第42卷，第6期
2022年6月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 6, pp1691-1697
June, 2022

近红外光谱的油页岩总有机碳快速检测

李泉伦[1]，陈争光[1*]，孙先达[2]

1. 黑龙江八一农垦大学信息与电气工程学院，黑龙江 大庆 163319
2. 东北石油大学“陆相页岩油气成藏及高效开发”教育部重点实验室，黑龙江 大庆 163318

摘 要 为了快速检测油页岩总有机碳(TOC)含量，以松辽盆地某区块所取岩芯为研究对象，测量230个岩石样本的TOC含量和近红外光谱数据。利用蒙特卡洛法剔除异常样本14个，剩余的216个样本进行去趋势加基线校正方法预处理，采用连续投影算法(SPA)、无信息变量消除算法以及竞争自适应算法选取特征波长。使用SPXY方法对样本按照2∶1的比例划分为144个校正集和72个验证集，然后建立线性的偏最小二乘(PLS)模型以及非线性的支持向量机(SVM)模型和随机森林(RF)模型对油页岩TOC含量进行预测。采用测定系数(R^2)和均方根误差(RMSE)作为模型的评价指标，探究不同特征波长选择方法对油页岩总有机碳建模的影响，比较不同建模方法对油页岩TOC含量预测的准确度。结果表明，特征波长提取能够起到优化模型的作用。SPA，UVE和CARS分别提取了16，253和65个波长，经过特征波长提取后模型测定系数均有提高，均方根误差均有下降，这说明进行特征波长优选对于简化模型、提高模型运算速度发挥着很重要的作用。此外，非线性的RF和SVM模型性能要优于线性模型PLS。这是因为油页岩中的碳存在于各类烃的中，不同类别含烃基团的吸收峰之间相互影响，使得油页岩总有机碳含量和近红外光谱数据之间存在着复杂的非线性关系，因此，非线性的SVM和RF模型能够表现出更好的效果。相比于其他模型，CARS-SVM模型验证集的测定系数(R_v^2)和均方根误差(RMSEV)表现出的结果较好，分别达到了0.906 6和0.222 0，该模型能够用于油页岩总有机碳含量的快速检测。研究结果说明，近红外光谱分析应用于油页岩TOC含量快速检测是可行的；建立的CARS-SVM模型能够表现出较好的预测效果，为我国油页岩TOC含量快速检测提供了一种新的方法和思路。

关键词 近红外；油页岩总有机碳；特征波长；支持向量机；随机森林

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)06-1691-07

引 言

作为一种非常规石油资源，油页岩可以作为石油的一种替代能源。世界油页岩储量丰富并分布较为广泛。油页岩的勘测和开采具有重要的战略意义[1]。油页岩中的总有机碳(total organic carbon，TOC)含量是评价油页岩含油率的重要指标之一。目前，用于检测油页岩总有机碳的方法主要是按照GB/T19145—2003沉积岩中总有机碳的测定方法，其检测方法主要采取灼烧法。该方法具有准确度高的优点，但是容易产生残留污染，检测过程复杂繁琐且耗时较多，操作难度大，技术要求高等不足，无法满足快速检测油页岩总有机碳的要求。

近红外光谱分析技术是用于物质成分定量分析的一种快速检测技术。它具有检测速度快、无需破坏样品、不需要检测试剂、仪器操作简便等优点。目前，近红外光谱分析技术已经应用到农业[2]、医学[3]、食品[4]、石油[5]等领域。

近红外是指波长在780～2 500 nm范围内的电磁波，能够反应含氢基团(比如N—H，C—H，O—H等)震动的倍频与合频响应情况。近年来，近红外光谱分析逐渐应用于岩石、土壤中有机碳等成分含量的定量检测。申燕[6]等利用近红外光谱分析法对东北黑土的有机碳进行了测定，所建模型的拟合效果良好。王赛亚[7]等对煤炭和岩石的近红外光谱曲线特征和吸收特征进行了研究，证明了近红外光谱分析用于测定天然岩石中各种矿物含量是可行的。李耀翔[2]等利用近红外光谱测定了森林土壤的有机碳含量，经过预处理后验证

收稿日期：2021-05-05，**修订日期**：2021-07-06

基金项目：国家重点研发计划项目(2016YFD0701300)，黑龙江省博士后科研启动金项目(LBH-Q18028)资助

作者简介：李泉伦，1995年生，黑龙江八一农垦大学信息与电气工程学院硕士研究生 e-mail：18663065663@163.com

*通讯作者 e-mail：ruzee@sina.com

集相关系数达到 0.849 4。Romeo[8]等用近红外光谱分析了来自昆士兰州中部斯图尔特矿床的 53 个油页岩样品的烃(干酪根)含量。利用二阶导数和多元散射校正预处理后建立 PLS 校正模型，评价指标相关系数 R^2 到达 0.73，证明了可以使用近红外光谱分析预测油页岩的含油率。王宏智[9]等使用实验室合成样本，研究了不同波长组合选择方法对油页岩含油率近红外光谱数据进行波长筛选并建立留一交互校验多元线性回归模型进行验证。赵振英[5]等利用近红外光谱分析对油页岩含油率的波长选择方法进行了研究。由此可见，基于近红外光谱的测量技术在岩石和土壤中碳含量检测具有可行性。以上研究的样本多为人工合成模拟的油页岩样本或经过处理后的样本，并非自然条件下的油页岩样本，所建立的模型以线性的偏最小二乘(partial least squares，PLS)模型为主，模型精度不高。由于近红外光谱数据和理化值之间存在非线性关系[4]，因此，使用线性的 PLS 进行建模不足以表达自变量和因变量之间的关系，而基于非线性建模的支持向量机(support vector machine，SVM)、随机森林(random forest，RF)等方法在近红外光谱建模中受到越来越多的重视。以大庆油田松辽盆地某区块所取岩芯为研究对象，建立基于 SVM 和 RF 的油页岩 TOC 含量非线性模型，并和经典的 PLS 方法进行比较，以期为油页岩 TOC 含量快速检测建立更加稳定高效的近红外光谱模型，为油页岩总有机碳的检测提供更加简便快速的方法。

1 实验部分

1.1 样本

实验用大庆油田松辽盆地某区块采集的岩芯样本共计 230 个，在对岩心样本进行分析和数据采集前使用液氮冷冻保存，按照 GB/T19145—2003 沉积岩中总有机碳的测定方法测量其总有机碳的含量。

1.2 光谱采集

使用傅里叶变换近红外光谱仪 TANGO(德国 BRUKER 公司)采集光谱数据，波数范围：11 542～3 940 cm^{-1}，分辨率为 8 cm^{-1}，扫描 32 次取平均值。图 1 所示是全部 230 个油页岩样品的平均光谱。从光谱曲线来看，岩石样本光谱曲线基线漂移较为严重，在 7 300，5 200和 4 500 cm^{-1}三个波数附近有明显的吸收峰，同时伴随一定量的噪声。8 800 和 4 200 cm^{-1}附近有吸收峰，但不是很明显。

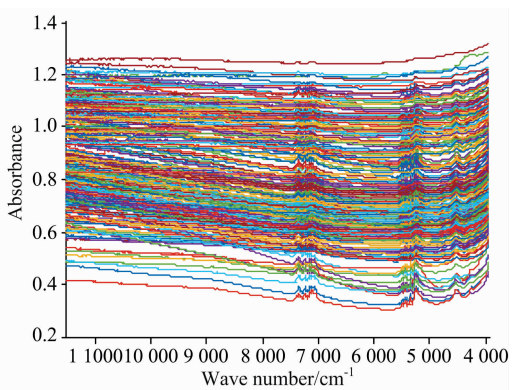

图 1 原始光谱图

Fig. 1 Original spectra

1.3 异常样本剔除

由于受到环境因素、样本来源多样性以及测量仪器等客观因素的影响，导致一些样本的光谱偏离样本的总体分布，这些所谓的异常样本引入将会导致模型的预测精度大幅下降[10]。因此，对异常样本进行剔除是保证定量模型可靠的必要条件。采用蒙特卡洛交叉验证算法对异常样本进行剔除。蒙特卡洛随机取样(Monte Carlo sampling，MCS)法每次随机抽取一定比例的样本(占样品量的 80%)构成校正集建立偏最小二乘模型，剩余的 20%作验证集对模型进行验证，计算验证集残差。经过多次抽样建模后能够得到多个预测残差，计算出这些预测残差的均值与方差，将样本预测值误差高于平均残差的样本标记为异常样本。最后通过校正集相关系数 R^2、交叉验证均方差 RMSECV、预测均方差 RMSEP 对模型进行评价，验证剔除异常样本是否有利于模型精度的提高。剔除异常样本前后的建模结果如表 1 所示，由表可以看出剔除异常样本后的模型的性能参数有明显改善。

表 1 剔除异常样本前后的建模结果

Table 1 Modeling results before and after removing outlier samples

剔除方法	样本个数	R_c^2	RMSECV	R_v^2	RMSEP
未剔除	230	0.660 8	0.419 0	0.624 6	0.443 6
MCCV	216	0.718 9	0.366 3	0.688 4	0.386 3

1.4 光谱预处理

为了消除噪声干扰和基线漂移对模型性能的影响，一般在建模之前对光谱数据进行预处理。目前比较常用的近红外光谱预处理方法有 Savitzky-Golay(S-G)卷积平滑、基线校正(baseline correction，BSC)、标准正态变量变换(sandard normal variate correction，SNV)、一阶导数(first derivative)、二阶导数(second derivative)、去趋势(detrend，DT)等。S-G 能有效提高光谱的平滑性，减少高频噪声干扰。SNV 主要是减少固体颗粒物大小不均和物体表面散射以及光程变换对光谱数据的影响，从而达到去除噪声的目的。DT 算法用来处理漫反射光谱基线漂移的问题，一般和 SNV 组合使用。导数方法用来消除背景干扰和基线校正，提高分辨率和灵敏度。通过对比不同预处理方法的效果最终确定适合油页岩样本的近红外光谱预处理方法。

1.5 特征波长选择

近红外光谱筛选波长后所建立的模型相比于全光谱模型，不仅模型变量数大幅度减少，而且模型性能也有大幅提升。进行特征波长选择可以通过简单的模型来提高模型的解释性，通过减少噪声或者干扰选择出效果更好的变量，并且能够提高模型的预测能力[11]，在众多的波长选择算法中，连续投影(successive projections algorithm，SPA)算法，无信息变量消除(uninformative variables elimination，UVE)算法和竞争自适应重加权(competitive adaptive reweighted sampling，CARS)算法具有一定的代表性，波长选择结果较优。

1.5.1　连续投影算法

连续投影算法是一种前向循环的特征变量选择算法，即从一个波长开始，其他波长向这个波长向量的法平面投影，投影长度最长的波长向量被选择为特征波长，然后以新选择的波长为基础，重复上述投影过程，直到达到指定的波长个数为止。该算法可以从众多光谱信息中筛选出重要样本变量的波长，用少数几列光谱数据来概括大部分光谱信息，降低了模型的复杂度，有效提高建模的速度和模型的稳定性。

1.5.2　无信息变量消除法

无信息变量消除算法是通过向样本光谱矩阵中人为引入随机噪声，并在此基础上建立偏最小二乘回归交叉验证模型，得到的偏最小二乘回归系数均值与标准差的商作为衡量波长重要性的关键指标，将噪声矩阵的最大值作为阈值，大于阈值的变量被作为优选的特征向量。UVE 算法可以去除没有贡献的变量，减少模型的运算量，增强模型的适应性。

1.5.3　竞争性自适应重加权算法

竞争自适应重加权算法是将每个波长变量看作是一个单位个体，将适应力弱的个体剔除，从而保留适应性强的个体。通过蒙特卡洛采样和 PLS 模型的测定系数进行特征波长选择，首先通过蒙特卡罗采样随机选择校正集样本建立 PLS 模型[12]。计算该模型各参数的系数权重，然后利用 CARS 算法筛选出 PLS 模型回归系数绝对值权重大的波长，去掉权重小的波长，模型交叉验证的均方根误差最小的波长组合即为选择的特征波长。该算法可以有效保留特征变量及相关影响变量，剔除冗余及噪声变量。

1.6　建模方法及评价指标

1.6.1　支持向量机

支持向量机是建立在统计学习理论的 VC 维理论和最小化结构风险基础上的一种有监督二分类机器学习算法。该算法的基本思想是将低维非线性问题转换成高维线性问题来分类，通过非线性变换将输入变量映射到一个高维的特征空间，并在新的空间中进行线性回归寻找一个最优的超平面，使得所有样本到超平面的距离最小，从而解决常规空间中样本线性不可分的问题[13]。支持向量机能够较好地解决高维空间中遇到的维数灾难问题，具有良好的泛化能力，在解决小样本，非线性样本分类以及高维模式识别中表现出很多特有的优势，并对异常样本及噪声具有很好的鲁棒性。

1.6.2　随机森林

随机森林算法是基于决策树和自助重采样法的一种集成学习算法。它的基本思想是利用自助重采样法不断生成训练变量集和检验变量集，由检验变量集随机生成多个分类决策树，每个决策树节点的分裂变量也是随机产生，从而形成随机森林。随机森林通过对产生的决策树进行投票得出最终预测结果，其结果具有较高的准确性。该算法优点体现在训练速度快且不易出现过拟合，能够很好的处理数量大的样本，并且对噪声具有较强的鲁棒性[14]。

1.6.3　模型评价指标

通过内部交叉验证和外部验证对模型效果进行检验，采用模型值和真实值的相关系数 R^2 和均方根误差(root mean square error，RMSE)作为模型的评价指标。R^2 越接近 1，模型 RMSE 值越小，说明模型的预测效果越好。

2　结果与讨论

2.1　光谱预处理和样本划分

针对光谱样本的消噪和基线矫正需求，分别采用 S-G 卷积平滑、SNV、BSC、DT、一阶导数和二阶导数 6 种方法以及组合处理共 8 种方法对光谱数据进行预处理，并以 PLS 模型的模型参数作为预处理效果评价光谱预处理方法。不同的光谱预处理方法的结果如表 2 所示。其中，基于 DT 方法(DT，DT＋SNV，DT＋BSC 等)的预处理方法模型性能优于其他方法。这是因为，岩石样本近红外光谱基线漂移较为严重(见图 1)，DT 预处理方法能在一定程度上消除漫反射光谱的基线漂移，因此效果较好。所有预处理方法中，在去趋势基础上进行基线校正的 DT＋BSC 预处理方法的 PLS 模型性能最优。经过预处理之后可以看出 11 541.94～8 872.71 cm^{-1}波段共 580 个波长点没有明显的吸收峰，且噪声明显。因此，在后续的研究中只对 DT＋BSC 处理后的 8 864.473～3 946.174 cm^{-1}波段共 1 265 个波长点的光谱数据进行分析。

表 2　不同预处理方法下的建模结果

Table 2　Modeling results under different pretreatment methods

光谱预处理方法	R_c^2	RMSEC	R_v^2	RMSEV
无	0.718 9	0.366 3	0.688 4	0.386 3
DT	0.746 0	0.341 4	0.621 6	0.382 0
Derivative-1	0.706 7	0.303 8	0.643 6	0.310 6
Derivative-2	0.719 8	0.365 8	0.663 9	0.342 3
SNV	0.548 7	0.464 2	0.376 0	0.426 5
S-G	0.729 3	0.359 6	0.695 3	0.382 4
BSC	0.658 5	0.403 8	0.570 8	0.449 4
DT＋SNV	0.739 3	1.352 7	0.578 0	0.334 8
DT＋BSC	0.759 1	0.315 9	0.734 8	0.357 3

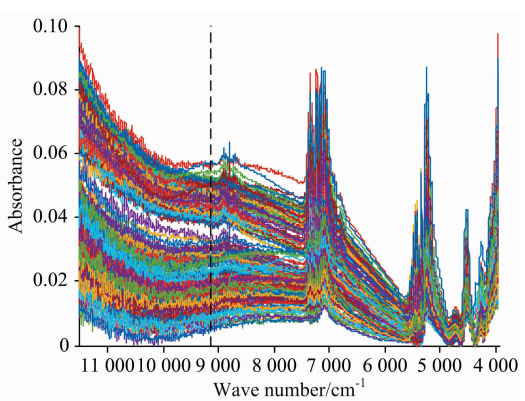

图 2　DT＋BSC 预处理的光谱

Fig. 2　Spectra after DT＋BSC pretreatment

预处理后的光谱数据采用 SPXY 算法按照 2∶1 的比例对剔除异常样本之后的 216 个油页岩样本进行划分，得到校正集样品 144 个，验证集样品 72 个。样本集划分结果见表 3，由表 3 可知，校正集和验证集样品的分布比较均匀，校正

集总有机碳样品含量基本涵盖了验证集。

表 3 样本集划分

Table 3 Sample set division

样本类型	样本数	最小值/%	最大值/%	均值/%	标准差
总体样本	216	0.053	3.922	1.835	0.693
校正集	144	0.053	3.922	1.897	0.747
验证集	72	0.699	2.783	1.713	0.551

2.2 波长选择结果

2.2.1 SPA 算法

图 3 为 SPA 算法提取不同数量特征波长对应的模型的 RMSE，从图中可以看出 RMSE 的值随波长数的增加而降低，当选取波长个数为 16 时，模型均方根误差到达稳定且最小。被选中的波长分别是 8 918，8 790，7 348，7 138，7 068，7 002，5 643，5 412，5 272，5 132，4 526，4 481，4 374，4 353，4 325 和 4 213 cm^{-1}，其分布情况如图 4 所示。这些波长与芳烃和 CH_3 伸缩振动的倍频峰以及 CH_2 变形伸缩振动的合频峰一致。

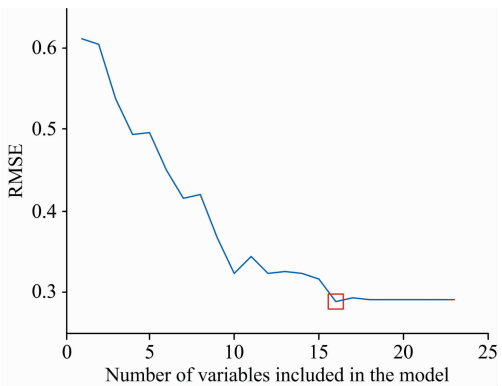

图 3 不同特征波长数的 RMSE 值

Fig. 3 RMSE values of models with different characteristic wavelength numbers

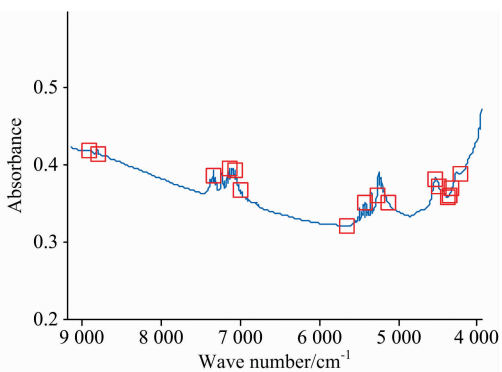

图 4 基于 SPA 算法筛选的特征波长

Fig. 4 Characteristic wavelengths selected based on SPA algorithm

2.2.2 UVE 算法

UVE 算法筛选的油页岩 TOC 特征波长结果如图 5 所示。其中竖线左侧为 1265 个光谱变量的稳定性指数分布曲线，右侧为 UVE 产生的同光谱变量相同数量的随机变量稳定性指数分布曲线。以随机变量稳定性指数绝对值的最大值作为筛选变量的阈值，即稳定性指数分布曲线在两条水平虚线以外的光谱变量被选中，共选择出 253 个特征波长，所有特征波长点位于光谱吸收峰附近，其分布如图 6 所示。

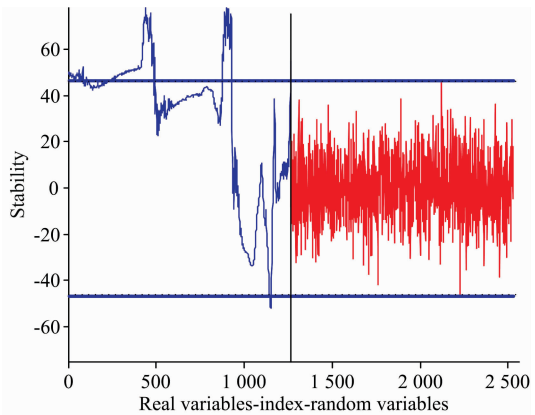

图 5 各波长变量和随机变量下的稳定性指数

Fig. 5 Stability index of each wavelength variable and random variable

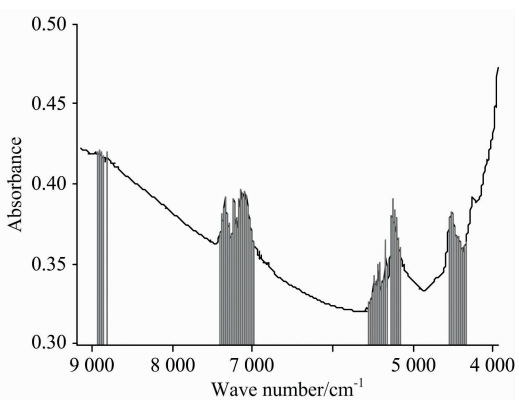

图 6 基于 UVE 算法筛选的特征波长

Fig. 6 Characteristic wavelengths selected based on UVE algorithm

2.2.3 CARS 算法

CARS 算法筛选的特征波长数、RMSE 以及回归系数随运行次数的变化如图 7 所示。从图中可以看出，当运行次数从 1 次增加到 24 次，特征波长数的下降由快变慢，RMSECV 逐渐降低，表明在 1～24 次运行过程中剔除了较多的无关光谱变量，模型精度不断提高。随着运行次数的继续增加，RMSECV 缓慢或者迅速增大，回归系数不断变大。在运行次数为 24 次时，RMSECV 值最低，此时有 65 个波长被保留下来，其中大部分波长位于光谱吸收峰附近，其分布如图 8 所示。

利用 SPA，UVE 和 CARS 进行特征波长筛选后，特征波长数明显少于全光谱波长的 1 265 个波长，分别是全谱波长的 1.26%，20%和 5.14%。说明特征波长筛选对于简化模型、提高模型效率能表现出较好的效果。此外从筛选出的波

长来看，有明显吸收峰或者吸收峰附近的波长被保留了下来。

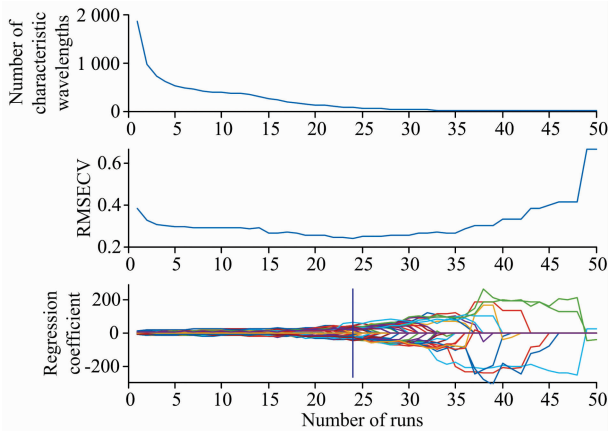

图 7　基于 CARS 算法筛选特征波长的过程

Fig. 7　Process of selecting characteristic wavelengths based on CARS algorithm

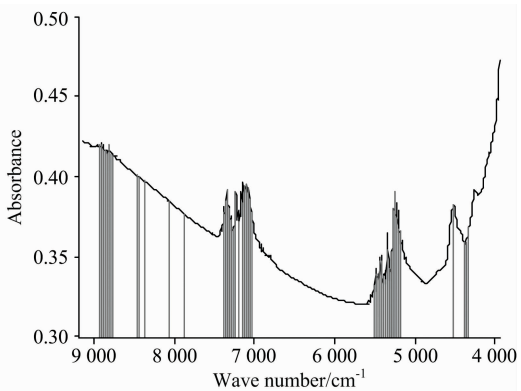

图 8　基于 CARS 算法筛选的特征波长

Fig. 8　Characteristic wavelengths selected based on CARS algorithm

2.3　不同模型的性能比较

为比较不同建模方法和不同特征波长提取方法对油页岩中有机碳含量的预测效果，分别采用 PLS、SVM 和 RF 建模方法对全光谱波段、CARS、SPA、UVE 筛选出的特征波长建立油页岩 TOC 含量的预测模型(表 3)。SVM 进行建模，以径向基函数作为模型核函数，根据网格搜索法优选惩罚因子和核函数参数。建立 RF 模型时，以不同特征波长提取方法下验证集相关系数最高时的决策树数量作为 RF 模型的最优决策树数目。

由表 4 可知，基于 CARS，UVE 和 SPA 三种特征波长的模型精度均高于全谱波长模型的精度。这是因为利用全光谱进行建模时，包含的变量较多，变量间存在有冗余信息的干扰，而特征波长提取可以有效去除冗余信息，提取后的波长能充分代表原始光谱的有效信息，从而提高模型质量。在三种特征波长选择方法中，基于 CARS 算法提取特征波长之后所建立的模型效果最好，尤其是 CARS-SVM 模型，其验证集测定系数由未进行特征波长选择时的 0.793 0 提升到 0.906 6，均方根误差由 0.286 8 降低到 0.222 0，是所有模型中最优的，该模型可以应用于油页岩总有机碳含量预测。

表 4　不同方法建模结果

Table 4　Modeling results of PLS, RF and SVM

建模方法	特征波段筛选方法	波段数	校正集		验证集	
			R_c^2	RMSEC	R_v^2	RMSEV
SVM	全谱波段	1 265	0.867 8	0.260 5	0.793 0	0.286 8
	CARS	65	0.912 4	0.204 6	0.906 6	0.222 0
	UVE	253	0.882 8	0.222 6	0.806 5	0.282 0
	SPA	16	0.883 5	0.244 0	0.815 3	0.278 0
RF	全谱波段	1 265	0.785 9	0.274 6	0.717 4	0.342 7
	CARS	65	0.812 5	0.297 6	0.765 9	0.316 9
	UVE	253	0.857 2	0.285 1	0.799 3	0.278 0
	SPA	16	0.842 5	0.293 7	0.812 7	0.323 5
PLS	全谱波段	1 265	0.765 5	0.314 8	0.742 1	0.342 5
	CARS	65	0.803 8	0.322 4	0.761 6	0.280 9
	UVE	253	0.834 5	0.284 3	0.754 1	0.310 0
	SPA	16	0.774 8	0.316 2	0.762 7	0.320 6

三种模型中，SVM 模型的效果优于 RF 模型和 PLS 模型，说明基于 SVM 法建立的预测模型能较好地应用于近红外光谱检测模型，其预测精度高、实用性强，对快速准确检测油页岩 TOC 含量有实际价值[15]。此外，相较于近红外光谱领域常用的线性的 PLS 模型而言，SVM 和 RF 两种非线性的建模方法得到模型在校正集上的表现均有不同程度的提高，其中 SVM 模型无论在校正集还是在验证集均有明显提升。这是因为油页岩样本中的碳存在于各类烃的中，由于不同类别含烃基团的吸收峰之间相互影响，使得油页岩 TOC 含量和近红外光谱数据之间存在着复杂的非线性关系。作为线性分析的 PLS 模型，其表达光谱数据和浓度数据之间的非线性关系能力不及 SVM 和 RF，后者能够有效地对自变量和因变量之间的非线性关系进行描述，从而增强了光谱数据和样本理化值之间的相关性，使得所建立的非线性模型效果要略优于线性的 PLS 模型。该结果与陈华舟[4]等基于近红外光谱针对鱼粉蛋白进行定量分析所建立的非线性模型 SVM 和 RF 效果优于线性模型 PLS 的效果结论一致。建立非线性模型比线性模型的 R^2 更好，准确度更高[16]。因此，可以通过非线性建模近红外光谱法实现油页岩总有机碳含量快速检测。

3　结　论

运用近红外光谱分析结合化学计量学方法对油页岩总有机碳含量进行了定量分析，研究结果表明：去趋势和基线校正组合的预处理方式针对油页岩总有机碳的近红外光谱数据建立的模型表现出了较好的效果。使用三种不同特征波长算法进行波长提取后建立的模型精度相比于全光谱模型均有所提高，说明了进行波长筛选对近红外光谱建模的重要性。对于油页岩有机碳含量预测模型而言，利用非线性模型进行预测分析得到的效果更好。使用 CARS-SVM 方法对油页岩总

有机碳含量进行预测，模型能够达到较好的效果，CARS-SVM 方法在对油页岩总有机碳含量检测方面有着巨大潜力。本研究为油页岩总有机碳的快速检测提供了一种新的思路和方法。

References

[1] ZHANG Xiao-ming, PAN Yi, YANG Shuang-chun, et al(章小明，潘 一，杨双春，等). Contemporary Chemical Industry(当代化工), 2012, 41(4): 377.

[2] LI Yao-xiang, WANG Hong-tao, GENG Zhi-wei, et al(李耀翔，汪洪涛，耿志伟，等). Journal of West China Forestry Science(西部林业科学), 2014, 43(3): 1.

[3] Bensaidance M R, Turgeon A F, Lauzier F, et al. BMJ Open, 2020, 10(11): e043300.

[4] CHEN Hua-zhou, CHEN Fu, SHI Kai, et al(陈华舟，陈 福，石 凯，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2015, 46(5): 233.

[5] ZHAO Zhen-ying, LIN Jun, ZHANG Fu-dong, et al(赵振英，林 君，张福东，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2014, 34(11): 2948.

[6] SHEN Yan, ZHANG Xiao-ping, LIANG Ai-zhen, et al(申 艳，张晓平，梁爱珍，等). Chinese Journal of Applied Ecology(应用生态学报), 2010, 21(1): 109.

[7] WANG Sai-ya, WANG Shi-bo, GE Shi-rong, et al(王赛亚，王世博，葛世荣，等). Journal of China Coal Society(煤炭学报), 2020, 45(8): 3024.

[8] Romeo M, Adams M, Hind A, et al. Journal of Near Infrared Spectroscopy, 2002, 10(3): 223.

[9] WANG Zhi-hong, ZHANG Fu-dong, TENG Fei, et al(王智宏，张福东，滕 飞，等). Optics and Precision Engineering(光学精密工程), 2015, 23(2): 371.

[10] YIN Bao-quan, SHI Yin-xue, SUN Rui-zhi, et al(尹宝全，史银雪，孙瑞志，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2015, 46(S1): 122.

[11] Yun Y H, Li H D, Deng B C, et al. Trends in Analytical Chemistry, 2019, 113: 102.

[12] Li H, Liang Y, Xu Q, et al. Analytica Chimica Acta, 2009, 648(1): 77.

[13] Suykens J A K, Vandewalle J. Neural Processing Letters, 1999, 9(3): 293.

[14] Breiman L. Machine Learning, 2001, 45(1): 5.

[15] ZHOU Yu-qing, QIN Meng-zhi, MA Zhi-hong(周宇晴，秦梦芝，马志宏). Journal of Tianjin Agricultural University(天津农学院学报), 2016, 23(2): 49.

[16] QIU Lan-lan, LI Ming-mei(裘兰兰，李明梅). China Pharmacy(中国药房), 2016, 27(27): 3850.

Rapid Detection of Total Organic Carbon in Oil Shale Based on Near Infrared Spectroscopy

LI Quan-lun[1], CHEN Zheng-guang[1*], SUN Xian-da[2]

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China
2. Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development, Ministry of Education, Northeast Petroleum University, Daqing 163318, China

Abstract To quickly detect the total organic carbon (TOC) content of oil shale, the TOC content and near-infrared spectrum data of 230 rock samples were measured in a certain block of the Songliao basin. The Monte Carlo method eliminates 14 abnormal samples, and the remaining 216 samples are pretreated by the method of detrended and baseline correction. The feature wavelength is selected by successive projections algorithm (SPA), uninformative variable elimination (UVE) algorithm and competitive adaptive reweighted sampling (CARS) method. The SPXY method divides the sample set into calibration set (144 samples) and validation set (72 samples) according to the ratio of 2∶1. Then linear partial least squares (PLS) model, nonlinear support vector machine (SVM) model and random forest (RF) model are adopted to predict the TOC content of oil shale. The determination coefficient (R^2) and root mean square error (RMSE) was used as the evaluation indexes of the model to explore the influence of different characteristic wavelength selection methods on TOC modeling of oil shale and to compare the accuracy of different modeling methods on TOC content prediction of oil shale. The results show that the feature wavelength extraction can optimize the model. SPA, UVE and CARS extract 16, 253 and 65 wavelength points respectively. After the feature

wavelength extraction, the model determination coefficient is improved, and the root means square error is decreased. This shows that the feature wavelength extraction plays an important role in simplifying the model and improving model efficiency. In addition, The performance of the nonlinear RF and SVM model is better than that of the linear PLS model. The reason is that the carbon in oil shale exists in all kinds of hydrocarbons, and the absorption peaks of different hydrocarbon groups interact with each other, which makes the complex nonlinear relationship between the TOC content of oil shale and the near-infrared spectroscopy data. Therefore, the nonlinear SVM and RF model can show better performance. Compared with other models, the coefficient of determination (R_v^2) and root mean square error (RMSEV) of the CARS-SVM model invalidation set show better results, reaching 0.906 6 and 0.222 0 respectively. This model can be used to rapidly detect TOC content in oil shale. The results of this study show that the application of near-infrared spectroscopy in the rapid detection of TOC content in oil shale is feasible, and the CARS-SVM model can show good prediction performance, which provides a new method and idea for the rapid detection of TOC content in oil shale in China.

Keywords Near-infrared; Total organic carbon in oil shale; Characteristic wavelength; Support vector machine; Random forest

* Corresponding author

(Received May 5, 2021; accepted Jul. 6, 2021)

(上接 1683 页)

会务组联系方式

毛慰明(会议稿件)
云南师范大学物理与电子信息学院
电话：0871-65941168；13529401604
e-mail：maoweiming3@126.com
欧全宏(会议咨询)
云南师范大学物理与电子信息学院
电话：0871-65941168；15908891183
e-mail：ouquanhong@163.com
王香凤(厂商联络)
北京师范大学分析测试中心
电话：010-58807981；13520034335
Email：xiangfeng@bnu.edu.cn
刘文广(厂商联络)
云南师范大学物理与电子信息学院
电话：0871-65941168；15987101479
e-mail：liuwgkm@qq.com

支持媒体

会议官网：光谱网：http://www.sinospectroscopy.org.cn(会议各类信息以光谱网发布为准)

主办单位：

中国光学学会
中国化学会
中国光学学会光谱专业委员会

承办单位：

云南师范大学物理与电子信息学院

第 22 届全国分子光谱学学术会议暨 2022 年光谱年会组委会
2021 年 12 月 28 日