

Ensemble-SISPLS 近红外光谱变量选择方法

李四海¹, 赵磊²

1. 甘肃中医药大学信息工程学院, 甘肃 兰州 730000

2. 甘肃省高校中(藏)药化学与质量研究省级重点实验室, 甘肃 兰州 730000

摘要 近红外光谱具有高维小样本的特点, 变量选择是提高定量分析模型稳健性和可解释性的一种有效方法。确定独立筛选(SIS)是一种基于边际相关性的超高维数据变量选择方法, 广泛用于基因微阵列数据的变量选择。SIS具有将数据维度降低至样本大小规模的能力, 其降维能力与 LASSO 相当, 在相当宽泛的近似条件下, 由于具有安全筛选性质, 所有重要变量被保留的概率趋于 1。基于确定独立筛选偏最小二乘(SIS-SPLS)的变量选择是一种迭代式的 SIS 变量选择方法, 首先利用 SIS 方法完成光谱重要变量的初选; 然后根据重要变量的边际相关性大小进行逐步前向选择: 建立偏最小二乘回归模型, 依据贝叶斯信息准则(BIC)确定最终的变量选择结果。SIS-SPLS 以逐步前向选择的方式实现对重要变量的增量式筛选, 随着潜变量个数的增加及因变量残差的逐步减小, SIS-SPLS 方法选择的变量个数将趋于稳定。然而仅以边际相关性对变量重要性进行评价, 当光谱变量个数远大于样本数时, 该方法也存在选择的变量过多、变量选择结果不够稳健等问题。为进一步提高小样本情况下变量选择的稳健性, 将集成学习引入 SIS-SPLS 方法之中, 提出了一种集成 SIS-SPLS 变量选择方法(Ensemble-SISPLS)。该方法首先对校正集样本进行自助重采样, 对采样得到的每一个校正子集分别使用 SIS-SPLS 方法进行变量筛选, 通过投票机制并设置频次阈值对所有校正子集的变量选择结果进行集成, 选择出现频次大于给定阈值的变量并建立偏最小二乘回归模型, 计算 5 折交叉验证均方根误差。对频次阈值和潜变量个数两个关键参数使用网格搜索法进行优选, 根据子模型的交叉验证均方根误差和变量个数对子模型性能进行综合评价, 以最优子模型包含的变量作为最终的变量选择结果。分别在 Corn 数据集和当归数据集上进行变量选择实验, 比较 Ensemble-SISPLS, SIS-SPLS 和 UVE-PLS 三种变量选择方法的性能。其中当归数据集共 77 个样本, 样本采自甘肃岷县和渭源县, 使用 Nicolet-6700 型近红外光谱仪扫描得到所有样本的近红外光谱并对当归中的阿魏酸含量进行预测。Ensemble-SISPLS 方法在 Corn 数据集上选择的变量个数、RMSEP 和决定系数分别为 22, 0.000 8 和 0.999 8; SIS-SPLS 方法在 Corn 数据集上选择的变量个数、RMSEP 和决定系数分别为 97, 0.007 3 和 0.998 8。Ensemble-SISPLS 方法在当归数据集上选择的变量个数、RMSEP 和决定系数分别为 24, 0.018 1 和 0.996 3; SIS-SPLS 方法在当归数据集上选择的变量个数、RMSEP 和决定系数分别为 38, 0.022 6 和 0.994 3。结果表明, 该方法进一步提高了变量选择结果的稳健性和预测能力。Ensemble-SISPLS 变量选择方法有效结合了 SIS-SPLS 较强的变量选择能力和集成学习良好的泛化能力, 提高了变量选择的稳健性。此外, 由于在子模型的预测能力和变量个数之间进行了折中, 一定程度上减少了选择变量的个数, 提高了模型的可解释性。

关键词 近红外光谱; 变量选择; 确定独立筛选; 偏最小二乘; 集成学习

中图分类号: TP391.4 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2019)04-1047-06

引言

近红外光谱是一种绿色新型的分析检测技术, 具有简

单、快速、无破坏性等特点, 已被广泛用于农产品检测、食品检测、石油化工及中药光谱的定性及定量分析等领域。近红外光谱的变量之间通常存在严重的多重共线性, 如果直接对全光谱进行定量分析会导致模型过于复杂, 计算代价大、

收稿日期: 2018-03-20, 修订日期: 2018-07-08

基金项目: 国家自然科学基金项目(81660577), 甘肃省自然科学基金项目(1506RJZA046), 甘肃省高校中(藏)药化学与质量研究省级重点实验室开放基金项目(ZZY-2018-05)资助

作者简介: 李四海, 1972 年生, 甘肃中医药大学信息工程学院副教授 e-mail: lshroom@163.com

泛化能力不足等问题。因此,在建立定量分析模型之前通常要进行重要变量选择,以提高模型的预测能力和稳健性^[1]。

目前,国内外学者提出了许多基于 PLS 的变量选择方法,包括 PLS-VIP, UVE-PLS, CARS-PLS, 稀疏偏最小二乘等^[2], 这些方法大致可分为三类: 过滤式(Filter methods)、封装式(Wrapper methods)和嵌入式(Embedded methods)^[3]。基于确定独立筛选的偏最小二乘(sure-independence-screening sparse partial least squares, SIS-SPLS)是一种封装式变量选择方法,该方法利用确定独立筛选(sure independence screening, SIS)方法的安全筛选性质,根据边际相关性大小完成变量初选,初选变量数一般不超过样本个数,然后使用偏最小二乘方法以逐步前向选择方式找到最优的初选变量组合。然而,当变量数远大于样本数时,仅仅以自变量和响应变量之间的边际相关性大小对变量重要性进行评价,会导致一些边际相关性虽然较大但并不重要的变量也被选择,其变量选择结果仍然存在冗余^[4]。

集成学习是一种较为新颖的机器学习方法,能够有效提高模型的预测能力和稳健性。文献[5-6]分别建立了 k -近邻和支持向量机分类器对药物动力学中的构效关系进行了研究,通过核方法和决策树集成,有效提高了分类模型的泛化能力。

为进一步压缩选择变量的数量并提高变量选择结果的稳健性,将集成学习方法^[7-8]引入 SIS-SPLS 中,提出了一种 Ensemble-SISPLS 变量选择方法。通过对样本的自助重采样,一方面能够更好地满足 $p \gg n$ 的安全筛选条件,另一方面,能够提高小样本情况下变量选择的稳健性,压缩选择变量的数量,提高模型的可解释性。分别运用 UVE-PLS, SIS-SPLS 及 Ensemble-SISPLS 方法对 Corn 数据集和当归数据集进行变量选择,然后建立 PLSR 模型对 Corn 中的湿度及当归中的阿魏酸含量进行预测,比较了三种变量选择方法的性能。

1 Ensemble-SISPLS 变量选择方法

1.1 确定独立筛选

假设 \mathbf{X} 为原始近红外光谱矩阵, $\mathbf{X} \in \mathbf{R}^{n \times p}$, $p \gg n$ 。共 n 个样本, p 个波长变量, $\mathbf{y}^{n \times 1}$ 为单变量响应变量。对 \mathbf{X} 进行中心化使每个波长变量均满足: 均值为 0、标准差为 1。建立光谱和响应变量之间的线性回归方程

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

式(1)中, $\boldsymbol{\beta}$ 为回归系数向量, $\boldsymbol{\varepsilon}$ 为随机误差项。

令 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$, 满足

$$\boldsymbol{\alpha} = \mathbf{X}^T \mathbf{y} \quad (2)$$

将 $\boldsymbol{\alpha}$ 中的 p 个值依据绝对值大小进行降序排列并选择值较大的前 d 个变量, d 的取值一般为 $n-1$ 或 $n/\log n$ 的整数部分,文献[9]将这种变量选择方法称为确定独立筛选。SIS 是一种硬阈值方法,在相当宽泛的近似条件下具有以概率 1 保留所有重要变量的安全筛选性质^[10]。

1.2 确定独立筛选偏最小二乘(SIS-SPLS)

确定独立筛选偏最小二乘是一种硬阈值稀疏偏最小二乘

变量选择方法,其变量选择分为两个阶段: 首先根据 SIS 方法对光谱数据按照变量边际相关性大小进行降序排列,然后进行偏最小二乘变量选择。

设 \mathbf{X} 为光谱矩阵,共 n 个样本, p 个光谱变量, \mathbf{y} 为单变量响应变量。将 \mathbf{X} 中心化为零均值、标准差为 1。记 H 为潜变量个数, I 为所有 p 个光谱变量的集合, A 为当前的变量选择结果集。SIS-SPLS 变量选择方法的具体过程如下^[11]:

1) 初始化 $A = \emptyset$, $\mathbf{X}_0 = \mathbf{X}$, $\mathbf{Y}_0 = \mathbf{Y}$

2) for $k=1 : H$, k 为潜变量:

(1) 计算 \mathbf{X}_{k-1} 和 \mathbf{Y}_{k-1} 在变量集 I 上的边际相关系数 w_k , 将 \mathbf{X}_{k-1} 按变量相关性大小进行降序排列。

$$w_k = \mathbf{X}_{k-1}^T \mathbf{Y}_{k-1} \quad (3)$$

(2) 进行 PLS 回归。依次选取前 j 个变量,根据贝叶斯信息准则(Bayesian information criterion, BIC),计算 BIC 值。

$$\text{BIC}(j) = n \log \frac{\text{RSS}(j)}{n} + \frac{j}{n} \log(n + \log p) \quad (4)$$

式(4)中, $\text{BIC}(j)$ 表示前 j 个变量的 BIC 值, $\text{RSS}(j)$ 表示 \mathbf{Y}_{k-1} 在前 j 个变量上的残差。选择 BIC 值最小的前 m 个变量得到 A_k ; 分别计算自变量和因变量在 A_k 上的残差 \mathbf{X}_k 和 \mathbf{Y}_k 。

(3) $A = A_k$, $I = I - A_k$ 。

3) 输出变量选择集 A 。

1.3 Ensemble-SISPLS 变量选择

Ensemble-SISPLS 方法在 SIS-SPLS 方法中引入集成学习,通过对校正集样本的自助重采样,得到一定规模的校正子集,然后对校正子集分别进行 SIS-SPLS 变量选择。统计变量在所有校正子集上被选择的总频次并对所有校正子集的选择结果进行集成。对频次阈值 α 和潜变量个数 k 两个关键参数使用网格搜索法进行优化。具体流程如图 1 所示。

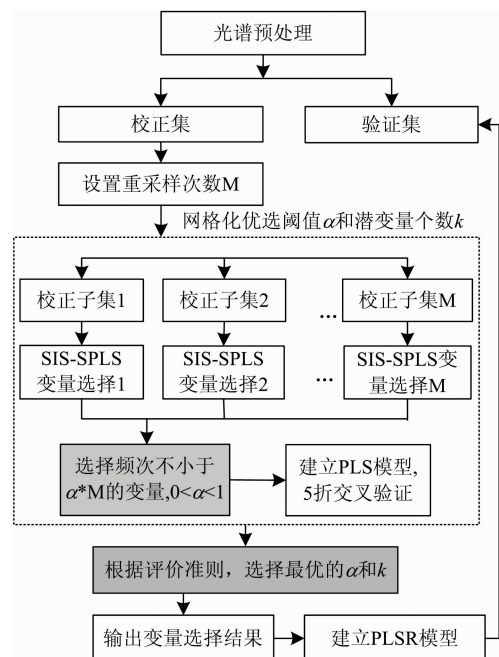


图 1 Ensemble-SISPLS 变量选择方法流程图

Fig. 1 Flowchart of the Ensemble-SISPLS method

Ensemble-SISPLS 方法的具体步骤如下:

Step1 光谱预处理, 划分校正集和验证集。

Step2 对校正集样本进行自助重采样, 抽样比为 2/3, 共抽样 M 次, 取 $M=100$ 。

Step3 生成频次阈值 α 和潜变量个数 k 的网格化数据, 对每一对 (k, α) , 执行以下两步:

(1) 对所有校正子集分别进行 SIS-SPLS 变量选择, 基于投票机制, 选择总次数不小于 αM 的变量, 得到变量选择结果 $\text{result}(k, \alpha)$ 。

(2) 在 $\text{result}(k, \alpha)$ 上建立 PLS 子模型, 计算 5 折交叉验证的 RMSECV (k, α) 的值。

Step4 根据式(5)对子模型性能进行评价, 选择最优的 (k, α) , 以最优子模型包含的变量作为最终的变量选择结果。

$$(k, \alpha) = \min_{k, \alpha} \text{RMSECV} \times |M| \quad (5)$$

其中, RMSECV 为网格法搜索得到的所有子模型的交叉验证均方根误差, $|M|$ 为所有子模型各自选择的变量个数。以上评价准则在子模型的预测能力和变量个数之间进行折中, 防止选择过多的变量。

2 结果与讨论

2.1 数据集

变量选择实验在两个数据集上进行: Corn 数据集和当归数据集。Corn 数据集共 80 个样本, 选择 m5spec 扫描得到的近红外光谱, 预测值为湿度(moisture)。当归数据集的 77 个样本采自甘肃岷县蒲麻镇、甘肃渭源县会川镇。使用 Nicolet-6700 型近红外光谱仪扫描得到所有样本的近红外光谱^[12], 预测值为当归中的阿魏酸含量, 其参考值通过 HPLC 法测定。当归数据集预处理方法为一阶导数结合正交信号校正, Corn 数据集未进行预处理。

2.2 Corn 数据集实验结果及分析

Corn 数据集选择校正集样本 60 个, 验证集样本 20 个。使用 Ensemble-SISPLS 方法在校正集上进行变量选择, 对频次阈值参数和潜变量个数进行网格化寻优, 搜索范围分别为 $[1, 10]$, $[5, 9]$, 步长均为 1, 子模型在校正集上的最优参数空间如图 2 所示。

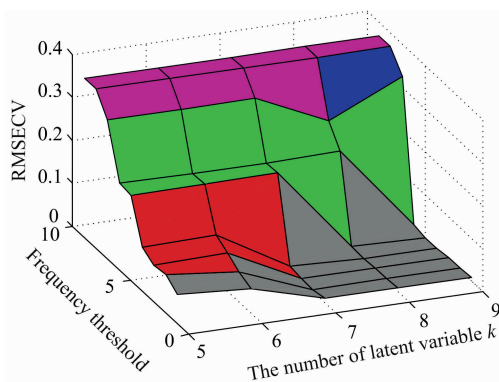


图 2 Corn 数据集网格化寻优结果

Fig. 2 The result of grid optimization on Corn dataset

从图 2 可知, 频次阈值参数对 RMSECV 的影响较大, 潜变量个数影响较小。随着频次阈值的减小, 选择的变量逐步增多, RMSECV 逐步减小, 且存在一个较为平坦的区域, RMSECV 保持在较低水平。分别选择频次阈值参数 $\alpha=0.5$, 潜变量个数 $k=9$, 最终选择 22 个变量, 所选变量大多具有较好的化学意义, 图 3 给出了变量选择结果。

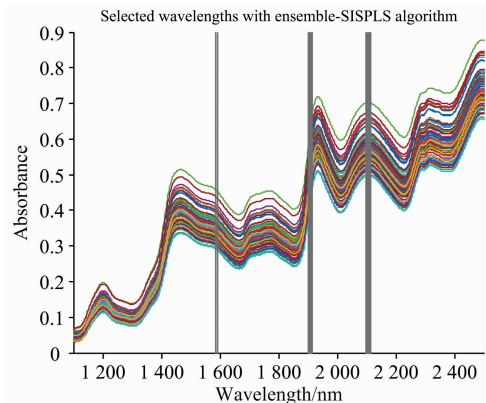


图 3 Corn 数据集变量选择结果

Fig. 3 Wavelength selection results with Ensemble-SISPLS method on Corn dataset

为说明 Ensemble-SISPLS 方法的有效性, 在相同的校正集上, 分别使用 UVE-PLS^[13], SIS-SPLS 变量选择方法对 Corn 数据集进行变量选择, 建立 PLSR 回归模型对相同的验证集样本进行预测, 三种方法的性能对比见表 1, 其中 RMSEP 为预测均方根误差。

表 1 Corn 数据集预测结果对比

Table 1 Comparisons of corn dataset prediction results

变量选择方法	变量个数	RMSEP	决定系数
UVE-PLS	336	0.011 6	0.996 9
SIS-SPLS	97	0.007 3	0.998 8
Ensemble-SISPLS	22	0.000 8	0.999 8

2.3 当归数据集实验结果及分析

当归数据集选择校正集样本 57 个, 验证集样本 20 个。校正集中阿魏酸含量最大值为 $1.548 7 \text{ mg} \cdot \text{g}^{-1}$ 、最小值为 $0.452 2 \text{ mg} \cdot \text{g}^{-1}$ 、平均值为 $0.853 6 \text{ mg} \cdot \text{g}^{-1}$ 。验证集中阿魏酸含量最大值为 $1.596 4 \text{ mg} \cdot \text{g}^{-1}$ 、最小值为 $0.420 1 \text{ mg} \cdot \text{g}^{-1}$ 、平均值为 $0.979 5 \text{ mg} \cdot \text{g}^{-1}$ 。使用 Ensemble-SIS-PLS 方法在校正集上进行变量选择, 对频次阈值参数和潜变量个数进行网格化寻优, 搜索范围分别为 $[1, 10]$, $[1, 10]$, 步长均为 1, 子模型在校正集上的最优参数空间如图 4 所示。

选择频次阈值参数 $\alpha=0.2$, 潜变量个数 $k=5$, 共选择 24 个重要变量, 如图 5 所示。

选择的 24 个变量中, 6 930.9, 6 934.8, 6 938.6, 6 942.5, 6 969.5, 6 973.3, 6 977.2, 6 981.0 和 6 984.9 cm^{-1} 共 9 个波数位于 O—H 基的二倍频峰附近; 5 800.8

cm^{-1} 位于 C—H 基的二倍频峰附近; 4 747.9, 4 751.7, 4 755.6, 4 759.5, 4 782.6, 4 786.5, 4 790.3 cm^{-1} 共 7 个波数可能与芳香环上的 CH 基团有关, 这表明所选的波数具有较好的化学意义^[14]。

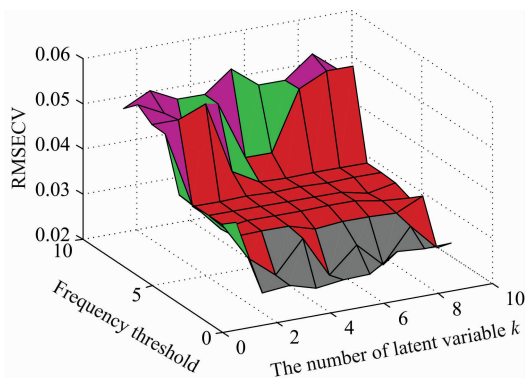


图 4 当归数据集网格化寻优结果
Fig. 4 The result of grid optimization on *Angelica sinensis* dataset

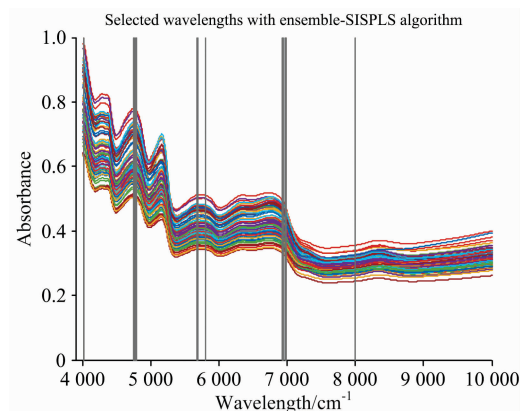


图 5 当归数据集变量选择结果
Fig. 5 Wavelengths selection results with Ensemble-SISPLS method on *Angelica sinensis* dataset

在相同的校正集上, 分别使用 UVE-PLS 和 SIS-SPLS 方法对当归光谱进行变量选择, 根据选择的变量在相同的验证集上分别建立 PLSR 模型, 三种方法的预测性能对比见表 2, 其中 RMSEP 为预测均方根误差。

References

- [1] SONG Xiang-zhong, TANG Guo, ZHANG Lu-da, et al (宋相中, 唐果, 张录达, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2017, 37(4): 1048.
- [2] Wang Z X, He Q P, Wang J. Journal of Process Control, 2015, 26: 56.
- [3] Mehmood T, Liland K H, Snipen L, et al. Chemometrics and Intelligent Laboratory Systems, 2012, 118: 62.
- [4] Kong X B, Liu Z, Yao Y, et al. Test, 2017, 26(1): 1.
- [5] Huang X, Xu Q S, Cao D S, et al. Analytical Methods, 2014, 6(17): 6621.
- [6] Huang X, Cao D S, Xu Q S, et al. Chemometrics and Intelligent Laboratory Systems, 2013, 120: 71.
- [7] Hu Y, Peng S, Peng J, et al. Talanta, 2012, 94(94): 301.
- [8] Qu F, Ren D, Wang J, et al. Sensors, 2016, 16(1): 89.

表 2 当归数据集预测结果对比

Table 2 Comparisons of *Angelica sinensis* dataset prediction results

变量选择方法	变量个数	RMSEP	决定系数
UVE-PLS	316	0.004 1	0.999 8
SIS-SPLS	38	0.022 6	0.994 3
Ensemble-SISPLS	24	0.018 1	0.996 3

2.4 性能比较分析

从表 1 和表 2 可以看出, UVE-PLS 方法在两个数据集上选择的变量均最多, 主要原因在于 UVE-PLS 只能去除部分噪声和冗余变量, 因此对其选择的变量还需要进一步进行筛选。SIS-SPLS 和 Ensemble-SISPLS 方法选择的变量较少, 预测均方根误差较低。与 SIS-SPLS 方法相比, Ensemble-SISPLS 方法选择的变量数进一步减少, 稳健性进一步提高。主要原因在于根据机器学习中的偏差方差 (Bias-Variance) 理论, 通过对校正集样本进行放回抽样并构建一定规模的 SIS-SPLS 子模型, 一定程度上减小了集成模型的方差, 增强了集成模型的泛化能力。通过设置频次阈值参数 α 对子模型的变量选择结果进行集成, 可以灵活控制选择变量的个数。根据网格搜索法得到的最优参数, 可以看到 Ensemble-SISPLS 方法选择的变量数进一步减少, 但选择的变量与响应变量之间具有更好的相关性, 预测能力有一定程度的提高, 表明本文方法进一步压缩了冗余变量, 保留了重要变量, 提高了变量选择的稳健性。

3 结论

变量选择对于提高近红外光谱定量分析模型的预测能力和可解释性具有重要作用。将集成学习引入 SIS-SPLS 方法, 提出了 Ensemble-SISPLS 变量选择方法, 通过对校正集的自助重采样, 构建一定规模的校正子集, 利用 SIS-SPLS 方法得到各校正子集的变量选择结果, 根据提出的投票机制对各校正子集的变量选择结果进行了有效集成。使用网格搜索方法对频次阈值及潜变量个数进行优选, 结果表明, 变量选择结果主要与频次阈值参数有关, 受潜变量个数影响较小。在两个数据集上的实验结果表明, 本方法提高了小样本情况下变量选择的稳健性, 进一步剔除了无效冗余信息, 减少了选择变量的个数, 提高了定量模型的预测能力和可解释性。

- [9] Fan J, Lv J. *Journal of the Royal Statistical Society*, 2008, 70(5): 849.
- [10] Huang X, Xu Q S, Liang Y Z. *Analytical Methods*, 2012, 4(9): 2815.
- [11] Xu X, Cheng K K, Deng L, et al. *Chemometrics and Intelligent Laboratory Systems*, 2017, 170: 38.
- [12] LI Si-hai, CHEN Jian-guo, REN Guo-jin(李四海, 陈建国, 任国瑾). *Transducer and Microsystem Technologies(传感器与微系统)*, 2017, 37(12): 114.
- [13] Zhang R, Chen Y, Wang Z, et al. *Chemometrics and Intelligent Laboratory Systems*, 2017, 163: 7.
- [14] Li B, Wang C, Xi L, et al. *Analytical Methods*, 2014, 6(24): 9691.

A Variable Selection Method Based on Ensemble-SISPLS for Near Infrared Spectroscopy

LI Si-hai¹, ZHAO Lei²

1. School of Information Science & Engineering, Gansu University of Traditional Chinese Medicine, Lanzhou 730000, China
2. Key Laboratory of Chemistry and Quality for Traditional Chinese Medicines of the College of Gansu Province, Lanzhou 730000, China

Abstract Near-infrared spectroscopy has the characteristics of high-dimensional small sample, which means the number of variables is by far larger compared to that of samples. Variable selection is an effective method to improve the robustness and interpretability of quantitative analysis models of near-infrared spectroscopy. Sure Independence Screening (SIS), an effective feature selection method for ultrahigh dimensional space based on marginal correlations between each predictor and response, is widely used for variable selection of gene microarray data. SIS has the ability to reduce the dimensionality of data to the size of the sample, which is comparable to the reduction ability of LASSO. In a fairly general asymptotic framework, the use of SIS with the sure screening property means that all the significant variables remain after employing the variable screening method with probability tending to one. The variable selection method, based on sure independence screening combined with partial least squares regression (SIS-SPLS), is an iterative SIS method. Firstly, the SIS method is used to complete the initial selection of significant variables, then the stepwise forward selection is carried out on the basis of the marginal correlation of selected significant variables; the partial least squares regression model is established, and the final variable selection result is determined according to the Bayesian Information Criterion (BIC). SIS-SPLS implements an incremental screening of important variables in the stepwise forward selection manner. As the number of latent variables increases and the residual decreases gradually, the number of variables selected by SIS-SPLS will stay steady. Whereas, the evaluation of the importance of variables only by the marginal correlation, when the number of spectral variables is much larger than that of samples, will make the selected variable still large in number, or make the robustness of the variable selection results unsatisfactory. To improve the robustness of variable selection results in the case of small samples, a new variable selection method based on ensemble learning, the SIS method and partial least squares regression (Ensemble-SISPLS) was developed in this paper. First, using the bagging ensemble strategy, the bootstrap method was adopted to resample at random on the calibration set. The variable selection was performed by SIS-SPLS on each calibration subset. The variable selection results of all the calibration subsets were aggregated together by the vote rule. The variable whose frequency was greater than the given threshold was selected and the partial least squares regression model was established to calculate the root mean square error of the 5-fold cross validation. The grid search method was utilized to optimize the two key parameters of the frequency threshold and the number of latent variables. Based on the cross-validation root mean square error and number of variables of the sub-models, the sub-model performance was comprehensively evaluated, and the variables included in the optimal sub-model were treated as the final variable selection result. The variable selection experiments were respectively performed on the Corn dataset and the *Angelica sinensis* dataset, several variable selection methods such as Ensemble-SISPLS, SIS-SPLS and UVE-PLS were compared in selected variable number and model robustness. A total of 77 *Angelica sinensis* samples were collected from Minxian and Weiyuan Counties in Gansu Province. Near infrared spectra of all samples were obtained through a Nicolet-6700 near-infrared spectrometer for the prediction of ferulic acid content in *Angelica sinensis*. The number of selected variables, RMSEP and the coefficient of determination of the Ensemble-SISPLS method on the Corn dataset were 22, 0.000 8 and 0.999 8 respectively; the number of selected variables, RMSEP and the coefficient of determination of the SIS-SPLS method on the Corn dataset were 97, 0.007 3 and 0.998 8 respectively. The number of selected variables, RMSEP and the coefficient of determination of the Ensemble-SISPLS method on *Angelica sinensis* dataset were 24, 0.018 1 and 0.996 3 respectively;

the number of selected variables, RMSEP and the coefficient of determination of the SIS-SPLS method on *Angelica sinensis* dataset were 38, 0.022 6 and 0.994 3. The results showed that the Ensemble-SISPLS method further improved the robustness and predictability of the variable selection result. The Ensemble-SISPLS method which combines the variable selection ability of the SIS-SPLS method and the good generalization capacity of ensemble learning can improve the robustness of variable selection. In addition, the evaluation criteria of sub-models manage to make an optimal compromise between the prediction performance and the number of selected variables, which reduces the number of selected variables to some extent and at the same time improves the interpretability of the model.

Keywords Near infrared spectroscopy; Variable selection; Sure independence screening; Partial least squares; Ensemble learning

(Received Mar. 20, 2018; accepted Jul. 8, 2018)

关于《光谱学与光谱分析》调整审稿费收费标准的通知

尊敬的《光谱学与光谱分析》广大作者、读者：本刊自 2018 年 7 月 1 日以后登记的稿件向投稿作者收取审稿费 200 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。
2. 没有国家级基金资助的稿件，请您不要投稿。
3. 不是光谱专业的稿件，请您不要投稿。
4. 与其他文章重合率超过 10% 的稿件，请您不要投稿。

所投稿件经初审通过后，作者会收到缴纳审稿费的通知。请作者及时从我刊网站(<http://www.gpxygpx.com>)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，本刊不再受理。交费后本刊开据增值税电子普通发票，并传至作者提供的电子邮箱，作者可自行打印。

联系电话：010-62181070，62182998

电子邮箱：chngpxygpx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2018 年 6 月 30 日