不同含水量的岩石近红外光谱的特征选择

张 芳[1,2]，户佐乐[1,2]，侯欣莉[3]，张秀莲[1,2]，付成功[1,2]，李英骏[1,4]，何满潮[1]

1. 中国矿业大学(北京)深部岩土力学与地下工程国家重点实验室，北京 100083
2. 中国矿业大学(北京)力学与建筑工程学院，北京 100083
3. 天津泰达绿化集团有限公司，天津 300457
4. 中国矿业大学(北京)理学院，北京 100083

摘 要 岩石含水量是影响岩石物理、化学和力学特性的一个重要指标。在岩土工程、隧道工程等领域，岩石含水量的大小是诱发灾变和病害的关键原因。与传统方法相比，利用近红外光谱(NIRS)特征检测岩石含水量，具有无损、定量的明显优势，其难点和关键是近红外光谱的特征选择。针对该问题，进行了室内实验，研究不同含水量下的岩石近红外光谱的特征选择。特征选择方法中的Filter法，利用样本数据内在的特点，评价特征的重要程度，增强了特征与类的相关性，同时削减了特征之间的相关性，具有复杂度低、直观、效率高、普适性强的优点，符合该研究的数据特点。因此，选用Filter型的依赖性度量法进行特征选择。室内实验中，首先制备11种不同含水量的砂岩试样，并分别采集了前后左右4个测试点处的共计44条近红外光谱曲线；然后，利用一阶导数法对光谱进行预处理，基于此，选择1 400和1 930 nm谱段进行光谱特征分析，并分别提取2个谱段处的峰面积、峰高、半高宽、左肩宽度、右肩宽度、左右肩宽比共计6个初始特征变量；考虑到6个初始特征变量的量纲不同，且变量之间的变化幅度不同，对原始数据进行正规化变换，消除量纲和变化幅度不同带来的影响；接着，根据自变量的筛选原则，去掉自变量之间具有强线性相关的冗余变量；然后，利用依赖性度量法中的统计相关系数作为相关程度的度量标准，分析了初始特征变量之间以及初始特征变量与含水量之间的相关程度，并得到了2个强相关谱段处的最优特征变量；最后，在强相关谱段处分别构建了多元回归模型，并对模型进行了检验分析。研究结果表明：(1)波长1 400和1 930 nm附近的近红外光谱吸收峰特征与岩石含水量有明显相关性；(2)波长1 400 nm处的峰高、右肩宽度、左肩宽度与含水量线性相关性明显；波长1 930 nm处的峰高、右肩宽度与含水量线性相关性明显；(3)多元线性回归模型能够较精确表达含水量与近红外光谱之间的相关性，利用该模型可实现基于近红外光谱特征的含水岩石含水量预测，为利用近红外光谱实现动态监测与评估岩石含水量提供基础建模数据。

关键词 含水岩石；近红外光谱；特征选择；相关性；含水量预测

中图分类号：TU458 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)11-3395-08

引 言

岩石含水量是影响岩石物理、化学和力学特性的一个重要指标。在岩土工程、隧道工程等领域，岩石含水量的大小是诱发灾变和病害的关键原因。与传统方法相比，利用近红外光谱(NIRS)特征检测岩石含水量，具有无损、定量的明显优势，近年来该方法逐渐引起人们的关注。

光谱特征选择是建立岩石含水量与近红外光谱特征之间的定量关系的难点和关键。如何从高维数据中剔除冗余或无关的特征变量是目前面临的难题。一个合适的特征选择方法，不仅可以有效简化推理规则，还可在不降低模型准确性和稳定性的前提下提高运行效率。目前国内外诸多学者已开展了大量的研究工作[1-7]，根据评估方法可将特征选择大致分为过滤型(Filter)、封装型(Wrapper)和嵌入型(Embedded)三类方法，其中Filter利用样本数据内在的特点，评价特征的重要程度，增强了特征与类的相关性，同时削减了特征之间的相关性，具有复杂度低、直观、效率高、普适性强的优点，符合本研究的数据特点，故采用此方法。

Filter法的度量标准有四类，分别是距离度量、信息度

收稿日期：2018-09-12，**修订日期**：2019-02-07

基金项目：国家自然科学基金青年基金项目(51604276)资助

作者简介：张 芳，1976年生，中国矿业大学(北京)高级工程师 e-mail：zhangf76@163.com

量、依赖性度量和一致性度量，其中距离度量常用欧氏距离、平方距离等参数度量；信息度量常用信息增益、互信息等参数度量；依赖性度量常用 Pearson 相关系数、Fisher 分数、平方关联系数等参数度量；一致性度量采用不一致率进行度量。这四类度量标准适用范围不同，也有局限性：距离度量函数要求满足单调性，信息度量(如 BIF 法)未考虑到所选特征间的相关性，会带来较大冗余，一致性度量对噪声数据比较敏感，不适用于近红外光谱。

综上所述，根据数据特点。选用 Filter 型的依赖性度量法进行特征选择，分析岩石在不同含水量下的近红外光谱的特征变量之间以及特征变量与含水量之间的相关关系，选择最优特征变量，为利用近红外光谱分析技术实现动态监测与评估岩石含水量提供基础建模数据。

1 实验部分

1.1 系统介绍

利用深部软岩气态水吸附智能测试系统[图 1(a)]、电子恒温水箱[图 1(b)]和真空干燥箱[图 1(c)]制备不同含水量的砂岩，数据采集系统为瑞士万通的 XDS SmartProbe 近红外光谱分析仪，如图 2 所示，其技术性能参数如表 1。

图 1 制备不同含水量岩石的试验系统

(a)：深部软岩气态水吸附智能测试系统；(b)：电子恒温水箱；(c)：真空干燥箱

Fig. 1 Test system for preparing rock with different water contents

(a)：Intelligent testing system for vapour water adsorption of deep soft rock;
(b)：Electronic constant temperature water tank; (c)：Vacuum drying oven

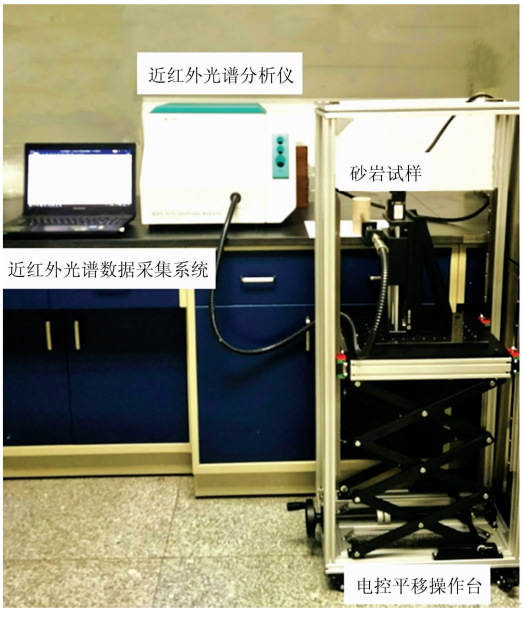

图 2 近红外光谱测试系统

Fig. 2 Test system of near infrared spectroscopy

表 1 技术性能参数

Table 1 Technical performance parameters

探头类型	反射探头
采样方式	漫反射-固体
光谱范围	400～2 500 nm
测样点个数	4
测样角度	90°
测样方式	直接与样品接触

1.2 方法

砂岩样品(图 3)取自陕西省榆林市神木县中部孙家岔乡的柠条塔井煤矿，其基本的物理力学参数如表 2。分别制备含水量为 0%，10%，20%，…，90%，100%共 11 个不同含水量的砂岩试样，其制备步骤为：①将砂岩试样放在真空干燥箱[图 1(c)]干燥 12 h 后，测量含水量为 0%的近红外光谱曲线；②将砂岩试样放入电子恒温水箱[图 1(b)]煮沸 8 h，使其达到饱和状态；③取出砂岩试样在室温下晾干，待表面自由水消失后，测量近红外光谱，得到含水量 100%的近红外光谱曲线；④将饱和砂岩试样放到深部软岩气态水吸附智能测试系统[图 1(a)]中，进行蒸发实验，观察含水量曲线，分别在理论计算含水量达到 36，32，…，4 g 时，中止含水量制备实验，分别测量含水量为 90%，80%，…，10%的近红外光谱曲线。

实验过程中将光纤探头分别垂直接触试样的前后左右共 4 个测试点进行测量，整个实验共采集 11 个不同含水量的共 44 条近红外光谱曲线。

1.3 分析软件参数设置

数据分析所用到的软件有 Origin8.0 和 Matlab7.0，其中利用 Origin 分别提取光谱吸收峰的 6 个初始特征变量值；利用 Matlab 计算吸收峰回归模型的各个参数和检验临界值，设置在显著性水平为 0.05 的水平下有意义，即认为该回归方程具有 0.95 的置信度。

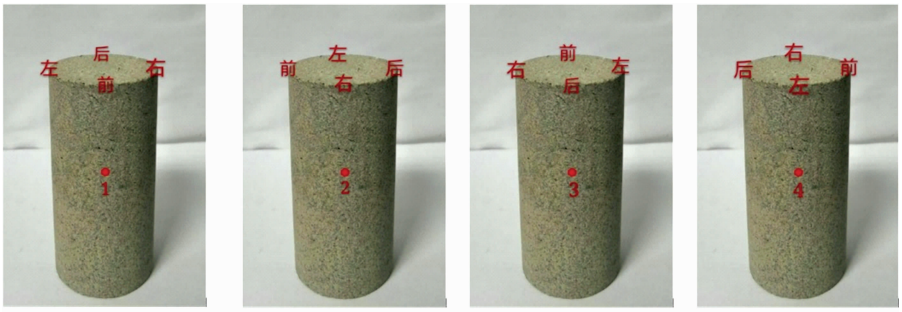

图 3　砂岩测试点位置

(a)：测点 1；(b)：测点 2；(c)：测点 3；(d)：测点 4

Fig. 3　Test point locations of sandstone

(a)：Point 1；(b)：Point 2；(c)：Point 3；(d)：Point 4

表 2　基本物理力学参数

Table 2　Basic physical and mechanical parameters

岩性	高度/mm	直径/mm	干燥后质量/g	吸水饱和后质量/g	岩样描述
砂岩	101.24	49.58	431.94	471.94	灰褐色，表面较粗糙，手摸有砂感，结构致密

2　近红外光谱特征选择

2.1　近红外光谱预处理

首先利用 XDS SmartProbe 近红外光谱分析仪配套软件将每个含水量的 4 条近红外光谱取平均值，然后再利用一阶导数法对该光谱进行预处理，消除背景的常数平移对近红外光谱的影响，使数据具有更好的连续性，处理前后光谱如图 4。

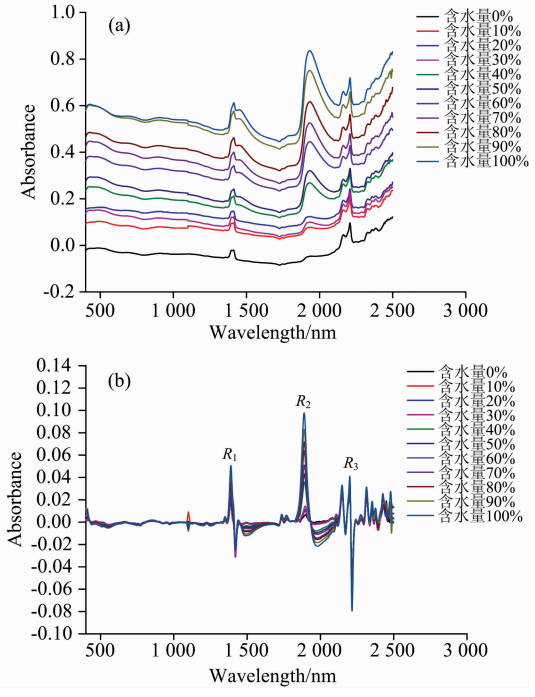

图 4　近红外光谱

(a)：原始光谱；(b)：一阶导数预处理后光谱

Fig. 4　Near-infrared spectrum

(a)：Original spectra；(b)：First-order derivative spectra

2.2　谱段选择和特征提取

分析预处理后的近红外光谱[图 4(b)]可知，在 400～2 500 nm 波长范围内有 3 个明显的吸收峰，分别位于波长 1 400，1 930 和 2 300 nm 附近，三处波长光谱的反射率随岩石含水量变化而变化，依次将其记作峰 R_1、峰 R_2、峰 R_3。随着含水量的不断增大，R_1 和 R_2 两个吸收峰的波峰越来越高，峰顶中心位置逐渐右移，最终 R_1 峰中心点位置停留在 1 400 nm 左右，R_2 峰中心位置停留在 1 930 nm 左右，而 R_3 吸收峰的波峰随着含水量增加逐渐减小，信号特征逐渐减弱，且受 2 400 nm 之后的噪音波段干扰强烈，故峰 R_3 不适合作为含水量信息的特征谱段。因此，选择峰 R_1、峰 R_2 所在的 1 400 和 1 930 nm 谱段进行含水岩石光谱特征分析，其具体提取的特征变量如图 5 所示，分别为峰面积(area)、峰高(height)、半高宽(FWHM)、左肩宽度(left half width，

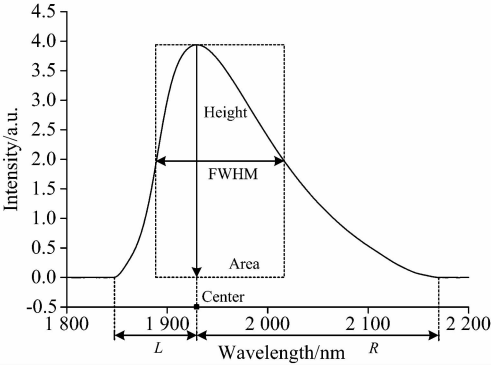

图 5　吸收峰的初始特征变量示意图

Fig. 5　Schematic diagram of the initial characteristic variables of the absorption peak

LHW)、右肩宽度(right half width，RHW)、左右肩宽比(LHW/RHW)共计 6 个初始特征变量，记作 $X=\{x_1, x_2, x_3, x_4, x_5, x_6\}$。

利用 origin8.0 分别计算峰 R_1、峰 R_2 的 6 个初始特征变量值，如表 2 和表 3 所示。

表 3 峰 R_1 处的初始特征变量值

Table 3 Initial characteristic variable values at the peak R_1

因变量 Y (含水量/%)	自变量 X(初始特征变量)					
	x_1：峰面积/nm	x_2：峰高	x_3：半高宽/nm	x_4：左肩宽度/nm	x_5：右肩宽度/nm	x_6：左右肩宽比
0	0.344 9	0.019 3	16.856 8	9.240 3	7.616 6	1.213 2
10	0.345 3	0.018 8	17.120 6	9.089 7	8.030 8	1.131 9
20	0.335 6	0.018 7	16.977 1	9.138 6	7.838 4	1.165 9
30	0.417 8	0.023 4	16.939 5	9.208 5	7.731 0	1.191 1
40	0.493 1	0.025 2	18.002 6	9.674 1	8.328 5	1.161 6
50	0.526 5	0.026 0	18.451 4	9.698 1	8.753 4	1.107 9
60	0.617 9	0.038 7	19.387 8	9.742 4	9.645 5	1.010 0
70	0.749 8	0.035 1	19.234 4	9.809 8	9.424 5	1.040 9
80	0.872 3	0.039 7	19.792 8	10.415 9	9.376 8	1.110 8
90	0.985 0	0.043 6	20.285 0	10.377 9	9.907 1	1.047 5
100	1.146 7	0.049 6	20.703 4	10.357 3	10.346 1	1.001 1
均值	0.621 3	0.029 8	18.522 9	9.704 8	8.818 1	1.107 4
最大值	1.146 7	0.049 6	20.703 4	10.415 9	10.346 1	1.213 2
最小值	0.335 6	0.018 7	16.856 8	9.089 7	7.616 6	1.001 1

表 4 峰 R_2 处的初始特征变量值

Table 4 Initial characteristic variable values at the peak R_2

因变量 Y (含水量/%)	自变量 X(初始特征变量)					
	x_1：峰面积/nm	x_2：峰高	x_3：半高宽/nm	x_4：左肩宽度/nm	x_5：右肩宽度/nm	x_6：左右肩宽比
0	0.204 2	0.005 9	35.519 8	23.128 9	12.390 9	1.866 6
10	0.300 7	0.009 5	31.421 0	19.026 1	12.394 9	1.535 0
20	0.333 5	0.010 8	29.556 9	17.199 9	12.357 0	1.391 9
30	0.456 1	0.013 7	30.708 8	17.507 9	13.200 9	1.326 3
40	1.422 1	0.035 5	36.016 3	18.708 6	17.307 7	1.080 9
50	1.698 5	0.042 0	36.341 2	18.888 9	17.452 3	1.082 3
60	2.044 7	0.050 2	36.569 8	18.944 6	17.625 2	1.074 9
70	2.573 7	0.063 1	36.599 1	18.958 1	17.640 9	1.074 7
80	2.897 6	0.070 8	36.714 8	19.259 0	17.455 7	1.103 3
90	3.418 1	0.082 7	37.140 4	19.481 0	17.659 3	1.103 2
100	4.001 0	0.096 9	37.002 3	19.368 2	17.634 2	1.098 3
均值	1.759 1	0.043 7	34.871 8	19.133 8	15.738 1	1.248 9
最大值	4.001 0	0.096 9	37.140 4	23.128 9	17.659 3	1.866 6
最小值	0.204 2	0.005 9	29.556 9	17.199 9	12.357 0	1.074 7

2.3 特征变量归一化

分析表 3 和表 4 可知，6 个初始特征变量的量纲不同，且变量之间的变化幅度不同，可能导致在分析计算过程中，一些数量级较小的变量作用无法体现，因此对原始数据进行正规化变换，即将所有特征变量转换成 0－1 内的数值，消除量纲和变化幅度不同带来的影响。

归一化方法是将原始数据矩阵的各元素减去该元素所在列的最小值后再除以该列元素的极差，公式如下

$$x'_{ij}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \quad (1)$$

归一化结果如表 5 和表 6 所示。

2.4 特征变量的相关性分析

2.4.1 筛选原则

对特征变量进行筛选和简化，即去掉自变量 X 之间具有强线性相关的冗余变量，以及自变量中与因变量相关性较小的变量。

自变量的筛选原则[8]：(1)变量的零值测试，即变量不应过于接近零值，否则将使变量的作用偏小；(2)变量的方差测试，标准方差越大的变量，其作用也越大，反之则越小，因此应删除标准方差接近零的变量；(3)自变量的相关性测试，即根据自变量间的相关系数可以判断自变量间的相关程度，取门槛值 0.95 作为特征变量取舍的界限，当两自变量相

表 5　峰 R_1 处归一化后的初始特征变量值

Table 5　Iniuial characteristic variable values after normalization at the peak R_1

因变量 Y（含水量/%）	自变量 X(初始特征变量)					
	x_1：峰面积	x_2：峰高	x_3：半高宽	x_4：左肩宽度	x_5：右肩宽度	x_6：左右肩宽比
0	0.011 5	0.020 0	0.000 0	0.113 5	0.000 0	1.000 0
10	0.011 9	0.004 2	0.068 6	0.000 0	0.151 8	0.616 6
20	0.000 0	0.000 0	0.031 3	0.036 9	0.081 3	0.777 0
30	0.101 3	0.152 1	0.021 5	0.089 6	0.041 9	0.896 0
40	0.194 1	0.211 2	0.297 9	0.440 6	0.260 8	0.756 6
50	0.235 4	0.235 4	0.414 6	0.458 7	0.416 5	0.503 7
60	0.348 0	0.323 5	0.658 0	0.492 1	0.743 3	0.042 3
70	0.510 7	0.531 5	0.618 1	0.543 0	0.662 4	0.187 6
80	0.661 6	0.678 4	0.763 3	1.000 0	0.644 9	0.517 4
90	0.800 6	0.805 3	0.891 2	0.971 3	0.839 2	0.218 9
100	1.000 0	1.000 0	1.000 0	0.955 8	1.000 0	0.000 0
均值	0.352 3	0.360 1	0.433 1	0.463 8	0.440 2	0.501 5
标准方差	0.346 3	0.345 4	0.373 5	0.381 4	0.354 0	0.346 3
最大值	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
最小值	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0

表 6　峰 R_2 处归一化后的初始特征变量值

Table 6　Initial characteristic variable values after normalization at the peak R_2

因变量 Y（含水量/%）	自变量 X(初始特征变量)					
	x_1：峰面积	x_2：峰高	x_3：半高宽	x_4：左肩宽度	x_5：右肩宽度	x_6：左右肩宽比
0	0.000 0	0.000 0	0.786 3	1.000 0	0.006 4	1.000 0
10	0.025 4	0.039 5	0.245 8	0.308 0	0.007 1	0.581 3
20	0.034 1	0.053 7	0.000 0	0.000 0	0.000 0	0.400 6
30	0.066 4	0.086 1	0.151 9	0.052 0	0.159 2	0.317 7
40	0.320 8	0.324 9	0.851 8	0.254 5	0.933 7	0.007 9
50	0.393 6	0.397 3	0.894 6	0.284 9	0.960 9	0.009 7
60	0.484 8	0.486 9	0.924 8	0.294 3	0.993 6	0.000 3
70	0.624 1	0.628 4	0.928 6	0.296 5	0.996 5	0.000 0
80	0.709 4	0.713 5	0.943 9	0.347 3	0.961 6	0.036 2
90	0.846 5	0.843 7	1.000 0	0.384 7	1.000 0	0.036 0
100	1.000 0	1.000 0	0.981 8	0.365 7	0.995 3	0.029 9
均值	0.409 5	0.415 8	0.700 9	0.326 2	0.637 7	0.219 9
标准方差	0.355 4	0.350 2	0.373 7	0.255 0	0.473 6	0.327 2
最大值	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
最小值	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0

关系数大于此值时，应剔除其一；(4)自变量与因变量的相关性测试，即根据自变量与因变量间的相关系数可以判断相关程度，本文取门槛值 0.85 作为特征变量取舍的界限，当自变量与因变量相关系数小于此值时，应剔除此特征变量。

2.4.2　特征选择与分析

(1)自变量零值和标准方差分析

由表 5 和表 6 可知，特征变量经归一化处理后没有出现零值；分析特征变量的标准方差，可知峰 R_1 各特征变量的标准方差比较接近，故全部保留，则峰 R_1 处筛选后的特征变量为 $X=\{x_1, x_2, x_3, x_4, x_5, x_6\}$；峰 R_2 中左肩宽度 x_4 的标准方差明显小于其他特征变量，根据筛选原则，剔除左肩宽度 x_4，峰 R_2 处筛选后的特征变量为 $X=\{x_1, x_2, x_3, x_5, x_6\}$。

(2)自变量间的相关性分析

将初筛后的特征变量输入 matlab 组成矩阵 X，然后输入命令 $R=$corrcoef(X)，计算矩阵 X 的相关系数矩阵 R，计算结果如表 7 和表 8 所示。

根据筛选原则，分析表 7 和表 8 可知，峰 R_1 自变量间的相关系数大于门槛值 0.95 的有 $r_1(x_1, x_2)$，$r_1(x_1, x_3)$，$r_1(x_3, x_4)$，$r_1(x_3, x_5)$，综合考虑其重要性，剔除峰面积 x_1、半高宽 x_3 和左右肩宽比 x_6，则筛选后的特征变量为 $X=\{x_2, x_4, x_5\}$；同理，峰 R_2 自变量间的相关系数大于门槛值 0.95 的只有 $r_2(x_1, x_2)$，综合考虑其重要性，剔除峰面积 x_1 和左右肩宽比 x_6，则筛选后的特征变量为 $X=\{x_2, x_3, x_5\}$。

(3)自变量与因变量间的相关性分析

将上述筛选后的特征变量及因变量(含水量)输入 mat-

lab，计算各自变量与因变量相关系数，计算结果如表 9 和表 10 所示。

表 7 峰 R_1 特征变量间的相关系数

Table 7 Correlation coefficients between characteristic variables at the peak R_1

$r_1(x_i, y_j)$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1	0.998	0.961	0.947	0.934	−0.784
x_2	0.998	1	0.950	0.944	0.918	−0.760
x_3	0.961	0.950	1	0.954	0.988	−0.878
x_4	0.947	0.944	0.954	1	0.895	−0.695
x_5	0.934	0.918	0.988	0.895	1	−0.941
x_6	−0.784	−0.760	−0.878	−0.695	−0.941	1

Note: $i=1, 2, \cdots, 6$; $j=1, 2, \cdots, 6$

表 8 峰 R_2 特征变量间的相关系数

Table 8 Correlation coefficients between characteristic variables at the peak R_2

$r_2(x_i, y_j)$	x_1	x_2	x_3	x_5	x_6
x_1	1	1.000	0.760	0.857	−0.730
x_2	1.000	1	0.750	0.855	−0.734
x_3	0.760	0.750	1	0.847	−0.493
x_5	0.857	0.855	0.847	1	−0.880
x_6	−0.730	−0.734	−0.493	−0.880	1

Note: $i=1, 2, \cdots, 6$; $j=1, 2, \cdots, 6$

表 9 峰 R_1 的因变量与自变量间的相关系数

Table 9 Correlation coefficients between dependent and independent variables at the peak R_1

$r_1(y, x_i)$	x_2	x_4	x_5
y	0.961	0.944	0.951

表 10 峰 R_2 处的因变量与自变量间的相关系数

Table 10 Correlation coefficients between dependent and independent variables at the peak R_2

$r_2(y, x_i)$	x_2	x_3	x_5
y	0.985	0.682	0.862

根据筛选原则，分析表 9 和表 10 可知，峰 R_1 自变量与因变量间的相关系数均大于门槛值 0.85，则筛选后的特征变量为 $X=\{x_2, x_4, x_5\}$；同理，峰 R_2 自变量与因变量间的相关系数小于门槛值 0.85 的有 $r_2(y, x_3)$，故剔除半高宽 x_3，则筛选后的特征变量为 $X=\{x_2, x_5\}$。

综上所述，峰 R_1 筛选后的特征变量为 $X=\{x_2, x_4, x_5\}$，按相关程度由大到小为峰高、右肩宽度、左肩宽度三个特征变量；峰 R_2 筛选后的特征变量为 $X=\{x_2, x_5\}$，按相关程度由大到小为峰高、右肩宽度两个特征变量。

3 回归模型的构建与检验

采用多元线性回归方法建立因变量 Y 与筛选后自变量 X 的数学模型并进行统计检验，对回归方程可信度进行判断。

3.1 峰 R_1 处的回归模型及其检验

应用 matlab 中的 regress 命令，计算峰 R_1 回归模型的各个参数，以及应用 finv 命令计算峰 R_1 回归模型的检验临界值。

经计算，得出峰 R_1 因变量 Y 与自变量 X 的数学回归模型为

$$y = 11.219\ 4 + 36.396\ 3x_2 + 20.344\ 0x_4 + 36.887\ 8x_5 \quad (2)$$

$F=55.417\ 5 > F_{3,7}(0.05)=4.346\ 8$，因此，线性回归效果显著，即 Y 与 X 之间存在线性相关关系，且 $p=0.000\ 0<0.05$，说明岩石含水量与峰 R_1 处的峰高、左肩宽度、右肩宽度之间的多元线性回归方程的模型参数为 0 的原假设是小概率事件，即式(2)的回归方程可以通过检验。

3.2 峰 R_2 处的回归模型及其检验

同理，可得出峰 R_2 因变量 Y 与自变量 X 的数学回归模型为

$$y = 10.397\ 7 + 87.446\ 8x_2 + 5.081\ 9x_5 \quad (3)$$

$F=139.811\ 1 > F_{2,8}(0.05)=4.459\ 0$，因此，线性回归效果显著，即 Y 与 X 之间存在线性相关关系，且 $p=0.000\ 0<0.05$，说明岩石含水量与峰 R_2 处的峰高、右肩宽度之间的多元线性回归方程的模型参数为 0 的原假设是小概率事件，即式(3)的回归方程可以通过检验。

综上所述，岩石含水量与光谱特征之间呈线性相关关系，其中峰 R_1 的峰高、右肩宽度、左肩宽度与含水量线性回归效果较显著，峰 R_2 的峰高、右肩宽度与含水量线性回归效果较显著。采用多元线性回归模型可建立含水岩石近红外光谱特征与含水量之间的因果关系，利用该回归模型可进行基于近红外光谱特征的含水岩石中的含水量预测。

4 结 论

在室内制备了 11 种不同含水量岩石，并分别测量了近红外光谱曲线，对其进行一阶求导之后，分析谱段特征进行初始特征提取，然后进行了光谱特征选择，最后利用最优特征变量构建多元回归模型，并对此模型进行了检验分析，得出如下结论：

(1)波长 1 400 和 1 930 nm 附近的近红外光谱吸收峰特征与岩石含水量有明显相关性；

(2)波长 1 400 nm 处的峰高、右肩宽度、左肩宽度与含水量线性相关性明显；波长 1 930 nm 处的峰高、右肩宽度与含水量线性相关性明显；

(3)采用含水岩石近红外光谱吸收峰特征建立多元线性回归模型，通过线性模型能够较精确表达含水量与近红外光谱之间的相关性，利用该回归模型可进行基于近红外光谱特征的含水岩石中的含水量预测。

致谢：非常感谢为本工作做出贡献的同事们，特别要感谢王东升、李鹏飞、胡臣、谢运鑫等学生参与实验模拟工作。

References

[1] CAI Zhe-yuan, YU Jian-guo, LI Xian-peng, et al(蔡哲元, 余建国, 李先鹏, 等). Pattern Recognition and Artificial Intelligence(模式识别与人工智能), 2010, 23(2): 235.

[2] DOU Gang, CHEN Guang-sheng, ZHAO Peng(窦 刚, 陈广胜, 赵 鹏). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(8): 2425.

[3] Herrera L J, Lafuente V, Ghinea R, et al. Lecture Notes in Engineering & Computer Science, 2015, 2215(1): 7.

[4] ZHAO Jun-yang, ZHANG Zhi-li(赵军阳, 张志利). Computer Application(计算机应用), 2009, 29(1): 109.

[5] Bonev B, Escolano F, Giorgi D, et al. Computer Vision & Image Understanding, 2013, 117(3): 214.

[6] Zheng K, Wang X. Pattern Recognition, 2018, 77: 20.

[7] Cláudia Pascoal, M Rosário Oliveira , António Pacheco, et al. Neurocomputing, 2017, 226(C): 168.

[8] NI Li-jun, ZHANG Li-guo(倪力军, 张立国). Basic Stoichiometry and Its Application(基础化学计量学及其应用). Shanghai: East China University of Science and Technology Press(上海: 华东理工大学出版社), 2011. 7.

Feature Selection of Near-Infrared Spectra of Rock with Different Water Contents

ZHANG Fang[1,2], HU Zuo-le[1,2], HOU Xin-li[3], ZHANG Xiu-lian[1,2], FU Cheng-gong[1,2], LI Ying-jun[1,4], HE Man-chao[1]

1. State Key Laboratory for Geomechanics & Deep Underground Engineering, China University of Mining & Technology, Beijing 100083, China
2. School of Mechanics and Civil Engineering, China University of Mining & Technology, Beijing 100083, China
3. Tianjin TEDA Greening Group Co. Ltd., Tianjin 300457, China
4. School of Science, China University of Mining & Technology, Beijing 100083, China

Abstract Water content of rock is an important index to affect the physical, chemical and mechanical properties of rock. In geotechnical engineering, tunnel engineering and other fields, water content is the key factor to induce disaster and disease. Compared with the traditional method, the determination of rock water content by using the feature of NIR spectrum (NIRS) has obvious advantages of nondestructive and quantitative analysis, and the difficulty and key is the feature selection of NIR spectrum. In order to solve this problem, laboratory experiments were carried out to study the feature selection of near infrared spectra of rock under different water content. The Filter method of feature selection, using the inherent characteristics of the sample data, evaluates the importance of the feature, enhances the correlation between the feature and the class, and reduces the correlation between the features, so it has the advantages of low complexity, being intuitionistic, high efficiency and strong universality and accords with the characteristics of the data studied in this paper. Therefore, this paper selects the Filter type dependency metric for feature selection. In the laboratory experiment, 11 kinds of sandstone samples with different moisture content were prepared, and 44 NIR spectra were collected respectively at 4 test points on the front, behind, left and right sides. Then, the first derivative method was used to preprocess the spectrum. Based on this, the spectral characteristics were analyzed at 1 400 and 1 930 nm, and six initial characteristic variables (the peak area, peak height, width of half height, width of left shoulder, width of right shoulder, the ratio of the width of the left shoulder to the width of the right shoulder)were extracted respectively. Considering the different dimensions and variation range of the six initial characteristic variables, the original data were normalized to eliminate the influence of different dimensions and variation ranges. And then, according to the principle of independent variable selection, redundant variables with strong linear correlation between independent variables were removed. Then, used the statistical correlation coefficient in the dependency metric as the measure of correlation degree, and the correlation among the initial characteristic variables and the correlation between the initial characteristic variables and water content were analyzed. The optimal characteristic variables at two strongly correlated spectral segments were obtained. Finally, multiple regression models were constructed at the strong correlation spectral segments, and the models were tested and analyzed. The results showed that: (1) the characteristics of the near-infrared spectral absorption peaks around the wavelengths of 1 400 and 1 930 nm are significantly correlated with the rock water content; (2) the peak height, right half width and left half width at the wavelength of 1 400 nm have linear correlation with the water content obviously, and the peak height and right half width at the

wavelength of 1 930 nm also have linear correlation with the water content obviously; (3) the multiple linear regression model can accurately express the correlation between the water content and the near-infrared spectrum, and the model can be used to predict the water content of water-bearing rock based on the characteristics of near-infrared spectrum. It provides basic modeling data for dynamic monitoring and evaluation of rock water content by using near infrared spectrum analysis technology.

Keywords Water-bearing rock; Near infrared spectroscopy; Feature selection; Correlation; Water centent prediction

(Received Sep. 12, 2018; accepted Feb. 7, 2019)

关于《光谱学与光谱分析》调整审稿费收费标准的通知

尊敬的《光谱学与光谱分析》广大作者、读者：我刊自 2018 年 7 月 1 日以后登记的稿件向投稿作者收取审稿费 200 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。

2. 没有国家级基金资助的稿件，请您不要投稿。

3. 不是光谱专业的稿件，请您不要投稿。

4. 与其他文章重合率超过 10%的稿件，请您不要投稿。

所投稿件经初审通过后，作者会收到缴纳审稿费的通知。请作者及时从我刊网站(http://www.gpxygpfx.com)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，我刊不再受理。交费后我刊开据增值税电子普通发票，并传至作者提供的电子邮箱，作者可自行打印。

联系电话：010-62181070，62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2018 年 6 月 30 日