基于多模型融合 Stacking 集成学习的油田产量预测

张庭婷[1]，潘美琪[2]，朱天怡[3]，曹 煜[2]，张站权[2]，刘单珂[2]，贺 兴[3]，于立军[2]

（1. 上海交通大学 碳中和发展研究院，上海 200230；2. 上海交通大学 智慧能源创新学院，上海 200240；
3. 上海交通大学 电子信息与电气工程学院，上海 200240）

摘要：基于机器学习前沿理论，提出一种基于多模型融合 Stacking 集成学习方式的组合预测方法，以国内某特高含水油田区块中多口水驱产油井历年生产历史数据为试验样本，预测其动态产油量。依据不同算法的训练原理，选取极限梯度提升树算法、长短记忆网络（LSTM）、时域卷积网络（TCN）等作为模型的基学习器，采用多元线性回归作为模型的元学习器。结果表明：融合后的 Stacking 模型充分发挥了各基学习器的优势，相比单一模型，融合后的 Stacking模型预测平均误差较小，预测鲁棒性较好。该模型的提出对融合模型在特高含水油藏开发方面具有重要的应用意义。

关键词：多模型融合；Stacking 集成学习；极限梯度提升树；长短期记忆网络；时域卷积网络；产量预测

中图分类号：TE341；TP181 **文献标志码**：A **文章编号**：1671－1807(2023)02－0263－09

油田开发指标的科学可靠预测是评价油田开发条件、编制油田开发计划、设计油田开发方案的基础和依据。油藏开发过程中，当油藏含水超过90％时，通常认为它处于特高含水期。相较中低含水油藏的开发，特高含水油藏阶段的显著特征主要表现在采油速度低、剩余油过于分散、措施效果变差、井况恶化、开发经济效益下降等方面[1-2]。目前中国大多数油田已进入高/特高含水期[3-4]，这给油田的高效开发提出了更高的要求，亟须科学可靠的开发指标预测（如产量预测等）来支撑油田的开发。

在对油田产量进行预测时，常规的方法是依据各种曲线理论，如递减曲线、增长曲线和水区曲线等[5-6]，但特高含水期利用水驱特征曲线等传统方法来预测产量时，常带来水驱特征曲线上翘的问题，使得预测误差增大。并且，进入特高含水期的油田，地层物性变化多样，开发情况复杂，原始数据获取困难，加上数据的残缺性和波动性，容易造成原有产量预测模型的结果与该阶段的真实生产状况偏离较大[7]。同时，传统的油藏数值模拟等方法花费高、时效性差。因此，对油藏开发工作而言，迫切需要开发出既能提高工作效率、又能提高预测精度的预测方法。

近些年，随着人工智能技术的发展，石油工业的智能化得以快速发展[8]，以机器学习为代表的人工智能技术贯穿于石油工业的上下游全产业链[9]，显著提高了油气全产业链的效率和效益[10]。在油气开发和生产中，机器学习被广泛应用于动态分析、生产指标预测、开发方案智能决策与优化及开发实验数据挖掘与应用等问题[10]。针对特高含水期油田产量预测，国内外学者都进行了大量的研究，使用的网络从 BP 神经网络拓展到结构相对复杂、层数较深的网络[11-13]。常被用于油田产量预测的机器算法有支持向量机（SVM）[14]、自回归（AR）、人工神经网络（ANN）[15-16] 和长短期记忆网络（LSTM）[17] 等，研究的重点包括数据预处理[18]、特征选择和学习器的选择[13]等。研究表明，学习器的选择会对预测结果产生很大影响，学习器选择不当会导致训练速度慢、预测精度低、拟合度过高等问题。文献[19]认为 LSTM 在油气生产率时间序列预测中优于 Simple RNN。此外，数据信息的多少也会影响预测结果，输入数据里涵盖不同的有用信息，其预测精度也往往不同。在对油田产量进行预测时，若仅凭预测误差的高低来舍弃误差较大的方法，往往会造成信息的丢失，从而降低预测的可靠性。基于多种方法综合集成的组合预测方法可有效解决这类问题。

收稿日期：2022-08-15

基金项目：中国石油化工股份有限公司科技重大专项（P20071-4）。

作者简介：张庭婷（1984—），女，重庆人，上海交通大学碳中和发展研究院，助理研究员，博士，研究方向为能源系统建模与大数据分析；通信作者于立军（1969—），男，吉林吉林人，上海交通大学智慧能源创新学院，院长助理，教授，博士，研究方向为能源大数据应用与风光发电智慧运维。

组合预测方法的本质是综合多个“好而不一样”的预测模型的有用信息，构成一个新的预测模型，旨在降低方差，提高指标的整体预测精度。平均法、投票法和学习法（Stacking）是 3 种常见的集成学习结合策略。其中，采用多折交叉验证的Stacking是相对更为稳健的一种模型融合策略。近两年，基于 Stacking 集成的模型融合预测被用于电力负荷预测[20-21]、交通流量预测[22]、火灾预测[23]、图像识别[24]等领域，取得了较好的效果。在石油开发领域，也出现了一些 Stacking 模型的应用探索。文献[25]在综合 Stacking、Boosting、Bagging 等集成方法优缺点的基础上，提出一种名为 B2S 的融合模型并将其应用于提高油气采收率，该模型在 426 个样本数据集上实现了 96.94%的测试精度。

特高含水油田开发系统错综复杂。在构建针对特高含水油田的动态产量预测模型时，需要在剖析常规多种预测方法的基础上，发展适合特高含水油田区块里不同产油井的组合预测方法，从而指导油田高效生产开发。本文将典型的和前沿的机器学习算法极限梯度提升树（XGBoost）、LSTM、时域卷积网络（TCN）等与集成学习方式进行融合，构建基于 Stacking 的组合预测方法，用于预测油田区块多个差异化油井的动态产量。首先对 Stacking 集成学习的理论框架以及各种算法的训练机理进行介绍；其次，考虑多口产油井的数据观测，建立 Stacking 框架下的多模型融合油田产量预测模型；最后，通过多口油井的产量预测验证算法的有效性。

1　算法理论介绍

1.1　模型框架

图 1 所示为 Stacking 集成学习的算法框架结构[26]。它首先将原始数据集按照一定的规则进行划分，划分后的若干子集输入到第 1 层预测模型中通过各基学习器进行训练，得到的预测结果为第 1 层模型的输出。然后将该预测结果输入到第 2 层模型，并利用该层的元学习器模型进行训练，得到最终的预测结果。Stacking 在训练基学习器时，需要使用交叉验证法，再在众多基学习器求出的预测结果基础上训练元学习器。集成学习方式可解决单个模型容易陷入欠拟合或过拟合的困境，得到泛化性能优良的学习器，从而改善模型的整体预测精度。

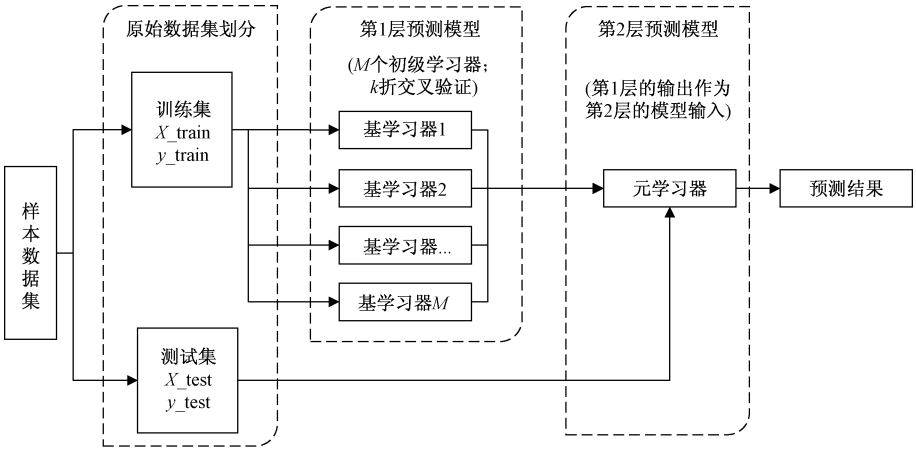


图 1　Stacking 集成学习算法框架结构

训练 Stacking 模型时，主要包括 3 个步骤[27]：

1）数据集划分与学习器确定。对于数据集 $A=\{(y_n,x_n),n=1,2,\cdots,N\}$，$x_n$ 代表样本 n 的特征向量，y_n 为样本 n 对应的预测值。借助随机采样的方式，将样本数据集划分为训练集 S 和测试集 T，同时确定基学习器的个数为 M 和元学习器的个数为 1。

2）进行 k 折交叉运算。将训练集 S 划分成 k 个大小相似的集合 $S_1,S_2,\cdots,S_k$，其中 $S_{-k}=S-S_k$，S_k 为 k 折交叉验证中的训练集，S_{-k} 则为测试集。在第 1 层预测模型中，对 S_k 采用基学习器算法 m 进行训练，得到基模型 L_m，$m=1,2,\cdots,M$。对 S_k 中的每个样本 x_n，用 z_{kn} 来表示基学习器 L_m 对它的预测。样本对应的预测值 y_n 和各基学习器预测得到的新结果 z_{kn} 一同构成新样本 S'，即 $S'=\{(y_n,z_{1n},z_{2n},\cdots,z_{kn}),n=1,2,\cdots,N\}$，新生成的 S' 将被用于训练元学习器。

3）元学习器训练。通过第 2 层预测模型对新生成的 S' 进行训练得到 L'，利用 M 个基学习器产生的 M 个测试指标集对元学习器进行检验和判断评价。

Stacking 融合方式的最大特征在于充分考虑第 1 层算法的特征,并通过第 2 层的结合策略发现第 1 层模型中各类算法的预测误差并及时纠正,从而可对模型的整体预测精度进行改善。

1.2 算法原理

1.2.1 极限梯度提升树模型

极限梯度提升树(XGBoost)是从传统梯度提升决策树改进和扩展而得到的集成树模型[28]。它的主要特色在于:不但可使用分类与回归树(CART),也能使用线性基础模型;可自定义损失函数,增加了模型构建的灵活性;在损失函数中加入了一个正则项,有助于模型泛化并防止模型过拟合;使用损失函数的二阶偏导数,可提供有关梯度方向的更多信息。

XGBoost 与 GBDT 类似,本质为加法模型,即是在每一步迭代中只优化当前步中的子模型,其目标函数为

$$\mathrm{Obj}=\sum_{i=1}^{N}L[F_m(x_i),y_i]+\sum_{j=1}^{m}\Omega(f_j)=\sum_{i=1}^{N}L[F_{m-1}(x_i)+f_m(x_i),y_i]+\sum_{j=1}^{m}\Omega(f_j)\quad(1)$$

式中:$F_m(x_i)$ 为第 m 步的子模型;$f_m(x_i)$ 为当前步的子模型;$F_{m-1}(x_i)$ 为训练完已经固定了的前 $m-1$ 个子模型;$\Omega(f)$ 表示子模型 f 的复杂度;$\sum_{j=1}^{m}\Omega(f_j)$ 为正则项,表示所有子模型复杂度之和。

运用二阶泰勒展开来可将 XGBoost 的损失函数近似表达,即

$$f(x_0+\Delta x)\approx f(x_0)+f'(x_0)\Delta x+\frac{f''(x_0)}{2}(\Delta x)^2\quad(2)$$

将式(1)中的 $F_{m-1}(x_i)$ 视作 x_0,$f_m(x_i)$ 视作 Δx,$L(\hat{y}_i,y_i)$ 视作 $\hat{y}_i$ 的函数,可得

$$\mathrm{Obj}=\sum_{i=1}^{N}L[F_{m-1}(x_i),y_i]+\frac{\partial L}{\partial F_{m-1}(x_i)}f_m(x_i)+\frac{1}{2}\frac{\partial^2 L}{\partial^2 F_{m-1}(x_i)}f_m^2(x_i)+\sum_{j=1}^{m}\Omega(f_j)\quad(3)$$

由于前 $m-1$ 个子模型已确定,故上式中除了 $f_m(x)$ 外都是常数,因此,目标函数可转为

$$\mathrm{Obj}=\sum_{i=1}^{N}[g_i f_m(x_i)+\frac{1}{2}h_i f_m^2(x_i)]+\Omega(f_m)\quad(4)$$

式中:$g_i=\frac{\partial L}{\partial F_{m-1}(x_i)}$,表示第 $m-1$ 轮损失函数对于第 i 个样本的一阶偏导;$h_i=\frac{\partial^2 L}{\partial^2 F_{m-1}(x_i)}$,表示第 $m-1$ 轮损失函数对于第 i 个样本的二阶偏导;L 代表损失函数,用来表征一次预测的好坏,损失函数越小代表模型越好。在 $F_{m-1}(x)$ 确定了的情况下,对每个样本点 i 都可以计算出一个 g_i 和 h_i。

对新的目标函数[式(4)]进行模型寻优,十分依赖 g_i 和 h_i。为了防止过拟合,有时会在模型更新公式 $F_m(x_i)=F_{m-1}(x_i)+f_m(x_i)$ 中在该轮训练的基学习器上乘以一个系数,称为学习速率(也称缩减系数)。学习速率的取值越小,每个轮次更新时"步长"越小,模型越保守。

1.2.2 Bagging

Bagging 是典型的并行类算法,其核心是基于自助采样(bootstrap sampling)[29]。假设数据集 D 有 m 个样本,需要在 D 中按照有放回的方式进行采样产生新的数据集 D',具体采用过程为:每次随机从 D 中挑选一个样本,将其拷贝放入 D',然后将该样本放回 D 中,再次进行采样,这次采样有可能采到上次放回的样本,如此重复执行 m 次后,得到数据集 D'。从以上过程中可以看出,D 中有一部分样本会在 D' 中多次出现,一部分样本始终不会在 D' 中出现。一般采用 D' 作为训练集,未被选中的作为测试集,测试结果亦称"包外估计"。"包外估计"的数据没有参与训练,可用来检验模型的泛化能力。

Bagging 算法是在自助采样的基础上完成的。首先在数据集中进行有放回的采样生产 m 个采样集;接着对每个采样集训练对应的基学习器;最后将这些基学习器进行按照一定的方式结合并输出结果。其中对分类任务采用简单投票法进行输出;对回归任务采用简单平均法进行结果输出。

1.2.3 长短期记忆网络

长短期记忆网络(LSTM)是在循环神经网络(RNN)的基础上改进得来[30],只是在结构上与一般的 RNN 略有不同。它的神经元在 RNN 单位元的基础上加了"门"结构机制(输入、遗忘、输出 3 种门),这些"门"可以捕获沿时间步的长期记忆和短期记忆,解决了标准 RNN 无法处理长期依赖信息的关键症结点——梯度消失和梯度爆炸问题[31]。

LSTM 网络的原理如图 2 所示。LSTM 的核心构成是细胞元组,其在时刻 t 的状态记为 C_t,计算公式为

$$C_t=f_t\times C_{t-1}+i_t\times\widetilde{C}_t\quad(5)$$

式中:f_t 为遗忘门,表示 $t-1$ 时刻单元状态 C_{t-1} 的哪些特征用于计算 C_t;i_t 为输入门;$\widetilde{C}_t$ 为新的记忆状态更新值,它是由当前状态输入数据 x_t 和上一时

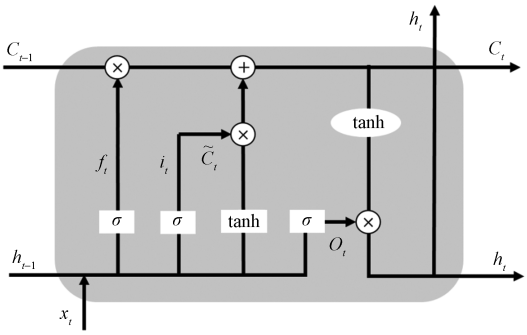


图 2　LSTM 的单元架构示意图[31]

刻单元组的状态 h_{t-1} 得到。f_t、i_t 和 $\widetilde{C}_t$ 等的计算公式如下：

$$f_t = \text{sigmoid}(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

$$i_t = \text{sigmoid}(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C) \quad (8)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \quad (9)$$

$$o_t = \text{sigmoid}(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t * \tanh(C_t) \quad (11)$$

式中：W_i、W_f、W_o 分别为输入门、遗忘门和输出门函数的权重项；b_i、b_f、b_o 分别为输入门、遗忘门和输出门的偏置项；* 为矩阵对应元素相乘。通常用 sigmoid 作为遗忘门、输入门的激活函数，其输出为一个介于[0,1]区间内的值。用 tanh 函数作为单元状态更新值的激活函数，将数值限制在 −1～1(关于原点对称)，适用于对信息的压缩。

可以看出，从 $t-1$ 时刻到 t 时刻 LSTM 传递的是长时记忆 C_{t-1} 和短时记忆 h_{t-1}，在本时刻中，各项参数主要依靠几个门控函数进行计算。遗忘门的作用是通过当前时刻输入和上一时刻传递的短时记忆信息对长时记忆信息进行选择，选择重要性较大的进行保留，否则将其舍弃。当前时刻的输入门用于决定当前时刻的输入有多少要被保存至长时记忆中。随后将经遗忘门和输入门的控制输出后的两部分新旧记忆相加得到当前时刻最终的长时记忆。

1.2.4　时域卷积网络

时域卷积网络(TCN)是在 CNN 和 RNN 的基础上发展起来的一种能够处理时间序列数据的网络结构，在特定条件下，处理效果优于传统的神经网络(RNN、CNN 等)，与 LSTM 类似，TCN 也避免了递归模型的常见缺陷——梯度爆炸/消失或缺乏内存保留等[32]。

TCN 的基本原理如图 3 所示。TCN 的显著特点是：①结构上的卷积是因果关系，即是没有从未来到过去的信息泄露；②体系结构可以获取任意长度的序列，并将其映射到相同长度的输出序列。对一维序列输入 $\boldsymbol{X} \in \mathbb{R}''$ 和卷积核 $f: \{0, \cdots, k-1\} \to \in \mathbb{R}''$，元素 s 的扩张卷积运算公式为

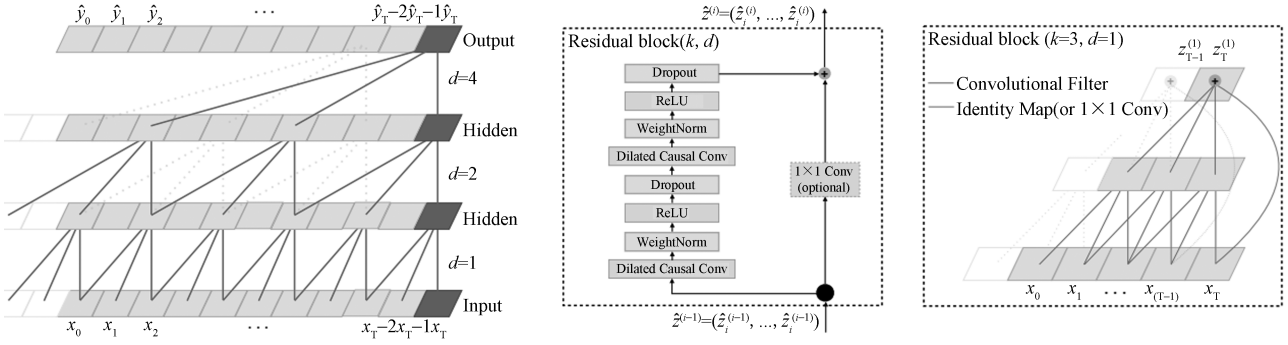
$$F(s) = (\boldsymbol{X} *_d f)(s) = \sum_{i=o}^{k-1} f(i) \cdot \boldsymbol{X}_{s-d \cdot i} \quad (12)$$

式中：d 为扩张系数；k 为卷积核大小；$s-d \cdot i$ 表示过去的方向。d 值的大小与神经网络的感受野有关，当 $d=1$ 时，扩张卷积为规则卷积，d 值越大，代表顶层的输出可代表更大范围的输入，从而有效扩展神经网络的感受野。因此，有两种方式可扩大 TCN 的感受野：一是选择大的卷积核 k，二是增加扩张系数 d。通常情况下，d 以指数形式增长，这确保有效命中历史记录中的每个输入。随着网络深度的增加，d 越大，代表可生成庞大的有效历史记录，如图 3(a)所示。

残差块[图 3(b)]包括一个分支，该分支引出一系列傅立叶变换 F，变换后的输出同 $\boldsymbol{X}$ 作为输入。

$$o = \text{Activation}[\boldsymbol{X} + F(x)] \quad (13)$$

在图 3(b)所示的残差块里，TCN 包括扩张因果卷积和非线性两层，对此采用修正线性单元(ReLU)，以防止梯度消失，提高模型的表达能力，使网络具有稀疏性。在归一化中，对卷积核采用权



(a) 扩张系数为1、2、4　　(b) TCN残差块　　(c) TCN中的残差连接示例

图 3　TCN 的网络架构示意图

重归一化。此外,每个扩张卷积后加入空间 Dropout 进行正则化,目的在于防止过拟合,提高模型的运输精度。TCN 的感受野取决于网络深度 n、卷积核大小 k 和扩张因子 d。稳定性对大而深的 TCN 网络而言至关重要。当残差输入和输出有不同的维度,会往其中添加一个 1×1 的卷积,这不仅可以使网络拥有跨层传递信息的功能,而且还能保证输入输出的一致性。

2 模型构建与实验设计

2.1 基于 Stacking 的融合模型构建

本文所研究的基于油田动态产量的预测属于回归问题,选用 Stacking 策略进行融合。根据 Krogh 和 Vedelsby 提出的"误差-分歧分解(error-ambiguity decomposition)"理论[33],在选择基学习器时应当遵循"好而不同"的原则,即考虑个体学习器性能好坏的同时,也要考虑个体学习器的两两不相似性。本研究先挑选 KNN、SVM、Bagging、Adaboost、XGBoost、随机森林、LSTM、TCN 等模型作为基学习器的候选模型,分别用来对特高含水油田的产量进行预测,然后挑选出预测性能较好的模型作为基学习器。经过初级筛选,发现 XGBoost、Bagging、LSTM、TCN 4 个模型在这 8 个模型中表现突出,本文选取这 4 个表现较优的模型作为基学习器。研究表明,元学习器为多元线性回归(MLR)时可以取得较好的预测效果[34],故本文选用 MLR 线性回归模型作为元学习器。图 4 为多模型 Stacking融合框架。

2.2 实验方案

选取中国东部某油田 6 口产油井(well A、well B、well C、well D、well E、well F)自开井以来的产油数据,建立动态产量预测模型。该油田为已进入进入特高含水阶段的水驱开发油田,含水率大于 95%。油田开发是复杂的过程,油井产量受影响的因素众多,包括注水方式、含水率、生产制度、剩余开采储量、措施增油量等[13]。初步研究表明,多个特征的加入可增加预测模型与工程实际的结合紧密度,但随着特征个数的增加,特征的冗余可能导致精度下降。在对特高含水油田进行动态产量预测时,常用油气的生产曲线作为预测的数据集。考虑到本研究是基于时间序列数据进行油井动态产量预测,主要目的是研究模型的融合方式对油井时序动态产量预测的影响,故本研究的样本数据集为选中油井的历史产油量、产水量、注入量等时间序列数据。

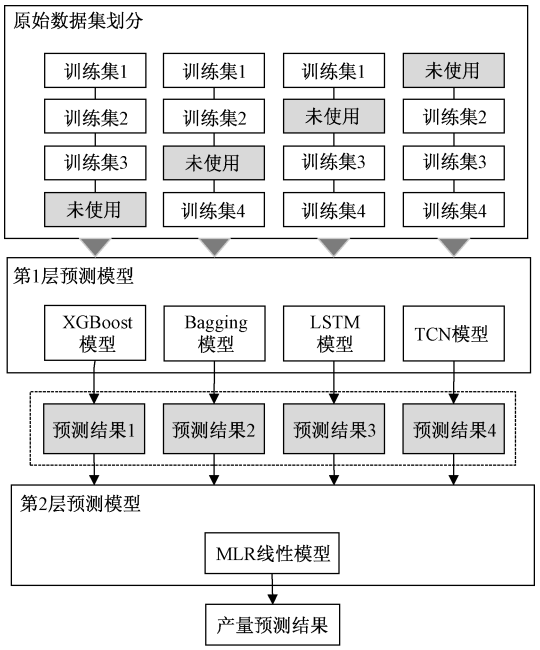


图 4 基于 Stacking 的多模型融合油井产量预测方法

因数据集中同一井生产过程中有个别点产量为零值的异常值,预测时先对数据集进行了零值消除。数据集时间跨度为 17 057 d(约 46.7 a),时间颗粒度前期为月、后期为日。数据集划分是选取数据集的前 85%为训练集,后 15%为测试集,采取 5 折交叉验证。模型训练时使用 Keras 搭建网络结构,采用 Python 3.3 作为实验程序编写的语言,同时使用了 Sklearn、Numpy 等第三方库计算技术指标。训练过程中的一些关键参数取值见表 1。

表 1 模型部分关键参数

基学习器	关键参数	取值
XGBoost	n_estimators	11
	gamma	0
	max_depth	3
	learning_rate	0.28
	reg_alpha	1×10^{-5}
	reg_lambda	1
Bagging	n_estimators	100
TCN	nb_filters	20
	kernel_size	6
	dilations	[2,4,8,16,32,64,128,256,512]
LSTM	units	100

为评判 Stacking 模型的预测效果,实验分别采用单一模型(XGBoost、Boosting、LSTM 和 TCN)以及 Stacking 融合模型对选定油井产量进行预测,并将预测结果进行对比分析。其中,Stacking 融合

模型考虑了两种融合策略：一是采用 XGBoost、Boosting、LSTM 和 TCN 作为基学习器的组合模型；二是采用随机森林（RF）、梯度提升决策树（GBR）、LightGBM、支持向量回归（SVR）、XGBoost、极端树回归（ETR）作为第一层模型的基学习器。具体设置见表 2。

表 2　实验组别设置

模型类别	编号	模型名称	基学习器种类
单一模型	1#	XGBoost	—
	2#	Bagging	—
	3#	LSTM	—
	4#	TCN	—
组合模型	5#	Stacking 模型 1（本文模型）	XGBoost、Bagging、LSTM、TCN
	6#	Stacking 模型 2（对照组）	RF、GBR、LightGBM、SVR、XGBoost、ETR

2.3　评价指标

为了对所设计的油田产量预测模型性能进行评估比较，本文采用平均绝对误差（MAE）和均方误差（MSE）两种常用的回归评价指标。假设测试集中样本的油井产油量为 y_i，预测模型得到的油井产油量为 $\hat{y}_i$，MAE 和 MSE 的评价方法如下：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{14}$$

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{15}$$

3　结果分析与讨论

3.1　单一模型预测结果对比分析

采用 XGBoost、Boosting、LSTM 和 TCN 4 种单一模型分别对 6 口产油井的日产油量进行预测，结果见表 3。从表中可知，XGBoost、Boosting、LSTM 和 TCN 4 个模型都表现出较好的预测性能。其中：XGBoost在产油井 A 和产油井 D 上表现最突出，预测的平均绝对误差最小；Boosting 在产油井 D 上表现出最低的均方误差 MSE；LSTM 整体表现优越，在产油井 C 和产油井 E 上的平均绝对误差和均方误差均最低，在产油井 A 上也表现出最低的均方误差；与 LSTM 类似，TCN 模型也表现突出，在产油井 B 和产油井 F 上都表现出最低的平均绝对误差和均方误差。可见，所选的 4 个基模型能实现油井产量的有效预测，不同模型在不同油井上的预测效果存在差异，这一方面说明了不同模型的差异化特征，同时也说明了油井样本数据的差异直接影响最优预测模型的选择。

表 3　6 口产油井单一模型预测效果评价　　单位：t

预测模型	Well A		Well B		Well C		Well D		Well E		Well F	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
1#	1.253	4.436	1.789	4.999	1.460	4.294	0.244	0.156	0.435	0.430	0.371	0.197
2#	1.581	6.647	1.834	5.001	1.430	4.292	0.257	0.130	0.537	0.473	0.342	0.227
3#	1.453	4.196	1.261	2.113	1.410	3.884	0.490	0.276	0.391	0.290	0.257	0.122
4#	1.555	4.594	1.005	1.341	1.413	4.012	0.291	0.158	0.443	0.433	0.224	0.098

注：1# 为 XGBoost；2# 为 Boosting；3# 为 LSTM；4# 为 TCN。

3.2　Stacking 融合模型预测性能分析

将构建的 Stacking 模型进行训练后用来预测选定油井的产油量。第 1 层模型选用 XGBoost、Boosting、LSTM 和 TCN 来进行构建，通过 5 折交叉运算以避免过拟合的发生，然后将基学习器训练得到的预测结果进行整合和维数转化输入到元学习器。第 2 层模型选择 MLR 作为元学习器进行训练，对第 1 层的基学习器集成学习，然后对最终结果进行预测。

由图 5 可知，采用 Stacking 模型对 6 口油井的日产量预测呈现较稳定的预测效果，但不同油井的产量预测曲线形态则有着明显差异。相比单一模型，Stacking 融合模型（5#）能够结合多种基学习器的优势，预测的平均误差较小（表 3），预测鲁棒性较好，但当预测准确度较差时，模型融合可能会增大误差。从理论层面上看，Stacking 集成模型的特色在于 Stacking 模型通过第 2 层模型的结合可摒弃第 1 层模型中各个基学习器预测效果较差的部分，从而将第 1 层模型中各基学习器的优势充分发挥。在对油井产量预测模型进行训练时，往往会给予一定的假设，这有可能引起预测的偏差，例如多个不同的假设在训练集上表现出性能相似的情况。此时，通过 Stacking 集成学习，可有效降低单一模型的泛化能力缺陷。此外，在对单一模型进行训练优化时，往往会陷入局部最小点，而局部最小并不代表模型的泛化能力好。通过多种基学习器进行结合，可有效避免模型陷入最小点。这也是采用 Stacking可显著提升预测精度的关键原因。

本研究设置了另一 Stacking 融合模型 2 对照组，以验证 Stacking 融合模型对油田产油井产量预

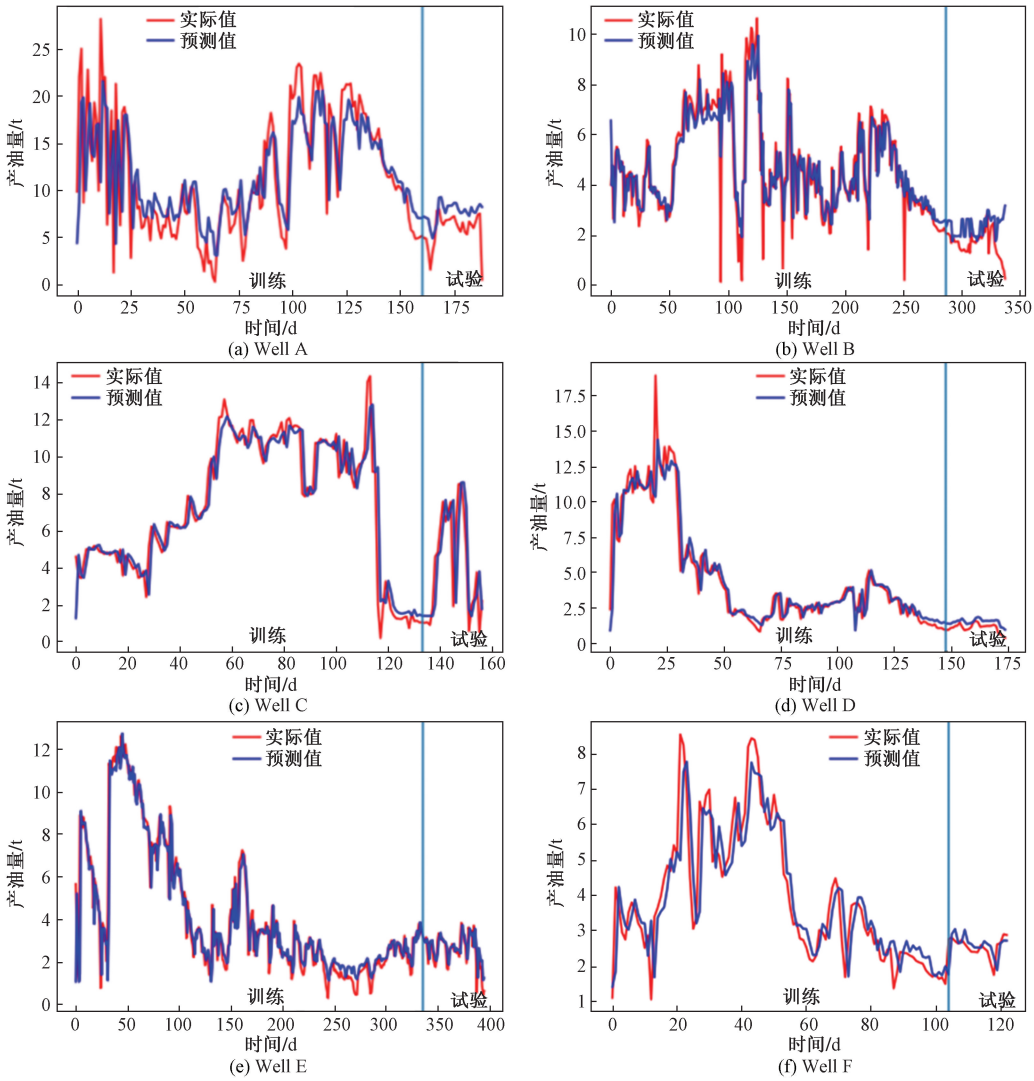


图 5　6 口产油井 Stacking 集成学习预测效果

测的有效性。在融合模型对照组(6 #)里,模型采用随机森林、梯度提升决策树、LightGBM、支持向量回归、XGBoost 和极端树回归作为基学习器,元学习器仍采用 MLR。表 4 为两组 Stacking 融合模型的预测结果。可以看到,除了油井 F 外,融合模型 1 在油井 A、油井 B、油井 C、油井 D 和油井 E 上都表现优异(6 #)。主要原因有:①模型融合模型 1 中选择了学习能力更强的基学习器,这些基学器提升了 Stacking 模型的整体预测能力;②所选产油井虽然位于同一区块,但油井在静态、动态等多项数据参数都呈现出不同的数据空间观测特征,需要多样性和差异性较大的算法模型来提升预测的稳健性和精确性。

4　结论与展望

利用 Stacking 集成方式将 Bagging、XGBoost、LSTM、TCN 等多个算法进行融合后对国内某特高含水水驱开发油田生产历史数据进行动态产油量预测,得到以下结论:

1) 采用 XGBoost、Boosting、LSTM 和 TCN 4 种所选的 4 个基模型能实现油井产量的有效预测,不同模型在不同油井上的预测效果存在差异。

2)本文提出的 Staking 融合模型充分考虑各个模型的优势,对 6 口油井的日产量预测呈现较稳定的预测效果,但不同油井的产量预测曲线形态则有着明显差异。

3)相比单一模型,Stacking 融合模型能够结合多种模型的优势,预测的平均误差较小,预测鲁棒性较好,但当预测准确度较差时,模型融合可能会增大误差。

上述研究成果可以丰富油田产量预测方法,支撑油田开发调整工作。今后的工作中将针对以下问题开展深入探讨。

表 4　6 口产油井不同基模型组合方式的 Stacking 算法产量预测误差分析　　单位:t

序号	Well A		Well B		Well C		Well D		Well E		Well F	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
5#	1.302	4.626	0.793	0.766	1.381	3.858	0.406	0.223	0.411	0.308	0.581	0.577
6#	1.868	5.776	1.017	1.482	2.379	9.288	0.491	0.270	0.510	0.397	0.270	0.137

1)提升 Stacking 融合模型的运算时间。Stacking 集成学习的框架设计比较复杂,对于基模型要训练多次,即使减少了模型训练时的若干数据量,仍存在计算时间较长等问题。针对这一问题,未来可考虑对不同基模型分别建模,从而有效减少运算时间。

2)提升 Stacking 融合模型的泛化性能。油田区块中产油井的产量受多种因素影响。本文的研究仅是考虑了时间序列的动态产量预测,后续研究可进一步对油藏开发相关资料进行深入分析,筛选出更多的静态和动态特征纳入模型训练,从而提升模型的泛化能力。

参考文献

[1] 钟仪华.特高含水油田开发规划动态预测方法研究[D].成都:西南石油大学,2008.

[2] 陈军,杨峻懿,刘启国.特高含水期油藏采油速度研究及影响因素分析[J].油气藏评价与开发,2018,8(4):17-21,32.

[3] CUI M L,WANG R,LV C Y,et al. Research on microscopic oil displacement mechanism of CO_2 EOR in extra-high water cut reservoirs[J]. Journal of petroleum science and engineering,2017,154:315-321.

[4] 王增林.胜利油田分层注水工艺技术研究与实践[J].油气地质与采收率,2018,25(6):1-6.

[5] 吴忠维,崔传智,王振,等.特高含水期产量递减规律及递减率的计算公式[J].中国科技论文,2018,13(15):1778-1782.

[6] 黄广庆.特高含水期产量递减分析及递减率表征公式[J].科学技术与工程,2019,19(15):99-104.

[7] 钟仪华,张志银,朱海双.特高含水期油田产量预测新方法[J].断块油气田,2011,18(5):641-644.

[8] 张凯,赵兴刚,张黎明,等.智能油田开发中的大数据及智能优化理论和方法研究现状及展望[J].中国石油大学学报(自然科学版),2020,44(4):28-38.

[9] 林伯韬,郭建成.人工智能在石油工业中的应用现状探讨[J].石油科学通报,2019,4(4):403-413.

[10] 杨勇.胜利油田勘探开发大数据及人工智能技术应用进展[J].油气地质与采收率,2022,29(1):1-10.

[11] 闵超,代博仁,张馨慧,等.机器学习在油气行业中的应用进展综述[J].西南石油大学学报(自然科学版),2020,42(6):1-15.

[12] 钟仪华,王淑宁,罗兰,等.用深度学习挖掘油田开发指标预测模型的知识[J].西南石油大学学报(自然科学版),2020,42(6):63-74.

[13] 王洪亮,穆龙新,时付更,等.基于循环神经网络的油田特高含水期产量预测方法[J].石油勘探与开发,2020,47(5):1009-1015.

[14] 王一淋,赵磊,梅海燕,等.GM(1,1)与支持向量机组合模型在特高含水期油田产量预测中的应用[J].石油天然气学报,2009,31(1):271-273.

[15] ALFARIZI M G,STANKO M,BIKMUKHAMETOV T. Well control optimization in waterflooding using genetic algorithm coupled with Artificial Neural Networks [J]. Upstream oil and gas technology,2022,9:100071.

[16] KALAM S,YOUSUF U,ABU-KHAMSIN S A,et al. An ANN model to predict oil recovery from a 5-spot waterflood of a heterogeneous reservoir[J]. Journal of petroleum science and engineering,2022,210:110012.

[17] SONG X,LIU Y,XUE L,et al. Time-series well performance prediction based on long short-term memory (LSTM) neural network model[J]. Journal of petroleum science and engineering,2020,186:106682.

[18] RASTOGI A,AGARWAL K,LOLON E,et al. Demystifying data-driven neural networks for multivariate production analysis[C]//SPE/AAPG/SEG Unconventional Resources Technology Conference 2020,URTeC 2020,2020:URTEC-2019-247-MS.

[19] SAGHEER A,KOTB M. Time series forecasting of petroleum production using deep LSTM recurrent networks [J]. Neurocomputing,2019,323:203-213.

[20] 史佳琪,张建华.基于多模型融合 Stacking 集成学习方式的负荷预测方法[J].中国电机工程学报,2019,39(14):4032-4042.

[21] HE Y,XIAO J,AN X,et al. Short-term power load probability density forecasting based on GLRQ-Stacking ensemble learning method[J]. International journal of electrical power & energy systems,2022,142:108243.

[22] OBASI T,SHAFIQ M O. CARD-B:a stacked ensemble learning technique for classification of encrypted network traffic[J]. Computer communications,2022,190:110-125.

[23] QU N,LI Z,LI X,et al. Multi-parameter fire detection method based on feature depth extraction and stacking ensemble learning model[J]. Fire safety journal,2022,128:103541.

[24] KARTHIK R,MENAKA R,KATHIRESAN G S,et al. Gaussian Dropout based stacked ensemble CNN for classification of breast tumor in ultrasound images[J]. IRBM,2022,43(6):715-733.

[25] PIRIZADEH M,ALEMOHAMMAD N,MANTHOURI

M,et al. A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods[J]. Journal of petroleum science and engineering,2021,198:108214.

[26] WOLPERT D H. Stacked generalization[M]. Boston: Springer,2017:6-10.

[27] 史佳琪. 区域综合能源系统供需预测及优化运行技术研究[D]. 北京:华北电力大学,2019.

[28] 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. San Francisco,California,USA:Association for Computing Machinery,2016:785-794.

[29] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[30] 黄颖,杨会杰. 基于 XGBoost 和 LSTM 模型的金融时间序列预测[J]. 科技和产业,2021,21(8):158-162.

[31] LIU K,ZHANG J. A dual-layer attention-based LSTM network for fed-batch fermentation process modelling-science direct[J]. Computer aided chemical engineering,2021,50:541-547.

[32] BAI S,KOLTER J Z,KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[EB/OL]. arXiv:1803.01271v2[cs. LG]. https://arxiv.org/abs/1803.01271(19 April,2018).

[33] KROGH A,VEDELSBY J. Neural network ensembles, cross validation,and active learning[C]// International Conference on Neural Information Processing Systems. MIT Press,1995:231-238.

[34] TING K M. Issues in stacked generalization[J]. Journal of artificial intelligence research,1999,10:271-289.

Production Prediction of Oilfield Based on Multi-model by Stacking Ensemble Learning

ZHANG Tingting[1], PAN Meiqi[2], ZHU Tianyi[3], CAO Yu[2], ZHANG Zhanquan[2], LIU Shanke[2], HE Xing[3], YU Lijun[2]

(1. Research Institute of Carbon Neutrality,Shanghai Jiao Tong University,Shanghai 200230,China;
2. College of Smart Energy,Shanghai Jiao Tong University,Shanghai 200240,China;
3. School of Electronic Information and Electrical Engineering,Shanghai Jiao Tong University,Shanghai 200240,China)

Abstract: An oil production prediction method based on multi-model combination under Stacking ensemble learning was proposed associated with the frontier theory of machine learning. The model was used to predict the dynamic oil production from the production data of a domestic ultra-high permeability oilfield in China developed by water flooding. Considering the differences in training principles of different algorithms, the XGBoost algorithm, long and short memory network (LSTM), temporal convolutional network (TCN) and other models are selected as base learners, the MLR algorithm is chosen as meta learner. The results show that the Stacking ensemble model has smaller average error and better prediction robustness compared with the traditional single model, since the ensemble model fully combined the advantages of each base learner. The proposed model is of great significance to the application in ultra-high water cut reservoirs.

Keywords: multi-model combination; Stacking ensemble learning; XGBoost; long-short term memory; temporal convolutional network; production prediction