文章编号:1009-9603(2023)01-0161-10 DOI:10.13673/j.cnki.cn37-1359/te.202203015

基于自动机器学习的采油井压裂效果预测方法

盖 建[1,2]

(1.国家能源陆相砂岩老油田持续开采研发中心,黑龙江 大庆 163712;
2.中国石油大庆油田有限责任公司 勘探开发研究院,黑龙江 大庆 163712)

摘要:目前大庆油田采油井压裂效果预测时多是凭借经验或者多元线性回归等简单模型,存在着预测结果稳定性差且预测精度不高的问题。以大庆油田N23区块为例,借助数理统计方法对采油井压裂效果与各项影响因素开展了相关性分析,并采用随机森林算法研究了各影响因素对N23区块采油井压裂效果的影响程度;阐述了自动机器学习中元学习、贝叶斯优化和模型集成这三项关键技术的原理以及实现方法,并利用自动机器学习建立了基于数据驱动的采油井压裂效果预测模型;同时,将自动机器学习预测模型与随机森林、支持向量机和神经网络这3种常见机器学习算法的预测性能进行了对比,并利用该自动机器学习预测模型对N23区块的水力压裂进行设计与优化。结果表明,压裂前的生产参数对预测采油井压裂效果有着重要的影响;自动机器学习预测模型比其他算法的精度更高,模型在测试集上的决定系数为0.695,预测结果相对误差的平均值为18.96%,比目前水平降低了57.53%;经过模型优化的压裂方案较原方案增加经济效益约3.2×10^4~27.4×10^4元/井次。

关键词:水力压裂;自动机器学习;元学习;贝叶斯优化;模型集成;大庆油田
中图分类号:TE357.1 **文献标识码**:A

Prediction method for hydraulic fracturing effect of oil production well based on automatic machine learning technology

GAI Jian[1,2]

(1.*R & D Center of Sustainable Development of Continental Sandstone Mature Oilfield, Daqing City, Heilongjiang Province*, 163712, *China*; 2.*Exploration and Development Research Institute of Daqing Oilfield Co., Ltd., PetroChina, Daqing City, Heilongjiang Province*, 163712, *China*)

Abstract: At present, the prediction of the hydraulic fracturing effect of oil production wells in Daqing Oilfield mostly relies on experience or simple models such as multiple linear regression, which leads to poor stability of prediction results and low prediction accuracy. With Block N23 of Daqing Oilfield as an example, the correlation between the fracturing effect of oil production wells and influencing factors is analyzed by the mathematical statistics. The influence of those factors on the hydraulic fracturing effect in Block N23 is studied by a random forest algorithm. Additionally, the principles and implementation methods of meta learning, Bayesian optimization, and model ensemble in automatic machine learning are presented, and a prediction model of a data-driven hydraulic fracturing effect based on the automatic machine learning technology is constructed. Meanwhile, the model is compared with three common machine learning algorithms: random forest, support vector machine and neural network. The proposed model is employed to design and optimize the hydraulic fracturing of Block N23. The results show that the production parameters before fracturing exert an important influence on predicting the effect of oil production well after fracturing. The model constructed by the automatic machine learning algorithm has higher accuracy than other algorithms. The determination coefficient on the test set is 0.695, and the average relative prediction error is 18.96%, which is 57.53% lower than the current level. Compared with the original one, the fracturing scheme opti-

收稿日期:2022-03-15。
作者简介:盖建(1990—),女,黑龙江双城人,工程师,硕士,从事油气田开发理论、大数据与人工智能工作。E-mail:gaijian@petrochina.com.cn。
基金项目:国家科技重大专项"大庆长垣特高含水油田提高采收率示范工程"(2016ZX05054),中国石油天然气集团有限公司重大科技专项"大庆油气持续有效发展关键技术研究与应用—特高含水后期水驱高效精准挖潜技术研究与规模应用"(2016E-0205)。

mized by the model can increase the economic benefit by about 3.2×10^4–27.4×10^4 yuan per well.

Key words: hydraulic fracturing; automatic machine learning; meta learning; Bayesian optimization; model ensemble; Daqing Oilfield

水力压裂是大庆油田开发过程中一种非常重要的增产方式[1-4]。水力压裂有着较为高额的成本，压裂后的产油效果直接决定了其经济效益。因此，亟需通过较为精确的预测模型对压裂后的产油效果进行提前预判，以达到压裂方案优化设计的目的。目前，对大庆油田压裂后产油效果的预测大部分是凭借经验或者多元线性回归等简单模型，导致了预测结果的不确定性强和准确度低。

机器学习是建立高精度预测模型的一种有效方法，正逐渐被应用到油气田勘探开发的各个领域[5-14]。诸多研究人员用机器学习来评估完井和增产措施的效果[15-18]。目前，机器学习在压裂中的应用多是针对致密油[7]与页岩气[19-21]等非常规储层；文献[22]利用机器学习对大庆油田油井压裂效果进行预测并取得了较高的精度，但存在一些问题：第一，在影响因素分析时，考虑因素不够全面，没有将可能影响压裂效果的压裂液、压裂类型等工程因素，以及沉积相、目的层深度等地质因素考虑在内。第二，在影响因素相关性分析与模型特征选择中，虽然采用了神经网络等方式，但仍局限于每个影响因素与目标变量之间的单因素分析，没有考虑影响因素之间的关联，导致特征选择不够客观。第三，研究中采用的基础数据集中样本数量少，可能造成预测模型的普适性与推广性不足。

机器学习算法的种类较多，每种算法适宜解决的问题不同。目前在使用机器学习算法解决问题时，会遇到以下2个问题：第一，没有一种机器学习算法能在所有数据集上都有最好的表现。第二，大部分机器学习算法性能的优劣在很大程度上依赖于超参数优化。以上2个问题会造成即使花费大量时间和精力去进行机器学习建模，仍然无法达到更高的精度。自动机器学习满足了不同数据集对不同机器学习流程的需求，能够较好地解决上述2个问题。目前流行的自动机器学习系统包括Auto-WEKA，Hyperopt-sklean，Auto-sklearn，TPOT和Auto-Keras等，它们能在不同的预处理器、分类器、超参数设置等流程之间执行组合优化，从而大大减少用户的工作量，并且降低机器学习使用者的门槛。应用自动机器学习建立了大庆油田N23区块的采油井压裂效果预测模型，并且利用研究成果指导了N23区块采油井压裂参数的优化设计。

1　数据准备

压裂措施数据来自大庆油田N23区块，该区块主要发育萨尔图、葡萄花和高台子3个油层。萨Ⅱ、萨Ⅲ和葡Ⅰ是区块的主力开发油层组，为河流-三角洲沉积。密闭取心资料显示，油层中的孔隙大部分互相连通，平均孔隙度为26.6%，平均渗透率为1 184.8 mD，以细砂岩为主，含量为47.3%，粒度中值为0.13 mm，分选系数为3.4。该区块采用5点法面积井网布井，开发井均为直井，通过向地层注水补充能量。

收集整理了该区块采油井压裂措施数据共887井次。数据集包含采油井的地质特征、措施前生产数据、压裂工程参数和措施效果。地质特征包括采油井坐标、措施目的层厚度、平均深度、渗透率、孔隙度、破裂压力和沉积相类型；措施前生产数据包括日产油量、含水率、日产液量以及沉没度；压裂工程参数包括压裂方式、压裂液体积、加砂量、压裂液类型、混砂比和裂缝条数。措施效果采用采油井压裂后稳定的日产油量，将该指标作为目标值开展研究。

2　研究方法

2.1　自动机器学习

自动机器学习工作流程（图1）包括3个主要部分，分别是元学习（meta learning）、贝叶斯优化（Bayesian optimization）和模型集成（build ensemble）。

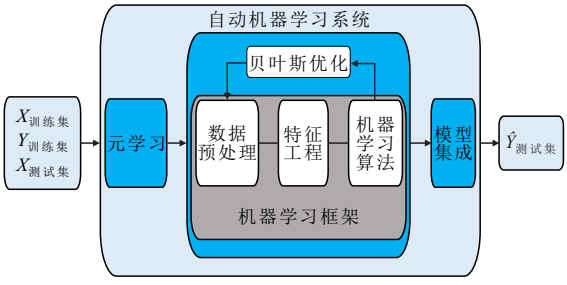


图1　自动机器学习工作流程示意

Fig.1　Automatic machine learning process

2.1.1　元学习

为了提高效率，自动机器学习采用元学习[23]来

预热贝叶斯优化流程。元学习可以实现从以前的任务中获得知识,应用该技术选择可能在目标数据集上表现良好的机器学习框架实例。从已有数据集库中选择与新数据集相似的数据集,将相似数据集的机器学习框架作为初始参数传递给贝叶斯优化流程,具体实现方法如下:收集OpenML存储库[24]的开源数据集,对于每一个数据集,评估一组包括常规、信息论相关和统计相关的元特征[25]。然后,在2/3的数据上采用k折交叉验证进行贝叶斯优化,将剩余的1/3数据作为测试集,将能使测试集获得最佳性能的机器学习框架作为最优实例储存。同时,计算本研究目标数据集D_{frac}的元特征,在元特征空间中分别计算所有数据集与D_{frac}的L1范数并排序。D_{frac}的L1范数表达式为:

$$d_{\text{p}}(D_{\text{frac}},D_j)=\sum_i\left|m_i^{\text{frac}}-m_i^j\right| \tag{1}$$

其中,L1范数能够定义2个数据集之间的相似度,L1范数越小,相似度越高。最后,将与目标数据集相似度最高的25个已储存的机器学习框架传送给贝叶斯优化流程。

2.1.2　贝叶斯优化

贝叶斯优化[26]的原理是通过拟合一个概率模型来捕捉超参数组合与其对应模型性能之间的关系,使用该模型选择最佳的超参数设置方向,计算超参数组合,用计算的结果更新模型,然后通过不断迭代使误差逐渐减小。基于树模型的贝叶斯优化在高维、结构化和部分离散的问题[27]上比基于高斯模型的贝叶斯优化[28]更为理想。而在基于树模型的贝叶斯优化方法中,基于随机森林的序列模型算法配置(SMAC)[29]比树状结构Parzen估计方法(TPE)[30]表现更好,因此本研究中使用SMAC。SMAC使用随机森林算法[31],通过每次评估1折并尽早丢弃性能较差的超参数组合,来实现快速交叉验证。本研究在数据预处理、特征预处理和算法工程3个部分通过贝叶斯优化实现了自动化。

2.1.3　模型集成

自动机器学习利用贝叶斯优化得到了很多性能较好的模型,如果仅保留性能最佳的一个模型而丢弃其他模型,那么在时间和计算力上都比较浪费。因此,储存性能较好的多个模型并构建一个集成模型。集成模型的效果通常优于单个模型[32-33],而当组成集成模型的各个基础模型单独性能很强且产生的误差不相关时,集成模型的表现会更好。另外,集成模型还会大大地提高模型的泛化能力,防止出现过拟合。采用集成选择(ensemble selection)来进行模型集成。集成选择[32]是一个贪婪的过程,它向一个空的集成中迭代地加入模型,力求使集成模型在验证集上的性能最好。

2.1.4　自动机器学习系统

本次研究采用的自动机器学习系统为Auto-sklearn2.0[34]。Auto-sklearn在2016年首次由FEURER等提出[35],它能够较好地实现上述3项技术。与其他机器学习算法和Auto-WEKA,Hyperopt-sklean等比较成熟的自动机器学习系统相比,其在多数数据集上性能更优[35]。Auto-sklearn2.0在老版本基础上,对模型选择、算法组合与策略自动化这3个方面进行了改善,这些算法的优化使得新版本的计算精度相比老版本提高了5倍[34],笔者采用该自动机器学习系统运算24 h的结果。

2.2　常规机器学习

2.2.1　数据预处理

为了消除特征之间数量级差异的影响,对特征集进行了标准化,即:

$$z_i=\frac{x_i-\mu}{\sigma} \tag{2}$$

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{3}$$

$$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)} \tag{4}$$

另外,按照75%和25%的比例将数据集随机地划分为训练集和测试集。其中,训练集用于模型训练和超参数优化,测试集不参与训练过程,仅用于评价模型的预测性能。

2.2.2　特征重要性评估及特征选择

评估特征重要性有助于特征选择,进而提高模型性能。封装法能够根据机器学习模型预测效果对特征的重要性进行评分,相比单因素分析更能体现特征对目标变量的影响程度。采用基于随机森林的封装法评估特征的重要性并进行特征选择,其具体实现方法为:对于每一个特征,将随机森林中每一个子决策树上该特征形成节点的Gini指数下降程度进行求和,用这个指标来衡量特征的重要性[36]。与此同时,为了使特征重要性的计算结果更加稳定,运算过程采取7折交叉训练的形式。然后,采用贪婪过程来进行特征选择,即根据特征重要性程度由大到小,向算法模型中逐一加入特征,进而得到特征数量与模型精度和稳定性的关系,最终选择使模型得分高且得分标准差低的特征组合作为下一步模型计算的输入变量。

2.2.3　模型训练与优化

采取7折交叉验证方式进行超参数优化，通过对比平均交叉验证误差来优选模型的所有超参数；然后，用优选的超参数在整个训练集上进行训练，并用从未参与模型训练的测试集来评价算法精度。7折交叉验证方式能够充分高效地利用数据，并能够稳健地评估超参数性能，减少因数据集随机划分而导致的模型不稳定性，避免模型的过拟合。

2.2.4　机器学习算法

为了与先进的自动机器学习进行比较，采取了随机森林[31]、支持向量回归[37]和神经网络[38]这3种较为成熟、在算法结构上差异较大且在大部分数据集上性能较好的机器学习算法，利用这几种算法进行建模，并对比算法之间的预测性能。

支持向量回归是支持向量机的一种形式，它使用不敏感损失系数作为损失函数：

$$L(z)=\max\left(0,\left|Y_i-\hat{Y}_i\right|-\varepsilon\right) \tag{5}$$

支持向量回归模型的主要超参数有核函数、不敏感损失系数、惩罚参数和宽度系数，将这几个参数作为待优化的超参数。

神经网络是一种由大量神经元相互联接构成的运算模型。当网络参数设置不当时，容易导致模型的过拟合现象。因此，为了提高模型的泛化能力，采用了早停技术（图2）、L2正则化、批量标准化和dropout的正则化方法。L2正则化表达式为：

$$L=E_{\text{in}}+\lambda\sum_{i=1}^{n_\omega}\omega_i^2 \tag{6}$$

神经网络的调整参数包括激活函数类型、优化器类型、学习率、神经元数和batch_size。

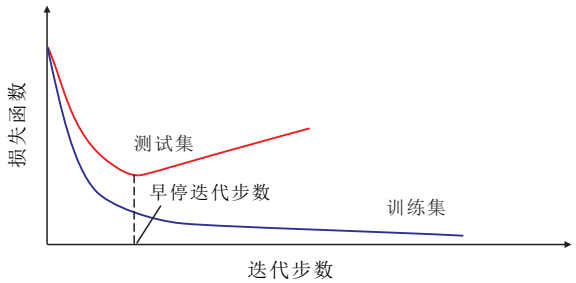


图2　早停技术示意

Fig.2　Early stopping technology

随机森林是由众多弱学习器（决策树）集合而成的一种强学习器。影响该算法精度的2个重要超参数分别为子模型的数量和节点分裂时参与判断的最大特征数，对这2个参数进行优化。

2.2.5　支持库、超参数调整及模型评估

采用Scikit-learn模型包[39]来实现支持向量回归和随机森林算法，用基于Python的Keras[40]来构建神经网络模型。超参数调整则通过网格搜索来完成。将决定系数和均方误差作为评估模型性能的指标，其表达式分别为：

$$R^2=1-\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2} \tag{7}$$

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \tag{8}$$

3　结果分析与讨论

3.1　数据统计分析

各个变量之间的线性相关性用Pearson相关系数进行描述（图3）。其中，0，1和-1分别为完全线性无关、完全线性正相关和完全线性负相关。另外，采用箱线图描述压裂后日产油量与影响因素之间的关系（图4）。每一个箱线的区间为该组分类数据的第1个四分位到第3个四分位。

从图3可以看出，措施前日产油量与压裂后日产油量的线性相关性最强，Pearson相关系数为0.70。措施前含水率、日产液量也与目标变量有着较强的相关性，同时，图4也表征出这些措施前生产数据对压裂后日产油量有着较为明显的影响。分析可知，油井实施压裂的大部分原因是由于近井地带储层堵塞等问题造成了产液能力下降，这类问题导致的产液能力下降是一个渐变的过程，而当油井日产液量有一定幅度的异常下降时，就会及时根据情况开展压裂等措施，很少会等待日产液量下降至原来的一半甚至更少时才采取补救措施，因此，油井日产液量、日产油量这2个指标能在很大程度上描述油井压裂后的产油潜力。另外，压裂前含水率对压裂后日产油量也有较大的影响，这是由于油井压裂前含水率能够在一定程度上描述油井周围储层的含油情况，部分油井周围储层含油情况较好，含水率较低，但由于日产液量下降等原因造成日产油量较低，需要通过压裂来增产，这类油井在压裂后产油效果也较好。

Pearson相关系数只能描述变量之间的线性关系，箱线图也仅能定性地查看影响因素与目标变量的变化趋势，而且2种分析方式均为单因素分析。然而，在实际的压裂问题中，压裂后日产油量与影响因素之间可能存在非常复杂的非线性关系，而且受多因素同时影响。因此，需要探究更加适合的方

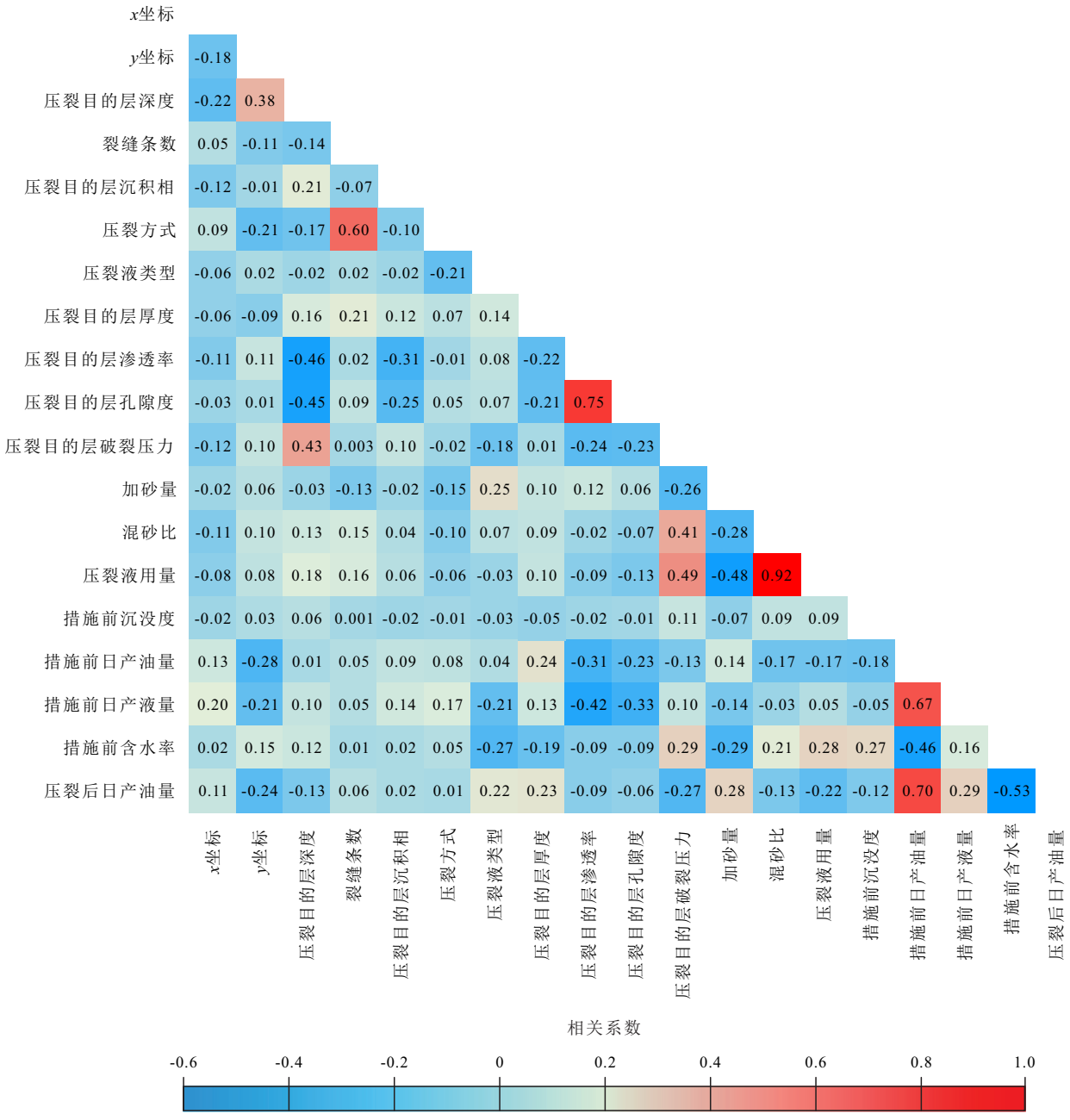


图3　影响压裂后日产油量各因素相关性热力图

Fig.3　Heat map of correlation between various factors affecting oil production rate after hydraulic fracturing

法来进一步评价各个特征对于目标变量的影响程度。

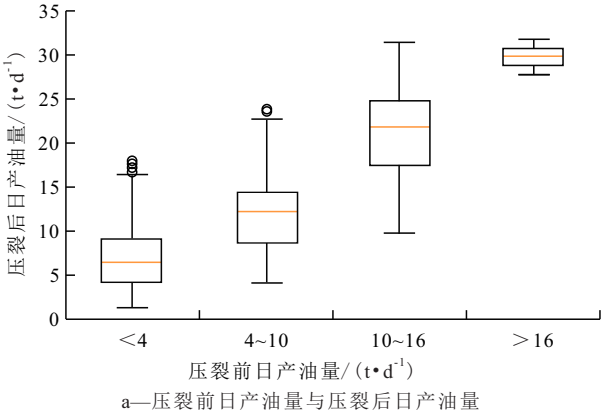
3.2　特征重要性分析及特征选择

基于随机森林封装法的特征重要性评价结果如图5所示。总体来看，压裂后日产油量受压裂前各项生产因素影响最大，其次是地质和工程因素。这意味着对采油井生产情况进行实时监测，并依据动态指标合理地选择压裂井和目的层对于措施效果更为关键。另外，加砂量和压裂液体积也对压裂后日产油量起了重要作用。这是因为这2个因素与形成裂缝的长度和导流能力有一定的关系。

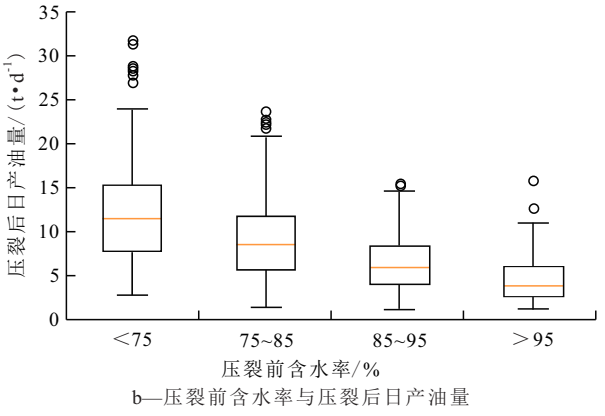
分析模型R^2可知（图6），模型的拟合精度在特征数量为6后逐渐平缓；特征数量达到15时，模型不但获得了较高的精度，且达到了非常低的标准差，即稳定的性能。因此，按照重要程度选取前15个特征作为3种常规机器学习算法的输入变量进行计算。

3.3　机器学习预测模型建立

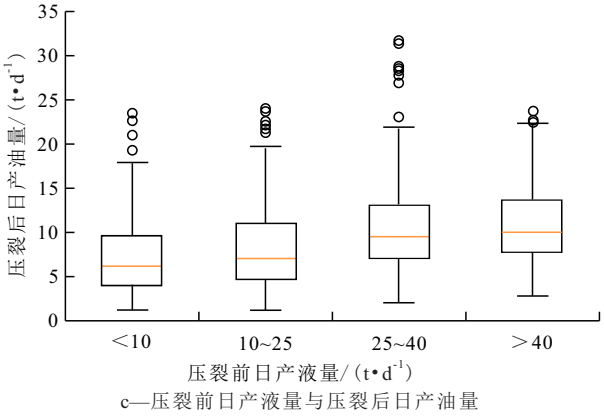
支持向量回归、神经网络、随机森林这3种常规机器学习算法与自动机器学习预测模型的预测性能对比结果（表1）显示，各种算法建立的模型在测试集上性能由好到差依次为自动机器学习、随机森



a—压裂前日产油量与压裂后日产油量



b—压裂前含水率与压裂后日产油量



c—压裂前日产液量与压裂后日产油量

图4　影响因素与压裂后日产油量关系箱线图

Fig.4　Box plot of relationship between influencing factors and oil production rate after hydraulic fracturing

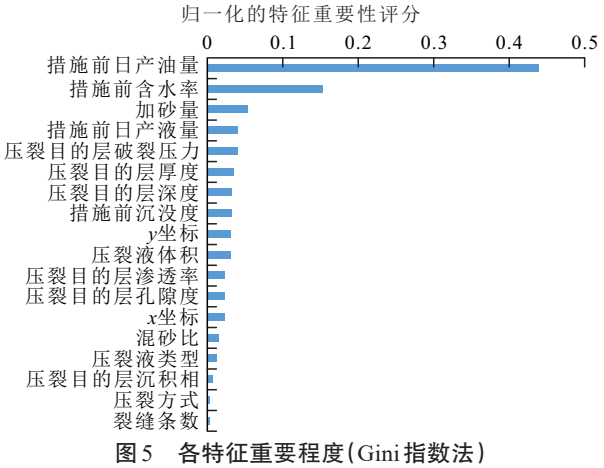


图5　各特征重要程度(Gini指数法)

Fig.5　Feature importance based on Gini index

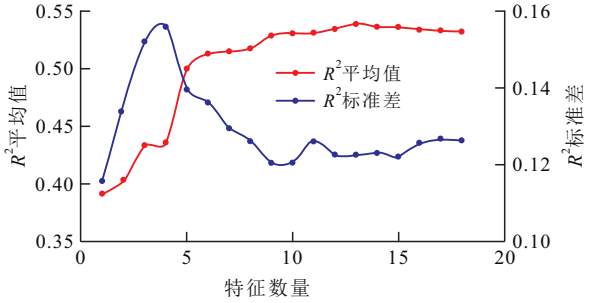


图6　交叉验证中模型R^2平均值与标准差随特征数量变化关系

Fig.6　Relationship between mean value and standard deviation of R^2 with feature number in cross validation

表1　各种算法在数据集上的预测性能

Table1　Prediction performance of algorithms on data set

算法	训练集			测试集		
	R^2	均方误差	相对误差/%	R^2	均方误差	相对误差/%
自动机器学习	0.728	6.82	16.18	0.695	7.81	18.96
随机森林	0.937	3.87	8.25	0.609	9.47	25.15
神经网络	0.607	9.75	26.82	0.584	10.09	29.97
支持向量回归	0.609	9.70	25.57	0.574	10.32	31.59

林、神经网络和支持向量回归。随机森林虽然在测试集上也展现了较好的性能，但是其在训练集上的R^2过高，说明该模型存在着较为严重的过拟合现象，模型的泛化能力较弱。从图7可以直观地看出，自动机器学习预测模型在训练集和测试集上的预测结果均较好。自动机器学习预测模型在测试集上的R^2为0.695，均方误差为7.81，预测结果相对误差的平均值为18.96%，标准差为16.97%，好于其他算法，因此优选其为最佳的压裂效果预测模型。为了对比该模型在现有预测水平上的提升效果，从采油与地面工程运行管理系统中提取压裂方案，查询压裂后日产油量预测值，计算实际压裂后日产油量和方案预测值之间的相对误差，统计得出测试集的

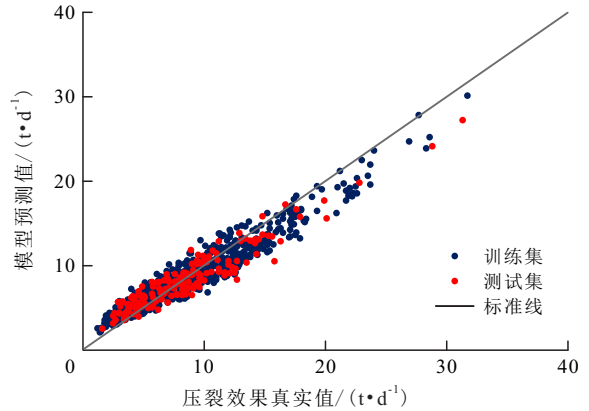


图7　自动机器学习预测模型的预测值与真实值对比

Fig.7　Comparison between predicted values of automatic machine learning prediction model and real values

方案预测相对误差平均值为76.49%，标准差为78.52%。对比可知，本研究建立的压裂效果预测模型可在目前水平上将预测相对误差的平均值降低57.53%，标准差降低61.55%，预测精度与稳定性均大幅提高。最优自动机器学习预测模型中各基础预测器参数信息见表2。

表2　最优自动机器学习预测模型中各基础预测器信息

Table2　Information of base regressors in optimal automatic machine learning model

编号	回归器类型	权重比例	模型参数	
			参数名称	参数值
1	libsvm_svr	0.4	C'	426.10
			ε	0.586
			kernel	sigmoid
2	ard_regression	0.4	$\alpha1$	4.79×10^{-4}
			$\alpha2$	2.51×10^{-4}
			$\lambda1$	1.58×10^{-9}
			$\lambda2$	4.07×10^{-6}
3	random forest	0.18	*n_iter*	300
			n_estimators	190
			max_*features*	12
			max_*depth*	9
			min_samples_*split*	21
4	libsvm_svr	0.02	C'	141.28
			ε	0.626
			kernel	sigmoid

4　应用情况

4.1　经济效益测算

利用本研究建立的模型可大幅提高压裂后日产油量预测精度，进一步辅助压裂方案的制定与优化，最大程度提高压裂投资产生的经济效益。具体经济效益测算过程如下：①选取8口已压裂井，利用自动机器学习预测模型对压裂参数进行重新优化，得到每口井的最优压裂方案参数与压裂效果预测值。②利用上述模型在测试集上的相对误差平均值来估算每口井措施效果的范围，并与原方案的实际效果进行对比，求取参数优化后相比于原方案增加的初期日增油量。③压裂有效期选取4个月，假设模型优化方案比原方案额外日增油量在压裂有效期内按指数关系递减（图8），有效期末额外日增油量趋于0，再通过求取积分估算有效期内总的额外增油量。④选取油价为70美元/bbl，汇率为6.37，通过计算即可得到模型优化方案相比原方案的额外经济效益。从模型优化增加经济效益测算结果（表3）可以看出，经过模型优化压裂参数后，选取的8口井相比原方案平均可额外增加经济效益3.2×10^4~27.4×10^4元/井次，平均为16.1×10^4元/井次。额外总增油量的表达式为：

$$V_{\text{总增油量}}=\int_0^{t_{\max}}ae^{bt}\,\mathrm{d}t \tag{9}$$

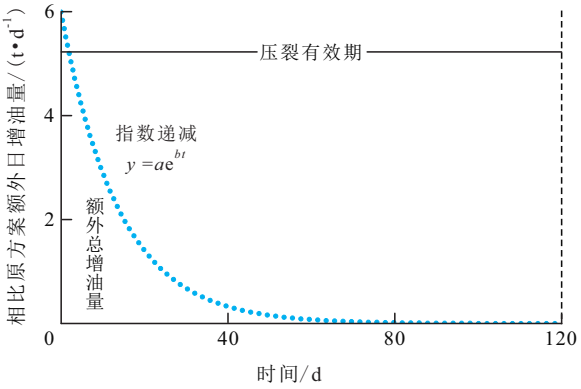


图8　自动机器学习预测模型优化后压裂方案较原方案额外增油量示意

Fig.8　Extra oil increment of optimal scheme compared with the original scheme

4.2　矿场应用案例

为了进一步验证模型的精度与现场应用的可行性，利用上述建立的自动机器学习预测模型对4口采油井进行压裂方案的设计与优化。

监测N23区块采油井日产液量、产液剖面等各项动态指标变化情况，结合井组注采关系、连通关系等地质条件与地层能量、剩余油饱和度等开发参数，选定4口采油井开展压裂措施。压裂方案设计优化涉及地质和工程2个方面的参数。首先，模型考虑的各项地质因素主要受压裂目的层位的影响。分析各采油井的生产层位、产液剖面和连通情况等信息，每口采油井初选3个目的层进行排列组合。压裂工程方面选取压裂液类型和加砂量这2个对模型效果影响较大的参数。加砂量选取12，15，18，21，24这5个数值，压裂液类型则将胍胶压裂液、缔合压裂液和其他压裂液作为待选。将每口井的所有选层、加砂量及压裂液类型进行逐一组合，可得到90个压裂方案，详见表4。将所有压裂方案参数输入自动机器学习预测模型进行计算，优选出各采油井的最佳方案如表4所示。2019年6月开始，按最优方案对这4口采油井实施了压裂改造，实际效果（表5）表明，压裂后采油井的日产油量和模型预测值非常接近，本文建立的自动机器学习预测模型对实际矿场的预测性能较好。该自动机器学习预测模型对大庆油田N23区块的水力压裂设计与优化

表3　模型优化增加经济效益测算结果

Table3　Economic benefits increased by model optimization

井号	按原方案压裂后实际日产油量/(t·d⁻¹)	经模型优化参数后预测值/(t·d⁻¹)						相比原压裂方案增加经济效益/(10⁴元·井次⁻¹)		
		压裂后日产油量（限值由±平均误差计算）			相比原压裂方案额外日增油量					
		平均值	上限	下限	平均值	上限	下限	平均值	上限	下限
N23-40-425	5.5	8.6	10.2	7.0	3.1	4.7	1.5	15.3	22.0	8.1
N23-41-423	13.8	17.4	20.7	14.1	3.6	6.9	0.3	17.4	31.0	2.0
N23-4-B127	14.5	17.9	21.3	14.5	3.4	6.8	0.0	16.6	30.6	0.0
N23-4-SB444	8.8	11.2	13.3	9.1	2.4	4.5	0.3	12.2	21.2	2.0
N23-4-W455	4.5	7.5	8.9	6.1	3.0	4.4	1.6	14.8	20.8	8.6
N23-51-429	17.1	20.8	24.7	16.9	3.7	7.6	−0.2	17.9	33.8	−1.5
N23-51-434	5.1	7.8	9.3	6.3	2.7	4.2	1.2	13.5	20.0	6.7
N23-D51-431	19.7	23.9	28.4	19.4	4.2	8.7	−0.3	20.0	38.1	−2.0
平均值	11.1	14.4	17.1	11.7	3.3	6.0	0.5	16.1	27.4	3.2

表4　压裂参数优化情况

Table4　Optimization results of hydraulic fracturing parameters

井号	压裂目的层		加砂量/m³		压裂液类型	
	待选参数	模型优化结果	待选参数	模型优化结果	待选参数	模型优化结果
N23-163-483	GⅡ28—GⅡ34,GⅢ2—GⅢ4,GIV4—GIV6	GⅡ28—GⅡ34,GⅣ4—GⅣ6	12,15,18,21,24	15	胍胶压裂液、缔合压裂液、其他压裂液	缔合压裂液
N23-51-430	SⅠ2—SⅠ4+5,SⅡ6—SⅡ8,SⅡ13—SⅢ1	SⅠ2—SⅠ4+5,SⅡ13—SⅢ1		24		胍胶压裂液
N23-D41-428	SⅠ2,SⅠ3—SⅠ4+5,SⅡ1—4	SⅠ2,SⅡ1—4		12		胍胶压裂液
N23-D41-436	SⅠ1—SⅠ3,SⅠ4+5—SⅡ1,SⅡ2+3—SⅡ4	SⅠ1—SⅠ3,SⅠ4+5—SⅡ1		12		胍胶压裂液

注:G代表高台子油层,S代表萨尔图油层,Ⅰ,Ⅱ,Ⅲ,Ⅳ表示油层组序号,GⅡ28表示高台子油层Ⅱ油层组28小层。

表5　最优压裂方案效果预测及施工后实际值

Table5　Effect prediction of optimal hydraulic fracturing scheme and actual values after hydraulic fracturing

井号	压裂后日产油量/(t·d⁻¹)		预测相对误差/%
	模型预测值	现场实际	
N23-163-483	10.3	9.4	9.23
N23-51-430	11.1	11.4	2.44
N23-D41-428	12.4	11.3	9.81
N23-D41-436	10.9	10.1	7.57
平均值			7.26

具有指导意义,模型经过简单的参数修改便可预测其他开发区块,简单易用,推广性强。

5　结论

依据矿场压裂统计数据分析了大庆油田N23区块的采油井压裂增产情况,采用自动机器学习建立了一个精度高、稳定性强的采油井压裂效果预测模型。各项统计结果表明,采油井水力压裂的各影响因素与压裂后日产油量存在定性关系;利用封装法进行了特征重要性评估,得到对模型影响较大的特征为:压裂前日产油量、含水率、加砂量等。利用自动机器学习建立的预测模型精度高于支持向量回归、神经网络和随机森林这3种常见机器学习算法,模型在测试集上的精度高达0.695,预测相对误差仅为18.96%,比目前降低了57.53%。经过模型优化的压裂方案较原方案平均可额外增加经济效益约3.2×10⁴~27.4×10⁴元/井次;另外,利用自动机器学习预测模型对N23区块采油井压裂增产方案进行了优化,结果显示,模型计算出的最优参数组合方案实际效果较好,而且现场的实施效果与模型预测值非常相近。自动机器学习预测模型对N23区块压裂措施参数设计具有指导作用,现场可行性高,模型推广性强。

符号解释

a,b——回归公式系数;

C'——支持向量回归算法中的惩罚因子;

$d_p(D_{frac},D_j)$——目标数据集与数据库中第j个数据集之

间的L1范数；

D_{frac}——目标数据集；

D_j——数据库中第j个数据集；

E_{in}——未包含正则化项的训练样本误差；

i——元特征个数；

j——数据集个数；

k——交叉验证的折数；

kernel——支持向量回归算法中的核函数类型；

L——损失函数；

$L(z)$——不敏感损失函数；

m_i^{frac}——目标数据集中第i个元特征值；

m_i^j——第j个数据集中第i个元特征值；

max_*features*——random forest算法中寻找最佳分割时要考虑的特征数量；

max_*depth*——random forest算法中树的最大深度；

min_samples_*split*——random forest算法中拆分内部节点所需的最少样本数；

MSE——均方误差；

n——样本的数量；

n_estimators——random forest算法中决策树的数量；

n_iter——ard_regression算法中的最大迭代次数；

n_ω——待学习参数的数量；

R^2——决定系数；

t——压裂后生产天数，d；

t_{max}——压裂有效期，d；

$V_{总增油量}$——额外总增油量，t；

x_i——第i个样本特征在标准化前的数值；

$X_{测试集}$——测试集特征；

$X_{训练集}$——训练集特征；

$\bar{Y}$——所有样本目标变量的平均值；

Y_i——第i个样本目标变量的实际值；

$\hat{Y}_i$——第i个样本目标变量的模型预测值；

$Y_{训练集}$——训练集目标变量值；

$\hat{Y}_{测试集}$——测试集目标变量预测值；

z_i——第i个样本特征在标准化后的数值；

$\alpha1$，$\alpha2$，$\lambda1$，$\lambda2$——ard_regression算法中的模型系数；

ε——不敏感损失系数；

λ——正则化参数；

μ——所有样本的平均值；

σ——所有样本的标准差；

ω_i——第i个网络层待学习参数。

参考文献

[1] 邓刚.压裂驱油试验邻井压窜风险预警方案及其应用[J].大庆石油地质与开发，2022，41(1)：91-96.
DENG Gang. Early warning scheme and its application for the pressure channeling risk in adjacent wells of the fracturing-flooding test[J].Petroleum Geology & Oilfield Development in Daqing，2022，41(1)：91-96.

[2] 马玉娟.大庆长垣油田三类油层压裂驱油提高采收率技术及其应用[J].大庆石油地质与开发，2021，40(2)：103-109.
MA Yujuan. Application of fracturing-flooding EOR technique in Type Ⅲ oil reservoirs in Daqing Placanticline Oilfield[J]. Petroleum Geology & Oilfield Development in Daqing，2021，40(2)：103-109.

[3] 任佳伟，王贤君，张先敏，等.大庆致密油藏水平井重复压裂及裂缝参数优化模拟[J].断块油气田，2020，27(5)：638-642.
REN Jiawei，WANG Xianjun，ZHANG Xianmin，et al. Refracturing and fracture parameters optimization simulation for horizontal well in Daqing tight oil reservoir[J]. Fault-Block Oil and Gas Field，2020，27(5)：638-642.

[4] 陈文将.聚驱后压堵结合工艺现场试验[J].大庆石油地质与开发，2020，39(5)：111-116.
CHEN Wenjiang. Field test of the fracturing-plugging combined technology after the polymer flooding[J]. Petroleum Geology & Oilfield Development in Daqing，2020，39(5)：111-116.

[5] HANGA K M，KOVALCHUK Y. Machine learning and multi-agent systems in oil and gas industry applications：A survey[J]. Computer Science Review，2019，34：100191.

[6] SHAHEEN M，SHAHBAZ M，REHMAN Z U，et al. Data mining applications in hydrocarbon exploration[J]. Artificial Intelligence Review，2011，35(1)：1-18.

[7] 马俊修，石胜男，陈进，等.基于机器学习的玛湖地区水平井压裂设计优化[J].深圳大学学报：理工版，2021，38(6)：621-627.
MA Junxiu，SHI Shengnan，CHEN Jin，et al. Optimization of fracture design for horizontal wells in Mahu region based on machine learning[J]. Journal of Shenzhen University：Science and Engineering，2021，38(6)：621-627.

[8] 李宁，徐彬森，武宏亮，等.人工智能在测井地层评价中的应用现状及前景[J].石油学报，2021，42(4)：508-522.
LI Ning，XU Binsen，WU Hongliang，et al. Application status and prospects of artificial intelligence in well logging and formation evaluation[J]. Acta Petrolei Sinica，2021，42(4)：508-522.

[9] 陈雁，焦世祥，程超，等.基于自编码器的半监督隔夹层识别方法[J].特种油气藏，2021，28(1)：86-91.
CHEN Yan，JIAO Shixiang，CHENG Chao，et al. Semi-supervised interlayer identification method based on self-encoder[J]. Special Oil & Gas Reservoirs，2021，28(1)：86-91.

[10] 王相，杨耀忠，何岩峰，等.基于深度学习的油井工况智能诊断技术研究及应用[J].油气地质与采收率，2022，29(1)：181-189.
WANG Xiang，YANG Yaozhong，HE Yanfeng，et al. Research and application of intelligent diagnosis technology of oil well working conditions based on deep learning[J]. Petroleum Geology and Recovery Efficiency，2022，29(1)：181-189.

[11] 翟亮.基于XGBoost算法的吸水剖面预测方法研究与应用[J].油气地质与采收率，2022，29(1)：175-180.
ZHAI Liang. XGBoost-based water injection profile prediction method and its application[J]. Petroleum Geology and Recovery Efficiency，2022，29(1)：175-180.

[12] 周恒，武中原，张欣，等.基于LSTM循环神经网络的横波预测方法[J].断块油气田，2021，28(6)：829-834.

ZHOU Heng, WU Zhongyuan, ZHANG Xin, et al. Shear wave prediction method based on LSTM recurrent neural network [J]. Fault-Block Oil and Gas Field, 2021, 28(6): 829-834.

[13] 史长林,魏莉,张剑,等.基于机器学习的储层预测方法[J].油气地质与采收率,2022,29(1):90-97.
SHI Changlin, WEI Li, ZHANG Jian, et al. Reservoir prediction method based on machine learning[J]. Petroleum Geology and Recovery Efficiency, 2022, 29(1): 90-97.

[14] 马陇飞,萧汉敏,陶敬伟,等.基于梯度提升决策树算法的岩性智能分类方法[J].油气地质与采收率,2022,29(1):21-29.
MA Longfei, XIAO Hanmin, TAO Jingwei, et al. Intelligent lithology classification method based on GBDT algorithm[J]. Petroleum Geology and Recovery Efficiency, 2022, 29(1): 21-29.

[15] SHELLEY B, GRIESER B, JOHNSON B J, et al. Data analysis of Barnett Shale completions[J]. SPE Journal, 2008, 13(3): 366-374.

[16] MONTGOMERY J B, O'SULLIVAN F M. Spatial variability of tight oil well productivity and the impact of technology [J]. Applied Energy, 2017, 195(C): 344-355.

[17] AWOLEKE O O, LANE R H. Analysis of data from the Barnett Shale using conventional statistical and virtual intelligence techniques [J]. SPE Reservoir Evaluation & Engineering, 2011, 14(5): 544-556.

[18] 蒋文超.基于机器学习与模型融合的大庆油田SN区块油井压裂效果预测技术[J/OL].大庆石油地质与开发:1-9[2022-07-11].https://doi.org/10.19597/J.ISSN.1000-3754.202109045.
JIANG Wenchao. Prediction model for production well hydraulic fracturing effect of Block SN in Daqing Oilfield based on machine learning and model ensemble[J/OL]. Petroleum Geology & Oilfield Development in Daqing: 1-9[2022-07-11]. https://doi.org/10.19597/J.ISSN.1000-3754.202109045.

[19] 周济民,张海晨,王沫然.基于物理经验模型约束的机器学习方法在页岩油产量预测中的应用[J].应用数学和力学,2021,42(9):881-890.
ZHOU Jimin, ZHANG Haichen, WANG Moran. Machine learning with physical empirical model constraints for prediction of shale oil production[J]. Applied Mathematics and Mechanics, 2021, 42(9): 881-890.

[20] 檀朝东,贺甲元,周彤,等.基于PCA-BNN的页岩气压裂施工参数优化[J].西南石油大学学报:自然科学版,2020,42(6):56-62.
TAN Chaodong, HE Jiayuan, ZHOU Tong, et al. A study on the optimization of fracturing operation parameters based on PCA-BNN [J]. Journal of Southwest Petroleum University: Science & Technology Edition, 2020, 42(6): 56-62.

[21] 严子铭,王涛,柳占立,等.基于机器学习的页岩气采收率预测方法[J].固体力学学报,2021,42(3):221-232.
YAN Ziming, WANG Tao, LIU Zhanli, et al. Machine-learning-based prediction methods on shale gas recovery[J]. Chinese Journal of Solid Mechanics, 2021, 42(3): 221-232.

[22] 周志军,薛江龙,李慧敏,等.压裂后增油量预测模型[J].特种油气藏,2013,20(3):76-78.
ZHOU Zhijun, XUE Jianglong, LI Huimin, et al. Study on prediction model of oil increment post-fracturing[J]. Special Oil & Gas Reservoirs, 2013, 20(3): 76-78.

[23] REIF M, SHAFAIT F, DENGEL A. Meta-learning for evolutionary parameter optimization of classifiers[J]. Machine Learning, 2012, 87(3): 357-380.

[24] VAN RIJN J N, BISCHL B, TORGO L, et al. OpenML: A collaborative science platform[C]. Berlin: Machine Learning and Knowledge Discovery in Databases, 2013.

[25] KALOUSIS A. Algorithm selection via meta-learning[D]. Geneve: University of Geneve, 2002.

[26] BROCHU E, CORA V M, DE FREITAS N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning[EB/OL]. [2010-12-14]. https://arxiv.org/abs/1012.2599.

[27] EGGENSPERGER K, FEURER M, HUTTER F, et al. Towards an empirical foundation for assessing Bayesian optimization of hyperparameters[C]. Lake Tahoe: NIPS Workshop on Bayesian Optimization in Theory and Practice, 2013.

[28] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[C]. Lake Tahoe: Proc. of NIPS'12, 2012.

[29] HUTTER F, HOOS H H, LEYTON-BROWN K. Sequential model-based optimization for general algorithm configuration [C]. Rome: Proc. of LION'11, 2011.

[30] BERGSTRA J, BARDENET R, KÉGL B, et al. Algorithms for hyper-parameter optimization[C]. Granada: Proc. of NIPS'11, 2011.

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

[32] LACOSTE A, MARCHAND M, LAVIOLETTE F, et al. Agnostic Bayesian learning of ensembles[C]. Beijing: Proc. of ICML'14, 2014.

[33] CARUANA R, NICULESCU-MIZIL A, CREW G, et al. Ensemble selection from libraries of models[C]. Banff: Proc. of ICML'04, 2004.

[34] FEURER M, EGGENSPERGER K, FALKNER S, et al. Auto-sklearn2.0: the next generation[EB/OL]. [2020-07-13]. https://www.automl.org/auto-sklearn-2-0-the-next-generation/.

[35] FEURER M, KLEIN A, EGGENSPERGER K, et al. Efficient and robust automated machine learning[C]. Barcelona: Proc. of NIPS'16, 2016.

[36] ROBIN Genuer, JEAN-MICHEL Poggi, CHRISTINE Tuleau-Malot. Variable selection using Random Forests[J]. Pattern Recognition Letters, 2010, 31(14): 2 225-2 236.

[37] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[38] LECUN Yann, BENGIO Yoshua, HINTON Geoffrey. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[39] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12(10): 2 825-2 830.

[40] CHOLLET F. Keras[EB/OL]. [2021-09-03]. https://github.com/fchollet/keras.

编辑 刘北羿