引用格式：黄家宸，张金川．机器学习预测油气产量现状[J]．油气藏评价与开发，2021，11(4)：613-620.
HUANG Jiachen, ZHANG Jinchuan. Overview of oil and gas production forecasting by machine learning[J]. Petroleum Reservoir Evaluation and Development, 2021, 11(4): 613-620.
DOI: 10.13809/j.cnki.cn32-1825/te.2021.04.018

机器学习预测油气产量现状

黄家宸[1]，张金川[2,3,4]
（1.中国石化石油勘探开发研究院，北京 100083；2.中国地质大学（北京）能源学院，北京 100083；3.自然资源部页岩气资源战略评价重点实验室，北京 100083；4.非常规天然气能源地质评价与开发工程北京市重点实验室，北京 100083）

摘要：机器学习是一种通用的数据驱动分析方法，也是一个重要的油气大数据分析利用手段。油气勘探开发作为具有悠久历史和庞大数据基础的重要领域，具有很大的数据挖掘潜力。利用油气田大数据分析技术可以帮助决策者进行投资分析、风险评估、生产优化，带来巨大的经济效益。机器学习方法早已被研究人员尝试应用于油气领域相关研究，随着机器学习算法的发展，许多应用场景被不断提出，但针对具体场景的通用方案仍在探索中。笔者从最基本原理着手介绍了机器学习的建模过程，梳理了用于油气田大数据分析的3类主要机器学习方法的发展历史，结合油气田大数据的特点，讨论了油气田大数据分析利用的核心内容、目标及优势，分析了机器学习在油气领域的主要应用场景，总结了目前典型油气产量预测中存在的问题及对策。

关键词：油气田大数据；数据驱动模型；产量预测；机器学习；智能油田
中图分类号：TE328 **文献标识码**：A

Overview of oil and gas production forecasting by machine learning

HUANG Jiachen[1], ZHANG Jinchuan[2,3,4]
(1. Sinopec Petroleum Exploration and Production Research Institute, Beijing 100083, China; 2. China University of Geosciences (Beijing), Beijing 100083, China; 3. Key Laboratory of Shale Gas Exploration and Evaluation, Ministry of Land and Resources, Beijing 100083, China; 4. Beijing Key Laboratory of Unconventional Natural Gas Geology Evaluation and Development Engineering, Beijing 100083, China)

Abstract: The machine learning is not only an important tool for oil and gas big data analysis, but also a general data-driven analysis method. As an important field with a long history and a large data base, oil and gas exploration and development has a great potential for data mining. The use of big data analysis technology for oil and gas field can help decision makers to conduct investment analysis, risk assessment and production optimization, which brings significant economic benefits. The machine learning method has been tried by the researchers applying to the researches on oil and gas. Nowadays, many application scenarios have been proposed with the development of machine learning algorithms, but general solutions for specific scenario are still divided. So that, we introduces the procedure of a machine learning modeling upon the most basic principles, and summarizes the development history of the main three kinds of machine learning methods that can be applied to oil and gas big data analysis. And then based on the characteristics of oil and gas field big data, the core contents, goals and advantages of oil and gas field big data analysis and utilization are discussed, the main application scenarios of machine learning in oil and gas field are analyzed, and the existing problems and countermeasures in typical oil and gas production prediction are summarized.

Keywords: big data of oil and gas field, data-drive model, production forecast, machine learning, intelligent oilfield

---

收稿日期：2021-05-18。
第一作者简介：黄家宸（1993—），男，博士，助理研究员，从事地球物理及油气田大数据研究工作。地址：北京市海淀区学院路31号中国石化石油勘探开发研究院，邮政编码：100083。E-mail：huangjiachen.hjc@163.com
基金项目：国家自然科学基金项目“页岩含气性关键参数测试及智能评价系统”（41927801）。

机器学习是一种泛化能力较强的数据驱动预测方法。李航[1]、周志华[2]在专著中全面、系统、详细地介绍了各类常见机器学习算法的原理,GOODFELLOW等[3]在专著中详细地介绍了机器学习的数学基础、使用经验以及现阶段的理论及发展。机器学习任务目标可以分为多类,包括分类和回归(有监督学习)、聚类(无监督学习)、时序分析、概率图模型、强化学习等,已在一些典型场景中大量应用。

机器学习在非互联网领域的应用通常被称为"AI+",它可以代替研究人员完成重复的、经验性的工作,也可以用来提取人工难以处理的复杂信息,从而对数据进行更深入地挖掘。由于机器学习技术的应用产生了巨大的效益,目前,科研工作者正致力于改良机器学习算法以适应实际应用场景。

要使用机器学习解决实际问题,首先要对问题进行描述。机器学习问题可以描述为:通过已知数据来预测未知数据的属性,其中每个样本可以包含多个属性。在有监督学习的分类问题中,监督学习的训练样本包含对应的"标签",样本标签属于两类或者多类,是离散型变量;在回归问题中,样本标签包括一个或者多个连续变量。在无监督学习中,训练样本的属性不包含标签,如聚类问题。

笔者介绍了目前常用于油气田大数据分析的机器学习方法的原理、分类及历史演变,充分调研后,总结了油气大数据分析利用的特点与方法,举例说明了具备一定的数据条件且适用机器学习的应用场景。以动态及静态产量预测为例,将目前存在的问题归纳为:①研究的普适性及一般性不足;②研究应用场景与价值不明确;③模型拓展应用不足。针对以上问题提出了对策,旨在把握相关问题未来的研究重点。

1 机器学习原理及分类

机器学习的流程如图1所示。首先,选择具有相同属性的、标签分布较为均衡的若干样本,并随机划分一部分为训练集、一部分为测试集。再进行数据的预处理,如归一化、标准化,然后对部分属性进行特征提取,将计算机不能直接识别的属性(称作"非结构化特征")转换为机器学习模型可以利用的属性(称作"结构化特征"),这一操作叫做"特征工程"。接下来建模的过程就是通过最优化过程寻找一个描述属性与标签之间关系的函数。有参模型中逼近函数的过程叫做优化过程,典型的方法有梯度下降法、牛顿法等;评价训练过程中函数是否接近最优化的指标叫做评估指标,如方差、交叉熵。随着模型训练迭代次数的增加,模型在训练集上的误差一定是越来越小的,但模型在测试集上的误差则可能会增大,这种情况叫做"过拟合"。过拟合的模型不能正确描述特征与标签之间关系,不具备实际价值。因此,模型完成训练后还需评价模型在测试集上的预测效果来评估模型是否可靠。若使用了多种不同的模型进行预测,最后还可以分析不同模型预测效果产生差异的原因,结合对数据及算法的认识,对模型中人为的设定(如:k-近邻算法的k值、神经网络的层数、循环神经网络的拓扑结构等)进行调整,这些设定被称作"超参数"。

按照样本是否有标签,机器学习可以分为有监督学习和无监督学习。如果样本的属性包含时间序列,也可以认为是时间序列分析问题,它既可以是有监督学习也可以是无监督学习,最常用的分析及预测方法为循环神经网络(RNN)。本研究主要讨论目前常用于油气田大数据分析利用的机器学习方法,分述如下:

1)有监督学习

有监督学习是机器学习中最重要的分支,它在已知样本标签情况下对学习器(机器学习中的基本模型)进行训练。

有监督学习最早可以追溯到线性判别分析(LDA)[3]。这是一种有监督的数据降维算法,可以解决二分类问题。之后,随着贝叶斯决策理论的发展,贝叶斯分类器——一种条件概率计算方法,开始被应用于文本分析等场景。逻辑回归同样具有悠久的历史,最初被应用于二元序列的分析[4]。逻辑回归使用激活函数将样本从低维空间映射到高维空间的思

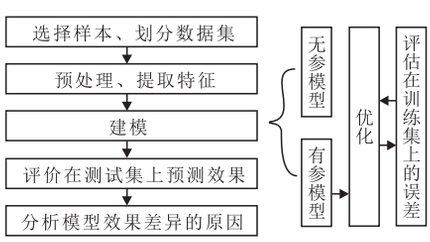

图1 机器学习流程

Fig. 1 Workflow of machine learning

想是神经网络的基础原理。此外，有的机器学习模型不需要进行模型训练，称作无参模型，如基于模板匹配思想的算法k-近邻（kNN）。

机器学习模型由于难以分析内在机理，通常被称作“黑箱模型”，但也有一些概率图方法相比其它模型更容易进行可视化分析，如基于条件判断和信息论的方法决策树（DT）。树形模型原理简单但可解释性强，在机理分析上被广泛使用。还有一些改进的概率图模型，如使用了自助采样法（Bootstrap）重新获取样本集的随机森林[5]，比简单的决策树能够获得更好的预测效果。再后来，诞生了一些具有很高效率和准确度的非概率图模型，既可以解决分类问题和也可以解决回归问题，如经典算法支持向量机（SVM）[6]。这是一种通过核函数将输入向量映射到高维空间，使得非线性的回归和分类问题变为线性的思维。

目前最受欢迎的机器学习算法要属深度学习，它宽泛地涵盖了各类具有多层神经元的机器学习模型。人工神经网络（ANN）是一种近年兴起的深度学习模型，它最早被称作感知机模型[7]。感知机模型被定义为一种组织自学习的神经系统，最初被划分为机器而非程序，但本质上是一种线性分类器。RUMELHART等[8]定义了第一个多层神经网络，实现了反向传播（BP）算法，奠定了神经网络走向完善的基础。

神经网络经过一些改良可以更好地提取不同尺度下的数据集特征，达到更好的预测效果。LECUN[9]设计出了第一个卷积神经网络，使用卷积等操作模仿人对图像的感知，最初被用于手写数字识别，是当今被广泛使用的深度卷积神经网络的基础。ELMAN[10]提出了循环神经网络，用于处理序列数据。循环神经网络的层数可以动态调整，能够处理可变长度的时间序列。GOODFELLOW[11]提出了生成对抗网络的思想，用于生成看似真实的数据，最初被用于图片的创作、人声和乐器声的重现等。

此外，将一些好而不同的异质模型组合起来还可以组成新的机器学习模型，从而提升预测效果，这些异质模型被称作组件学习器，通过算法组合起来的模型被称作集成模型。例如AdaBoost是一种将组件学习器线性集成起来组成一个新模型的方法，可以整合不同学习器的优点从而实现更精确的预测[12]。

2）无监督学习

相对于有监督学习，无监督学习应用范围较窄，因此，发展较为缓慢，现存算法较少。无监督学习主要可以分为聚类和数据降维两类问题。

聚类算法起源于Ward提出的层次聚类模型[13]。这是一种符合人的直观思维的算法，包含若干不同的实现方式。聚类算法中最著名的是k均值算法，可以将样本分为k类，它也是所有聚类算法中变化种类最多的[14]。此外，还有可以解决带有缺失数据的极大似然估计问题的最大期望算法（EM）[15]，基于密度的聚类算法Mean Shift[16]以及将聚类问题转化成图切割问题的谱聚类算法[17]等。

3）时间序列分析的机器学习方法——循环神经网络

循环神经网络（RNN）通常被用于有监督学习，具有普通神经网络没有的“记忆”功能，专用于实现时间序列的训练和预测。虽然一些常规的机器学习方法可以用来分析预测时间序列，但这些模型往往会忽略数据在历史观测中的自相关性。因此，我们需要使用研究序列的机器学习方法来进行时间序列的预测，将数据在时间上的自相关性纳入考虑。循环神经网络中每一时间步的输出都会影响下一时间步的输出，能够很好地使用数据的历史观测信息。

循环神经网络目前主要应用于自然语言处理中。语言文本中单词具有先后顺序，是时间序列。MIKOLOV等[18]建立了基于循环神经网络的语言模型，实现了循环神经网络的应用。GREGOR等[19]实现了深度循环神经网络的注意力机制，模仿了人在语言阅读中的特点。

循环神经网络对数据的使用有别于其它神经网络，例如为了充分地使用数据样本，模型在训练和预测时使用不同的权重传递方法。Teacher Forcing方法通常用于循环神经网络的训练阶段，该方法不使用上一个状态的输出作为下一个状态的输入，而是直接使用训练数据的真实值对应的上一项作为下一个状态的输入[20]。该方法在训练中会更正模型训练过程中的统计属性，使模型训练变得稳定。与之对应的使用上一时刻预测值预测当前时刻预测值的方法叫做Free Running方法，它不依赖于真实数据，更多地被用于预测阶段标签或值的输出。

循环神经网络所有结构的集合叫做网络的主体结构，也称作网络拓扑结构，对主体结构进行改进可以使循环神经网络更好地应用于不同的场景。例如：在机器翻译中，循环神经网络的输出依赖于前面

所有时间步的输出，同时也可能依赖于未来的输出。因此，双向循环神经网络拓扑结构能够对问题更好地描述[21]。拓扑结构的改进设计有很多，如：多维循环神经网络[22]，网格长短期记忆网络[23]等。这些结构在不同的场景中均有良好的表现。

除了主体结构，循环神经网络的另一个重要特征是单元（Cell），它是每个时间步节点对应的非循环神经网络。通常，没有对单元进行优化的简单循环神经网络（Simple RNN）在较长的时间序列预测问题中存在诸多不足，如梯度在传播中会消失的梯度弥散现象。这种现象导致的问题被称作梯度下降中的“长期依赖”问题[24]。为了解决此问题，HOCHREITER等[25]提出了著名的长短期记忆（Long short-memory, LSTM）网络。LSTM网络是RNN的一个变体，它能够学习长期依赖特征，在网络较深时表现优于Simple RNN。LSTM网络的单元具有特殊的“门”的结构，可分为控制门单元和输出门单元，而控制门单元是由可以用来去除网络传播中不需要的信息的“遗忘门”[26]和保留有效信息的“记忆门”组成的。

循环神经网络建模时，不仅可以在网络结构上进行优化，还可以在梯度传播、优化器和激活函数的选择等方面进行改进。如：使用Relu激活函数代替Sigmoid激活函数来防止梯度弥散[27]；使用梯度裁剪来避免“梯度爆炸”[28]；使用Dropout方法来减少过拟合[29]。根据实际问题，同时采用这些不同的优化方案可以大幅提升循环神经网络的预测效果。

2 油气田大数据分析利用的特点与方法

数据是机器学习的基础，只有在数据基础强大的前提下，才能使用机器学习方法对数据进行挖掘。因此，要将机器学习应用于某行业领域，必须保证该行业领域有充分的数据积累，而油气行业满足这一条件。全世界上百年对油气的勘探开发积累了宝贵的油气田大数据，而使用现今机器学习技术对其分析利用可以进一步提炼数据的价值，甚至可以用纯粹的统计学原理对油气生产进行分析（数据驱动方法），来实现油气生产的降本增效。油气大数据的分析利用具有浓烈的专业色彩，因此，不能完全照搬互联网领域的数据挖掘模型，需要先明确其内容、目标及优缺点，再定制最佳的分析方法。

2.1 油气田大数据的特点及其分析利用的内容与目标

油气田大数据分析利用的意义在于通过数据驱动的方法，预测开发效益、减少人工成本、监控油藏动态、优化生产参数，从而达到指导投资和工程决策、降本增效的目的。油气的大数据分析是一种自上而下的分析方法，先通过数据驱动方法得到分析对象之间的定性或定量关系，再根据数据统计特点分析其内在机理，与传统数值模型分析法过程相反。其优势在于，可以在对研究区专业认知不足或者已有数据质量较差的情况下使用。

传统大数据技术已在互联网行业广泛成功使用，但在油气勘探开发领域，大数据技术还没有非常成功的应用实例，这是因为两者的数据特点和任务目标都不相同。虽然油气田大数据研究方法是建立在互联网大数据技术基础上的专业化应用，但是油气田大数据分析不仅仅是一个统计学问题，它的最终目标不是挖掘数据之间的关系，而是通过分析数据之间的关系进行决策。因此，油气田大数据分析利用不能照搬互联网大数据技术，不能终止于得到表面的因果关系，而是要深刻挖掘现象的内在机理。笔者通过研究总结了油气田大数据与传统互联网大数据的不同以及油气田大数据分析利用的核心内容（表1）。

2.2 数据驱动模型在油气田大数据分析中的优势

在油气田勘探开发领域，传统分析油气生产的过程通常建立在物理解析模型之上，而油气田大数据分析则是基于统计学方法，通过实际勘探开发和生产数据来得到数据驱动模型。二者各有优缺点和不同的适用场景，主要研究内容与目标也不同（表2），在实际应用时应当根据对目标区的了解程度和数据条件选择最适用的方法。

3 机器学习在油气领域的主要应用场景

只要有大数据存在的地方，就有应用机器学习方法的可能性。一个新的机器学习算法被提出后，将很快被应用于具有一定数据量基础的各行各业。机器学习在油气领域的应用也依赖于机器学习算法

表1 油气田大数据的特点及其分析利用的核心内容

Table 1 Characteristics of big data in oil and gas field and key points of data analysis and utilization

数据特点	互联网大数据	油气田大数据	油气田大数据分析利用核心内容
4V特征(数据量Volume、数据类型Variety、数据价值Value、产生速度Velocity)	数据量大;类型繁多;价值密度低;产生快、失效快	数据量大;类型繁多;价值密度高;产生快、失效慢	尽快尽早地挖掘数据之间的关系,如:地质和开发条件与产量的关系
数据产生基础	被动无意识产生,需要数据挖掘才能找到数据的价值	主动设计产生,所有测量都与油气生产有关	根据应用场景和所选分析方法,进一步进行原始数据采集
数据之间的相关性	事先未知,需要进行统计分析,统计意义通常不明确	知道定性关系,但有时难以准确定量地描述	挖掘内在机理与观测数据的本质关系而非统计关系
处理技术	使用通用的大数据处理方法,这些机器学习算法通常是基于特定问题而被发明的	带专业约束关系的处理。需要建立适合油气田大数据分析的改进过的机器学习新模型	结合专业知识选择合适的数据驱动模型,使用数学模型对数据驱动模型进行约束和优化
数据结构化特征	可用数据较为完整;对于图像、自然语言等非结构化数据,有较为成熟的预处理算法和相应机器学习模型	历史数据通常不完整,特别是数据关系缺少;一些主导产量的信息难以结构化,如:构造特征、沉积环境等	如何进行数据预处理(包括数据筛选、特征工程等)。数据预处理工作量最大,并且对最终分析效果影响最大
数据预测的准确性	信息通常完备,特征中包含预测目标的全部信息,数据量足够大时弱学习器等价于强学习器,预测准确性高	信息通常不完备,很多重要信息不能被获取或特征化,数据噪声较大,预测效果存在理论上限,预测准确性低	在样本和可选特征有限、数据随机性成分较多的情况下,选择最适合预测场景模型,尽量提高预测准确性和稳定性

表2 数据驱动模型与传统解析模型对比及数据驱动建模工作要点

Table 2 Comparison between data-driven model and traditional analytical model and key points of data-driven modeling

模型	优势	劣势
传统解析模型	遵从实际机理,结果易于解释,更容易用于指导开发设计; 解析模型对计算机算力要求低,数值模拟对已知信息使用充分; 用于新的目标区时不需要先验的认识	模型应用条件苛刻,实际应用中必须先获得模型中所有参数; 无法处理一些非结构化的数据,如:井的层位信息等
数据驱动模型	普遍适用性高,稳定性好,需要后期人工调整的部分较少; 基于模糊逻辑,对专业知识依赖少; 可以使用通用的数据特征处理方法,容易处理非结构化数据	基于统计学原理,不易解释机理; 需要很多先验的训练样本才能得到可靠的模型,并且模型不能推广到其他研究区
数据驱动建模工作要点	充分利用已知信息,结合专业知识进行数据特征工程; 根据已知数据的特点,建立对目标区最适用的、通用的预测模型,最终实现动态生产监测。 样本中对其标签产生影响的、但未作为机器学习或统计学建模特征的若干属性要尽量一致,排除非确定性信息的影响	

的发展,越来越多的应用场景随着算法的提出而产生。总结目前较为成熟的方法,机器学习可被应用于以下油气勘探开发场景:

1)地球物理分析

地球物理学家通常需要大量的时间进行地震和测井解释,在庞大的地球物理资料中评估构造和地层的不确定性。基于机器学习的地球物理分析方法对地震、测井资料进行自动的整理和演算,例如从多维地震数据集中得到不断演化的地震属性结合,可大大减少地球物理学家的工作量。

2)地质分析

地质学家的工作与机器学习非常相似,主要任务是综合现有的资料,结合丰富的地质知识和经验,重新还原油气藏的形成过程。使用机器学习方法可以在短时间内处理大量的信息,一个好的机器学习模型对一些精细地质特征的观察甚至超过地质学家,结合机器学习方法可以提升地质分析的效果。

3)储层评价

储层包含着大量与油气生产相关的信息,使用机器学习方法可以快速进行矿物的显微识别、提取岩石物理特征、确定烃类的体积、分析流体在储层岩石中的状态等。

4)油藏工程分析

机器学习可以为油藏提供快速、准确的动态预

测,及时监测产量以及与产量相关的属性,对油藏进行动态调整,最终实现“智能油田”。由于油藏的部署及开发策略具有复杂性,数据驱动方法有时比传统的油藏模拟方法更高效、准确。

5)生产监测

油井在生产中,各类监测传感器会源源不断地制造生产数据。这些数据可以帮助我们诊断生产故障,建立油井生产快速预警系统。由于不同研究区油井的故障响应可能不同,根据研究区样本建立的机器学习模型可以更加贴合实际地描述油井的生产状态。

6)经济预测

油气田开发的核心目标就是在开发周期内获得最大的经济价值,因此在不同生产阶段对油气产量及成本的预测是必不可少的。可以根据已有的生产资料及多属性预测的结果,来评估下一步拟定的开发方案是否能产生较大的经济效益。

4 基于机器学习的油气产量预测的现状、问题及对策

产量预测是机器学习在油气田大数据分析利用中一个重要的、典型的应用,基于机器学习的产量预测包含了油气数据挖掘的主要元素。在已有的研究中,通常是通过老井的数据来训练机器学习模型,并预测新井的产量。若预测目标(标签)是井在固定一段时间内的总产量或稳产产量,则称为静态产量预测;若预测目标是井的生产曲线,则称为动态产量预测。

4.1 研究现状

1)静态产量预测

井的产量变化往往基于井所处区块的层位特征、地质和工程参数等静态属性的变化,静态产量预测就是对井的静态属性的挖掘。LOLON等[30]使用3个多变量统计模型评估Bakken和Three Forks地层的井参数与产量之间的关系。评估得出,主要影响致密储层产量的工程变量是水力压裂过程中的总压裂液量和支撑剂量。CLAR等[31]使用人工神经网络预测Eagle Ford页岩工区水平井的产量,发现总产量与侧向长度、垂直深度、孔隙度和压裂液量均有显著关系。

然而,使用机器学习进行产量预测的研究还处于起步阶段。目前,页岩、致密油气藏等非常规油气藏的静态产量预测模型并没有取得很好的效果[32],其中一个主要原因是油气产量不仅受量化属性的影响,而且还受储层岩性、构造、钻井压裂等非量化增产措施的影响,而这些属性往往是非结构化的数据,导致机器学习建模时此类信息难以被使用。例如:在地质因素的考虑上,以往的产量预测研究要么忽略了区块特征[30],要么在样本井较少时假设储层具有均质性[33],但实际上即使在很小的范围内,致密地层或页岩地层的地质特征变化也会很大,不使用区块特征信息或非均质性信息会导致预测不准确[34]。虽然很难直接将这些属性进行充分的特征工程,但可以找到替代的方法,一种折中的方法就是使用井位(井距、井位坐标信息)来表示井之间的地质差异[35]。因此,当一些非结构化数据信息难以纳入模型时,我们应当采用一些替代手段或增加适用条件限制,并适当结合物理机理来加强对数据的提取和模式识别,从而减少这些因素的影响,如:根据数据的产生过程设计特殊架构的端到端模型,在不进行特征工程的情况下进行机器学习,减少在特征编码时对油气专业知识的依赖[36]。

2)动态产量预测

油气的生产曲线预测是一个典型的时间序列分析问题,使用动态生产数据预测井的生产曲线就是动态产量预测。生产曲线的预测研究包括对其自相关性、趋势或周期性变化的分析,这些性质可以帮助我们预测未来产量。因此,基于数据驱动的生产曲线数据挖掘分析方法将对油气井产量评估及预测至关重要。

在已有的研究中被用来预测油气生产曲线的时间序列学习器主要为简单循环神经网络(Simple RNN)和长短时记忆网络(LSTM)。SAGHEER等[37]比较了Simple RNN和LSTM两种循环神经网络在油气生产预测中的应用效果,认为LSTM的性能优于Simple RNN。因此,在进行生产曲线预测分析时应当尝试不同的循环神经网络,选择对数据集最适用的模型。

目前油气产量时间序列预测的研究大多是同步时间序列预测,问题描述为:已知属性(Source)和预测目标(Target)在每一步相对应,且已知属性时间序列长度和预测目标时间序列长度相等时计算产油量,例如基于循环神经网络适用油藏的产量曲线预

测同期的月产油、产水、产气量研究[38-39]。同步时间序列预测实际上是对未测量数据的计算，因此，能够应用的场景有限，例如：不能实现使用井的静态参数及已知生产曲线预测未来生产曲线，也不能结合训练好的模型来优化生产设计。一种解决方案是更改循环神经网络的拓扑结构，建立滞后时间序列的循环神经网络模型，从而实现在更重要的场景下的产量预测。滞后时间序列目前主要应用在机器翻译和文本生成问题中，在油气产量预测场景中尚未得到应用。

4.2 存在问题

目前，基于机器学习的水平井产量预测研究在建模及应用层面还存在不足，主要可以归纳为以下3个方面。

1）机器学习在产量预测中的一般性研究较少，多为针对某一数据集的模型优化和比较研究，应用价值有限。每个油田的产量主控因素可能有较大差别，且很多参数是无法量化的，无法加入到机器学习中，带来了较大的不确定性。因此，通过某一油田数据集训练出的机器学习模型往往不能在其他油田使用。并且，过量的调参往往会导致严重的过拟合。因此，使用机器学习方法进行产量研究的重点不在于调参，而在于总结具有普适性的产量预测过程。

2）对产量预测问题在机器学习模型中的描述欠缺，实际应用场景不明确。目前的动态产量预测建模通常是根据当前生产的产液量、井底流压等实时参数进行产量的推算，不能实现对未来长期的预测。换言之，没有一个能够使用的静态参数或已知生产曲线能够预测未来生产曲线问题的时间序列分析模型。

3）基于机器学习产量预测的建模方法研究较多，但拓展应用研究较少，难以深刻挖掘机器学习模型在产量影响因素分析及生产优化设计中的价值。产量预测模型不仅可以用于预测，还可以用于产能主控因素分析、生产决策和优化等，但目前相关研究不够深入。

4.3 解决对策

针对以上问题，应当着重以下3个方面的研究。

1）致力于总结基于机器学习的产量预测建模的一般性过程，并结合油气领域知识寻找提取与产量有关的非结构化属性的模式，得到可泛化的特征工程方法。使用多种机器学习模型在不同特征选择下进行产量预测，对比不同数据条件下，选用不同模型时在测试集上的预测结果，并结合油气田大数据的特点，分析方法的适用性及产生误差的原因。

2）研究产量的动态预测问题在神经网络预测模型中有哪些应用场景，从而正确地将产量预测问题描述为机器学习问题。例如：在产量动态预测时，使用井的静态参数或已知生产曲线预测未来生产曲线，实际上是一个滞后时间序列预测问题，已知量和预测目标在时间上并非互相对应，应当参照机器翻译模型来确定循环神经网络的拓扑结构。建立循环神经网络模型前需要明确已知属性和预测目标，不能混淆其与训练集、测试集的概念。

3）建模后充分挖掘模型价值，根据现有的产量预测结果分析产能主控因素，进一步认识了解研究区特点；找到可优化的工程参数，结合网格搜索、贝叶斯优化等优化方法，实现在新增生产井前对工程参数进行合理的调整，达到降本增效的目的。利用机器学习模型能够剔除数据噪声的特性，排除井在生产过程中的随机性产量波动，描述确定的产量波动，从而用于分析老井生产措施对产量的短期影响，实现生产措施的定量优化。

参 考 文 献

[1] 李航.统计学习方法[M].北京：清华大学出版社，2012.
LI Hang. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012.

[2] 周志华.机器学习[M].北京：清华大学出版社，2016.
ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[3] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning [M]. Boston: MIT Press, 2016.

[4] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(2): 179-188.

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

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

[7] ROSENBLATT F. The perceptron: A probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6): 386.

[8] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[9] CUN Y L. Generalization and network design strategies[J]. Connectionism in Perspective, 1989, 19: 143-155.

[10] ELMAN J L. Distributed representations, simple recurrent networks, and grammatical structure[J]. Machine Learning,

1991, 7(2): 195-225.

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[12] FREUND Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121(2): 256-285.

[13] WARD J H. Hierarchical grouping to optimize an objective function[J]. Journal of the American Statistical Association, 1963, 58: 236-244.

[14] MACQUEEN J. Some methods for classification and analysis of multivariate observations[J]. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, 1(14): 281-297.

[15] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39 (1): 1-22.

[16] CHENG Y. Mean shift, mode seeking, and clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790-799.

[17] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[18] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]// Eleventh annual conference of the international speech communication association, Chiba, 2010.

[19] GREGOR K, DANIHELKA I, GRAVES A, et al. Draw: A recurrent neural network for image generation[C]// International Conference on Machine Learning, PMLR, Lille, 2015.

[20] DOYA K. Bifurcations of recurrent neural networks in gradient descent learning[J]. IEEE Transactions on Neural Networks, 1993, 1(75): 218.

[21] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[22] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Multi-dimensional recurrent neural networks[C]// International conference on artificial neural networks. Springer, Berlin, 2007.

[23] KALCHBRENNER N, DANIHELKA I, GRAVES A. Grid long short-term memory[C]// International Conference on Learning Representations, London, 2016.

[24] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 2(5): 157-166.

[25] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[26] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: Continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.

[27] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// ICML, Haifa, 2010.

[28] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks[C]// International conference on machine learning, PMLR, Atlanta, 2013.

[29] SRIVASTAVA N. Improving neural networks with dropout[D]. Toronto: University of Toronto, 2013, 182: 566.

[30] LOLON E, HAMIDIEH K, WEIJERS L, et al. Evaluating the relationship between well parameters and production using multivariate statistical models: A middle bakken and three forks case history[C]// SPE hydraulic fracturing technology conference, Woodlands, 2016.

[31] CLAR F H, MONACO A. Data-driven approach to optimize stimulation design in Eagle Ford formation[C]//Unconventional Resources Technology Conference, Denver, 2019.

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

[33] ZHOU Q, KLEIT A, WANG J, et al. Evaluating gas production performances in marcellus using data mining technologies[C]// Unconventional Resources Technology Conference. Denver, 2014.

[34] CLARKSON C R, JENSEN J L, PEDERSEN P K, et al. Innovative methods for flow-unit and pore-structure analyses in a tight siltstone and shale gas reservoir[J]. AAPG bulletin, 2012, 96(2): 355-374.

[35] SCHUETTER J, MISHRA S, ZHONG M, et al. Data analytics for production optimization in unconventional reservoirs[C]// Unconventional Resources Technology Conference, San Antonio, 2015.

[36] 陈云天.基于机器学习的测井曲线补全与生成研究[D].北京:北京大学,2020.
CHEN Yuntian. Research on well log completion and generation based on machine learning[D]. Beijing: Peking University, 2020.

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

[38] 周于皓,刘慧卿,祁鹏,等.基于循环神经网络的缝洞型油藏油井产量预测[J].计算物理,2018,35(6):668-674.
ZHOU Yuhao, LIU Huiqing, QI Peng, et al. Forecast of oil production in fractured-vuggy reservoir by using recurrent neural networks[J]. Chinese Journal of Computational Physics, 2018, 35(6): 668-674.

[39] 谷建伟,周梅,李志涛,等.基于数据挖掘的长短期记忆网络模型油井产量预测方法[J].特种油气藏,2019,26(2):77-81.
GU Jianwei, ZHOU Mei, LI Zhitao, et al. Oil well production forecast with long - short term memory network model based on data mining[J]. Special Oil and Gas Reservoirs, 2019, 26(2): 77-81.

(编辑 李颖洁)