光谱数据解析中的变量筛选方法

李艳坤[1*]，董汝南[1]，张　进[2]，黄克楠[3]，毛志毅[4]

1. 华北电力大学(保定)环境科学与工程系，河北省燃煤电站烟气多污染物协同控制重点实验室，河北 保定　071003
2. 贵州医科大学食品科学学院，贵州 贵阳　550025
3. 中国人民解放军陆军第八十二集团军医院，河北 保定　071000
4. 天津市建筑材料科学研究院有限公司，天津　300110

摘　要　如何从海量或高维数据中“提纯”出有用的信息，这是当前数据分析面临的一个巨大的挑战，也是当前研究的一个热点。变量筛选技术能够从众多、复杂的量测数据中提取出特征信息变量，达到简化多元模型乃至提高模型预测性能等目的。在光谱分析中，来自噪声等诸多因素的影响，量测数据会不可避免地包含干扰和无关信息变量，以及变量间存在的多重共线性，这些都会影响模型的稳健性和预测能力。近年来变量(波长)筛选方法在光谱解析领域的研究与应用中取得了较大的进展。结合国内外相关研究文献和作者的研究体会，不仅仅综述了近红外光谱，还综述了中红外光谱、拉曼光谱等众多筛选变量的方法的提出、特点、发展、类别、比较和近五年来在不同领域的应用进展。其中，评价变量重要性的参数及其标准或阈值的选择、搜索变量的策略和途径是变量筛选方法的关键。而且每种方法都具有各自的优势和局限性，实际使用中要根据方法自身特点结合目标体系的特征选择合适的方法。重点内容：(1)对比了光谱数据分析中常用的波长筛选和波段筛选方法；(2)对比了基于PLS模型参数的不同变量筛选方法的原理和特点；(3)根据搜索和筛选变量策略的不同将变量筛选方法进行分类评述。最后，围绕在解析实际复杂体系中变量筛选方法出现的过拟合、不稳定等问题进行了讨论并提出相应的解决措施，同时对变量筛选方法的研究趋势、发展前景和应用方向进行了展望。其中，新的评价变量重要性的判据和搜索变量的策略等工作仍需要展开深入地研究。期望本综述能够对光谱变量筛选的后续研究及应用起到积极的推动作用。

关键词　变量筛选；光谱数据；特征变量；冗余信息

中图分类号：O657.3　　**文献标识码**：R　　**DOI**：10.3964/j.issn.1000-0593(2021)11-3331-08

引　言

随着测量技术的飞速发展，现代分析仪器的多个分析通道可提供丰富的数据，从而获取海量及高维数据变得愈加容易。然而在数据的多元模型构建中，不是所有的变量都适合进入最终的模型。冗余及干扰变量的存在都会影响模型准确性；或者有时获取某些变量的成本过高，从而需要摒弃某些变量；当然对因变量影响显著的自变量若未进入模型，也会影响模型的准确性。而通过筛选变量能够提取出代表体系组成和特点的信息变量，从而达到数据降维、模型简化、提高预测效率乃至提高模型解释或预测性能的目的。所以，变量筛选已成为目前多元模型构建中的一个重要步骤。近年来，光谱领域中变量筛选方法的研究取得了很大的进展。图1显示出变量(波长)筛选相关出版论文数量从2004年—2019年呈逐年增长趋势(来源：SCI-EXPANDED with searchtopic “variable-selection” or “wavelength-selection”)。光谱数据大都存在量大、波段数多等特点。例如，使用傅里叶变换近红外(near-infrared，NIR)分析仪时，6 000 cm^{-1}光谱范围内可获得1 557个光谱点(变量)[1]。包括噪声、基线漂移、谱带重叠、背景干扰、杂散光等诸多因素的影响，会不可避免地导致数据中包含冗余和干扰变量。变量间存在的多重共线性也会影响建模，使得数据分析的结果变得不可靠。因此，变量筛选已经广泛地应用在光谱分析中。尤其近红外光谱(NIRs)吸收强度较弱，信噪比低，灵敏度低，谱峰宽且数目多、严重重叠。因此，借助化学计量学包括变量筛选技术从

收稿日期：2020-11-01，**修订日期**：2021-02-16

基金项目：中央高校基本科研业务费(2017MS135)，国家自然科学基金项目(21305043)资助

作者简介：李艳坤，女，1977年生，华北电力大学(保定)环境科学与工程系副教授

*通讯作者　　e-mail：liyankun@ncepu.edu.cn；liyankun_ncepu@foxmail.com

其光谱中提取出表征成分、结构等特征信息，克服了分析技术的难点，才使得该技术得以迅猛发展和应用。所以，目前变量筛选方面的综述[1-4]大都聚焦于其最得意的应用—近红外光谱领域。而本文结合作者的研究体会与文献调研，全面地综述了近/中红外光谱、拉曼光谱等光谱分析中常用的变量筛选算法的提出、发展、特点、分类及近年来的应用。

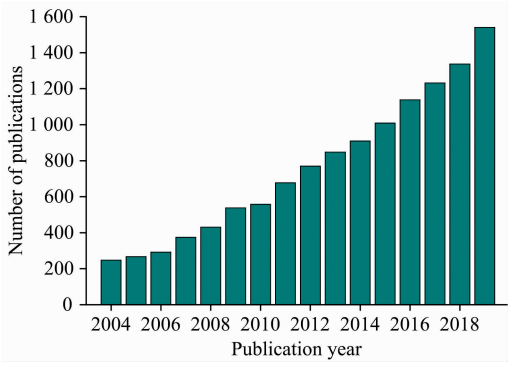

图 1 变量/波长筛选相关论文出版数量

Fig. 1 An overview of related works on variable/wavelength selection

1 基于 PLS 模型参数的变量筛选方法及其应用

偏最小二乘(partial least squares，PLS)是由 Wold 等于 20 世纪末提出的一种经典的多元校正算法[5]。在最初始的非线性迭代偏最小二乘基础上发展了 PLS-SVD 算法、简化的偏最小二乘(SIMPLS)、非线性的 Kernel PLS 等。PLS 求得模型的预报残差平方和较小，且适用于变量多、样本少的问题，得到广泛应用。同时基于 PLS 模型参数(回归系数、变量稳定性、变量投影重要性、光谱载荷权重)用于筛选变量的算法也在不断地应用在数据分析中。表 1 总结了代表性算法的提出及特点，并对其原理、发展及近五年来的应用进行综述。

表 1 基于 PLS 模型参数的变量筛选方法

Table 1 PLS parameter-based variables selection methods

Method	First appearance[Ref.]	Characteristic (Merit and Drawback)
UVE (uninformative variable elimination)	Massart, 1996[6]	Intuitive and practical, effectively eliminate the influence of non-objective factors; Random noise variables make the result unstable, and LOOCV makes calculation efficiency low.
MC-UVE (Monte Carlo-UVE)	Shao, 2008[7]	MC technique instead of LOOCV, does not add noise variables, high stability; Needs to define a threshold, tends to select more variables.
iPLS (interval PLS)	Norgaard, 2000[8]	Focus on a choice of better sub-intervals; Just testing a series of adjacent but nonoverlapping intervals, which would miss some more informative ones.
MWPLS (moving window PLS)	Jiang, 2002[9]	Considers all the possible continuous intervals but maybe not the optimized intervals.
CARS (competitive adaptive reweighted sampling)-PLS	Liang, 2009[10]	With fewer variables and latent variables; The reliability of PLS model parameters based on full spectra needs to be strengthened, low stability.
VIP (variable importance in projection)	Wold, 1993[11]	Accumulate the importance of each variable reflected by loading weight from each component; It can be used when the independent variables number is more than the sample size; Require probabilistic considerations regarding VIP.
RT (randomization test)-PLS	Fisher, 1935[12]	Combines permutation and statistical test, the result is more reliable; When the dataset is large, it has low efficiency and time consumption.
IVS (interactive variable selection)	Lindgren & Wold, 1994[13]	Dimension-wise instead of model-wise, variable selection is carried out for each PLS component, an interactive variable selection approach; Large elements in sometimes suppress smaller values.
IPW (iterative predictor weighting)[15]	Forina, 1999[14]	The importance measure is used both to re-scale the original X-variables and to eliminate the least important variables; Time-consuming for too many variables.

1.1 无信息变量消除(UVE)-PLS

UVE 通过留一交叉验证建立一系列 PLS 模型，计算每个变量的稳定性“stability”(回归系数平均值与其标准偏差的比值：“c_j”)。通过在数据中添加数值较小的随机变量(噪声)的“c_{artif}”作为阈值来删除无信息变量。邵学广课题组基于 UVE 融合蒙特卡洛(Monte Carlo，MC)思想，用 MC 技术代替 LOOCV，提出蒙特卡洛无信息变量消除(MC-UVE)[7]。该方法对烟草样品的 NIRs 波长进行筛选，与全谱 PLS 及 UVE-PLS 相比，在保留变量数目最少的情况下取得最好的预测精度。并进一步和小波变换(WT)结合，得到更加精简的定量模型。此后，MC-UVE 开始成功地应用于各种光谱数据的分析中[16-18]。

1.2 区间偏最小二乘(iPLS)和移动窗口偏最小二乘(MW-PLS)

区间(间隔)PLS 将光谱均分成若干个连续等宽子区间，在每个子区间内分别建立 PLS 模型，将交叉验证均方根误差

(root mean square error of cross validation，RMSECV)最小的子区间确定为最佳模型波段。由于子区间的位置随着全谱划分区间数目的确定而固定，然而这些区间不一定恰与成分相关的信息区间吻合。为此，产生了采用一个窗口沿整个波谱逐步移动的策略，即 MWPLS。

牛晓颖等[19]利用 iPLS 筛选出了猪、牛、羊肉等鲜肉中多种不饱和脂肪酸适宜建模的近红外光谱波段。Zhao 等[20]运用 iPLS 分析猪肉皮下脂肪的拉曼光谱，实现了其含碘值的检测。Yu 等[21]采用激光诱导击穿光谱技术结合 iPLS，实现了含石油土壤中金属的定量分析。为了优化子区间的组合，出现一些 iPLS 改进方法：向前/向后区间偏最小二乘(BiPLS/FiPLS)、区间协同 iPLS(SiPLS)和确定独立性筛选 iPLS(SIS-iPLS)等[22-23]。许良等[24]采用近红外漫反射光谱结合 MWPLS 筛选克霉唑的特征波长区域，得到测定克霉唑粉末药品的最佳模型。谢军等[25]将 MWPLS 用于人血清葡萄糖的衰减全反射红外光谱分析中。Wang 等[26]发展了深度协同-自适应移动窗口偏最小二乘-遗传算法用于煤样 NIRs 分析，得到水分、灰分、挥发物的最佳校准模型。还出现了窗口尺寸可变的 CSMWPLS (changeable size MWPLS)、移动窗口组合搜索的 SCMWPLS (searching combination MW-PLS)、对称收缩循环固定窗口 PLS (SCRUWPLS)[27]等。

1.3　竞争性自适应重加权采样(CARS)-PLS

CARS 通过蒙特卡洛采样，利用指数衰减函数(EDP)和自适应重加权采样(ARS)策略选出 PLS 回归系数绝对值大的波长点，去除权重小的波长点，最终选出 RMSECV 最低值对应的变量子集。蒋雪松等用 CARS-PLS 建立了植物油反式脂肪酸的拉曼光谱定量模型，筛选出特征变量。Nie[28]等用 CARS 分析愈风宁心滴丸原料中葛根素，提高了检测精度。石岩等[29]用 CARS 研究人工牛黄的 NIRs，用于建模的变量数大幅减少。Hu 等[30]测定了葡萄酒中总酸、总糖和酒精含量，结果明显优于全谱 PLS。融合 MC-UVE 和 CARS 优势产生了基于变量稳定性的 SCARS(stability CARS)，用于咖啡因、尼古丁、玉米中水分的检测[31-32]。Zheng 等[33]提出双竞争自适应重加权采样(double CARS)。

1.4　变量投影重要性(VIP)分析

VIP 分析中变量通过主成分传递对目标值的解释能力。若主成分对目标值的解释作用很强，而变量对主成分的作用又很大，则该变量会具有较大的 VIP 值，即被认为是贡献大的变量而被保留。Ferreira 等[34]用 VIP 对巴西大豆的 NIRs 进行筛选，剔除了冗余变量，对大豆膳食纤维进行了准确分析。Gosselin 等[35]提出 PLS-bootstrap-VIP 用于可见光-NIRs 分析聚合物薄膜、木材/塑料复合材料、柴油参数，验证了筛选特征波长的有效性。

1.5　随机检验(RT)-PLS

邵学广课题组将随机检验(RT)思想引入 PLS 模型用于变量筛选。将数据对应的自变量值打乱，将随机化的自变量值与光谱响应值之间建立多个 PLS 模型。统计随机模型的回归系数值超过正常(真实)模型回归系数的比例(P)，具有较小 P 值的变量即为重要变量。RT-PLS 用于谷物和烟草 NIRs 波长筛选中，RMSEP 平均值及标准偏差都小于全谱 PLS[36]。同样构建了烤烟中三种多酚的近红外漫反射光谱模型，与采用高效液相色谱法测得的参考值一致[37]。

2　其他变量筛选方法

其他常见的光谱变量筛选算法按搜索及筛选变量的策略可以分为五类：(1)智能寻优算法：利用进化和群集智能算法，搜寻使目标函数较优的变量子集。其中遗传算法(GA)应用较广泛，在其基础上引入生物免疫系统原理，发展了免疫遗传算法(IGA)。结合 GA、SA(模拟退火)算法优势，提出 GSA 算法[38]；(2)基于模型集群分析算法[39]：采用随机采样(Monte Carlo，Bootstrap，Binary matrix)产生的变量子集建立系列子模型，挑选出 RMSECV 较低的子模型，采用统计检验评价变量的重要性，在下次迭代中赋予较高的取样权重。梁逸曾课题组基于模型集群策略提出一系列的算法；(3)基于变量空间共线性最小化算法：降低被选中变量建模的严重多重共线性，保留最小冗余信息的变量子集。例如，序列前进筛选法中的连续投影算法通过构建变量的正交矩阵来选择变量，降低了多重共线性变量对模型的影响；(4)基于分类模型的变量筛选：利用分类模型的内部参数作为评价，筛选出对分类模型有重要意义变量的同时，计算得到目标样本的得分用于分类判别。其中 LDA[40]是降维和提取特征信息的有效方法之一。不相关变量投影分析(ULDA)考虑了基于 LDA 的变换矩阵列向量间的不相关性，减少降维后的数据冗余，在寻找疾病生物标志物中得到好的应用[41]。李艳坤等[42]用 ULDA 解析人体血清多肽质谱，从15 154个变量中挑选出 7 个特征变量较好地区分了良性和恶性肿瘤。不仅简化模型提高了诊断效率，而且 7 个特征变量所对应的多肽可作为潜在卵巢癌标志物；(5)正则化回归算法：在原有的损失函数的基础上增加惩罚回归系数的正则项，收缩回归系数，减少所有特征变量回归系数估计值的数量级，自动将无关变量的回归系数置接近于 0。这几类算法包含的代表性算法总结于表 2。

此外，还出现了基于其他原理的算法。通过将光谱投影到局部线性嵌入(locally Linear Embedding)[69]空间后，依次移除变量后引起样本位置的变化而提出一种用于变量筛选的方法。还提出潜变量投影图(latent projective graph)[70]等算法。

3　变量筛选方法的类别与比较

3.1　波长筛选和波段筛选

根据筛选出光谱变量的分布特征，分为波长筛选和波段(波长区间)筛选。波长选择(wavelength selection，WS)以波长点为单位(即一个变量)，因此所选择的变量是离散的。波段筛选(wavelength interval selection，WIS)通常考虑相邻变量的连续和协同作用(正协同和副协同)，可能增加选择变量的复杂性。而对光谱分段处理本质上降低了变量选择的难度，以划分的波长区间(若干连续变量组成)为单位寻找最优区间(组合)。但波段的划分很关键，图2展示了划分波段的

表 2 其他光谱变量筛选方法

Table 2 Other common methods of spectral variables selection

Selection strategy	Representative methods[Ref.]	First appearance[Ref.]	Characteristic(Merit and Drawback)
Intelligent optimizing algorithms (IOA)-based	GA(Genetic algorithm)	Holand, 1975[43]	Return to the mathematical essence of variable combination optimization, retain advantages of the combination of variables; Too many combinations of variables to optimize, usually need more preset parameters, sometimes easy to fall into local optimum.
	SA(Simulated Annealing)	Metropolis, 1953[44]	
	PSO(Particle swarm optimization)	Eberhart&Kennedy, 1995[45]	
	ACO(Ant colony optimization)	Colorni, 1991[46]	
	GWO(Gray wolf optimizer)	Mirjalili, 2014[47]	
Model population analysis (MPA)-based	BOSS (Bootstrapping soft shrinkage)	Liang, 2016[48]	The traditional strategy of rigidly eliminating variables according to a single index is transformed into a flexible strategy of changing weight, which can preserve the effective variables more safely; The introduction of random algorithm helps to preserve the combination effect among spectral variables, however, it also makes the calculation more complicated.
	VCPA (Variable combination population analysis)	Liang, 2015[49]	
	VISSA (Variable iterative space shrinkage approach)	Liang, 2014[50]	
	ICO (Interval combination optimization)	Xiong & Min, 2016[51]	
	iRF (internal Random frog)	Liang, 2013[52]	
Collinearity minimization-based	SPA (Successive projection algorithm)[53, 54]	Araujo, 2001[55]	Minimizing the influence of multi-collinearity variables on the model; In the optimization, each variable is used as the starting point, the calculation amount is too large to be suitable for small-size sample.
	SR (Stepwise regression)[56]		
Category model-based	LDA (Linear discriminant analysis)	Fisher, 1936[57]	The correlation between variables and model is preserved, and the overall prediction accuracy is improved by combining different classification algorithms. The computational complexity is small, but the result is limited by the performance of the classification model.
	ULDA (Uncorrelated lineardiscriminant analysis)[58]	Jin, 2001[59]	
	RF (Random forest)[60-62]	Breiman, 2001[63]	
	SVM (Support vector machine)	Vapnik, 1995[64]	
Regularization method	LASSO (Least absolute shrinkage and selection operator)[65]	Tibshirani, 1996[66]	Parameter estimation and variable selection are realized simultaneously, fast. When the number of variables is large, the over-fitting can be avoided; The suitable parameter value should be chosen.
	EN (Elastic net)	Zou, 2003[67]	
	RR (Ridge regression)	Hoerl & Kennard, 1998[68]	

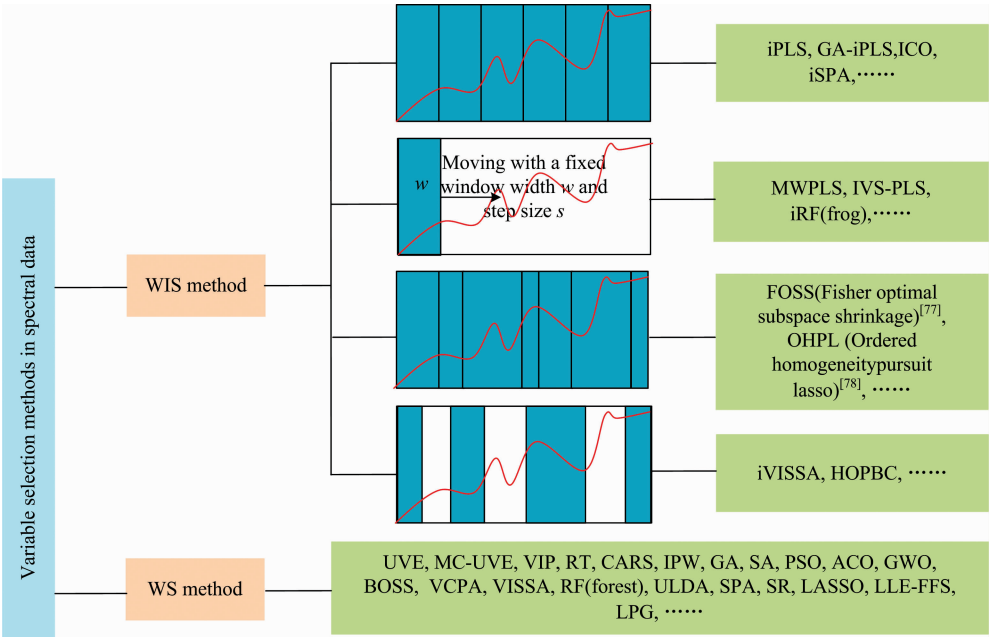

图 2 波段和波长筛选方法

Fig. 2 The methods of WS and WIS

四种方式[1]。其中，张进等提出的启发式最优波段组合(heuristic optimal partner band combination)[71]通过 SPA 选择冗余信息最小的变量，以此为中心向两侧扩展一定宽度，然后采用排列组合策略选出具有协同效应的波段组合，提高了基

于变量直接排列组合的选择方法的效率[72]。

两类方法筛选出变量的分布虽具有相似性，但对模型预测能力的影响有一定的差别。一组谷物样本的近红外光谱（http://software.eigenvector.com/Data/Corn/corn.mat）和相应的蛋白质含量模型用于考察。首先利用小波变换结合多元散射校正对光谱进行预处理，然后采用 UVE，iPLS 和 MWPLS 等筛选出变量，如图 3 所示[73]。将筛选后的变量用于 PLS 建模，结果中 MWPLS 和 iPLS 的 RMSEP 值较低，也就是 WIS 的结果要优于 WS。随后考察了光谱中变量间的相关系数，发现强相关变量分布比较连续。同时考察了一组强相关变量分布比较分散的烟草数据，筛选变量后用 PLS 建模预测尼古丁含量，发现 WS 比 WIS 方法的预测结果有优势。而两组数据无论使用波长筛选还是波段筛选，都优于全谱模型的预测结果。因此，筛选变量在光谱分析中非常有必要。而选择合适的波长或波段筛选方法，需在一定程度上考虑强相关变量的分布情况。

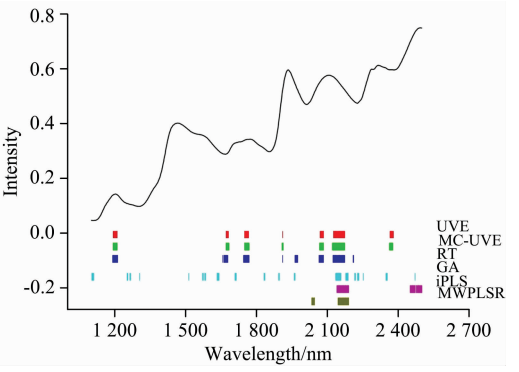

图 3　谷物 NIR-蛋白质模型变量筛选方法比较

Fig. 3　Comparisons of variable selection methods in NIR-protein model for corn data

3.2　过滤式、封装式、嵌入式方法

从变量子集选择标准与学习算法的关系角度，变量筛选方法又可以分为：独立于学习算法的过滤式、依赖于学习算法的封装式和与学习算法集成的嵌入式方法。过滤式(Filter)通过引入阈值来实现对变量的选择与否，方法与后续学习器无关。计算简便，但筛选结果受阈值影响较大；封装/缠绕式(Wrapper)使用迭代的过滤方法，将学习器的性能作为变量筛选的评价标准，直到选出最优变量组合；嵌入式(Embedded)利用模型的内部参数作为评价，保留变量和模型间的相互关系。变量选择方法自身即算法组成的一部分，嵌入到算法中。三类算法的流程及包含的常用算法如图 4 所示。

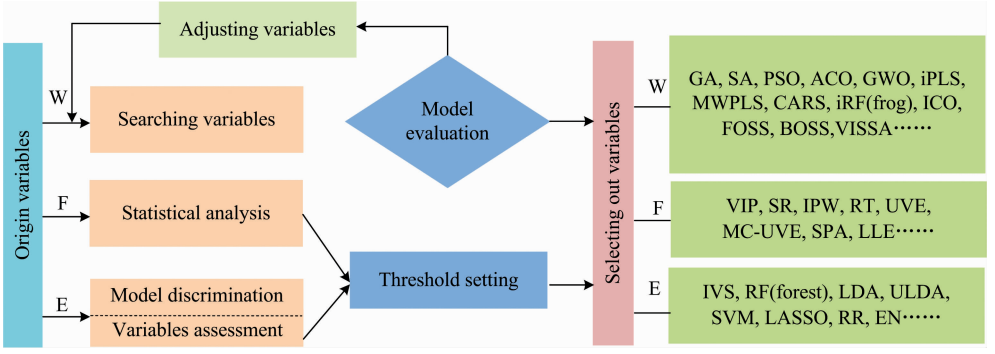

图 4　过滤式、封装式和嵌入式方法

Fig. 4　Illustration for filter (F), wrapper (W) and embedded (E) methods

4　结　论

通过作者多年来在该领域的研究体会结合文献分析表明，每种变量筛选方法都具有各自的优势和局限性。实际中需要根据分析目标和算法本身两方面的特点选择合适的方法。下面围绕在解析实际体系中存在的若干问题进行讨论。

(1) 有时单一的变量筛选方法往往达不到分析的要求，此时需要方法的联合使用，例如 CARS-SPA 和 CARS-GA 等等。联合方法不等同于几种方法的简单耦合，而是协同发挥优势。通常前一种粗选以消除无信息变量，后一种精选以挑选典型特征信息变量或降低变量间的多重共线性。最终选择的效果取决于不同算法的逻辑结合方式和综合利用度。

(2) 某些筛选方法中由于采取随机抽样的变量子集和迭代方式进行优化，会导致模型筛选出的变量不稳定，进而产生不稳定的结果，降低模型的可信度。例如 CARS 尽管具有筛选出变量数目少的优点，但每次重复运行选中的变量个数及位置都会发生变化。对此，李艳坤等采取多模型共识策略[16, 18]综合多个成员模型的预测结果，或保留被选择频率较高的变量，得到更准确、稳健的结果。尤其应用在近红外光谱中，虽然其对应于结构信息的特征较差，但选定的波长与目标物的功能基团之间仍然得到了合理的解释。

(3) 在寻找重要变量过程中，存在过拟合风险。由于数据包含大量的变量，总会有一些不相关变量由于偶然性而变得很重要[76]，某些奇异样本的加入也会影响模型的构建。此外，采用 RMSECV 作为评价指标，或基于 PLS 回归系数的筛选方法[75]较多地利用自变量信息，这些都可能带来模型过拟合的风险。所以，尽可能地在建模前对数据进行奇异值的识别和剔除、采用独立的外部样本评估变量或大规模的数据集验证。而新的评价变量重要性的参数及其判据，以及搜索变量的策略和途径等工作仍需展开深入地研究。

(4) 与尺度缩放、基线校正等光谱预处理方法(平滑、导数等)联合。李艳坤等曾利用小波变换系数替代原始光谱输入 MC-UVE 模型中，在保留更少变量时取得相当或优于原

MC-UVE 筛选模型的预测结果[7, 16]。此外，方法和数据属性之间可能存在相互作用，因此“不存在总是最好的方法，而存在最合适的方法”。所以，面对如此多的方法及方法的组合，可以发展集成(汇集预处理和变量筛选方法)智能化选择(根据数据特征或建模性能优劣，后者更简单直接)算法，并开发适用于测量仪器或独立使用的计算软件，使用起来会更加快捷，尤其对于非专业人士处理数据将会非常有用。

致谢：感谢邵学广教授(南开大学)对本稿件提供的宝贵建议与指导。

References

[1] Yun Y H, Li H D, Deng B C, et al. Trac-Trend Anal. Chem., 2019, 113: 102.

[2] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.

[3] Nie M P, Meng L W, Chen X J, et al. J. Chemometr., 2019, 33(4): e3113.

[4] Mehmood T, Ahmed B J. Chemometrics, 2016, 30(1): 4.

[5] Wold S, Albano C, Dunll M. Pattern Regression Finding and Using Regularities in Multi-variate Data. London: Analysis Appfied Science Publication, 1983.

[6] Centner V, Massart D L. Denoord O E, et al. Anal. Chem., 1996, 68: 3851.

[7] Cai W S, Li Y K, Shao X G. Chemom. Intell. Lab. Syst., 2008, 90(2): 188.

[8] Norgaard L, Saudland A, Wagner J, et al. Appl. Spectrosc., 2000, 54(3): 413.

[9] Jiang J H, Berry R J, Siesler H W, et al. Anal. Chem., 2002, 74(14): 3555.

[10] Li H D, Liang Y Z, Xu Q S, et al. Anal. Chim. Acta, 2009, 648(1): 77.

[11] Wold S, Johansson E, Cocchi M. 3D-QSAR in Drug Design, Theory, Methods, and Applications. Leiden: ESCOM Science Publishers, 1993.

[12] Fisher R A. The Design of Experiments. Edinburgh: Oliver and Boyd. 1935.

[13] Lindgren F, Geladi P, Rännar S, et al. J. Chemometr., 1994, 8(5): 349.

[14] Forina M, Casolino C, Millan C P. J. Chemometr., 1999, 13(2): 165.

[15] Chen D, Hu B, Shao X, et al. Analyst, 2004, 129(7): 664.

[16] Li Y K, Jing J. Chemom. Intell. Lab. Syst., 2014, 130(130): 45.

[17] Li C, Zhao T L, Li C, et al. Food Chem., 2017, 221(4): 990.

[18] Li Y K. Anal. Methods, 2012, 4(1): 254.

[19] NIU Xiao-ying, SHAO Li-min, ZHAO Zhi-lei, et al(牛晓颖, 邵利敏, 赵志磊, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(2): 443.

[20] ZHAO Fang, PENG Yan-kun(赵 芳, 彭彦昆). Chinese Journal of lasers(中国激光), 2017, 44(11): 243.

[21] Ding Y, Xia G Y, Ji H W, et al. Anal. Methods, 2019, 11(29): 3657.

[22] Miao X X, Miao Y, Gong H R, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 257: 119700.

[23] Pereira Rainha K, Tristão do Carmo Rocha J, Tavares Rodrigues R R, et al. Anal. Lett., 2019, 52(18): 2914.

[24] XU Liang, YAN Liang-liang, SAI Jilahu, et al(许 良, 闫亮亮, 塞击拉呼, 等). Computers and Applied Chemistry(计算机与应用化学), 2016, 33(4): 415.

[25] XIE Jun, MA Hui, PAN Tao(谢 军, 马 辉, 潘 涛). Chinese Journal of Analysis Laboratory(分析试验室), 2015, 34(3): 255.

[26] Wang S H, Zhao Y, Hu R, et al. Chinese J. Anal. Chem., 2019, 47(4): e19034.

[27] Cramer J A, Kramer K E, Johnson K J, et al. Chemom. Intell. Lab. Syst., 2008, 92(1): 13.

[28] Nie L X, Dai Z, Ma S C. Analytical Letters, 2016, 49(14): 2259..

[29] SHI Yan, SUN Dong-mei, XIONG Jing, et al(石 岩, 孙冬梅, 熊 婧, 等). Chinese Pharmaceutical Journal(中国药学杂志), 2018, 53(14): 1216.

[30] Hu L Q, Yin C L, Ma S, et al. Spectrochim. A, 2018, 205: 207.

[31] Zhang X, Li W, Yin B, et al. Spectrochim. Acta A, 2013, 114: 350.

[32] Li W, Zhang X, Zheng K Y, et al. J. AOAC Int., 2015, 98(1): 183.

[33] Zheng K Y, Feng T, Zhang W, et al. Chemom. Intell. Lab. Syst., 2019, 191: 109.

[34] Ferreira D S, Poppi R J, Lima Pallone J A. J. Cereal Sci., 2015, 64: 43.

[35] Gosselin R, Rodrigue D, Duchesne C. Chemom. Intell. Lab. Syst., 2010, 100(1): 12.

[36] Xu H, Liu Z C, Cai W S, et al. Chemom. Intell. Lab. Syst., 2009, 97(2): 189.

[37] Mao Z Y, Shan R F, Wang J J, et al. Spectrochim. Acta A, 2014, 128: 711.

[38] XIE Huan, CHEN Zheng-guang(谢 欢, 陈争光). Analytical Chemistry(分析化学), 2019, 47(12): 1987.

[39] YUN Yong-huan, DENG Bai-chuan, LIANG Yi-zeng(云永欢, 邓百川, 梁逸曾). Chinese Journal of Analytical Chemistry(分析化学), 2015, 43(11): 1638.

[40] Ma X P, Pang J F, Dong R N, et al. J. Food Compos. Anal., 2020, 91: 103509.
[41] Li Y K, Ma X P, Huang K N, et al. Indian J. Biochem. Bio., 2019, 56(1): 53.
[42] Li Y K, Zeng X C. Anal. Methods, 2016, 8: 183.
[43] Holland J H. Adaptation in Natural and Artificial Systems. Ann Arbor, Mich: University of Michigan Press, 1992.
[44] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. J. Chem. Phys., 1953, 21(6): 1087.
[45] Kennedy J, Eberhart R. Particle Swarm Optimization, IEEE International Conference on Neural Networks, Perth, 1995, 4: 1942.
[46] Colorni A, Dorigo M, Maniezzo V, et al. Distributed Optimization by Ant Colonies, Proceedings of the First European Conference on Artificial Life. Paris, 1991: 134.
[47] Mirjalili S, Mirjalili S M, Lewis A. Adv. Eng. Software, 2014, 69: 46.
[48] Deng B C, Yun Y H, Cao D S, et al. Anal. Chim. Acta, 2016, 908: 63.
[49] Yun Y H, Wang W T, Deng B C, et al. Anal. Chim. Acta, 2015, 862: 14.
[50] Deng B C, Yun Y H, Liang Y Z, et al. Analyst, 2014, 139(19): 4836.
[51] Song X Z, Huang Y, Yan H, et al. Anal. Chim. Acta, 2016, 948: 19.
[52] Yun Y H, Li H D, Wood L R E, et al. Spectrochim. Acta A, 2013, 111: 31.
[53] Moreira E D T, Pontes M J C, Galvão R K H, et al. Talanta, 2009, 79(5): 1260.
[54] Gomes A D, Galvao R K H, de Araújo M C U, et al. Microchem J. 2013, 110: 202.
[55] Araujo M C U, Saldanha T C B, Galvao R K H, et al. Chemom. Intell. Lab. Syst., 2001, 57(2): 65.
[56] Yu Q, Li J, Yao L, et al. J. Appl. Remote Sens., 2018, 12(3): 036019.
[57] Fisher R A. Annals of Eugenics, 1936, 7: 179.
[58] PANG Jia-feng, TANG Chen, LI Yan-kun, et al(庞佳烽, 汤 谌, 李艳坤, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(10): 3235.
[59] Jin Z, Yang J Y, Hu Z S, et al. Pattern Recognit., 2001, 34(7): 1405.
[60] Zhang S P, Tan Z L, Liu J, et al. Spectrochim. Acta A, 2019, 227: 117551.
[61] WU Li-zhou, WANG Xiao-hui, WANG Zhi-hui, et al(吴立周, 王晓慧, 王志辉, 等). Journal of Zhejiang A&F University(浙江农林大学学报), 2020, 37(1): 136.
[62] LI Guan-wen, GAO Xiao-hong, XIAO Neng-wen, et al(李冠稳, 高小红, 肖能文, 等). Chinese Journal of Luminescence(发光学报), 2019, 40(8): 1030.
[63] Breiman L. Mach. Learn., 2001, 45(1): 5.
[64] Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, MD: ACM Press, 1992: 144.
[65] Zhang R Q, Zhang F Y, Chen W C, et al. Chemom. Intell. Lab. Syst., 2019, 184: 132.
[66] Tibshirani R. J. R. Stat. Soc. B, 1996, 58(01): 267.
[67] Zou H, Hastie T. Regression Shrinkage and Selection via the Elastic Net, With Application to Microarrays, 2003: 1.
[68] Hoerl A E, Kennard R W. Technometrics, 1970, 12(1): 55.
[69] Shan R F, Cai W S, Shao X G. Chemom. Intell. Lab. Syst., 2014, 131: 31.
[70] Shao X G, Du G R, Jing M, et al. Chemom. Intell. Lab. Syst., 2012, 114: 44.
[71] Zhang J, Cui X Y, Cai W S, et al. J. Chemom., 2018, 32(11): e2971.
[72] Zhang J, Cui X Y, Cai W S, et al. Sci. China Chem., 2019, 62(02): 271.
[73] Xu H, Cai W S, Shao X G. Anal. Methods, 2010, 2: 289.
[74] Lin Y W, Deng B C, Wang L L, et al. Chemom. Intell. Lab. Syst., 2016, 159: 196.
[75] Lin Y W, X N, Wang L L, et al. Chemom. Intell. Lab. Syst., 2017, 168: 62.
[76] Mehmood T, Liland K H, Snipen L, et al. Chemom. Intell. Lab. Syst., 2012, 118: 62.

Variable Selection Methods in Spectral Data Analysis

LI Yan-kun[1*], DONG Ru-nan[1], ZHANG Jin[2], HUANG Ke-nan[3], MAO Zhi-yi[4]

1. Department of Environmental Science and Engineering, North China Electric Power University, Hebei Key Lab of Power Plant Flue Gas Multi-Pollutants Control, Baoding 071003, China
2. School of Food Science, Guizhou Medical University, Guiyang 550025, China
3. The 82nd Army Group Hospital of the Chinese People's Liberation Army, Baoding 071000, China
4. Tianjin Building Material Science Research Academy, Tianjin 300110, China

Abstract How to extract useful information from massive or high-dimensional data is a huge challenge for current data analysis and a hot spot of current research. Variable selection technology can extract feature information variables from numerous and complex measurement data, and achieve the purpose of simplifying multivariate model and even improving the model's prediction performance. In spectral analysis, the measurement data will inevitably contain interference and irrelevant information variables and the multicollin earity among variables, which will affect the robustness and prediction ability of the model. Therefore, the variable(wavelength) selection methods have progressed greatly in the research and application of spectral analysis. Based on the related pieces of literature and the author's research experiences, this paper summarizes the proposals, characteristics, developments, categories, comparisons and applications in recent five yearsof methods for selecting variables not only in near-infrared spectra area but also in fields of mid-infrared spectra, Raman spectra and other spectra. The parameters as their criteria or thresholds for evaluating the importance of variables and the strategies or tracks of selecting variables are vital. Moreover, each method has its advantages and limitations. In practice, it is necessary to select the appropriate method according to the characteristics of boththe method and the object. Key contents: (1) Compared the wavelength selection, and wavelength interval selection methods; (2) Summarized the different variable selection methods based on PLS model parameters; (3) Classified and overviewed the variable selection methods according to the strategiesof searching and selection of variables. Finally, we discuss the problems of variable selection methods (such as overfitting and instability etc.) appearing in the actual system and the corresponding solutions. Meantime, there look forward to the research trend, development prospect and application direction of the variable selection methods. Among them, new criteria for evaluating the importance and new selection strategy of variables still require further research. It is expected that this paper will play a positive role in promoting the follow-up researches and applications of variable selection technology.

Keywords Variable selection; Spectral data; Characteristic variable; Redundant information

(Received Nov. 1, 2020; accepted Feb. 16, 2021)

* Corresponding author