第40卷，第10期
2020年10月

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

Vol. 40, No. 10, pp3254-3259
October, 2020

不同平滑集成 CARS 算法在红茶等级光谱判别中的应用

袁 荔，施 斌，于建成，唐天宇，袁 园，唐延林*

贵州大学物理学院，贵州 贵阳 550025

摘 要 移动窗口平滑集成 CARS 算法(MWS-ECARS)是一种稳定的特征变量提取算法。在前人研究的基础上，提出了两种基于不同窗口平滑算法改进的 MWS-ECARS 对红茶光谱降维，并与原始的 MWS-ECARS、常用的连续投影算法(SPA)、竞争性自适应重加权算法(CARS)、移动窗口偏最小二乘法(MW-PLS)比较，建立偏最小二乘算法回归模型(PLSR)，选择出最优红茶等级判别模型。两种改进的 MWS-ECARS 方法分别是窗口高斯滤波平滑集成 CARS(gaussian filter ECARS，GF-ECARS)、窗口中值滤波平滑集成 CARS(median filter ECARS，MF-ECARS)。CARS 算法运行 n 次(该研究 $n=1\ 000$)，整合波长及其对应的挑选频率并用不同的窗口平滑算法对挑选频率进行平滑，窗口宽度均为 3～31，窗口步长均为 2；将通过不同窗口宽度和平滑算法平滑过的挑选频率进行阈值的设定，起始阈值及步长均为 20；最后选择出挑选频率大于阈值的波长，建立 PLSR 模型，以预测集相关系数(R_p^2)为判断因子，R_p^2 越接近 1，说明建立的模型预测能力更为准确。结果表明：改进后的 GF-ECARS 算法提取的特征变量建立红茶等级判别模型的结果最好，R_p^2 达到 0.969 2。原因是在窗口高斯滤波平滑算法中，随着窗口宽度增大，其曲线上各点的振幅差距会变小。在高斯算法加权平均的过程中，不容易出现将低频的波长与高的权值相联系。在实际应用中，往往会出现有效波段的挑选频率较低的情况，可以通过选择窄窗口宽度的高斯滤波对其进行平滑。另外，高斯曲线的特征能使高斯滤波很好的保护窗口边缘图像的细节。虽然 MF-ECARS 算法的建模结果比原始 MWS-ECARS 略差，但其 R_p^2 仍然达到了 0.96 以上，表明改进后的算法能提高原始模型的预测能力。不同窗口平滑算法的 MWS-ECARS 提取特征变量不同，但随着平滑窗口宽度的增加，特征变量区间连续性都在增强，数目均在减少。三种 MWS-ECARS 算法的预测集相关系数都显示出它们比常用的 SPA，CARS 和 MWPLS 三种降维算法更有效，更稳定。为光谱数据的选择性降维算法研究提供参考。

关键词 移动窗口平滑集成 CARS；可见-近红外光谱；红茶；等级

中图分类号：O433 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)10-3254-06

引 言

化学计量学是光谱分析技术中的重要组成部分，它包括光谱预处理，光谱降维，光谱定量、定性模型建立等内容。在实际应用中，由于光谱数据可能具有信号强度弱、信号重叠、外界噪声干扰大等问题，导致分析结果精度低，稳定性差[1]。在此背景下，运用光谱降维算法与日俱增。常用的降维方法有主成分分析(principal component analysis，PCA)、竞争自适应重加权采样(the competitive adaptive reweighting algorithm，CARS)、连续投影算法(successive projections algorithm，SPA)、移动窗口偏最小二乘法(the moving window partial least squares method，MWPLS)等。Omar 等使用 PCA 算法对烟草光谱降维，结合偏最小判别分析算法(PLS-DA)对烟草商标进行判别[2]。Leqian 等基于可见-近红外光谱利用蚁群算法和 CARS 算法检测与分类葡萄酒的品质参数[3]。Dong 等采用协同区间偏最小二乘法(partial least square regression，PLSR)和极值学习机结合自适应增强算法将红茶的发酵质量与近红外光谱建立定量分析模型，结果表明该算法能够对红茶发酵品质实行在线监测[4]。Song 等利用 Haar，Sym，Coif 和 Bior 小波对遗传算法选择的光谱数据再次压缩，然后根据各小波函数压缩的变量建立 PLS 模型[5]。

红茶是全球范围内饮用最广的茶叶之一，遍及亚洲、非洲、欧洲等各个国家。红茶带有独特的物质成分(茶黄素，茶

收稿日期：2019-08-22，**修订日期**：2019-12-26

基金项目：国家自然科学基金项目(11164004，11864006)和贵州省光子科学与技术创新人才团队项目(20154017)资助

作者简介：袁 荔，女，1995 年生，贵州大学物理学院硕士研究生 e-mail：zhengguxu@foxmail.com

*通讯联系人 e-mail：tylgzu@163.com

红素，茶褐素等)，使其受到医药、食品和各个相关领域的关注与研究[6-7]。Dey 等发现口服红茶提取物(BTE)会改变实验性白化大鼠妊娠期和哺乳期大鼠血液和肝脏的参数[8]。Ji 等发现红茶多糖可以显著抑制 H22 肿瘤细胞的生长，有效保护肿瘤小鼠的胸腺和脾脏[9]。Lantano 等通过对不同茶的冷热浸泡，研究出提高绿茶与红茶中活性物质含量的新的浸渍方法[10]。Dash 等在海水体系中利用首次冲泡后产生的红茶残渣生产生物乙醇，以减少生物乙醇工业对淡水的消耗[11]。

尽管现有的光谱数据降维算法层出不穷，研究人员仍然不断在改善这些算法的不利之处[12-13]。在中国农业大学宋相中提出的基于移动窗口平滑集成策略的特征波段挑选算法(moving window smoothing ensemble CARS，MWS-ECARS)基础上，考虑在该算法中使用不同的窗口平滑算法来挑选特征变量，并筛选出用于红茶等级判别效果最优的光谱数据降维算法，为光谱数据的选择性降维提供参考。

1　实验部分

1.1　红茶样品的制备及可见-近红外光谱的采集

微型植物粉碎机，天津市泰斯特仪器有限公司生产。荷兰 Avantes 公司生产的 Avaspec-2408 标准型光纤光谱仪，测定范围为 350～1 100 nm，光谱采样间隔为 4 cm^{-1}，扫描次数为 10 次，探头视场角为 15°。

5 个等级的红茶样本均购自贵州省太升茶行，分别为特级、一级、二级、三级、四级，每个等级茶叶样品数分别为 40 个，每个等级人为随机以 3∶1 的比例划分为校正集与预测集，最后得到校正集 150 个，预测集 50 个样品。校正集用于建模，预测集用于验证模型的可靠性。将 200 个红茶样本通过微型植物粉碎机粉碎后，用 40 目标准分样筛筛滤，最后盛放在高为 0.4 cm、直径为 2.2 cm 的黑色培养皿中，压平样品表面，减少粗糙表面造成的光能量损失。在密不透光的环境中测样品光谱，保持光纤头距离样品表面 1.6 cm，每次测量先进行“白板”校正，后采集样品光谱，以减少环境和仪器带来的误差。

使用 The Unscrambler X(CAMO Software AS 公司)及 Matlab(2015)(MathWorks 公司)软件进行数据的处理与分析，MWS-ECARS 计算过程由自编 Matlab 程序和 The Unscrambler X 软件共同实现。

1.2　MWS-ECARS 原理

MWS-ECARS 算法原理是：采用窗口平滑算法对多次重复运行 CARS 得到的波长累积被选频率做平滑处理，以保留高频波长点及其附近的有效波长；通过设定频率阈值，将大于阈值的波长选出作为特征波长，由于特征波长点邻近的有效波长频率往往略低于高频特征波长，也会保留，所以最后被挑选出来的特征变量通常会形成特征波段[9]。

2　结果与讨论

2.1　红茶样本的可见-近红外光谱

200 个红茶样本的可见-近红外光谱如图 1 所示，光谱范围为 350～1 100 nm。由于在 350～400 和 1 000～1 100 nm 波段内的光谱受噪声影响较大，选取 400～1 000 nm 范围的波段参与后续的鉴别建模。

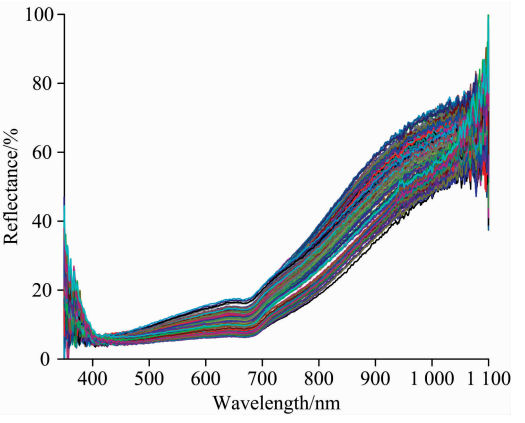

图 1　红茶的可见-近红外光谱

Fig. 1　Visible-near infrared spectra of black tea

2.2　光谱数据预处理

在采集光谱的过程中，为了尽可能地减小噪声的影响，实验选择移动均值平滑(MA-Smoothing)，高斯滤波平滑(GF-Smoothing)，中值滤波平滑(MF-Smoothing)，卷积平滑(SG-Smoothing)，去趋势(De-trending)和多元散射校正(MSC)6 种平滑算法对原始光谱平滑，GF-Smoothing 的 R_p^2 最大，为 0.947 7，所以用 GF-Smoothing 预处理后的光谱数据进行后续的数据分析。

表 1　不同预处理方式与 PLSR 建模结果

Table 1　The PLSR model result of different pretreatments

预处理	训练集		预测集	
	RMSEC	R_c^2	RMSEP	R_p^2
Raw	0.246 2	0.970 0	0.410 4	0.916 0
MA Smoothing	0.191 4	0.964 8	0.238 5	0.945 9
GF Smoothing	0.308 3	0.962 3	0.364 9	0.947 7
MF Smoothing	0.233 4	0.947 4	0.275 2	0.927 9
SG Smoothing	0.328 8	0.962 9	0.384 8	0.932 7
De-trending	0.412 8	0.951 9	0.509 5	0.927 8
MSC	0.453 8	0.967 4	0.588 5	0.945 8

2.3　光谱数据降维

2.3.1　基于 MWS-ECARS 的光谱数据降维

设定 CARS 运行次数为 1 000，三种平滑算法的频率平滑窗口宽度均为 3～31，宽度步长为 2，频率阈值为 20～700，阈值步长为 20。由于篇幅限制，仅列出特征变量挑选变化明显的结果，黑色曲线是预处理以后的光谱曲线，彩色柱形图是特征变量区域，柱形图与黑色曲线重叠区域是算法选择的特征变量。图 2 是基于 MA-ECARS 挑选的特征变量，平滑窗口宽度分别是 3，17 和 31，阈值均为 140，窗口宽度较小时，特征变量区间小且数目多，覆盖范围广。随着平滑窗口宽度增加，特征变量离散程度逐渐降低，特征波段区间变大，且大窗口宽度的区间数目比小窗口少。在三种 MWS-

ECARS 算法中(MA-ECARS，MF-ECARS，GF-ECARS)都不同程度上展现出这种规律。

图 3 是以窗口中值滤波为平滑算法的 MF-ECARS 提取的特征变量部分情况，平滑窗口宽度为 5，15 和 23，阈值均为 80。从图中可知，随着窗口宽度的增加，提取的变量区间数目减少，连续性增强。但选择的平滑算法不同，挑选的特征变量仍与 MA-ECARS 提取有所不同。

图 4 是以窗口高斯滤波为平滑算法的 GF-ECARS 挑选特征变量部分情况，平滑窗口宽度分别为 5，19 和 31，阈值均为 200。GF-ECARS 提取特征变量的情况也有所不同。

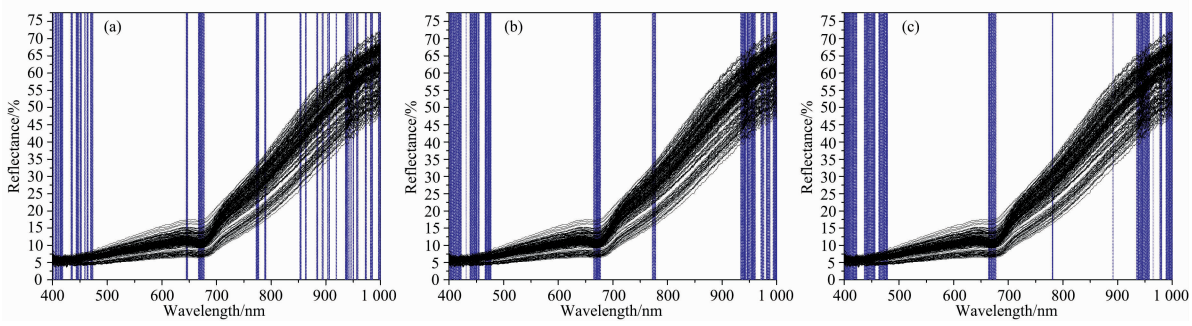

图 2 基于不同窗口宽度的 MA-ECARS 挑选的特征变量

(a)：窗口宽=3；(b)：窗口宽=17；(c)：窗口宽=31

Fig. 2 Characteristic variables selected by MA-ECARS based on different window widths

(a)：Window width=3；(b)：Window width=17；(c)：Window width=31

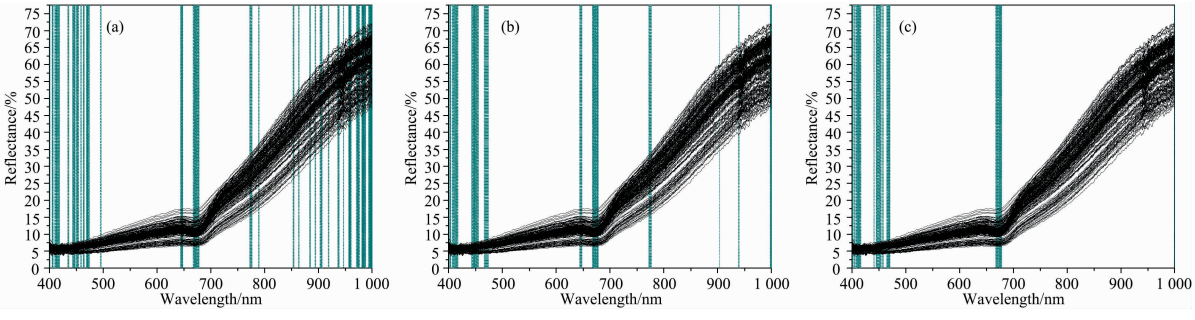

图 3 基于不同窗口宽度的 MF-ECARS 挑选的特征变量

(a)：窗口宽=5；(b)：窗口宽=15；(c)：窗口宽=23

Fig. 3 Characteristic variables selected by MF-ECARS based on different window widths

(a)：Window width=5；(b)：Window width=15；(c)：Window width=23

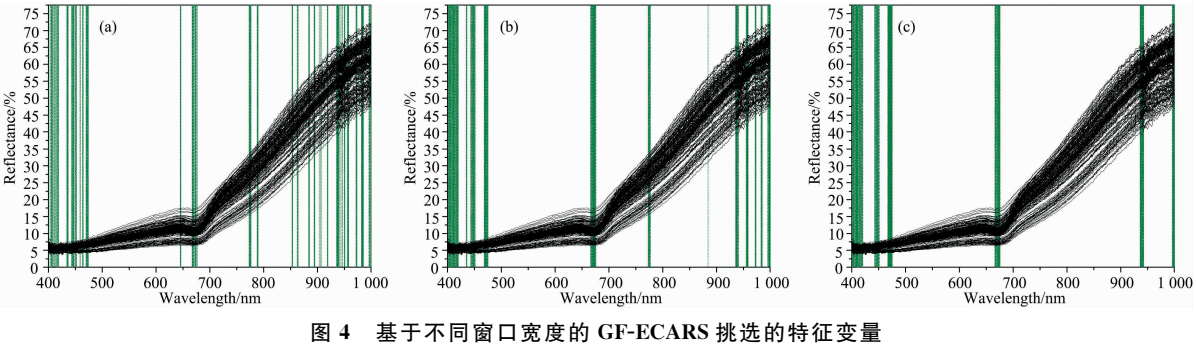

图 4 基于不同窗口宽度的 GF-ECARS 挑选的特征变量

(a)：窗口宽=5；(b)：窗口宽=19；(c)：窗口宽=31

Fig. 4 Characteristic variables selected by GF-ECARS based on different window widths

(a)：Window width=5；(b)：Window width=19；(c)：Window width=31

2.3.2 基于连续投影算法(SPA)和竞争自适应重加权算法(CARS)的光谱数据降维

使用 SPA 算法和 CARS 算法从预处理后的光谱数据中挑选出特征波长，分别如图 5，图 6 所示。SPA 挑选出 5 个特征波长：400.29，430.90，472.54，673.17 和 943.50 nm。CARS 挑选出 93 个特征波长，几乎分布在光谱变化明显的位置。

2.3.3 移动窗口偏最小二乘法(MWPLS)挑选特征波段

基于 MWPLS 算法挑选的特征波段如表 2 所示。设定窗口宽度为 90～210，窗口步长取 10，主成分数目为 4～10。对

于每一个特定宽度的窗口，在主成分数为 10 时，交叉验证均方根误差为最小值。由表中知道，选择的特征变量为 796.69～913.73 nm 区间，因为此时预测集均方根误差(RMSEP)最小。

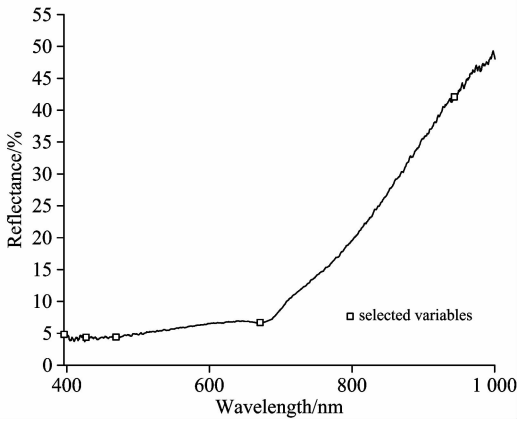

图 5　SPA 挑选的特征波长

Fig. 5　Characteristic wavelengths selected by SPA

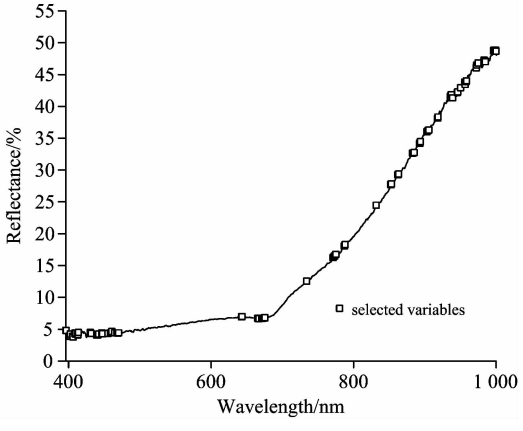

图 6　CARS 挑选的特征波长

Fig. 6　Characteristic wavelengths selected by CARS

表 2　基于移动窗口偏最小二乘法挑选(MWPLS)的特征波段

Table 2　Characteristic bands selected by moving window partial least squares (MWPLS)

窗口宽度	光谱范围	主因子数	RMSEP
90	862.14～912.07	10	0.288 3
100	861.58～917.04	10	0.276 8
110	859.91～920.91	10	0.270 8
120	854.34～920.91	10	0.264 1
130	839.83～912.07	10	0.255 9
140	833.68～911.53	10	0.246 7
150	833.68～917.04	10	0.237 2
160	833.68～922.57	10	0.230 1
170	828.09～922.57	10	0.227 3
180	816.89～917.04	10	0.220 3
190	811.29～917.04	10	0.214 8
200	810.16～921.46	10	0.210 8
210	796.69～913.73	10	0.204 2

2.4　建模结果分析

基于四种降维方法挑选出的有效波长(波段)建立偏最小二乘回归模型(PLS)，结果如表 3 所示。同时可以从 R_p^2 看出，GF-ECARS 相关系数最大，说明改进的 MWS-ECARS 算法能提高红茶等级的预测能力。尽管基于三种不同的窗口平滑算法的 MWS-ECARS 挑选的特征变量不一样，但是建模的效果都是优秀且稳定的，表明不同的 MWS-ECARS 均能够不同程度上弥补具有随机参数 CARS 而导致模型不稳定的不足。从变量选择的情况来看，SPA 算法提取的变量数过少，且是离散的波长点，MWS-ECARS 提取的几乎是特征波段。有效波长附近的波长也具有物质的信息，相较于特征波长来说，波段建模的结果会更好。MWPLS 提取的有效波段处于近红外区域，仅为一段有效波段，相较而言，更突出 MWS-ECARS 降维效果好，覆盖信息全面的优点。所以实验选择 MWS-ECARS 中的 GF-ECARS 作为实验模型。图 7 是 GF-ECARS-PLSR 模型对 50 个预测集样本的预测等级与实际等级的对比情况，从斜率和截距可以看出所建红茶等级模型是有效的。

表 3　不同特征变量挑选方法与 PLSR 建模

Table 3　The PLSR model of different selection methods of characteristic variables

特征波长选择方法	变量数目	训练集		预测集	
		RMSEC	R_c^2	RMESP	R_p^2
RAW	1 057	0.308 3	0.962 3	0.364 9	0.947 7
SPA	5	0.145 4	0.913 7	0.622 3	0.808 1
CARS	120	0.181 9	0.983 5	0.321 2	0.9 470
MWPLS	211	0.368 1	0.932 3	0.429 7	0.910 2
MA-ECARS	86	0.246 3	0.969 7	0.265 9	0.965 5
MWS-ECARS MF-ECARS	142	0.242 3	0.970 6	0.267 7	0.964 4
GF-ECARS	96	0.232 2	0.973 1	0.251 7	0.969 2

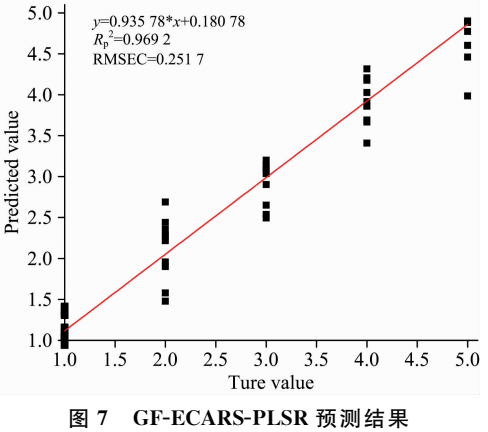

图 7　GF-ECARS-PLSR 预测结果

Fig. 7　The prediction results of GF-ECARS-PLSR

3　结　论

为了较大程度上消去外界因素对模型建立的干扰，将获得的 200 个样本光谱进行 6 种方法预处理，其中高斯滤波平

滑的建模效果最好，预测集相关系数最高，所以选择经高斯滤波平滑后的数据进行后续的实验处理。其次，使用 MWS-ECARS，SPA，CARS 以及 MWPLS 4 种数据降维方式对预处理后的数据提取特征变量。结果显示，MWS-ECARS 算法中的 GF-ECARS 算法提取的特征变量建立的偏最小二乘回归模型结果最好，相关系数达到 0.969 2。

对于四种不同的降维算法，SPA 由于选择的特征变量数目过少，失去了样品光谱中部分细节信息。通过 CARS 得到的特征变量尽管建模的效果不错，但由于该算法中引入了随机参数，每次运行后得到的特征变量和数目都不相同，所以建立的定性定量模型稳健性较差。同时，从光谱建模的角度上发现用特征波段的建模结果通常比用特征波长好，因为具有样品信息的某一波长点邻近的部分波长也具有样品的光谱信息，所以，用 MWPLS 和改进的 MWS-ECARS 提取特征波段建模效果相对较好。MWPLS 选择的特征变量仅为某段光谱区域，不够全面，建模效果不会十分出色。改进的三种 MWS-ECARS 虽然提取的特征变量情况不同，但都在很大程度上覆盖了光谱信息，提取的特征波段区间大小可变，具有特征波长与波段同时选择，在一定程度上降低了变量的冗余性和保留了有效信息的连续性。尽管窗口平滑算法不同，但它们建模的结果都显示出 MWS-ECARS 的稳定性和优异性，对于基于可见-近红外光谱的红茶样品等级判别 GF-ECARS 算法是最合适的。在前人的基础上，提出基于不同窗口平滑算法的两种 MWS-ECARS 算法对红茶等级进行光谱判别是可行的。

References

[1] SONG Xiang-zhong, XIONG Yan-mei, ZHANG Lu-da, et al(宋相中，熊艳梅，张录达，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(S1): 181.
[2] Omar J, Slowikowski B, Boix A. Forensic Science International, 2019, 294: 15.
[3] Leqian H, Chunling Y, Shuai M, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 205: 574.
[4] Dong C, Zhu H, Wang J, et al. Food Science and Biotechnology, 2017, 26(4): 853.
[5] Song J, Li G, Yang X. J. Sci. Food Agric., 2019, 9: 4898.
[6] Li D X, Wang H, Wan X C. China Tea Scienc Society, 2005: 13.
[7] Singh B N, Prateeksha, Rawat A K S, et al. Critical Reviews in Food Science and Nutrition, 2017, 57(7): 1394.
[8] Dey A, Gomes A, Dasgupta S C. Pharmacognosy Magazine, 2017, 13(52): S769.
[9] Ji H Y, Dong X D, Yu S S, et al. Journal of Food Measurement and Characterization, 2019, 13: 1620.
[10] Lantano C, Rinaldi M, Cavazza A. Journal of Food Science and Technology, 2015, 52(12): 8276.
[11] Dash I, Bhaskar D, Bhawsar H, Sahoo M. Bioresource Technology Reports, 2018, 4: 209.
[12] CHEN Yue-yang, GAO Zhi-shan, YU Xiao-hui, et al(陈玥洋，高志山，郁晓晖，等). Journal of Applied Optics(应用光学), 2017, 38(1): 99.
[13] Wang S H, Zhao Y, Hu R, et al. Chinese Journal of Analytical Chemistry, 2019, 47(4): e19034.

Application of Different Smoothing Ensemble CARS Algorithm in Spectral Discrimination of Black Tea Grade

YUAN Li, SHI Bin, YU Jian-cheng, TANG Tian-yu, YUAN Yuan, TANG Yan-lin*
School of Physics, Guizhou University, Guiyang 550025, China

Abstract Moving window smoothing ensemble CARS (MWS-ECARS) is a stable algorithm for extracting characteristic variables. Based on the previous studies, two improved MWS-ECARS are proposed to reduce the dimension of black tea spectrum based on different window smoothing algorithms in this paper, and compared with the original MWS-ECARS, the commonly used successive projections algorithm (SPA), the competitive adaptive reweighting algorithm (CARS) and the moving window partial least squares method (MWPLS). A partial least square regression model (PLSR) was established to select the best black tea grade discrimination model. Two improved MWS-ECARS methods are Gaussian filter ECARS (GF-ECARS) and Median filter smoothing ECARS (MF-ECARS), respectively. The CARS algorithm runs n times (n=1 000 in this paper). The wavelength and its corresponding selected frequency are sorted out and different window smoothing algorithms are used to smooth the selection frequency. The window widths are all 3～31, and the window step sizes are all 2. The threshold is set through the selection frequency smoothed by different window widths and smoothing algorithm, and the starting threshold and step size are both 20. Finally, the wavelength whose selection frequency is higher than the threshold is selected and the PLSR model is established. The correlation coefficient of prediction set (R_p^2) is taken as the judgment factor. The closer R_p^2 is to 1, the

more accurate the established model is. The results show that the black tea grade discrimination model established by the extracted characteristic variables with the improved GF-ECARS algorithm is the best. The R_p^2 reaches 0.969 2. The reason is that the amplitude difference of each point on the curve will become smaller in the window Gaussian filtering smoothing algorithm as the window width increases. In the weighted average process of Gaussian algorithm, it is not easy to associate the low frequency wavelength with the high weight. In practical applications, the selection frequency of effective band is often low, which can be smoothed by selecting a Gaussian filter with narrow window width. In addition, due to the characteristics of the Gaussian curve, the Gaussian filtering algorithm can well protect the details of the window edge image. Although the modeling result of the MF-ECARS algorithm is slightly worse than the original MWS-ECARS, its R_p^2 still reaches over 0.96. This shows that the improved algorithm can improve the prediction ability of the original model. MWS-ECARS extraction feature variables are different based on different window smoothing algorithms. However, as the smoothing window width increases, the continuity of the extracted characteristic variables is enhanced and the number of extracted characteristic variables is reduced. The R_p^2 of the three MWS-ECARS algorithms all show that they are more effective and stable than the commonly SPA, CARS and MWPRS algorithms. This study can provide ideas for selective dimensionality reduction of spectral data.

Keywords Moving window smoothing ensemble CARS (MWS-ECARS); Visible-near infrared spectroscopy; Black tea; Grades

(Received Aug. 22, 2019; accepted Dec. 26, 2019)

* Corresponding author