极端梯度提升与随机森林融合的天然气露点预测方法

熊伟[1]，何彦霖[2]，宋伟[1]，张厚望[1]，尹爱军[2]

（1.中国石油西南油气田分公司 重庆气矿，重庆 400021；
2.重庆大学 机械与运载工程学院，重庆 400044）

摘要：目的 解决目前水露点数据多为人工采用测量仪器测得，时效性低且成本高昂的问题。**方法** 建立一种基于极端梯度提升（XGBoost）和随机森林（RF）的天然气水露点预测方法。采用 XGBoost 方法对所有监测工艺参数进行分析，筛选出主要影响水露点的关键工艺特征参数，以排除无关特征参数对预测的干扰。建立 RF 预测模型，输入关键特征集参数，实现对水露点的实时预测。以重庆气矿某脱水监测系统监测数据与生产数据为例，对所提预测方法进行对比分析验证。**结果** 相较于 XGBoost、SVM 等预测方法，RF 模型具有最佳的预测性能，且经过 XGBoost 特征选择后，RF 预测结果的 MAE 值降低了 0.016 9 ℃，RMSE 值降低了 0.014 6 ℃。**结论** 基于极端梯度提升与随机森林融合的水露点预测方法具有更优的预测精度与鲁棒性，对指导脱水现场生产具有积极作用。

关键词：三甘醇脱水装置；天然气水露点；极端梯度提升（XGBOOST）；特征提取；随机森林（RF）
中图分类号：TB115 **文献标识码：**A **文章编号：**1672-9242(2022)06-0133-08
DOI：10.7643/ issn.1672-9242.2022.06.000

Prediction Method of Natural Gas Water Dew Point Based on the Fusion of eXtreme Gradient Boosting and Random Forest Regression

XIONG Wei[1], *HE Yan-lin*[2], *SONG Wei*[1], *ZHANG Hou-wang*[1], *YIN Ai-jun*[2]

(1. Chongqing Gas District, Southwest Oil and Gasfield Company, Chongqing 400021, China; 2. College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China)

ABSTRACT: Aiming at the problems that the current water dew point data are mostly manually measured with measuring instruments, the timeliness is low at the time with the high cost, this paper establishes a prediction method natural for gas water

收稿日期：2021–03–23；**修订日期：**2021–05–07
Received：2021-03-23；**Revised：**2021-05-07
基金项目：重庆市科技重大主题专项重点研发项目（cstc2018jszx-cyztzxX0032）；中国石油重庆气矿科研项目（K20-15）
Fund：Key Science and Technology Research Project of Chongqing under Grant (cstc2018jszx-cyztzxX0032); Petro China Chongqing Gas Mine Scientific Research Project (K20-15)
作者简介：熊伟（1982—），男，工程师，主要研究方向为天然气脱水及集输工艺。
Biography：XIONG Wei (1982-), Male, Engineer, Research focus: natural gas dehydration and gathering and transportation technology.
通讯作者：尹爱军（1978—），男，博士，教授，主要研究方向为智能测试与仪器、智能运维与健康管理。
Corresponding author：YIN Ai-jun (1978-), Male, Doctor, Professor, Research focus: intelligent test and instrument, intelligent operation and health management.
引文格式：熊伟，何彦霖，宋伟，等. 极端梯度提升与随机森林融合的天然气露点预测方法[J]. 装备环境工程, 2022, 19(6): 133-140.
XIONG Wei, HE Yan-lin, SONG Wei, et al. Prediction Method of Natural Gas Water Dew Point Based on the Fusion of eXtreme Gradient Boosting and Random Forest Regression[J]. Equipment Environmental Engineering, 2022, 19(6): 133-140.

dew point based on extreme gradient boosting (XGBoost) and random forest (RF). This paper analyzes all the monitored process parameters by using the XGBoost method, and filtrates the pivotal process characteristic parameters that mainly affect the water dew point to eliminate the interference of irrelevant typical parameters to the prediction; the RF prediction mode is established, and the typical characteristic parameters are inputted to realize the real-time prediction of the water dew point. Taking the monitoring data and production data of a dewatering monitoring system in the Chongqing gas mine as an instance, this paper compares and analyzes the proposed prediction method. Compared with the other prediction methods, such as XGBoost and SVM, RF model has the best prediction performance, and after XGBoost feature selection, the MAE value and RMSE value of RF prediction results are reduced by 0.016 9 ℃ and 0.014 6 ℃ respectively. The results show that the water dew point prediction method based on the fusion of eXtreme Gradient Boosting and Random forest regression has better prediction accuracy and robustness. What's more, it has a positive effect on guiding dehydration on-site production.

KEY WORDS: triethylene glycol dehydration unit; gas water dew point; extreme gradient boosting (XGBOOST); feature extraction; random forest (RF)

天然气作为一种优质的清洁能源，承担着国民经济快速发展的重要职责，保证天然气在集输过程中的安全以及产品质量至关重要。天然气在集输过程中，由于时变温度及压力的作用会析出游离的液态水，极易与天然气中的碳、硫等酸性物质形成酸性溶液，进而导致集输管线腐蚀穿孔、阀门堵塞等危害现象[1]。天然气水露点反映了天然气中液态水的含量，是衡量脱水装置脱水性能及天然气产品质量的一项重要技术指标。目前，对于水露点的测量方法大多停留在人工采用冷却镜面露点仪测量[2]的阶段，成本高，且检测仪易受到外界因素的影响，从而导致检测结果与实际值存在误差，同时结果存在一定的时延性，不能实时准确地反映产品质量。因此，对天然气水露点进行实时准确地评估是天然气集输过程中一项重要的任务。

考虑到对天然气水露点影响最大的工艺参数为三甘醇（Triethylene Glycol，TEG）浓度与吸收塔接触温度，基于此，研究人员已提出多种利用相关均衡性的天然气水露点估算方法[3-6]。然上述平衡式无法精确估计气相 TEG 溶液上方的平衡水浓度[7]。Twu 等[7]提出使用 TST（Twu-Sim-Tassone）状态方程来模拟含有 TEG 与水的二元系统，同时提出 TST 状态方程与相关均衡关联的天然气水露点预测方法，但该方法泛化能力较弱。文献[8]构建了水露点关于吸收塔接触温度与 TEG 浓度的一个平衡多项式，尽管该多项式相关工艺参数的有效覆盖范围较广，如吸收塔接触温度范围为 10~80 ℃、TEG 浓度范围为 90%~99.999%，但仍有必要开发更高精度的水露点预测模型。随着数据驱动方法的兴起，Ahmadi 等[9]提出应用基于粒子群优化（Particle Swarm Optimization，PSO）的人工神经网络（Artificial Neural Network，ANN）预测不同 TEG 浓度和吸收塔接触温度下的水露点。Afshin 等[10]根据 TEG 浓度与吸收塔接触温度，分别利用多层感知网络（Multilayer Perceptron，MLP）与径向基神经网络（Radial Basis Function Neural Network，RBF-NN）对水露点进行预测，结果表明，MLP 模型具有更好的表现。文献[11]采用基因表达式编程（Gene Expression Programming，GEP）构造水露点关于 TEG 浓度与吸收塔接触温度的函数，结果显示，所构建的函数较文献[8]更简单、更准确。Ahmad 等[12]将贝叶斯正则化训练的前反馈人工神经网络（Feedforward Artificial Neural Network，FANN）用于预测 TEG 脱水过程中天然气平衡水露点。尽管上述各方法均能实现对水露点的评估，但仅考虑了 TEG 浓度、吸收塔接触温度 2 个与天然气水露点关联紧密的工艺参数，忽略了天然气 TEG 脱水与再生过程中其余重要工艺参数的影响。

由于在解释变量与响应之间复杂非线性时所具备的优秀能力[13]，基于分类回归树（Classification And Regression Tree，CART）的算法被证明是有效可靠的方法，但仍存在过拟合、预测能力差等问题[14]。随机森林（Random Forest，RF）作为其中最具代表性的算法，它克服了上述缺点，拥有极佳的预测能力，在医学[15]、航空[16]、电力[17]等领域均得到了有效应用。将 RF 算法引入石化领域，构建天然气水露点预测模型，为天然气脱水装置关键参数预测提供参考。实际 TEG 脱水装置中，在天然气脱水与 TEG 再生流程中涵盖多个反映装置状态的工艺参数，如各部位的流量、液位、压力、温度以及控制阀开度等，各工艺参数相互耦合以维持化工过程平衡。提取对水露点敏感性较高的影响参数，减少无关或冗余参数对水露点预测结果的影响，提高预测模型的预测精度是本文重点研究内容之一。极端梯度提升（eXtreme Gradient Boosting，XGBoost）[18]是一种性能优秀的特征选择算法，在各领域[19-20]获得了较好的效果。

本文将 XGBoost 算法引入至水露点预测领域，对脱水过程中各工艺参数与水露点间的重要性进行评分，筛选最优特征参数集，以最优特征参数集作为 RF 模型特征变量，提出将 XGBoost 算法、RF 算法有机结合的天然气水露点预测方法。利用 XGBoost

约简样本指标，提取关键特征参数，充分利用脱水流程中各工艺参数对预测对象的影响。再使用 RF 建立水露点预测模型，实现对水露点进行有效且实时的预测。以天然气脱水监测系统监测数据与生产数据为例，验证所提方法的有效性。

1　XGBoost 特征选择

XGBoost 是一种基于梯度提升决策树（Gradient Boosting Decision Tree，GBDT）的机器学习框架[21]。常规 GBDT 模型在优化过程中仅使用一阶导数信息，难以实施分布式训练。XGBoost 则对损失函数执行二次泰勒展开式，同时利用一阶及二阶导数信息，可在训练时自动使用 CPU 的多线程并行计算。此外为防止过拟合，在损失函数中增加正则惩罚项降低模型复杂度，并采用行列采样的方式进行采样。模型如式（1）所示。

$$\hat{y}_i = m_J(\boldsymbol{x}_i) = \sum_{j=1}^{J} f_j(\boldsymbol{x}_i), f_j \in F \tag{1}$$

式中：J 为 CART 树的数量；f_j 为表示第 j 棵 CART 树；F 为所有 CART 树的集合空间；$\boldsymbol{x}_i$ 表示第 i 个数据点的特征向量。

对应的目标函数为：

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_j \Omega(f_j) \tag{2}$$

式中：l 代表损失函数，表征预测值与观测值之间的误差；Ω 为用于防止过拟合的正则惩罚项函数，可有效限制叶子节点的数量。

$$\Omega(f_j) = YT + \frac{\lambda}{2} w^2 \tag{3}$$

式中：Y、λ 为惩罚系数；T 表示给定 CART 树的叶节点数目；w 为每棵 CART 树上叶子节点的权重。

设 $\hat{y}_i^{(t)}$ 为第 i 个样本在第 t 次迭代后的预测值，有：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i) \tag{4}$$

对目标函数进行二阶泰勒展开，有：

$$\mathcal{L}^{(t)} = \sum_i \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f) \tag{5}$$

其中，g_i、h_i 分别表示一阶与二阶梯度，见式（6）。

$$\begin{cases} g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i = \partial_{\hat{y}_i^{(t-1)}}{}^2 l(y_i, \hat{y}_i^{(t-1)}) \end{cases} \tag{6}$$

将式（3）与（6）代入目标函数的二阶泰勒展开式，当其导数为 0 时，最佳权重及目标函数为：

$$w_j^* = -\frac{\sum g_i}{\sum h_i + \lambda} \tag{7}$$

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum g_i\right)^2}{\sum h_i + \lambda} + \gamma T \tag{8}$$

基于 XGBoost 进行特征选择时，特征变量的平均增益反映了以当前特征为分支节点进行分裂所提升的准确率，以该指标量化特征变量在模型中的重要程度。每次分裂后模型的增益表示为：

$$G_{\text{ain}} = \frac{1}{2} \left[\frac{\left(\sum\sum_{i \in I_{\text{L}}} g_i\right)^2}{\sum_{i \in I_{\text{L}}} h_i + \lambda} + \frac{\left(\sum_{i \in I_{\text{R}}} g_i\right)^2}{\sum_{i \in I_{\text{R}}} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{9}$$

式中：I_{L}、I_{R} 分别表示分裂后所有左、右侧叶子节点的集合。对特征变量在每棵 CART 树的增益进行加权平均计算即可得到其在模型中的重要性得分。

2　水露点预测模型

2.1　随机森林

随机森林（Random Forest，RF）是在一种在决策树基础上所构建的集成学习方法，其在决策树训练过程中引入了随机子空间与随机属性特征，有效提高了模型的抗噪能力、抗过拟合性，如今广泛应用于分类与回归问题中。单棵决策树在面对数据中的微小变化时容易产生极大误差，RF 结合多棵决策树进行分类或回归处理，克服了单棵决策树容易出现的结果不稳定现象。RF 的基本原理是通过使用 Boostrap 重抽样法，从总体训练数据集中有放回地随机抽取多个与样本容量相同的样本子集，最大限度地构建多棵决策树，且每棵决策树随机选择特征进行节点分裂，构建一个全局集成学习器，而后取每棵决策树的回归均值作为回归预测结果。RF 模型的构造过程如图 1 所示。

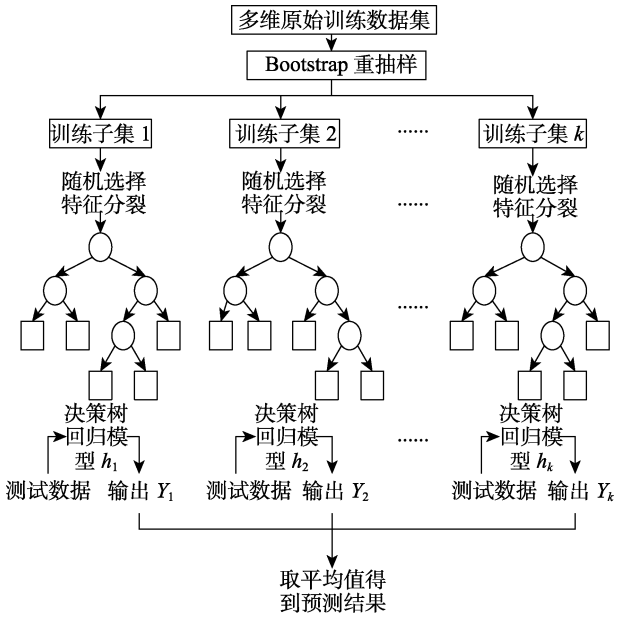

图 1　RF 构造过程

Fig.1 The Construction process of RF

构造随机森林的一般步骤有：

1）记原始训练数据集为 $S=(\boldsymbol{x}_i,y_i)_{i=1}^{N}$，生成随机向量序列 $\boldsymbol{\theta}_i,i=1,2,\ldots,k$。利用 Bootstrap 重抽样法对 S 进行随机抽样，进而得到 k 个样本容积与 S 相同的样本子集。

2）对于特征参数集 X，针对每个样本子集，建立决策树回归模型 $h(X,\boldsymbol{\theta}_i),i=1,2,\ldots,k$。随机选择 m 个特征，m 应小于总的特征数，使得每个叶节点选择最大信息增益的特征进行分裂，同时不进行剪枝处理。其中，信息增益表示为：

$$G_{\text{Gain}}(A)=E_{\text{Entropy}}(D)-\sum_{j=1}^{w}\frac{|D_j|}{|D|}E_{\text{Entropy}}(D_j) \tag{10}$$

$$E_{\text{Entropy}}(D)=-\sum_{i=1}^{m}p_i log_2 p_i \tag{11}$$

式中：i 为回归或分类值；p_i 表示对应值发生的概率；w 为划分节点的个数；$\frac{|D_j|}{|D|}$ 为第 m 个划分叶节点的权重值。

3）所有样本子集训练完成后，得到决策树回归模型，将所有决策树组合形成随机森林，取各决策树的均值作为随机森林回归预测模型的最终结果。

2.2　预测流程

在三甘醇脱水装置天然气水露点预测中，用 XGBoost 筛选出的影响水露点的关键特征参数建立特征变量集 X，输入至 RF 模型得到水露点的预测结果。其流程如图 2 所示。

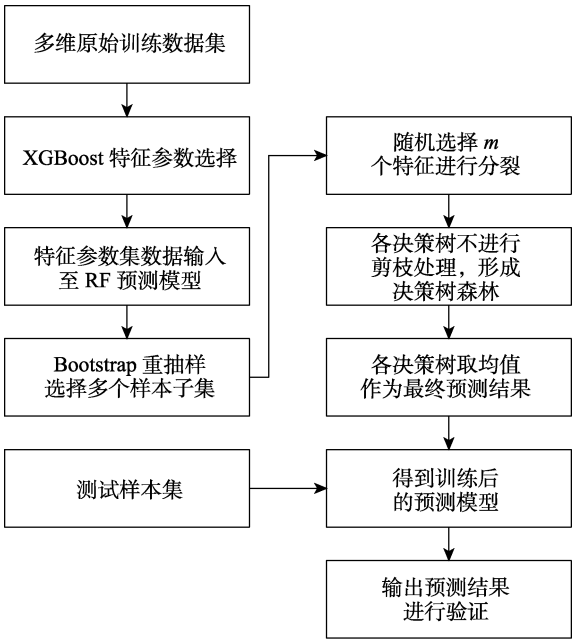

图 2　天然气水露点预测流程

Fig.2 Natural gas water dew point prediction process

采用均方根误差（δ_{RMSE}）与平均绝对误差（δ_{MAE}）指标评价预测方法的有效性及准确性。

$$\delta_{\text{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{12}$$

$$\delta_{\text{MAE}}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{13}$$

式中：n 为样本总数；$\hat{y}_i$ 为测试样本预测值；y_i 为观测值。δ_{RMSE} 表征了预测值与观测值之间的偏差，而 δ_{MAE} 表示预测值与观测值绝对误差，当这 2 个指标的数值越小时，表征方法拟合效果越好、性能越优。

3　实验验证

3.1 数据来源

西南油气田公司重庆气矿某脱水装置现场如图 3 所示。该套脱水装置采用三甘醇脱水工艺实现对天然气脱水，分为天然气脱水与三甘醇再生 2 部分子系统，包括原料气分离器、吸收塔、重沸器及精馏柱等众多处理子设备。

图 3　脱水现场

Fig.3 Dehydration site map

根据已建立的脱水装置实时数据监测与采集系统，对原料气分离器液位、三甘醇循环量及精馏柱温度等 33 个工艺参数数据进行采集，并统计三甘醇贫液浓度、三甘醇富液浓度及天然气水露点等 3 个生产参数数据，详细监测参数见表 1。

本文以该脱水装置生产数据对所提方法进行实验验证，并在全部特征及特征选择后的数据集上对比验证方法的优越性。以脱水装置监测工艺参数为特征参数集，收集该场站 2017—2019 年共计 495 条监测数据及天然气水露点，部分工艺参数原始数据如图 4 所示。

表 1　脱水装置工艺参数

Tab.1 Process parameters of dehydration unit

序号	参数名称	序号	参数名称	序号	参数名称
P_1	进装置压力	P_{13}	压力控制阀开度	P_{25}	燃料气压力
P_2	原料气分离器液位	P_{14}	出吸收塔富甘醇温度	P_{26}	精馏柱顶部温度
P_3	过滤分离器差压	P_{15}	进闪蒸罐富甘醇温度	P_{27}	缓冲罐液位
P_4	吸收塔差压	P_{16}	闪蒸罐压力	P_{28}	出缓冲罐贫甘醇温度
P_5	三甘醇循环量	P_{17}	闪蒸罐压力控制阀开度	P_{29}	三甘醇入泵前温度
P_6	吸收塔磁浮子液位	P_{18}	闪蒸罐液位	P_{30}	循环泵变频器给定值
P_7	吸收塔雷达液位	P_{19}	闪蒸罐液位控制阀开度	P_{31}	灼烧炉炉膛温度
P_8	吸收塔液位控制阀开度	P_{20}	板式换热器富甘醇温度	P_{32}	灼烧炉顶部温度
P_9	计量静压	P_{21}	重沸器中部温度	P_{33}	灼烧炉温度控制阀开度
P_{10}	计量差压	P_{22}	重沸器后端温度	P_{34}	三甘醇贫液浓度
P_{11}	计量温度	P_{23}	重沸器前端温度	P_{35}	三甘醇富液浓度
P_{12}	瞬时处理量	P_{24}	重沸器温度控制阀开度	P_{36}	天然气水露点

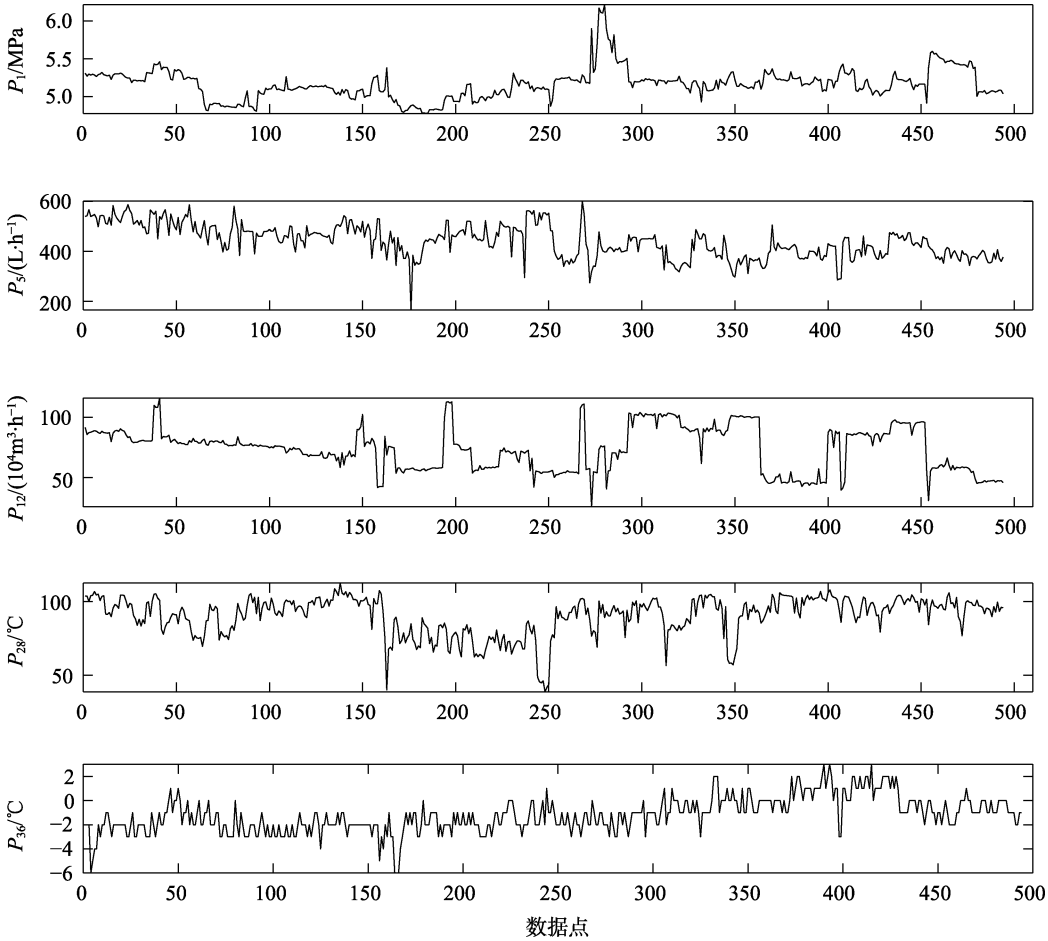

图 4　部分监测参数原始数据

Fig.4 Raw data of some monitoring parameters

3.2　特征提取

针对该三甘醇脱水装置多维原始训练数据集，以所有工艺参数为自变量、天然气水露点为因变量，设定 XGBoost 模型损失函数正则项的叶节点复杂性系数 γ=0.0，惩罚系数 λ=1，决策树的数量为 100，决策树的最大深度为 8，最小叶子点权重和为 2，学习率为 0.01。得到所有工艺特征参数对于天然气水露点的重要性得分，如图 5 所示。

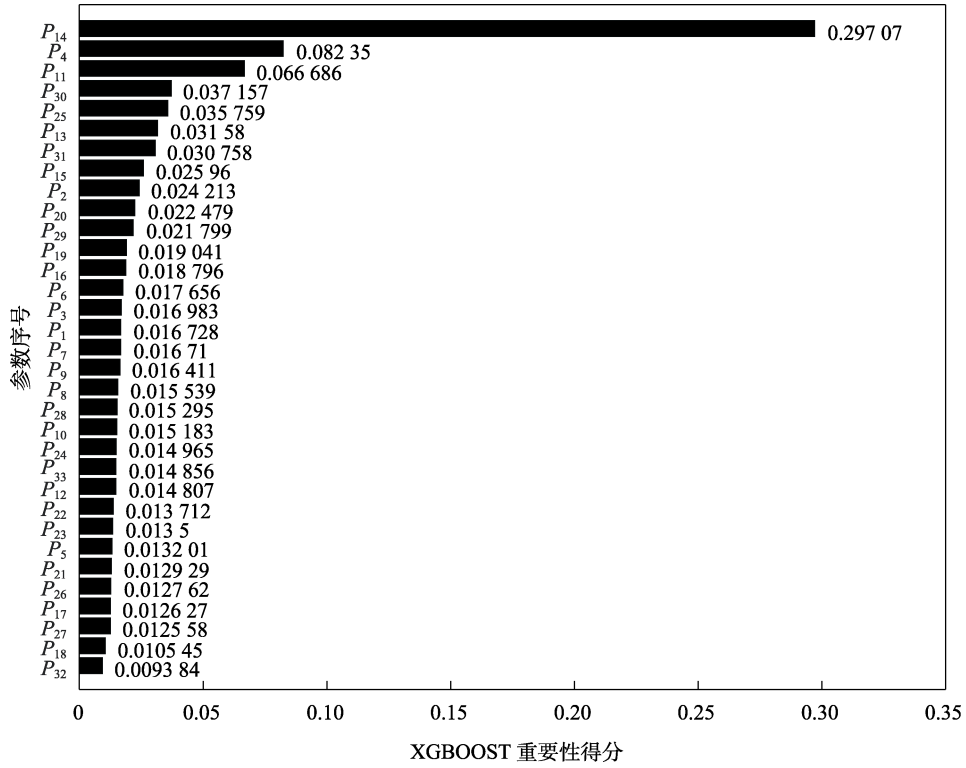

图 5　工艺参数重要性排序

Fig.5 Process parameter importance ranking

从图 5 可以看出，出吸收塔富甘醇温度的特征重要性最高，达到 0.297 07，说明出吸收塔富甘醇温度对天然气水露点具有较大的影响作用。该参数同时反映了三甘醇进吸收塔温度与湿天然气进塔温度 2 个影响因素，与实际情况吻合[22]。在模型特征选择中，过多或过少的特征数都会导致模型的预测失效，根据重要性得分进行排序，由排序后不同特征个数所对应模型预测准确率如图 6 所示。随着特征个数的增加，在特征集为前 9 个特征时 δ_{RMSE} 及 δ_{MAE} 均达到最小值，故选择前 9 个工艺参数作为后续 RF 预测模型的特征参数集，见表 2。

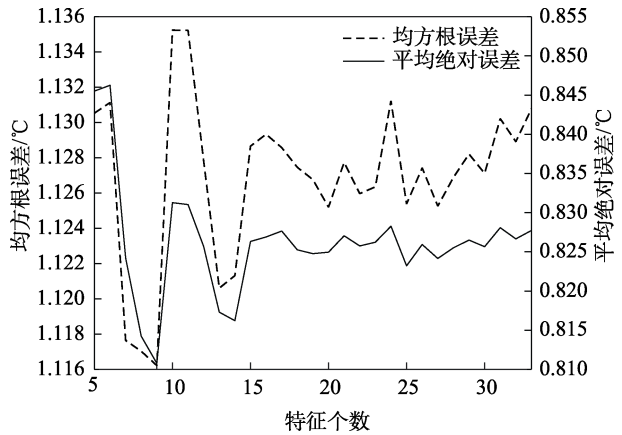

图 6　不同特征个数预测评价

Fig.6 Prediction and evaluation of the number of different features

表 2　特征参数集

Tab.2 Characteristic parameter set

序号	参数名称
P_{14}	出吸收塔富甘醇温度
P_4	吸收塔差压
P_{11}	计量温度
P_{30}	循环泵变频器给定值
P_{25}	燃料气压力
P_{13}	压力控制阀开度
P_{31}	灼烧炉炉膛温度
P_{15}	进闪蒸罐富甘醇温度
P_2	原料气分离器液位

3.3　对比分析

设置测试集与训练集的比例为 0.25，为了更好地验证所提方法的优越性，在全部特征以及特征选择后的参数集中，运用 RF、XGBoost、支持向量机（Support Vector Machine，SVM）进行对比分析验证，结果如图 7、8 所示。其中，采用网格搜索对 RF 模型进行寻优处理，设定 RF 模型的决策树数量为 100，叶子节点最小样本数为 15，不限制树的最大深度，内部节点再划分样本数为 2。图 7 表示了以全部工艺参数作为后续预测模型的特征集时各种预测模型的预测结果对比，图 8 表示了采用 XGBoost 选择的特征参数作为预测模型特征集时各种预测模型的预测结果

对比。可以看出，无论是以全部参数为特征还是进行特征选择后，RF 的预测值相较于其余 2 个模型而言更接近真实值，吻合效果更佳，说明 RF 用于天然气水露点预测领域具有较强的可行性。

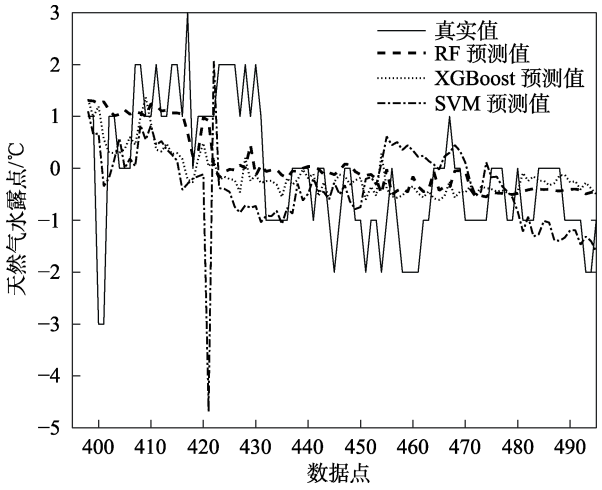

图 7　全部工艺参数为特征集

Fig.7 The feature sets of all process parameters

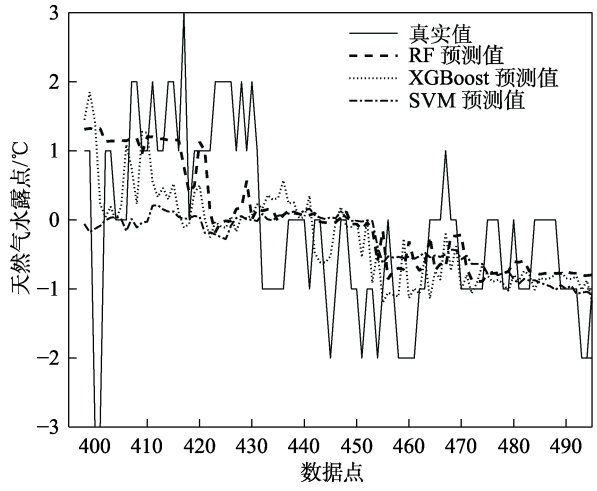

图 8　选择特征参数集数

Fig.8 Select the data of feature parameter sets

为进一步评价模型预测性能，采用式（12）、（13）所示的均方根误差、平均绝对误差对 3 组预测模型进行误差分析，结果如图 9 所示。对于全部参数作为特征集时，RF 的δ_{MAE}=0.827 7 ℃分别低于 XGBoost 模型的 0.896 2 ℃和 SVM 模型的 1.005 3 ℃，同时其δ_{RMSE}均低于其余 2 组模型；以 XGBoost 选择参数作为特征集时，RF 相较于其余 XGBoost 与 SVM 模型，具有最低的δ_{RMSE}、δ_{MAE} 值。进一步说明了无论是否进行特征选择，RF 的预测效果更好。对比特征选择前后的评价指标，可以看出，δ_{RMSE}、δ_{MAE} 值均有一定程度的降低。从δ_{MAE} 值来看，RF、XGBoost 与 SVM 特征选择后，分别减少了 0.016 9、0.031 8、0.082 1 ℃；从δ_{RMSE} 值来看，特征选择后，RF 与 SVM 预测模型分别降低了 0.014 6、0.230 8 ℃，而 XGBoost 则增加了 0.020 4 ℃。出现该现象的原因是 XGBoost 模型中存在个别观测值与实际值有较大偏离程度的离群点，导致δ_{RMSE} 指标变差。根本原因在于按照特征重要性选择特征参数数目时，所选择的基准模型为 RF，从而针对 XGBoost 模型可能去除了部分有效信息，导致预测结果出现个别奇异点。从δ_{MAE} 指标可以看出，XGBoost 模型在特征选择后，整体预测效果得到了提升。综上所述，经过特征选择后的预测模型，预测性能均得到了一定程度的提升，且本文所提方法均有更好的预测效果。

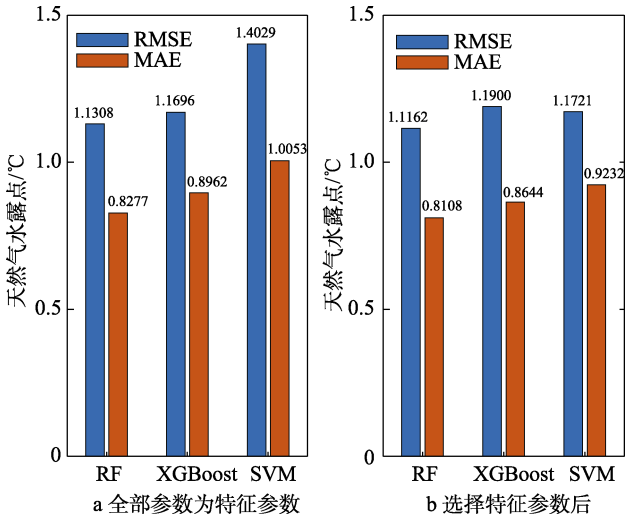

图 9　预测结果评价

Fig.9 Evaluation of prediction results:
a) all parameters are characteristic parameters; b) after selecting characteristic paramenters

将原始脱水监测数据集划分为 5 等份，任取其中 1 份作为测试数据集，交叉验证本文所提方法的泛化能力，取评价结果的平均值作为交叉验证评价指标，结果见表 3。可以看出，经过特征选择后，与 XGBoost 模型及 SVM 模型相比，RF 模型具有更优的δ_{RMSE}及δ_{MAE} 指标性能，说明由 XGBoost 特征提取后运用 RF 模型对天然气水露点预测具有较好的泛化能力，预测精度和可靠性更高。

表 3　交叉验证评价结果

Tab.3 Cross-validation evaluation results

℃

评价指标	RF	XGBoost	SVM
δ_{RMSE}	1.175 17	1.367 23	1.732 35
δ_{MAE}	0.890 82	1.065 25	1.402 46

4　结论

针对三甘醇脱水装置实际生产中天然气水露点数据多为人工采用检测仪获得，易受到外界因素的影响，同时检测费用高昂、时效性低等问题，将 XGBoost

与 RF 进行有机融合，建立了天然气水露点 XGBoost-RF 预测方法。以实际生产数据对比分析，结果表明：

1）以脱水系统实时监测工艺参数数据为特征，可有效实现对天然气水露点的预测，实时性高，且避免了外界因素的影响。

2）采用 XGBoost 算法，对特征参数进行重要性排序，并选择对目标参数敏感特征，降低了冗余特征的影响，提高了预测模型的预测性能。特征选择前后，RF 预测结果的平均绝对误差值降低了 0.016 9 ℃，均方根误差值降低了 0.014 6 ℃；

3）对比分析 RF 与 XGBoost、SVM 预测模型结果，RF 预测模型具有更好的预测能力与工程实用性，可为天然气集输处理现场提供积极的指导作用。

参考文献：

[1] 陈赓良. 天然气三甘醇脱水工艺的技术进展[J]. 石油与天然气化工, 2015, 44(6): 1-9.
CHEN Geng-liang. Technical Progress of TEG Dehydration Process in Natural Gas Industry[J]. Chemical Engineering of Oil & Gas, 2015, 44(6): 1-9.

[2] 邹伟. 冷却镜面凝析湿度计法检测天然气水露点的不确定度评定[J]. 计量学报, 2019, 40(S1): 150-153.
ZOU Wei. Evaluation of Uncertainty in Detecting Dew Point of Natural Gas by Cooling Mirror Condensation Hygrometer[J]. Acta Metrologica Sinica, 2019, 40(S1): 150-153.

[3] SCAUZILLO F. Equilibrium Ratios of Water in the Water-Triethylene Glycol-Natural Gas System[J]. Journal of Petroleum Technology, 1961, 13(7): 204-211.

[4] ROSMAN A. Water Equilibrium in the Dehydration of Natural Gas with Triethylene Glycol[J]. Society of Petroleum Engineers Journal, 1973, 13(5): 297-306.

[5] HERSKOWITZ M, GOTTLIEB M. Vapor-Liquid Equilibrium in Aqueous Solutions of Various Glycols and Poly(ethylene glycols). 3. Poly(ethylene glycols)[J]. Journal of Chemical & Engineering Data, 1985, 30(2): 233-234.

[6] WON K W. Thermodynamic Basis of the Glycol Dew-Point Chart and Its Application to Dehydration[C]//73rd GPA Annual Convention New Orleans, LA: [s. n.], 1994.

[7] TWU C H, TASSONE V, SIM W D, et al. Advanced Equation of State Method for Modeling TEG-Water for Glycol Gas Dehydration[J]. Fluid Phase Equilibria, 2005, 228-229: 213-221.

[8] BAHADORI A, VUTHALURU H B. Rapid Estimation of Equilibrium Water Dew Point of Natural Gas in TEG Dehydration Systems[J]. Journal of Natural Gas Science and Engineering, 2009, 1(3): 68-71.

[9] AHMADI M A, SOLEIMANI R, BAHADORI A. A Computational Intelligence Scheme for Prediction Equilibrium Water Dew Point of Natural Gas in TEG Dehydration Systems[J]. Fuel, 2014, 137: 145-154.

[10] AFSHIN T, ALI B H, HOSSEIN M, et al. Prediction of Water Formation Temperature in Natural Gas Dehydrators Using Radial Basis Function (RBF) Neural Networks[J]. Natural Gas Industry B, 2016, 3(2): 173-180.

[11] ROSTAMI A, SHOKROLLAHI A. Accurate Prediction of Water Dewpoint Temperature in Natural Gas Dehydrators Using Gene Expression Programming Approach[J]. Journal of Molecular Liquids, 2017, 243: 196-204.

[12] AHMAD Z, BAHADORI A, ZHANG Jie. Prediction of Equilibrium Water Dew Point of Natural Gas in TEG Dehydration Systems Using Bayesian Feedforward Artificial Neural Network (FANN)[J]. Petroleum Science and Technology, 2018, 36(20): 1620-1626.

[13] HASTIE T, FRIEDMAN J, TIBSHIRANI R. The Elements of Statistical Learning[M]. New York: Springer New York, 2001.

[14] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[15] 曹文哲, 应俊, 陈广飞, 等. 基于 Logistic 回归和随机森林算法的 2 型糖尿病并发视网膜病变风险预测及对比研究[J]. 中国医疗设备, 2016, 31(3): 33-38.
CAO Wen-zhe, YING Jun, CHEN Guang-fei, et al. Risk Prediction and Comparitive Research of Type 2 Diabetes Mellitus Complicated with Retinopathy Based on Logistic Regression and Random Forest Algorithm[J]. China Medical Devices, 2016, 31(3): 33-38.

[16] HAN Te, JIANG Dong-xiang, ZHAO Qi, et al. Comparison of Random Forest, Artificial Neural Networks and Support Vector Machine for Intelligent Diagnosis of Rotating Machinery[J]. Transactions of the Institute of Measurement and Control, 2018, 40(8): 2681-2693.

[17] 吴潇雨, 和敬涵, 张沛, 等. 基于灰色投影改进随机森林算法的电力系统短期负荷预测[J]. 电力系统自动化, 2015, 39(12): 50-55.
WU Xiao-yu, HE Jing-han, ZHANG Pei, et al. Power System Short-Term Load Forecasting Based on Improved Random Forest with Grey Relation Projection[J]. Automation of Electric Power Systems, 2015, 39(12): 50-55.

[18] CHEN T, GUESTRIN C. Xgboost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. California: Association for Computing Machinery, 2016.

[19] WHITE C, ISMAIL H D, SAIGO H, et al. CNN-BLPred: A Convolutional Neural Network Based Predictor for B-Lactamases (BL) and Their Classes[J]. BMC Bioinformatics, 2017, 18(Suppl 16): 577.

[20] 宋国琴, 刘斌. 基于 XGBoost 特征选择的幕课翘课指数建立及应用[J]. 电子科技大学学报, 2018, 47(6): 921-926.
SONG Guo-qin, LIU Bin. The Establishment and Application of Drop-out-Index of MOOCs Based on XGBoost Feature Selection[J]. Journal of University of Electronic Science and Technology of China, 2018, 47(6): 921-926.

[21] MA Xiao-jun, SHA Jing-lan, WANG De-hua, et al. Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning[J]. Electronic Commerce Research and Applications, 2018, 31: 24-39.

[22] 曾自强, 张育芳. 天然气集输工程[M]. 北京: 石油工业出版社, 2001.
ZENG Zi-qiang, ZHANG Yu-fang. Natural Gas Gathering and Transportation Project[M]. Beijing: Petroleum Industry Press, 2001.

责任编辑：刘世忠