近红外光谱的选择比率竞争群体分析的变量选择算法

王玉喜[1]，贾振红[1*]，杨　杰[2]，Nikola K Kasabov[3]

1. 新疆大学信息科学与工程学院，新疆 乌鲁木齐　830046
2. 上海交通大学图像处理与模式识别研究所，上海　200240
3. Knowledge Engineering and Discovery Research Institute，Auckland University of Technology，Auckland　1020，New Zealand

摘　要　光谱分析是化学计量学的一个重要应用方向，并已被广泛应用到各个领域，其中光谱变量选择又是光谱分析的重要环节。研究不同的变量选择方法客观地识别有用的信息变量和消除无关或干扰变量十分关键。提出了一种新的变量选择方法，命名选择比率的竞争性群体分析法(SRCMPA)。该算法采用选择比率，自适应加权采样和模型群体分析的思想，并结合了变量排列和指数递减函数方法。关键波长定义为多元线性回归模型中得分值较大的波长，将线性模型PLS下的选择比率的得分值作为评价各波长重要性的指标，然后，根据每个波长的重要性，SRCMPA依次从蒙特卡罗采样中选择N个波长子集，以迭代和竞争的方式运行。在每一次采样运行中，以固定比率的样品以建立校准的PLS模型并计算每个变量的选择比率值，基于排序选择比率的得分值和作为权重的归一化的SR(选择比率)得分值，采用指数递减函数的强制选择和自适应加权采样竞争选择的两步过程来选择关键变量。最后，应用交叉验证(CV)方法来选择具有最低交叉验证均方根(RMSECV)的子集作为最优子集。该算法已在小麦蛋白数据集和啤酒数据集上进行了测试，并使用三种高效算法作对比。通过对实验结果来评估算法优越性，该算法能够找到数据集的关键波长变量的最佳组合，并能用于解释感兴趣的化学特性，通过建模后的评价结果也是最佳的。该算法在啤酒光谱数据集的运行结果，相较于啤酒数据集的全光谱PLS模型，变量个数由567个减少到42个左右。并且模型的RMSECV由0.622下降到0.115，RMSEP由0.823减少到了0.263左右，预测精度分别提高了81.5%和68.1%。Q2_CV和Q2_test也分别由0.940，0.852提高到了0.994和0.995。在小麦蛋白数据集的运行结果，相较于于小麦蛋白光谱数据集的全光谱PLS模型，变量个数由175个减少到18个左右。并且模型的RMSECV由0.607下降到0.292，RMSEP由0.519减少到了0.234左右，预测精度分别提高了51.9%和54.9%。Q2_CV和Q2_test也分别由0.748，0.774提高到了0.931和0.839。

关键词　变量选择；选择比率；自适应加权采样；群体模型分析；蒙特卡罗采样

中图分类号：O65　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2020)04-1056-07

引　言

近几年来，近红外光谱(NIR)分析在石化、制药、环境、临床、农业、食品和生物医学等领域得到了广泛的应用；有时，不同样品的光谱包含的信息非常相近，变量提取困难。灵敏、快速和准确的提取相关变量来预测样品的化学成分是化学计量学的重要内容。一般来说，近红外光谱技术与多变量技术结合用于对相关物质的定性或定量分析。在光谱化学计量学中通常遇到的是具有大量波长变量和相对较少样本的光谱数据情况，在这种情况下建模具有过度拟合的高风险，并导致多变量校准模型不良或低效的预测结果。多变量分析中的变量选择是一个非常重要的步骤，因为消除无关或无信息变量和降低数据维度不仅可以简化校准建模，并在准确性和鲁棒性方面也能改进预测结果。

鉴于变量选择带来的益处，基于不同策略的变量选择方法已被大量提出。这不仅包括传统经典的方法，如前向选择和后向消除[1]；惩罚性方法，如最小绝对收缩和选择算子

收稿日期：2019-04-08，**修订日期**：2019-09-02

基金项目：国家自然科学基金项目(U1803261和61665012)和中华人民共和国教育部国际科技合作项目(2016-2196)资助

作者简介：王玉喜，1989年生，新疆大学信息科学与工程学院研究生　　e-mail：18865279382@163.com

*通讯联系人　　e-mail：jzhh@xju.edu.cn

(LASSO)[2]，弹性网和最小角度回归(LARS)[3-4]；智能学习算法，如遗传算法(GA)[5]，蚁群优化(ACO)[6]和粒子群优化(PSO)[7]。还有一些基于不同的变量排列标准的方法，如回归系数[8]，投影中的变量重要性(VIP)和选择性比率(SR)[9]，蒙特卡罗无信息消除(MC-UVE)[10]和子窗口置换分析(SPA)[11]。随着模型群体分析(MPA)思想的发展，在此基础上提出了一些新的算法如竞争自适应重加权采样(CARS)[12]，变量迭代空间收缩法(VISSA)[13]，搜索空间的交替通缩和膨胀法(ADISS)[14]，自加权变量组合集群分析法(AWVCPA)[15]，变量组合群体分析(VCPA)[16]，自举软收缩法(BOSS)[17]等。

本算法继续了 MPA(模型集群分析)策略算法的优点，首先从大量的子模型中提取有用信息，避免单个模型的结果或参数不可靠性。其次保留变量间的协同与组合效应，在随机采样优化中产生随机变量的组合。并通过收缩策略逐步消除无关变量，保留信息变量。同时还规避掉了此策略算法需要大量的迭代和循环、算法效率低、收敛速度慢的缺点。本算法将时间效率和变量选择效果考虑在内，即降低时间成本，同时能够保证选择出近红外光谱中的信息变量，消除数据集变量中的无信息和干扰变量，增加光谱模型的可靠性与稳定性。还考虑了关键变量以回归系数绝对值定义的问题，MPA 策略下的算法大部分以回归系数绝对值作为变量重要性的依据，以采样技术(如二进制重采样)通过收缩策略逐步消除无关变量，由于回归系数的绝对值并不总是反映变量重要性的真实信息，会受到噪声等诸多因素的影响[18]，从而会对变量选择算法造成不良影响，而以 SR(选择比率)得分值定义的重要变量会更有优势，可以消除噪声诸多因素对光谱数据的影响。啤酒酵母底物数据集在采集光谱时在 1 100～2 500 nm 处存在噪声，本算法可以消除噪声的影响，即采用选择比率可以定位到信息变量区域，减弱噪声因素和无关变量对变量选择算法的影响，减少噪声和无关变量被选入关键变量的可能。

1　实验部分

1.1　数据来源

1.1.1　啤酒数据集

啤酒近红外光谱数据集[19]是使用 NIR Systems Inc. 收集 25 ℃下的分散近红外数据(包括视觉区域)。并以 2 nm 的间隔在 400～2 250 nm 范围内收集。对于该研究，选择了 NIR 区域 1 100～2 250 nm(576 个数据点)。原始提取物浓度表明酵母发酵成酒精的底物被认为是研究感兴趣的化学性质，并用化学方法测量其浓度。通过对提取值进行分类，运用 Kennard-Stone 分类法选取其中的 40 个样本的近红外光谱数据和化学值数据作为校正预测模型集，剩余的 20 个样本的近红外光谱数据和化学值数据作为预测集检验模型。

1.1.2　小麦蛋白数据集

小麦蛋白近红外光谱数据集[20]由 100 个小麦样品组成，其中蛋白质值是感兴趣的研究性质，并用化学方法测量其蛋白质浓度。在 1 100～2 500 nm 区间，具有 701 个光谱点，间隔为 2 nm。由于“大 p，小 n”问题，原始光谱被适当的窗口大小压缩，将窗口大小设置为 4，此数据集减少到 175 个变量，每四个原始变量的平均值作为一个变量值。根据 Kennard-Stone(KS)方法将数据集分成 80 个样品的校准集和 20 个样品的独立测试集。

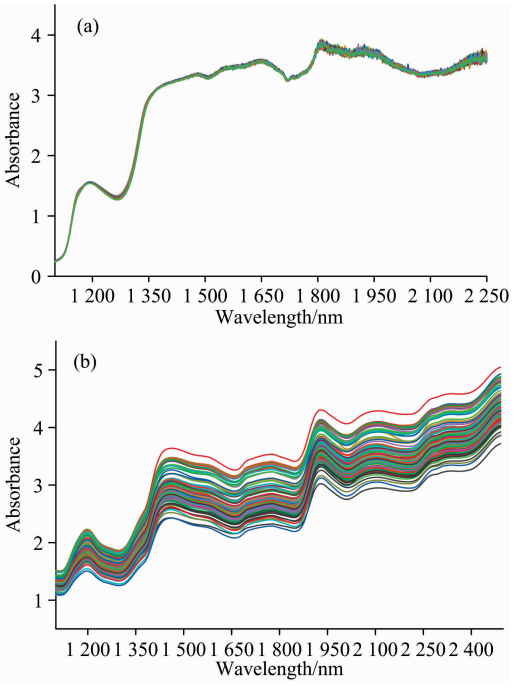

图 1　(a)啤酒光谱；(b)小麦蛋白光谱

Fig. 1　(a) Raw NIR spectra of beer; (b) Raw NIR spectra of wheat protein

1.2　模型校验

假设大小为 $n\times p$ 的数据矩阵 $\boldsymbol{X}$ 包含行中 n 样本和列中的 p 变量，并且大小为 $n\times 1$ 的向量 $\boldsymbol{y}$ 表示所测量的感兴趣属性。在建立 PLS 模型时，$\boldsymbol{X}$ 和 $\boldsymbol{y}$ 都以均值中心化处理。

模型评价参数的作用是评价通过校正集样本建立的预测模型可靠性。在近红外光谱多元校正建模过程中，由相关系数 Q_2、预测均方根误差(RMSEP)和交叉验证均方根误差(RMSECV)对模型评价。模型相关系数 Q_2 越高，即越接近 1 越好。RMSECV 和 RMSEP 越小，即越接近 0 模型预测能力越强。

1.3　使用软件说明

使用的是一台通用联想计算机，内核为 i5 3.2 GHz CPU，内存为 4g，操作系统为 Microsoft Windows 7。所有计算均在 MATLAB 2016a 中进行。数据可视化处理用 Origin2016。

2　SRCMPA 算法原理

将功能指数递减函数(EDF)的迭代次数和蒙特卡罗采样(MCS)次数设置为 N。每次随机 MCS 采样的采样比率为 R。使用上述设置，SRCMPA 可以在每次迭代中分为四个步骤：(1)变量的子集使用固定选择比率的蒙特卡罗抽样随机建立。

(2)计算每个变量的 *SR* 得分值，为一个 p 维的得分值向量，并对其值排序，然后使用 EDF 强制消除排列靠后面的非信息或冗余变量，以变量保留比例 $r_i = ae^{-ki}$，即以 EDF 消除 $r_i \times p$ 数量以外的靠后变量。(3)标准化的 *SR* 分数作为每个波长和自适应加权抽样方法进一步消除变量的权重。有较大权重的变量被保留的概率更大，而权重弱的变量竞争力较弱，并且在变量的群体内逐渐被消除。(4)N 次迭代后会获得 N 变量子集，并应用交叉验证以评估每个子集。其中交叉验证的最小均方根误差的子集被选为最佳子集。

2.1 蒙特卡罗采样

蒙特卡罗抽样是一个用于分析复杂(多元变量)问题十分有效且应用广泛的重要统计工具，在每次采样运行中，样本和变量都分别随固定数量随机选择。MCS 在样本空间和校准集的可变空间中实现，以此获得若干个子数据集，并利用 PLS 等一些回归方法在每个子数据集建立子模型，进而形成模型空间。利用统计分析方法可对每个子模型输出参数分析，来评价每个子数据集感兴趣的未知参数。

2.2 选择比率和回归系数

模型解释是偏最小二乘法(PLS)的大多数应用中的重要任务。从作为潜在回归方法的性质看，偏最小二乘回归提供了一种多对多线性回归建模的方法，能够处理具有严重多重相关性高维度数据。然而，使用潜在变量也会给模型解释带来困难。这种困难是由于 PLS 构造的潜在变量不仅是为了最大化数据矩阵 $\boldsymbol{X}$ 和响应 $\boldsymbol{y}$ 的相关性，而且还同时尝试 $\boldsymbol{X}$ 解释方差的最大化。因此我们无法使用诸如权重和负荷之类的模型参数来直接解释模型。尤其是在受多种变异来源影响的分析数据中，当主要变异源与 Y 无关时，所解释的 $\boldsymbol{X}$ 方差的最大化可能会将无关信息带入 PLS 模型。因此基于这些参数对 PLS 模型和变量重要性解释并不容易。

2.2.1 回归系数

对于回归系数(Beta)重要变量的选择，直接的策略是量化回归系数周围的置信区间，但在 PLS 线性模型下，响应向量 $\boldsymbol{y}$ 依赖其正交投影到由 $\boldsymbol{X}$ 的列向量所生成的子空间上的投影矩阵，即帽子矩阵。PLS 回归系数也没有用于不确定性的封闭分析形式。因此，重采样技术通常用于确定置信区间。各种重采样技术可用于 PLS 回归系数，但并没有一种方法可以在模型中提供变量重要性的直接排序。通常以回归系数的绝对值作为指导，但回归系数的绝对值并不总是反映变量重要性的真实信息，还会受到噪声等诸多因素的影响。

2.2.2 选择比率

关于选择比(SR)[21]，在给定 PLS 的回归系数向量 $\boldsymbol{b}_{\text{pls}}$ 条件下，*TP* 分数是通过以 $\boldsymbol{X}$ 的行在归一化回归系数向量上的投影来实现的，t_{TP} 是与预测值成比例的。对于载荷 P_{TP} 是通过投影 $\boldsymbol{X}$ 的列到分数向量得到的

$$t_{\text{TP}} = \boldsymbol{X}b_{\text{pls}} / \| b_{\text{pls}} \| \quad (1)$$

$$P_{\text{TP}} = \boldsymbol{X}^{\text{T}} t_{\text{TP}} / t_{\text{TP}}^{\text{T}} t_{\text{TP}} \quad (2)$$

解释和残差方差可以通过变量矩阵 $\boldsymbol{X}$ 和投影(TP)分数和载荷来计算

$$\boldsymbol{X} = t_{\text{TP}} p_{\text{TP}}^{\text{T}} + E_{\text{TP}} \quad (3)$$

$$S_{i,\,\text{exp}} = \| t_{\text{TP}_i} p_{\text{TP}_i}^{\text{T}} \|^2,\ i = 1, 2, \cdots, p \quad (4)$$

$$S_{i,\,\text{res}} = \| e_{\text{TP}_i} \|^2,\ i = 1, 2, \cdots, p \quad (5)$$

由式(4)和式(5)确定选择比被定义为对于第 i 个变量的解释的方差 $S_{i,\,\text{exp}}$ 与每个变量的残差方差 $S_{i,\,\text{res}}$ 之比

$$SR_i = S_{i,\,\text{exp}} / S_{i,\,\text{res}},\ i = 1, 2, \cdots, p \quad (6)$$

F 检验定义为高辨别能力的可变区域间的边界和非兴趣区域。为了确定哪一个变量具有高辨别能力和拒绝零假设(解释和剩余方差是相同)，其值必须超过 F 分布的临界值 F_{crit}

$$SR_i > F_{\text{crit}} = F(\alpha, N-2, N-3) \quad (7)$$

应用 *SR* 来重新量化 $\boldsymbol{X}$ 方差，以通过目标旋转或正交滤波策略改进对变量重要性的解释。目的是分配与 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 之间的协方差成比例的信息，同时隔离正交无关变化。参考文献中提出了确定变量重要性的临界阈值。在 SR_i 中评估 F 分布的 $N-2$ 和 $N-3$ 自由度。这项工作中，选择了 F 检验(95%)标准选择候选目标。

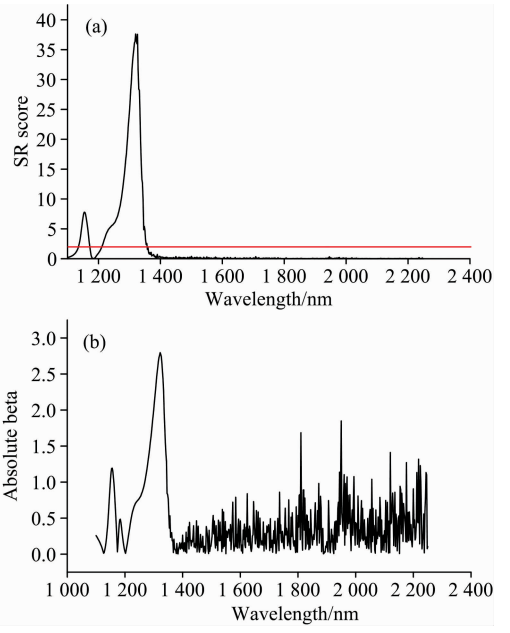

图 2 啤酒光谱数据集以选择比率和回归系数绝对值的变量重要性图示

红色线是代表着变量重要性的阈值

(a)：选择比率得分值；(b)：回归系数绝对值

Fig. 2 The variable importance of the beer spectral data set with the selection ratio and the absolute value of the regression coefficient

The red line represents the threshold of important variable

(a): Selectivity ratio scores;

(b): Absolute value of regression coefficient

用含有噪声的真实的啤酒光谱数据来比较 SR 和回归系数 Beta 绝对值来定义变量重要性的情况，以表明 SR 作为重要变量选择标准的可行性。

图 2 中 *SR* 定义的重要变量的曲线比较平滑，干扰较少。而回归系数的绝对值定义的重要变量还包括了大量无关变量的存在，曲线出现大量的干扰变量，这会对以此为变量重要性的变量选择算法会造成非常大影响，会大大增加无关和干扰变量被选入关键变量的可能性。并且 *SR* 定义的重要变量

区域与啤酒数据集酵母底物化学性质的重要变量吻合，在啤酒光谱数据集中 1 100～1 350 nm 区域对应 O—H 拉伸键振动的第一倍频和 C—H 拉伸键的第二倍频。它符合啤酒光谱集所要研究的感兴趣的酵母底物的化学性质。所以 *SR* 作为变量选择方法的重要变量定义更具有优势，它可以将噪声影响剔除掉。

$SR=[sr_1, sr_2, \cdots, sr_p]^{\mathrm{T}}$ 是 p 维 *SR* 分数向量，其中 *SR* 向量里的值都大于临界阈值，*SR* 分数中第 i 个元素 sr_i 反映第 i 个波长对于 y 贡献。我们评估每个波长的重要性，将 *SR* 进行排序，排名越靠前的变量越重要。我们在这里对于评估每个变量，还要定义归一化的权重用于自适应抽样来竞争选择重要变量

$$w_i = \frac{sr_i}{\sum_{i=1}^{p} sr_i},\ i = 1, 2, \cdots, p \tag{8}$$

另外注意的是被消除的波长的权重被强制变为零，并使得权重向量总是 p 维的。

2.3　功能递减函数和自适应竞争抽样

EDF 被用来模仿“物竞天择”原则。EDF 的选择可分为两个阶段[12]，第一阶段被名为“快速筛选”，有很多不重要的变量会被迅速消除，对于指数递减函数在开始阶段对应的消除比率比较大，消除无信息力度比较大。第二阶段被名为“精细筛选”，随着无信息和不重要的变量的减少，指数递减函数对应的消除比率越来越小，且接近于 0，是为了避免错误的消除关键变量。

在基于 EDF 的强制波长减少之后，SRCMPA 中采用自适应重加权采样(ARS)以竞争方式进一步消除波长。采用自适应采样进一步消除较弱权重的变量，这类似于进化论中的“适者生存”。权重越大的变量具有较大的概率被保留，而其较弱权重的变量竞争性比较差，在变量种群会被逐渐淘汰。

3　结果与讨论

基于 Kennard-Stone(KS)方法将所有数据集分成校准集和独立测试集。KS 方法旨在通过最大化每对所选样本之间的欧几里德距离来覆盖多维空间。校准集用于变量选择和拟合优度，独立测试集用于验证校准模型以进行预测。校准集进行变量选择时，用交叉验证。此外，为了评估 SRCMPA 的性能，我们将与优秀方法 CARS，BOSS，VISSA 进行比较。通过交叉验证与蒙特卡罗采样次数之间的参数优化选择，对于 CARS 和 SRCMPA 的蒙特卡罗采样运行的次数都选择为 300，并且蒙特卡罗采样比率都为 0.9。BOSS 算法的二进制采样次数为 1000，优秀子集占优比率为 0.1。VISSA 算法二进制采样次数为 5000，子集选择比率为 0.05。对于所有方法，最大潜在变量限制为 10，潜在变量的数量由 10 倍交叉验证确定。在建模之前，每个数据集将被均值中心化。所有方法进行 50 次运行以获得统计结果并公平地比较这些方法。

3.1　SRCMPA 算法的变量提取结果

在图 3(a)中，啤酒近红外光谱所选中的信息变量区域主要分布在 1 100～1 350 nm 之间，这个区域与 O—H 键伸缩振动第一倍频区一致。这与本研究感兴趣的酵母底物的化学性质相一致，说明本方法 SRCMPA 能够很好地消除无信息或干扰变量，达到较好的选择信息变量的目的。

在图 3(b)中小麦蛋白数据集中所选的波长变量集中在 1 100～1 400 nm 的区域，这部分区域属于 C—H 拉伸模式的第二倍频和 O—H 的拉伸模式的第一倍频。光谱特征和官能团的振动模式有关。样品中存在的有机物在 NIR 区域具有明显的光谱特征，对应于几个官能团相对强烈的组合模式的吸收强度。本算法选择了相关的信息区域变量，达到消除无关或无信息变量的目的，这也与我们选择研究的小麦蛋白化学有机物的性质相一致，说明本算法 SRCMPA 有很好的选择特性。

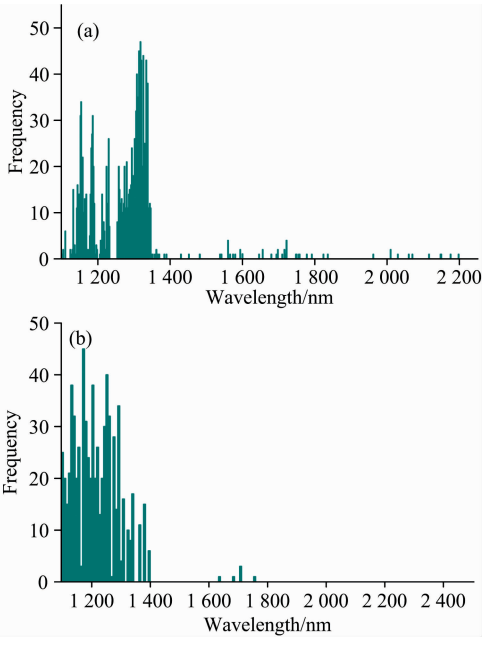

图 3　SRCMPA 运行 50 次后(a)啤酒光谱变量被选取的频率和(b)小麦光谱变量被选取的频率

Fig. 3　(a) Frequency of beer spectral variables selected and (b) frequency of wheat spectral variables selected after running SRCMPA for 50 times

3.2　不同方法建模结果与统计分析

将均值中心化的啤酒和小麦近红外光谱数据在相同条件下分别采用 4 变量选择方法(CARS，VISSA，BOSS，SRCMPA)进行 50 次变量选择选取特征波长，然后利用 PLS 建立预测模型。对模型输出结果平均值和标准差来说明。表 1 和表 2 分别是啤酒中酵母浓度和小麦蛋白以不同方法建模后的结果。本算法在啤酒数据集的运行结果，相较于全光谱 PLS 模型，变量个数已由 567 个减少到 42 个左右。并且模型的 RMSECV 由 0.622 下降到 0.115，RMSEP 由 0.823 减少到了 0.263 左右，预测精度分别提高了 81.5%和 68.0%。Q2_CV 和 Q2_test 也分别由 0.940，0.852 提高到了 0.994 和 0.995，啤酒酵母底物数据集在 1 100～2 500 nm 内采集时存在噪声，本算法消除了噪声的影响，使得建模效果要比其他的算法更有优势。本算法在小麦蛋白数据集的运行结果，相

较于全光谱 PLS 模型，变量个数已由 175 个减少到 18 个左右。并且模型的 RMSECV 由 0.607 下降到 0.292，RMSEP 由 0.519 减少到了 0.234 左右，预测精度分别提高了 51.9% 和 54.9%。Q2_CV 和 Q2_test 也分别由 0.748，0.774 提高到了 0.931 和 0.839。

表 1 不同建模方法对啤酒中酵母浓度的预测结果

Table 1 Results of different methods for beer yeast dataset

Characteristics	PLS	CARS-PLS	VISSA-PLS	BOSS-PLS	SRCMPA-PLS
nVAR	567	41.0±17.8	50.7±0.98	47.7±18.5	42.1±9.3
nLVs	10	5.7±1.1	9.5±0.7	7.9±1.9	7.5±1.3
RMSECV	0.622	0.168±0.030	0.125±0.016	0.110±0.006	0.115±0.010
RMSEP	0.823	0.540±0.009	0.515±0.421	0.591±0.049	0.263±0.029
Q2_CV	0.940	0.986±0.091	0.996±0.003	0.995±0.002	0.994±0.001
Q2_test	0.852	0.933±0.081	0.934±0.002	0.923±0.012	0.995±0.001
T/s	N/A	1.13	162.1	56.2	1.02

注：nVAR：选择变量数；nLVS：潜在变量数；RMSECV：交叉验证均方根误差；RMSEP：预测均方根误差；Q2_CV：交叉验证相关系数；Q2_test：测试集的相关系数；T/s：运行 50 次的平均时间；所有的统计结果均为 50 次运行的平均值±标准差，下同

Note：nVAR：Number of variables；nLVs：Number of latent variables；RMSECV：Root-mean-square error of cross-validation；RMSEP：Root-mean-square error of prediction；Q2_CV：Coefficient of determination of cross-validation；Q2_test：Coefficient of determination of test set；T/s：Average time for 50 runs；All statistical results are the mean values±standard deviations over 50 runs，the same below

表 1 和表 2 说明所有变量选择方法的建模结果都优于全光谱建模，变量选择是十分必要的，可以剔除无信息或干扰变量，消除全光谱建模时的过拟合或不可靠的问题。对比本算法 SRCMPA 与 CARS-PLS，VISSA-PLS，BOSS-PLS 可知，本算法在建模的预测与交叉验证的统计结果上，总体都有最佳的结果，并且在算法运行时间效率上也是最佳的。可以通过节省大量的时间成本，来达到快速建模的目的，预测结果也同时得到保障。SRCMPA-PLS 在啤酒数据集的预测均方根误差(RMSEP)0.263，比 CARS-PLS，VISSA-PLS，BOSS-PLS 的 RMSEP 都要低，预测的相关确定系数(Q2_test)0.995，比 CARS-PLS，VISSA-PLS，BOSS-PLS 的都要高，凸显了本算法的优势。同样在小麦蛋白数据集上模型预测也都有良好的结果。VISSA-PLS 和 BOSS-PLS 虽然可以达到选择信息变量建模提高效果的目的，但效率低，需要非常多的时间消耗在选择变量步骤上面。近红外光谱分析也要考虑到时间成本问题，快速有效的分析模型对现实应用十分重要。

表 2 不同方法小麦蛋白的预测结果

Table 2 Results of different methods on wheat protein dataset

Characteristics	PLS	CARS-PLS	VISSA-PLS	BOSS-PLS	SRCMPA-PLS
nVAR	175	16.9±4.4	10.9±1.0	6.6±0.7	18.2±0.75
nLVs	10	8.4±1.3	8.5±1.1	7.9±0.5	8.2±1.2
RMSECV	0.607	0.277±0.034	0.329±0.019	0.325±0.032	0.292±0.032
RMSEP	0.519	0.418±0.019	0.269±0.012	0.305±0.004	0.234±0.005
Q2_CV	0.748	0.877±0.017	0.921±0.001	0.925±0.001	0.931±0.011
Q2_test	0.774	0.827±0.052	0.829±0.028	0.852±0.039	0.839±0.012
T/s	N/A	0.826	162.09	36.3	0.805

4 结 论

提出了一种新的变量选择方法 SRCMPA，该算法结合了选择比率，自适应加权采样和模型群体分析(MPA)，变量排列和指数递减函数(EDF)竞争的方法。CARS，VISSA 和 BOSS 都以 PLS 的回归系数作为重要信息变量思路，在啤酒和小麦蛋白两种真实光谱的建模情况下，总体效果都不具备 SRCMPA 算法的优势。本算法规避掉了从 PLS 模型以回归系数作为提取重要信息思路的弊端，而采用新的重要变量表示方法选择比率。并且 VISSA 和 BOSS 算法都会在变量选择时花费较多时间，效率比较低，而本算法同样解决了时间效率上的问题。证明了 SRCMPA 能够消除无信息变量和进行波长选择以构建高性能校准模型。

References

[1] QU Fang-fang, REN Dong, HOU Jin-jian, et al(瞿芳芳, 任 东, 侯金健, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(2): 593.
[2] Zhang Ruoqiu, Zhang Feiyu, Chen Wanchao, et al. Chemometrics & Intelligent Laboratory Systems, 2018, 175: 47.
[3] Huang X, Luo Y P, Xu Q S, et al. Anal. Methods, 2017, 9(4): 672.
[4] Alfons A, Croux C, Gelper S. Computational Statistics & Data Analysis, 2016, 93(C): 421.
[5] Ge T, Wei B, Wu D, et al. Journal of Applied Spectroscopy, 2018, 85(1): 109.
[6] Ranzan C, Trierweiler L F, Hitzmann B, et al. Chemometrics and Intelligent Laboratory Systems, 2015, 142: 78.
[7] Cao H, Wang Y, Yang S, et al. Journal of Chemometrics, 2015, 29(5): 289.
[8] Huang X, Luo Y P, Xu Q S, et al. Anal. Methods, 2017, 9(4): 672.
[9] Farrés Mireia, Platikanov S, Tsakovski S, et al. Journal of Chemometrics, 2015, 29(10): 528.
[10] Li C, Zhao T, Li C, et al. Food Chemistry, 2017, 221: 990.
[11] Bin J, Ai F, Fan W, et al. Chemometrics & Intelligent Laboratory Systems, 2016, 158: 1.
[12] Wang Y, Jiang F, Gupta B B, et al. IEEE Access, 2017, (99): 1.
[13] Deng B C, Yun Y H, Liang Y Z, et al. The Analyst, 2014, 139(19): 4836.
[14] Mahanty Biswanath. Chemometrics and Intelligent Laboratory Systems, 2018, 174: 45.
[15] ZHAO Huan, HUAN Ke-wei, SHI Xiao-guang, et al(赵 环, 宦克为, 石晓光, 等). Chinese J. Anal. Chem.(分析化学), 2018, 1(46): 136.
[16] Yun Y H, Wang W T, Deng B C, et al. Analytica Chimica Acta, 2015, 862: 14.
[17] Deng B C, Yun Y H, Cao D S, et al. Analytica Chimica Acta, 2016, 908: 63.
[18] Jiang H, Zhang H, Chen Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 149: 1.
[19] Norgaard L, Saudland A, Wagner J, et al. Applied Spectroscopy, 2000, 54: 413.
[20] Wang Weiting, Yun Yonghuan, Deng Baichuan, et al. RSC Advances, 2015, 5: 95771.
[21] Farrés Mireia, Platikanov S, Tsakovski S, et al. Journal of Chemometrics, 2015, 29(10): 528.

A Variable Selection Method of the Selectivity Ratio Competitive Model Population Analysis for Near Infrared Spectroscopy

WANG Yu-xi[1], JIA Zhen-hong[1*], YANG Jie[2], Nikola K Kasabov[3]

1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
2. Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China
3. Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland 1020, New Zealand

Abstract Spectral analysis is an important application of chemometrics and has been widely used in various fields. Spectral variable selection is a key part of spectral analysis. Therefore, it is critical to study different variable selection methods to objectively identify useful information variables or eliminate irrelevant and interfering variables. In our study, a new variable selection method of the selectivity ratio competitive population analysis (SRCMPA) is proposed. This algorithm adopts the idea of selection ratio, adaptive weighted sampling and model population analysis, and combines the method of variable arrangement and exponential decline function. The key wavelength is defined as the wavelength with a high score value in the regression model. In this paper, the score value of the selection ratio under the PLS model is used as an index to evaluate the importance of each wavelength. Then, according to the importance of each wavelength, SRCMPA sequentially selects N wavelength subsets from Monte Carlo sampling, and runs in an iterative and competitive manner. In each sampling operation, the PLS model is built with a fixed ratio samples and the selection ratio value of each variable is calculated. Based on the score value of the ranking selection ratio and the normalized SR (selection ratio) score value as the weight, the key variables are selected by two steps: the compulsory selection of exponential decline function and the competitive selection of adaptive weighted sampling. Finally, cross validation (CV) method is applied to select the optimal subset with the lowest cross validation mean square root (RMSECV). The algorithm has been tested on wheat protein data set and beer data set, and compared with three efficient algorithms. Through the experimental results to evaluate the superiority of the algorithm, this algorithm can find the best combination of the key wavelength

variables of the data set, and can be used to explain the chemical characteristics of interest, the evaluation results after modeling are also the best. Compared with the PLS model of full-spectrum beer data set, the number of variables in this algorithm has been reduced from 567 to about 42. And the RMSECV of model decreased from 0.622 to 0.115, RMSEP decreased from 0.823 to 0.363, and the prediction accuracy increased by 81.5% and 55.9%, respectively. Q2_CV and Q2_test also increased from 0.940, 0.852 to 0.994 and 0.995. For wheat protein data sets, Compared with the PLS model of full-spectrum wheat protein spectral data set, the number of variables has been reduced from 175 to about 18. And the RMSECV of the model decreased from 0.607 to 0.292, the RMSEP decreased from 0.519 to 0.234, and the prediction accuracy increased by 51.9% and 54.9%, respectively. Q2_CV and Q2_test also increased from 0.748, 0.774 to 0.931 and 0.839.

Keywords Variable selection; Selection ratio; Adaptive weighted sampling; Population model analysis; Monte Carlo sampling

(Received Apr. 8, 2019; accepted Sep. 2, 2019)

* Corresponding author