第39卷，第5期
2019年5月

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

Vol 39，No. 5，pp1428-1432
May，2019

窗口竞争性自适应重加权采样策略的近红外特征变量选择方法

李 跑[1]，周 骏[2]，蒋立文[1]，刘 霞[1]，杜国荣[1,2*]

1. 湖南农业大学食品科学技术学院食品科学与生物技术湖南省重点实验室，湖南 长沙 410128
2. 上海烟草集团有限责任公司技术中心北京工作站，北京 101121

摘 要 通过消除光谱中的冗余信息变量，挑选出代表样品性质的特征变量代替全谱建立定量模型，可以提高近红外分析结果的准确性。基于进化论中适者生存原理的竞争性自适应重加权采样(CARS)算法因具有计算速度快、筛选得到的特征波长少等优点，在近红外特征变量筛选方面得到了广泛的应用。然而该方法在计算过程中容易出现校正集和验证集结果不一致情况。这是因为算法过于强调校正集交叉验证结果，且并未考虑相邻变量之间的协同作用。为了建立更加稳健的变量筛选方法，通过结合“窗口”以及CARS算法的优势，提出了一种基于窗口竞争性自适应重加权采样(WCARS)策略的近红外特征变量筛选方法，并将其应用于复杂植物样品近红外光谱与其化学成分含量之间的建模分析。采用WCARS方法可以实现准确定量分析，且通过与竞争性自适应重加权采样(CARS)方法结果相比较，WCARS方法得到的校正集和预测集结果一致，在一定程度上减少了过拟合问题的出现。该策略能有效增强特征变量选择的稳健性，提高了定量模型的可信度，具有一定的应用价值。

关键词 近红外光谱仪；化学计量学；窗口竞争性自适应重加权采样

中图分类号：O657.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)05-1428-05

引 言

随着近红外仪器和化学计量学方法的飞速发展，近红外光谱得到了广泛应用。近红外光谱反应的是物质吸收的倍频与合频信息，包含了绝大多数类型有机物组成和分子结构的丰富信息。与传统方法相比，近红外光谱具有穿透力强，无需复杂前处理操作，不破坏样品，可通过光纤进行远距离在线检测等优点，因此被广泛用于食品、医药、烟草和环境等多个领域复杂样品的快速分析[1-3]。由于近红外光谱谱峰较宽，实际样品中各种成分的吸收峰重叠严重，近红外光谱定性和定量分析必须通过建立多元校正模型来实现。然而，其光谱中通常包含大量冗余信息的波长变量，且变量之间存在较为严重的共线性关系，如果直接采用全谱建立模型，不仅会增大模型的计算量，且冗余变量会降低模型的准确性[4]。因此需要通过消除光谱中的冗余信息变量，挑选出代表样品性质特征的变量代替全谱去建立定量模型，以提高近红外分析的准确性[5-7]。

现阶段常见的近红外特征变量选择方法可以分为波段选择和波长点筛选两大类。前者主要是通过采样窗口或间隔方式将数据进行分段处理以得到连续的波长数据，主要包括间隔偏最小二乘(interval partial least squares，iPLS)[8]和移动窗口偏最小二乘(moving window partial least squares，MW-PLS)[9]；后者通过搜索方法及评价变量重要性程度的标准以得到最优的变量子集，主要有竞争性自适应重加权采样(competitive adaptive reweighted sampling，CARS)[10]、蒙特卡罗-无信息变量消除(Monte Carlo uninformative variable elimination，MC-UVE)[11]、随机检验-偏最小二乘回归(randomization test-partial least squares regression，RT-PLS)[12]和改进CARS算法(modified competitive adaptive reweighted sampling，MCARS)[13]等。其中基于进化论中适者生存原理的CARS算法因具有计算速度快、筛选得到的特征波长少等优点，在近红外特征变量筛选方面得到了广泛的应用[14-15]。然而该方法在计算过程中容易出现校正集和验证集结果不一致情况。这是因为CARS算法过于强调校正集交叉验证结果，且并未考虑相邻变量之间的协同作用。

收稿日期：2018-03-06，**修订日期**：2018-07-26

基金项目：国家自然科学基金项目(31601551，31671931)，湖南农业大学引进人才科学基金项目(15YJ08)和湖南农业大学青年科学基金项目(16QN24)资助

作者简介：李 跑，1989年生，湖南农业大学食品科学技术学院讲师 e-mail：lipao@mail.nankai.edu.cn

*通讯联系人 e-mail：nkchem09@mail.nankai.edu.cn

为了建立更加稳健的特征变量选择方法，考虑相邻变量之间的协同作用，通过结合“窗口”以及 CARS 算法的优势，提出了一种基于窗口竞争性自适应重加权采样(window competitive adaptive reweighted sampling，WCARS)策略的近红外特征变量选择方法。针对复杂植物样品近红外光谱数据集，将 WCARS 方法应用于近红外有效特征变量的筛选，以建立简洁、稳定的定量模型，并将其模型结果与传统 CARS 算法的结果进行了比较。

1　实验部分

1.1　WCARS 算法原理

如果一个波长与目标组分相关，那么与之相邻的波长也应该与目标组分相关。因此 WCARS 算法把近红外变量沿着波长方向均等分为多个“窗口”。对数据进行偏最小二乘(partial least squares，PLS)[16]计算，统计每个“窗口”内的回归系数，使用其绝对值的均值作为是否保留该窗口变量的依据，选择剩余变量建立 PLS 模型并计算交互验证均方根误差(root mean squared error of cross validation，RMSECV)，将 RMSECV 值最小的变量集合作为最优变量子集。算法具体步骤如下：

Step 1：沿着波长方向将变量分为 N 个“窗口”，初始化 $i=0$；

Step 2：$i=i+1$；

Step 3：使用保留变量($i=1$ 时为所有变量)建立 PLS 模型，得到回归系数；

Step 4：统计保留的($N-i+1$)个“窗口”区间内回归系数绝对值的均值；

Step 5：删除回归系数绝对值均值最小对应窗口内的变量；

Step 6：使用 K 折交叉验证，检验保留窗口变量，计算其对应模型的 $RMSECV_i$；

Step 7：如果 i 小于 $N-1$，返回 Step 2，重复 Step 2—7，否者结束循环；

Step 8：使用 $RMSECV_i$ 最小值对应的保留窗口变量建立 PLS 模型。

我们公开了 WCARS 的源程序以及相关数据，具体下载地址如下：https://github.com/nkchem09/MVCtools

1.2　数据处理

研究使用的数据集为上海烟草集团有限责任公司技术中心北京工作站提供的晒红烟近红外数据以及其常见化学成分包括糖、还原糖、总植物碱、总氮和蛋白质的含量。样品共 131 个，使用带漫反射配件的布鲁克 MPA 型光谱仪(Bruker Optics Inc.，Ettlingen，Germany)采集光谱。采用漫反射模式，波数范围为 3957～11 995 cm^{-1}，间隔约为 4 cm^{-1}，共 2 085个波长点。扫描次数为 64 次。图 1 为样品的近红外漫反射光谱图。从图 1 可以看出原始光谱中存在明显的基线漂移和噪声，因此在建立模型之前采用 Savitzky-Golay 一阶导数(窗口为 17)方法对原始光谱数据进行预处理。

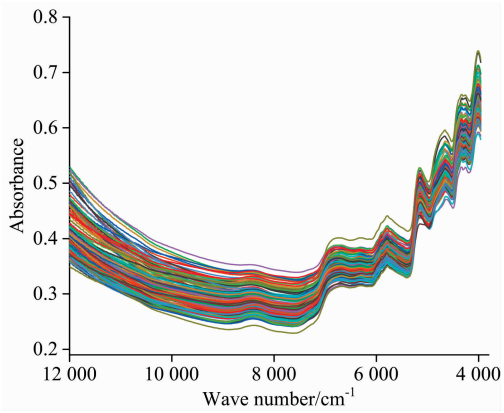

图 1　样品的近红外光谱

Fig. 1　Near-infrared spectra of samples

2　结果与讨论

2.1　CARS 计算结果

利用 Kennard-Stone 算法把 131 个晒红烟样品分建模集样品和预测集样品，使用 87 个建模集样品建立化学指标的模型，再用剩余 44 个预测集样品用于模型的验证。采用 PLS 方法建立定量模型。此外，为了提高建模结果的准确性，采用 CARS 算法选择合适变量以提高定量结果的准确性。表 1 总结了 100 次重复计算的 CARS-PLS 模型的结果。从表中数据可以看出，所有化学指标 RMSECV 均小于预测误差均方根(root mean square error of prediction，RMSEP)，表明校正集结果要优于更为重要的预测集结果。如果直接采用 CARS-PLS 建立的模型，容易导致模型的过拟合问题出现以及定量结果可信度的下降。

表 1　CARS-PLS 模型的计算结果

Table 1　Calculation results of quantitative models built with CARS-PLS method

	Factor	Number of variables (σ)[a]	RMSECV (σ)	RMSEP (σ)
Reducing sugar/%	10	43 (25.895)	0.473 (0.011)	0.703 (0.021)
Total sugar/%	10	52 (23.289)	0.516 (0.024)	0.779 (0.044)
Total alkaloid/%	10	38 (10.584)	0.317 (0.008)	0.412 (0.021)
Total nitrogen/%	10	33 (9.696)	0.126 (0.002)	0.228 (0.004)
Protein/%	10	28 (17.229)	0.568 (0.013)	0.672 (0.023)

[a]σ is the standard deviation of the 100 results

2.2 WCARS 参数确定

在 WCARS 算法中引入了“窗口”概念。首先需要选择一个合适的“窗口”大小。由于每一个近红外光谱信号包含有 2 085个波长点，因此对于所有化学指标数据我们分别取 10～190 个“窗口”建立模型。图 2 分别显示了 RMSECV 值随着“窗口”数的变化情况以及 25～75 分位范围、1.5 倍四分位距、中位数、平均值和奇异值。从图中可以很明显地看出，当“窗口”数大于 10 时，RMSECV 值基本不随着“窗口”数的改变而发生显著性变化，表明窗口数不会影响到结果的准确性。此外，虽然不同“窗口”结果分布范围、四分位间距略有差异，但这种差异较小，偏度及中位线接近，说明不同“窗口”数据总体特征基本相似。因此，我们选取“窗口”数为 100 作为下一步研究的参数。

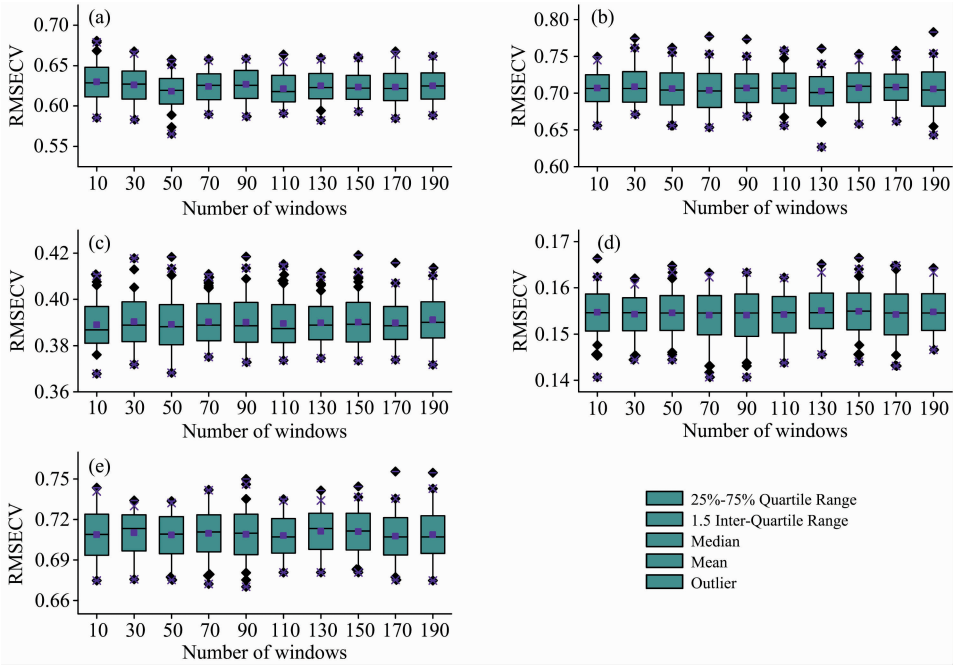

图 2 RMSECV 值随着不同“窗口”数目的变化情况

(a)—(e)分别为还原糖、总糖、总植物碱、总氮和蛋白质

Fig. 2 Variation of RMSECV with the different numbers of the window

(a)—(e) are the results of reducing sugar, total sugar, total alkaloid, total nitrogen and protein, respectively

2.3 因子数的选择

在建立 PLS 模型之前需要选择合适的因子数。使用留一交叉验证方法得到 RMSECV 随因子数变化的走势图，然后找到 RMSECV 值最低点对应的因子数。以还原糖含量的测定为例，对近红外光谱进行留一交叉验证，得到该光谱的 RMSECV 随因子数变化的趋势图(图 3)。从图中可以看出，RMSECV 最低点对应的因子数为 10～11，因此在测定还原糖含量时，建立定量模型所用的最佳因子数应该为 10。依据同样的方法我们选择了其他 4 个指标的最佳因子数。

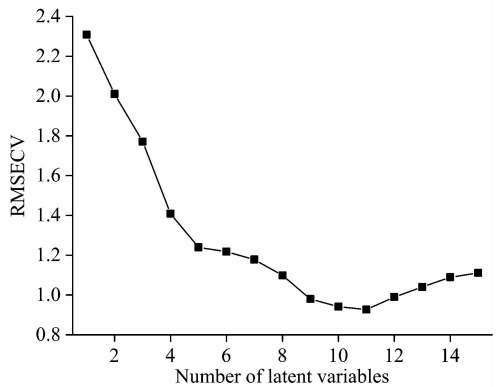

图 3 RMSECV 随因子数变化的趋势图

Fig. 3 A plot of RMSECV versus factors

2.4 WCARS 定量结果

分别利用 CARS 和 WCARS 算法筛选出与化学指标相关的特征波长变量建立 PLS 检测模型，并对验证集样本进行预测，比较两种检测模型的预测效果。图 4 为 CARS 算法和 WCARS 算法筛选的变量分布。由图可以看到两种方法筛选的变量分布较为相似，且 WCARS 所选窗口包含了所有 CARS 所选变量。此外，CARS 所选变量的相邻波长也纳入到 WCARS 所筛选的变量中。

表 2 总结了 100 次重复计算的 WCARS-PLS 模型交叉验证结果。从表中可以看出，虽然 WCARS 方法比传统 CARS 方法增加了“窗口”数这一个参数，增加了其操作难度，但是通过结合“窗口”和 CARS 算法的优势，WCARS 方法能很好地筛选出特征波长变量。虽然该方法在某些组分的预测效果上不如 CARS 方法，但是 WCARS 方法得到的所有化学指标 RMSECV 与 RMSEP 值较为一致，避免了传统 CARS 方法过拟合的风险。此外，为了更好说明方法的准确性，我们将

WCARS 方法结果与 MC-UVE 方法结果进行了对比。表 3 总结了 100 次重复计算的 MC-UVE-PLS 模型交叉验证结果。由表 2 和表 3 结果可知，WCARS 方法结果与 MC-UVE 结果较为相似，在总糖和总氮指标上 WCARS 方法结果略优于 MC-UVE 结果。综上所述，WCARS 方法建立的校正集模型对验证集样品的预测效果较好，模型可以用于定量分析。

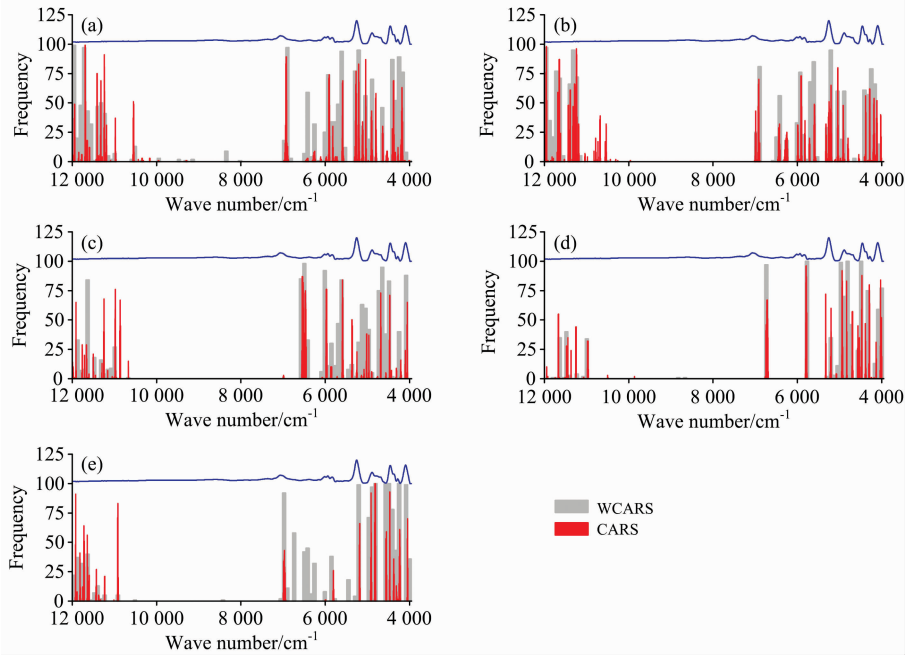

图 4　CARS 算法和 WCARS 算法筛选的变量分布

(a)—(e)分别为还原糖、总糖、总植物碱、总氮和蛋白质

Fig. 4　Variable distribution of CARS and WCARS algorithms

(a)—(e) are the results of reducing sugar, total sugar, total alkaloid, total nitrogen and protein, respectively

表 2　WCARS-PLS 模型的计算结果

Table 2　Calculation results of quantitative models built with WCARS-PLS method

	Factor	Number of windows (σ)[a]	RMSECV (σ)	RMSEP (σ)
Reducing sugar/%	10	19 (5.116)	0.621 (0.017)	0.683 (0.034)
Total sugar/%	10	17 (4.824)	0.706 (0.022)	0.707 (0.085)
Total alkaloid/%	10	15 (5.052)	0.392 (0.010)	0.424 (0.018)
Total nitrogen/%	10	12 (3.015)	0.155 (0.004)	0.177 (0.007)
Protein/%	10	15 (4.974)	0.708 (0.014)	0.619 (0.031)

[a]σ is the standard deviation of the 100 results

表 3　MC-UVE-PLS 模型的计算结果

Table 3　Calculation results of quantitative models built with MC-UVE-PLS method

	Factor	Number of windows (σ)[a]	RMSECV (σ)	RMSEP (σ)
Reducing sugar/%	10	195 (39.319)	0.624 (0.019)	0.666 (0.025)
Total sugar/%	10	178 (32.035)	0.706 (0.050)	0.774 (0.023)
Total alkaloid/%	10	294 (36.404)	0.393 (0.005)	0.394 (0.003)
Total nitrogen/%	10	181 (49.422)	0.160 (0.009)	0.237 (0.003)
Protein/%	10	294 (34.836)	0.706 (0.001)	0.582 (0.004)

[a]σ is the standard deviation of the 100 results

3　结　论

为了建立稳健的特征波长选择方法，结合“窗口”与 CARS 方法的优势，提出了一种 WCARS 方法用于近红外特征变量的筛选，并应用于复杂植物样品近红外光谱与其化学成分含量之间的建模分析。通过与 CARS 方法结果对比表明，WCARS 方法得到的校正集和预测集的结果较为一致，在一定程度上避免了过拟合问题的出现。该策略能有效增强特征波长变量选择的准确性和稳定性，提高了模型的可信

度，具有一定的应用价值。

References

[1] ZHANG Yi-ting, WANG Cui-cui, FAN Meng-li, et al(张伊挺，王翠翠，樊梦丽，等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2017, 36(12): 4100.
[2] Wang C C, Cai W S, Shao X G. Analytical Letters, 2017, 51(4): 537.
[3] Wei H Y, Li H, Liu P, et al. Spectroscopy Letters, 2017, 50(7): 470.
[4] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立，袁洪福，陆婉珍). Progress Chemistry(化学进展), 2004, 16(4): 528.
[5] Tan Z, Lou T T, Huang Z X, et al. Journal of Agricultural and Food Chemistry, 2017, 65: 6274.
[6] Han X, Huang Z X, Chen X D, et al. Fuel, 2017, 207: 146.
[7] Zhang C X, Bian X H, Liu P, et al. Chemometrics and Intelligent Laboratory Systems, 2016, 161: 43.
[8] Saudland A, Wagner J, Nielsen J P, et al. Applied Spectroscopy, 2007, 54(3): 413.
[9] Ranzan C, Trierweiler L F, Hitzmann B, et al. Chemometrics and Intelligent Laboratory Systems, 2015, 142: 78.
[10] Li H D, Liang Y Z, Xu Q S, et al. Analytica Chimica Acta, 2009, 648(1): 77.
[11] Cai W S, Li Y K, Shao X G. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 188.
[12] Xu H, Liu Z C, Cai W S, et al. Chemometrics and Intelligent Laboratory Systems, 2009, 97(2): 189.
[13] Han X, Tan Z, Huang Z X, et al. Analytical Methods 2017, 9: 3720.
[14] Nie L X, Dai Z, Ma S C. Analytical Letters, 2016, 49(14): 2259.
[15] Wu L J, Wang B X, Yin Y F, et al. Analytical Letters, 2016, 49(14): 2290.
[16] Li H D, Xu Q S, Liang Y Z. Chemometrics and Intelligent Laboratory Systems, 2018, 176: 34.

A Variable Selection Approach of Near Infrared Spectra Based on Window Competitive Adaptive Reweighted Sampling Strategy

LI Pao[1], ZHOU Jun[2], JIANG Li-wen[1], LIU Xia[1], DU Guo-rong[1,2*]

1. College of Food Science and Technology, Hunan Agricultural University, Changsha 410128, China
2. Beijing Work Station, Technology Center, Shanghai Tobacco Group Co., Ltd., Beijing 101121, China

Abstract Variable selection plays an important role in the quantitative analysis of near infrared spectra. The accuracy of near infrared spectroscopy can be improved by eliminating the redundant variables and selecting the characteristic variables. Competitive adaptive reweighted sampling (CARS) method is a newly developed strategy for wavelength selection by employing the principle "survival of the fittest" on which Darwin's Evolution Theory is based. The number of selected wavelengths by CARS is much smaller than those of other methods with fast calculating speed and high accuracy. However, it is easy to get inconsistent results between the calibration and validation set due to the excessive attention on the cross validation results. In order to develop a robust variable selection method, by combining the advantages of CARS and "window", a new tactic called window competitive adaptive reweighted sampling (WCARS) is employed to select characteristic variables and applied to the analysis of the near infrared spectra of the complex plant samples and the contents of the chemical components. Compared with the results of CARS method, accurate quantitative results can be obtained by the WCARS method. Furthermore, the results of correction set are consistent with those of the prediction set, and the problem of overfitting can be avoided. The results show that WCARS tactic can efficiently improve the accuracy and stability of variables selection and optimize the precision of prediction model, which has a certain application value.

Keywords Near infrared spectroscopy; Chemometrics; Window competitive adaptive reweighted sampling

(Received Mar. 6, 2018; accepted Jul. 26, 2018)

* Corresponding author