 DOI: https://doi.org/10.12345/smg.v6i4.19374

Screening of Main Controlling Factors for Gas Production in Water-Bearing Gas Reservoirs Based on Machine Learning

Zhaolong Zhu Xinyu Feng Wei Wang Lu Wang

College of Petroleum and Natural Gas Engineering, Chongqing University of Science and Technology, Chongqing, 400000, China

Abstract

In recent years, machine learning has been widely applied in the petroleum field, and machine learning algorithms are now used to screen the main control factors that affect gas production in gas wells. Taking the CS1-1 well group in a certain gas reservoir as an example, three machine learning algorithms, namely decision tree, random forest, and gradient boosting regression tree, were used to select seven influencing factors such as oil pressure, casing pressure, and daily water production as feature variables, and daily gas production as the target variable for main control factor screening. The results showed that random forest performed the best in accuracy and reliability, and the selected main control factors were tubing pressure, co fluid flow rate, and daily water production. Screening the main control factors is of great significance for improving the accuracy of gas production prediction.

Keywords

gas reservoir with water; gas production; controlling factors; machine learning

基于机器学习的有水气藏产气量主控因素筛选

朱兆龙 冯心雨 王伟 王璐

重庆科技大学石油与天然气工程学院，中国 · 重庆 400000

摘　要

近年来，机器学习在石油领域中得到了广泛应用，现采用机器学习算法筛选影响气井产气量的主控因素。以某气藏CS1-1井组为例，使用决策树、随机森林和梯度提升回归树这三种机器学习算法，选取了油压、套压、日产水量等7种影响因素作为特征变量，以日产气量作为目标变量进行主控因素筛选。结果表明，随机森林在准确性和可靠性方面表现最佳，筛选出的主控因素为油管压力、协液流速和日产水量。筛选主控因素对提高产气量预测精度具有重要意义。

关键词

有水气藏；产气量；主控因素；机器学习

1 引言

有水气藏是指随着开发过程的进行，除了会产出天然气，还会产出一定量的水的气藏，其产量和开发效果受到多种复杂因素的影响。准确筛选影响有水气藏气井产量的主控因素对于提高气井产能、优化开发策略以及延长气井的经济寿命至关重要。传统的主控因素筛选方法往往面临难以适应数据量大和变量多的情况，以及结果可靠性不强等局限性。近年来，机器学习方法因其优秀的数据处理能力和模式识别能力，在石油领域得到了广泛的应用。相对于传统方法，机器学习可以有效识别数据中的复杂模式和非线性关系，通过适当选择和调整机器学习算法，构建适应于有水气藏特性的预测模型，从而提高主控因素筛选的准确性和效率。

本研究选用三种机器学习算法分析同一个井组经预处理的数据，通过建立算法模型，对比了各算法在抽取主控因素方面的精度与可靠性，筛选出对产气量影响较大的主控因素。

【作者简介】朱兆龙（1999–），男，中国河南许昌人，在读硕士，从事油气田开发研究。

2 数据提取与处理

2.1 数据获取

为了验证有水气藏气井生产指标与产气量之间相关性，本研究从 YC 气藏中获取到了 CS1-1 井组的实际生产数据，共计 1246 条数据。选取特征字段有油管压力、套管压力、日产水量、天然气偏差因子、天然气密度、协液流速、协液流量。

2.2 数据预处理

数据预处理是构建气井产量预测模型中极为关键的一

个步骤，其工作量在整个工程中占比超过一半，对原始数据进行预处理，能够达到清洗数据、提高数据质量的目的，为有水气藏产气量主控因素筛选奠定基础。

处理缺失数据是数据预处理的关键步骤之一。对于有水气藏气井，当数据缺失是由于气井作业关井等原因导致的，可以直接删除这些缺失数据，当数据是由于操作原因引起时，可以通过计算缺失值前后 5 天数据的平均值来填补，这样可以在一定程度上保持数据的完整性，处理后的数据如表 1 所示。

表 1 YC 气藏 CS1–1 井组部分数据

套压	油压	日产水量	天然气密度	天然气偏差因子	临界协液流速	临界流量
9.90	11.40	1.20	164.50	0.5869	1.16	5.0508
9.90	11.40	1.00	164.50	0.5869	1.16	5.0508
9.90	11.40	1.20	164.50	0.5869	1.16	5.0508
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.20	5.80	4.90	23.60	0.9090	3.18	1.9833
2.20	5.80	4.50	23.60	0.9090	3.18	1.9833
2.20	5.80	3.50	23.60	0.9090	3.18	1.9833

3 主控因素筛选

3.1 利用决策树筛选主控因素

决策树是一种常用的监督学习算法，用于构建分类器和预测模型。决策树的基本原理是通过一系列的是或否决策进行学习判断，从根节点到叶节点，最终得出结论的过程。

将训练集和测试集按照 8 ： 2 的比例进行划分，得到 997 条训练数据，249 条测试数据，特征为：临界协液流量、临界协液流速、天然气密度、天然气偏差因子、日产水量、套管压力、油管压力，预测目标为日产气量。通过 R^2 评估方法对决策树算法模型进行评估，Train_set 得到的值为 0.9622，Test_set 的值为 0.9412。利用 trait_proportion 方法得出特征重要性占比，将其可视化，如图 1 所示。

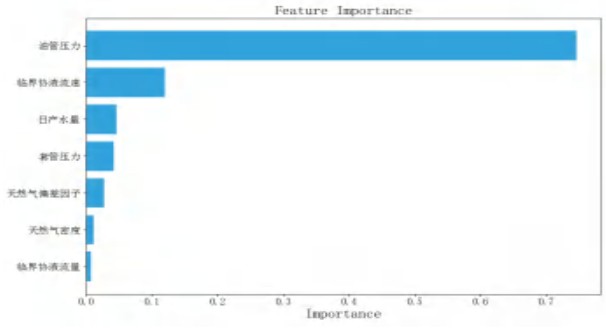

图 1 决策树算法筛选主控因素

由图 1 可以看出，决策树算法对影响有水气藏气产气量的众多因素进行筛选之后，油管压力、协液流速和日产水量这三项特征对气井产量影响程度最显著。

3.2 利用随机森林筛选主控因素

基于决策树算法筛选结果，能够明显地看出其存在一大缺陷就是有时会对训练数据产生过拟合。为了防止筛选结果过拟合，因此引入随机森林算法，其是解决筛选结果过拟合的方式之一。随机森林的基本原理是一种集成学习方法，它通过构建多个决策树并将它们的结果进行投票或平均，以得到最终的预测。这种方法的主要优点是，通过集成多个模型，它可以有效地处理拟合问题，提高模型的泛化能力。另外，随机森林算法特有的一个优势是能够轻松评估各个特征对模型预测结果的贡献程度。

机森林算法预测模型创建训练之后，通过 R^2 评估方法对随机森林算法模型进行评估，Train_set 的值为 0.9862，Test_set 的值为 0.9834。同样，利用 trait_proportion 方法得出各特征重要性占比，可视化效果如图 2 所示。

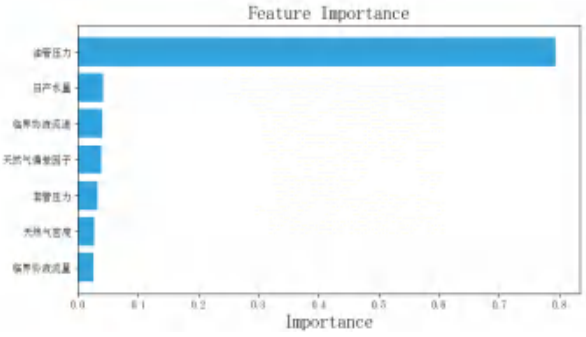

图 2 随机森林算法筛选主控因素

在图 2 中可以看出油管压力、日产水量和协液流速是随机森林算法分析得出的重要特征，同样，油压的占比超过其他 6 个指标之和。

3.3 利用梯度提升回归树筛选主控因素

梯度提升回归树是的原理基于集成学习中的 Boosting 算法。其原理是逐步构建一个加法模型，通过迭代方式来组合多个弱学习器，逐渐减小模型在训练数据上的损失函数。梯度提升回归树与随机森林的方法不同，它是一种序列化方法，每棵树都依赖于前一棵树的结果。这意味着梯度提升树的每一轮迭代都在尝试纠正前一轮的错误，因此每棵树都更加关注数据中的错误部分。

梯度提升回归树模型创建训练之后，使用 R^2 评估方法

对其进行评估，得到 Train_set 的值为 0.9685，Test_set 的值为 0.9582。使用 trait_proportion 方法得出各特征重要性占比，将其可视化如图 3 所示。

在图 3 中可以看出油管压力、套管压力、天然气密度和日产水量是梯度提升树算法给出的重要特征，其中油压对产气量的影响是最显著的。

3.4 筛选结果对比

通过对决策树、随机森林和梯度提升回归树这三种算法的结果进行对比，我们可以清晰地了解到各特征在以日产气量为预测目标时的重要性等级。每种算法根据特征重要性评估得出的得分在图 4 中有详细展示，这些结果为我们提供了有关不同算法性能和特征重要性的深入见解。

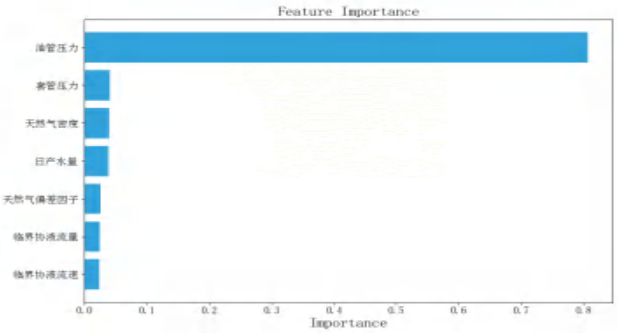

图 3 梯度提升回归树算法筛选主控因素

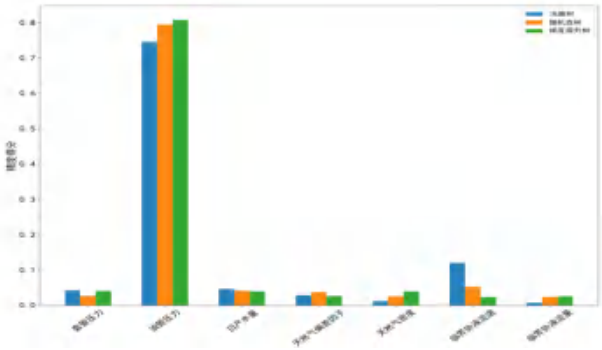

图 4 三种算法中各特征的重要性得分

4 结论

通过对比三种算法在筛选有水气藏产气量主控因素时的表现来看，随机森林算法在预测的准确度上的要领先于其他两种方法，表明了它在处理类似问题上的适用性和可靠性。

根据三种算法的筛选结果可以看出，油管压力、协液流速和日产水量的特征重要性占比相对较大，符合现场生产逻辑。因此，选取油管压力、协液流速和日产水量作为影响有水气藏气井产气量的主控因素。

参考文献

[1] 王建琴.基于改进决策树的数据挖掘与分析算法设计[J].电子设计工程,2024,32(4):84-88.

[2] Loreti D, Visani G. Parallel approaches for a decision tree-based explainability algorithm[J].Future Generation Computer Systems,2024,158:308-322.

[3] 宋佳,唐善杰.基于机器学习的火驱产油量主控因素筛选[J].精细石油化工进展,2023,24(2):44-48.

[4] 李明钊,李熠胥,王佳.基于梯度提升回归树模型的烟草产量预测方法[J].云南化工,2023,50(9):109-111.

[5] 石磊.一种基于随机森林算法的探明储量预测新方法[J].中国石油勘探,2023,28(3):167-172.

[6] 张棪,曹健.面向大数据分析的决策树算法[J].计算机科学,2016,43(S1):374-379.