ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 572－581
DOI:10. 16337/j. 1004-9037. 2020. 03. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多层集成学习的岩性识别方法

段友祥[1]，赵云山[1]，马存飞[2]，姜文煊[1]

(1. 中国石油大学(华东)计算机与通信工程学院,青岛,266580; 2. 中国石油大学(华东)地球科学与技术学院,青岛,266580)

摘　要：岩性识别是油藏地质解释中的关键问题和难点问题,人工智能特别是机器学习技术的发展和应用为岩性识别问题解决提供了新的技术途径。本文利用支持向量机(Support vector machine,SVM)、多粒度级联森林(Multi-grained cascade forest,GCForest)、随机森林(Random forest,RF)以及XGBoost(eXtreme gradient boosting)等机器学习模型建立一个异构多层集成学习模型,该集成学习模型克服了单一模型对数据集要求高、泛化能力差以及识别精度低等缺点。本文分别利用集成模型和单一模型进行了岩性识别实验。实验结果表明,本文集成模型在岩性分类测试集上平均精度达到96.66%,高于SVM的平均精度75.53%、GCForest的平均精度96.21%、随机森林的平均精度95.06%和XGBoost的平均精度95.77%。该集成模型能有效地用于油藏地质分析中的岩性识别和分类任务,适应性强,识别精度高。

关键词：岩性识别;SVM;GCForest;随机森林;XGBoost;集成模型
中图分类号：TP391　　**文献标志码**:A

Lithology Identification Method Based on Multi-layer Ensemble Learning

DUAN Youxiang[1], ZHAO Yunshan[1], MA Cunfei[2], JIANG Wenxuan[1]

(1. College of Computer and Communication Engineering, China University of Petroleum, Qingdao, 266580, China; 2. School of Geosciences, China University of Petroleum, Qingdao, 266580, China)

Abstract: Lithology identification is a key and difficult problem in reservoir geological interpretation. The development and application of artificial intelligence, especially machine learning technology, provides a new technical way to solve lithology identification problems. This paper uses machine learning models such as support vector machine (SVM), multi-grained cascade forest (GCForest), random forest (RF) and eXtreme gradient boosting (XGBoost) to build a heterogeneous multi-layer integrated learning model. The integrated learning model overcomes the shortcomings of single model such as high requirement for data sets, poor generalization ability and low recognition accuracy. In this paper, lithology recognition experiments are carried out using integrated models and single models. The experimental results show that the average accuracy of the integrated model is 96.66%, higher than that of SVM (75.53%), GCForest (96.21%), random forest (95.06%) and XGBoost (95.77%). The integrated model can be effectively applied to lithology identification and classification tasks in reservoir geological analysis with strong

基金项目:国家"十三五"科技重大专项(2017ZX05009-001)资助项目。
收稿日期:2019-07-08;**修订日期**:2019-10-31

adaptability and high recognition accuracy.

Key words: lithology identification; SVM; GCForest; random forest; XGboost; integrated model

引　　言

岩性识别是油藏描述、实时钻井监控、储层参数求解及储层评价的基础。传统上进行岩性识别时,常用的方法是钻井取心,但是取心过程复杂、成本高,因此油田的取心井一般较少[1]。测井是利用岩层的电化学特性、导电特性、声学特性以及放射性等地球物理特性,测量地球物理参数的方法,即获得的测井数据是地下岩石的物理响应信息,这些信息最典型的表现形式就是测井曲线。基于测井信息的岩心分析、解释和识别,就是通过研究和分析岩性差异表现敏感的测井曲线,建立相应的解释模型(例如响应方程、图版等),然后通过人工或计算机对获得的测井数据进行处理,应用解释模型对岩性进行解释和识别,基于测井信息的岩性分析(地层评价)是油藏储层评价的重要手段之一,一直是人们研究的重点[2-3]。

多年来,人们提出和发展了很多利用测井数据进行岩性解释的方法,归纳总结可分为传统的基于测井曲线响应特征的定性解释方法,基于测井响应方程的定量解释方法、图版法,基于机器学习的支持向量机(Support vector machine, SVM)、人工神经网络以及随机森林(Random forest,RF)方法等。其中定性解释方法和图版法的可靠性取决于解释人员的实践经验和剖面的复杂程度,而且需要人工方式进行处理,受人为因素影响大,解释效率较低,另外图版法只能利用部分测井资料,不能实现全部测井信息的有效利用;定量解释方法通过建立测井响应方程实现岩性识别,相比于定性方法可靠性更高,但其受限于地层矿物成分种类,对致密储层的适用性较差;以人工神经网络为代表的基于机器学习的岩性解释方法具有数据处理高度自动化、岩性识别智能化等特点,很早就受到研究者的关注,也取得了较好的研究成果[3-5]。近几年机器学习迎来了新的发展机遇,基于机器学习新技术的岩心识别方法成为新的研究热点[6-8]。机器学习发展出了很多方法,主要包括:(1)有监督学习,如决策树、回归、感知器、K近邻、贝叶斯、人工神经网络以及卷积神经网络等;(2)无监督学习,如降维、聚类等;(3)强化学习,如策略迭代、蒙特卡洛等。这些方法的模型和原理不同,在问题求解时的适用条件和表现差异比较大。因此,单一学习器很难在数据量不足的情况下拟合为一个稳定、可靠的模型。为此人们提出将多个有偏好的模型(在某些方面表现的较好)组合起来,以期得到一个更好更全面的模型,即集成学习。本文面向油藏地质分析中的岩性识别实际问题,在分析和研究已有机器学习方法的基础上,提出了一种基于多层集成学习思想的岩性识别模型。

1　预处理

使用机器学习方法所构建的岩性识别模型都需要多维数据作为样本进行训练,而SVM和随机森林作为一种有监督的机器学习方法还需要给样本加上准确可靠的标签,并且为模型提供合理的输入特征。输入特征的好坏直接影响模型训练和预测精度[9]。各个属性通常有着不同的量纲,并且数据特征往往属于不同的正态分布,这会降低模型的训练速度,增加过拟合风险,因此需要对数据进行标准化处理。

(1)数据的中心化。对于原始测井样本数据集,利用式(1)进行处理,可得到均值为0、标准差为1的数据,有

$$x_1 = \frac{x - \mu}{\sigma} \tag{1}$$

式中：x为样本数据；μ为样本数据均值；σ为样本标准差。

(2)数据的归一化。将测井数据样本集的所有属性值利用式(2)进行处理，将其归一化到(0,1)之间，消除量纲的影响，有

$$x_2 = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} \tag{2}$$

特征选择是指从原始特征集中按照某种评估标准找出最优的特征子集，即找到与标签高度相关的特征，从而达到用较少的特征变量得到更好的实验结果的目标。随机森林是一种优秀的特征选择算法，在进行包裹式特征选择时，随机森林中决策树生成阶段需要进行采样，所以可以使用袋外数据(Out of bag，OOB)进行验证，不用单独划分测试集[10]。本文使用随机森林算法进行特征选择。

(1)计算特征重要性

使用随机森林计算特征f重要性的计算过程如下。首先，随机森林中的每一棵决策树，使用对应的袋外数据计算它的袋外数据误差，记为e_1；然后随机地对袋外数据所有样本特征加入噪声干扰，再次计算它的袋外数据误差，记为e_2。假设随机森林中有N棵树，那么特征f的重要性I_f的计算公式为

$$I_f = \frac{\sum_{n=1}^{N}(e_2 - e_1)}{N} \tag{3}$$

若给某个特征随机加入噪声之后，袋外的准确率大幅度降低，则说明该特征对于样本的分类结果影响很大，重要程度比较高。

(2)特征选择

前面虽然评估出单个特征的重要性，但是并没有去除冗余特征。基于特征重要性进行的特征选择有很多方法，本文使用的是一种较为简单的方法，步骤如下：

(a) 对随机森林中的特征变量按照式(1)计算其重要性，并进行降序排序；

(b) 确定删除比例，然后剔除一定比例的特征，得到新的特征集；

(c) 重复上述过程，直到剩下m个特征为止。

2 多层集成学习模型

通过深入分析和研究目前常用的单一或集成的机器学习模型，将它们有机融合，设计出适用于岩性识别的多层集成学习模型。通过阅读岩性识别研究的相关文献，并结合实际需要，选择了SVM、多粒度级联森林(Multi-grained cascade forest，GCForest)、随机森林和XGBoost(eXtreme gradient boosting)4种机器学习模型组成新的集成学习模型[11-12]。

2.1 SVM模型

SVM是20世纪90年代中期发展起来的基于统计学习理论的一种可用于分类和回归的机器学习模型[13-14]。SVM用于分类时的基本原理是根据结构风险最小化原则寻找最优线性超平面将样本分开，使其在未知的测试样本集上误差最小(即具有良好的泛化性能)，且对于线性可分的数据，最优的分类超平面一定存在；对于数据线性不可分数据，不能找到分类超平面，这时可以使用核函数将输入数据映射到高维的特征空间，形成线性可分的稀疏数据。SVM是一种有坚实理论基础的机器学习模型。该模型的优点：(1)稀疏性好，即少量样本就可以获得较好的分类效果；(2)适应复杂的非线性问题；(3)无局

部极小值问题(相对于神经网络等算法);(4)可以很好地处理高维数据集;(5)算法简单,泛化能力比较强,有较好的鲁棒性。该模型的缺点:(1)对于核函数的高维映射解释力不强,尤其是径向基函数;(2)解决多分类问题存在困难;(3)对缺失数据敏感;(4)大规模数据计算复杂度大。

2.2 GCForest模型

深度森林(Deep forest)又叫GCForest[15],是南京大学周志华教授提出的一种可以媲美深度神经网络的基于决策树的集成方法,其结构主要包括级联森林和多粒度扫描。级联森林的每一层由若干个不同类型的森林(即随机森林和完全随机森林)组成,每个森林由多个决策树组成,每个决策树都会得到一个决策结果,综合所有决策树的决策结果,取其均值,得到每个森林的决策结果,然后将得到的每个森林的决策结果通过k折交叉验证形成类向量与样本的原始特征向量进行拼接,将其作为下一层的输入。多粒度扫描通过使用多个大小不同的滑动窗口扫描原始特征,获得不同粒度的特征向量,以此来增强级联森林的差异性。该模型的优点:(1)容易训练,计算开销小,效率高;(2)模型超参数较少,并且对参数设置不敏感;(3)适用范围广,即使是小规模数据集,也可以获得较好的训练结果;(4)每个级联的生成使用了交叉验证,降低过拟合风险;(5)相比深度神经网络更容易进行理论分析。该模型的缺点:运行时间比较长;受计算资源限制,并未具体应用在大规模数据集上。

2.3 RF模型

RF模型是Breiman等提出的一种集成学习模型,它由多个决策树基分类器并行组合而成[16]。它基于BootStrap抽样原理从原始数据集中有放回的随机选择有差异的若干个训练样本子集;以无剪枝决策树为基分类器对每抽样样本子集进行建模,并将生成的多个决策树集成,利用所构建的决策树森林对测试数据进行分类投票,最后由投票结果决定样本的最终分类或预测结果。

随机森林中的随机性主要体现在:(1)随机采样。随机森林在计算每棵树时,从全部n个训练样本(样本数为n)中选取一个可能有重复的、大小同样为m的数据集进行训练。(2)特征选取的随机性。在每个节点随机选取所有特征中的部分特征,用来计算最佳分割方式。该模型的优点:(1)所有的数据都能够有效利用,不用人为的分出一部分数据来做交叉验证;(2)在少量参数的情况也可以实现很高的精确度,对于分类和回归问题都适用;(3)不用担心过拟合的问题;(4)不需要事先做特征选择,每次只需随机的选取几个特征来训练树。该模型的缺点:(1)解决回归问题的表现没有在分类问题中好;(2)难以控制内部的运行;(3)对于小数据或者低维数据,效果可能不是很好;(4)决策树个数较多时,空间和时间复杂度较大。

2.4 XGBoost

XGBoost是Chen在梯度提升树(Gradient boosting decision tree,GBDT)模型基础上提出的一种基本的集成学习模型[17-18],因此其主要思想也是在降低残差的方向上训练一个新的模型,即不断地添加树,不断地进行特征分裂来生长一棵树,每次添加一个树,就是学习一个新函数,去拟合上次预测的残差。但XGBoost模型将GBDT模型的代价函数进行了改进,将表达误差近似的一阶泰勒展开改进为二阶的泰勒展开,同时使用代价函数的一阶导数和二阶导数来表达误差近似,并且在代价函数中加入了正则化项。二阶导数的应用加快了模型的收敛速度,正则化项的加入降低了模型的复杂程度,从而避免过拟合。该模型的优点:(1)使用许多策略防止过拟合,如正则化项、特征抽样等;(2)目标函数优化利用了损失函数,只要满足二阶连续可导都可作为损失函数;(3)支持并行化,树与树之间是串行关系,但是同层级节点可并行,训练速度快;(4)添加了对稀疏数据的处理;(5)支持设置样本权重。该模型的缺点:时间和空间的复杂度高。

通过上面的分析可以看出每个单一机器学习模型都有自己的优势,同时也有不足。SVM可以在小

数据集上工作良好,选取合适的核函数可使SVM在非线性数据上有很好的表现;GCForest是一种基于决策树的集成方法,易进行理论分析,适用范围广,无论是大数据集还是小数据集都适用;随机森林是一种决策树并行集成学习模型,直观易于理解,可解释性好,适用于方差大、偏差小的数据集;XGBoost是一种决策树串行集成学习模型,适用于方差小、偏差大的数据集。

2.5 多层集成学习模型

融合SVM稀疏性好、GCForest适用范围广、随机森林适合高方差的数据和XGBoost适合高偏差的数据等单一机器学习模型的各自优势[19-25],本文基于多层(随机森林和XGBoost是决策树模型的两种基本集成模型)集成的思想,设计了一个岩性识别多层集成学习模型,该模型具有更强的抗干扰能力和特征组合能力。模型结构如图1所示。

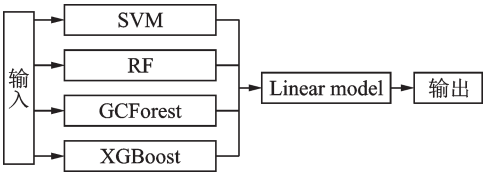

图1 多层集成学习模型结构

Fig.1 Multi-layered ensemble learning model structure

该多层模型集成多个异质模型(其中包括Bagging模型和Boosting模型),可以根据不同模型的特点对数据特征有不同侧重,具有很好的泛化能力,同时这几种模型也是数值特征数据进行机器学习最常用的模型。SVM适用于各种小样本状态,同时核函数使得它能拟合复杂数据;GCForest对于数据量没有要求,只需要很少的训练数据就能够获得较好的性能,并且对参数设置不敏感;随机森林是并行集成模型,适用于方差大、偏差小的数据集;XGBoost是串行集成模型,适用于方差小、偏差大的数据集;这样集成可以减少数据分布震荡造成的模型效果不稳定。

多层集成模型在独立训练几个异质模型后,已经能拟合出较好的结果,因此最后只需采用一个简单、高效的传统加权平均方法获得模型最终结果。

3 实验及分析

本文根据岩石物理测井响应机理和解释方法,对曲线稳定性较好的井段进行数据录入,优选出声波时差AC,补偿中子CNL,感应电导率COND,体积密度DEN,自然伽马GR,2.5 m底部梯度电阻率R25,4 m底部梯度电阻率R4,高分辨率深感应HRID,高分辨率中感应HRIM,电阻率RT,冲洗带电阻率RXO,自然电位SP等16种对岩性特征响应敏感的岩石物探参数进行岩性识别。

本文的岩性样本包括4类,分别是砂岩、泥岩、断层核部和诱导裂缝带,样本数共计4 181个(由于篇幅限制,这里只显示部分岩性类别数据),如图2所示,将其划分为训练集和测试集,其中训练集样本

	深度	AC	CNL	COND	DEN	GR	R25	R4	RFOC	HRID	HRIM	RILD	RILM	RL	RN	RT	RXO	SP	岩性
2	2655.3	93.85	25.933	438.113	2.469	65.243	5.825	5.79	1.831	2.588	2.61	2.307	2.132	2.131	1.616	2.442	2.47	67.112	砂岩
3	2655.325	93.942	25.849	439.464	2.47	65.405	5.779	5.765	1.837	2.551	2.575	2.301	2.121	2.067	1.516	2.424	2.453	67.108	砂岩
4	2655.35	94.485	26.069	440.812	2.471	66.248	5.82	5.832	1.858	2.477	2.504	2.297	2.1	1.82	1.122	2.379	2.41	67.13	砂岩
5	2655.375	94.351	25.833	442.164	2.472	66.068	5.732	5.761	1.857	2.46	2.487	2.29	2.095	1.849	1.169	2.375	2.406	67.113	砂岩
6	2655.4	94.218	25.596	443.516	2.474	65.889	5.643	5.69	1.855	2.442	2.47	2.283	2.09	1.877	1.215	2.371	2.402	67.096	砂岩
7	2682.9	113.027	33.548	234.484	2.437	77.54	12.336	7.842	4.439	5.21	5.099	4.412	4.357	2.241	2.13	5.051	4.93	64.636	泥岩
8	2682.925	112.999	33.452	236.901	2.439	77.892	12.291	7.936	4.409	5.17	5.057	4.375	4.316	2.207	2.117	5.008	4.888	64.671	泥岩
9	2682.95	112.997	33.383	238.085	2.439	78.285	12.263	8.095	4.373	5.136	5.013	4.349	4.282	2.111	2.07	4.974	4.852	64.72	泥岩
10	2682.975	112.957	33.275	241.117	2.441	78.616	12.209	8.156	4.346	5.093	4.971	4.307	4.238	2.108	2.073	4.927	4.808	64.748	泥岩
11	2683	112.916	33.166	244.149	2.443	78.947	12.155	8.218	4.319	5.051	4.93	4.265	4.195	2.104	2.077	4.881	4.764	64.776	泥岩
12	2661.05	77.056	18.558	260.765	2.483	45.836	8.661	5.541	3.547	5.148	5.124	4.043	4.255	1.735	1.77	4.863	4.789	66.151	核部
13	2661.075	77.103	18.594	263.956	2.485	46.164	8.692	5.517	3.57	5.068	5.039	3.993	4.194	1.812	1.843	4.767	4.701	66.215	核部
14	2661.1	77.15	18.631	267.147	2.487	46.492	8.723	5.492	3.593	4.989	4.954	3.943	4.133	1.89	1.916	4.671	4.613	66.279	核部
15	2667.2	63.511	17.607	179.363	2.519	35.511	9.252	4.226	4.748	8.115	9.158	5.676	6.328	2.932	2.343	7.882	9.099	65.539	诱导裂缝带
16	2667.225	63.357	17.554	180.422	2.52	35.348	9.343	4.223	4.692	8.057	9.143	5.651	6.183	2.837	2.267	7.727	8.9	65.535	诱导裂缝带
17	2667.25	62.744	17.441	181.034	2.522	34.971	9.48	4.216	4.675	8.191	9.496	5.639	5.979	2.85	2.158	7.668	8.906	65.517	诱导裂缝带
18	2667.275	62.82	17.419	182.317	2.522	34.915	9.547	4.215	4.599	8.036	9.298	5.607	5.864	2.702	2.1	7.466	8.605	65.521	诱导裂缝带
19	2667.3	62.896	17.397	183.6	2.521	34.858	9.614	4.214	4.524	7.882	9.099	5.575	5.748	2.554	2.041	7.264	8.304	65.524	诱导裂缝带
20	2667.325	62.972	17.375	184.883	2.52	34.801	9.681	4.213	4.448	7.727	8.9	5.543	5.632	2.406	1.982	7.062	8.003	65.528	诱导裂缝带
21	2667.35	62.624	17.291	185.755	2.52	34.495	9.817	4.211	4.374	7.668	8.906	5.521	5.453	2.087	1.825	6.836	7.638	65.511	诱导裂缝带

图2 岩性类别数据

Fig.2 Lithology category data

3 135个,测试集样本1 046个,具体分布见表1—2所示。

表1 训练样本集岩性类别分布

Table 1 Distribution of lithology categories in training samples

岩性类型	砂岩	泥岩	断层核部	诱导裂缝带	总计
样本数	754	1 584	400	397	3 135
占比/%	24.05	50.53	12.76	12.66	100

表2 测试样本集岩性类别分布

Table 2 Distribution of lithology categories in test samples

岩性类型	砂岩	泥岩	断层核部	诱导裂缝带	总计
样本数	261	540	124	121	1 046
占比/%	24.95	51.63	11.85	11.57	100

对单一学习模型和本文提出的多层集成学习模型分别使用训练样本集进行训练,得到岩性识别模型,然后用训练得到的模型在测试样本集上进行岩性识别实验,验证模型进行岩性识别的准确性。

(1)使用SVM模型进行识别,在测试集上的识别结果见表3所示。

表3 SVM模型测试集结果矩阵表

Table 3 Result matrix of SVM model test set

模型:SVM		识别类型			
	类型	砂岩	泥岩	断层核部	诱导裂缝带
实际类型	砂岩	214	35	6	6
	泥岩	3	532	2	3
	断层核部	39	0	76	9
	诱导裂缝带	32	1	15	73

从表3中可以看出:测试集中砂岩样本共261个,识别正确(即识别为砂岩)214个,识别错误47个(其中识别为泥岩35个,识别为断层核部6个,识别为诱导裂缝带6个)。识别准确率为81.99%。测试集中有泥岩样本共540个,识别正确532个,识别错误8个(其中识别为砂岩3个,识别为断层核部2个,识别为诱导裂缝带3个),准确率为98.52%。测试集中有断层核部样本共124个,识别正确76个,识别错误48个(其中识别为砂岩39个,识别为诱导裂缝带9个),准确率为61.29%。测试集中有诱导裂缝带样本共121个,识别正确73个,识别错误48个(其中识别为砂岩32个,识别为泥岩1个,识别为断层核部15个),准确率为60.33%。

(2)使用GCForest模型,在测试集上的识别结果矩阵见表4所示。

同样可以计算得出:砂岩识别准确率为96.55%,泥岩识别准确率为98.89%,断层核部识别准确率

表4 GCForest模型测试集结果矩阵表

Table 4 Results matrix of GCForest model test set

模型:GCForest		识别类型			
	类型	砂岩	泥岩	断层核部	诱导裂缝带
实际类型	砂岩	252	3	5	1
	泥岩	2	534	2	2
	断层核部	4	0	117	3
	诱导裂缝带	3	1	2	115

为94.35%,诱导裂缝带识别准确率为95.04%。

(3)使用随机森林模型,在测试集上的识别结果矩阵见表5所示。

表5 随机森林模型测试集结果矩阵表

Table 5 Result matrix of random forest model test set

模型:随机森林		识别类型			
	类型	砂岩	泥岩	断层核部	诱导裂缝带
实际类型	砂岩	249	8	3	1
	泥岩	2	533	2	3
	断层核部	7	0	114	3
	诱导裂缝带	5	0	2	114

可以计算得出:砂岩识别准确率为95.40%,泥岩识别准确率为98.70%,断层核部识别准确率为91.94%,诱导裂缝带识别准确率为94.21%。

(4)使用XGBoost模型,在测试集上的识别结果矩阵见表6。

表6 XGBoost模型测试集结果矩阵表

Table 6 Result matrix of XGBoost model test set

模型:XGBoost		识别类型			
	类型	砂岩	泥岩	断层核部	诱导裂缝带
实际类型	砂岩	251	5	3	2
	泥岩	1	535	1	3
	断层核部	6	0	113	5
	诱导裂缝带	2	0	2	117

同样可以计算得出:砂岩识别准确为96.17%,泥岩识别准确率为99.07%,断层核部识别准确率为91.13%,诱导裂缝带识别准确率为96.69%。

(5)使用本文提出的模型,在测试集上的识别结果矩阵见表7所示。

表7 本文多层集成学习模型测试集结果矩阵表

Table 7 Result matrix of multi-layer integrated learning model test set

模型:本文模型		识别类型			
	类型	砂岩	泥岩	断层核部	诱导裂缝带
实际类型	砂岩	252	5	3	1
	泥岩	1	535	0	4
	断层核部	4	0	118	2
	诱导裂缝带	3	0	2	116

可以计算得出:砂岩识别准确为96.55%,泥岩识别准确率为99.07%,断层核部识别准确率为95.16%,诱导裂缝带识别准确率为95.87%。

图3—6是在识别不同的岩性时各模型的表现,图中SVM代表支持向量机模型、Xgb代表XGBoost模型、GCF代表GCForest模型、RF代表随机森林模型、MIEL代表本文提出的多层集成学习模型。其

中,图3是5个模型在砂岩上的识别准确率对比;图4是5个模型在泥岩上的识别准确率对比;图5是5个模型在断层核部上的识别准确率对比;图6是5个模型在诱导裂缝带上的识别准确率对比。

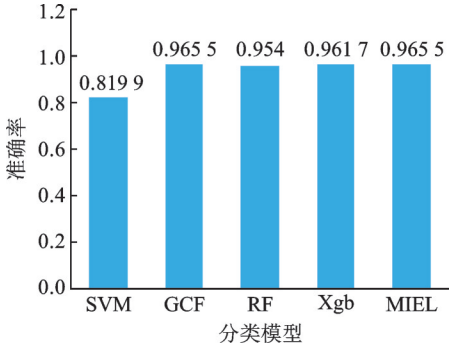

图3 砂岩准确率对比图

Fig.3 Comparison of sandstone accuracy

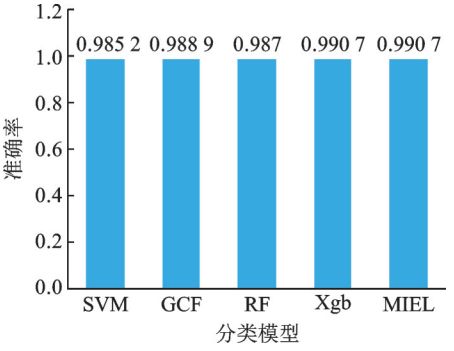

图4 泥岩准确率对比图

Fig.4 Mudstone accuracy comparison chart

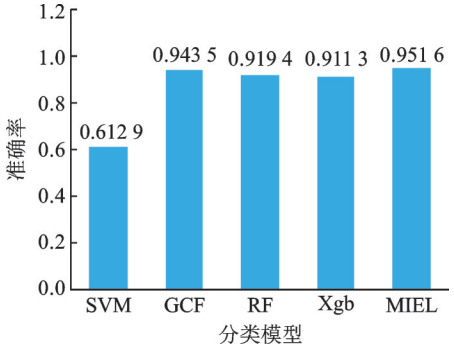

图5 断层核部准确率对比图

Fig.5 Comparison of accuracy of the fault core

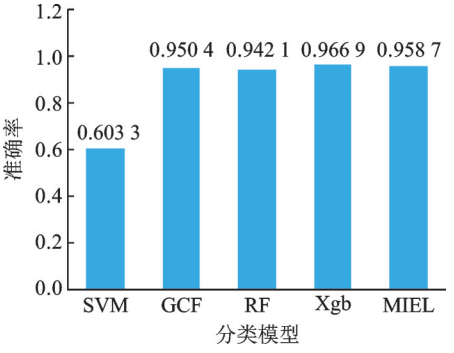

图6 诱导裂缝带准确率对比图

Fig.6 Comparison of accuracy of the induced crack zone

图7是5个模型进行岩性识别的平均准确率对比。从图中可以看出,本文提出的多层集成学习模型在4种岩性类别的识别上都具有较优表现,平均准确率为96.66%,高于SVM的75.53%,GCForest模型的96.21%,随机森林模型的95.06%,XGBoost模型的95.77%。

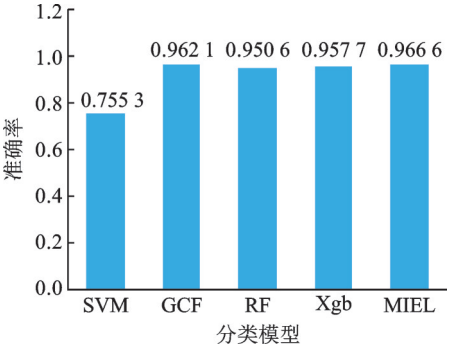

图7 集成模型平均准确率对比图

Fig.7 Comparison of average accuracy of integrated models

4 结束语

本文研究机器学习用于油藏地质分析岩性识别,在分析目前常用单一学习模型和基本集成模型优点和不足的基础上,面向测井数据,利用多个异构的机器学习模型,采用多层集成思想,建立了一个用于岩性识别的多层集成学习模型。该模型具有泛化性能好、适应性强、稳定性和可靠性高以及准确率高等优点。利用该模型进行了包括泥岩、砂岩、断层核部和诱导裂缝带4种岩性类别的岩性识别实验,并与其他模

型的识别结果进行了对比分析。实验分析结果表明,本文的多层集成模型与其他模型相比具有更好的表现,在测试集上识别准确率达到了96.66%,为机器学习技术应用于复杂油藏地质分析提供了有益的探索。

参 考 文 献:

[1] SILVA A A, LIMANETO I A, MISSÁGIA R M, et al. Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information [J].Joumal of Applied Geophysics, 2015, 117: 118-125.

[2] TSANG S, YUN S K, DOBBIE G, et al. Detecting online auction shilling frauds using supervised learning[J]. Expert Systems with Applications an International Journal, 2014, 41(6): 3027-3040.

[3] BEYAN C, FISHER R. Classifying imbalanced data sets using similarity based hierarchical decomposition[J]. Pattern Recognition, 2015, 48(5): 1653-1672.

[4] 赵显令,王贵文,周正龙,等.地球物理测井岩性解释方法综述[J].地球物理学进展,2015,30(3): 1278-1287.
ZHAO Xianling, WANG Guiwen, ZHOU Zhenglong, et al. A review of lithologic interpretation methods for geophysical logging [J]. Progress in Geophysics, 2015, 30(3): 1278-1287.

[5] 范宜仁,黄隆基,代诗华.交会图技术在火山岩岩性与裂缝识别中的应用[J].测井技术,1999,23(1): 53-56.
FAN Yiren, HUANG Longji, DAI Shihua. Application of intersection graph technology in lithology and fracture identification of volcanic rocks [J]. Logging Technology, 1999, 23(1): 53-56.

[6] 钟仪华,李榕.基于主成分分析的最小二乘支持向量机岩性识别方法口[J].测井技术,2009,33(5): 425-429.
ZHONG Yihua, LI Wei. Least squares support vector machine lithology identification method based on principal component analysis[J]. Logging Technology, 2009, 33(5): 425-429.

[7] 朱怡翔,石广仁.火山岩岩性的支持向量机识别[J].石油学报,2013,34(2): 312-322.
ZHU Yixiang, SHI Guangren. Support vector machine identification of volcanic lithology [J]. Journal of Petroleum, 2013, 34 (2): 312-322.

[8] 李洪奇,郭海峰,郭海敏,等.复杂储层测井评价数据挖掘方法研究[J].石油学报,2009,30(4): 542-549.
LI Hongqi, GUO Haifeng, GUO Haimin, et al. Research on data mining method for complex reservoir logging evaluation [J]. Journal of Petroleum, 2009, 30(4): 542-549.

[9] 张国英,王娜娜,张润生,等.基于主成分分析的BP神经网络在岩性识别中的应用[J].北京石油化工学院学报,2008(3): 43-46.
ZHANG Guoying, WANG Nana, ZHANG Runsheng, et al. Application of BP neural network based on principal component analysis in lithology identification[J]. Journal of Beijing Institute of Petrochemical Technology, 2008(3): 43-46.

[10] 梁炉方.基于随机森林和支持向量机的癌症基因数据分析[D]. 济南: 山东大学,2017.
LIANG Lufang. Analysis of cancer gene data based on random forest and support vector machine [D]. Jinan: Shandong University, 2017.

[11] GÉRON A. Hands-on machine learning with scikit-learn and tensorflow[M]. Sebastopol, CA: O′Reilly Media, 2017.

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

[13] 韩启迪,张小桐,申维.基于决策树特征提取的支持向量机在岩性分类中的应用[J].吉林大学学报(地球科学版),2019,49(2): 611-620.
HAN Qidi, ZHANG Xiaotong, SHEN Wei. Application of support vector machine based on decision tree feature extraction in lithology classification[J]. Journal of Jilin University(Earth Science Edition), 2019, 49(2): 611-620.

[14] 牟丹,王祝文,黄玉龙,等.基于SVM测井数据的火山岩岩性识别——以辽河盆地东部坳陷为例[J].地球物理学报,2015,58 (5): 1785-1793.
QI Dan, WANG Zhuwen, HUANG Yulong, et al. Identification of volcanic lithology based on SVM logging data—Taking the eastern depression of the Liaohe Basin as an example [J]. Chinese Journal of Geophysics, 2015, 58(5): 1785-1793.

[15] ZHOU Z H, FENG J. Deep forest: Towards an alternative to deep neural networks[C]//Proceedings of Processing of the Twenty-sixth International Joint Conference on Artifical Intelligence (IJCAI-17). [S.l.]: [s.n.], 2017: 3553-3559.

[16] 马超.半监督随机森林分类算法及其并行化研究[D].重庆:重庆大学,2017.
MA Chao. Semi-supervised random forest classification algorithm and its parallelization research [D]. Chongqing: Chongqing University, 2017.

[17] 江凯,王守东,胡永静,等.基于Boosting Tree算法的测井岩性识别模型[J].测井技术,2018,42(4):395-400.
JIANG Kai, WANG Shoudong, HU Yongjing, et al. Logging identification model based on boosting tree algorithm[J]. Logging Technology, 2018, 42(4): 395-400.

[18] 杨笑,王志章,周子勇,等.基于参数优化AdaBoost算法的酸性火山岩岩性分类[J].石油学报,2019,40(4):457-467.
YANG Xiao, WANG Zhizhang, ZHOU Ziyong, et al. Classification of acidic volcanic rocks based on parameter optimization AdaBoost algorithm[J]. Editorial office of ACTA Petrolei Sinica, 2019, 40(4): 457-467.

[19] 马峥,张春雷,高世臣.主成分分析与模糊识别在岩性识别中的应用[J].岩性油气藏,2017,29(5):127-133.
MA Zheng, ZHANG Chunlei, GAO Shichen. Application of principal component analysis and fuzzy recognition in lithology identification[J]. Lithologic Reservoirs, 2017, 29(5): 127-133.

[20] 范存辉,梁则亮,秦启荣,等.基于测井参数的遗传BP神经网络识别火山岩岩性——以准噶尔盆地西北缘中拐凸起石炭系火山岩为例[J].石油天然气学报,2012,34(1):68-71.
FAN Cunhui, LIANG Zeliang, QIN Qirong, et al. Identification of volcanic lithology based on genetic BP neural network based on logging parameters: A case study of the Carboniferous volcanic rocks in the central auger of the northwestern margin of the Junggar Basin [J]. Journal of Oil and Gas, 2012, 34(1): 68-71.

[21] 刘明军,李恒堂,姜在炳.GA-BP神经网络模型在彬长矿区测井岩性识别中的应用[J].煤田地质与勘探,2011,39(4):8-12.
LIU Mingjun, LI Hengtang, JIANG Zaibing. Application of GA-BP neural network model in logging lithology identification in Binchang mining area [J]. Coalfield Geology and Exploration, 2011, 39(4): 8-12.

[22] 张莹,潘保芝.基于主成分分析的SOM神经网络在火山岩岩性识别中的应用[J].测井技术,2009,33(6):550-554.
ZHANG Ying, PAN Baozhi. Application of SOM neural network based on principal component analysis in lithology identification of volcanic rocks [J]. Logging Technology, 2009, 33(6): 550-554.

[23] 曹莹,苗启广,刘家辰,等.AdaBoost算法研究进展与展望[J].自动化学报,2013,39(6):745-758.
CAO Ying, MIAO Qiguang, LIU Jiachen. et al. Research progress and prospects of AdaBoost algorithm[J]. Journal of Automation, 2013, 39(6): 745-758.

[24] 瞿晓婷,张蕾,冯宏伟,等.面向复杂储层的非均衡测井数据的岩性识别[J].地球物理学进展,2016(5):2128-2132.
QU Xiaoting, ZHANG Lei, FENG Hongwei, et al. Lithology identification of unbalanced logging data for complex reservoirs [J]. Progress in Geophysics, 2016(5): 2128-2132.

[25] 何羽飞,万金彬,刘淼,等.基于测井多参数的复杂储层岩性综合识别[J].测井技术,2015(1):48-51.
HE Yufei, WAN Jinbin, LIU Miao, et al. Comprehensive identification of complex reservoir lithology based on multi-parameter logging[J]. Logging Technology, 2015(1): 48-51.

作者简介:

段友祥(1964-),男,教授,研究方向:油气领域智能信息处理,E-mail: yxduan@upc.edu.cn。

赵云山(1993-),男,硕士研究生,研究方向:机器学习及应用。

马存飞(1987-),男,博士,研究方向:非常规油气地质与油藏描述。

姜文煊(1995-),女,硕士研究生,研究方向:机器学习及应用。

(编辑:刘彦东)