第42卷，第5期
2022年5月

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

Vol. 42, No. 5, pp1601-1606
May, 2022

牛奶蛋白质含量的SSA-SVM高光谱预测模型

刘美辰，薛河儒*，刘江平，代荣荣，胡鹏伟，黄 清，姜新华

内蒙古农业大学计算机与信息工程学院，内蒙古 呼和浩特 010000

摘 要 牛奶中包含着很多人体需要的营养元素，如脂肪、蛋白质、钙等；对牛奶营养元素进行分析是牛奶安全检测关键的一部分。高光谱技术可以有效地结合图像和光谱数据识别牛奶种营养元素。为了实现对牛奶中蛋白质含量快速、精确的预测，采用竞争性自适应重加权(CARS)算法选取特征波长，并提出一种基于麻雀搜索算法(SSA)优化支持向量机(SVM)实现对牛奶蛋白质含量预测。利用高光谱仪获取牛奶反射光谱(400～1 000 nm)。通过选取归一化(N)、标准化(Standardization)和多元散射校正(MSC)对原始的牛奶数据进行光谱降噪处理提高光谱利用率；利用竞争性自适应重加权算法和连续投影算法(SPA)对经过处理的牛奶光谱数据提取特征波长，求取蛋白质和光谱间的相关系数并进行重要性排序，获取重要的特征波段；最后，通过遗传算法(GA)优化SVM，粒子群算法(PSO)优化SVM和偏最小二乘法(PLS)算法对牛奶蛋白质进行预测并比较预测结果，为了提高蛋白质预测的精度和模型稳定性，提出利用SSA对SVM的核函数 g 和惩罚参数 c 进行优化，以均方根误差(RMSE)作为适应度函数，通过迭代选择最优的回归参数训练模型。牛奶数据预测结果表明最优组合模型为：MSC-CARS-SSA-SVM。模型测试集的决定系数 R^2 为0.999 6，均方根误差RMSE为0.001 1，耗时4.112 1 s。结果表明：使用CARS算法能实现特征波段的提取和冗余信息的剔除，从而提高模型效率，简化了算法的复杂度；SSA算法优化SVM的参数，通过迭代更新麻雀最优位置，可以快速得到全局最优解，与SVM，GA-SVM，PSO-SVM和PLS相比，牛奶蛋白质的预测准确度和模型稳定性都得到了明显提高，满足了对乳品检测的精确度要求，是快速检测牛奶蛋白质的一个可行新方法。为光谱模型的优化及预测模型精度的提高提供参考。

关键词 高光谱；牛奶蛋白质；竞争性自适应重加权算法；支持向量机；麻雀算法

中图分类号：TP79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)05-1601-06

引 言

牛奶作为天然乳液，包含着很多人体需要的营养元素。近些年，对牛奶中蛋白质的含量检测日益得到社会的重视。最常见的乳品检测方法为化学分析法，虽然准确度高，但检验过程复杂且成本大。高光谱利用400～1 000 nm内的光谱波段对牛奶进行检测，不仅操作简单、预测快速无损，而且效果好。

国内外很多学者对牛奶中的营养元素光谱检测做了研究。Laverroux等使用液相色谱结合荧光法对牛奶中的维生素B2进行分析，展示了在浓度区间内对营养元素预测的一种新方法[1]；范睿等基于近红外光谱，建立主成分回归模型对牛奶中的掺假蛋白质检测[2]；Lin等基于近红外光谱检测蛋白质，通过运用支持向量机(support vector machine，SVM)和BP神经网络(back-propagation artificial neural network，BP-ANN)以及偏最小二乘法(partial least-regression，PLS)建模，结合多种预处理方法，最终得分最高的模型DOSC-KPLS的 R^2 达到了0.974[3]。使用高光谱对牛奶进行分析的研究也有报道，近年，Munir等将光谱处理的奶粉图像空间属性与化学属性相结合，使用PLS建立回归模型实时显示质量，证明高光谱技术可以用于乳制品检测[4]。赵紫竹等为了对牛奶中脂肪含量进行检测使用了NPLS算法[5]；张倩倩等采用主成分回归和最小二乘支持向量机两个方法对牛奶中的蛋白质进行定量分析。目前基于高光谱技术的牛奶定量分析中，由于牛奶成分复杂，原始光谱数据中变量间的相关性低，模型的预测精度仍需要提高，分析方法也需要提升。

收稿日期：2021-04-16，**修订日期**：2021-07-16

基金项目：国家自然科学基金项目(61461041，31960494)，内蒙古科技厅关键技术攻关项目(2020GG0169)资助

作者简介：刘美辰，1997年生，内蒙古农业大学计算机与信息工程学院硕士研究生 e-mail：lmcdrr@163.com

*通讯作者 e-mail：xuehr@imau.edu.cn

综上，针对牛奶分析精度和方法的需求，实验以五种蛋白质含量不同的牛奶为研究样本，测量其 400～1 000 nm 范围的高光谱反射数据，结合蛋白质含量数据，建立 SSA-SVM 模型，通过综合考察多种预处理和特征波长提取方法对建模的作用。并采用常应用于牛奶分析中的偏最小二乘法回归(PLS)算法、具有强非线性拟合能力的支持向量机回归(SVM)算法以及两个基于 SVM 的优化算法 GA-SVM 和 PSO-SVM 作为对比，找出测量牛奶蛋白质含量的最优方法，为牛奶营养元素含量的定量检验提供参考。

1 实验部分

1.1 样品采集

所用的样品是从市面上购买的五种蛋白质含量不同的牛奶，蛋白质含量分别为 3.0，3.2，3.3，3.4 和 3.6 g·(100 mL)$^{-1}$。培养皿的直径为 90 mm 高度为 8 mm。分别取不同批次的五种牛奶置于培养皿中，用玻璃棒搅拌。共 250 个样本，其中训练集 175 个，测试集 75 个。

1.2 高光谱数据采集

利用图 1 所示 HyperSpec VNIR 高光谱仪采集样本的牛奶反射光谱。光谱扫描范围在 400～1 000 nm 区间内，光谱的波段数为 125 个，平均间隔为 0.8 nm，曝光时间选择 10 ms，像元混合次数为 6 次，分辨率为 4.8 nm，采集的光谱图像像素为 777×1 004。采集条件：室温(23～25 ℃)，牛奶样本置于光谱仪探头垂直下方的吸光黑色绒布上，于暗室中测量。卤素灯至样本 40 cm 的距离垂直照射，扫描探头到样本的垂直距离为 15 cm。测量前，对光谱图像标定，分别采集白板和全黑的标定图像进行校正以消除因光源和暗电流存在导致的噪声[6]。

图 1 高光谱仪

Fig. 1 High spectrometer

1.3 数据处理

图 2(a)为高光谱仪采集图像后经 ENVI 导出的牛奶样本原始高光谱曲线。

因测量时室内光线、角度、温度等因素可能对光谱数据造成误差，光谱的噪声很大，所以需对低质量的数据进行优化，减小对目标结果的影响。采取的预处理方法有：归一化(normalization，N)、标准化(standardization)和多元散射校正(multiplicative scatter correction，MSC)。

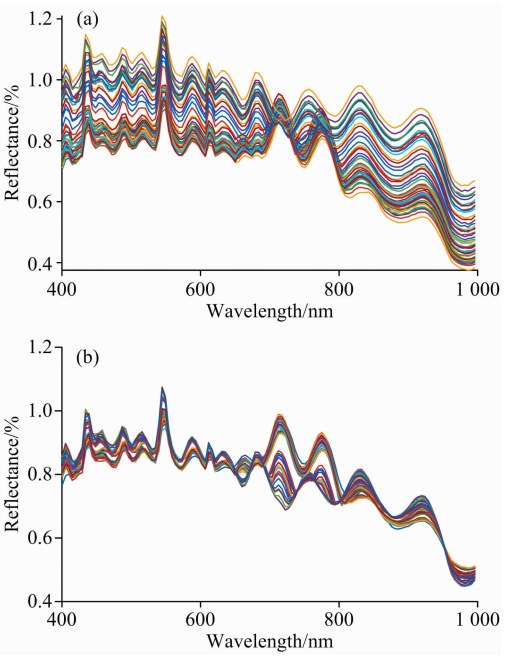

图 2 原始光谱和预处理光谱

Fig. 2 The original and pretreated spectra

1.4 建模方法

1.4.1 竞争性自适应重加权算法采样算法

通过竞争性自适应重加权算法(competitive adaptive reweighted sampling，CARS)选取特征波段，算法基于蒙特卡洛采样结合偏最小二乘法(PLS)，由衰减指数法(EDF)来决定波长个数[7]，通过自适应重加权采样(adaptive reweighted sampling，ARS)保留波长回归系数的权重值大的集合，创建 PLS 模型，引入交叉验证，不断优化计算均方根误差(root mean squard error，RMSE)，选择 RMSE 最小的子集，即模型精度最高的特征波长组合[8]。

1.4.2 基于优化算法的 SVM 参数优化

SVM 作为一个经典的预测算法具有着较强的非线性拟合能力和鲁棒性，并且因预测精度高和复杂度低的优势常被应用于牛奶定量分析的研究中，但据预测结果表明其对于参数选取的敏感度高，若选取不当会使性能降低[9]。故需要选取合适的优化算法对 SVM 优化提高计算精度。目前已有的对其改进的优化算法均可实现参数选取的要求，但由于大多欠缺于局部性能，所以计算精度仍需改进。

1.4.3 基于 GA 的 SVM 参数优化

遗传算法(genetic algorithm，GA)是一种常被应用于预测模型的全局优化算法。其原理是通过生物作用机制，对当前研究的种群进行筛选，逐步选出适应度最高的个体[10]。实验最大迭代次数和设置为 200，基因位置为参数值，将参数带入 SVM 模型中，对牛奶数据进行训练并计算个体的适应度值，如达最大迭代次数，则停止搜索。输出全局最优的参

数值，实现对 SVM 的参数优化。

1.4.4　基于 PSO 的 SVM 参数优化

粒子群算法(particle swarm optimization，PSO)的原理是利用种群中的各粒子通过学习不断调整位置和速度来实现优化[11]。实验最大迭代次数为 200，SVM 参数取值范围一定，利用 SVM 对牛奶数据做出预测，计算粒子的适应度值并更新其速度和位置，当达到最大迭代次数时，跳出循环，输出全局的最佳参数并训练 SVM 模型，实现 PSO 对 SVM 的参数优化。

1.4.5　基于 SSA 的 SVM 参数优化

麻雀搜索算法(sparrow search algorithm，SSA)是通过麻雀的捕食、追随、侦查三种行为为启发而创建[12]。算法将麻雀种群分为发现者和跟随者，麻雀属性为位置，对应优化的解，适应度值对应觅食位置。其中发现者发现者和跟随者的位置是动态变化的，凡是发现者会选择最好的位置觅食。

追随者会在发现者周围觅食或与发现者争夺食物，当追随者发现更好觅食位置则更新发现者的解，否则不变。一旦发现危险，边缘麻雀会迁移到安全区躲避危险，同时就处于最佳位置的麻雀会在周围位置走动。SSA 基于以上步骤循环搜索最优解。其对支持向量机算法进行优化的主要思路为用麻雀位置表示 SVM 的参数 c 和 g，通过对全局的适应度值排序，求最优值和最优位置，得知最优参数[13]。

SSA 优化 SVM 的流程图见图 3。

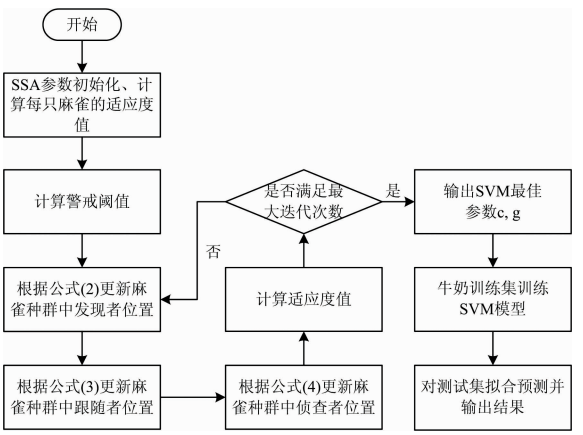

图 3　SSA 优化 SVM 参数流程图

Fig. 3　Flow chart of SVM parameters optimized by SSA

主要步骤如下：

(1) 参数初始化。对麻雀种群初始化，最大迭代参数 $\mathrm{iter}_{\max}$，麻雀总数、发现者和跟随者比例以及参数 c 和 g 的取值范围。

(2) 适应度设置。适应度函数设置为 SVM 训练后的 MSE 误差

$$\mathrm{fitness} = \mathrm{argmin}(\mathrm{MSEpridect}) \tag{1}$$

对全局的适应度值排序，求最优值和最优位置。

(3) 更新发现者位置。

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \exp\left(\dfrac{-i}{\alpha \mathrm{iter}_{\max}}\right), & R_2 < ST \\ X_{i,j}^{t} + QL \end{cases} \tag{2}$$

式(2)中，迭代次数为 t 时，第 i 只麻雀的 j 维位置信息，α 为(0，1]的随机数，$\mathrm{iter}_{\max}$ 为最大迭代次数，R^2 是[0，1]中的一个随机数，代表安全值，ST 为警戒阈值，Q 为一个标准正态分布随机数。ST 为警戒阈值。

(4) 更新追随者位置。

$$X_{i,j}^{t+1} = \begin{cases} Q\exp\left(\dfrac{x_{\mathrm{worst}}^{t} - x_{i,j}^{t}}{i^2}\right), & i > \dfrac{n}{2} \\ X_{p}^{t+1} + \left| x_{i,j}^{t} - X_{p}^{t+1} \right| A^{+} L \end{cases} \tag{3}$$

式(3)中，x_{worst} 为最坏位置，x_b 则为最好的位置。$\boldsymbol{A}^{+} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$，其中，$\boldsymbol{A}$ 为一个大小为 $1 \times D$ 的矩阵，当 $i > n/2$ 时，其值会收敛于 0，第 i 只获取食物少的麻雀需要更新位置获取食物。当 $i \leqslant n/2$ 时，值收敛于最优位置。

(5) 更新侦查者位置。

$$x_{i,j}^{t+1} = \begin{cases} x_{\mathrm{best}}^{t} + \beta \left| x_{i,j}^{t} - x_{\mathrm{best}}^{t} \right|, & f_i > f_g \\ x_{i,j}^{t} + K\left(\dfrac{\left| x_{i,j}^{t} - x_{\mathrm{worst}}^{t} \right|}{(f_i - f_w) + \varepsilon}\right), & f_i = f_g \end{cases} \tag{4}$$

式(4)中，x_{best} 代表最优位置，f_g 表示全局的最优位置上的麻雀的最佳适应度值，f_w 为最坏位置上麻雀的适应度值，K 是区间[−1，1]内的一个随机值，ε 为一个为了防止分母为 0 的极小数。

(6) 计算各适应度值。对麻雀的位置更新，直到最大迭代次数。

(7) 如满足最优解，则输出最优解 c 和 g。否则。继续执行步骤(2)—(6)。

1.4.6　评价指标

预测模型的好坏依据以下几个参数评估：决定系数(determination coefficients，R^2)、均方根误差(root mean squared error，RMSE)、耗时。RMSE 越小则表明预测的数据与原始数据重合度越高、R^2 越趋近于 1 回归效果越好。

2　结果与讨论

2.1　三类预处理的 PLS 模型精度

三类预处理的 PLS 模型精度如表 1 所示。选用多元散射校正(MSC)预处理的模型的拟合程度最高，预处理后的光谱曲线如图 2(b)所示，可以看出经过 MSC 处理后，可以有效的消除一些由散射、光程突变、乳液不均匀带来的影响，消除了牛奶光谱数据的噪声干扰。从而提高了光谱的灵敏度和实用性。

表 1　预处理模型精度对比

Table 1　Precision comparison of preprocessing models

建模方法	RMSEP	R_{p}^2	RMSEC	R_{c}^2
N-PLS	0.313 6	0.924 4	0.129 3	0.963 5
Standardization-PLS	0.286 9	0.927 7	0.120 7	0.965 4
MSC-PLS	0.261 6	0.931 6	0.105 4	0.968 9

2.2　SPA 算法特征选择

连续投影算法(successive projections algorithm，SPA)在食品检测分析中常被用于特征选择，能从众多波段中选出冗

余信息少的波段，从而使共线性问题达到最小化，减小模型变量，提高运算效率[14]。以经过 MSC 处理后的牛奶光谱值为输入变量，通过反复迭代投影值，对最终得到的特征变量进行回归分析，以均方根误差 RMSE 最小的集合为特征波段集合。选出的特征波段有 17 个。

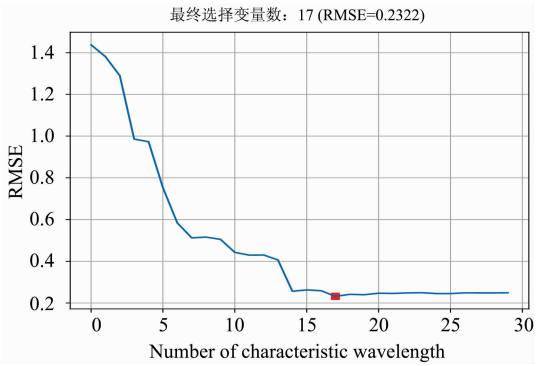

图 4 不同特征波长个数对应的 RMSE 分布曲线

Fig. 4 RMSE *vs* number of characteristic wavelengths

2.3 CARS 算法特征选择

图 5 为采用 CARS 提取牛奶的全部波长变量进行特征提取后的运算结果图。设置蒙特卡罗采样次数为 20[15]。图 5(a)显示随着采样次数增加，由于衰减指数 EDF 的作用，开始在 CARS 的光谱粗选阶段中会用较大的学习率，选择的波长数量快速下降，到较优解的时候逐渐降低学习率来训练模型，即 CARS 的精选阶段。结果显示选取的变量数在前 7 次中下降很快，之后下降速度减慢至平稳。同时，随着采样次数的增加，部分无关波段被剔除导致 RMSECV 值整体呈现下降趋势，图 5(b)可知开始迭代时，由于牛奶数据中大量和

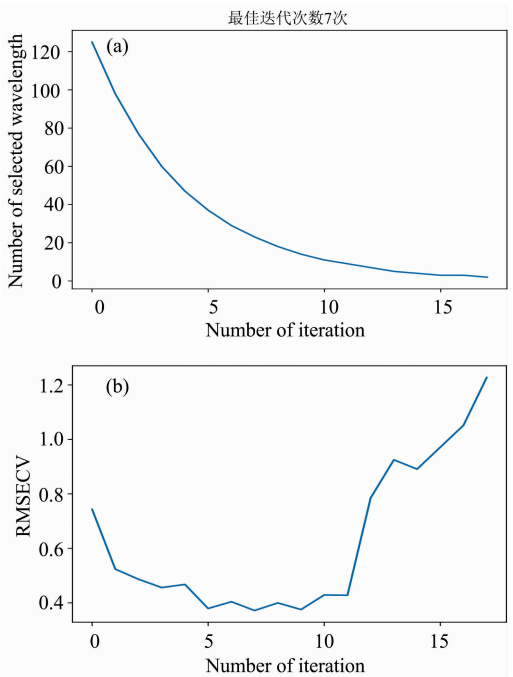

图 5 CARS 特征变量选择

Fig. 5 Characteristic variable selection by CARS

蛋白质无关的波段被消除，故 RMSECV 值快速减小，当采样次数到第 7 次时均方根误差达到最低值，后由于部分重要信息被剔除，RMSECV 值又整体呈上升趋势。故当迭代次数为 7 时所选择的波长集合为最优集合。7 次采样得到 27 个特征波长。

2.4 模型预测

惩罚参数 c 和核函数 g 是支持向量机中两个重要参数，影响着预测模型的精度；因此，选取最优的 c 和 g 是 SVM 优化的关键[16]，也是本工作提出用 SSA 算法优化 SVM 的初心。经反复调试，麻雀算法的参数设置为：麻雀数量 $N=20$，最大迭代次数 Max_iteration=200；SVM 的惩罚参数 c 和核函数 g 取值范围均为$[2^{-5}, 2^{5}]$。最终得到的最佳参数为：$c=3.482\ 5$，$g=0.031\ 2$。也就是说，用此最佳的 c 和 g 值，SSA-SVM 模型的精度最高。

为了考察 SSA-SVM 模型的预测性能，选取 PLS 和用遗传算法(GA)、粒子群算法(PSO)优化的 SVM 模型比较。将 CARS 和 SPA 分别选取出的 27 和 17 个特征波段作为几个模型的输入变量；GA-SVM 的最佳参数为 $c=7.221\ 8$，$g=0.043\ 5$，PSO-SVM 的最佳参数为 $c=17.187\ 2$，$g=0.051\ 2$。将最佳参数代入模型，预测牛奶的蛋白质含量，通过对比各个模型的精度和耗时，选取出最优的牛奶蛋白质预测模型。

对比蛋白质含量预测精度：全波段数据作为输入变量时，SSA-SVM 算法较 PLS 算法准确率提高了 5.83%，较 SVM 算法准确率提高 6.03%，较 GA-SVM 算法和 PSO-SVM 算法准确率分别提高了 2.68%和 5.2%。表明 SSA-SVM 为牛奶蛋白质最佳的预测模型，可以有效的预测牛奶中蛋白质含量。从图 6 可以看出 SSA-SVM 算法的适应度值随适应度迭代次数增加逐渐趋近最优值，到达最低点后保持稳定。

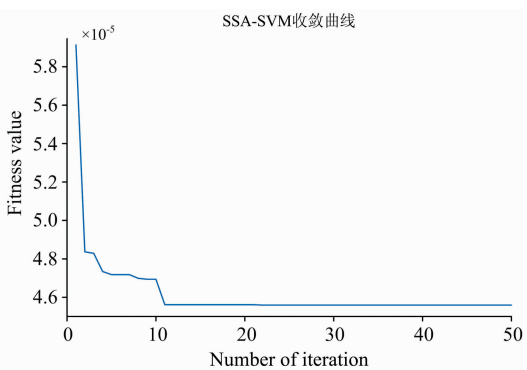

图 6 适应度曲线

Fig. 6 Fitness curve

表 2 显示将全部 125 个波段、CARS 提取的 27 个特征波段和 SPA 提取的 17 个特征波段分别作为输入变量时，利用五种不同的预测模型对牛奶蛋白质预测的精度对比。当全波段作为输入时，SSA-SVM 模型对蛋白质含量预测可以满足预测要求，$R^2=0.988\ 9$，RMSE=0.009 2，耗时 6.056 9 s。SVM、PLS 两种经典回归模型耗时虽短，但准确率较低。

GA-SVM 和 PSO-SVM 虽然精度较传统 SVM 得到提高，但仍低于本文提出的 SSA-SVM 模型，并且耗时过长。由表 2 可以更直观的看出不同模型针对牛奶数据的预测效果。预测效果排序为：SSA-SVM＞GA-SVM＞PSO-SVM＞PLS＞SVM。且在特征波段选择时，CARS 的效果好于 SPA，当以 SSA-SVM 为预测模型时，CARS 选择的波段作为输入变量准确度比 SPA 选择波段高 0.12%，比全波段高 1.07%。

结果表明，SSA 算法较于 GA、PSO 算法对 SVM 的优化更具有优势，不容易陷入局部最优，有效提高了牛奶蛋白质预测的准确率，且迭代时间得到明显缩短。PLS 和 SVM 算法虽耗时短，但预测准确率低，在牛奶蛋白质预测中不具有优势。

表 2　各模型精度对比

Table 2　Precision comparison of different models

建模方法	变量数	RMSE	R^2	Time/s
Full-PLS	125	0.091 4	0.930 6	6.132 1
CARS-PLS	27	0.037 6	0.959 8	5.755 4
SPA-PLS	17	0.052 7	0.956 1	5.269 7
Full-SVM	125	0.105 4	0.928 6	4.231 1
CARS-SVM	27	0.040 1	0.957 1	3.987 1
SPA-SVM	17	0.053 3	0.948 6	3.236 7
Full-GA-SVM	125	0.029 7	0.962 1	16.326 9
CARS-GA-SVM	27	0.015 3	0.979 8	13.143 2
SPA-GA-SVM	17	0.023 1	0.968 7	10.052 7
Full-PSO-SVM	125	0.088 4	0.936 9	13.241 1
CARS-PSO-SVM	27	0.031 6	0.964 9	10.231 1
SPA-PSO-SVM	17	0.037 6	0.959 8	8.534 1
Full-SSA-SVM	125	0.009 2	0.988 9	6.056 9
CARS-SSA-SVM	27	0.001 1	0.999 6	4.112 1
SPA-SSA-SVM	17	0.004 8	0.998 4	4.052 7

表 2 可知：CARS-SSA-SVM 的预测效果最好，R^2＝0.999 6，RMSE＝0.001 1，耗时为 4.112 1 s。预测结果如图 7 所示。

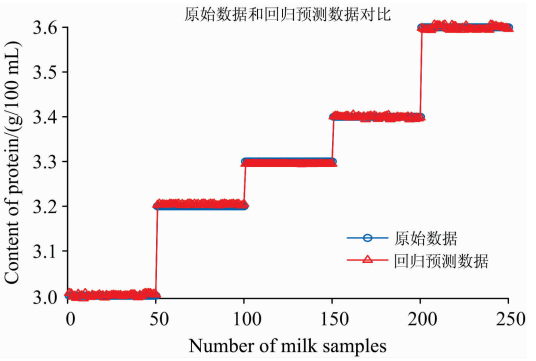

图 7　CARS-SSA-SVM 验证结果

Fig. 7　CARS-SSA-SVM verification results

3　结　论

以五种蛋白质浓度不同的牛奶作为实验对象建立 SVM，PLS，GA-SVM，PSO-SVM 和 SSA-SVM 五种不同的模型，研究了利用牛奶的高光谱反射率对蛋白质含量进行预测，结论如下：

（1）MSC 算法可以有效的对原始牛奶光谱数据预处理，降低了噪声干扰，消除了光谱差异，增强了光谱和数据的相关性，有效降低了牛奶蛋白质预测模型的误差，提高了准确率。

（2）使用 CARS 和 SPA 算法对牛奶数据进行特征提取，CARS 算法较 SPA 算法更能增强了光谱和蛋白质含量之间的相关性，消除了冗余信息，大大降低了模型输入变量多导致的复杂性，同时模型的预测效率也得到了有效的提高。

（3）采用 SVM，PLS，GA-SVM，PSO-SVM 和 SSA-SVM 五种回归算法对比实验。其中利用 SSA-SVM 算法选择最佳参数的模型具有较高的精度。在 CARS 的特征波长作为输入变量时，验证集 R^2＝0.999 6，RMSE＝0.001 1，相较于另外四个模型，稳定性和准确性得到明显提高，满足了预测要求。

（4）SSA-SVM 牛奶营养含量分析方法精度高、实验过程简便、耗时小。是牛奶中蛋白质含量预测的新方法。下一步目标是有关牛奶营养元素的含量建立更加理想化的预测模型，对有效波长的选择进行更深入的研究分析。

References

[1] Laverroux S, Picard F, Andueza D, et al. Food Chemistry, 2021, 342: 128310.

[2] FAN Rui, SUN Xiao-kai, CHEN Jie, et al (范　睿, 孙晓凯, 陈　杰, 等). Modern Food Science and Technology(现代食品科技), 2017, 33(11): 264.

[3] HUANG Bao-ying, SHE Zhi-yun, WANG Wen-min, et al(黄宝莹, 佘之蕴, 王文敏, 等). China Brewing(中国酿造), 2020, 39(7): 16.

[4] Munir M T, Vilson D I, Yu W, et al. Journal of Food Engineering, 2018, 221: 1.

[5] ZHAO Zi-zhu, WEI Yong, ZHANG Nai-qian, et al(赵紫竹, 卫　勇, 张乃迁, 等). China Dairy Industry(中国乳品工业), 2018, 46(2): 45.

[6] LUO Hao-dong, LIU Cui-ling, SUN Xiao-rong, et al(罗浩东, 刘翠玲, 孙晓荣, 等). China Brewing(中国酿造), 2021, 40(4): 183.

[7] Xuan Wang, Wang Yanxue. Journal of Physics: Conference Series, 2021, 1820(1): 012078.

[8] ZHANG Xu-hui, ZHANG Kai-xin, ZHANG Chao, et al(张旭辉, 张楷鑫, 张　超, 等). Journal of Xi'an University of Science and Technology(西安科技大学学报), 2020, 40(5): 760.

[9] Yang J, Sun L J, Xing W, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 253: 119585.
[10] SHEN Zhi-chao, ZHAO Zhi-heng, LU Lei, et al(申志超, 赵志衡, 卢 雷, 等). Journal of the Chinese Cereals and Oils Association(中国粮油学报), 2021, 36(3): 140.
[11] TAO Zhi-yong, YU Zi-jia, LIN Sen(陶志勇, 于子佳, 林 森). Journal of Electronic Measurement and Instrumentation(电子测量与仪器学报), 2021, 35(1): 18.
[12] LI Ya-li, WANG Shu-qin, CHEN Qian-ru, et al(李雅丽, 王淑琴, 陈倩茹, 等). Computer Engineering and Applications(计算机工程与应用), 2020, 56(22): 1.
[13] LÜ Xin, MU Xiao-dong, ZHANG Jun(吕 鑫, 慕晓冬, 张 钧). Systems Engineering and Electronics(系统工程与电子技术), 2021, 43(2): 318.
[14] ZHOU Meng-ran, YU Dao-yang, HU Feng, et al(周孟然, 余道洋, 胡 锋, 等). Journal of Henan Normal University · Natural Science Edition(河南师范大学学报 · 自然科学版), 2021, 49(2): 46.
[15] YUAN Zi-ran, WEI Li-fei, ZHANG Yang-xi, et al(袁自然, 魏立飞, 张杨熙, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(2): 567.
[16] XIAO Hai-jun, LU Chang-jing, HE Fan(肖海军, 卢常景, 何 凡). Journal of South-Central University for Nationalities · Natural Science Edition(中南民族大学学报 · 自然科学版), 2017, 36(3): 90.

Hyperspectral Analysis of Milk Protein Content Using SVM Optimized by Sparrow Search Algorithm

LIU Mei-chen, XUE He-ru*, LIU Jiang-ping, DAI Rong-rong, HU Peng-wei, HUANG Qing, JIANG Xin-hua
College of Computer and Information Engineering, Inner Mongolia Agricultural University, Huhhot 010000, China

Abstract Milk contains many nutritional elements needed by the human body, such as fat, protein, calcium, etc. Therefore, analysing nutritional elements in milk is a key part of milk safety detection. Hyperspectral technology can effectively identify nutritional elements in milk by combining image and spectral data. In order to quickly and accurately predict protein content in milk, the Competitive Adaptive Reweighted Sampling (CARS) algorithm was used to select characteristic wavelengths. A method based on Sparrow Search Algorithm (SSA) to optimize Support Vector Machine (SVM) was proposed to predict milk protein content. The reflectance spectra of milk (400～1 000 nm) extracted by the hyperspectral spectrometer were used for the experiment. During Normalization (N), Standardization and Multiplicative Scatter Correction (MSC), the original milk data are used for spectral noise reduction to improve spectral utilization. The successive projections algorithm (SPA) and the competitive adaptive re-weighting algorithm were used to extract the feature wavelengths from the processed milk spectral data. The correlation coefficients between proteins and the spectrum were calculated and ranked by importance to obtain the important feature wavelengths. In the end, through SVM, the Genetic Algorithm (GA)-SVM, Particle Swarm Optimization (PSO)-SVM and Partial Least-Regression (PLS) algorithm was used to predict milk proteins and compare the prediction results. In order to improve the accuracy of protein prediction and model stability, SSA was proposed to optimize the kernel function G and penalty parameter C of SVM. Root Mean Squared Error (RMSE) was used as the fitness function, and the optimal regression parameters were selected through iteration to train the model. The results of milk data prediction showed that the optimal combination model was MSC-CARS-SSA-SVM. The determination coefficient R^2 of the model test set was 0.999 6, the root means square error RMSE was 0.001 1, and the time was 4.112 1 seconds. The results show that the CARS algorithm can extract the characteristic bands and eliminate redundant information, thus improving the efficiency of the model and simplifying the algorithm's complexity. The SSA algorithm optimizes SVM's parameters and can quickly obtain the global optimal solution by iteratively updating the optimal position. Compared with SVM, GA-SVM, PSO-SVM and PLS, the prediction accuracy and model stability are significantly improved, which meets the accuracy requirements of milk detection, and is a feasible new method for fast detection of milk protein. It provides a theoretical reference for the optimization of spectral models and the improvement of prediction model accuracy.

Keywords Hyperspectral; Milk protein; Competitive adaptive reweighted sampling; Support vector machine; The sparrow search algorithm

* Corresponding author

(Received Apr. 16, 2021; accepted Jul. 16, 2021)