文章编号：2095-2163(2020)03-0091-04 中图分类号：TP183 文献标志码：A

基于集成学习的测井岩性识别模型的构建

邹 琪，何月顺，杨 希，章 权
（东华理工大学 信息工程学院，南昌 330013）

摘 要： 岩性识别不论是在地层评价，还是在油藏描述、钻井监控等地质勘察工作中有着重要的作用。针对传统基于测井响应方程的测井岩性识别方法效率低并且信息冗余等问题，本文提出了一种基于 Stacking 集成学习的测井岩性识别方法。该方法建立了一种以朴素贝叶斯、随机森林、支持向量机三种模型作为初级训练器，逻辑回归模型作为次级训练器的融合模型。该融合模型提高了测井岩性识别效率，实现了测井数据自动化处理，提高了地质勘察人员的工作效率。文中以鄂尔多斯盆地的钻孔测井数据为训练样本，通过与其他机器学习模型的结果比较，该模型的预测结果与实际结果相比具有较好的一致性，识别率优于其他学习模型。

关键词： 岩性识别；集成学习；融合模型；Stacking

Construction of lithology identification model by well logging based on ensemble learning

ZOU Qi, HE Yueshun, YANG Xi, ZHANG Quan
(School of Information Engineering, East China University of Technology, Nanchang 330013, China)

【Abstract】 Lithology identification plays an important role in geological surveys such as stratigraphic evaluation, reservoir description and drilling monitoring. Aiming at the low efficiency and information redundancy of traditional logging lithology identification methods based on logging response equation, this paper proposes a logging lithology identification method based on Stacking in ensemble learning.This method establishes a fusion model with Naive Bayes, Random Forest, Support Vector Machine and Logistic Regression.Three machine learning models, Naive Bayes, Random Forest and Support Vector Machine are used as a primary training device to separately train the data, and then Logistic Regression model is used as a secondary learning device to predict.The fusion model improves the logging lithology identification efficiency, realizes automatic processing of logging data, and improves the working efficiency of geological survey personnel. In this paper, the borehole logging data of the Ordos Basin is used as the training sample.Compared with the results of other machine learning models, the prediction results of the model are better than the actual results, and the recognition rate is better than other learning models.

【Key words】 lithology identification; ensemble learning; fusion model; Stacking

0 引 言

地层岩性是指岩石颜色、成分、结构、特殊矿物等特征的总和，岩性识别是通过一些特定的方法来判定和区别岩性的过程。目前，测井岩性识别方法主要可以分为基于测井曲线响应特征的定性解释方法[1]、基于测井响应方程的定量解释方法[2]、图版法[3]和基于机器学习的智能化方法[4-8]。定性解释方法和图版法的实施主要依赖于人员的实践经验和剖面的复杂度，人为因素影响较大；定量解释方法相比于定性解释方法可靠性更高，但其受限于地层矿物成分数量，对复杂岩性储层的适用性较差[2]；基于机器学习的岩石识别方法主要有聚类分析法、支持向量机方法和决策树方法。聚类分析法对训练样本的要求为趋于无穷大，才会取得良好的效果，所以对于小样本来说，该方法在识别中并不实用。支持向量机方法能较为准确地识别过渡岩性，且在实际岩性识别中有较好的效果，决策树方法是一种符号学习方法，易于直观理解，但上述机器学习方法都是单一学习方法，不能对错误样本进行再学习。

本文提出一种基于 Stacking 集成学习方法的测井岩性识别模型，该模型融合随机森林、支持向量机、朴素贝叶斯三种机器学习方法，并对鄂尔多斯盆地地层进行岩性识别，结果表明，该模型在识别准确率上与其他模型相比有明显提升。

基金项目： 国家自然科学基金(41872243)。

作者简介： 邹 琪(1996-)，男，硕士研究生，主要研究方向：数据挖掘、机器学习；何月顺(1971-)，男，博士，教授，主要研究方向：无线传感网络与物联网技术、大数据分析与智能信息处理；杨 希(1993-)，女，硕士研究生，主要研究方向：大数据分析、机器学习；章 权(1994-)，男，硕士研究生，主要研究方向：机器学习、进化算法。

收稿日期： 2019-11-10

1 岩性识别现状

岩性识别技术自 20 世纪 90 年代引入国内,其方法包括重磁、地震、遥感、测井、地球化学、电磁、手标本及薄片分析。岩石物性是指岩石三相组成部分的相对比例关系不同所表现的物理状态,同时也代表着岩石的力学、热学、电学、声学、放射学等特性参数和物理量。区分和识别岩性的主要步骤就在于岩石物性的研究,密度、电导率、磁化率、波阻抗等在地质勘察工作中是经常用的岩石物性。测井资料往往存在着大量的地层岩性信息,这些信息是岩性识别的基本信息[9]。因此,在众多岩性识别方法中,测井岩性识别方法是目前比较成熟的一种方法。

刘昊等人[10]针对实际储层非均匀性,利用 K-means 聚类算法和 DBSACN 聚类算法对某盆地具有十维特征量的测井数据建立了岩性识别模型,提高了分类识别的准确度,识别效果更加接近储层的真实特性。陈华等人[11]采用最小二乘支持向量机对孔隙度、渗透率和饱和度进行了预测,取得了良好的预测效果。胡剑策[12]将最小二乘支持向量机和主成分分析方法引入油气储层的识别和预测,提出了一种基于主成分分析的最小二乘支持向量机的预测模型,该模型的性能优于一些其他模型。温志平等人[13]针对传统神经网络岩性识别模型存在收敛速度慢、难以选择合适的网络拓扑和学习参数问题,提出一种采用递阶遗传染色体编码方式并将具有非线性的 Sigmoid 函数引入到遗传操作算子的自适应递阶遗传优化神经网络模型,从而减少了遗传算法陷入早熟的几率。江凯等人[14]以录井资料和测井资料为基础,优选自然伽马、自然电位、冲洗带电阻率、侵入带电阻率、原状地层电阻率、密度、补偿中子、声波时差 8 个测井属性,使用 Boosting Tree 算法建立了岩性识别模型,并使用该模型对玛北油田岩石进行识别,正确率优于决策树、支持向量机等传统机器学习方法。杨笑等人[15]为提高长岭气田火山岩岩性识别的准确率,采用决策树、支持向量机、逻辑回归、AdaBoost - 决策树、AdaBoost - 支持向量机和 AdaBoost-逻辑回归 6 种算法对酸性火山岩岩性识别进行分类和识别,通过交叉验证进行参数优化及模型评价,对比不同算法发现 AdaBoost-决策树算法的准确率最高。

目前基于集成学习的方法在岩性识别上的应用并不广泛,大部分研究学者还是在单一机器学习模型之上进行研究和改进的。集成学习中的 Stacking 思想首先训练出多个不同的模型,然后再以之前训练的各个模型的输出作为输入来新训练一个新的模型,换句话说,Stacking 算法根据模型的输出是允许改其他分类器的参数甚至结构的。

2 集成学习

集成学习是通过多个基分类器组合来完成学习任务并提高准确率的一种技术[16-17]。通过集成学习,集成学习器能获得比单一学习器更优越的泛化性能,其原理是使用一定量的样本来训练多个弱学习器,再采用“少数服从多数”的投票法来选择分类结果[18]。这样即使一些学习器有错误时,也能通过多数学习器来纠正。集成学习一般可以分为用于减少方差的 Bagging、用于减少偏差的 Boosting 和用于提升预测结果的 Stacking 三大类。

本文所采用的是用于提升预测结果的 Stacking 方法,其通过一个元分类器或元回归器来整合多个分类模型或回归模型。Stacking 的工作流程如下:

(1)将训练样本分为 N 份训练集和 1 份测试集来进行 N 折交叉验证。

(2)用初级分类器对 $N-1$ 份训练集进行训练,训练之后的模型再对剩下的 1 份验证集进行预测生成数据集 $a_i(i <= N)$,此模型同时对测试集进行预测产生数据集 $b_j(j <= N)$。

(3)重复步骤(2) N 次,产生 $a_1,a_2,\ldots,a_N$ 和 $b_1,b_2,\ldots,b_N$,将 $a_1,a_2,\ldots,a_N$ 拼凑起来,记为 $A_i(i \leqslant N)$,并对 $b_1,b_2,\ldots,b_N$ 这部分数据相加取平均值,记为 $B_j(j \leqslant N)$。

(4)对每一个初级分类器进行步骤(2)和步骤(3)操作,得到新的训练集 $A_1,A_2,\ldots,A_N$ 和新的测试集 $B_1,B_2,\ldots,B_N$。

(5)让次级分类器对从步骤(4)中得到的训练集和测试集分别进行训练和预测,得到最后的预测结果。

3 模型的构建

本文选取了随机森林、支持向量机、朴素贝叶斯三种机器学习模型为初级训练器,以逻辑回归模型为次级训练器来进行样本的学习训练。文中采用了 3 折交叉验证方法,将训练集等分为 3 份,其中 2 份用来训练学习,剩下 1 份进行验证。文中使用随机森林、支持向量机、朴素贝叶斯模型依次对训练集中的样本进行 3 折交叉验证训练后对测试集进预测,得出新的训练集和测试集,然后使用逻辑回归模型对新的训练集学习训练,最后将训练后的模型对测

试集进行预测，具体流程如图 1 所示。

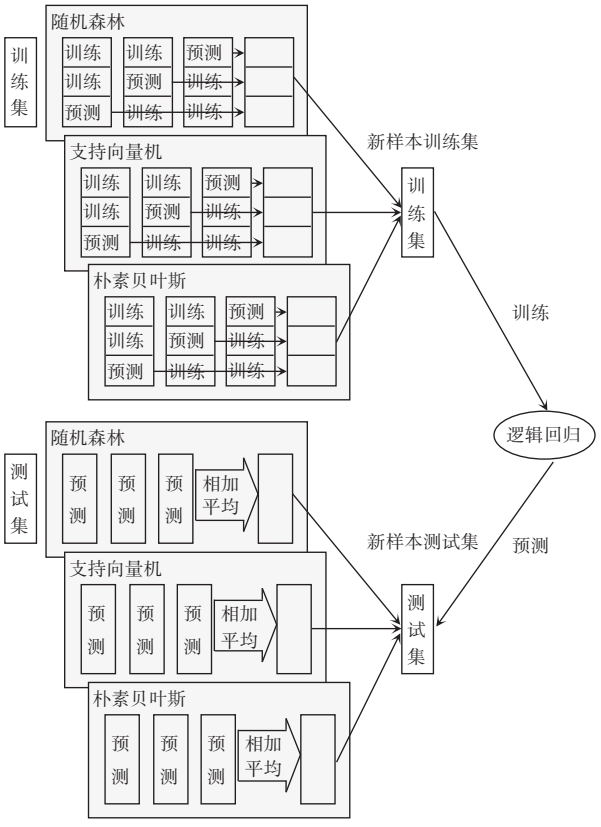

图 1　集成学习流程

Fig. 1　Ensemble learning process

4　应用实例

4.1　样本构建

本文选取鄂尔多斯盆地的钻孔测井数据来验证本文提出的融合模型的准确率。收集盆地中的 1 729个样本，其中包含泥岩、泥质粉砂岩、砂质泥岩三大类，这三类的样本比例分别为 59%，11.6%，29.4%，见表 1。本文提取了自然伽马（GR）、阵列感应电阻率（AT10、AT20、AT30、AT60、AT90）、纵横波方式单极纵波时差（DT4P）、光电吸收截面指数（PEFL）、岩性密度（RHOM）和自然电位（SP）这几条测井曲线作为分类参数，并将全部样本中的1 210个样本作为训练集用于训练岩性识别模型，519 个样本作为测试集来检验融合模型的岩性识别效果，同时将其结果与使用朴素贝叶斯、随机森林、支持向量机的识别结果进行对比。

表 1　训练样本集岩性类别分布

Tab. 1　The lithology category distribution of training sample set

	岩性			总计
	泥岩	泥质粉砂岩	砂质泥岩	
样本数	1 019	202	508	1 729
占比/%	59.0	11.6	29.4	100

4.2　结果分析

样本的每个特征属性来源于不同的测量方法，量纲有所不同，因此，本文采用 Sklearn 库中的 StandardScaler 类来进行数据的标准化和归一化操作。本文实验是在 Python3.7 下使用 Sklearn 和 Pandas 等库进行实现的，IDE 为 PyCharm professional edition。硬件环境为 Intel（R）Core（TM）i5-3230M CPU@2.60 GHz、8 GBRAM 设备。本文采用 3 折交叉验证方法依次对朴素贝叶斯、随机森林、支持向量机三种岩性识别模型进行训练，并对本文提出的融合模型训练，不同模型的交叉验证准确率见表 2。对训练后的模型在测试集上进行预测，不同模型的岩性预测结果见表 3，最后对不同模型进行评估检验，结果见表 4。

表 2　交叉验证准确率

Tab. 2　Cross-validation accuracy

模型	第一次验证	第二次验证	第三次验证	平均值
朴素贝叶斯	0.617	0.615	0.572	0.602
随机森林	0.788	0.814	0.776	0.793
支持向量机	0.837	0.814	0.823	0.825
融合模型	0.842	0.868	0.838	0.850

表 3　不同模型岩性预测结果

Tab. 3　Different model lithology prediction results

岩性	样本数	准确率			
		朴素贝叶斯	随机森林	支持向量机	融合模型
泥岩	307	0.706	0.844	0.897	0.892
泥质粉砂岩	62	0.491	0.636	0.690	0.907
砂质泥岩	150	0.447	0.866	0.801	0.865
合计/准确率	519	0.609	0.823	0.842	0.886

表 4　不同模型评价指标

Tab. 4　Different model evaluation indicators

模型	评价指标	泥岩	泥质粉砂岩	砂质泥岩	平均值
朴素贝叶斯	精确率	0.706	0.491	0.447	0.548
	召回率	0.726	0.419	0.447	0.531
	F_1 值	0.716	0.452	0.447	0.538
随机森林	精确率	0.844	0.636	0.866	0.782
	召回率	0.919	0.677	0.687	0.761
	F_1 值	0.880	0.656	0.766	0.767
支持向量机	精确率	0.897	0.690	0.801	0.796
	召回率	0.883	0.790	0.780	0.818
	F_1 值	0.890	0.737	0.791	0.806
融合模型	精确率	0.892	0.907	0.865	0.888
	召回率	0.941	0.790	0.813	0.848
	F_1 值	0.916	0.845	0.838	0.866

表 2 给出了不同模型在验证集上的准确率,可以看出本文提出的融合模型在验证集上的准确率基本稳定在 0.85 左右,准确率高于其他三种模型,证明该模型具有较强的稳定性。从表 3 中可以看出,本文提出的融合模型在泥岩识别的准确率到达了 0.892,高于朴素贝叶斯的 0.706 和随机森林的 0.844,略低于支持向量机的 0.897。在泥质粉砂岩这种小样本的识别上,融合模型的准确率达到了 0.907,精确率远高于其他 3 种模型。对于砂质泥岩的识别准确率来说,融合模型的0.865高于朴素贝叶斯的0.447和支持向量机的 0.801,稍微低于随机森林的0.866。对于不同类别的岩性来说,融合模型的准确率基本维持在0.886左右,表明了融合模型有着较好的泛化能力,其准确率更是高于随机森林、支持向量机、朴素贝叶斯三种模型的准确率。

表 4 比较了不同模型的评价指标,本文提出的融合模型与随机森林、支持向量机、朴素贝叶斯比较得出泥岩、泥质粉砂岩、砂质泥岩最佳分类 F_1 值分别为 0.916、0.845、0.838,这些最佳 F_1 值均来自融合模型,并且可以看出融合模型的平均 F_1 值高于朴素贝叶斯 30%左右,高于随机森林 10%左右,高于支持向量机 6%,分类效果显著提升。

5 结束语

本文主要研究了以地质大数据为背景下的基于集成学习中 Stacking 思想的测井岩性识别方法。首先介绍了岩性识别的相关方法,其中有传统的基于矿物物性的测井岩性识别方法,也有基于机器学习的一些识别方法,比如支持向量机、神经网络等。接着详细叙述了随机森林、支持向量机、朴素贝叶斯和集成学习等相关机器学习知识,并提出了一种基于集成学习中 Stacking 思想的融合模型,该模型融合了随机森林、支持向量机和朴素贝叶斯三种机器学习模型。最后通过实验,将本文提出的融合模型与随机森林、支持向量机和朴素贝叶斯三种机器学习模型的岩性识别结果作对比,结果表明融合模型的岩性识别准确率高于其他三种模型,并且有着较强的泛化能力和稳定性。

本文提出的模型虽然在岩性识别率上优于其他三种基本机器学习模型,但没有去尝试融合多种优化过后的机器学习算法,这也为其他研究者提供了一个参考。

参考文献

[1] 叶涛,韦阿娟,邓辉,等. 基于常规测井资料的火山岩岩性识别方法研究一以渤海海域中生界为例[J]. 地球物理学进展,2017,32(4):1842.

[2] 洪有密. 测井原理与综合解释[M]. 北京:中国石油大学出版社,2008.

[3] 黄布宙,潘保芝. 松辽盆地北部深层火成岩测井响应特征及岩性划分[J]. 石油物探,2001,40(3):42.

[4] SEBTOSHEIKH M A, MOTAFAKKERFARD R, RIAHI M A,et al.Support vector machine method, a new technique for lithology prediction in an Iranian heterogeneous carbonate reservoir using petrophysical well logs[J].Carbonates and Evaporites, 2015,30(1):59.

[5]石广仁. 支持向量机在多地质因素分析中的应用[J]. 石油学报,2008,29(2):195.

[6] KONATÉ A A,PAN Heping,FANG Sinan,et al. Capability of self-organizing map neural network in geophysical log data classification: Case study from the CCSD-MH[J]. Journal of Applied Geophysics,2015,118:37.

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

[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.

[9] 付光明,严加永,张昆,等. 岩性识别技术现状与进展[J]. 地球物理学进展,2017,32(1):26.

[10]刘昊,朱丹丹,陈冬,等. 基于聚类算法的岩性预分类方法研究[C]//2018 IPPTC 国际石油石化技术会议论文集. 北京:西安华线网络信息服务有限公司,2018:387.

[11]陈华,邓少贵,范宜仁. 基于 LS-SVM 的测井物性参数的预测方法[J]. 计算机工程与应用,2007,43(23):208.

[12]胡剑策. 基于 PCA 的 LS-SVM 预测模型应用[J]. 计算机系统应用,2012,21(6):167.

[13]温志平,方江雄,刘军,等. 自适应递阶遗传神经网络测井岩性识别方法研究[J]. 东华理工大学学报(自然科学版),2017,40(4):368.

[14]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.

[15]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.

[16]杨草原,刘大有,杨博,等. 聚类集成方法研究[J]. 计算机科学,2011,38(2):166.

[17]张莉婷. 基于集成学习的工业产品质量控制方法研究[D]. 广州:华南理工大学,2018.

[18] XUE Di, LI Jingmei, WU Weifei, et al. Homology analysis of malware based on ensemble learning and multifeatures[J]. PloS one,2019,14(8):e0211373.