第37卷，第4期
2017年4月

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

Vol. 37, No. 4, pp1048-1052
April, 2017

近红外光谱分析中的变量选择算法研究进展

宋相中，唐 果，张录达，熊艳梅，闵顺耕*

中国农业大学理学院，北京 100193

摘 要 随着人们对近红外光谱分析技术了解的深入，人们发现通过剔除近红外光谱中的冗余变量不仅可以简化近红外光谱分析模型，提高模型的可解读性，通常还可以提高模型的预测效果及稳健性。变量选择的有效性已经在各种近红外光谱应用体系中得到了广泛的验证，发展成为了近红外光谱分析建模过程中一个越来越重要的步骤。为此，化学计量学家们近些年来开发了大量原理不同的新型变量选择算法，基于各种原理的衍生算法也层出不穷。为了让近红外光谱分析研究人员能够较为迅速地对这些算法的特点有所认识，对目前常见的各种变量选择算法的算法原理和优缺点进行了梳理。根据各种算法依据的原理不同，将目前近红外光谱领域常见的变量选择算法大致分为基于偏最小二乘模型参数，基于智能优化算法，基于连续投影策略，基于模型集群分析策略和基于变量区间等五类。在梳理的过程中，我们发现变量选择算法的发展趋势目前主要集中在以下两点：第一，算法的复杂程度不断提高；第二，不同变量选择算法之间的联用开始逐渐增多。此外，作者结合自身在应用变量选择算法时的体会和思考，还总结了变量选择算法在应用层面上存在的一些问题。例如光谱预处理方法对变量选择算法使用效果的影响，以及部分算法存在的稳定性较差，选择变量的可靠性存疑等。

关键词 近红外光谱；变量选择算法；综述

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2017)04-1048-05

引 言

传统的近红外光谱分析模型，一般均采用全谱参与建模[1]。由于近红外光谱分析技术一般作为快速无损分析工具使用，全谱中不可避免地会含有大量噪声、无信息甚至是干扰的变量。这些变量的存在不仅增加了多元校正模型的复杂程度，有时还会严重影响模型的预测性能[2-3]。近30年来化学计量学家们发明了大量的变量选择算法，用以从近红外光谱中提取有效变量。大量文献资料表明采用合适的变量选择算法从全谱中选择出有效变量后，建立的近红外分析模型不仅更加简约，模型的稳健性和预测的精密度也通常比全谱模型有所提高，同时模型的可解释性也大为提高[4]。因此，变量选择已经发展成为了近红外光谱分析中的一个重要步骤，是近年来近红外光谱分析技术和化学计量学的研究热点[5-7]。

此前虽有文献[5-9]对已有变量选择算法进行过综述，但是这些文章大多针对某一类别的算法进行综述。变量选择算法的发展一直非常迅猛，尤其是近些年来，涌现出了一大批新型算法，在近红外光谱分析中的应用也越来越多。然而，最近一篇关于近红外光谱分析中常用的变量选择算法综述已经是在多年之前发表的[6]。因此本文将从算法原理特点出发，重点介绍近年来在近红外光谱分析中常用的变量选择算法。此外，作者还结合自身使用相关算法及阅读相关文献时的体会和思考，归纳总结了目前变量选择算法的一些发展趋势，并提出了变量选择算法在近红外光谱分析实际应用中可能会遇到的一些问题。

1 主要类别及原理特点介绍

由于近红外光谱分析领域常用的变量选择算法的种类较多，难以一一予以具体介绍。根据各种变量选择算法原理的异同，将目前比较常用的变量选择算法大体分为以下几类。

第一类，主要根据全谱偏最小二乘(partial least squares，PLS)模型的一些参数作为变量选择的判据，这些常用于变量选择的参数详见表1。在有些应用中人们直接将处于参数曲线极值点位置的变量视为有效变量，例如Liu fei等[10]详

收稿日期：2015-11-29，修订日期：2016-04-06

基金项目：国家自然科学基金-青年基金项目(31301685)资助

作者简介：宋相中，1990年生，中国农业大学博士研究生 e-mail：sxzfhn@163.com

*通讯联系人 e-mail：minsg@cau.edu.cn

细地比较了全谱 PLS 模型的一些参数在选择有效变量方面的应用效果。结果表明采用这些参数作为依据选择的变量均分布在有信息的变量区域内，其中回归系数是最为常用的参数，因为其计算简单而且选择变量的化学意义易于解释。另外一些算法则通过设置阈值来选择有效变量，例如无消息变量剔除(uninformative variables elimination，UVE)[11]将变量稳定性大于人工噪声变量相应参数最大值的变量视为有信息变量。然而，更多的基于全谱 PLS 模型参数的变量选择算法则依据这些参数中的一种或几种的组合进行变量排序后，按照排序逐步剔除变量建立一系列子模型，然后通过比较子模型的预测效果间接选取变量，一般认为模型交叉检验预测误差最小的子模型对应的变量子集即为最优的变量子集。这类算法目前常见的有预测参数排序变量剔除(predictive property ranked variable reduction，PPRVR)[12]，蒙特卡洛无消息变量剔除(Monte Carlo uninformative variables elimination，MCUVE)[13]，竞争适应再加权抽样方法(competitive adaptive reweighted sampling，CARS)[14-15]，变量子集迭代优化(iteratively variable subset optimization，IVSO)[16]等。总体来说，此类算法的最大特点是选择有效变量的速度较快，选中变量的化学意义通常也易于解释。然而由于全谱 PLS 模型是在噪声，无信息甚至是干扰变量存在的情况下建立的，其模型参数难免会受到这部分变量的干扰，因此这类算法选中变量的可靠性仍有待加强[17]。

表 1　常用于变量选择的 PLS 模型参数

Table 1　Parameters of the PLS model common used for variable selection

参数名称	英文名称
回归系数	Regression Coefficients(RC)
光谱载荷权重	Loading Weights (LW)
变量稳定性	Stability
变量投影重要性	Variable Importance in Projection (VIP)
选择比	Selectivity Ratio

第二类，通常采用智能优化算法进行变量组合优化，从而筛选出有效变量组合。目前常用于变量选择的智能优化算法有遗传算法(genetic algorithm，GA)[18-20]、模拟退火算法(simulated annealing，SA)[21]、粒子群优化算法(particle swarm optimization，PSO)[22]、蚁群优化算法(ant colony optimization，ACO)[23-24]等。此类算法的最大特点是将变量选择问题回归了变量组合优化的数学本质，可以较好地保留变量间的组合优势。由于要优化的变量组合数量太多，人们往往在此类算法之前对采集到的近红外光谱进行窗口化处理，即将固定个数的相邻光谱变量平均后作为一个输入变量或者采用变量区间代替单个变量参与组合优化[24-26]。此外，智能优化算法的使用往往需要预设较多的参数，而且优化效果容易受到适应度函数的影响。例如，遗传算法运行前需要设置的参数就有染色体数目，交叉概率，变异概率，迭代次数等。

第三类，一般采用连续投影策略进行变量排序产成一系列变量子集，然后通过比较变量子集模型的预测能力筛选出最优的变量子集，目前主要以连续投影算法(successive projections algorithm，SPA)[9, 27]为代表。此类算法的主要特点是连续投影策略可以最大程度地降低被选中变量间的共线性问题。然而，实际上对于近红外光谱而言，有效变量之间的投影距离并不一定最大，SPA 筛选出的变量子集中可能包含有一些无信息甚至是干扰变量。此外，由于优化过程中每个变量都要作为起点，进行连续投影排序，得到一组变量子集，往往导致该算法的计算量偏大。SPA 常被用于降低 UVE，MCUVE，CARS 等快速算法粗选的有效变量之间的共线性[28-30]，在一定程度上实现了算法间的优势互补。

第四类，是基于中南大学梁逸曾教授课题组提出的模型集群分析策略开发出的一系列变量选择方法[8]，目前以迭代保留有效变量(iteratively retains informative variables，IRIV)[31]，变量空间迭代收缩(variable iterative space shrinkage approach，VISSA)[32-33]，变量组合集群分析(variable combination population analysis，VCPA)[34]，自举柔性收缩算法(bootstrapping soft shrinkage，BOSS)[35]等算法为代表。此类算法一般采用随机算法产生的变量子集建立一系列子模型，然后从中挑选出预测效果较好的一部分子模型，统计这一部分子模型对应的变量情况。如果一个变量出现在优秀子模型中的比率较高，则认为这个变量较为重要，并在下次迭代中赋予较高的保留权重。此类算法从模型集群分析策略出发，将传统的根据单一指标硬性剔除变量的策略转化为柔性的改变权重策略，一般可以更加稳妥地保留有效变量。随机算法的引入则帮助此类算法可以更好地保留光谱变量间的组合效应。然而模型集群分析策略也使得此类算法的计算量较大。

第五类，前面四类的变量选择算法选择的对象一般是单一光谱变量，而此类算法的选择对象变成了光谱区间，因此被称之为变量区间选择算法。此类算法又可以分为区间偏最小二乘(interval PLS，iPLS)[36]和移动窗口偏最小二乘(moving windows PLS，MWPLS)[37]两类。其中 iPLS 及其衍生算法向后区间偏最小二乘(backward interval partial least-squares，BiPLS)，向前区间偏最小二乘(forward interval partial least-squares，FiPLS)，区间组合偏最小二乘(synergy interval partial least-squares，SiPLS)，区间随机蛙跳(interval random frog，iRF)[38]，区间组合优化算法(interval combination optimization，ICO)[39]，ACO-iPLS[24]，GA-iPLS[26]，SPA-iPLS[40]等，首先将光谱均分为若干子区间，然后再根据不同的策略进行区间选择。由于近红外光谱中各种基团的吸收峰宽窄不等，变量区间宽窄的设置也通常需要使用者根据采集到的近红外光谱和待测性质的具体情况进行优化。MWPLS 及其衍生算法移动窗口大小可调偏最小二乘(changeable size moving window partial least qquares，CSMWPLS)和移动窗口组合搜索偏最小二乘(searching combination moving window partial least squares，SCMW-PLS)等[41]，则通过使用移动窗口建立 PLS 模型的方式寻找信息较为丰富的光谱区间。此外，由于近红外光谱高度重叠性，相邻变量之间往往具有较强的共线性特征，采用区间选择算法选择变量的稳定性一般要高于常规的单一变量选择算

法。尽管变量区间选择的策略不同，最终选中的变量区间一般都可以采用单一变量选择算法进行精简[26, 42]。

除了上述五类变量选择算法外，近年来还出现了一些基于其他原理的变量选择算法，例如随机检验(randomization test, RT)[43]，潜在投影图(latent projective graph, LPG)[44]，局部线性嵌入(locally linear embedding, LLE)[45]，弹性网络分组(elastic net grouping, ENG)[46]等。

2 发展趋势

变量选择算法发展的一个重要趋势是各种新型算法层出不穷，算法复杂程度逐步提高。为了提高选中变量的预测效果，新型变量选择算法往往不再依靠单一指标对变量进行硬性剔除，而是广泛采用较为柔性的剔除或折中策略，例如模型集群分析策略[31-34]和模型共识策略[47-50]。然而，这些策略的应用不仅增大了算法的复杂程度，也大幅度提高了算法需要消耗的计算时间。

变量选择算法发展的另一个趋势是不同算法的联合使用开始出现并增多。联合使用的策略往往利用了算法之间优势的互补性，精选出的变量一般更少，建立模型的预测效果与算法单独使用通常有所提升或相当[29-30, 42]。然而，变量选择算法联用需要使用者对所要联用算法的特点有较深理解，以确保联用算法各自的优势能够形成互补，并且避免其缺点的互相叠加。由于变量选择算法已经有几十种，且各种算法的原理差异较大，如何优化出更好的算法联用组合仍需要大量的研究。

3 实际应用中存在的问题

虽然变量选择算法在大多数应用中都能得到优于全谱模型的预测效果，并起到简化模型的效果。但是变量选择算法在实际应用中仍然存在着一些问题。(1)不同的光谱预处理方法对变量选择算法的影响；(2)部分变量选择算法的稳定性如何提高；(3)部分变量选择算法选中变量的可靠性存疑等。

尽管变量选择算法本身可以剔除近红外光谱中的大部分噪声、无信息以及干扰变量。光谱预处理技术作为一种传统消除样品间光程变动引起的光谱差异，校正颗粒大小不同造成的光谱散射差异，提高光谱信噪比的有力工具，依然是近红外光谱分析一个不可或缺的环节[51-52]。不同光谱预处理方法对于模型的预测性能有很大的影响。然而在大多数文献中，一般只对原始光谱或一阶导数处理后光谱进行变量筛选。不同光谱预处理方法对变量选择算法的影响很少受到关注，因此需要研究不同光谱预处理方法对变量选择算法的影响，以了解不同光谱预处理对于变量选择算法选出变量分布和建立模型预测性能的影响。

一些变量选择算法还存在着稳定性不强的问题，例如在文献[31-34]中，变量选择算法往往需要重复运行几十甚至是几百次，对重复运行结果进行统计以反映变量选择算法对模型优化的最终效果。MCUVE 和 CARS 等常用的变量选择算法，每次重复运行选中的变量分布情况都会发生变化，只有部分变量每次运行均被选中，有时甚至会出现没有任何变量每次运行均能被选中的情况。

变量选择算法选中变量的可靠性也开始成为人们关注的重点，化学计量学目的在于从复杂体系中提取出有效的化学信息。因此，我们不能仅仅依靠各种变量选择算法提供模型的预测性能对其进行评价。例如，虽然 CARS 选择的变量通常可以给出较好的预测模型，但是其选择的变量子集中也容易掺杂进一些噪声变量[17]。因此，如何提高类似算法选择变量的可靠性也是一个不容忽视的问题。

4 结 论

经过近几十年的发展，变量选择算法已经成为了建立简洁高效近红外光谱分析模型的一种重要工具。为此，化学计量学家和近红外光谱研究人员开发了多达几十种的变量选择算法，依据各种算法原理的差异，本文将近红外光谱分析中常见的变量选择算法大致分为基于偏最小二乘模型参数，智能优化算法，连续投影策略，模型集群分析策略和变量区间等五类，并分类概述了各类算法的原理和主要特点。通过文献调研，我们还发现变量选择算法的发展趋势主要集中在以下两点：第一，新出现的变量选择算法复杂程度不断提高，第二，不同变量选择算法之间的联用开始出现并逐渐增多。尽管变量选择算法的发展越来越先进，在具体的应用中依然存在着一些问题，希望广大近红外光谱同行在今后的应用中能够给予相应的关注，进而更好地发挥变量选择算法的应用效果。

References

[1] Manley M. Chemical Society Reviews, 2014, 43(24): 8200.
[2] Li H D, Liang Y Z, Long X X, et al. Chemometrics and Intelligent Laboratory Systems, 2013, 122: 23.
[3] Spiegelman C H, McShane M J, Goetz M J, et al. Analytical Chemistry, 1998, 70(1): 35.
[4] Yun Y H, Liang Y Z, Xie G X, et al. Analyst, 2013, 138(21): 6412.
[5] ZHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立，袁洪福，陆婉珍). Progress in Chemsitry(化学进展), 2004, 16(4): 528.
[6] Zou Xiaobo, Zhao Jiewen, Povey M J, et al. Analytica Chimica Acta, 2010, 667(1): 14.
[7] Mehmood T, Liland K H, Snipen L, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 118: 62.
[8] Li H D, Liang Y Z, Cao D S, et al. TrAC. Trends in Analytical Chemistry, 2012, 38: 154.
[9] Soares S F C, Gomes A A, Araujo M C U, et al. TrAC Trends in Analytical Chemistry, 2013, 42: 84.

[10] Liu F, He Y, Wang L. Analytica Chimica Acta, 2008, 615(1): 10.
[11] Centner V, Massart D L, de Noord O E, et al. Analytical Chemistry, 1996, 68(21): 3851.
[12] Andries J P, Vander Heyden Y, Buydens L M. Analytica Chimica Acta, 2013, 760: 34.
[13] Cai W, Li Y, Shao X. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 188.
[14] Li H, Liang Y, Xu Q, et al. Analytica Chimica Acta, 2009, 648(1): 77.
[15] Zheng K Y, Li Q Q, Wang J J, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 112: 48.
[16] Wang W T, Yun Y H, Deng B C, et al. RSC Advances, 2015, 5(116): 95771.
[17] Lin Z Z, Pan X N, Xu B, et al. Journal of Chemometrics, 2015, 29(2): 87.
[18] Yun Y H, Cao D S, Tan M L, et al. Chemometrics and Intelligent Laboratory Systems, 2014, 130: 76.
[19] Arakawa M, Yamashita Y, Funatsu K. Journal of Chemometrics, 2011, 25(1): 10.
[20] Zou X B, Zhao J W, Huang X Y, et al. Chemometrics and Intelligent Laboratory Systems, 2007, 87(1): 43.
[21] Hörchner U, Kalivas J H. Journal of Chemometrics, 1995, 9(4): 283.
[22] Cao H, Wang Y X , Yang S C, et al. Journal of Chemometrics, 2015.
[23] Allegrini F, Olivieri A C. Analytica Chimica Acta, 2011, 699(1): 18.
[24] Huang X W, Zou X B, Zhao J W, et al. Food Chemistry, 2014, 164: 536.
[25] Leardi R, Nørgaard L. Journal of Chemometrics, 2004, 18(11): 486.
[26] Zou X B, Zhao J W, Mao H P, et al. Applied Spectroscopy, 2010, 64(7): 786.
[27] Paiva H M, Soares S F C, Galvão R K H, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 118: 260.
[28] Ye S F, Wang D, Min S G. Chemometrics and Intelligent Laboratory Systems, 2008, 91(2): 194.
[29] Li J B, Zhao C J, Huang W Q, et al. Analytical Methods, 2014, 6(7): 2170.
[30] Tang G, Huang Y, Tian K D, et al. Analyst, 2014, 139(19): 4894.
[31] Yun Y H, Wang W T, Tan M L, et al. Analytica Chimica Acta, 2014, 807: 36.
[32] Deng B C, Yun Y H, Liang Y Z, et al. Analyst, 2014, 139(19): 4836.
[33] Deng B C, Yun Y H, Ma P, et al. Analyst, 2015, 140(6): 1876.
[34] Yun Y H, Wang W T, Deng B C, et al. Analytica Chimica Acta, 2015, 862: 14.
[35] Deng B C, Yun Y H, Cao D S, et al. Analytica Chimica Acta, 2016, 908: 63.
[36] Norgaard L, Saudland A, Wagner J, et al. Applied Spectroscopy, 2000, 54(3): 413.
[37] Jiang J H, Berry R J, Siesler H W, et al. Analytical Chemistry, 2002, 74(14): 3555.
[38] Yun Y H, Li H D, Wood L R, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013, 111: 31.
[39] Song X Z, Huang Y, Yan H, et al. Analytica Chimica Acta, 2016, 948:19.
[40] de Araújo Gomes A, Galvło R K H, de Araújo M C U, et al. Microchemical Journal, 2013, 110: 202.
[41] Du Y P, Liang Y Z, Jiang J H, et al. Analytica Chimica Acta, 2004, 501(2): 183.
[42] Ranzan C, Trierweiler L F, Hitzmann B, et al. Chemometrics and Intelligent Laboratory Systems, 2015, 142: 78.
[43] Xu H, Liu Z C, Cai W S, et al. Chemometrics and Intelligent Laboratory Systems, 2009, 97(2): 189.
[44] Shao X G, Du G R, Jing M, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 114: 44.
[45] Shan R F, Cai W S, Shao X G. Chemometrics and Intelligent Laboratory Systems, 2014, 131: 31.
[46] Fu G H, Xu Q S, Li H D, et al. Applied Spectroscopy, 2011, 65(4): 402.
[47] Shahbazikhah P, Kalivas J H. Chemometrics and Intelligent Laboratory Systems, 2013, 120: 142.
[48] Li Y K, Jing J. Chemometrics and Intelligent Laboratory Systems, 2014, 130: 45.
[49] Liu K, Chen X J, Li L M, et al. Analytica Chimica Acta, 2015, 858: 16.
[50] Han Q J, Wu H L, Cai C B, et al. Analytica Chimica Acta, 2008, 612(2): 121.
[51] Engel J, Gerretzen J, Szymańska E, et al. TrAC Trends in Analytical Chemistry, 2013, 50: 96.
[52] Tong P J, Du Y P, Zheng K Y, et al. Chemometrics and Intelligent Laboratory Systems, 2015, 143: 40.

Research Advance of Variable Selection Algorithms in Near Infrared Spectroscopy Analysis

SONG Xiang-zhong, TANG Guo, ZHANG Lu-da, XIONG Yan-mei, MIN Shun-geng*
College of Science, China Agricultural University, Beijing 100193, China

Abstract Researchers begin to realize that near infrared spectroscopy analysis model can be simplified by removing some redundant variables from the full-spectrum with the growing understanding of near infrared spectroscopy. It is obvious that the simplified model constructed with retained informative variables can be interpreted more easily. Moreover, both prediction performance and robustness of calibration model can be improved wi hvariable selection, which has been proved in numerous applied examples. Therefore, variable selection has become a critical step in the process of constructing near infrared spectroscopy analysis models, and various kinds of variable selection algorithms and their derivative algorithms have been developed by chemometrics scientists. In order to help the researchers in near infrared spectroscopy analysis field to have a fast overview on variable selection algorithms, we try to review some variable selection algorithms commonly used in near infrared spectroscopy area in this article, including their main rationales and characteristics. These variable selection algorithms are divided into five categories according to their different features. These algorithms are based on parameters of partial least squares (PLS) model, intelligent optimization algorithms, successive projections strategy, model population analysis strategy, and spectral intervals respectively. During the process of carding literatures, we find that the development trends of variable selection algorithms mainly focus on two points: firstly, complexity of new proposed algorithms increaces continually; secondly, the combination of different algorithms becomes more and more popular. Furthermore, we also summarized several specific applied problems that may be occurred when variable selection algorithms are applied in near infrared spectroscopy analysis area. For example, how do different spectral pretreatment methods affect the performance of variable selection algorithm? How to address the poor stability and reliability of some variable selection algorithms?

Keywords Near infrared spectroscopy; Variable selection; Review

(Received Nov. 29, 2015; accepted Apr. 6, 2016)

* Corresponding author