石油科学通报
Petroleum Science Bulletin

基于机器学习方法的多采样点储层粒度剖面预测

刘珊珊，汪志明*

中国石油大学(北京)石油工程学院，北京 102249
* 通信作者, wangzm@cup.edu.cn

收稿日期: 2021−04−11
创新研究群体科学基金复杂油气井钻井与完井基础研究(编号 51821092) 资助

摘要　地层砂的粒度特征值d50(筛析曲线累计质量分数 50%对应的粒径值，μm)是防砂设计中的关键参数，为获得粒度纵向分布剖面，开展了基于机器学习方法的储层粒度与测井曲线响应关系研究。经典机器学习往往缺少模型内部的特征提取过程，而且采用单一采样点作为输入，缺失相邻数据关联关系反映层位信息。考虑到储层的地质连续性，利用测井曲线趋势和背景信息，将深度相邻数据点作为机器学习特征值，提出了一种基于多采样点的粒度剖面预测方法，构造和训练了基于随机森林(Random Forest)、支持向量机(Support Vector Machine)、Xtreme Gradient Boosting Tree(XGBoost)、人工神经网络(Artificial Neural Network)的预测模型。研究结果表明，与单点映射模型相比，考虑储层纵向地质连续性的各模型预测精度均高于单点预测，其中五点映射的ANN模型(ANN-5)预测效果最好，测试集d50预测相关系数最高为 0.819，误差MAE最小为 9.59，证实了多个采样点作为输入隐含利用了部分地层信息，有效地提高了预测精度。研究了特征点密度对模型准确率的影响，对训练集二维输入空间中样本的特征点高斯核密度分布以及测试集样本点处的训练集特征点密度进行估算，得出在高密度区域中的测试集样本点的RMSE普遍较低。当增加训练样本数量时，模型预测精度将进一步提高。采用层次分析法确定影响模型选择各因素的权重，通过模糊综合评判法优选机器学习模型，根据优选出的模型对临近区块储层粒度剖面进行预测，预测结果很好地捕捉了粒度变化趋势，模拟了其峰值。

关键词　机器学习；粒度剖面预测；测井曲线；地质纵向连续性

Reservoir grain size profile prediction of multiple sampling points based on a machine learning method

LIU Shanshan, WANG Zhiming

College of Petroleum Engineering, China University of Petroleum-Beijing, Beijing 102249, China

Abstract　The particle size characteristic (d50, the particle size value corresponding to 50% of the cumulative mass fraction of the sieve analysis curve, μm) of formation sand is a key parameter in sand control design. In order to obtain the vertical distribution profile of particle size, the response relationship between reservoir particle size and logging curve based on a machine learning method is studied. Classical machine learning often lacks a feature extraction process inside the model. Moreover, when a single sampling point is used as the input, the adjacent data association relationship is missing to reflect the horizon information. Considering the geological continuity of reservoirs, using the trend and background information of logging curves,

引用格式：刘珊珊, 汪志明. 基于机器学习方法的多采样点储层粒度剖面预测. 石油科学通报, 2022, 01: 93−105
LIU Shanshan, WANG Zhiming. Reservoir grain size profile prediction of multiple sampling points based on a machine learning method. Petroleum Science Bulletin, 2022, 01: 93-105. doi: 10.3969/j.issn.2096-1693.2022.01.009

taking the depth adjacent data points as machine learning eigenvalues, a grain size profile prediction method based on multiple sampling points is proposed. A prediction model based on random forest, support vector machine, Xtreme gradient boosting tree and artificial neural networks is constructed and trained. The results show that, compared with the single point mapping model, the prediction accuracy of each model considering the vertical geological continuity of reservoir is higher than that of single point prediction. The five point mapping ANN model (ANN -5) has the best prediction effect, with the highest correlation coefficient 0.819 and the least error measures 9.59 of the testing set. It is proved that multiple sampling points are used as input to implicitly utilize part of the stratum information and effectively improve the prediction accuracy. The influence of feature point density on the accuracy of the model is also studied. The Gaussian kernel density distribution of the feature points of the samples in the two-dimensional input space of the training set and the feature point density of the training set at the sample points of the test set are calculated. It is concluded that the RMSE of the sample points of the test set in the high-density area is generally low. The prediction accuracy of the model will be further improved as the number of training samples increases. AHP is used to determine the weight of each factor affecting the model selection, and fuzzy comprehensive evaluation is used to optimize the machine learning model. According to the optimized model, the grain size profile of the reservoir in adjacent blocks is predicted. The predictions capture well the trend of grain size change and simulate its peak value.

Keywords machine learning; grain size profile prediction; logging curve; geological vertical continuity

doi: 10.3969/j.issn.2096-1693.2022.01.009

0 引言

地层砂粒度分布PSD(Particle Size Distribution)在储层描述、沉积学，特别是在智能完井防砂技术中有重要应用[1]。其中粒度中值d50即筛析曲线上累重百分数50%对应的粒径，是油气开采地层评价和储层粒度分布特征参数之一，可为防砂方法的选择提供理论依据。粒度测量最常用的两种技术是筛析法和激光法，两种方法均需要通过岩心粒度测试来获取数据，钻井过程中储层取心费用昂贵，取心间隔有限，因此开发井取心数据较少，在制定开发井的完井防砂措施时往往没有实际开采层位的岩心，一般参照探井粒度数据进行设计，这种情况忽略了储层非均质性的影响。在某些情况下，设计是基于非常少的筛分数据，储层的纵横向各向异性和非均质性给防砂方案设计带来了困难和较大的风险。由于砂体内粒度的变化，在一口井中选择的d50不一定适用于同一油田的另一口井。而且对于分段分级防砂完井选择筛管或砾石粒径而言，获得整个储层连续粒度剖面具有重要意义。

近年来机器学习方法在科学和工程领域广泛应用，很多研究者也尝试使用数据驱动方法来解决地质问题[2–3]，例如利用支持向量机(SVM)、模糊逻辑模型(FLM)和人工神经网络(ANN)等方法来处理估算地球物理参数。储层的粒度特征是在漫长的历史过程中形成的，与沉积物的形成环境有很好的相关性。其中地层压实程度、孔隙度、以及黏土的含量等均在某一程度上可反应地层颗粒大小，可根据能够反映地层这些特性的测井曲线建立其与颗粒大小的映射关系。由于测井资料反映储层信息，不同的测井曲线实质上是同一储层在不同物理量下的反映，测井资料与储层颗粒特征之间存在映射关系。有学者建立了储层粒度分布预测的神经网络模型[4–7]。Oyeneyin和Faga[8]首先介绍了利用神经网络对粒度分布进行建模的概念，建模所需的数据是测井(电缆或随钻测井)和粒度数据。采用多层反向传播神经网络(BPNN)实现，最优拓扑结构为三层BPNN，三个神经元的隐藏层由sigmoid传递函数激活。Oluyemi[9]综合了统计和神经网络两种方法来预测定向井粒度分布(水平和垂直粒度分布)。粒度预测中主要使用的测井数据是伽玛曲线，因为伽玛曲线通常反映粒度－泥质含量的关系。其他测井曲线，如密度、中子、声波和电阻率等，可根据储层中流体的类型进行选择。在气藏中，仅结合伽玛曲线和密度曲线通常是最佳的。使用神经网络对粒度分布进行建模，将有助于更好地估计整个储层段的粒度分布。Faga[10]研究了成岩作用对神经网络粒度预测的影响，砂岩中粘土矿物的成岩作用影响其原生物性，包括对颗粒大小和形状、矿物成分、孔隙度、渗透率和沉积结构的影响。储层砂岩中自生粘土的分布广泛，成岩矿物的小规模变化会导致孔隙度和渗透率的大幅度波动[11]，这种变化在测井曲线上有所体现。Siron和Segall[12]对南卡罗来纳州沿海平原研究中指出，高岭石占主导地位的粘土含量较高对应高电阻率信号和低伽玛射线值；某些岩相内的高含量粘土降低了有效孔隙度和渗透率，对应较高电阻率。他们强调了将沉积学技术与测井数据结合起来，对地下岩性单元进行综合评价具有重要意义。

虽然有研究者尝试使用神经网络解决粒度预测问题，但是大多关注在全连接ANN的应用上，未有尝试其他机器学习模型。传统的ANN中，描述的是一种点对点的映射关系，以单点测井数据作为特征值来预测d50，也就是说ANN中生成的预测结果在空间上是完全相互独立的。换言之，通过ANN预测某一深度处储层粒度数值，仅与输入变量(作为输入的测井曲线)中相同深度处的不同物理量的测量值相关。因此ANN忽略了测井曲线随深度变化的趋势性信息以及数据的前后文(空间)关联中所蕴含的信息。本文提出了基于模糊数学综合评判优选机器学习模型方法，通过对比分析多种机器学习模型预测结果，优选出隶属度最高的方法用于新井预测储层粒度剖面，为防砂设计提供数据支撑。本文提出的方法为相关研究提供新思路，可用于根据钻井时获得的测井数据对地层粒度分布进行实时预测[13]。

1 方法原理

数据预处理对于机器学习获得准确的预测模型具有重要意义。测井数据作为模型的输入，由于各技术服务公司采用不同的数据编码格式采集软件，测井数据具有不同的数据格式，本文将所有LAS数据文件都转换为CSV文件的数据集。对测井数据进行平滑滤波处理、缺失值处理，剔除异常数据和空数据。根据井径测井消除井眼不规则性和冲洗段可能产生错误读数。实验采用位于南海北部湾海域某油田WZ11−4区块一口井伽玛射线(GR)与密度测井(Den)及实测d50数据作为训练集[14−15]，实验用水平井的测井段总长为44.7米，对应测深为962.6 m到1007.3 m。该构造位于南海北部北部湾盆地涠西南凹陷2号断裂带上升盘中部。整体埋深较浅，埋深1000 m左右，储层位于新近系角尾组二段地层，油藏类型为构造油藏。以该井作为训练样本，建立预测模型实现临近区块相同层位另一口井的粒度剖面预测。图1显示用于训练的测井曲线，表1列出了数据集不同统计特征，如计数、平均值、标准差、最小值、中值和最大值。体积密度在2.04和2.44 g/cm^{3}之间，GR在40.39和70.42 API度之间。还需可视化两个输入变量或一个输入变量与目标变量之间的关系，绘制了成对散点图(如图2)和单变量直方图(如图3)。图2显示了特征变量和预测值之间的关系，从图上可知没有异常值对模型预测产生影响，图3直方图及核密度估计分布曲线表明数据的分布规律接近正态分布。图4相关性矩阵表明用于训练模型的测井曲线与实验室测量的d50值之间的相对重要性。d50与GR和Den相对重要性分别为−0.45和0.0048。从相关性矩阵可以看出密度测井曲线与粒度中值具有正相关性。伽玛测井曲线与粒度中值有负相关性，伽玛测井反应泥质含量，伽玛值较高的层位泥质含量较高，对应储层粒度值较低。

在每个输入特征值的相同范围内缩放数据，可以最大限度地减少特征之间的偏差，加快模型的训练时间。在将图1特征参数引入模型训练之前，需对训练数据集进行标准化，同时，测试数据集采用训练集的均值和方差进行标准化。对于某一特征数据，可以采用公式(1)、(2)进行归一化。

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \tag{1}$$

$$x_j = \frac{x_j - \mu_j}{S_j} \tag{2}$$

式中，μ_j是归一化参数，x_j是实际参数，S_j是实际参数的标准偏差。

除了相关性分析外，本文还提出了多点映射的观

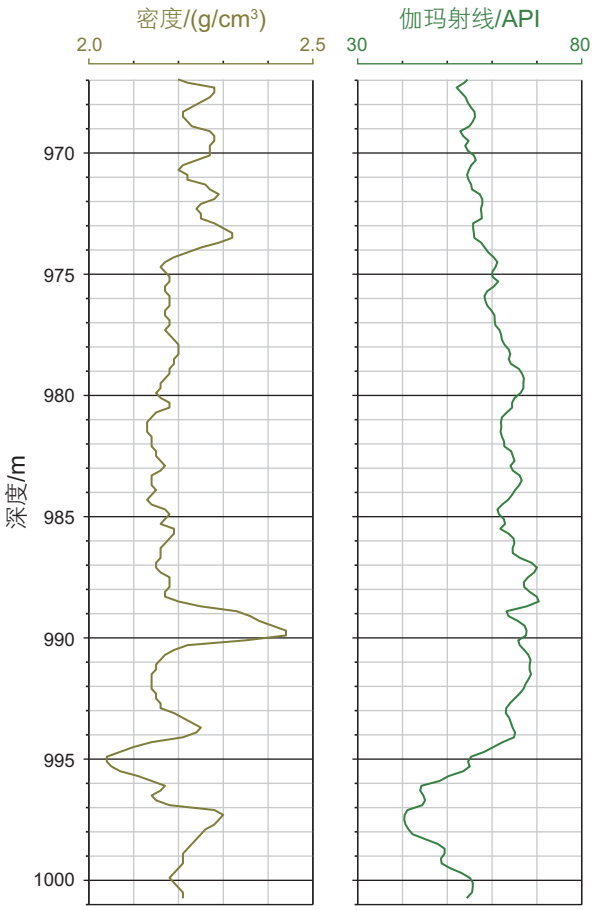

图1 电缆测井曲线

Fig. 1 Wireline logging curve

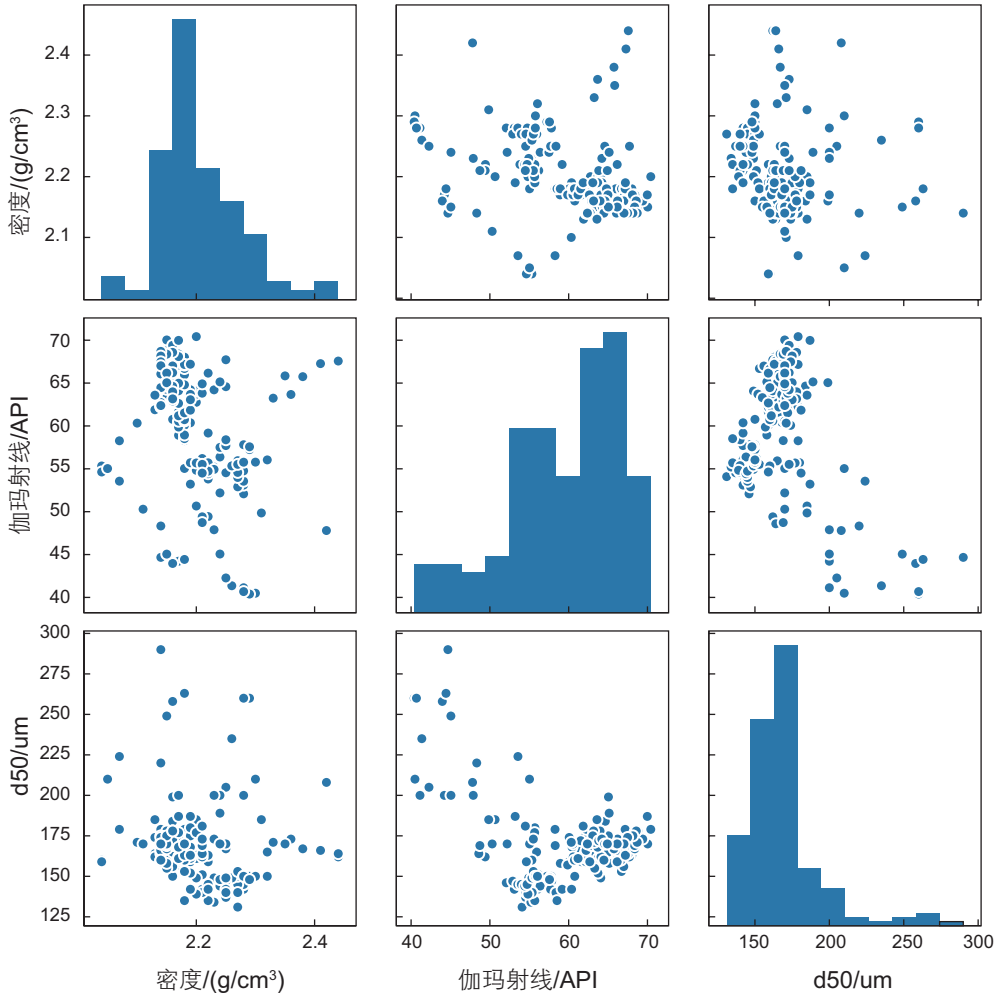

图 2 特征变量和预测值之间成对散点图

Fig. 2 Paired scatter plot between characteristic variable and predicted value

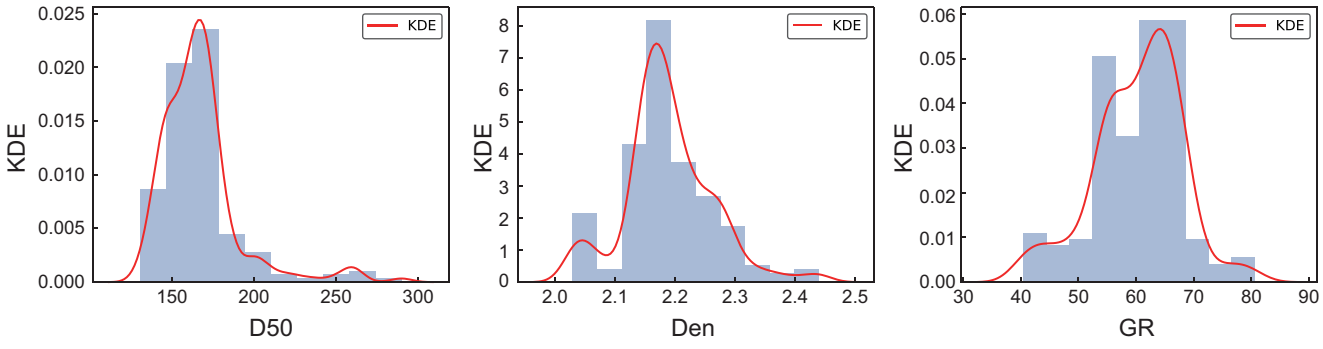

图 3 特征单变量的直方图

Fig. 3 Histogram of characteristic single variable

点来构建特征工程，其原理是基于地层的纵向连续性，考虑测井数据与粒度随测深的变化而变化的特性。测井采样间隔通常很小(0.1m)，测井仪器获得的不同深度地层的测井数据在纵向深度上相互影响。因此，在测井曲线中，每个数据点周围相互有影响的范围所包含的数据点有多个，这意味着d50 的预测可以看作是一个具有空间相关性的序列数据分析问题。为了更好地利用测井曲线的纵向连续性，选取深度上相邻多点特征作为训练特征，生成粒度剖面的过程综合考虑了测井曲线间的内在联系和不同测井曲线随深度的变化

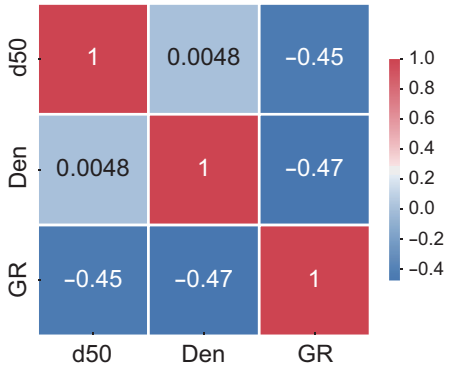

图 4　特征 Pearson 相关矩阵

Fig. 4　Characteristic Pearson correlation matrix

表 1　现场数据统计分析

Table 1　Statistical analysis of field data

	Den/(g/cm³)	GR/(API)
个数	172	172
均值	2.201	59.494
标准偏差	0.068	7.111
最小值	2.04	40.39
第一四分位数	2.16	55.168
第二四分位数	2.18	61.29
第三四分位数	2.24	64.828
最大值	2.44	70.42

趋势，更加符合地质学思想。特征选择的原理如图 5 所示。训练样本由特征参数和标签组成，根据相关系数矩阵，将 GR 和 Den 分别记录为特征，d50 作为训练标签。定义符号语言描述，输入变量表示为

$$X_K^N=\left[x_1^T,x_2^T,\cdots,x_{K-1}^T,x_K^T\right],\left(X^n\in R_x^K,n\in[1,N]\right) \quad (3)$$

式中，R_x^K 为 K 维输入空间，该空间每一维度的范围限制了特征工程的范围，且包括所有可能的取值组合；X_K^N 为输入变量，K 为特征的总数量，N 代表样本数量，每个列向量 $x_k^T k\in[1,K]$ 代表某一个特征(Den,GR)，$X^n n\in[1,N]$ 表示第 n 个样本的特征，是一个行向量，可用 $x_k^n k\in[1,K],n\in[1,N]$ 表示输入变量 X 中第 k 个特征在第 n 个样本中的值。

输出变量表示为：

$$y^{[N]}=\mathrm{y}^T,\left(y^n\in R_y,n\in[1,N]\right) \quad (4)$$

式中，R_y 为输出空间，$y^{[N]}$ 为输出变量，是输出空间的一个子向量，由 d50 样本组成，N 与输入变量的样本数量对应；列向量 y^T 代表 d50；$y^n n\in[1,N]$ 示第 n 个样本所对应的 d50。

因而，训练集可表示为：

$$D_s=\left\{\left(X^1,y^1\right),\left(X^2,y^2\right),\cdots,\left(X^N,y^N\right)\right\} \quad (5)$$

式中 D_s 为训练集的样本集合，该式表示一个拥有 N 个样本的训练集。

通过机器学习算法训练所获得的模型可表示 R_x^K 到 R_y 的映射：

$$H:R_x^K\rightarrow R_y \quad (6)$$

(6)式的意义为通过将单目标训练集 D_s 输入到机器学习算法中所训练得的模型 H，可描述特征空间和输出空间之间的关系，当有特征 $X=[x_1,x_2,\cdots,x_K]\left(X\in R_x^K\right)$ 的新测井数据输入到模型 H 中时，便可得到该储层 d50。

N 个样本按行存储为 $N\times(m+1)$ 矩阵，m 为特征点数。使用单点特征建模时，某个深度点的训练样本是单点特征和标签，如图 6 所示。奇数的采样点有中心点，选取不同奇数采样点数据特征建模。如图 7 所示，当选取 3 个点时一个样本的构造特征包含 3 个相邻点的 6 个特征，即 3×2 个特征，标签为该深度的 d50。x_q^P 中 P 代表点，q 代表单点的特征，选取 5 个点时原理同 3 点。可以看出多点参数的组合增加了与 d50 相关的信息量。

2　模型开发

模型配置一般指超参数，如随机森林算法中的 n 值、支持向量机中的不同核函数等，在大多数情况下，超参数的选择是无限的。通过绘制训练集和测试集的学习曲线，可以寻找模型的最优参数，以测试集上的泛化误差作为模型的最优参数，原理见图 8。实验采用 WZ11−4 井为数据集，70% 的数据用于训练，30% 用于测试，数据集邻域采样点分别取 1,3,5，实验环境：CPU 配置为 intel(R)Core(TM) i7−8565U @ 1.80 GHz 1.99 GHz，RAM 为 8 G，基于 Python 第三方

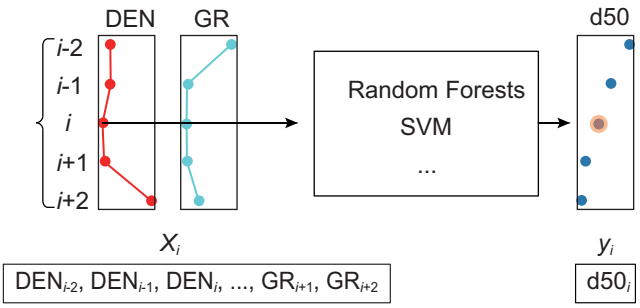

图 5　基于多采样点的构造特征

Fig. 5　Construction features based on multi sampling points

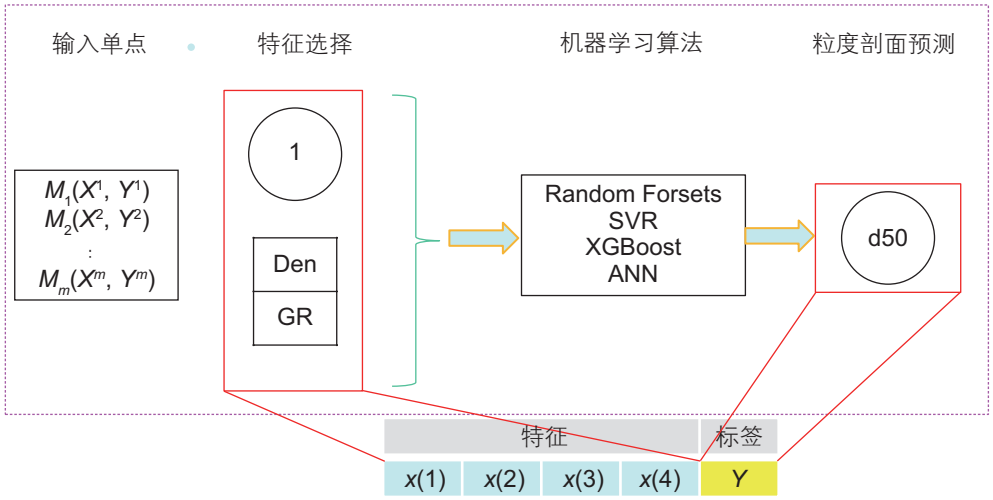

图 6 单个采样点预测原理

Fig. 6 Prediction principle of single sampling point

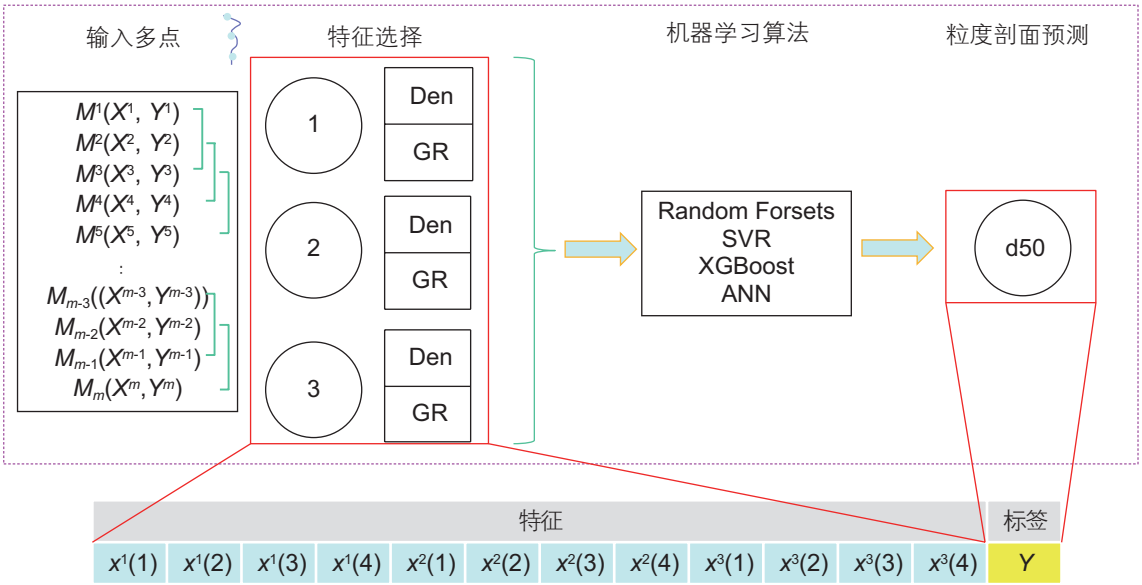

图 7 三个采样点预测原理

Fig. 7 prediction principle of three sampling points

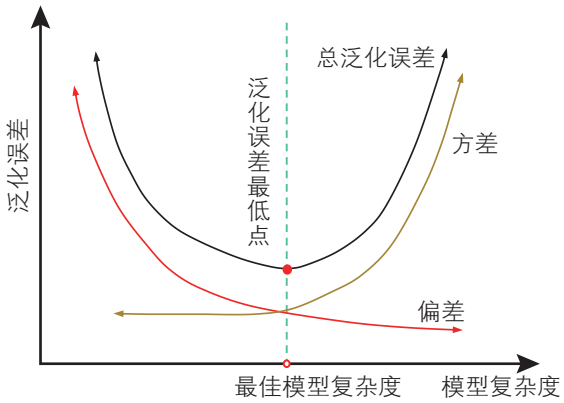

图 8 泛化误差与模型复杂度的关系

Fig. 8 Relationship between generalization error and model complexity

模块 SKlearn 实现随机森林 Random Forest，人工神经网络 ANN，XGBoost，支持向量机 SVR 四种回归模型训练，最优超参数解如表 2 所示。RF−1 代表单个采样点，RF−3 代表三个采样点，以此类推。在每个机器学习模型通过训练数据后，用最优超参数生成相应的拟合模型，粒度实际值与各模型预测值剖面图和交会图如图 9 所示，展示了 WZ11−4 井的学习效果，预测值与实测值两者交汇点较为集中，证明两者具有较高的相关性。

2.1 模型结果对比

表 3 列出了每个模型在训练和测试期间的性能

表 2　模型超参数最优解

Table 2　optimal solution of model super parameter

算法	超参数				
RF−1	弱评估器个数=40	树最大深度=7	发生分支后最小样本数=4	发生分支最小样本数=11	
RF−3	弱评估器个数=9	树最大深度=7	发生分支后最小样本数=5	发生分支最小样本数=9	
RF−5	弱评估器个数=61	树最大深度=5	发生分支后最小样本数=5	发生分支最小样本数=10	
XG−Boost−1	线性回归的损失函数	复杂度的惩罚项=1	学习率=0.3	树最大深度=1	弱评估器个数=116
XG−Boost−3	线性回归的损失函数	复杂度的惩罚项=0	学习率=0.2	树最大深度=1	弱评估器个数=143
XG−Boost−5	线性回归的损失函数	复杂度的惩罚项=0.3	学习率=0.2	树最大深度=3	弱评估器个数=74
SVR−1	高斯径向基核函数	核函数系数=0.5	多项式核函数次数=3	松弛系数惩罚项系数=1.0	核函数常数项=0
SVR−3	高斯径向基核函数	核函数系数=0.2	多项式核函数次数=3	松弛系数惩罚项系数=1.0	核函数常数项=0
SVR−5	高斯径向基核函数	核函数系数=auto	多项式核函数次数=3	松弛系数惩罚项系数=1.0	核函数常数项=0
ANN−1	激活函数='relu'	权重优化器='lbfgs'	神经元个数=(6)	最大迭代次数=4000	随机数种子=1
ANN−3	激活函数='relu'	权重优化器='lbfgs'	神经元个数=(3,4)	最大迭代次数=4000	随机数种子=1
ANN−5	激活函数='relu'	权重优化器='lbfgs'	神经元个数=(4,5)	最大迭代次数=4000	随机数种子=234

比较，通过模型训练和测试阶段的评价指标R^2(决定系数)、均方误差(MSE)、平均绝对误差(MAE)、均方根误差(RMSE)对各模型进行评估。根据模型误差的分析，所有训练集与测试集差异较小，这表明训练过程是可靠的(即没有过度拟合)。考虑到数据采集本身带有测量误差，因此可认为训练集误差相对较高的原因在于样品集合较大。这样的预测结果也能够证明该方法在训练集外仍能够取得可靠的结果。ANN 模型的预测效果最好，XGBoost模型次之，其次是RF和SVR。使用ANN建立的 5 点预测模型在预测方面优于其他模型。训练集和测试集的R^2分别为 0.891 和 0.819。各模型多点预测结果均高于单个采样点，用单个采样点作为输入对于噪声较为敏感。使用多个临近的采样点作为输入，降低了噪声对模型的影响，使模型具有更强的鲁棒性。由于数据间存在局部相关的特点，在邻域采样点数为 5 时，训练误差和测试误差均为最小，可以更准确预测储层粒度剖面。图 10 显示了使用五个点的ANN模型，给出了训练集中实际数据和预测数据之间的最高R^2和最低RMSE。由于模型有更多的信息在参数和目标之间建立更可靠的关系，因此使用更多的变量比存在于一个点的变量更好地预测了d50剖面。分析结果表明充分利用测井数据序列向前、向后两个方向的上下文关系可以取得更好的预测效果，证明了该方法在储层粒度预测方面的有效性。

2.2　特征点密度对模型准确率的影响

对机器学习模型来说，当往模型中输入一条地质与工程特征实例时，最终输出的粒度数据往往由与该实例相接近的训练集特征所决定，可以看出训练集样本的特征点的密集程度对最终的预测准确率会有一定的影响。理论上，若在输入空间中某一片区域聚集了大量的特征点，则该区域会被更好的覆盖，从而模型能更好的描述该范围内输入与输出空间的映射关系。

本文所选的特征组成了一个二维输入空间，以神经网络模型预测为例，利用高斯核密度估计[16]算法(Gaussian kernel density estimation)对训练集二维输入

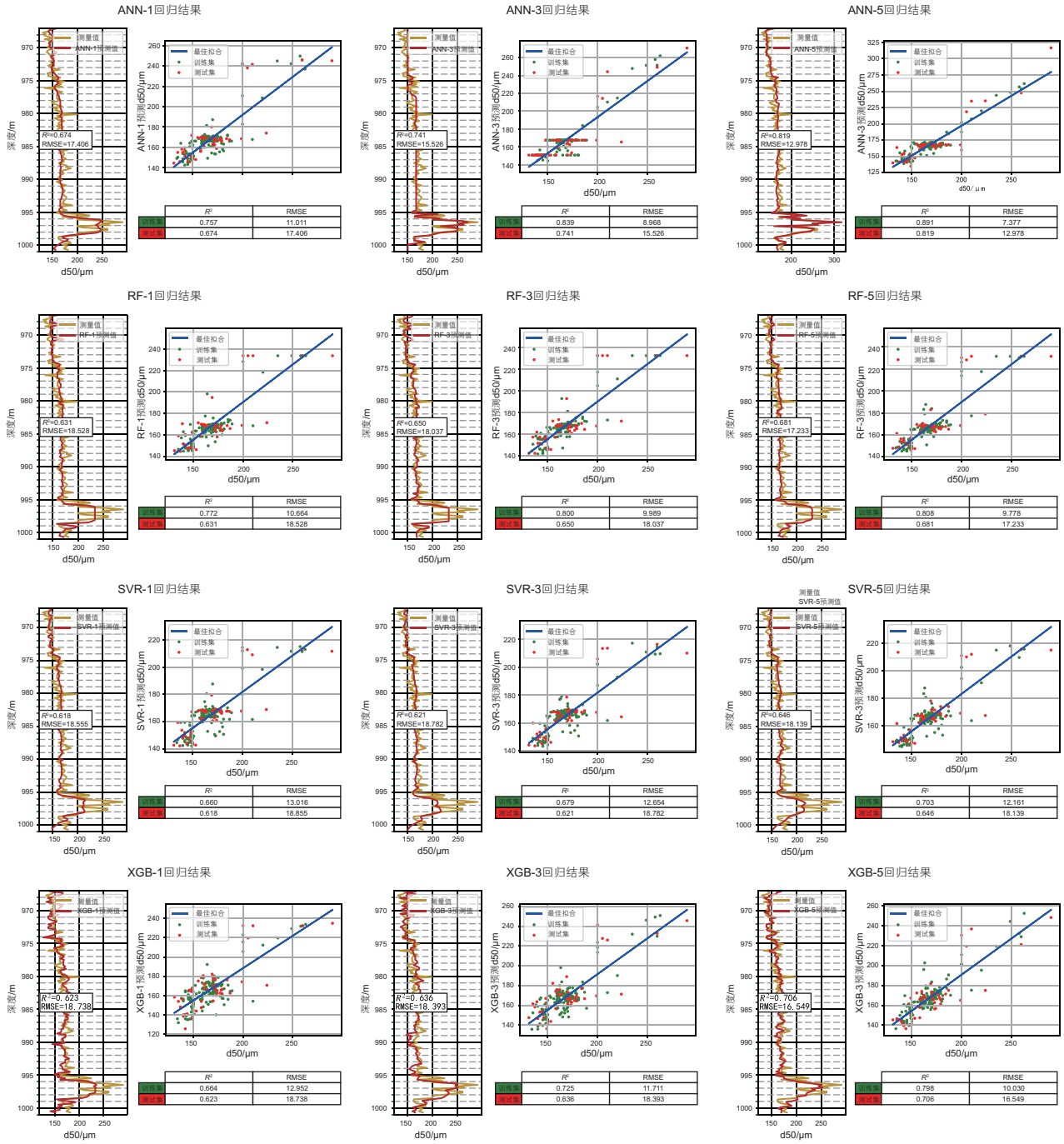

图 9 WZ11–4 井不同方法测量与预测的粒度剖面(左)和交叉图(右)的对比

Fig. 9 Comparison of grain size profile (left) and cross plot (right) measured and predicted by different methods in WZ11–4 well

空间中样本的特征点高斯核密度分布进行计算，并对测试集样本点处的训练集特征点密度进行估算。将计算结果与RMSE投影至“Den–GR”空间中并绘制成二维散点图，如图 11 所示，其中的“蓝–红”散点为 43 个测试集特征点及d50 均方根误差的分布，“蓝–黄”圆圈为当前点的高斯核密度，黑色散点是训练集样本的分布。可见，在高密度区域中的测试集样本点的 RMSE 普遍较低。

由于实验条件限制，本文只采用一口井作为训练数据，当数据集增大时，其输入空间的特征具有一定范围。此时，搜集一定量的样本使输入空间特征点达到一定的密集程度，便可以在预测该范围内d50 时达到较好的准确率。此外，还可以结合聚类算法，将输入空间中地质与工程特征相接近的高密度区域划分至同一区块，而后用不同区域的数据训练多个粒度预测模型。当预测新井时，先使用分类算法将其划分为某

表 3　模型预测结果对比

Table 3　Comparison of model prediction results

学习算法	训练集				测试集				训练时间 μs
	R^2	MSE	MAE	RMSE	R^2	MSE	MAE	RMSE	Train time
RF−1	0.772	113.715	7.404	10.664	0.631	343.274	13.581	18.528	152578
RF−3	0.800	99.778	7.189	9.989	0.650	325.324	13.344	18.037	112742
RF−5	0.808	95.612	7.059	9.778	0.681	296.981	12.942	17.233	137632
SVR−1	0.660	169.423	8.620	13.016	0.618	355.506	12.078	18.855	2993
SVR−3	0.679	160.126	8.057	12.654	0.621	352.746	11.445	18.782	998
SVR−5	0.703	147.893	7.778	12.161	0.646	329.032	11.250	18.139	1995
XGB−1	0.664	167.746	9.727	12.952	0.623	351.116	14.785	18.738	791513
XGB−3	0.725	137.151	9.047	11.711	0.636	338.306	14.502	18.393	45879
XGB−5	0.798	100.594	7.000	10.030	0.706	273.864	12.296	16.549	33910
ANN−1	0.757	121.252	7.993	11.011	0.674	302.965	12.908	17.406	130945
ANN−3	0.839	80.417	6.578	8.968	0.741	241.067	11.712	15.526	168717
ANN−5	0.891	54.425	5.759	7.377	0.819	168.422	9.590	12.978	100349

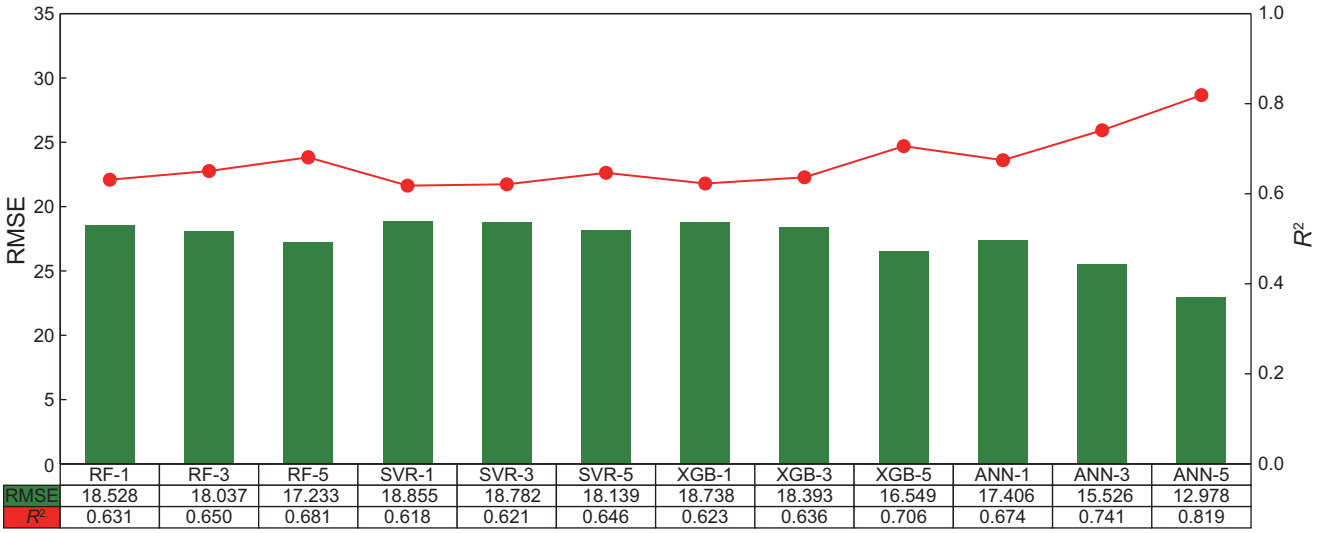

图 10　模型预测结果对比直方图

Fig.10　Comparison histogram of model prediction results

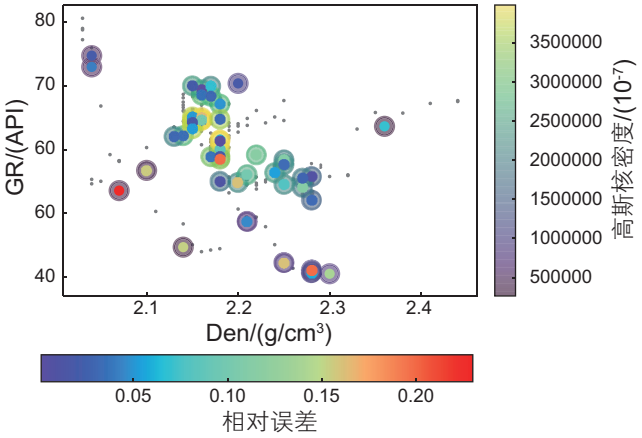

图 11　“密度－均方根误差”关系二维热力散点图

Fig. 11 Two dimensional thermal scatter diagram of “density root mean square error” relationship

一类的储层，并用所对应的预测模型开展预测，进一步提升预测的准确率。

3　机器学习模型优选

机器学习模型的选择在一定程度上会影响数据分析效果，为确定统一比较标准，更好的选择模型，采用层次分析法确定影响模型选择各因素的权重，然后利用模糊综合评判法[17]选择最优机器学习模型用于预测应用。

假设备选方案中有 m 个机器学习模型，每个模型有 n 个评价指标，由此建立特征向量矩阵(式 7)。

$$A=\left(a_{ij}\right)_{m\times n}(i=1,2,\ldots,m\ j=1,2,\ldots,n) \tag{7}$$

其中：a_{ij} 是属于第 i 个决策方案的第 j 个索引的值。

追求模型预测准确是机器学习的核心目标，能够同时处理大量数据，可以在超短时间内极速学习，是机器学习的重要优势。根据上文分析，选取训练集，测试集 R^2、RMSE、Traintime 为评价因素。获得这些评价指标的数据(表 4)，根据各方案具体指标求得特征向量矩阵 Y。

采用梯形分布与半梯形分布函数确定隶属度函数，进行归一化处理，R^2 越大、RMSE 与 Traintime 越小模型效果越好，根据(8)和(9)式建立隶属度矩阵 R[17]。

偏小型(越小越好)，见图 12a。

$$A(x)=\begin{cases}1 & x<a\\ \dfrac{b-x}{b-a} & a\leqslant x\leqslant b\\ 0 & x>b\end{cases} \tag{8}$$

偏大型(越大越好)，见图 12b。

$$A(x)=\begin{cases}0 & x<a\\ \dfrac{x-a}{b-a} & a\leqslant x\leqslant b\\ 1 & x>b\end{cases} \tag{9}$$

式中，a 为评价指标最小值，b 为评价指标最大值，x 为评价指标。

对各指标打分，建立判断矩阵(表 5)，从而计算出指标的权重值(表 6)。

采用加权平均算法，根据指标隶属度矩阵和各指标总权重，计算出四种预测模型的隶属度数值，如表 7 所示，根据最大隶属度法选择了预测效果最好的模型为 ANN-5。

4 模型验证与应用

为了进一步验证该模型泛化能力，使用邻近区块相同层位另一口井 WZ11-1E 现场数据进行粒度特征值 d50 的纵向剖面连续预测，该数据集没有参加模型训

表 4 评价指标

Table 4 Evaluation index

算法	训练集 R^2	训练集 RMSE	测试集 R^2	测试集 RMSE	训练时间/us
RF−1	0.772	10.664	0.631	18.528	152578
RF−3	0.800	9.989	0.650	18.037	112742
RF−5	0.808	9.778	0.681	17.233	137632
SVR−1	0.660	13.016	0.618	18.855	2993
SVR−3	0.679	12.654	0.621	18.782	998
SVR−5	0.703	12.161	0.646	18.139	1995
XGBoost−1	0.664	12.952	0.623	18.738	791513
XGBoost−3	0.725	11.711	0.636	18.393	45879
XGBoost−5	0.798	10.030	0.706	16.549	33910
ANN−1	0.757	11.011	0.674	17.406	130945
ANN−3	0.839	8.968	0.741	15.526	168717
ANN−5	0.891	7.377	0.819	12.978	100349

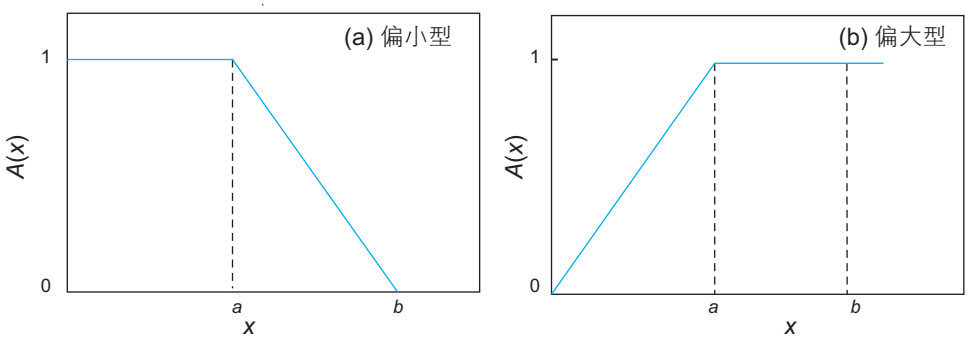

图 12 隶属度函数计算依据

Fig. 12 Calculation basis of membership function

练。图 13 显示了该井的电缆测井数据，其中包括伽玛射线和体积密度测井以及应用优选出训练后的 ANN-5 模型预测结果。从岩芯测得的实际中值粒度也被标绘出来，该井包含 13 个实验室测量 d50 值。将实际岩心

表 5　判断矩阵

Table 5　Judgment matrix

	训练集 R^2	训练集 RMSE	测试集 R^2	测试集 RMSE	训练时间 /us
训练集 R^2	1	2	0.2	0.2	3
训练集 RMSE	0.5	1	0.2	0.2	3
测试集 R^2	5	5	1	2	3
测试集 RMSE	5	5	0.5	1	3
训练时间 /us	1/3	1/3	1/3	1/3	1

表 6　指标权重

Table 6　Index weight

指标	权重
训练集 R^2	0.1207
训练集 RMSE	0.0982
测试集 R^2	0.4024
测试集 RMSE	0.304
运算时间 /s	0.0737

表 7　不同模型隶属度

Table 7　Membership degrees of different models

隶属度			
RF−1,3,5	0.203	0.296	0.405
SVR−1,3,5	0.074	0.1	0.204
XGBoost−1,3,5	0.019	0.187	0.491
ANN−1,3,5	0.335	0.641	0.991

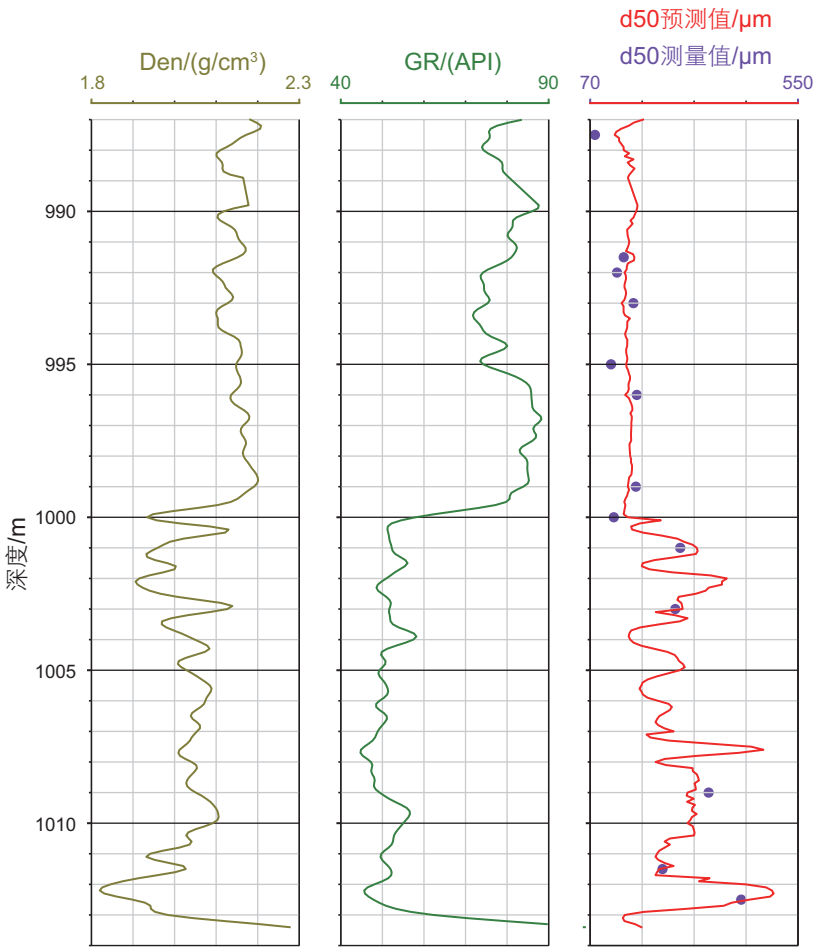

图 13　WZ11−1E 井测井曲线与 d50 预测结果

Fig. 13　Logging curve and D50 prediction results of wz11−1E well

粒度与预测值进行比较，预测结果很好地捕捉了粒度变化趋势，模拟了其峰值。

5 结论

(1)本文提出了考虑储层纵向连续性的地层砂粒度中值机器学习预测方法，该方法充分利用测井曲线随深度变化的趋势信息和以往数据空间关联所包含的信息，从储层沉积连续性角度兼顾了粒度预测问题研究中的空间尺度效应。通过选取合适的邻近采样点，确定测井数据的输入样本，有效利用地层的层位信息，符合地质沉积的有序性，具有比传统模型更高的准确性。机器学习是建立非线性关系智能模型的有效手段，目前，机器学习与工程实践的结合过于直接，主要是单向应用，较少涉及具体领域的知识。已有研究表明，将领域知识转化为模型的约束或先验信息加以利用可以突破提高模型效果的瓶颈，进一步提高模型的预测精度。

(2)采用 4 种机器学习方法(ANN、RF、SVR、XGBoost)建立了数据驱动的d50 预测模型并进行了对比试验，根据测井曲线趋势和背景信息，提取纵向连续点作为机器学习特征参数，并讨论了特征点密度对模型精度的影响。研究结果表明，无论采用哪种机器学习方法，多个采样点预测的精度都高于单点预测。基于 5 个采样点的ANN模型在训练集和测试集中具有最高的R^2 和最低的RMSE。提出了模糊数学综合评判优选机器学习模型方法，考虑到计算时间和精度，实际应用中，根据优选出的ANN-5 模型对临近区块储层粒度剖面进行预测，取得了良好效果。

参考文献

[1] BIXENMAN, P.W., TOFFANIN, E.P., and M.A. SALAM. Design and Deployment of an Intelligent Completion with Sand Control[C]. Paper presented at the SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, September 2001.

[2] LIU S S, ZHAO Y P, WANG Z M. Artificial Intelligence Method for Shear Wave Travel Time Prediction considering Reservoir Geological Continuity[J]. Mathematical Problems in Engineering, vol. 2021. , Article ID 5520428, 18 pages, 2021.

[3] 倪维军, 李琪, 郭文惠, 等. 基于支持向量机的页岩储层横波速度预测[J]. 西安石油大学学报(自然科学版), 2017,32(4) : 46−49,54. [NI W J,LI Q, GUO, W H, et al. Prediction of shear wave velocity in shale reservoir based on support vector machine[J]. Journal of Xi ' an Shiyou University (Natural Science Edition), 2017, 32(4) :46−49,54.]

[4] RIDER M.H. Gamma-ray log shape used as a facies indicator: critical analysis of an oversimplified methodology[J]. Geological Society London Special Publications, 1990, 48(1), 27−37.

[5] 李国和, 郑阳, 李莹, 等. 基于深度信念网络的多采样点岩性识别[J]. 地球物理学进展, 2018,33(4):1660−1665. [LI G H, ZHENG Y, LI Y, et al. Lithology recognition of multi-sampling points based on deep belief network[J]. Progress in Geophysics, 2018,33(4) :1660−1665.]

[6] 程希, 程宇雪, 程佳豪, 等. 基于机器学习与大数据技术的地球物理测井系统[J]. 西安石油大学学报(自然科学版), 2019,34(6):108−116. [CHENG X ,CHENG Y X,CHENG J H, et al. Geophysical logging system based on machine learning and big data technology[J]. Journal of Xi ' an Shiyou University (Natural Science Edition), 2019,34(6): 108−116.]

[7] 陈云天. 基于机器学习的测井曲线补全与生成研究[D]. 北京大学, 2020. [CHEN Y T. Research on Well Log Completion and Generation Based on Machine Learning[D]. Peking University,2020.]

[8] OYENEYIN, B.M., FAGA, A.T. Formation-Grain-Size Prediction Whilst Drilling: A Key Factor in Intelligent Sand Control Comple−tions[C]. SPE Paper No. 56626, 1999.

[9] OLUYEMI, GBENGA , OYENEYIN, BABS , and CHRIS MACLEOD. Prediction of Directional Grain Size Distribution: An Integrated Approach[C]. Paper presented at the Nigeria Annual International Conference and Exhibition, Abuja, Nigeria, July 2006.

[10] FAGA, A.T., and B.M. OYENEYIN. Effects of Diagenesis on Neural-Network Grain-Size Prediction[C]. Paper presented at the SPE Rocky Mountain Regional/Low-Permeability Reservoirs Symposium and Exhibition, Denver, Colorado, March 2000.

[11] GRIGSBY, JEFFRY D., LANGFORD, RICHARD P. Effects of diagenesis on Enhanced-Resolution Bulk Density Logs in Tertiary Gulf Sandstones: An Example from the Lower Vicksburg Formation, McAllen Ranch Field, South Texas[J]. AAPG Bulletin, V.80 No. 11 (Nov 1996), P. 1801−1819.

[12] SIRON, DONALD L., SEGALL, MARYLIN P. Influences of depositional environment and diagenesis on geophysical log response in the South Carolina Coastal Plain: effects of sedimentary fabric and mineralogy[J]. Sedimentary Geology, 1997,108(1−4):163−180.

[13] KANFAR, RAYAN, SHAIKH, OBAI, YOUSEFZADEH, MEHRDAD, TAPAN MUKERJI. Real-Time Well Log Prediction from Drilling Data Using Deep Learning[C]. Paper presented at the International Petroleum Technology Conference, Dhahran, Kingdom of

Saudi Arabia, January 2020.

[14] 李萍. 弱固结砂岩机械防砂优化设计研究[D]. 中国石油大学(北京), 2012. [LI P. A Study of the Optimization Design of Mechanical Sand Control in Weakly Consolidated Reservoirs [D]. China University of Petroleum (Beijing), 2012.]

[15] 王利华, 楼一珊, 马晓勇, 等. 储层粒度神经网络预测模型研究[J]. 西南石油大学学报(自然科学版), 2016,38(1):53−59.[WANG L H, Lou Y S, MA X Y, et al. Research on Neural Network Prediction Model of Reservoir Particle Size[J]. Journal of Southwest Petroleum University (Science&Technology Edition), 2016,38(1):53−59.]

[16] PARZEN E. On estimation of a probability density function and mode[J]. The Annals of Mathematical Statistics, 1962, 33(3):1065−1076.

[17] ZENG Q S, WANG Z MI, YANG G, WEI J G. Selection and Optimization Study on Passive Inflow Control Devices by Numerical Simulation[C]. Paper presented at the SPE Middle East Intelligent Energy Conference and Exhibition, Manama, Bahrain, October 2013.

(责任编辑 李世远 编辑 马桂霞)