

基于机器学习方法的数字岩芯电导率预测

郭晨, 陈龙

(长安大学信息工程学院, 陕西西安 710064)

摘要: 岩芯电导率是地下油气判断及岩石物理分析的重要参数。传统的岩石物理特性分析一般通过对岩芯样本的实验测量分析或对数字岩芯图像数据的有限元数值计算等方法, 来获得岩石样本的电导率等宏观物理特性。但这两类分析方法均需耗费较高的人工、时间成本或大量计算资源。随着人工智能等一代数字化分析方法的发展, 文中拟开发一种采用机器学习对岩石电特性进行预测的新方法。在研究中, 将针对岩芯的三维图像数据使用集成学习(ensemble learning)和人工神经网络(ANN)来预测电导率。其中传统机器学习与多层神经网络的输入特征是几何参数, 而三维卷积神经网络输入特征是三维二元分割图像。在研究中比较各种机器学习方法的优劣。实验表明, 以三维二元分割图像作输入特征的三维卷积神经网络(3DCNN)比采用Minkowski泛函作输入特征的学习模型在预测岩芯电导率上性能更优。

关键词: 机器学习; 岩石物理; 神经网络; 集成学习; 电导率

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2020)07-0100-04

doi:10.3969/j.issn.1673-629X.2020.07.022

Prediction of Digital Core Electrical Conductivity Using Machine Learning Method

GUO Chen, CHEN Long

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: The electrical conductivity of digital core is an important parameter for underground oil and gas evaluation and rock physics analysis. Traditional analysis of rock physical properties usually obtains macroscopic physical properties such as conductivity of rock samples by means of experimental measurement and analysis of core samples or finite element numerical calculation of digital core image data. However, both methods require high labor, time cost or a large amount of computing resources. With the development of artificial intelligence and other digital analysis methods, we intend to develop a new method for predicting rock electrical properties by machine learning. In this study, the integrated learning (ensemble learning) and artificial neural network (ANN) are used to predict the conductivity of three-dimensional image data of core. The input features of traditional machine learning and multi-layer neural network are geometric parameters, while the input features of three-dimensional convolution neural network are three-dimensional binary segmentation images. In the research, the advantages and disadvantages of various machine learning methods are compared. Experiment shows that the three-dimensional convolution neural network (3DCNN) with three-dimensional binary segmentation image as input feature has better performance than the learning model with Minkowski functional as input feature in predicting core conductivity.

Key words: machine learning; digital rock physics; neural network; ensemble learning; electrical conductivity

0 引言

数字岩石物理(digital rock physics, DRP)是近些年兴起的一种计算岩石物理的方法。通过CT扫描技术对岩石样本进行三维成像,建立能模拟岩石孔隙空间分布的数字化岩芯数据,在此基础上进行数值模拟,计算岩石等效物理参数^[1]。数字岩石物理包含数字岩芯模型建立和等效物理场模拟两个大的部分,其数字岩芯微观结构可见,微观影响因素可控,能进行多物理

场模拟,相比较实验岩石物理而言,这种分析方法较为精确且效率高^[2]。数字岩石物理可广泛应用在模拟流动性^[3]、渗透性^[4]、电特性^[2]、声学^[5]及弹性^[6]等领域。其中岩芯的电特性对于确定岩性,划分油水层、确定油水饱和度等物理参数具有重要参考作用。然而在实际应用中,由于CT扫描得到的岩芯三维数据庞大,对其进行精确物理场数值计算时耗费资源巨大。为了提高数字化分析方法的效率,本研究将机器学习方法

收稿日期:2019-08-18

修回日期:2019-12-20

基金项目:国家自然科学基金(41874140)

作者简介:郭晨(1984-),女,博士,副教授,研究方向为电磁地球物理应用;陈龙(1994-),男,硕士研究生,研究方向为机器学习与应用。

创新性地应用到了数字岩石物理之中。机器学习就是通过编码使得计算机具有像人一样的学习能力,进而对新的数据进行预测^[7]。本研究的目标为:通过机器学习模型对标定电导率的岩芯数据进行特征化学习,最终免去复杂耗时的物理场数值计算,直接通过岩芯数据预测其电导率。传统的机器学习算法有 K 近邻算法、决策树、线性回归、支持向量机、朴素贝叶斯分类器和逻辑回归等,而神经网络是一种特殊的机器学习方法,被广泛应用于图像分类、目标检测、语音识别、图像和语音生成等方面^[8]。目前神经网络相较于传统机器学习方法而言,其性能最优,但它也需要更多的数据,可解释性差,且计算成本高,因此两者是各有所长。在本研究中传统的机器学习方法和神经网络都被应用到岩芯电导率预测方面,通过学习岩石的几何特征、图像空间分布与电特性的映射关系等,最终达到优化预测效果、降低计算开销的目的。

1 数据处理

在研究中使用了 400×400×400 体素的贝雷亚砂岩岩芯(Berea Sandstone)。岩芯图像是三维二进制数据,在三维空间中,单个位置的数值为 1 或 0,分别代表岩石固体或空隙。为了获得足够多的实验岩芯样本,实验中对数字岩芯图像进行子样截取。由于考虑到岩芯的几何特性和计算资源,该实验采用 64×64×64 作为截取尺寸。研究使用了一个“移动立方体取样器”的方法对原始图像进行子样截取^[9]。以贝雷亚砂岩岩芯图像上沿着 XYZ 各个正方向每次移动 24 个单位进行子样截取,子样与子样之间可能会有重合部分。最终实验共获得了 3 375 个体素 64×64×64 尺寸的岩芯样本。取样后使用有限元的方法计算出每个岩芯样本的电导率,用电导率作为机器学习的目标。

2 特征提取

2.1 Minkowski 泛函

Minkowski 泛函是几何形态特征的统计量。对于 n 维空间,有 $n+1$ 个的 Minkowski 测量量。3 维图像具有 4 个 Minkowski 测量量,分别是体积(V)、表面积(S)、平均积分曲率(平均宽度)(M)和欧拉特性(X)^[10]。欧拉特性又常称作高斯曲率积分,公式如下:

$$V = V(y)$$

$$S = \int_{\partial y} ds$$

$$M = \int_{\partial y} \left[\frac{1}{2} \left(\frac{1}{r_1(s)} + \frac{1}{r_2(s)} \right) \right] ds$$

$$X = \int_{\partial y} \frac{1}{r_1(s)r_2(s)} ds$$

其中, $r_1(s)$ 表示表面上的最大曲率, $r_2(s)$ 表示表面上的最小曲率。Minkowski 泛函可以综合反映岩石几何参数,如孔隙度、比表面积、颗粒的圆度等^[11]。岩芯几何特性直接影响到岩芯的电特性。因此使用 Minkowski 泛函在预测岩芯电导率上展示出更好的优势。为了更直观地表现岩芯样本的几何特征,将体积转为孔隙率(空隙体积/总体积),表面积转换为比表面积(面积/体积)。实验中岩芯样本的孔隙率、比表面积、平均曲率积分和欧拉特性是由三维二元分割图像中直接计算获得。

2.2 三维二元分割图像

三维二元分割岩芯图像作为卷积神经网络的输入。实验岩芯的体素为 64×64×64,单个空间位置上的数值为 1 或 0。实验中将原为 0 到 1 的数值区间投影到 -1 到 1 的区间。具体操作是将原来图像上的 0 替换成 -1。

3 方法

3.1 机器学习

机器学习(machine learning, ML)是通过算法使得计算机具有学习能力,从历史数据中学习规律,进而预测新的数据。机器学习的步骤主要分为选择数据、搭建模型、验证模型、测试模型、使用模型、优化模型。

3.2 集成学习

集成学习(ensemble learning)是一种学习算法,它构造一组分类器,通过对预测进行(加权)表决来对新的数据进行分类^[12]。常见的集成技术有 Bagging、Boosting 和 Stacking。

3.2.1 加权平均法

加权平均法:

$$f(x) = \sum_{i=1}^k \omega_i x_i$$

其中, $\omega \geq 0$, $\sum_{i=1}^k \omega_i = 1$ 。

加权平均法是一种集成策略。通过模型训练可以得到多个不同的学习器,加权平均法是对每个使用的学习器的结果赋予不同的权值,实现结果的融合。对于验证结果较好的学习器赋高权值,而验证结果较差的学习器赋低权值,在赋值的同时保证总权值为 1^[13]。

3.2.2 堆叠法

Stacking 不使用琐碎的函数(如硬投票,加权平均)执行集成,而是训练一个学习器来执行这个集成的过程。在 Stacking 中个体学习器被称为初级学习器,而用于结合的学习器称为次级学习器。Stacking 先从初始数据集训练出初级学习器,将初级学习器的输出当作特征构成新数据集,利用新数据集再训练次

级学习器^[14]。Stacking 使用的初级学习器可以是同一种算法,也可以是不同的算法。

3.3 多层神经网络(MNN)

多层神经网络包含输入层、隐含层和输出层,结构如图 1 所示。每一层的神经元与下一层神经元都相互连接,同层神经元不连接。对于多层神经网络一般采用误差逆传播来更新权值^[15]。

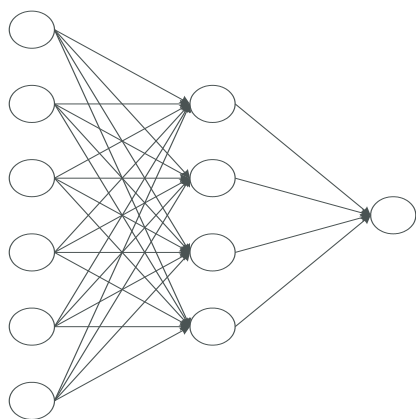


图 1 神经网络结构

3.4 3D 卷积神经网络(3DCNN)

3D 卷积神经网络是一种深度神经网络,主要包括卷积层、子采样层和全连接层。在卷积层中实现卷积操作。使用卷积核进行卷积运算,卷积核的大小是远小于图像矩阵的。具体操作是对于图像矩阵的每一个像素计算它周围像素与卷积核的乘积。为了避免维数灾难,通常在卷积层之后用一个子采样层来降低特征维数。全连接层是一种下一层的每一个神经元都与上一层的每个神经元相连接的网络结构。卷积神经网络通过卷积层、子采样层和全连接层组合的框架降低计算成本,防止过拟合,且提高了图像识别的准确性。

三维卷积如图 2 所示。

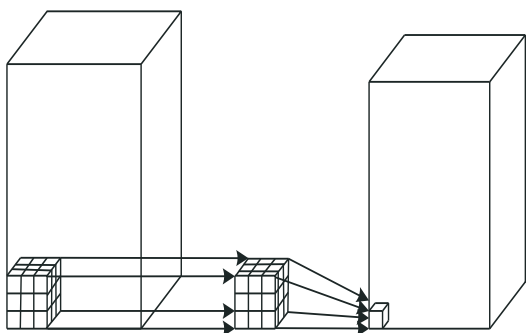


图 2 三维卷积

4 模型搭建与训练

4.1 集成学习基础学习器算法选取

本研究按 4 : 1 的比例将数据划分成训练数据和测试数据。训练数据用于训练模型,测试数据用于最后测试模型性能。在训练中,以归一化后 Minkowski

参数作为输入特征,以岩电导率为输出目标,按不同的机器学习算法训练多个学习器。同时,用五折交叉验证来验证模型性能。该实验使用了线性回归 (LinearRegression)、岭回归 (RidgeRegression)、Lasso 回归 (Lasso)、随机森林回归 (RandomForestRegressor)、梯度提升回归 (GradientBoostingRegressor)、支持向量回归机 (SVR)、线性支持向量回归机 (LinearSVR)、弹性网络 (ElasticNet)、随机梯度下降回归 (SGDRegressor)、贝叶斯岭回归 (BayesianRidge)、核岭回归 (KernelRidge)、极端随机森林回归 (ExtraTreesRegressor)、极限梯度提高回归 (XGBRegressor) 等算法来搭建基础学习器。

4.2 加权平均模型集成

表 1 给出了不同算法搭建出的学习器的性能对比,最终选定了按 LinearRegression、GradientBoostingRegressor、KernelRidge、XGBRegressor 算法搭建的模型为基础学习器。接着对各个模型进行参数调优,以求得更好的性能。最后按 0.2, 0.2, 0.3 和 0.3 的权值赋予各个学习器,进行加权平均,集成一个新的学习器。

表 1 各机器学习方法性能对比

算法学习器	标准误差	标准偏差
LinearRegression	0.006 828	0.000 159
RidgeRegression	0.006 957	0.000 285
Lasso	0.024 999	0.001 159
RandomForestRegressor	0.007 315	0.000 104
GradientBoostingRegressor	0.006 826	0.000 152
SVR	0.036 332	0.001 988
LinearSVR	0.006 898	0.000 177
ElasticNet	0.008 354	0.000 566
SGDRegressor	0.011 737	0.000 724
BayesianRidge	0.006 828	0.000 159
KernelRidge	0.006 744	0.000 214
ExtraTreesRegressor	0.007 246	0.000 093
XGBRegressor	0.006 799	0.000 141

4.3 Stacking 模型集成

在 Stacking 的策略中,仍使用 LinearRegression、GradientBoostingRegressor、KernelRidge、XGBRegressor 算法搭建初级学习器,选择了 LinearRegression 算法搭建次级学习器。训练中用初始训练数据集训练第一层初级学习器,产生新的数据集训练第二层的次级学习器,第一层初级学习器的输出是第二层次级学习器的输入特征,同时原始标签仍被作为新数据集的标签。

4.4 多层神经网络(MNN)和三维卷积神经网络(3DCNN)搭建训练

以 Minkowski 参数为输入特征搭建多层神经网络,以三维二值分割图像为输入特征搭建一个三维卷

积神经网络。两个模型都使用均方根误差 (RMSE) 作为损失函数,Adam 作为优化器来进行模型训练。实验将数据按 3 : 1 : 1 将数据集划分成训练集,验证集和测试集。训练集用于模型权重更新,验证集用于验证模型效果和提早结束训练,防止过度拟合,测试集用于评估最终模型效果。

5 结果与分析

表 2 总结了不同学习模型训练集和测试集的平均

绝对百分比误差 (MAPE)。如预期一致,一般情况下学习模型训练集上的 MAPE 小于测试集上的 MAPE。两种集成方式加权平均模型和 Stacking 模型均在预测效果上比单一机器学习方法有提升。对于多层神经网络而言,其预测效果与传统的机器学习方法效果并无大异。使用三维图像的 3DCNN 比使用几何参数的学习模型预测效果要好,这表明卷积操作比几何参数更好地捕获到岩芯的特征。

表 2 各模型测试性能对比

模型算法	特征	特征项个数	训练 MAPE	测试 MAPE
LinearRegression	Minkowski 参数	4	2.609 181 031 0	2.585 527 156 8
GradientBoostingRegressor	Minkowski 参数	4	2.301 006 277 3	2.517 798 802 6
KernelRidge	Minkowski 参数	4	2.560 741 821 2	2.503 187 917 3
XGBRegressor	Minkowski 参数	4	2.321 866 279 9	2.510 918 041 7
加权平均模型	Minkowski 参数	4	2.419 882 783 7	2.488 904 791 2
Stacking 模型	Minkowski 参数	4	2.470 594 745 8	2.490 279 546 2
多层神经网络	Minkowski 参数	4	2.627 999 203 3	2.651 129 301 7
3DCNN 模型	三维二元分割图像	64×64×64	0.886 797 757 5	1.849 176 499 5

6 结束语

文中提出了机器学习的方法在实验岩芯电导率的预测,其预测结果达到了较好的准确性。研究结果证明机器学习在岩石物理领域的强大潜力,为数字岩芯电导率的研究提供了可供参考的新方法。同时实验对比表明卷积神经网络比几何参数学习模型对数字岩芯电导率预测效果更优。计划在下一步研究中采用更深更复杂的卷积网络结构以谋求更好的效果,但几何参数仍然具有潜力,将添加更多的几何参数。在研究中初始样本只有一个岩芯样本,子采样的方式可能会导致图像和图像之间关联度强,造成实验误差。为了解决该问题,可能会采用对抗生成网络生成样本或者收集更多的岩芯样本。

参考文献:

[1] 刘向君,朱洪林,梁利喜. 基于微 CT 技术的砂岩数字岩石物理实验[J]. 地球物理学报,2014,57(4):1133-1140.

[2] 孔强夫,周灿灿,张 艳,等. 基于数字岩心岩石电性数值模拟方法综述[J]. 地球物理学进展,2015,30(2):718-724.

[3] SUN H,TAO G,VEGA S,et al. Simulation of gas flow in organic-rich mudrocks using digital rock physics[J]. Journal of Natural Gas Science and Engineering,2017,41:17-29.

[4] RAMANDI H L,MOSTAGHIMI P,ARMSTRONG R T. Digital rock analysis for accurate prediction of fractured media permeability[J]. Journal of Hydrology,2017,554:817-826.

[5] KNACKSTEDT M A,LATHAM S,MADADI M,et al. Digital rock physics:3D imaging of core material and correlations to

acoustic and flow properties[J]. The Leading Edge,2009,28(1):28-33.

[6] 印兴耀,秦秋萍,宗兆云. 数字岩石物理中弹性参数的有限差分计算方法[J]. 地球物理学报,2016,59(10):3883-3890.

[7] 黄 岩. 机器学习及其策略研究[J]. 中国新通信,2013(2):5-6.

[8] 余 凯,贾 磊,陈雨强,等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展,2013,50(9):1799-1804.

[9] MOSSER L, DUBRULE O, BLUNT M J. Reconstruction of three-dimensional porous media using generative adversarial neural networks [J]. Physical Review E, 2017, 96(4): 043309.

[10] MECKE K R. Additivity, convexity, and beyond: applications of minkowski functionals in statistical physics [J]. Lecture Notes in Physics,2000,554:111-184.

[11] 杨 博,刘钰洋,潘 懋. 基于 Minkowski 泛函和 K-means 聚类算法的岩石类型划分[J]. 科学技术与工程,2017,17(22):63-67.

[12] DIETTERICH T G. Ensemble methods in machine learning [C]//International workshop on multiple classifier systems. Berlin:Springer,2000:1-15.

[13] 唐小我. 最优组合预测方法及其应用[J]. 数理统计与管理,1992(1):31-35.

[14] 李 珩,朱靖波,姚天顺. 基于 Stacking 算法的组合分类器及其应用于中文组块分析[J]. 计算机研究与发展,2005,42(5):844-848.

[15] 刘 凯,仲 亮,张景忠. 基于 BP 神经网络算法的可见光图像热目标自动识别技术[J]. 科技创新与应用,2014(29):41.