-2-butene mass fraction/%	Semi-regenerated catalyst-carbon mass fraction/%
Prevalue interpolation	0.24	0.23	0.20	0.19
Linear interpolation	0.26	0.18	0.15	0.26
Spline interpolation	0.25	0.33	0.38	0.27
Average value	0.25	0.25	0.24	0.24

Index name	Raw hydrocarbon-*cis*-2-butene mass fraction/%	Raw hydrocarbon-*iso*-pentane mass fraction/%	Light diesel oil-95% temperature/℃	Stable gasoline of FCCU-10% recovery temperature/℃
Prevalue interpolation	0.20	0.19	0.24	0.22
Linear interpolation	0.14	0.26	0.27	0.47
Spline interpolation	0.38	0.33	0.13	−0.06
Average value	0.24	0.23	0.21	0.21

Index name	Mixed raw material-residual carbon mass fraction/%	Light diesel oil-90% temperature/℃	Turbine oil flash point (opening)/℃	Light diesel oil-10% temperature (on line)/℃
Prevalue interpolation	0.15	0.21	0.24	0.05
Linear interpolation	0.24	0.25	0.19	0.27
Spline interpolation	0.24	0.15	0.17	0.28
Average value	0.21	0.20	0.20	0.20

Index name	Mixed raw material-asphaltene mass fraction/%
Prevalue interpolation	0.22
Linear interpolation	0.18
Spline interpolation	0.18
Average value	0.19

表 4　LIMS 数据中与汽油收率负相关性较高的指标

Table 4　Factors with high negative correlations for gasoline yield in LIMS data

Index name	Sulfide mass concentration/(mg·L^{-1})	Equilibrium catalyst-vanadium mass fraction/(μg·g^{-1})	Ethanolamine lean solution-thermostable salt mass fraction/%	Equilibrium catalyst-nickel mass fraction/(μg·g^{-1})
Prevalue interpolation	−0.56	−0.53	−0.51	−0.48
Linear interpolation	−0.57	−0.52	−0.52	−0.49
Spline interpolation	−0.57	−0.53	−0.52	−0.49
Average value	−0.56	−0.53	−0.52	−0.48

Index name	Equilibrium catalyst-iron mass fraction/(μg·g^{-1})	Mixed raw materials-vanadium mass fraction/(μg·g^{-1})	Raw hydrocarbon-fused butane mass fraction/%	After-dry specific gravity(calculation)
Prevalue interpolation	−0.38	−0.41	−0.29	−0.32
Linear interpolation	−0.42	−0.36	−0.32	−0.30
Spline interpolation	−0.43	−0.33	−0.33	−0.28
Average value	−0.41	−0.36	−0.31	−0.30

Index name	Equilibrium catalyst nickel mass concentration/(mg·m^{-3})	Raw hydrocarbon-maleic mass fraction/%	Mixed raw material-end distillation point/℃	Rich gas *n*-pentane mass fraction/%
Prevalue interpolation	−0.27	−0.24	−0.27	−0.36
Linear interpolation	−0.30	−0.28	−0.28	−0.23
Spline interpolation	−0.29	−0.33	−0.29	−0.22
Average value	−0.29	−0.28	−0.28	−0.27

Index name	Equilibrium catalyst-pore volume/(mL·g^{-1})	Alkali liquor-disulfide mass fraction/(μg·g^{-1})	Mixed raw materials-nitrogen mass fraction/(μg·g^{-1})	Before-dry hydrogen sulfide mass concentration/(mg·m^{-3})
Prevalue interpolation	−0.29	−0.32	−0.28	−0.21
Linear interpolation	−0.25	−0.28	−0.21	−0.25
Spline interpolation	−0.25	−0.16	−0.26	−0.27
Average value	−0.26	−0.25	−0.25	−0.24

Index name	Condensate-chloride ion mass concentration/(mg·L^{-1})	Crude gasoline-10% recovery temperature/℃	Alkali liquor-alkali mass fraction/%	Mixed raw materials-sodium mass fraction/(μg·g^{-1})
Prevalue interpolation	−0.19	−0.15	−0.22	−0.20
Linear interpolation	−0.27	−0.14	−0.22	−0.21
Spline interpolation	−0.25	−0.32	−0.18	−0.20
Average value	−0.24	−0.21	−0.21	−0.20

Index name	Stable gasoline-initial distillation point/℃	Rich gas-hydrogen mass fraction/%	Distillation volume of mixed raw material at 350 ℃/mL	Sulphur containing sewage mass concentration CODcr/(mg·L^{-1})
Prevalue interpolation	−0.27	−0.18	−0.20	−0.13
Linear interpolation	−0.22	−0.20	−0.23	−0.22
Spline interpolation	−0.11	−0.21	−0.16	−0.22
Average value	−0.20	−0.20	−0.20	−0.19

Index name	Nickel mass fraction in mixed raw material/(μg·g^{-1})
Prevalue interpolation	−0.21
Linear interpolation	−0.21
Spline interpolation	−0.09
Average value	−0.17

表 5 DCS 数据中参照指标与汽油收率的相关性

Table 5 The correlations between reference factors and gasoline yield in DCS data

Index name	Correlation coefficient
Riser outlet temperature/℃	0.01
Settler top pressure/MPa	0.04
Front pressure of slide valve in second regenerator/MPa	−0.06
Pressure in front of regeneration valve/MPa	−0.02
Pressure of stand by control valve/MPa	0.01
Mass flow rate of refining oil into riser/(t·h^{-1})	0.23
Mass flow rate of refining oil into tower/(t·h^{-1})	0.19
Stripping steam (middle) mass flow rate/(t·h^{-1})	−0.07
Accumulated steam flow rate at lower part of reactor/(t·h^{-1})	0.05
Mass flow rate control of stripping steam(middle)/(t·h^{-1})	0.14
Accumulated steam mass flow of stripping in reactor/t	0.05
Mass flow rate control of stripping steam (upper)/(t·h^{-1})	0.15
Accumulated mass flow of stripping steam in the upper part of reactor/t	0.05
Pre-lifting steam mass flow accumulation/t	0.17
Pre-lifting steam mass flow rate/(t·h^{-1})	0.18

表 6 LIMS 数据中参照指标与汽油收率的相关性

Table 6 The correlations between reference factors and gasoline yield in LIMS data

Index name	Density (20 ℃) of mixed raw material/(kg·m^{-3})	Mixed raw material-saturated hydrocarbon mass fraction/%	Mixed raw material-aromatics mass fraction/%	Mixed raw material-glue mass fraction/%
Prevalue interpolation	−0.05	−0.11	0.14	0.05
Linear interpolation	−0.02	−0.12	0.13	0.17
Spline interpolation	−0.03	0.08	0.21	−0.12
Average value	−0.03	−0.05	0.16	0.03

Index name	Vanadium mass fraction in mixed raw materials/(μg·g^{-1})	Spent catalyst-carbon mass fraction/%
Prevalue interpolation	−0.41	−0.02
Linear interpolation	−0.36	0.01
Spline interpolation	−0.33	0.01
Average value	−0.36	0

验参考指标，共同作为候选原始特征。通过对所得的原始特征做尺度变换、多项式交叉、差分等特征工程处理，得到可以应用于 GBDT 算法模型的新的特征。

2.3 预测模型的构建

目前普遍用于构建产品收率预测模型的算法都为神经网络算法，少有研究使用树类模型对收率进行预测。相比之下，树类模型的可解释性与对模型特征的筛选作用，使得其在解释特征在模型中的重要性与工业优化方面更有潜力与优势。笔者选择树类模型中的 GBDT 算法构建预测模型，模型构建的框架如下所示，其中模型输入为训练数据集

$\boldsymbol{T}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\in\boldsymbol{R}^n$，$y_i\in\boldsymbol{R}$，迭代的次数为 t，损失函数为 $L(y,f(x))$，输出 GBDT 模型：

(1)特征选择：根据特征的权重 w 从特征集中抽取 p 比例的特征。

(2)初始化基学习器：

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

(3)对于迭代次数 $t=1,2,\cdots,T$：

a. 对训练样本 $i=1,2,\cdots,N$，计算负梯度 (r_{ti})：

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{2}$$

b. 利用 r_t 拟合 1 个回归树，得到第 t 棵回归树 (Tree_t)，并对于叶子区域计算最佳拟合值。

c. 更新模型 $f_t(x)=f_{t-1}(x)+\mathrm{Tree}_t$。

(4)得到模型：

$$\hat{f}(x)=f_T(x) \tag{3}$$

GBDT 算法的主要特点在于在训练中将损失函数的负梯度在当前模型的值作为残差估计，并利用线性搜索估计回归树叶结点区域的值，使损失函数最小化，从而更新回归树并得到最终的模型。它的每一次迭代都会在残差减少的梯度方上建立新模型，因此 GBDT 算法会更关注梯度比较大的样本。

笔者采用 GBDT 模型的开源模块实现 lightGBM 回归方法进行学习。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 10 月 21 日至 2018 年 3 月 20 日的数据作为整体数据集，选择前 4 个月的特征数据和实测收率值作为训练样本，剩余的数据作为预测样本用以验证模型的准确性。通过经验与局部网格搜索的方式调整其超参数并对比其交叉验证的结果，最终使用平均绝对误差 MAE 作为目标函数进行训练，设置回归树棵数为 106，对应学习率为 0.065，其余参数使用默认数值。其中，可由式(4)计算 MAE：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \tag{4}$$

使用现有采集监控数据，对相对于训练集的未来时间节点进行预测，并对比预测值与真实值，可以有效检验构建模型的拟合程度。如果需要对真实未来的产品收率进行预测，同样需要对相同时间段内的指标数据进行采集。

3　结果与讨论

为了评估回归模型的效果，采用式(5)所示方法确定测试集准确率。其中，$\hat{y}_i$ 表示预测结果；y_i 表示真实值；N 表示测试样本的个数。

$$\mathrm{Pre}=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|\hat{y}_i-y_i|}{|y_i|}\right) \tag{5}$$

2016 年 9 月至 2017 年 11 月之间真实汽油收率和处理掉其中的异常值后汽油收率的分布如图 5 所示，图 6 为图 5(b)的数值分布。由图 6 可以看出，真实汽油收率主要分布在 47%左右，基本呈现左右平衡的态势，近似正态分布，且分布相对较为集中，说明汽油收率的整体输出范围较小。在这种情况下，即使使用汽油收率的均值来进行预测，其准确率也能够达到 98%左右。

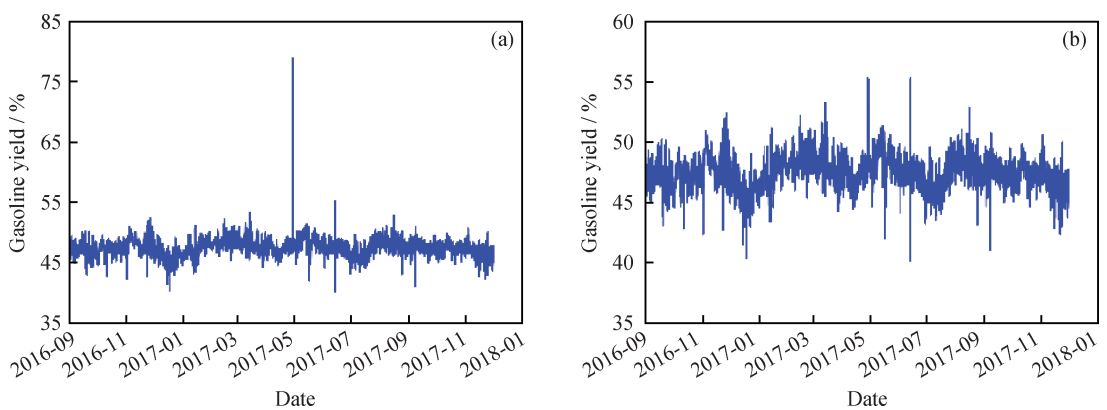

图 5　2016 年 9 月至 2017 年 11 月之间实际的汽油收率和去掉异常值后的汽油收率

Fig. 5　Actual gasoline yield and gasoline yield without outliers between September 2016 and November 2017

(a) Actual gasoline yield; (b) Gasoline yield without outliers

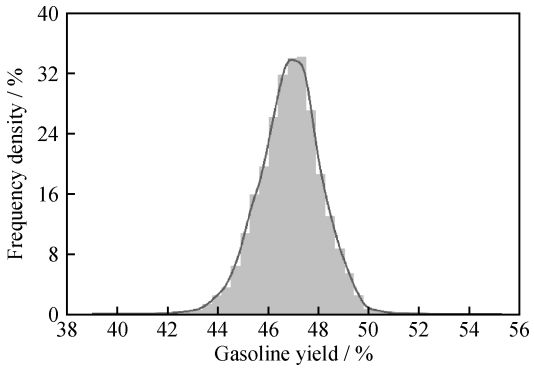

图 6 对图 5(b)的统计收率分布

Fig. 6 Statistical yield distribution for Fig. 5(b)

针对真实汽油收率的整体输出特点，通常意义的回归准确率，并不能很好地反映预测模型的拟合效果。结合原评估方法，考虑去掉收率的均值对变化程度的影响来考察模型对收率变化的预测能力。笔者选择同时使用决定系数 R^2 作为评估标准，R^2 是对模型进行回归后，评价回归模型系数的拟合优度，其计算方法如式(6)所示。

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y}_i)^2} \tag{6}$$

式(5)中，$\bar{y}_i$ 为收率在预测样本中的均值。

R^2 的取值范围一般为负无穷到 1，预测值与真实值的残差平方和越小，该值越接近 1，表明预测值对真实值的拟合优度越大，可解释程度越高。该标准可以反映模型输出对真实产率的拟合程度。与 Pearson 相关系数不同的是，相关系数一般用来描述变量间的线性关系，其绝对值越接近 1，表明变量间的相关性越显著；但 R^2 可以用于描述非线性的相关关系。当 R^2 小于 0 的时候，需要借助其他评估方法来评价拟合程度。

利用 GBDT 算法构造的预测模型对催化裂化的汽油收率进行预测，得到的汽油收率预测结果与实际工业数据的对比如图 7 所示。由图 7 可以看出，模型的预测值总体趋势与工业数据吻合较好，少有出现偏差较大的预测值。

由式(5)和式(6)计算预测结果的准确率和 R^2 系数，并与参考准确率对比分析。计算得到，预测模型的准确率达到 98.9%，明显高于 98%即参考准确率，验证了该模型的可行性和有效性；预测模型的 R^2 系数为 0.236，而该指标的参考值为 0，表明预测模型对汽油收率的拟合程度较好，分析得到的特征指标可以用来解释汽油收率的变化程度。同时，

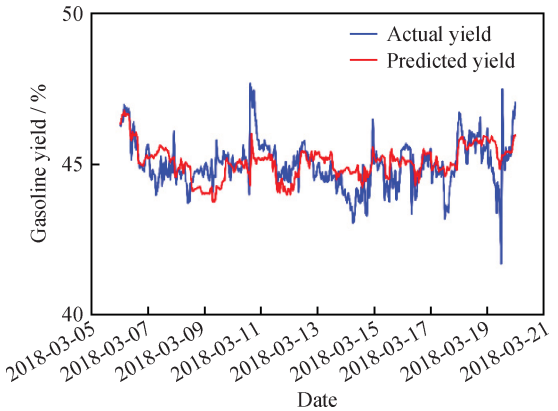

图 7 利用 GBDT 算法得到的汽油收率预测结果与实际工业数据的对比

Fig. 7 Comparisons between gasoline yield predictions of GBDT and actual gasoline yield

根据以上结果，由式(4)计算平均绝对误差(MAE)。

计算可得，基于模型计算得到的汽油产率预测值和实际值的平均绝对误差为 0.531%。因此，无论是从预测结果的准确性还是拟合度上来看，由 GBDT 构建的预测模型对汽油产率能够起到良好的预测效果。

4 结 论

基于某石化企业的 LIMS 及 DCS 系统中的工业生产数据，通过分析监控指标与实际汽油收率的相关性，筛选出了相关性高的分析指标，进一步明确了影响催化裂化装置汽油收率的因素。在此基础上，利用梯度提升决策树 GBDT 算法构建了催化裂化汽油收率的预测模型，并预测了相应的汽油收率。结果发现，由 GBDT 算法构建的汽油收率预测模型预测结果的准确率为 98.9%，R^2 系数为 0.236，平均绝对误差为 0.531%。模型预测结果与实际汽油收率相比，误差率小于 1%，表明构建的模型能精确预测催化裂化装置中汽油等产品收率，对装置操作工况的优化改进具有良好的指导意义，有助于在实际生产中进一步提升催化裂化装置的经济性。

参 考 文 献

[1] SOUZA N L A, TKACH I, JR E M, et al. Vanadium poisoning of FCC catalysts: A quantitative analysis of impregnated and real equilibrium catalysts[J]. Applied Catalysis A General, 2018, 560(12): 206-214.

[2] SHAH M T, UTIKAR R P, PAREEK V K, et al. Computational fluid dynamic modelling of FCC riser: A

review[J]. Chemical Engineering Research & Design, 2016, 111(7): 403-448.

[3] 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(7): 186-231.

[4] 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.)

[5] 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：自然科学版，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 (Edition of Natural Science), 2017, 41(6): 171-177.)

[6] ALI A E, HADIS M, HAMID B, et al. Nine-lumped kinetic model for VGO catalytic cracking; using catalyst deactivation[J]. Fuel, 2018, 231(21): 118-125.

[7] SANI A G, EBRAHIM H A, AZARHOOSHM J. 8-Lump kinetic model for fluid catalytic cracking with olefin detailed distribution study[J]. Fuel, 2018, 225(15): 322-335.

[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] 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.)

[11] 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.)

[12] 全石峰. 云计算环境下大数据处理对电子商务发展的作用[J]. 电脑知识与技术，2013，19(20)：4762-4770. (QUAN Shifeng. 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.)

[13] 赵云山，刘换换. 大数据技术在电力行业的应用研究[J]. 电信科学，2014，30(1)：57-62. (ZHAO Yunshan, LIU Huanhuan. Research on application of big data technique in electricity power industry[J]. Telecommunications Science, 2014, 30(1): 57-62.)

[14] 旷典，付尧明，房丽瑶. 大数据挖掘分析在航空发动机状态监控与故障诊断中的应用[J]. 西安航空学院学报，2017，35(5)：42-46. (KUANG Dian, FU Yaoming, FANG Liyao. Application of big data mining analysis in aircraft engine condition monitoring and fault diagnosis[J]. Journal of Xi'an Aeronautical University, 2017, 35(5): 42-46.)

[15] THOMAS L. Big data in forensic science and medicine[J]. Journal of Forensic and Legal Medicine, 2017, 57(7): 1-6.

[16] 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.)

[17] 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.

[18] 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.)

[19] 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.

[20] 孔金生，张娓娓，王爱玲. 催化裂化粗汽油干点的神经网络质量模型[J]. 湖南工业大学学报，2009，23(5)：77-80. (KONG Jinsheng, ZHANG Weiwei, WANG Ailing. Neural network quality model of FCC crude gasoline end-boiling-point[J]. Journal of Hunan University of Technology, 2009, 23(5): 77-80.)

[21] 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.