

变量选择方法对近红外光谱校正模型 优化极限的研究

潘正豪¹, 王 鹏², 陈昆燕², 李秋潼², 唐 杰², 杨 俊¹, 邵利民^{1*}

(1. 中国科学技术大学 化学与材料科学学院, 安徽 合肥 230026; 2. 重庆中烟工业有限责任公司
技术中心, 重庆 400060)

摘要: 变量选择经常被用于优化近红外光谱线性校正模型, 消除冗余信息, 提升回归的准确性和可解释性。该文研究并设计了一种基于蒙特卡洛的方法, 用于评估不同线性校正方法在变量选择的子空间中能达到的最优程度, 寻找变量选择对线性校正模型的优化极限。该方法通过获得验证指标——预测均方根误差(RMSEP)的分布图, 揭示变量选择方法在数据集上的优化效果与优化极限。将该方法应用于3组样品的近红外光谱建模研究, 结果表明: 在烟草-果胶数据集上的可优化率约为24.98%, RMSEP降低了15.2%; 在小麦-蛋白质数据集上的可优化率约为13.90%, RMSEP降低了9.5%; 在玉米-淀粉数据集上的可优化率约为14.05%, RMSEP降低了57.1%。应用该方法可以快速得到变量选择方法在模型上的优化极限, 为变量选择方法的设计、应用和评估提供参考。

关键词: 化学计量学; 近红外光谱; 化学校正; 线性模型; 变量选择; 蒙特卡洛方法

中图分类号: O657.3; TB9 **文献标识码:** A **文章编号:** 1004-4957(2023)12-1659-07

Study on the Optimization Limit of Near-infrared Spectroscopy Calibration Model by Variable Selection Methods

PAN Zheng-hao¹, WANG Peng², CHEN Kun-yan², LI Qiu-tong², TANG Jie²,
YANG Jun¹, SHAO Li-min^{1*}

(1. School of Chemistry and Material Science, University of Science and Technology of China, Hefei 230026, China; 2. Technical Center, China Tobacco Chongqing Industrial Co., Ltd., Chongqing 400060, China)

Abstract: Variable selection is often used to optimize linear calibration models for NIR spectroscopy, eliminating redundant information and improving the accuracy and interpretability of regression. In this paper, a Monte Carlo-based method is studied and designed to find the optimal limit of variable selection for linear calibration models, and can evaluate the optimal degree that different linear calibration methods (such as partial least squares regression PLSR) can achieve in the subspace of variable selection. In this study, the distribution plot of the root mean square error prediction (RMSEP) of the validation metric is obtained, and the optimization effect and optimization limit of the variable selection methods on the dataset are revealed. The method was applied to the near-infrared spectroscopy modeling of three datasets, and the results showed that the optimizable rate on the tobacco-pectin dataset was about 24.98%, and the prediction error RMSEP was reduced by 15.2%. The optimizable rate on the wheat-protein dataset is about 13.90%, and the prediction error RMSEP is reduced by 9.5%. The optimizable on the corn-starch dataset is about 14.05%, and the prediction error RMSEP is reduced by 57.1%. The application of this method can quickly obtain the optimization limit of the variable selection methods on the model, and provide reference for the design, application and evaluation of the variable selection methods.

Key words: chemometrics; near-infrared spectroscopy; chemical calibration; linear model; variable selection; Monte Carlo method

近红外光谱(Near-infrared spectroscopy, NIRS)是一种介于可见光和中红外之间的吸收光谱, 主要包含C—H、N—H和O—H化学键的信息, 几乎所有有机物和大部分含氢的分子都可以使用NIRS进行

表征。近红外分析方法具有分析速度快、效率高、成本低等特点,被广泛应用于农业^[1]、食品^[2]、医疗^[3]等行业。但是,NIRS包含很多泛频和合频,这使光谱解析变得复杂,难以在特定波段上找到特定化学键的信息。多变量方法诸如主成分回归(PCR)、偏最小二乘回归(PLSR)、人工神经网络(ANN)等可以用来建立化学校正模型,解析光谱中的有效化学信息。

建立校正模型时,通过变量选择方法使用不同光谱波段数据可以较大程度降低模型的预测误差,提高模型的可解释性和泛化能力^[4-6]。变量选择方法可以按照选择策略分为以下几类:基于过滤器的方法如RC^[7](Regression coefficients)和RT^[8](Randomization test),基于极值搜索的方法如连续投影算法(Successive projections algorithm, SPA),基于顺序搜索的方法^[9]如fiPLS(Forward interval PLS)和biPLS(Backward interval PLS),基于详尽搜索的方法如VCPA^[10](Variable combination population analysis)和siPLS^[11](Synergy interval PLS),基于智能优化的算法如遗传算法^[11](Genetic algorithm, GA)和模拟退火算法^[12](Simulated annealing, SA),基于模型集群分析的方法如竞争自适应重加权采样^[13](Competitive adaptive reweighted sampling, CARS)等。常用的变量选择方法还有蒙特卡洛-无用信息变量消除^[14](Monte Carlo-uninformative variable elimination, MC-UVE)、随机跳蛙^[15](Interval random frog, iRF)、间隔偏最小二乘法^[16](Interval partial least squares, iPLS)等。随着学科交叉的深入和计算机算法的开发,变量选择方法愈来愈多。面对新的数据集,优化其线性回归模型时,经常会遇到两个问题^[5, 17-18]:(1)若无相关经验,则可能需要逐个尝试常见的算法或开发新的算法;(2)在线性关系较弱的数据集中,有时即使应用变量选择方法,也无法得到较好的回归和预测结果。

笔者在研究烟草样品近红外光谱的PLSR模型时,发现无论应用何种方法,均无法有效提升模型在预测集上的预测误差,这可能说明几乎没有变量选择方法能够在本研究小组的烟草数据集上获得较好的回归效果。理论上虽然可以通过光谱全排列得到最优的变量选择,但是受限于计算速度,实际中无法完成,故优化算法变量选择成为可行的策略^[19]。

基于此,本文研究并设计了一种寻找变量选择优化极限的方法,基于蒙特卡洛随机重复试验,生成一系列变量子集,并对这些子集建模和验证,考察子模型在验证集上的预测效果,并记录其预测均方根误差(Root mean square error of prediction, RMSEP),绘制RMSEP的分布图,从而获得关于数据集和模型的相关信息。

对于特定数据集来说,尝试常见算法与开发新算法或许无法避免,但对于整体优化目标来说,有时需要快速了解一个数据集和模型是否适用变量选择方法,如果适用,该变量选择方法又表现出何种效果。基于该思想,这种通过大量随机试验得到的结果能在一定程度上接近数据集和模型的内在属性,因此可以高效探究变量选择算法对该数据集和模型的适用性:变量选择可以多大程度地优化此模型,变量选择方法在此模型上的优化极限是多少。与其他变量选择方法相比,本方法具有较强的泛化能力,可以获得更低的预测误差。

1 实验及原理

1.1 数据来源

使用3个近红外光谱数据集验证本文提出的方法。

数据集1来自实验室测量烟草样品的近红外光谱及其果胶含量。样品为重庆中烟和安徽中烟提供的2014~2015年云南、贵州、湖北、重庆、福建、安徽等产区的烤烟样品共203个。在Thermo Antaris II型傅里叶变换-近红外光谱仪(Thermo Scientific热电公司,美国)上按DB53/T 497-2013《烟草及烟草制品主要化学成分指标 近红外校正模型建立与验证导则》^[20]规定的方法采集样品光谱,近红外波长范围为1 000~2 500 nm,共获得1 557个变量。按照间羟基联苯法测量样品的果胶质量百分含量。

数据集2、3来自网络上公开的近红外光谱数据:数据集2来自https://www.cNIRS.org/content.aspx?page_id=22&club_id=409746&module_id=239453。该数据集由496个小麦样本组成,近红外波长范围为730~1 100 nm,间隔为0.5 nm(741个变量),包含小麦近红外光谱及其蛋白质含量。数据集3来自<https://www.eigenvector.com/data/Corn>,选用m5仪器的光谱数据集,该数据集由80个玉米样本组成,近红外波长范围为1 100~2 498 nm,间隔为2 nm(700个变量),包含玉米近红外光谱及其淀粉

含量。

3套数据集的基本信息列于表1。

表1 数据集概况
Table 1 Basic information on the datasets

		Number of samples	Maximum/%	Minimum/%	Mean/%	Standard deviation/%
Tobacco-pectin(烟草-果胶)	Calibration set	136	11.76	5.11	7.16	1.31
	Validation set	46	10.80	6.02	6.79	0.91
	Prediction set	21	8.28	5.01	6.65	0.81
Wheat-protein(小麦-蛋白质)	Calibration set	334	18.69	7.97	13.55	2.19
	Validation set	112	17.53	10.70	14.04	1.56
	Prediction set	50	18.50	8.21	13.39	2.51
Corn-starch(玉米-淀粉)	Calibration set	54	65.90	62.83	64.63	0.87
	Validation set	18	65.84	63.26	64.86	0.59
	Prediction set	8	66.47	63.10	64.75	0.94

1.2 算法步骤

(1)以浓度梯度法在各个浓度区间内选择10%的样品作为预测集;

(2)运用SPXY方法^[21]将剩余数据集划分为建模集和验证集,样本个数比为3:1。在初始全谱模型上以PLS方法建立回归模型,得到初始模型及其预测误差RMSEP₀;

$$\text{RMSEP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

式中, n 为样品总数, y_i 表示样品的真实值, $\hat{y}_i$ 表示样品的预测值。

(3)设置循环次数echotime和变量选择数 s ;

(4)在每次循环中,随机选择 s 个变量点,形成一个变量选择后的新子集;

(5)为每个建模集的变量选择建立PLS回归模型,代入验证集得到评价指标RMSEP;

(6)保留每个RMSEP值,绘制其分布图,得到可优化率 r_0 和优化极限 L_0 。 r_0 定义为RMSEP值小于RMSEP₀的部分的变量选择数占总循环次数的比值, L_0 定义为所有循环中出现的最小RMSEP₀;

(7)RMSEP值小于RMSEP₀的部分的变量选择是可优化变量选择,统计这些变量选择中每个变量出现的次数,作为每个变量波数的相关性度量;

(8)按照变量波数相关性降序逐个加入最终变量选择子集,在RMSEP值达到最小值时结束。

算法中第2~6步循环过程的流程图如图1所示。

2 结果与讨论

将本研究提出的方法应用于烟草果胶、小麦蛋白质和玉米淀粉3个数据集的处理。首先详细讨论了本方法获取烟草果胶数据集变量选择优化极限的过程;其次,绘制3个数据集的RMSEP分布图,对比变量选择在不同数据集上的优化效果;最后,比较了其他各种常用变量选择方法在数据集上的表现。

2.1 烟草-果胶数据集变量选择的优化极限

烟草-果胶数据集(数据集1)含有203个样品,在4 000 cm⁻¹至10 000 cm⁻¹波数上可获得具有1 557个吸光度的NIR光谱矩阵 $X_{203 \times 1557}$ 和果胶百分含量的化学测量值矩阵 y 。从数据集1中以浓度梯度抽取21个样品(10%)作为预测集备用,采用SPXY方法将剩余数据按照3:1划分成建模集和验证集,样品个数分别为136与46。经10折交叉验证方法得到校正集的主

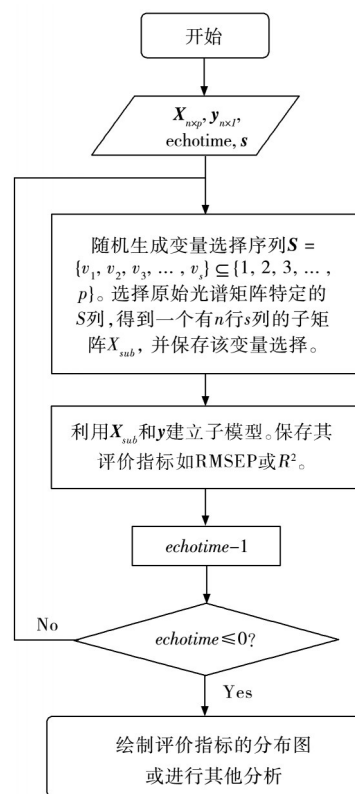


图1 算法步骤的流程图

Fig. 1 The flowchart of the proposed algorithm

成分数为10, 后续子模型的主成分数均取10。

设定循环次数为 10^6 次, 每次循环中随机选择 $s=30$ 个变量随机产生一个变量子集 $\mathbf{S} = \{v_1, v_2, v_3, \dots, v_s\} \subseteq \{1, 2, 3, \dots, p\}$, $\mathbf{S}$ 中含有 s 个变量序号, 选择过后的 $\mathbf{X}_{203 \times s}$ 与 $\mathbf{y}$ 以主成分数 $A=10$ 进行建模验证, 每次循环得到的以 RMSEP 为验证指标的分布图如图 2 所示。 s 的取值为 30~50 时, 能在子模型上取得最佳效果。

图 2 为间距较小的条形频数分布直方图。竖线表示初始模型的 RMSEP 值, 小于该值的变量选择占比 r_o 约 25.0%, 是可优化的变量选择比率。从图 2 可以看出, 变量选择方法在该烟草-果胶数据集中存在优化空间, 即任意选择 30 个变量, 约有 25.0% 的概率得到比全谱更好的子模型。但同时变量选择方法对其优化也存在上限, 任意方法产生的变量选择均不会低于其最小值 L_o (约 0.40)。

以初始模型 RMSEP 值作为截断阈值, 统计每一个变量被选中的次数, 绘制得到如图 3 所示的频数谱。在所有可优化的变量选择中, 每个光谱波数点被选中的频率, 表示了这些波数点处的变量在被选中之后对验证集的影响, 较好的变量选择被选中, 特定波段的频数会高于或低于均值; 高频数部分被选中后, 更容易组合成较好的变量选择, 而低频数部分更不容易。这表明高频数部分 (如 $4\ 200\text{ cm}^{-1}$) 比低频数部分 (如 $4\ 800\text{ cm}^{-1}$) 包含较多与烟草-果胶相关的光谱信息。

图 3 所示的频数图揭示了 NIR 光谱中每一个波数点与待测组分浓度之间的线性校正关系强度。由于蒙特卡洛方法引入的随机噪声, 频数谱中存在比较尖锐的锯齿, 利用 Savitzky-Goaly 滤波将其平滑后按照频数自高至低排列, 并逐个加入待选变量集中, 每加入一个变量后的子模型均可建模并验证, 其 RMSEP 值随着加入到变量选择中的变量个数的变化如图 4 所示。从图 4 可以看出, 选择 111 个变量后, 相应模型的 RMSEP 值降至最小值 (0.38), 算法步骤结束, 得到一个含有 111 个波数的最优变量选择。

2.2 变量选择优化效果

对 3 个数据集, 统一循环次数 (10^6), 循环中每次选择 30 个变量, 子模型使用各自基础模型的主成分数, 分别应用本方法研究其 PLSR 模型的变量选择优化效果, 得到的 RMSEP 分布图如图 5 所示。

图 5A 表明, 对于烟草-果胶数据集, 变量选择能在一定程度上提升预测的精度, 可优化率约 25.0%, 但提升上限有限 (0.40)。图 5B 表明, 对于小麦-蛋白质数据集, 可优化率约 13.90%, 说明样品之间本就具有较好的线性相关度, 全谱 PLS 处理之后即可得到较好的效果, 能够完成可接受误差范围内的回归预测, 变量选择方法几乎完全不能使其预测误差降低, 优化极限 (0.18) 与初始值 (0.22) 接

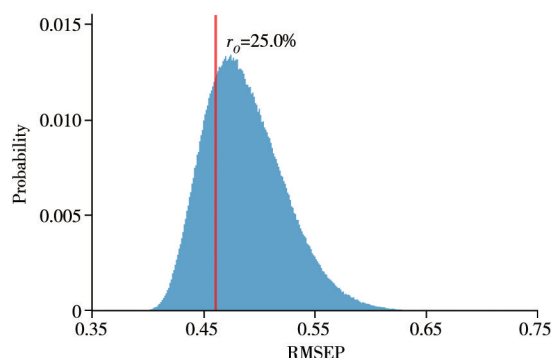


图 2 10^6 次循环结果的 RMSEP 值分布图

Fig. 2 Distribution of RMSEP after 10^6 cycles

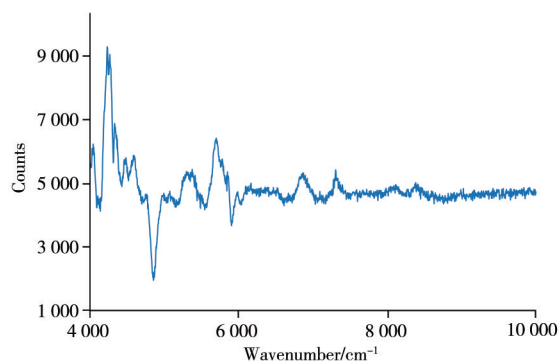


图 3 对优化有贡献的波数选择频数

Fig. 3 Frequency for optimization after variable selection

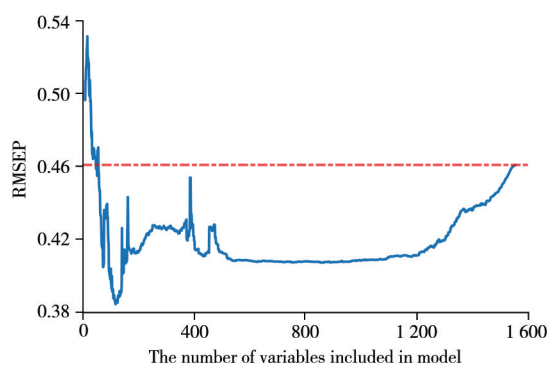


图 4 RMSEP 随加入变量数的变化图

Fig. 4 RMSEP changes with the number of variables included in model

the horizontal line represents the RMSEP of the model built with the data in the whole spectral range (水平虚线表示初始全谱模型的 RMSEP 值)

近，因此变量选择方法不适用于该数据集。图 5C 表明，对于玉米-淀粉数据集，可优化率约 14. 1%，其原始模型的 RMSEP 为 0. 15，提升上限约 0. 05，这说明变量选择方法在玉米-淀粉数据集上可以有较好的表现，较有可能找到一种变量选择方法使该模型的预测误差降到更低水平。可见，对于不同数据集来说，变量选择方法被应用之后的效果截然不同。

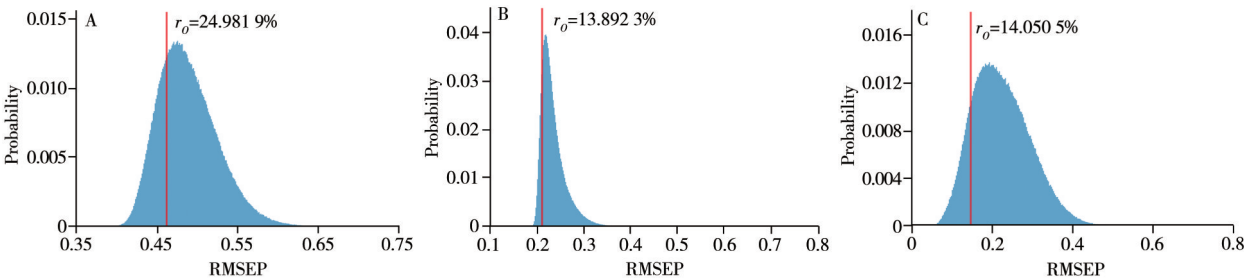


图 5 各数据集的 RMSEP 分布图
Fig. 5 RMSEP distribution of each dataset
A. tobacco-pectin dataset, B. wheat-protein dataset, C. corn-starch dataset; the vertical line represents the RMSEP value of the initial model, and the line-side value is r_0

结合表 2 可知，玉米中的淀粉含量较高(建模集 62. 83%)，因此相对误差较低，玉米-淀粉模型的相对误差(0. 003)远优于烟草-果胶模型(0. 063)，说明该玉米-淀粉数据集更适合使用线性回归模型，且变量选择方法可以在一定程度上优化模型，提升其预测精度。而烟草-果胶数据集的线性回归模型，使用变量选择方法仅能将预测精度从 $RMSEP_0$ (0. 46)提升至 L_0 (0. 39)，最多优化约 15. 2%，若要进一步提升该模型的预测精度，应尝试其他方法或其他模型。

2. 3 不同变量选择方法对比

对 3 个数据集应用 SPA、CARS、MC-UVE、iRF、GA、iPLS 6 种常用变量选择方法进行变量选择并以 PLS 建模，与本方法得到的变量选择极限进行对比，其校正集和验证集的决定系数、预测均方根误差列于表 2。所有数据均由 MATLAB 2021b 运行于搭载 AMD Ryzen 5 5600 和 Windows 10 操作系统的微机得到。

表 2 不同变量选择方法下各数据集 PLS 回归模型参数
Table 2 PLS regression model of each dataset under different variable selection methods

	Method	Number of variables	R_c^2	RMSEC	R_p^2	RMSEP	Time/s	A
Dataset 1 (Tobacco-pectin)	Full spectrum	1 557	0. 81	0. 56	0. 74	0. 46	0. 1	10
	SPA	35	0. 80	0. 57	0. 76	0. 44	68. 2	10
	CARS	51	0. 81	0. 55	0. 76	0. 43	0. 5	7
	MC-UVE	40	0. 79	0. 59	0. 76	0. 44	15. 5	12
	iRF	40	0. 81	0. 56	0. 74	0. 45	131. 4	8
	GA	116	0. 81	0. 55	0. 79	0. 41	53. 0	8
	iPLS	208	0. 80	0. 56	0. 78	0. 41	1. 0	9
Dataset 2 (Wheat-protein)	This study	111	0. 81	0. 59	0. 81	0. 39	9. 4	10
	Full spectrum	741	0. 99	0. 20	0. 98	0. 21	0. 03	15
	SPA	15	0. 99	0. 26	0. 98	0. 22	502. 8	15
	CARS	75	0. 99	0. 20	0. 98	0. 23	0. 3	12
	MC-UVE	200	0. 99	0. 20	0. 98	0. 21	16. 6	14
	iRF	132	0. 99	0. 19	0. 98	0. 23	71. 9	15
	GA	117	0. 99	0. 20	0. 98	0. 24	76. 8	13
Dataset 3 (Corn-starch)	iPLS	97	0. 99	0. 22	0. 98	0. 25	2. 0	12
	This study	95	0. 99	0. 19	0. 99	0. 19	5. 9	14
	Full spectrum	700	0. 99	0. 09	0. 83	0. 14	0. 01	25
	SPA	50	0. 99	0. 09	0. 83	0. 23	5. 4	24
	CARS	45	1. 00	0. 04	0. 97	0. 10	0. 1	24
	MC-UVE	54	0. 99	0. 08	0. 91	0. 16	9. 9	20
	iRF	74	1. 00	0. 04	0. 97	0. 09	20. 9	22

(续表 2)

Method	Number of variables	R_c^2	RMSEC	R_p^2	RMSEP	Time/s	A
GA	117	1.00	0.04	0.98	0.07	58.9	21
iPLS	93	1.00	0.04	0.97	0.10	0.7	20
This study	123	1.00	0.01	0.99	0.06	1.7	22

R_c^2 , fitted correlation coefficient of the calibration set; RMSEC, predicted root mean square error for modeling sets; R_p^2 , the fitting correlation coefficient of the validation set; RMSEP, predicted root mean square error for validation sets; A, the number of principal components in the PLSR algorithm(R_c^2 , 建模集的拟合相关系数; RMSEC, 建模集的预测均方根误差; R_p^2 , 验证集的拟合相关系数; RMSEP, 建模集的预测均方根误差; A, PLSR 算法选用的主因子数)

从表 2 可以看出, 与不作变量选择相比, 大部分变量选择方法均能提高模型预测能力, 但提升程度有限。利用本方法对不同数据集探查后发现, 变量选择方法对 PLS 方法在这些数据集上的预测能力提升效果存在上限。对于数据集 1~3, 各种变量选择方法的差别较为明显。同一个数据集上, 有的方法能够提升模型的回归效果和预测能力, 而该方法在另一个数据集上却没有相同的效果。这说明, 这些变量选择方法对数据集有不同的适用条件和场景, 同时变量选择方法的应用效果也受到数据集来源、品质、特性等的影响。这也就意味着, 在利用变量选择方法提升线性回归方法的准确度和预测能力时, 如果没有先验经验, 无法选择某一特定的方法达到对数据集的优化。

对于不同的数据集, 由于实验设计、样品选取等方法的不同, 会引入相当大的非线性因素, 最终导致近红外数据集在线性回归模型上的效果不能令人满意, 如数据集 1 中的烟草样品来自全国不同烟草产区, 其果胶含量受限于气候、土壤等种植条件而表现出较大的差异性。而数据集 2 中的小麦来自同一产地和种植周期, 仅仅在麦粒蛋白含量上存在差异, 其近红外光谱对蛋白质的线性校正模型具有非常好的拟合度和预测性。

3 结 论

本文提出的基于蒙特卡洛方法的变量选择模式, 可以用于检验变量选择算法在各类线性回归模型上的优化极限。较好的回归模型, 预测时的相对误差和绝对误差本身就较小, 这类模型的可优化率 r_o 和优化极限 L_o 均较小, 可以不必采用变量选择方法, 直接作为稳健模型进行应用。较差的回归模型, 若优化极限 L_o 较小而可优化率 r_o 较大, 可以尝试采用变量选择方法得到一些较好的子模型; 若优化极限 L_o 较大, 则应该放弃使用该线性模型和变量选择方法, 转而采用其他方法建模。因此, 对于各类数据集及其线性回归模型, 采用本方法确定变量选择方法对样品集提升效果的极限是有必要的。

变量选择方法仍存在一些需要解决的问题, 如预测模型会对噪声更敏感、妨碍异常值的检测, 方法选择存在一定困难等。在面对没有先验经验的样品数据集时, 本方法可以在一定程度上提供参考。通过较多次数的随机变量选择, 绘制子模型 RMSEP 的分布图, 一方面可以判断分布曲线最左侧的 RMSEP 极限极小值, 率先了解变量选择方法在模型上的提升上限; 另一方面可以根据极小值至初始模型的 RMSEP 值截断的分布曲线覆盖面积比、斜率, 判断变量选择方法对该数据集的提升效果。最后, 在可优化的变量选择中, 将不同变量波数被选中的频率, 逐一加入变量选择子集中, 可以得到一个兼具拟合效果、预测能力、解释性的变量选择。

参考文献:

- [1] Yukihiro O, Christian H, Satoru T, Søren B E. *Near-infrared Spectroscopy*. Singapore: Springer Singapore, **2021**: 331–345.
- [2] Grassi S, Casiraghi E. *Foods*, **2022**, 11: 1250.
- [3] Callery E L, Rowbottom A W. *Appl. Spectrosc. Rev.*, **2022**, 57(5): 411–440.
- [4] Chu X L, Huang Y, Yun Y H, Bian X H. *Chemometric Methods in Analytical Spectroscopy Technology*. Singapore: Springer Nature Singapore, **2022**: 169–207.
- [5] Yun Y H, Li H D, Deng B C, Cao D S. *Trends Anal. Chem.*, **2019**, 113: 102–115.
- [6] Wang H P, Chen P, Dai J W, Liu D, Li J Y, Xu Y P, Chu X L. *Trends Anal. Chem.*, **2022**, 153: 116648.
- [7] Wold S, Sjostrom M, Eriksson L. *Chemom. Intell. Lab. Syst.*, **2001**, 58(2): 109–130.
- [8] Xu H, Liu Z C, Cai W S, Shao X G. *Chemom. Intell. Lab. Syst.*, **2009**, 97(2): 189–193.
- [9] Zou X B, Zhao J W, Li Y X. *Vib. Spectrosc.*, **2007**, 44(2): 220–227.

- [10] Yun Y H, Wang W T, Deng B C, Lai G B, Liu X B, Ren D B, Liang Y Z, Fan W, Xu Q S. *Anal. Chim. Acta*, **2015**, 862: 14–23.
- [11] Leardi R. *J. Chemom.*, **2000**, 14: 643–655.
- [12] Kirkpatrick S, Gelatt C D, Vecchi M P. *Science*, **1983**, 220(4598): 671–680.
- [13] Li H D, Liang Y Z, Xu Q S, Cao D S. *Anal. Chim. Acta*, **2009**, 648(1): 77–84.
- [14] Cai W S, Li Y K, Shao X G. *Chemom. Intell. Lab. Syst.*, **2008**, 90(2): 188–194.
- [15] Yun Y H, Li H D, Leslie E W, Fan W, Wang J J, Cao D S, Xu Q S, Liang Y Z. *Spectrochim. Acta A*, **2013**, 111: 31–36.
- [16] Norgaard L, Saudland A, Wagner J, Nielsen J P, Munck L, Engelsen S B. *Appl. Spectrosc.*, **2000**, 54(3): 413–419.
- [17] Ciurczak E W. *Handbook of Near-infrared Analysis*. 4th ed. Boca Raton: CRC Press/Taylor & Francis Group, **2021**.
- [18] Brown S. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. Second edition. Amsterdam, Oxford, Cambridge, MA: Elsevier, **2020**.
- [19] Zhang J, Hu Y, Zhou L X, Li B Y. *J. Instrum. Anal.* (张进, 胡芸, 周罗雄, 李博岩. 分析测试学报), **2020**, 39(10): 1196–1203.
- [20] DB53/T 497–2013. Tobacco and Tobacco Products–Main Chemical Composition Indicators–Guidelines for the Establishment and Validation of Near-infrared Calibration Models. Local Standards in Yunnan Province(烟草及烟草制品 主要化学成分指标 近红外校正模型建立与验证导则. 云南省地方标准).
- [21] Galvão R K H, Araujo M S U, José G M, Pontes M J C, Silva E C, Saldanha T C B. *Talanta*, **2005**, 67(4): 736–740.

(责任编辑: 盛文彦)