基于互信息的含水岩石近红外光谱特征选择

张秀莲[1,2]，张 芳[1,2*]，周 暖[1,2]，张靖婕[1,2]，刘文芳[3]，张 帅[1,2]，杨晓杰[1,2]

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. 东北大学资源与土木工程学院，辽宁 沈阳 110004

摘 要 传统的相关分析方法无法准确刻画含水岩石的近红外光谱与其含水量之间的非线性关系。针对这个问题，首先进行了莫高窟崖壁砾岩水分运移的室内试验，分别采集了砾岩样品3个不同位置从初始干燥状态到饱和状态的全过程，共计51条近红外光谱信息；然后采用多点平滑与基线校正相结合(NPS+B-corr)的方法对原始近红外光谱进行预处理，根据强吸收谱段1 450和1 950 nm处的光谱曲线特征提取峰高，半高宽，峰面积，左肩宽度，右肩宽度，左右肩比共6个初始特征变量建立初始特征集，同时对所提取的光谱特征变量进行归一化处理，根据处理之后的结果绘制各光谱特征参数与含水量变化的曲线，确定含水量级别；接着，进行初始特征集各光谱特征变量间相关性筛选，以便去掉冗余特征，将初始特征集简化为，即由峰高(Height)，左肩宽度(LHW)，右肩宽度(RHW)三个特征变量构成特征集；最后利用互信息作为相关程度的度量标准，分别采用BIF(best individual feature)法和MIC(maximal information coefficient)法，研究了光谱特征变量与含水量级别之间依赖关系的强弱程度，结果表明：(1)砾岩的近红外光谱在波长1 450和1 930 nm附近有明显的吸收峰，且吸收峰随着含水量的变化表现出较强的关联变化，说明岩石近红外光谱反射率与岩石含水量有着明显的相关性；(2)初选光谱特征变量与岩石含水量的动态规律曲线呈S形，含水量可划分为干燥、吸水、饱和状态三个级别；(3)两种信息法选取的近红外光谱特征不完全一致，基于BIF法，波长1 450 nm处特征变量与岩石含水量级别之间的相关性从高到低排序为右肩宽度，峰高，左肩宽度；1 900 nm处为峰高，右肩宽度，左肩宽度。基于MIC法，1 450和1 900 nm处的特征变量与岩石含水量级别之间的相关性从高到低排序均为左肩宽度，峰高，右肩宽度。(4)利用决策树评估MIC和BIF法的有效性，MIC法比BIF法对含水量级别的识别精度更高。

关键词 含水岩石；近红外光谱；互信息；水分运移

中图分类号：TU458 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)07-2028-08

引 言

近红外光谱分析(near infrared spectroscopy，NIRS)具有无损、环保、快速、准确的特点，已被广泛应用于农业、食品、医药、生物化工等领域中[1]。近几十年来，利用该技术以检测岩质工程中的水分含量已成为研究的热点。David等[2]测量了四种不同土壤在不同含水量下的近红外光谱，发现近红外光谱与土壤体积含水量具有良好的相关性。Mouazen等[3]采用近红外仪测量了含水率为0.5%～26%的土壤样品的光谱特征，并用偏最小二乘分析法建立了光谱与土壤含水率的定量反演模型，并成功地对土壤样品的光谱进行了含水量的定量检测及其分类。

水的近红外光谱信息主要是由其物质分子中的O—H基团的振动吸收引起的[4]，其在近红外光谱上有两个比较宽的吸收峰，分别在O—H伸缩振动的一级倍频6 944 cm^{-1}(1 440 nm)和水分子的组合频5 154 cm^{-1}(1 940 nm)附近，岩石样品中的水分含量通过这些特征吸收就可以来测定。水的O—H谱带位置和宽度随氢键形成程度的变化而改变。含水岩石中的O—H基团发生振动吸收，吸收量与O—H基团数量成正比，可见近红外光谱吸收量与岩石含水量有密切关系。

收稿日期：2020-05-26，**修订日期**：2020-09-12

基金项目：国家自然科学基金青年科学基金项目(51604276)资助

作者简介：张秀莲，1985年生，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博士研究生
e-mail：xiulian1985@163.com *通讯作者 e-mail：zhangf76@163.com

不同含水量下的岩石近红外光谱反射率不同，通过构建岩石的近红外光谱特征与其含水量之间的定量关系，可以检测岩石含水量。光谱的特征选择是构建该定量关系的关键和难点，目前近红外光谱特征选择常用的方法主要有主成分分析法、多元线性回归、偏最小二乘法、遗产算法以及基于互信息的特征选择方法，其中前四种方法多用评价特征变量与目标属性变量之间的线性相关关系，主要应用于预测生物柴油燃料性能、烟草质量评价、土壤有机物等方面[5]。而基于互信息的特征选择方法属于信息论的相关度量法，可描述变量之间非线性相关关系。Vinh L T 等[6]通过对最大相关最小冗余方法(mRMR)中互信息度量进行规范化处理，以消除相关性和冗余度的影响，提出了基于归一化互信息的特征选择方法，并通过识别模型对此方法与原 mRMR 方法进行比较，结果表明基于归一化互信息的特征方法在分类准确性方面优于传统方法。由此可见，基于互信息的特征选择方法在变量之间的相关性评价方面具有明显的优势，且目前主要应用于人工智能、农业、医药等领域，还未见含水岩石近红外光谱特征选择方面的有关报道。

利用近红外光谱分析仪监测岩石水分动态运移过程，并测得其近红外光谱信息。整个过程随着时间变化，岩石含水量也处在动态变化中，易知，相应的近红外光谱在与 O—H 基团相关的谱段内也发生了变化。因此，采集岩石从干燥状态到饱和状态全过程的光谱特征曲线，提取特征谱段处不同光谱特征参数，采用互信息作为光谱特征参数之间相关程度的度量标准，分析出岩石材料中的 O—H 基团数量变化与光谱特征参数之间的相关程度，以期得到近红外光谱特征参数与岩石含水量之间的相关性以及相关程度大小。

1　特征选择理论基础

1.1　特征选择过程

设初始数据集记作 $T=(O, F, C)$，其中 $O=\{o_1, o_2, \cdots, o_n\}$为近红外光谱样本数据的集合，$o_i$ 表示某含水量岩石的近红外光谱，$F=\{f_1, f_2, \cdots, f_m\}$为初选的近红外光谱特征矩阵，存储提取的特征变量，$f_k$ 为某特征变量，$C=\{c_1, c_2, \cdots, c_l\}$为岩石含水量级别的集合，$c_j$ 表示某含水量级别(某一小区间)。

特征选择过程由产生初始特征子集、搜索策略、评价函数、终止条件四个步骤组成。本工作采用互信息作为相关程度的评价标准，对特征变量与含水量级别之间依赖关系的强弱程度进行度量，并根据其值按降序排序，然后，选择前 k 个特征(或一致性因子等其他终止条件)组成选择子集 S。

1.2　基于互信息的特征选择

信息度量法具有无参、非线性的优势，且不需要预先知道样本数据的分布，在特征选择算法中得到广泛应用。

1.2.1　BIF 法

BIF(best individual feature)法[7]直接利用互信息作为评价标准，是一种简单而直观的信息度量方法，利用该方法度量含水量级别 C 和某特征变量 f 的相关性，其评价标准可表示为

$$I(C; F)=\sum_{c\in C}\sum_{f\in F}p(c, f)\log\frac{p(c, f)}{p(c)p(f)} \tag{1}$$

其中，$p(c)$和 $p(f)$分别为含水量级别 c 和某特征值 f 的边缘概率分布；$p(c, f)$为含水量级别 C 与某特征值 f 的联合概率分布。

由此定义式可知，当含水量级别 C 和某特征值 f 完全无关或相互独立时，互信息 $I(C; f)=0$；二者间相互依赖程度越高，互信息 $I(C; f)$的值就越大。可见，BIF 选择算法计算步骤首先按式(1)分别计算特征矩阵 $F=\{f_1, f_2, \cdots, f_m\}$中的特征 f 与含水量级别 C 的互信息 $I(C; f)$，然后根据互信息值大小按降序顺序排序，最后选择前 k 个特征组成近红外光谱特征集 S，完成特征选择。

1.2.2　MIC 法

最大信息系数(maximal information coefficient，MIC)由 Resher[8]于 2011 年首次提出，他利用互信息定义了两个变量之间的最大信息系数，并利用该系数衡量两个变量之间的相关性。利用该方法度量含水量级别 C 和特征变量 f 相关性的具体步骤是：首先对含水量级别 C 与特征变量 f 的散点图进行网格划分，在固定行数 x_i、列数 y_i 情况下，有多种划分方式，对应着不同的网格 $G_{x_iy_i}(j)$。然后，用各散点在网格的子格内的频率来代替变量的概率，计算含水量级别 C 与特征变量 f 之间的互信息，并取其最大值，记为最大互信息

$$I_{\max}(C, f, x_i, y_i)=\max I((C, f)|_{G_{x_iy_i}(j)}),\ j=1, 2, \cdots \tag{2}$$

其中，$I_{\max}(C, f, x_i, y_i)$表示在固定行列数 x_i 与 y_i 情况下，不同的网格划分方式下的互信息最大值(最大互信息)；$I((C, f)|_{G_{x_iy_i}(j)})$表示含水量级别 C 与特征变量 f 的散点图在网格 $G_{x_iy_i}(j)$ 下的互信息。

最后，对于任意的行数 x_i、列数 y_i，$i=1, 2, \cdots$，由式(2)，计算最大信息系数

$$\mathrm{MIC}=\max\left\{\frac{I_{\max}(C, f, x_i, y_i)}{\log\min(x_i, y_i)}\right\} \tag{3}$$

其中 $x_i, y_i<B(n)$，$B(n)$为网格分割细度[9]。

以最大信息系数为评价标准时，每个特征变量的 MIC 值视为其权重，MIC 值越大，表明该特征变量 f 与含水量级别 C 之间的相关性越大。最后选择满足预先给定的阈值的特征变量，组成近红外光谱特征集 S，完成特征的选择。

2　实验部分

2.1　样品制备

砾岩岩样采自敦煌莫高窟北区崖壁，基本信息如表 1 所示，其中干燥后质量为砾岩样品在设定温度为 105～110 ℃真空干燥箱里烘干 24 h 后冷却至室温所测得质量。为了进一步了解砾岩所含矿物成分，采用日本理学电机公司(Rigaku)生产的 D/MAX2500 X 射线衍射仪对砾岩样品进行 X 射线衍射实验，其矿物成分分析结果见表 2 所示。因为砾岩结构疏松，呈半胶结状态，遇水易崩解，不易加工成标准试件，将其加工成尺寸约为 80 mm×90 mm×60 mm 的不规

表 1 试样基本信息表

Table 1 Basic information of the conglomerate samples

岩性	外貌描述	颜色	干燥后质量/g	尺寸/mm
砾岩	疏松多孔，块状结构	青灰色	922.60	约为 80×90×60

表 2 砾岩矿物成分分析结果

Table 2 Mineral composition of the conglomerate samples

岩性	石英/%	钾长石/%	钠长石/%	方解石/%	白云石/%	角闪石/%	石盐/%	黏土矿物总量/%
砾岩	30.4	2.2	23.5	11.8	9.7	2.0	6.9	13.5

则形状[图 2(a)]，以供实验所需。

2.2 仪器

利用中国矿业大学(北京)深部岩土力学与地下工程国家重点实验室何满潮[10]自主研发的“深部软岩水理作用智能测试系统”实现岩样吸水模拟。该系统主要由主体实验箱、称重系统和数据采集三部分组成，同时采用由瑞士万通(Metrohm)生产的 NIRS XDS OptiProbe Analyzer(近红外光谱分析仪)采集岩样的近红外光谱信息(实验参数见表 3)，该近红外仪器由光谱分析仪、电控平移操作台以及计算机组成，其中光谱分析仪可精确快速地捕捉波长在 400～2 500 nm 范围的光谱信号，光纤探头可采集样品固定位置处的反射光谱信号；电控平移操作台可实现 0～1 500 mm 行程内的精确定位，确保了不同岩石样品之间或同一试样不同时刻取样点位置的一致性。实验测试装置和测试原理如图 1 所示。

表 3 近红外光谱仪采集系统的实验参数

Table 3 Experimental parameters of near infrared spectrometer acquisition system

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
测样角度	90°
测试方式	直接与样品接触

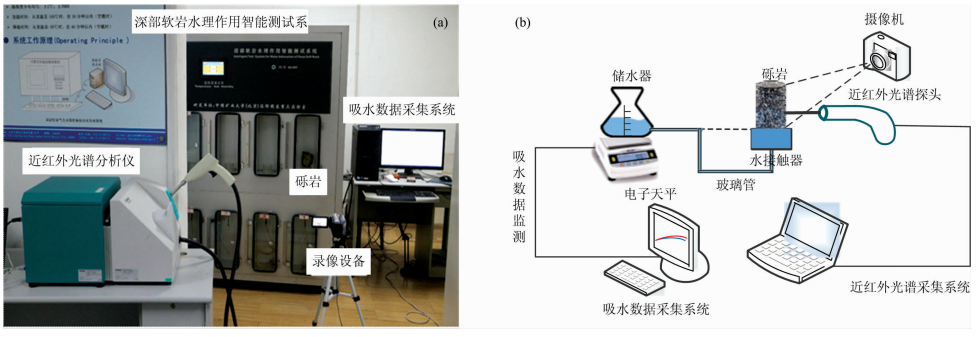

图 1 (a)吸水实验系统及近红外测试装置；(b)实验原理示意图

Fig. 1 (a) Water sorption test system and near-infrared testing device; (b) schematic diagram of the experimental setup

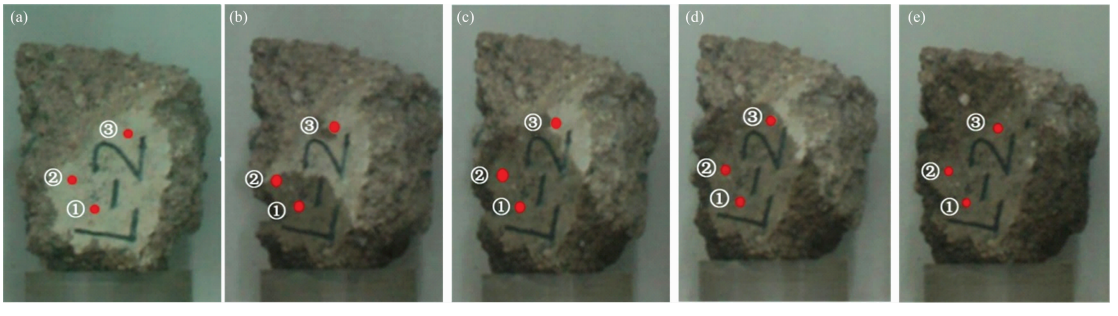

图 2 砾岩吸水过程可见光图

(a)：409 s；(b)：558 s；(c)：704 s；(d)：775 s；(e)：1 224 s

Fig. 2 Visible light map of the conglomerate water absorption process

(a)：409 s；(b)：558 s；(c)：704 s；(d)：775 s；(e)：1 224 s

2.3 方法与光谱采集

考虑到岩石材料的不均匀性，沿着砾岩高度方向分别采集了样品 3 个不同位置(图 2 所示红点位置)，共 51 条近红外光谱信息，这些光谱涵盖了砾岩 3 个测试点从初始干燥状

态到饱和状态的全过程，表征了砾岩材料中 O—H 基团数量变化与光谱特征参数之间的相关程度。将干燥后的砾岩岩样放置于图 1 所示位置，岩样开始吸水。在岩样吸水过程中，采用近红外光谱分析仪采集岩样的近红外光谱。测试时将光纤探头分别接触岩样的 3 个测量点，自下而上依次测量，测量的频率随岩石的吸水速度适时调整。

整个砾岩样品历时 1.15 h 从干燥状态达到饱和状态。在 1，2 和 3 号点分别采集近红外光谱曲线 17 条、18 条、16 条，分别记作 O1-1，O1-2，…，O1-17；O2-1，O2-2，…，O2-18；O3-1，O3-2，…，O1-16，共 51 条近红外光谱。不同采集点位置达到吸水饱和状态的时间是不一样的，整个实验过程 3 个采集点位置从干燥到饱和状态所用的时间分别为 2 998，3 048和3 099 s。

3　结果与讨论

3.1　光谱特征分析和谱段选择

3.1.1　光谱预处理

通过对比常用预处理方法的处理效果，选择多点平滑与基线校正相结合的方法(NPS+B-corr)对原始近红外光谱进行预处理，既去除了无效噪音，又校正了基线漂移；限于篇幅，文中仅列出砾岩 1 号点处的近红外光谱图，如图 3 所示，其中(a)为砾岩原始近红外光谱图，(b)为经过 NPS+B-corr 法预处理得到的砾岩近红外光谱图。

3.1.2　谱段选择

由图 3 可知，在 400～2 500 nm 波长范围内分别在 1 450，1 930 和 2 300 nm 附近有明显的吸收峰，依次命名为峰 R_1、峰 R_2、峰 R_3；光谱的吸光度随岩石含水量变化明显。

分析 1 号点光谱曲线(图 3)可知，R_1 和 R_2 两个吸收峰的波峰随含水量的增大而增高，峰顶中心位置逐渐右移，当含水量达 2.467%之后峰形变化越来越小，峰值波动减弱，最后峰 R_1 的中心点位于 1 450 nm 左右，峰 R_2 中心点位于 1 930 nm 左右。随着含水量的增加，峰 R_1、峰 R_2 的光谱信号越来越强，峰 R_3 的信号则不断减弱，加上临近噪音波段信号干扰，峰 R_3 的特征信号越来越难提取。综上分析，本文主要分析含水岩石在峰 R_1(1 450 nm)、峰 R_2(1 930 nm)谱段处的光谱特征。

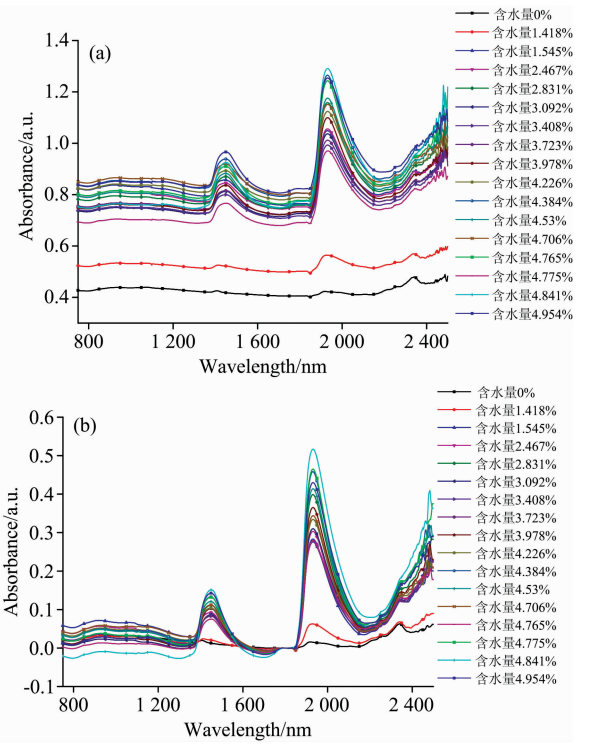

图 3　砾岩 1 号点处近红外光谱

(a)：原始光谱；(b)：经 NPS+B-corr 预处理后光谱

Fig. 3　Near-infrared spectrum of conglomerate point 1

(a)：Original spectrum；(b)：NPS+B-corr pre-processed spectrum

3.2　基于互信息的特征选择

3.2.1　初始特征集 F 的建立

构建特征集是建立岩石含水量与近红外光谱之间的数学模型中至关重要的一环，它是后续特征选择的基础，决定了特征选择的范围和质量。

分析经 NPS+B-corr 法预处理后的近红外光谱图的峰 R_1、峰 R_2 特点，提取下列特征变量：峰高(Height)，半高宽(FWHM)，峰面积(Area)，左肩宽度(Left Half Width，LHW)，右肩宽度(Right Half Width，RHW)，左右肩比(LHW/RHW)共 6 个初始特征变量，其几何示意图见图 4，因此设定的初始特征集 $F=\{f_1, f_2, f_3, f_4, f_5, f_6\}=$ {Height，Area，FWHM，LHW，RHW，LHW/RHW}，其具体数值如表 4 所示。限于篇幅，此处只列出 1 450 nm 谱段处部分数据。

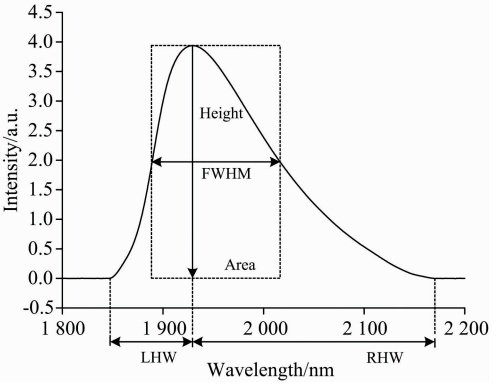

图 4　初始特征变量示意图

Fig. 4　Schematic diagram of initial characteristic variables

3.2.2　特征变量归一化

由表 4 可以看出 6 个初始特征变量的量纲不同，且变量之间的变化幅度也不同，因此需要对原始数据进行归一化处理，来消除量纲和变化幅度不同带来的影响。

归一化的方法是将原式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(详见本刊 40 卷 3 期 971 页)，有

$$x'_{ij}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \tag{4}$$

归一化结果如表 5 所示。

表 4 1 450 nm 谱段近红外光谱的初始特征变量(只列出部分)

Table 4 Initial characteristic variables of near-infrared spectra at 1 450 nm (only part of data listed)

特征值 F	f_1 /nm	f_2 /nm	f_3 /nm	f_4 /nm	f_5 /nm	f_6
O1-1	0.006	36.702	0.294	17.641	19.061	0.925
O1-2	0.014	103.242	1.628	21.971	81.271	0.270
O1-3	0.071	118.620	9.1463	52.161	66.460	0.784 8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
O3-16	0.157	119.694	20.562	53.409	66.285	0.806
均值	0.079	110.146	10.439	47.652	62.494	0.770
标准方差	0.041	26.540	5.512	12.990	14.850	0.120
最大值	0.177	123.649	23.952	54.434	81.271	0.938
最小值	0.005	36.032	0.279	16.305	18.589	0.270

3.2.3 含水量级别 C 的确定

含水量级别 $C=\{c_1, c_2, \cdots, c_l\}$ 为岩石含水量的级别集合，c_j 表示某含水量级别。实验中砾岩不断动态吸水，根据总吸水量与测量时得到的近红外光谱对应关系，绘制各特征变量随含水量变化曲线，这里仅列出峰高随含水量的变化曲线，如图 5 所示。

分析图 5 可知，含水量与各光谱特征参数的关系曲线有相似的变化趋势，曲线大致呈“S”型，可将其划分为三个阶段：低含水量时特征参数平稳变化，中间段特征参数迅速增长，高含水量时特征参数基本稳定。当高含水量达到一定程度时(图 5 中各点含水量超过 5%)，特征参数有所增加，这一现象可以解释为：在岩石自下而上动态吸水的过程中，虽然监测点的含水量已经达到饱和，但水分仍然在持续不断地路经监测点，向上吸附至试件上端未饱和区域，由此，最终在测试点表面形成一层水分薄膜，使得该处岩石表面具有较高反射率，对应的光谱特征参数出现进一步增加的现象。

结合吸水过程分析可知，这三个阶段水分尚未到达测试点，水分接触到测试点并开始浸润该位置，最终水分使测试点区域内岩石吸水饱和。测试点吸水过程是短暂迅速的，利用近红外光谱分析技术可以很精确的对岩石上各点的含水量大小进行监测识别。可见，测试点的光谱特征随含水量变化规律，有着非常明显的三个阶段，对应着各测试点的三个含水量级别，据此可设置含水量级别 C：$C=\{c_1, c_2, \cdots, c_l\}=$ {干燥状态，吸水状态，饱和状态}，该级别属于文本类别，为了计算最大相关系数 MIC 评价变量间相关程度，将其定义为：$C=\{c_1, c_2, c_3\}=$ {干燥状态，吸水状态，饱和状态}$=$ {1，2，3}。

表 5 1 450 nm 谱段近红外光谱的初始特征变量归一化值(只列出部分)

Table 5 Normalized initial characteristic variables of near-infrared spectra at 1 450 nm (only part of data listed)

特征值 F	f_1/nm	f_2/nm	f_3/nm	f_4/nm	f_5/nm	f_6
O1-1	0.005 8	0.007 6	0.000 6	0.035 0	0.007 5	0.980 5
O1-2	0.052 3	0.767 1	0.057 0	0.148 6	1.000 0	0.000 0
O1-3	0.383 7	0.942 6	0.374 6	0.940 4	0.763 7	0.770 7
⋮	⋮	⋮	⋮	⋮	⋮	⋮
O3-16	0.883 7	0.954 9	0.856 8	0.973 1	0.760 9	0.802 4
均值	0.432 6	0.845 9	0.429 2	0.822 1	0.700 4	0.747 9
标准方差	0.040 6	26.540 2	5.511 8	12.989 6	14.850 3	0.120 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

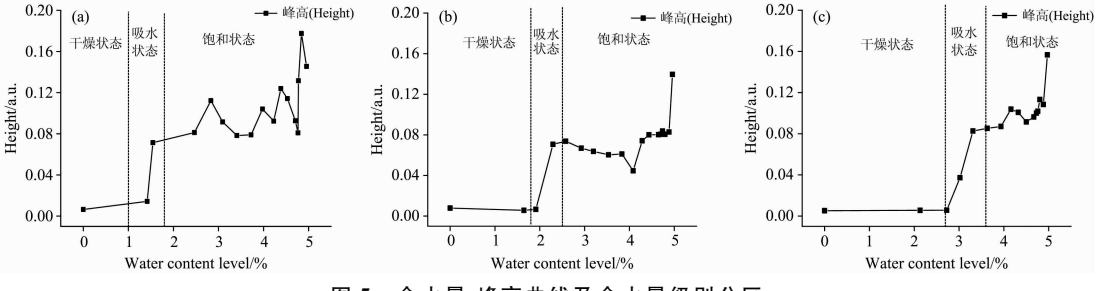

图 5 含水量-峰高曲线及含水量级别分区

(a)：1 号点；(b)：2 号点；(c)：3 号点

Fig. 5 Peak height of water content and water content level partition

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

3.2.4 特征选择计算与结果分析

在进行特征选择之前，首先要进行初始特征集各特征变量间相关性筛选，以便去掉冗余特征，根据文献[11]中计算方法，进行特征变量筛选。将初始特征集 F 简化为 $S=\{f_1, f_4, f_5\}$，即由峰高(Height)，左肩宽度(LHW)，右肩宽度(RHW)三个特征变量构成特征集。

(1)BIF 法的特征选择

利用式(1)计算特征 $S=\{f_1, f_4, f_5\}$与含水量级别 $C=\{c_1, c_2, c_3\}$之间的互信息 $I(C, f)$。首先根据 3.2.2 小节中光谱特征变量归一化之后的计算结果，将各特征变量值按照[max, min]进行分组，并计算每组所得的概率，然后再根据 3.2.3 小节含水量级别 C 的分类结果分别计算每一级别所占的概率，最后根据各特征变量和含水量级别的分组结果计算两者的联合分布概率，再代入到 1.2.1 小节中的公式，具体计算结果见表 6。

特征变量与含水量级别之间的互信息值越大，表示二者之间的相关性越大，分析表 6 可知，按互信息 $I(C, f)$由大到小排序，则峰 R1 的特征变量排序结果为 $I(C, \text{RHW})>I(C, \text{Height})>I(C, \text{LHW})$，可知相关性由高到低为右肩宽度，峰高，左肩宽度。同理峰 R_2 的特征变量排序结果为 $I(C, \text{Height})>I(C, \text{RHW})>I(C, \text{LHW})$，相关性由高到低为峰高，右肩宽度，左肩宽度。

根据 BIF 法得出特征变量的重要性排序，据此确定最终的近红外光谱特征选择，可以选择相关性最大的特征，也可以选择相关性较大的多个特征的组合作为最优特征子集。

表 6　特征变量与含水量级别的互信息值

Table 6　Mutual information values of characteristic variables and water content levels

吸收峰	$I(C, \text{Height})$	$I(C, \text{LHW})$	$I(C, \text{RHW})$
峰 R_1	0.618 8	0.553 8	0.675 3
峰 R_2	0.584 0	0.529 8	0.530 6

(2)MIC 法的特征选择

基于第 1.2.2 节中的计算方法，首先将特征 f 与含水量级别 $C=\{$干燥状态，吸水状态，饱和状态$\}=\{1, 2, 3\}$组成散点图 D，以峰 R_1 为例，各特征变量与含水量级别 C 之间的散点图 D_i 如图 6 所示，然后依次计算不同网格划分方式 $G_{x_i y_i}(j)$下最大互信息值 $I_{\max}(C, f, x_i, y_i)$，接着组成最大互信息标准化矩阵 $\left\{\frac{I_{\max}(C, f, x_i, y_i)}{\log\min(x_i, y_i)}\right\}$，再取这个矩阵的最大值，确定最大相关系数 $\text{MIC}=\max\left\{\frac{I_{\max}(C, f, x_i, y_i)}{\log\min(x_i, y_i)}\right\}$，计算结果如表 7 所示。

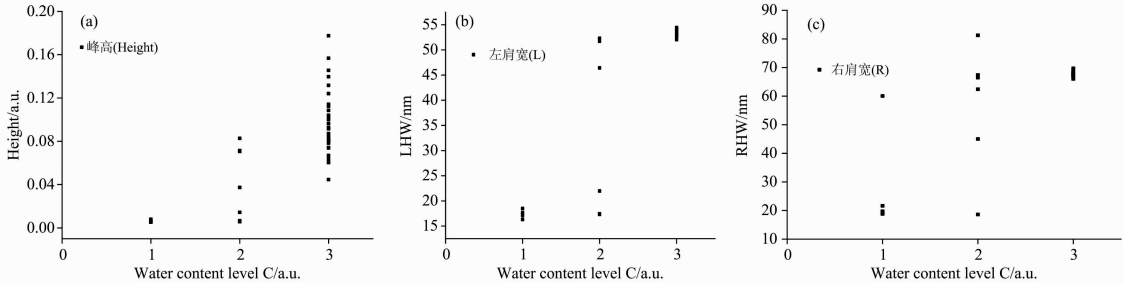

图 6　含水量级别 C-特征 f 散点图

(a)：Height-C 散点图；(b)：LHW-C 散点图；(c)：RHW-C 散点图

Fig. 6　Water content level C-feature f scatter plot

(a)：Height-C；(b)：LHW-C；(c)：RHW-C

表 7　特征变量与含水量级别间的 MIC 值

Table 7　MIC values for correlation between the characteristic variables and water content levels

吸收峰	MIC(C, Height)	MIC(C, LHW)	MIC(C, RHW)
峰 R_1	0.692 3	0.787 1	0.609 5
峰 R_2	0.716 3	0.787 1	0.370 6

由表 7 可知，峰 R_1 各特征变量 f 与含水量级别 C 之间的 MIC 值，由大到小排序为 MIC(C, LHW)＞MIC(C, Height)＞MIC(C, RHW)，其中任一特征值对应的 MIC 值均＞0.5，说明通过峰分析提取的特征变量，随着岩石含水量的变化，均会发生一定程度的变化，特征值 f 与岩样吸水状态 C 之间存在着一定的相关性，特别地 LHW 的 MIC 值接近于 0.8，说明其与岩石含水量存在较强的相关性，可作为有效表征岩石含水量的特征变量。综上，根据 MIC 法得到特征变量排序依次为左肩宽度，峰高，右肩宽度。

同理，对峰 R_2 的特征值 f 进行最优选择，MIC 值由大到小排序为 MIC(C, LHW)＞MIC(C, Height)＞MIC(C, RHW)，除特征变量右肩宽 RHW 外，任一特征变量对应的 MIC 值均＞0.5，其中，LHW 的 MIC 值接近于 0.8，表明其与岩石含水量有较强的相关性。综上，依 MIC 值为权重进行特征值重要性排序由高到低依次为左肩宽度，峰高，右肩宽度。

根据 MIC 法得出特征变量的重要性排序及其相关权重，据此可以选择相关性最大的特征变量，也可以选择相关性较强的多个特征变量的组合作为最有特征子集。

3.2.5　BIF 与 MIC 方法性能评估

为了评估 BIF 和 MIC 法的性能，我们使用决策树来评估所选特征的有效性。根据所采集到的 3 个点共 51 条近红外光谱特征，利用 BIF 法和 MIC 法分别计算峰 R_1(1 450 nm)和峰 R_2(1 930 nm)处 6 个特征变量的平均值，为了计算方便，我们对含水量取插值平均得到决策树中的 k 值，通过软件 matlab 导入决策树数据流，计算各级别的平均分类准确度。从图 7 看出，MIC 的准确度较高。

基于互信息的特征选择算法的评价标准是属于信息论的相关度量，主要用于描述两个随机变量之间相互依存关系的

强弱。采用 BIF 方法和 MIC 方法对含水岩石近红外光谱进行特征选择，均得到了近红外光谱特征参数与岩石含水量级别之间的相关程度强弱，且各个特征参数具有明显几何意义，通过 BIF 法与 MIC 法的特征选择结果，后期在利用近红外光谱特征建立砾岩含水量检测模型中，可以将峰高、左肩宽、右肩宽三个特征参数直接参与模型的构建中。总体来说互信息方法思路简单，概念清晰，计算过程简单，利用此方法获得的与含水岩石水分信息最相关的最优光谱特征，可以为建立更稳定、精确、高效的预测模型做基础。

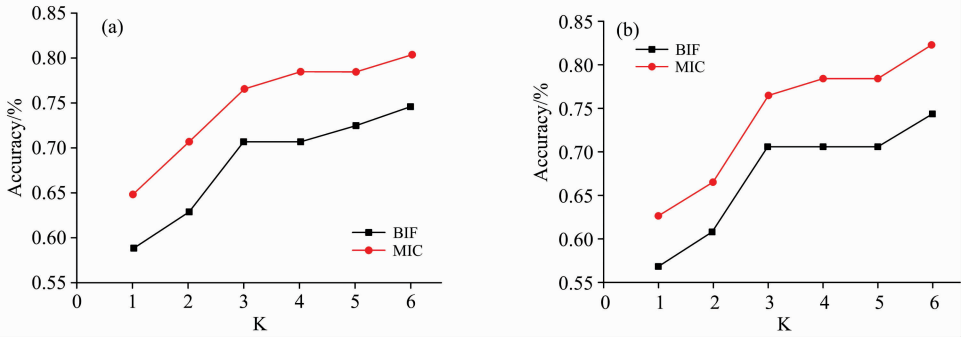

图 7 峰 R_1(a)和峰 R_2(b)在决策树分类器下的准确率

Fig. 7 The accuracies of the decision tree classifier for peak R_1(a) and peak R_2(b)

此外两种方法在评价含水岩石近红外光谱特征变量与含水量级别相关程度强弱的结果并不完全一致，而且 MIC 方法比 BIF 方法对含水量级别的识别精度更高些。其实两种方法都是通过计算变量间的概率来获得最优相关子集，所不同的是 BIF 法直接计算两个变量的边缘概率和联合概率分布，而 MIC 方法先对变量间通过不同网格划分的形式，用变量在网格内出现的频率来代替概率，然后用不同网格划分方式下的互信息最大值计算最大相关系数，MIC 方法所获得的样本数量更为精确，进一步考虑了所选变量间的相关性，因此计算结果更为准确。本文目前只针对砾岩一种岩性进行研究，对于 MIC 方法的准确性还需要在更多的样本集上来进行验证，同时基于上述两种特征选择算法，得出特征变量重要性排序，关于最优特征子集中变量类型与个数的确定，需要后期构建预测模型的精度来佐证，是一个反复验证与选择的过程。

4 结 论

利用互信息理论分析水分运移过程的岩石近红外光谱特征，并进行了光谱特征选择，得出以下结论：

(1)含水岩石的近红外光谱在 1 450 和 1 900 nm 附近有 2 处明显的吸收峰，且吸收峰随着含水量变化，表现出较强的关联变化，可作为分析光谱特征的基本谱段；

(2)根据各特征变量与岩石含水量特征曲线呈“S”型变化的特点，确定岩石含水量级别 $C=\{c_1, c_2, \cdots, c_l\}=\{$干燥状态，吸水状态，饱和状态$\}$；

(3)基于 BIF 法，以互信息为权重，评价特征变量与岩石含水量级别之间的相关性，峰 R_1 处相关性从高到低排序为右肩宽度，峰高，左肩宽度。峰 R_2 为峰高，右肩宽度，左肩宽度。

(4)基于 MIC 法，以最大相关系数 MIC 值作为权重，评价特征变量与岩石含水量级别之间的相关性，峰 R_1、峰 R_2 处相关性从高到低排序均为左肩宽度，峰高，右肩宽度。

(5)利用决策树评估 MIC 和 BIF 法的有效性，MIC 法比 BIF 法对含水量级别的识别精度更高。

References

[1] Cen H, He Y. Trends in Food Science & Technology, 2007. 18(2): 83.
[2] David B L, Gregory P A. Soil Science Society of America Journal, 2002, 66: 722.
[3] Mouazen A M, Karoui R, De Baerdemaeker J, et al. Soil Science Society of America Journal, 2006, 70(4): 1295.
[4] CHU Xiao-li(褚小立). Molecular Spectroscopy Analytical Technology Combined With Chemometrics and Its Applications(结合化学计量学的分子光谱分析技术及其应用). Beijing: Chemical Industry Press(北京：化学工业出版社), 2011. 264.
[5] Balabin R M, Smirnov S V. Analytica Chimica Acta, 2011, 692(1-2): 63.
[6] Vinh L T, Lee S, Park Y T, et al. Applied Intelligence, 2012, 37(1): 100.
[7] YAO Xu, WANG Xiao-dan, ZHANG Yu-xi, et al(姚 旭，王晓丹，张玉玺，等). Control and Decision(控制与决策), 2012, 27(2): 161.
[8] Reshef D N, Reshef Y A, Finucane H K, et al. Science, 2011, 334(6062): 1518.
[9] LIANG Ji-ye, FENG Chen-jiao, SONG Peng(梁吉业，冯晨娇，宋 鹏). Chinese Journal of Computers(计算机学报), 2016, 39(1): 1.

[10] HE Man-chao, ZHANG Guo-feng, ZHAO Jian(何满潮, 张国锋, 赵 健). Chinese Patent(中国专利): CN 102253181, 2011.
[11] ZHANG Fang, HU Zuo-le, HOU Xin-li, et al(张 芳, 户佐乐, 侯欣莉, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(11): 3395.

Near-Infrared Spectral Feature Selection of Water-Bearing Rocks Based on Mutual Information

ZHANG Xiu-lian[1,2], ZHANG Fang[1,2*], ZHOU Nuan[1,2], ZHANG Jing-jie[1,2], LIU Wen-fang[3], ZHANG Shuai[1,2], YANG Xiao-jie[1,2]

1. State Key Laboratory for Geomechanics and Deep Underground Engineering, China University of Mining and Technology, Beijing 100083, China
2. School of Mechanics and Civil Engineering, China University of Mining and Technology, Beijing 100083, China
3. College of Resources and Civil Engineering, Northeastern University, Shenyang 110004, China

Abstract The relationship between near-infrared spectroscopic measurements of rock and its water content does not follow simple linear correlations, preventing the direct use of classical correlation analysis. In the present paper, an experiment on the water migration in cliff conglomerates from the Mogao Grottoes was performed, and collected 51 pieces of near-infrared spectra from three different positions sample. These spectra cover the whole process of the conglomerate from the initial dry state to the saturated state; then we selected a combined N point smooth and baseline correction method (NPS+B-corr) to preprocess the original near-infrared spectrum. According to the spectral curve features at 1 450 and 1 950 nm of the strong absorption spectrum, six initial feature variables, namely Height, Full Width at Half Maximum (FWHM), Area, Left Half Width (LHW), Right Half Width (RHW), and (LHW/RHW), were extracted to establish the initial feature set. Simultaneously, the extracted spectral characteristic variables were normalized, and the curve of each spectral characteristic parameter and the change of water content were drawn according to the result of the processing, determine the water content level. Then, the correlation among the feature variables of the initial feature set should be screened to remove redundant features. The initial feature set is simplified to three characteristic variables: Height, LHW, RHW. Finally, based on mutual information, the Best Individual Feature and Maximal Information Coefficient methods were used to evaluate the relationship between samples' spectral characteristic parameters and water content. We found that: (1) at wavelengths between approximately 1 450 and 1 930 nm, the near-infrared spectrum of the conglomerate has obvious absorption peaks, and the absorption peaks show a strong correlation with the change of water content, which indicates that spectral reflectance was significantly correlated with water content; (2) the relationship of primary spectral characteristic parameters with total water content can be described by an S-shaped function, water content can be divided into three states of dry, water-absorbing, and saturated; (3) The near-infrared spectral characteristics selected by the two information methods are not completely consistent. Based on the BIF method, the correlation between the characteristic variable at 1 450 nm and the rock moisture content ranks from right to left as right shoulder width, peak height, and left shoulder width; at 1 900 nm, the peak height, right shoulder width, and left shoulder width. Based on the MIC method, the correlation between the characteristic variables at 1 450 and 1 900 nm and the rock water content level from the highest to the lowest in the left shoulder width, peak height, and right shoulder width; (4) Decision tree analysis suggests that the MIC method achieves higher accuracy in identifying water content level than the BIF method.

Keywords Water-bearing rock; Near-infrared spectroscopy; Mutual information; Water migration

(Received May 26, 2020; accepted Sep. 12, 2020)

* Corresponding author