第42卷，第3期
2022年3月

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

Vol. 42, No. 3, pp719-725
March, 2022

最小相关系数的多元校正波长选择算法

程介虹[1]，陈争光[1,2*]，衣淑娟[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. 黑龙江省水稻生态育秧装置及全程机械化工程技术中心，黑龙江 大庆 163319

摘 要 在近红外光谱的定量分析中，由于仪器的精密程度越来越高，采集的光谱数据通常具有很高的维度。因此，波长选择对于剔除噪声及冗余变量，简化模型，提高模型的预测性能是必不可少的。近红外光谱特征波长选择方法众多，但变量间的多重共线性问题仍是导致模型效果较差的一个关键问题。变量间共线性可以通过相关系数进行分析，当相关系数高于0.8，表明存在多重共线性。据此，以变量间相关系数为选择标准，提出一种以所选变量之间共线性最小化的波长选择方法，称之为最小相关系数法(MCC)。该方法以光谱数据的相关系数矩阵为基础，挑选出与其他波长相关系数平均值和标准差均较小的波长为候选建模波长集合，使得集合内波长之间线性相关性最小，进而消除模型变量之间共线性。然后通过标准回归系数优选对因变量影响较大的波长，获得预测模型。为了验证所提出算法的有效性，对该方法进行了测试。利用两组公开的近红外光谱数据集(柴油数据集、土壤数据集)，通过MCC算法进行波长选择，并与常用的几个波长选择方法，如：连续投影算法(SPA)、竞争性自适应重加权采样法(CARS)、随机蛙跳算法(RF)、迭代保留信息变量法(IRIV)进行比较。实验结果表明，MCC算法获得了良好的预测性能，MCC算法的预测精度相比于SPA，CARS和RF三种算法具有明显的优势，而MCC算法的预测精度与IRIV算法不相上下。因此，最小相关系数法可实现高效降维，提高模型的预测精度，是一种有效的波长选择算法。

关键词 波长选择；近红外光谱；多元校正；最小相关系数法

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)03-0719-07

引 言

近红外光谱区(800～2 500 nm)的含氢基团的倍频和合频吸收带较宽且严重重叠，全谱建模定量分析会存在多重共线性或无信息变量过多导致模型性能不佳。因此，对全谱进行特征波长选择在一定程度上可以减少数据冗余和多重共线性，提高模型的预测精度和预测效率。特征波长选择一直是近红外光谱分析中的热点。

常用的近红外光谱特征波长选择方法包括：无信息变量消除法[1]、竞争性自适应重加权采样法[2]、间隔偏最小二乘法[3]、遗传算法[4]、连续投影算法[5]、随机蛙跳[6]、迭代保留信息变量法[7]等等。所有这些研究表明，使用特征波长代替全谱建模可以获得更好的预测精度，这说明波长选择的重要性。

众所周知，在近红外光谱的定量分析中，变量间的多重共线性问题是导致模型效果较差的一个关键问题。这是因为如果两个向量之间存在共线性，意味着两个向量具有相似的趋势并且有可能携带相似信息，这类变量的存在会降低模型的性能。变量间共线性可以通过相关性分析判断，相关系数高于0.8，表明存在多重共线性。所以，以变量间相关系数为选择标准，提出一种以所选变量之间共线性最小化，并且输入变量对响应变量影响最大化为主要目的的波长选择方法，称之为最小相关系数法(minimal correlation coefficient，MCC)。

我们将该算法应用于两组公开的近红外光谱数据集，并与其他常用波长选择方法，如：连续投影算法(successive projections algorithm，SPA)、竞争性自适应重加权采样法(competitive adaptive reweighted sampling，CARS)、随机蛙跳算法(random frog，RF)、迭代保留信息变量法(iteratively

收稿日期：2021-02-10，**修订日期**：2021-05-07

基金项目：国家重点研发计划项目(2016YFD0701300)，黑龙江省农垦总局重点科研计划项目(HKKYZD190804)，黑龙江省省属高校基本科研业务费科研项目(ZRCPY201913)资助

作者简介：程介虹，1997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：1024212535@qq.com

*通讯作者 e-mail：ruzee@sina.com

retains informative variables，IRIV)进行对比，以说明本方法的有效性。

1 算法原理

1.1 符号说明

矩阵用粗斜体大写字母表示，向量由粗体小写字母表示，变量(标量)用斜体字母表示。光谱数据矩阵表示为 $\boldsymbol{X}_{N\times K}$，响应(浓度)向量表示为 $\boldsymbol{y}_{N\times 1}$，其中，$N$ 为样本数，K 为波长数。下标变量 i，j 表示光谱矩阵的第 i 列、第 j 列，“~”符号代表预测值。

1.2 相关概念

1.2.1 向量间最小相关系数

在近红外光谱的定量分析中，多重共线性问题是导致模型效果较差的一个关键问题。对于多元线性回归模型来说，当变量之间不存在线性相关性，即参与建模的变量中的任一变量与其他变量间均不存在线性相关性时，模型的预测性能较优。反之，如果参与建模的变量中的任一变量与其他变量间存在线性相关性，则模型存在多重共线性。

皮尔森相关系数(Pearson correlation)是一种衡量向量之间线性相关性的指标，计算方法如式(1)所示。

$$\rho_{x,y}=\frac{\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})}{\sigma_x\sigma_y}=\frac{E[(\boldsymbol{x}-E(\boldsymbol{x}))(\boldsymbol{y}-E(\boldsymbol{y}))]}{\sigma_x\sigma_y} \tag{1}$$

式中，$\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})$为向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的协方差，σ_x，σ_y 为向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的标准差，$E(\boldsymbol{x})$，$E(\boldsymbol{y})$为向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的期望。

皮尔森相关系数 ρ 取值区间为[-1，1]，其绝对值是表征两个向量相关性的一个度量(表 1)。如果两个向量之间的皮尔森相关系数 ρ 绝对值高于 0.8，即可判定存在一定程度的多重共线性；而两个向量之间的皮尔森相关系数 ρ 绝对值小于 0.4，那么两个向量弱相关。因此，向量之间皮尔森相关系数 ρ 的大小可以作为判断向量间共线性程度的一个度量。

表 1 变量间相关系数所对应的相关强度

Table 1 The strength of correlation corresponding to the correlation coefficient between variables

相关系数绝对值	相关强度
0.8～1.0	极强相关
0.6～0.8	强相关
0.4～0.6	中等程度相关
0.2～0.4	弱相关
0.0～0.2	极弱相关或无相关

光谱数据矩阵 $\boldsymbol{X}_{N\times K}$ 可以看作是一个包含 K 个 N 维向量的向量组，每个向量代表了在特定波长下近红外光谱的吸光度特性。为了后面算法描述方便，在此用运算符 $r(\cdot)$表示两个向量之间的相关系数绝对值，$r(\boldsymbol{x}_i,\boldsymbol{x}_j)$与皮尔森相关系数 ρ 的关系如式(2)所示。$r(\cdot)$具有特性：$r(\boldsymbol{x}_i,\boldsymbol{x}_j)=r(\boldsymbol{x}_j,\boldsymbol{x}_i)$，$r(\boldsymbol{x}_i,\boldsymbol{x}_i)=1$，如果光谱矩阵进行了标准化处理，那么，$r(\boldsymbol{x}_i,\boldsymbol{x}_j)=|\mathrm{cov}(\boldsymbol{x}_i,\boldsymbol{x}_j)|$。

$$r(\boldsymbol{x}_i,\boldsymbol{x}_j)=|\rho(\boldsymbol{x}_i,\boldsymbol{x}_j)| \quad (i,j=1,2,3,\cdots,K) \tag{2}$$

由上述理论可知，如果某一个波长与其余波长之间的相关系数 $r(\cdot)$均较小，则意味着该波长不能由其他波长线性表示，需要作为建模变量保留下来，这个波长即为关键波长。反之，如果一个波长与其余波长间的相关系数均较大，则意味着该波长与其余波长间存在多重共线性，不能作为建模变量，需要剔除。

那么，如何判断某一波长与其余所有波长之间的相关性均较小呢？首先计算矩阵 $\boldsymbol{X}_{N\times K}$ 中各列向量之间相关系数的绝对值，得到相关系数矩阵 $\boldsymbol{R}_{K\times K}$，其中第 i 列向量为 $\boldsymbol{X}_{N\times K}$ 中第 i 个波长 $\boldsymbol{x}_i$ 与其他波长之间的相关系数(图 1)。然后计算矩阵 $\boldsymbol{R}_{K\times K}$ 中各列向量的平均值 r 和标准差 σ，选取平均值 r_i 和标准差 σ_i 均较小的列对应的波长作为候选波长。这是因为，对于 $\boldsymbol{R}_{K\times K}$ 的第 i 列向量，其对应的 r_i 较小不一定说明该列向量中的所有元素均较小，只有同时满足标准差 σ_i 较小时，才能说明 $\boldsymbol{R}_{K\times K}$ 中第 i 列相关系数向量各元素值均较小，且分布较为集中。此时 $\boldsymbol{X}_{N\times K}$ 中第 i 个波长应该保留下来作为建模候选波长。

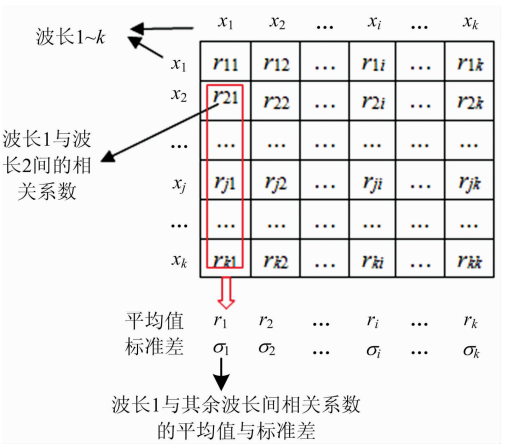

图 1 相关系数平均值与标准差

Fig. 1 Average value and standard deviation of correlation coefficient

1.2.2 最大标准回归系数

由于通过 MCC 方法消除了变量之间的相关性，因此，可以建立基于 MCC 方法的波长选择结果的线性回归方程。假设线性回归方程的形式如式(3)

$$\hat{\boldsymbol{y}}=\boldsymbol{\beta}_0+\sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{x}_i \tag{3}$$

式中，$\hat{\boldsymbol{y}}$ 为估计值，n 为回归变量个数，$\boldsymbol{\beta}_i$ 为最小二乘回归系数。

为了使不同变量的效应大小具有可比性，通常对回归系数 $\boldsymbol{\beta}_i$ 进行无量纲化处理，将其按式(4)转化为一个无量纲的标准回归系数。

$$\boldsymbol{b}_i=\boldsymbol{\beta}_i\times\mathrm{Var}(\boldsymbol{x}_i)/\mathrm{Var}(\hat{\boldsymbol{y}}) \tag{4}$$

式中，$\mathrm{Var}(\boldsymbol{x}_i)$和 $Var(\hat{\boldsymbol{y}})$分别是变量 $\boldsymbol{x}_i$ 和 $\hat{\boldsymbol{y}}$ 的方差。

标准回归系数 $\boldsymbol{b}_i$ 消除了变量 $\boldsymbol{x}_i$ 量纲的影响，具有可比性，$|\boldsymbol{b}_i|$的值越大，相应的变量 $\boldsymbol{x}_i$ 对 $\hat{\boldsymbol{y}}$ 的影响越大。所以，优先选择对 $\hat{\boldsymbol{y}}$ 的影响较大的波长建模，然后通过向前选择法

建立线性回归模型。

向前选择法是一种回归模型的自变量选择方法，其特点是把候选的自变量逐个引入回归方程，故称向前法。首先对符合最小相关系数准则的波长按如前所述计算得到标准回归系数 $\boldsymbol{b}_i$，并按标准回归系数绝对值$|\boldsymbol{b}_i|$大小进行降序排列，然后按排列顺序每次将一个标准回归系数所对应的波长引入模型，建立线性回归模型，计算每个模型的均方根误差，以获得最小均方根误差所引入的变量为准得到一个最优模型。

1.3　算法步骤

根据上述理论基础，MCC 具体算法步骤如下：

步骤 1：利用通过式(2)计算光谱矩阵 $\boldsymbol{X}_{N\times K}$ 中各列向量之间相关系数的绝对值，得到相关系数矩阵 $\boldsymbol{R}_{K\times K}$；

步骤 2：计算矩阵 $\boldsymbol{R}_{K\times K}$ 中各列数据除对角线元素之外的其他元素的平均值 r_i 和标准差 σ_i(图 1)；

步骤 3：通过网格寻优法(后面详细说明)，选择满足某一阈值条件的波长组成候选波长集 $\boldsymbol{S}$；

步骤 4：以集合 $\boldsymbol{S}$ 中波长建立的 MLR 方程[式(3)]，然后通过式(4)计算标准回归系数 $\boldsymbol{b}_i$。对集合 $\boldsymbol{S}$ 中的波长按 $\boldsymbol{b}_i$ 的绝对值降序排序，得到集合 $\boldsymbol{S}'$；

步骤 5：对集合 $\boldsymbol{S}'$中的波长通过向前选择法建模获得该阈值条件下最小均方根误差模型；

步骤 6：选择网格中的下一对阈值，重复步骤 3—步骤 5，直至完成网格寻优。

END：所有阈值条件下最小均方根误差对应的变量集合即为最终所选波长。

1.4　网格寻优

算法步骤 3 中的网格寻优法，具体操作表述如下：

设阈值范围$(r_{\min}, r_{\max})=(\min(r_i), \max(r_i))$，$(\sigma_{\min}, \sigma_{\max})=(\min(\sigma_i), \max(\sigma_i))$，$i=1, 2, \cdots, K$，将阈值范围$(r_{\min}, r_{\max})$和$(\sigma_{\min}, \sigma_{\max})$划分为 $q\times q$ 均匀网格 T，如图 2 所示。以网格 T 中每一个点$(t_{rm}, t_{\sigma n})(m, n=1, 2, \cdots, q)$作为阈值对，从矩阵 $\boldsymbol{R}$ 中选择满足条件$(r<t_{rm}$ 和$\sigma<t_{\sigma n})$的波长组成候选波长集 $\boldsymbol{S}$，按照上述步骤中的方法，得到该阈值条件下的模型的均方根误差。对网格中的每一个阈值对进行上述算法操作(算法步骤 3—步骤 5)。挑选出模型精度最高对应的$(t_{rm}, t_{\sigma n})$作为最优阈值条件。

注意，当 $m=1$ 或 $n=1$ 时，$t_{r1}=r_{\min}$，$t_{\sigma 1}=\sigma_{\min}$，因为 $r_{\min}$ 和 $\sigma_{\min}$分别是相关系数的平均值的最小值和标准差的最小值，此时满足条件$(r<t_{r1}$ 或 $\sigma<t_{\sigma 1})$的波长集 $\boldsymbol{S}$ 为空。

同样，当 $m=q$ 或 $n=q$ 时，$t_{rq}=r_{\max}$，$t_{\sigma q}=\sigma_{\max}$时，则所有的波长都满足条件$(r<t_{rq}$ 和$\sigma<t_{\sigma q})$，导致波长间最小相关准则得不到保障。

因此，取两个阈值系数 $thr1>1$ 和 $thr2<1$，令$(r_{\min}, r_{\max})=(thr1*\min(r_i), thr2*\max(r_i))$，$(\sigma_{\min}, \sigma_{\max})=(thr1*\min(\sigma_i), thr2*\max(\sigma_i))$，在此基础上生成网格矩阵 $\boldsymbol{T}$。这样，当满足条件$(r<t_{rm}$ 和$\sigma<t_{\sigma n})$时，待选波长子集不为空，同时所选波长之间的相关系数较小。

根据经验，阈值系数 $thr1$ 和 $thr2$ 的取值范围分别为 1.3～1.6 和 0.7～0.9，这主要是根据数据的相关性确定的，并通过实验得到，阈值系数对波长选择的影响将在后面讨论部分进行详细阐述。

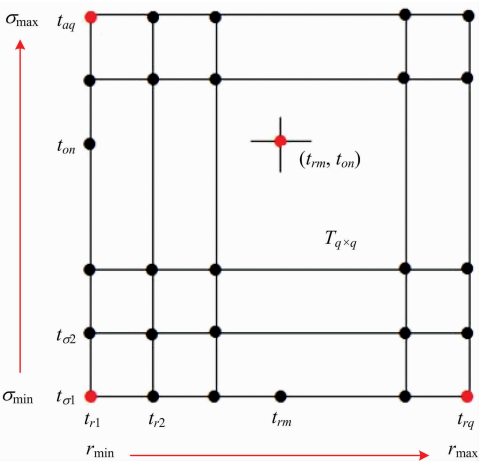

图 2　网格 T 阈值对图解

Fig. 2　Graphic analysis of threshold in grid T

2　实验部分

2.1　样本数据集

选取两个公共数据集进行试验：数据集 1 为一组柴油样本的近红外光谱数据，来自于 Eigenvector 网站，包含 246 个柴油样本，每条光谱有 401 个波长变量，光谱范围为 750～1 550 nm。响应变量为柴油 50%回收率下的沸点(BP50)值。

数据集 2 为一组土壤样本近红外光谱数据，来自于 Quality & Technology 网站，包含 108 个土壤样本，每条光谱有 1 050 个波长变量，光谱范围为 400～2 500 nm。响应变量为土壤有机质(soil organic matter，SOM)的含量。

2.2　样本集划分

近红外光谱数据集被分为一个校正集和一个独立预测集。校正集用于建立校正模型，将校正集进一步划分为训练集(calibration set)和验证集(validation set)，利用验证集对训练集的标定模型的误差进行评估，即验证集是用来指导选择候选波长子集变量个数的集合。独立的预测集(prediction set)用来评估所生成模型的性能，它不用于校正和验证程序的任何步骤。通过 SPXY(sample set portioning based on joint x-y distance)算法将数据集划分为一个校正集(60%的样本)、一个验证集(20%的样本)和一个独立的预测集(20%的样本)。各数据集的样本集划分数及样本化学性质统计结果如表 2 所示。

2.3　模型建立与评价

偏最小二乘回归(partial least squares regression，PLSR)是光谱分析中常用的多元统计数据分析方法。PLSR 适用于预测变量高度共线性，特别是当预测变量大于样本数时，该方法特别有效。然而 PLSR 通常存在潜在变量与原始变量相比难以解释的问题。

多元线性回归(multiple linear regression，MLR)是一种简单、常用的校正方法。MLR 采用最小二乘法进行回归计算，其优点是方便解释波长变量对因变量的影响，利用该方法所得到的模型比 PLSR 模型更易于解释。但其缺点是存在

表 2 样本化学性质含量统计结果

Table 2 Results of sample chemical property

性质	样本集	样本数	最小值	最大值	平均值	标准差	变异系数/%
Dataset1(BP50)	Whole set	246	197	293	258.38	17.06	6.6
	Calibration set	148	197	293	256.32	19.19	7.5
	Validation set	49	201	280	260.86	12.83	4.9
	Prediction set	49	198	283	262.12	12.62	4.8
Dataset2(SOM)	Whole set	108	42.91	95.85	85.43	10.82	12.7
	Calibration set	65	42.91	95.85	82.52	12.82	15.5
	Validation set	22	76.74	93.24	88.30	4.79	5.4
	Prediction set	21	86.65	93.46	91.40	1.63	1.8

多重共线性，当变量数大于样本数时无法实现，因此 MLR 一般在提取特征波长之后再建模。因为本文提出的 MCC 算法消除了变量之间的多重共线性，MCC 波长选择结果适合通过 MLR 方法建模。

波长选择的一个最主要的目的就是提高模型的预测能力，而模型的预测能力主要通过模型的决定系数(R^2)和均方根误差(root mean squared error，RMSE)指标来评价。其中，决定系数又称为拟合优度，取值范围为[0－1]，R^2 越接近 1，自变量对因变量的解释程度越高。RMSE 用来衡量实测值与模型预测值之间的偏差，RMSE 的值越小，说明模型预测精确度越高。

3 结果与讨论

3.1 柴油数据结果

柴油数据的原始近红外光谱图如图 3(a)所示，通过窗口宽度为 11 的 Savitzky-Golay(S-G)一阶求导法进行预处理，预处理后的近红外光谱如图 3(b)所示，后续的波长选择和建模均在图 3(b)数据基础上进行。对原始光谱进行 S-G 导数预处理能够提高光谱分辨率，减少变量间的线性相关性，为后续的波长选择奠定基础。

对于柴油数据，设置 $thr1=1.45$，$thr2=0.85$，通过 MCC 方法选择 20 个变量，选择出的变量分布如图 4 所示。分别为 1 272，1 264，1 266，1 232，1 288，1 294，1 078，1 040，1 274，1 196，1 076，1 230，1 234，1 330，1 316，1 036，1 286，1 338，1 262 和 1 080 nm。20 个波长中，任一波长与其他 19 个侯选波长间的相关系数的平均值和标准差如图 5 所示，各波长与其余所选波长间的相关系数平均值在 0.3～0.4 之间，处于弱相关水平。标准差在 0.2～0.3 之间，相关系数相对集中。由此可见，所选变量之间弱相关，变量间的共线性较小。并且 MCC 选择的波长集中在 1 100 和 1 200～1 300 nm 附近，柴油是轻质石油产品，复杂烃类混合物，而烃类化合物是由碳与氢原子所构成的化合物。MCC 方法选取的特征波长都处于 C—H 键的三倍频和二倍频吸收区域。

为证明 MCC 波长选择算法的有效性，我们对数据集通过 SPA，CARS，IRIV 及 RF 算法进行波长选择，建立相应回归模型，比较预测精度。通过文献[8]可知，各方法的变量初始化、建模方法各不相同，如 MCC 是通过计算所有变量间的最小相关系数，建立 MLR 模型，选择最优变量；SPA 是计算全部变量正交子空间上的最大投影值，建立 MLR 模型，选择最优变量；CARS 是通过蒙特卡洛采样随机抽取 80%的样本建立 PLS 模型；IRIV 是通过二进制矩阵取样，建立 PLS 模型；RF 是通过蒙特卡洛采样，建立 PLS 模型。根据各波长选择算法特点，对 SPA 和 MCC 建立 MLR 模型，对 CARS，IRIV 和 RF 建立 PLSR 模型，会获得较优的预测能力。

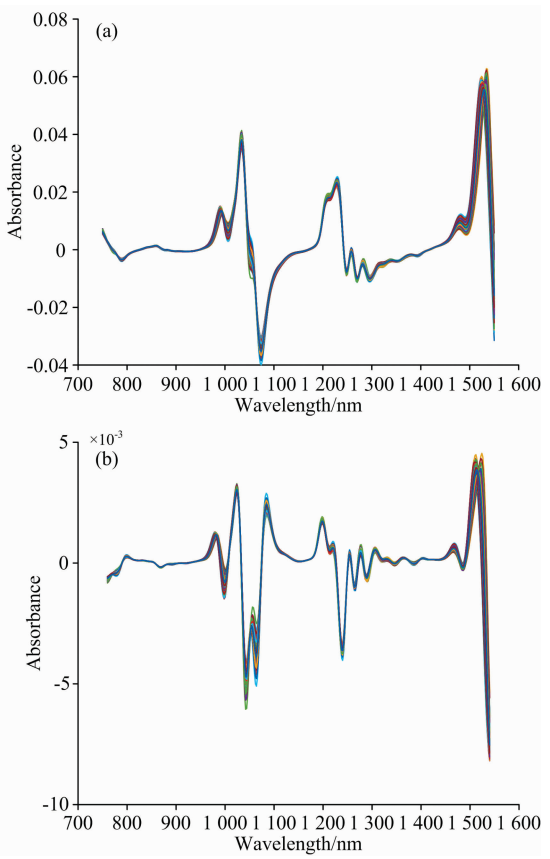

图 3 柴油样本原始近红外光谱图(a)和预处理后的近红外光谱图(b)

Fig. 3 Original (a) and Preprocessed (b) NIR spectra of diesel fuels sample

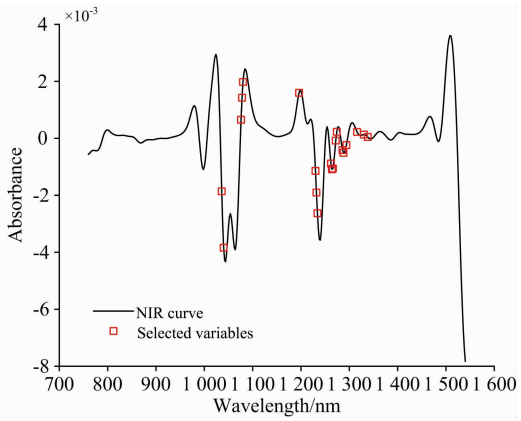

图 4　MCC 波长选择结果

Fig. 4　Wavelength selection results by MCC

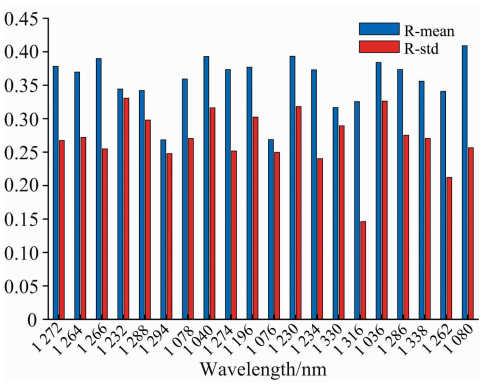

图 5　所选波长间的相关系数平均值与标准差

Fig. 5　Mean and standard deviation of correlation coefficient between selected wavelengths

表 3 为 MCC、全谱及四种常用的波长选择算法(SPA，CARS，IRIV 和 RF)所建模型验证集和测试集的模型结果。由表 3 可知，基于波长选择的回归模型的参数均不同程度地优于基于全谱的 PLSR 模型(FULL-PLSR)，说明对全谱进行特征波长选择的重要性。对比可以发现，MCC 相比于 SPA，CARS 和 RF 三种算法具有明显的优势，IRIV 较 MCC 的预测集均方根误差略优，但在选择的变量个数上，MCC 算法略有优势。这个结果进一步表明，MCC 变量选择方法具有一定的有效性。

表 3　基于柴油数据集不同波长选择方法下模型的参数

Table 3　Model results based on different wavelength selection methods for diesel fuels datasets

No.	建模方法	波长数	Validation set		Prediction set	
			R^2	RMSE	R^2	RMSE
1	FULL-PLSR	401	0.865 0	4.677 8	0.905 2	3.946 8
2	SPA-MLR	25	0.953 6	2.834 5	0.953 9	2.844 9
3	CARS-PLSR	18	0.928 9	4.760 1	0.927 9	3.544 5
4	IRIV-PLSR	60	0.966 4	3.274 7	0.956 8	2.639 6
5	RF-PLSR	34	0.948 0	4.071 9	0.943 4	3.009 9
6	MCC-MLR	20	0.953 9	2.760 1	0.956 0	2.779 2

3.2　土壤数据

将土壤数据的原始近红外光谱数据通过窗口宽度为 11 的 S-G 一阶求导法预处理后，分别通过 MCC，SPA，CARS，IRIV 及 RF 进行特征波长提取并建立回归模型，验证集和测试集模型结果如表 4 所示。由表 4 可知，MCC 的预测集均方根误差相比于 SPA，CARS 和 RF 三种算法具有较优的预测精度，IRIV 与 MCC 的预测效果不相上下。与数据集 1 所得结果一致，可以证明 MCC 方法的有效性。

表 4　基于土壤数据不同波长选择方法下模型参数

Table 4　Model results based on different wavelength selection methods for soil datasets

No.	建模方法	波长数	Validation set		Prediction set	
			R^2	RMSE	R^2	RMSE
1	FULL-PLSR	850	0.585 2	3.229 9	0.656 2	2.917 6
2	SPA-MLR	23	0.953 9	0.908 9	0.810 1	1.870 9
3	CARS-PLSR	9	0.975 5	1.803 9	0.908 8	1.368 2
4	IRIV-PLSR	33	0.912 4	1.182 5	0.910 6	1.027 9
5	RF-PLSR	25	0.898 8	1.282 7	0.807 4	2.108 4
6	MCC-MLR	12	0.936 1	1.144 7	0.926 5	1.032 3

3.3　MCC 方法优势分析

Zhang 等[9]在土壤数据集的基础上，利用 SiPLS，SiPLS-GA 和 SiPLS-GA-SPA 三种方法进行波长选择后的数据建立 PLSR 模型，其中基于 SiPLS-GA-SPA 波长选择结果建立的 PLSR 模型具有最佳预测精度，RMSEP＝1.42。用 MCC 波长选择法建立的模型的 RMSEP 为 1.032 3。相较而言，基于 MCC 的波长选择算法更具有优势。根据以往的研究结果可知，基于 SPA 方法的波长选择模型优于 GA[10]，CARS[11]和 IPLS[12]等其他波长选择方法。由此可见，MCC 算法在波长选择方面具有一定的优势。

从 MCC 算法可以看出，MCC 和 SPA 的基本原理都是选择最少冗余信息和最小共线性的变量组合，不同的是，SPA 是利用向量的投影分析，通过将多个向量投影到某个超平面(前一个投影向量的法平面)上，比较投影向量的模，选择模最小的向量作为待选向量，以此达到该向量与前一个向量的共线性最小。但是，SPA 波长选择只考虑相邻两次选择波长之间的相关性，而不考虑所选择的波长和其他所选波长之间的相关性，即，SPA 波长选择算法并没有考虑所有特征波长之间的共线性。

MCC 算法以某一波长和其他所有波长间相关系数最小为准则选择特征波长，使得所选特征波长和其他所有波长之间相关性最小，从而达到降低共线性的目的。由于减少了所有特征波长之间的共线性，因此 MCC 波长选择的结果适合使用线性回归方法建模，所以 MCC-MLR 模型的预测精度较优。从前面两个数据集建模所得的结果来看，相较于其他波长选择算法所建模型的预测精度，MCC-MLR 模型的预测精度较优，是一种有效的模型。MCC 算法在应用于近红外光谱数据的波长选择具有一定的优势，本方法的设计思想可以为其他高维数据的特征选择和降维提供一定的参考。

3.4 阈值系数的选取对 MCC 方法的影响

根据算法步骤 3，为了保证候选波长集 $\boldsymbol{S}$ 在某一阈值下不为空集，阈值系数 $thr1$ 应大于 1。另外，为了保证所选择的波长更有效，阈值系数 $thr2$ 应小于 1。为了获得更好的 $thr1$ 和 $thr2$，通过选择 $thr1$ 和 $thr2$ 的不同组合，得到 MCC-MLR 模型的 RMSEV。图 6 为以土壤数据集为例得到的不同的 $thr1$ 和 $thr2$ 阈值系数取值情况下的模型 RMSE 分布情况。

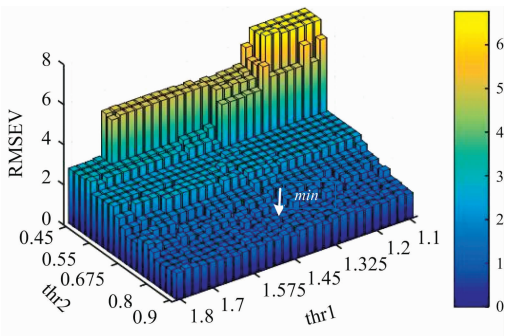

图 6 不同阈值系数下 MCC-MLR 模型的 RMSEV 值变化

Fig. 6 Changes in RMSEV value of MCC-MLR model under different threshold coefficients

从图 6 可以看出，随着 $thr2$ 变大，RMSEV 逐渐变小，当 $thr2$ 大于 0.8 时，RMSEV 趋于平缓，MCC 算法的 $thr2$ 取值为 0.85。随着 $thr1$ 的降低，RMSEV 有下降趋势，在 $thr1=1.45$ 附近取得最小值，以此，MCC 算法的 $thr1$ 的取值为 1.45。对于柴油数据集来说，当 $thr1$ 取值 1.45，$thr2$ 取值为 0.85 时，MCC-MLR 方法也能得到最小 RMSE 模型。因此，MCC 设置 $thr1$ 和 $thr2$ 的默认值分别为 1.45 和 0.85。对于其他数据集，两个阈值系数的取值可能有细微波动，可以通过实验确定。

4 结 论

针对变量间的多重共线性问题，提出一种基于最小相关系数的近红外光谱波长选择方法，该方法以所选变量之间共线性最小化，并且输入变量对响应变量的影响最大化为主要目的，消除模型间的共线性，提高模型预测精度。为了验证所提出算法的有效性，利用两组公开的近红外光谱数据集对该方法进行了测试。结果表明，MCC 算法获得了良好的预测性能，是一种有效的波长选择算法。MCC 方法可以为其他类型高维数据降维提供参考。

References

[1] Weng S, Yu S, Dong R, et al. International Journal of Food Properties, 2020, 23(1): 269.
[2] Vohland M, Harbich M, Ludwig M, et al. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9(9): 4011.
[3] Rahman A, Kondo N, Ogawa Y, et al. Biosystems Engineering, 2016, 141: 12.
[4] WANG Bing-yu, SUN Wei-jiang, HUANG Yan, et al(王冰玉，孙威江，黄 艳，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(4): 1100.
[5] Acebal C C, Grünhut M, Lista A G, et al. Talanta, 2010, 82(1): 222.
[6] Li H D, Xu Q S, Liang Y Z, Analytica Chimica Acta, 2012, 740: 20.
[7] Yun Y H, Wang W T, Tan M L, et al. Analytica Chimica Acta, 2014, 807: 36.
[8] Yun Y H, Li H D, Deng B C, et al. TrAC Trends in Analytical Chemistry, 2019, 113: 102.
[9] ZHANG Xiao-ming, TANG Ning(张小鸣，汤 宁). Modern Electronics Technique(现代电子技术), 2018, 41(22): 126.
[10] WANG Tao, BAI Tie-cheng, YU Cai-li, et al(王 涛，白铁成，喻彩丽，等). Jiangsu Agricultural Sciences(江苏农业科学), 2018, 46(19): 269.
[11] WU Long-guo, WANG Song-lei, HE Jian-guo, et al(吴龙国，王松磊，何建国，等). Chinese Journal of Luminescence(发光学报), 2017, 38(10): 1366.
[12] Chen J, Ren X, Zhang Q, et al. Journal of Cereal Science, 2013, 58(2): 241.

Wavelength Selection Algorithm Based on Minimum Correlation Coefficient for Multivariate Calibration

CHENG Jie-hong[1], CHEN Zheng-guang[1,2*], YI Shu-juan[2]

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China
2. Heilongjiang Engineering Technology Research Center for Rice Ecological Seedlings Device and Whole Process Mechanization, Daqing 163319, China

Abstract In the quantitative analysis of near-infrared spectroscopy, as the instrument's precision is getting higher and higher. The collected spectral data usually has a very high dimension. Therefore, wavelength selection is essential for eliminating noise and redundant variables, simplifying the model, and improving the model's predictive performance. There are many methods for selecting characteristic wavelengths in NIR spectroscopy, but the problem of multicollinearity among variables is still a key issue that leads to poor model effects. Collinearity between variables can be analyzed by correlation coefficient. When the correlation coefficient is higher than 0.8, it indicates that there is multicollinearity. Therefore, this paper takes the correlation coefficient between variables as the selection criteria and proposes a wavelength selection method that minimizes the collinearity between the selected variables, called the Minimal Correlation Coefficient (MCC) method. This method is based on the correlation coefficient matrix of the spectrum data. It selects the wavelength with the smaller average and standard deviation of the correlation coefficients of other wavelengths as the candidate modeling wavelength set so that the linear correlation between the wavelengths in the set is minimized, and the model has eliminated Collinearity between variables. Then use the standard regression coefficient to select the wavelength that has a greater impact on the dependent variable to obtain the prediction model. In order to verify the effectiveness of the proposed algorithm the method is tested. Using two sets of opening NIRS data sets (diesel dataset and soil dataset), wavelength selection was carried out by MCC algorithm, and compared with several other commonly used wavelength selection methods, including successive projections algorithm (SPA), competitive adaptive reweighted sampling (CARS), random frog (RF) and iteratively retains informative variables (IRIV). The experimental results show that the MCC algorithm has good prediction performance, the model prediction accuracy based on MCC is better than that of SPA, CARS, RF, and is roughly the same as that of IRIV. Therefore, the minimum correlation coefficient method is an effective wavelength selection algorithm, which can reduce the dimension efficiently and improve the prediction precision of the model.

Keywords Wavelength selection; Near-infrared spectroscopy; Multivariate calibration; Minimal correlation coefficient

(Received Feb. 10, 2021; accepted May 7, 2021)

* Corresponding author