中国煤炭行业知识服务平台www.chinacaj.net

移动扫码阅读

李泽辰,杜文凤,胡进奎,等.基于测井参数的页岩有机碳含量支持向量机预测[J].煤炭科学技术,2019,47(6):199-204.doi:10.13199/j.cnki.cst.2019.06.030

LI Zechen,DU Wenfeng,HU Jinkui,*et al*.Prediction of shale organic carbon content support vector machine based on logging parameters[J].Coal Science and Technology,2019,47(6):199-204.doi:10.13199/j.cnki.cst.2019.06.030

基于测井参数的页岩有机碳含量支持向量机预测

李泽辰[1,2],杜文凤[1],胡进奎[1,2],李 冬[1,2]

(1.中国矿业大学(北京) 煤炭资源与安全开采国家重点实验室,北京 100083;2.中国矿业大学(北京) 地球科学与测绘工程学院,北京 100083)

摘 要:为了解决传统的有机碳含量 TOC 测量方法成本高和无法获得 TOC 含量连续分布的问题,提出了一种 TOC 含量的统计预测方法。由于地层的岩性的不同,TOC 含量的差异非常大,因此,首先对原始的测井数据聚类,通过聚类的方法将不同岩性的地层区分开,对不同的地层分别建立 TOC 含量的预测模型,再通过聚类的方法提高了各测井参数和 TOC 含量的相关性,这不仅提高了模型的准确性,而且使得模型更有说服力;然后通过粒子群算法优化 SVM 模型参数,避免了因人工选择参数带来的模型不稳定的问题,依此建立测井参数优选的 SVM-RFE 模型,对每一类分别进行测井参数筛选,有效的规避了各测井参数之间的信息冗余和不相关参数带来的模型性能降低和训练时间增加的问题;最后利用优选后的测井数据和 SOM 的分类结果,对不同的地层岩性分别建立 SVR 模型进行预测。结果表明:通过与其他 TOC 含量预测模型对比,SOM-SVR 模型更加稳定,更有说服力,预测误差小,平均相对误差约 6%,平均绝对误差不超过 0.2。由此,可以通过 SOM 算法对不同岩性的地层进行聚类之后再建立 TOC 含量的预测模型,更有利于提高模型的精度。

关键词:总有机碳含量;测井;SOM 聚类;粒子群算法;递归特征消除算法;支持向量机算法

中图分类号:P631.815 **文献标志码**:A **文章编号**:0253-2336(2019)06-0199-06

Prediction of shale organic carbon content support vector machine based on logging parameters

LI Zechen[1,2],DU Wenfeng[1],HU Jinkui[1,2],LI Dong[1,2]

(1. *State Key Lab of Coal Resources and Safe Mining*, *China University of Mining and Technology(Beijing)*,*Beijing* 100083,*China*;

2. *College of Geoscience and Surveying Engineering*, *China University of Mining and Technology(Beijing)*,*Beijing* 100083,*China*)

Abstract:In order to solve the problem that the traditional TOC (Total Organic Carbon) content measurement method is costly and unable to obtain continuous distribution of TOC content, a statistical prediction method of TOC content is proposed.Due to the difference in lithology of the stratum, the difference in TOC content is very large.Therefore, the original log data was first clustered, and the strata of different lithologies were separated by clustering, and predictive models of TOC content were established for different strata.The correlation between the logging parameters and the TOC content was improved by the clustering method, which not only improved the accuracy of the model, but also made the model more convincing.Then the particle swarm optimization algorithm optimized the SVM model parameters, avoiding the manual parameter selection.The model was unstable, and then the SVM-RFE model with well logging parameters was established.The logging parameters were screened for each type, effectively avoiding the information redundancy and irrelevant parameters between the logging parameters. The performance of the model was reduced and the training time was increased.Finally, the SVR model was established for different formation lithology by using the optimized logging data and SOM classification results.Compared with other TOC content prediction models, the results show that the SOM-SVR model is more stable, more convincing, and the prediction error is small, the average relative error is about 6%, and the average absolute error is less than 0.2.It can be concluded that the SOM algorithm is used to cluster the different lithology strata and then establish the TOC content prediction model, which is more conducive to improve the accuracy of the model.

Key words:total organic carbon content; logging; SOM clustering; particle swarm optimization; recursive feature elimination; support vector machine algorithm

收稿日期:2018-09-01;**责任编辑**:曾康生

基金项目:国家科技重大专项资助项目(2016ZX05066-001)

作者简介:李泽辰(1994—),男,山东德州人,硕士研究生。E-mail:2690145249@qq.com

通信作者:杜文凤(1963—),女,吉林扶余人,教授,博士生导师,博士。E-mail:duwf66@126.com

中国煤炭行业知识服务平台www.chinacaj.net

0 引　　言

总有机碳含量(TOC)是烃源岩地球化学评价中的重要参数之一。该参数不仅被用来研究烃类化合物的化学性质,在研究烃源岩的分布中也起着至关重要的作用。在一定深度处,TOC 含量较多表明可能有烃源岩存在,TOC 含量较少表明烃源岩存在的可能性不大。岩石热解是广泛用于生烃潜力评价的标准方法之一,这种方法可以给出岩石中 TOC 含量的精确值。然而,通过地球化学分析得到 TOC 含量通常是昂贵的且耗时长,而且不能获得 TOC 含量在纵向上连续的分布值,因此,开发一种快速低成本的方法来获得 TOC 含量的连续分布值显得尤为重要。

在这方面,杜文凤等[1]以实测 TOC 含量数据为基础,分析了煤系烃源岩中煤岩,炭质泥岩和泥岩有机碳含量的测井响应特征,发现电阻率,声波时差,自然伽马,密度和中子孔隙度等测井参数与有机碳含量相关性较高,并基于此建立了 TOC 含量预测模型;闫建平等[2]通过深入剖析烃源岩石物理特性方法,对以后研究与工作中建立有效,易操作的测井解释模型并精确地提取烃源岩层段的各种有机地化参数奠定了基础;李延钧等[3]依据页岩在不同测井曲线上的测井响应特征,利用电阻率,声波时差,自然伽马能谱等测井信息,采用 $\Delta \lg R$ 法以及自然伽马能谱法回归计算页岩有机碳含量;王濡岳等[4]从矿物组分、地球化学参数、物性参数、含气性和可压裂性评价等方面全面深入论述了页岩气储层测井评价方法及评价参数的计算,认为矿物组分、地球化学参数、物性参数、含气性和可压裂性是页岩气储层测井评价的重点,并结合实际资料改进了 TOC 和脆性指数计算公式。

由于页岩气储层总有机碳含量受多种地质因素影响,它们之间存在着非常复杂的非线性关系,传统单参数或多参数的方法很难表达各测井参数与 TOC 含量之间内在的复杂的非线性关系。孟召平等[5-6]提出的 BP 神经网络的方法虽然可以解决测井参数和 TOC 含量之间复杂的线性关系的问题,但是由于测井数据量大,不同的数据之间差别较大,这就会导致相同 TOC 含量的地层但其测井曲线差别较大,这种差别会引起神经元响应的不同,神经网络的模型因此而变得不稳定,从而使预测误差增大。

针对上述各种问题,笔者给出了一种 TOC 含量的预测模型。首先根据页岩原始测井数据将具有相似特征的页岩进行归类,降低页岩测井数据差别较大对 TOC 含量预测精度的影响;在此基础上,对每一类页岩通过支持向量机的方法进行测井参数优选,获得 TOC 含量预测的优势测井曲线;最后对每一类分别构建神经网络模型进行预测,从而提高模型的稳定性,降低预测误差,提高模型性能。

1 基于测井参数的页岩分类

1.1 聚类目的

为了降低页岩测井数据差别较大对 TOC 含量预测精度的影响;采用自组织特征映射网络(简称 SOM)将测井曲线具有相似特征的页岩进行归类。通过聚类把具有相似响应的测井数据划分为一类,把具有不同测井响应的数据划分为不同类,有效规避了测井数据之间的差异过大,带来的模型不稳定,预测误差大的问题[7-8]。另一方面,通过聚类提高各测井参数与 TOC 含量之间的相关性,进一步提高预测的准确性。

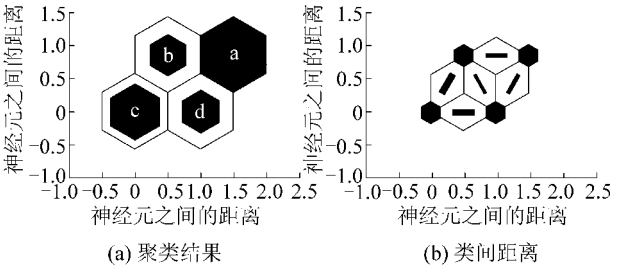
1.2 SOM 原理

自组织特征映射网络(Self-Organizing Feature Map,简称 SOM)是由 T.Kohonen[9-10]于 1981 年首次提出的,是一种无监督聚类方法,它能将输入模式在输出层映射成一维或二维离散图形,识别变量特征并自动聚类。SOM 的基本原理是:当某类模式输入时,输出层某节点得到最大刺激而获胜,获胜节点周围的节点因侧向作用也受到刺激。这时网络进行一次学习操作,获胜节点及周围节点的连接权值向量朝输入模式的方向做相应的修正。当输入模式类别发生变化时,二维平面上的获胜节点也从原来节点转移到其他节点。这样,网络通过自组织方式用大量样本数据来调整其连接权值,最后使得网络输出层特征图能够反映样本数据的分布情况。

1.3 聚类结果

根据 SOM 的原理通过 Matlab 2017a 编程实现,对 150 组测井数据进行了 SOM 聚类,聚类结果如图 1 所示。图 1a 表示将 150 组测井数据分为 a、b、c、d 类,a 类有 68 组数据,b 类有 21 组数据,c 类有 23 组数据,d 类有 38 组数据;图 1b 表示神经元之间距离,每个神经元都代表一个分类,由于这种类间差距很难用数字表示,因此,用相邻神经元之间连接线的粗细表示类间差距的大小,相邻神经元之间连接线越粗表示神经元之间的距离越大,也就是说类间差距越大,从图 1b 可以看出相邻神经元之间的连接线都较粗,说明类间差异性很大,表明这些类之间的测井数据存在较大差异,通过分类使得类间数据足够的不相似,而每一类足够的相似,说明分类结果较好。

中国煤炭行业知识服务平台www.chinacaj.net



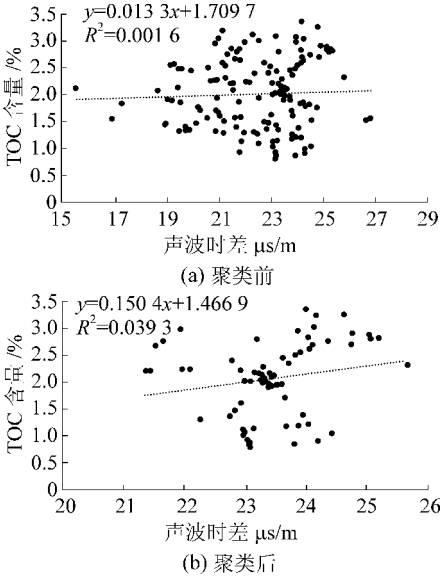
(a) 聚类结果　　(b) 类间距离

图 1　SOM 聚类结果

Fig.1　SOM clustering results

1.4　聚类前后对比

本次聚类的目的是为了把具有相似性的测井数据归为一类。通过图 1b 可以看出,类间连接线都较粗,表明类间差距非常大,说明测井数据之间存在非常大的差异,通过聚类成功把这种差异缩小,使得同类数据之间更具相似性,对体提高模型的稳定性具有极大的帮助。另一方面通过图 2 聚类前后对比,发现聚类后测井参数和 TOC 含量之间的相关性得到了显著提高,从而可以进一步提高预测的准确性。从这 2 方面可以得出本次研究聚类方法的可行性。



(a) 聚类前

(b) 聚类后

图 2　SOM 聚类前后对比

Fig.2　Comparison before and after SOM clustering

2　预测方法及原理

2.1　支持向量机原理

支持向量机(Support Vector Machine,SVM)是由 Vapnik[11] 等于 20 世纪 90 年代初提出的一种新型机器学习算法。其原理是把在样本空间中线性不可分的样本通过核函数映射到特征空间,在特征空间中使得原来线性不可分的样本变得线性可分,从而实现分类的目的。定义最优分类超平面的方程 $f(x)$ 为

$$f(x) = \boldsymbol{W}^{\mathrm{T}}\varphi(x) + \boldsymbol{b} \tag{1}$$

式中:$\boldsymbol{W}$ 为权值系数矩阵;φ 为归一化函数;x 为样本数据;$\boldsymbol{b}$ 为系数矩阵。

支持向量回归机(SVR)不是把样本分为 2 类,而是使得最优分类平面两侧的样本到平面的总距离最小,从而实现拟合。定义误差可接受范围为 ε 、某 i 个松弛因子为 ξ_i 、惩罚系数为 C 、n 为样本个数,于是问题就转化为求目标函数的最小值,目标函数为

$$C \parallel W \parallel^2/2 + C\sum_{i=1}^{n}\xi_i \tag{2}$$

本次研究就是根据支持向量回归机的原理,通过各测井参数与 TOC 含量之间复杂的非线性关系进行 TOC 含量预测。

2.2　支持向量机参数优化

参数的优化对提高模型的性能至关重要。根据前人[12-13] 研究经验,支持向量机核函数选择为高斯核函数,因此,支持向量机回归模型的拟合精度主要取决于不敏感损失系数 ε 、惩罚系数 C 、宽度系数 σ 等 3 个参数的选取。本次研究采用粒子群算法进行参数优化。

粒子群算法(Particle Swarm optimization,PSO)是由 Kenned 和 Ebermart 于 1995 年根据鸟群和鱼群的社会行为提出的一种群智能优化算法粒子群算法是基于群体与适应度的。粒子群的每个粒子的位置都代表问题的一个潜在最优解,每个粒子具有位置和速度两个特征,粒子位置坐标对应的目标函数值作为该粒子的适应度值,算法通过适应度来衡量粒子位置的优劣。算法首先随机初始化一群粒子,然后通过迭代找到最优解。在每次迭代中,粒子通过跟踪两个极值来更新自己的位置,一个是粒子本身所找到的最有位置,即个体极值 p ;另一个是整个粒子群目前找到的最优解,即全局极值 g 。粒子找到上述两个极值之后更新自己的位置和速度。重复以上过程直到算法收敛。如图 3 所示,此时的全局最优粒子就是参数的最优解,把粒子群算法得出的最优解加到 SVR 模型中,建立预测模型。

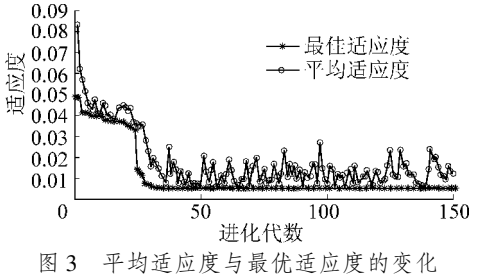


图 3　平均适应度与最优适应度的变化

Fig.3　Average fitness and change of optimal fitness

中国煤炭行业知识服务平台www.chinacaj.net

3　建立 TOC 含量预测模型

3.1　测井数据预处理

本次研究数据来源于鲁西南含煤区页岩气资源潜力调查的测井实测记录和实验分析数据。本次研究测井记录包括:自然电位(SP)、井径(CAL)、自然伽马(GR)、钍(TH)、钾(K)、铀(U)、声波时差(AC)、深浅三侧向电阻率(LLD、LLS)、录井 C_1 含量(C_1)、孔隙度(POR)、渗透率(PERM)。数据的预处理主要是剔除异常值和数据归一化。

3.2　测井参数优选

为了去除测井数据之间的相关性对模型带来的影响,本次研究采用递归特征消除算法对测井曲线进行优选。SVM-RFE 算法由 Guyon 等[15] 于 2002 年提出,SVM-RFE 利用 SVM 的权值作为 RFE 的特征排序标准, SVM-RFE 具体流程如图 4 所示,根据 SOM 分类结果,用 SVM-RFE 对 SOM 分类结果分别进行测井参数优选,各类测井参数的 SVM-RFE 的排序结果见表 1。表中数字代表在不同类中各测井参数的排序位置,数字越小说明这个测井参数排名越靠前,表明其对预测结果的影响也越大。通过对各参数的排序,由于排名靠后的测井参数对结果的影响较小,相反可能会带来某些错误的信息,因此,每一类中选择排名前八位的测井参数作为模型的输入。

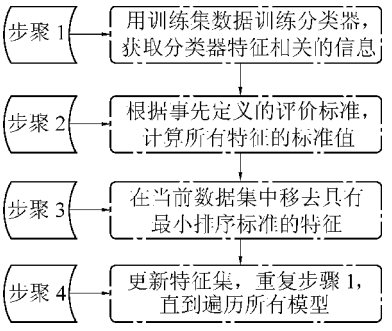


图 4　SVM-RFE 流程

Fig.4　SVM-RFE flow chart

表 1　各测井参数的 SVM-RFE 排序结果

Table 1　SVM-RFE sorting results for each logging parameters

类别	排序号												
	SP	CAL	GR	TH	K	U	AC	LLD	LLS	C_1	POR	PERM	KTH
a	6	11	4	7	8	2	5	10	9	3	12	1	13
b	6	4	8	11	12	13	9	2	1	7	5	3	10
c	13	3	6	2	5	1	7	12	10	11	9	8	4
d	11	1	13	3	12	4	10	9	10	6	5	2	7

3.3　TOC 含量预测模型建立

对本次研究的 150 组 TOC 含量数据进行预测。首先,对原始的测井数据进行 SOM 聚类,对每一类分别构建模型,每一类构建模型的步骤相同;然后对聚类后的每一类分别进行测井参数优选,其中测井参数优选模型 SVM-RFE 也涉及到 SVM 模型,模型的核函数选择依然为高斯核函数,同样的使用粒子群算法进行 SVM 模型参数优化,最终得到各测井参数的排序结果;然后把优选后的数据作为最终的输入数据,输入到 SVR 模型中,预测结果如图 5 所示。

3.4　交叉验证与误差分析

本次研究误差分析采用平均绝对误差(MAE)、均方根误差(RMSE)和平均相对误差(MRE)来评价模型性能。平均绝对误差是所有单个观测值与算数平均值的偏差的绝对值的平均;平均相对误差可以很好的反映预测值误差的实际情况;均方根误差又叫标准误差,它是观测值与真值偏差的平方和观测次数 n 比值的平方根,标准误差对一组测量中的特大或特小误差反映特别敏感,所以,标准误差能够很好的反映出测量的精密度。根据误差计算公式,4 类预测结果的误差见表 2。

表 2　SOM-SVR 模型预测误差

Table 2　Error of SOM-SVM model prediction

类别	SOM-SVR 模型误差/%		
	RMSE	MAE	MRE
a	0.116 5	0.063 6	3.45
b	0.299 6	0.183 4	9.70
c	0.203 6	0.144 8	7.64
d	0.220 0	0.116 8	8.96

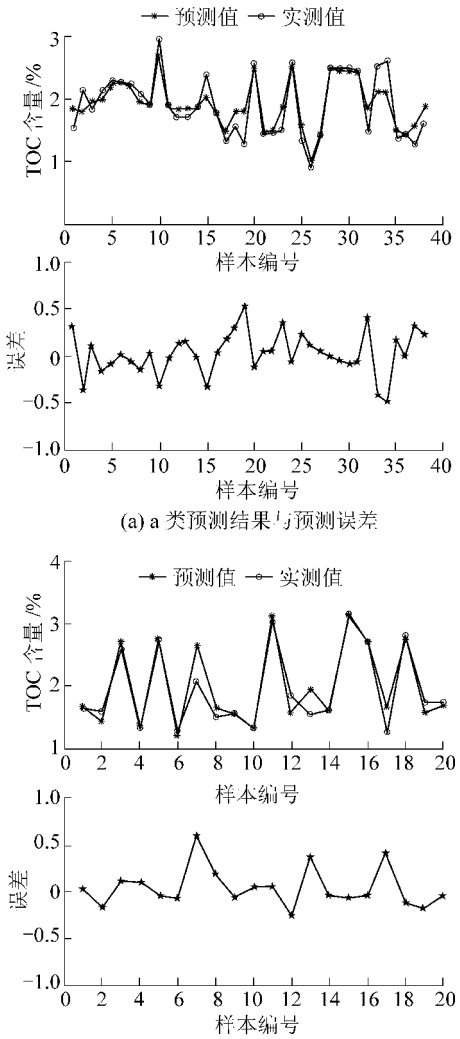
4　不同 TOC 预测模型应用效果分析

为了更好地验证模型的性能,把 SOM-SVR 模型和目前使用比较多的 2 种 TOC 含量计算法进行了对比。

4.1　$\Delta \lg R$ 模型

$\Delta \lg R$ 模型是由 Exoon 和 Esso 公司[16] 于 1997 开发的一种利用测井参数计算有机碳含量的技术。

中国煤炭行业知识服务平台www.chinacaj.net



(a) a 类预测结果与预测误差

(b) b 类预测结果与预测误差

图 5　预测结果与预测误差

Fig.5　Errors of prediction results and prediction

该方法将声波时差和电阻率 2 条曲线反向叠加,二者的坐标轴方向相反,电阻率采用对数坐标,声波时差采用算术坐标。采用声波时差和电阻率曲线叠合的 $\Delta\lg R$ 模型的 TOC 含量 c 计算公式为

$$\Delta\lg R = \lg(R/R_0) + K(\Delta t - \Delta t_0) \quad (3)$$

$$c = (\Delta\lg R) \times 10^{(2.297-0.168\,8LOM)} \quad (4)$$

式中:$\Delta\lg R$ 为电阻率和声波时差叠合后的间距;R 为实测电阻率值;Δt 为实测声波时差值;R_0 和 Δt_0 为对应的电阻率和声波时差基线值;K 为叠合系数;LOM 为有机质成熟度,通过镜质体反射率来确定。通过式(3)和式(4),使用 $\Delta\lg R$ 模型对研究地区的 TOC 含量进行了计算。

4.2　BP 神经网络模型

根据前人研究经验[5,6],建立了 TOC 含量的 BP 神经网络预测方法,主要包括隐含层个数的确定、激励函数的选择和权值阈值的优化。根据文献(6)选取了隐含层个数为 3,激励函数选择 Sigmoid 函数作为激励函数,根据文献(6)选择遗传算法来优化网络的权值和阈值,由此建立了 TOC 含量的 BP 神经网络预测模型。

4.3　对比结果分析

相比于 $\Delta\lg R$ 法和神经网络法,SOM-SVM 模型的预测结果能够更好的拟合实测值,预测误差小,模型精度明显提高。为了更好的比较模型的性能,分别计算了 $\Delta\lg R$ 和神经网络法的均方根误差和平均绝对误差以及平均相对误差见表 3。对比表 2 和表 3 可以看出,SOM-SVM 的均方根误差最小,说明其预测值与真实值偏离最小;平均相对误差最小不超过 10%,平均绝对误差最小不超过 0.2,相比于其他模型精度明显提高。

表 3　其他模型预测误差

Table 3　Errors of other model prediction

模型	RMSE/%	MAE	MRE/%
BP 神经网络	0.645	0.481 3	26.30
$\Delta\lg R$	1.118 6	0.903 1	55.60

5　结　　论

1)为去除测井数据之间的差异,通过聚类分析的方法,有效规避了岩石、岩相、流体含量等非有机质含量变化引起的测井曲线的变化的问题,同时使得各测井数据与 TOC 含量之间的相关性更加显著。

2)采用粒子群算法,优化 SVR 模型。分别建立了 BP 神经网络模型、$\Delta\lg R$ 模型和聚类 SVR 模型并做试算分析,结果表明:聚类的 SVR 模型对 TOC 含量的预测效果最好,且具有极强的预测稳定性。

3)采用 SVM-REF 算法进行测井曲线优选,去除测井参数之间的相关性给模型带来的影响,提高了模型的稳定性和模型精度。致谢:感谢由斯伦贝谢公司提供的 Petrel 软件,为本次论文的完成提供了有力的技术支持。

参考文献(References):

[1]　杜文凤,王　攀,梁明星,等.煤系烃源岩有机碳含量测井响应特征与定量预测模型[J].煤炭学报,2016,41(4):954-963.
DU Wenfeng,WANG Pan,LIANG Mingxing,*et al*.Well logging response characteristics and quantitative prediction model of organic carbon content of hydrocarbon source rocks in coal-bearing strata measures[J].Journal of China Coal Society,2016,41(4):954-963.

[2]　闫建平,蔡进功,赵铭海,等.运用测井信息研究烃源岩进展及其资源评价意义[J].地球物理学进展,2009,24(1):270-279.
YAN Jianping,CAI Jingong,ZHAO Minghai,*et al*.Advances in the study of source rock evaluation by geophysical logging and its sig-

中国煤炭行业知识服务平台www.chinacaj.net

nificance in resource assessment[J]. Progress in Geophysics, 2009,24(1):270-279

[3] 李延钧,张烈辉,冯媛媛,等.页岩有机碳含量测井评价方法及其应用[J].天然气地球科学,2013,24(1):169-175
LI Yanjun,ZHANG Liehui,FENG Yuanyuan *et al*.Logging evaluation method and its application for measuring the total organic carbon content in shale gas[J]. Natural Gas Geoscience, 2013, 24(1):169-175.

[4] 王濡岳,丁文龙,王 哲,等.页岩气储层地球物理测井评价研究现状[J].地球物理学进展,2015,30(1):228-241.
WANG Ruyue,DING Wenlong,WANG Zhe,*et al*.Progress of geophysical well logging in shale gas reservoir evaluation[J].Progress in Geophysical,2015,30(1):228-241.

[5] 孟召平,郭彦省,刘 尉.页岩气储层有机碳含量与测井参数的关系及预测模型[J]煤炭学报,2015,40(2):247-253.
MENG Zhaoping,GUO Yansheng,LIU Wei.Relationship between organic carbon content of shale gas reservoir and logging parameters and its prediction model[J].Journal of China Coal Society,2015,40(2):247-253.

[6] 王 攀,彭苏萍,杜文凤,等.基于测井参数的煤系烃源岩总有机碳含量预测模型[J].煤炭学报,2017,42(5):1266-1276.
WANG Pan,PENG Suping,DU Wenfeng,*et al*.Prediction model of total organic carbon content on hydrocarbon source rocks in coal measures based on geophysical well logging[J].Journal of China Coal Society,2017,42(5):1266-1276.

[7] EBRAHIM Sfidari,ALI Kadkhodaie Ilkhchi,SAEID Najjari.Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems[J].Computers & Geosciences,2012,87:190-205.

[8] AHMED Amara Konaté ,HEPING Pana,SINAN Fanga,*et al*.Capability of self-organizing map neural network in geophysical log data classification:a case study from the CCSD-MH[J].Computers & Geosciences,2015,118:37-46.

[9] KOHONEN T. The self-organize maps[J]. Proceeding of the IEEE,1990,78(9):585-622.

[10] SONG Chengyun,LIU Zhining,WANG Yaojun,*et al*.Multi-waveform classification for seismic facies analysis[J]. Computers & Geosciences,2017,101:1-9.

[11] VAPNIK V N.The nature of statistical learning theory[M].New York:Springer Verlag,1995.

[12] ANAZI A F Al,GATES I D.Support vector regression for porosity prediction in a heterogeneous reservoir:a comparative study[J]. Computers & Geosciences,2010,36(2):1494-1503.

[13] MAOJIN Tan,Xiaodong SONG,Xuan YANG,*et al*. Support-vector-regression machine technology for totalorganic carbon content prediction from wire line logs in organic shale:a comparative study[J].Computers & Geosciences,2015,26:792-802.

[14] JIAN Liu,ZHONG Liu,YING Xiong.Method of parameters optimization in SVM based on PSO[J].Transactionson Computer Science & Technology,2013,2(1):9-16.

[15] ISABELLE Guyon,JASON Weston,STEPHEN Barnhill.Gene selection for cancer classification using support vector machines[J]. Machine Learning,2002,46(3):389-422.

[16] PASSEY Q R,CREANEY S,KULLA J B,*et al*.A practical model for organic richness from porosity and resistivity logs[J].AAPG Bulletin,1990,74(12):1777-1794.