第42卷，第7期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 7，pp2262-2268
2022年7月 Spectroscopy and Spectral Analysis July，2022

基于蚁群-遗传算法的光谱选择方法与应用

黄 清[1]，薛河儒[1*]，刘江平[1*]，刘美辰[1]，胡鹏伟[1]，孙德刚[2]

1. 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
2. 山东华宇工学院信息工程学院，山东 德州 253000

摘 要 脂肪作为牛奶中的重要营养成分，是评价牛奶质量的一项重要指标。高光谱图像技术能够提供几十到数千波长的数据，能够反映牛奶中不同组成成分细微的光谱差异；另一方面，相邻波段之间往往具有很强的相关性，不仅增加了计算量，而且容易造成维数灾难等问题，因此对高光谱数据进行波段选择非常重要。工作中提出了PLS-ACO特征波段选择方法，并与遗传算法结合，组合成了PLS-ACO-GA的特征波段选择新方法。提出的两种方法以蚁群算法为基础，PLS回归模型回归系数的绝对值作为评价波长重要性的主要依据，以此作为蚁群算法的启发式信息，利用蚁群算法进行智能搜索，结合遗传算法，产生更多优秀的特征波段组合，避免PLS-ACO算法得到的只是局部最优解，得到的最优波段组合能够更好的反映牛奶中脂肪成分的信息；通过计算波长贡献率，筛选出最优波段组合，并与遗传算法，CARS算法和基本蚁群算法光谱特征选择方法比较，最后比较不同特征选择方法下的PLS回归模型预测效果。PLS-ACO，PLS-ACO-GA，CARS，GA和ACO分别筛选了牛奶样品光谱中的18，16，40，43和42个特征波段。其中PLS-ACO-GA筛选波段后的PLS预测模型效果最好，预测集 R_p^2 和RMSEP分别为0.997 6和0.062 2，PLS-ACO次之，预测集 R_p^2 和RMSEP分别为0.997 0和0.077 8。PLS-ACO和PLS-ACO-GA不仅减少了特征波段数量，而且提高了模型的精度。对PLS-ACO-GA进行特征波段选择后的数据，建立MLR，RFR和PLS回归预测模型。MLR预测模型的 R_p^2 和RMSEP分别为0.997 6和0.062 3。RFR回归模型 R_p^2 和RMSEP分别为0.999 9和0.003 0，PLS回归模型的 R_p^2 和RMSEP分别为0.997 6和0.062 2。RFR模型在三种回归预测模型中表现最好。研究结果表明PLS-ACO和PLS-ACO-GA这两种方法可以实现光谱数据特征波段选择，高光谱技术可以实现牛奶中脂肪含量的检测，为牛奶脂肪含量检测提供了一种新的、快速无损的方法。

关键词 高光谱；牛奶脂肪；遗传算法；蚁群算法；特征波段；偏最小二乘

中图分类号：TP391 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)07-2262-07

引 言

随着日常生活水平的提高，牛奶已经逐渐成为了人们生活中必不可少的一种营养品，牛奶中含有丰富的乳脂肪，蛋白质等营养物质，其中，乳脂肪是特别优质的一种脂肪，其组成成分是一分子的甘油和三分子的脂肪酸。全脂牛奶的脂肪含量约为3%，低脂牛奶的脂肪含量为0.5%～2%，脱脂牛奶中脂肪含量低于0.5%。不同人群对牛奶脂肪含量有着不同的需求，儿童、青少年因成长需要可以选择全脂牛奶。而一些患有高血脂、高血压等疾病的人群，应当选择脱脂牛奶。牛奶中对脂肪含量有着及其严格的把控标准，对牛奶脂肪含量的精准检测显得愈发重要。随着光谱技术的发展，光谱分析法凭借着其快速，无损等优点，逐渐应用到牛奶品质检测中来[1]。高光谱图像具有极高的光谱分辨率，高光谱数据能够反映牛奶中不同组成成分的光谱信息，然而高光谱图像的众多波段之间，尤其是相邻或相近波段存在严重的信息冗余。如何提取高光谱数据的有效信息，去除冗余信息是决定高光谱数据应用效果的关键要素之一[2]。波段选择是进行光谱数据分析时常用的一种技术手段，与其他方法相比，波段选择只是从全光谱波段中选取与目标信息相关的波段，并不改变光谱的物理信息。偏最小二乘(partialleast squares，PLS)是目前应用最广泛的线性建模方法，Li等提出了竞争性自适应重加权算法(competitive adaptive reweighted sam-

收稿日期：2021-07-29，**修订日期**：2021-10-27

基金项目：国家自然科学基金项目(61461041)资助

作者简介：黄 清，1998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：1905550194@qq.com

*通讯作者 e-mail：xuehr@imau.com；liujiangping@imau.edu.cn

pling，CARS)，以 PLS 回归系数绝对值大小作为衡量指标，采用竞争性自适应加权采样法对光谱数据进行变量筛选[3]，但是 CARS 算法随机性强，稳定性差，需要经过运行多次才可能选出相对较好的特征波段组合。有研究提出将变量投影重要性系数(variable importance in the projection，VIP)加 PLS 回归系数结合蚁群算法(ant colony optimization，ACO)的波长筛选新方法。张小鸣等提出了一种基于变量有效性精英蚁群系统，将 PLS 回归模型系数作为蚁群的初始信息素，在信息素更新过程中引入 VIP 系数[4]。在张小鸣发明的方法中，均以蚁群算法为基础，而在蚁群算法中，以正反馈机制为基础，这两种方法都容易陷入局部最优，且计算 VIP 系数增加了算法的复杂度，针对这些问题，本工作将蚁群算法与 PLS 回归系数相结合(简称为 PLS-ACO)，以 PLS 模型回归系数单独作为蚁群算法的启发式信息，预测集相关系数(correlation coefficient of prediction，R_p^2)和预测均方根误差(root mean square error of cross of prediction，RMSEP)的商作为个体适应度值的评价标准，仅以 PLS 回归模型系数，R_p^2 和 RMSEP 的商作为信息素更新的依据，简化算法，降低算法复杂度，利用蚁群算法进行智能搜索，为了避免陷入局部最优，结合遗传算法，组合成一种新的特征波段选择方法(简称为 PLS-ACO-GA)，利用蚁群算法中不同蚂蚁所寻找的最优路径作为遗传算法的初始种群。通过遗传算法产生更多的个体。从所有产生的个体中筛选出适应度较高的个体作为优秀个体，统计每个波段在优秀个体中出现的次数，以此作为依据筛选出最优的特征组合。在蚁群算法迭代的基础上，遗传算法的初始种群不再是随机生成，而遗传算法，通过选择交叉，产生了更多优秀的个体，为计算波长的贡献率提供了更多的优秀个体范本，加入遗传算法后，不仅解决了蚁群算法容易陷入局部最优，更极大加强了算法的稳定性和通用性。

1　实验部分

1.1　材料

在天猫、京东等平台购买了 6 种不同品牌的牛奶，其中，特仑苏脂肪含量为 4.4 g·(100 mL)$^{-1}$，QQ 星和蒙牛高钙脂肪含量为 3.7 g·(100 mL)$^{-1}$，The land 脂肪含量为 3.3 g·(100 mL)$^{-1}$，伊利甄浓脂肪含量为 4.6 g·(100 mL)$^{-1}$，伊利脱脂脂肪含量为 0 g·(100 mL)$^{-1}$。

1.2　样本反射光谱数据采集

高光谱采集系统由高光谱成像仪、高精度扫描云台、石英卤钨灯照明电源、等部件组成。高光谱测量仪为美国 Headwall 公司 2010 年生产的型号为 1003B—10141 的高光谱成像仪，光谱范围 400～1 000 nm，共 125 个波段，采集到的光谱图像分辨率为 777×1 004 像素。

在实验的过程中，尽量保持牛奶样本与摄像头的距离约 20 cm，为保证不受其他反射光源的影响，在牛奶样本盛器下垫一块黑色绒布。在光谱采集过程中，光谱摄像头中的传感器中会产生暗电流，每次采集到的光谱数据会伴随着一定的噪声，影响高光谱图像的质量，为了避免客观因素影响，在每次采集光谱图像之前，需要对该系统进行黑白校正处理，在一定程度上提高图像的质量[5]。实验中，首先盖上相机盖，点击暗电流，采集全黑的光谱校正图像 I_B；然后取下相机盖，用白板采集全白校正图像 I_W；最后，采集牛奶样本的光谱图像，每个样本平行采集三次，最后根据公式 $I_C=(I_R-I_B)/(I_W-I_B)$对采集到的牛奶光谱图像进行黑白校正。其中，I_C 为校正后的牛奶光谱图像，I_R 为实验过程中采集到的牛奶光谱图像。利用 ENVI 软件从三张图片中选取效果最好的光谱图像，从中选取 40 个光线清晰均匀的感兴趣区域，导出每个区域的光谱反射率数据作为该品牌牛奶的 40 个样

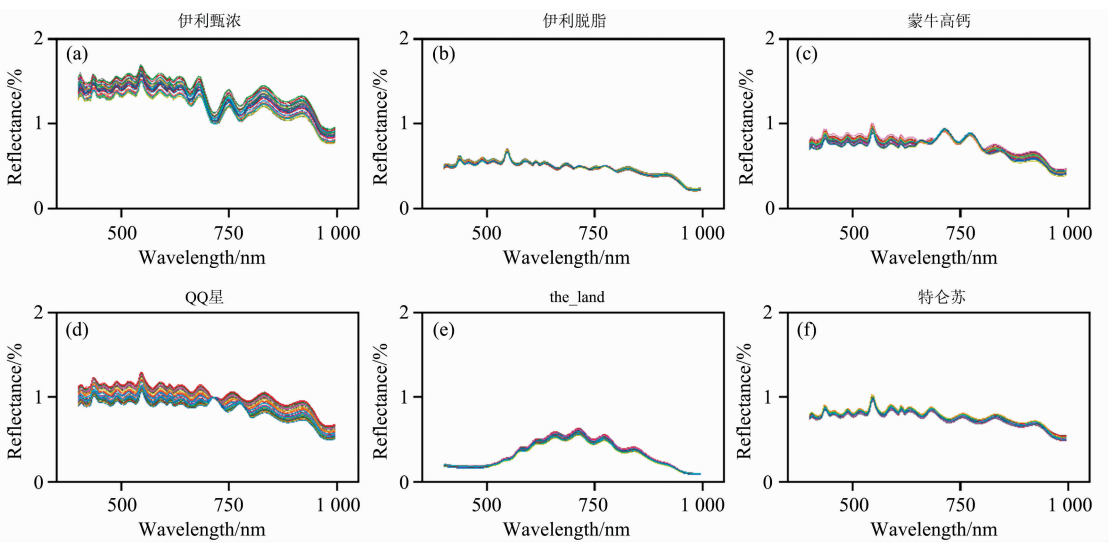

图 1　六种不同品牌样本原始光谱图

(a)：伊利甄浓；(b)：伊利脱脂；(c)：蒙牛高钙；(d)：QQ 星；(e)：The land；(f)：特仑苏

Fig. 1　Original spectra of samples of six different brands

(a)：Yilizhennong；(b)：Yiliskim milk；(c)：Mengniu high calcium milk；(d)：QQ star；(e)：The land；(f)：Telunsu

本进行实验。一共分别采集了 6 种不同品牌的牛奶样品，合计 240 个样本。图 1(a—f)是六个不同品牌牛奶的原始光谱数据反射曲线。受采集环境和仪器设备的影响，虽然原始高光谱图像经过了黑白校正处理，但是数据中仍然存在一些无用的信息和噪声，通过图 1 可以看出，伊利甄浓和 QQ 星的光谱反射图像在采集的过程中受到的干扰较大。为了降低散射光和噪声的干扰，需要对采集到的光谱数据进行数据预处理。

2 结果与讨论

2.1 光谱数据预处理

光谱分析中常用的预处理方法包括导数校正法(其中包括一阶导数（first derivative，1st Der)，二阶导数(second derivative，2st Der)，和多阶导数等)，多元散射校正(multiplicative scatter correction，MSC)，标准正态变换(standard normoal variate transformation，SNV)，卷积平滑(savitzky-golay，S-G)等。其中导数校正法和 S-G 平滑可以有效消除基线平滑和平移引起的噪声，MSC 和 SNV 则可以减少因光的散射等现象造成的噪声。

使用上述不同的方法对原始光谱数据进行数据预处理，对预处理后的数据建立 PLS 回归模型，将校正集相关系数(correlation coefficient of calibration R_c^2)，校正集交叉验证均方根误差(RMSECV)和预测集 R_p^2 和 RMSEP 作为模型评价的指标。实验中得到不同预处理方法下的模型预测结果数据如表 1 所示。通过表 1 实验数据可知，对比原始数据，通过导数校正后，模型精度反而降低了，这是因为整数阶微分变换会忽略渐变的分数阶微分信息，可能造成某些信息丢失，影响有效的信息检测，建模的精度会受到一定制约[6]。通过 S-G 平滑，MSC 和 SNV 这三种方法处理后，模型精度都有一定的提高。其中，经过 MSC 预处理后的数据建模效果最好，因此，将其作为后续进行特征波段选择所需的数据集。

表 1 不同预处理方法的牛奶脂肪含量 PLS 回归模型预测结果

Table 1 The prediction results of the PLS regression model for the milk fat content of different pretreatment methods

预处理方法	R_c^2	R_p^2	RMSECV	RMSEP
Raw data	0.999 1	0.987 3	0.046 6	0.143 7
1st Der	0.998 3	0.973 6	0.064 7	0.207 4
2st Der	0.999 1	0.954 6	0.046 7	0.270 0
S-G	0.998 8	0.988 8	0.053 4	0.135 1
MSC	0.999 9	0.993 4	0.015 5	0.103 8
SNV	0.999 0	0.988 0	0.049 4	0.139 6

2.2 特征波段提取

蚂蚁在寻找食物的过程中，会根据路径的长短，判断该路径的好坏，留下不同浓度的信息，方便自己的同伴判断觅食的路径，随着时间的延长，较优的路径上积累的信息元素比例越来越高，越来越多的蚂蚁选择此路径，慢慢的形成了一种正反馈机制。蚁群算法就是模拟蚂蚁觅食这一行为，来寻找最优特征波段组合。在基本蚁群算法中，初始信息素浓度都是人为设定的(大多数都设定为 1)，初始信息素的匮乏，使得算法收敛时间过长，执行效率低下[7]。受到 CARS 算法的启发，将 PLS 回归模型系数的绝对值作为评价波段好坏的标准，首先在全波段数据下建立 PLS 回归模型，每个波段对应的回归系数绝对值作为信息素的初始值。避免了在第一次迭代中，蚂蚁随机寻找特征波段的缺点，加快了算法的收敛速度。将蚁群算法与 PLS 回归系数相结合，虽然解决了传统蚁群算法收敛速度慢，模型复杂的缺点，但是仍然无法解决蚁群算法由于正反馈机制容易陷入局部最优的缺点，因此将 PLS-ACO 算法与遗传算法相结合，经过遗传算法迭代产生更多优秀的个体，为计算波段的贡献率，提供更多的样本，以提高算法的稳定性和通用性。研究提出了全局阈值(threshol)的概念，传统的遗传算法和蚁群算法，都是选取某一次迭代最好的个体或几个个体来进入贡献矩阵中，但是在前几次的迭代中的最优个体，很有可能不及最后几代里的最差个体。通过全局阈值，对产生的所有个体来进行筛选。大于全局阈值的个体将进入到贡献矩阵中，计算每个波段的贡献率，也就是贡献矩阵里每个波段组合中每个波段出现的次数，依次剔除贡献率最小的波段，对筛选后的波段组合建立 PLS 回归预测模型，选择适应度值最大的波段组合作为最终的最优波段组合。

2.2.1 PLS-ACO

(1) 初始化算法的参数：用全光谱波段数据建立 PLS 回归模型，得到对应波段的回归系数(β)的绝对值作为蚁群的初始信息素矩阵(init-pheromone)。最大迭代次数(max-iters)，蚂蚁个数(max-ants)，每只蚂蚁选取最大特征个数(max-features)，以及信息素启发因子(Q)，信息素挥发因子(ρ)，入选特征路径矩阵(has)，未选路径矩阵(have)，全局阈值，贡献矩阵(contribution)。

(2) 蚂蚁选择路径：在每一次迭代中，每只蚂蚁随机选择一个特征波段作为路径起始点，并将其储存在 has 矩阵，have 矩阵择则删除该特征波段，init-pheromone 矩阵删除对应特征波段的信息素。由 init-pheromone 矩阵，计算每一个特征波段的被选概率，用轮盘赌算法选择下一个特征波段，直到所选特征波段达到 max-features。

(3) 选取最优蚂蚁：对所有蚂蚁选择的特征波段建立 PLS 回归模型，计算出预测集真实值和预测值之间的 R_p^2 和 RMSEP，每只蚂蚁的适应度值(ant-fitness，AF)等于 R_p^2 除以 RMSEP。如果该蚂蚁的适应度值大于 threshol，则将该蚂蚁选择的特征波段组合存入 contribution 矩阵中。并将适应度值最大的蚂蚁，作为此次迭代中的最优蚂蚁。

(4) 信息素更新：仿造生物界蚂蚁觅食行为更新信息素，选取每一次迭代中，适应度值最高的蚂蚁进行信息素的更新。将最优蚂蚁所选择的特征波段对应的回归系数绝对值，以及适应度值作为信息素更新的依据，对应波段的信息素按照信息素更新公式得到加强。没有被选择的特征波段，信息素会因为挥发，浓度慢慢变小。具体信息素更新公式如式(1)所示。

$$\tau_n = (1-\rho)x\tau_{n-1} + \text{dela} \quad (1)$$

$$\begin{cases}\beta_i + Q \times \mathrm{AF} & \text{波段 } i \text{ 被选中} \\ 0 & \text{波段 } i \text{ 未被选中}\end{cases}$$

重复步骤 2—步骤 4，直至达到设置的最大迭代次数，通过设置 threshol，得到最终 contribution 矩阵。剔除 contribution 矩阵中，选取的特征波段完全相同的特征波段组合。计算每个波段的贡献率，依次剔除贡献率最小的波段，对选择的特征波段组合建立 PLS 回归模型，选择适应度值最大的波段组合作为最终所选择的最优特征波段组合。

2.2.2　PLS-ACO-GA

具体步骤：

在 PLS-ACO 的基础下，结合遗传算法[8]，进行如下步骤：(1) 初始化算法的参数：遗传算法交叉概率(pc)，变异概率(pm)。

(2) 初始种群生成：在 PLS-ACO 算法中，将每一次迭代中，每一只蚂蚁所选择的特征波段组合作为遗传算法中的个体，所有蚂蚁选择的特征波段组合构成了遗传算法中的初始种群。

(3) 选择运算：用轮盘赌算法淘汰那些适应度值低的个体。为避免最优个体的遗失，每一次轮盘赌选择过后，如果最优个体被过滤了，则将最优个体代替适应度值最低的个体，重新保留在种群中，以便于通过交叉产生更多优秀的个体。

(4) 交叉运算：随机生成一个 0 到 1 之间的数，如果该数大于 pc，随机选择一个个体作为母代，进行交叉操作。本研究使用单点交叉操作，即随机生成一个交叉点，交换该点前后的特征波段。结果生成两个新的个体，将其建立 PLS 回归模型，计算其适应度值，并与父代母代进行比较，若子代比父代母代更加优秀，则进行替换，对应的父代或者母代淘汰。

(5) 变异运算：本工作使用单点随机变异，随机生成一个 0～1 之间的数，若该数大于变异概率 pm，随机选择一个特征，改变其是否被选择的状态。

(6) 个体选取：如果该个体的适应度值大于全局阈值，则将该个体的路径存入贡献矩阵中。

重复步骤(1)—步骤(6)，直至达到设置的遗传最大迭代次数，将通过全局阈值的特征波段组合存入贡献矩阵中，进入下一次 PLS-ACO 算法迭代中。

PLS-ACO-GA 算法流程如图 2 所示。

2.2.3　特征波段选择方法结果与比较

通过多次反复实验，根据经验将蚁群算法中最大迭代次数设置为 50，蚂蚁个数设置为 80，ρ 设置为 0.4，Q 设置为 0.2，遗传迭代次数为 30，pc 设置为 0.5，pm 设置为 0.01，threshol 最小设置为 9.564 5(全波段 PLS 建模下 R_p^2 除以 RMSEP 所得)，以步长 0.5 递增，当贡献矩阵个体数小于 100 时不再增加。max-features 的大小需要考虑到特征波段总数的大小，如果设定过小，算法因随机性会降低模型的精度，设定过大，一些无关的特征波段会伴随着好的波段点混入，不仅会影响模型精度，也大大增加了计算复杂度和时间。最后将 max-features 分别设置为 10，20，30。经过实验发现，当 max-features 大于 30 时，对模型的精度不会有任何改善，且筛选出的特征波段过多。统计不同阈值筛选后的个体中每个波长出现的次数，除以入选贡献矩阵个体总数，计算对应每个波段的波段贡献率，以此作为波段重要性的依据来筛选最优波段组合。图 3 和图 4 分别是 PLS-ACO 和 PLS-ACO-GA 将 max-features 设置为 20 时的波段贡献率，图中的六条曲线分别是六种不同品牌的牛奶其中一个样本的光谱反射率曲线图。从图中可以看出，频率较高的波段一般处于某个波峰或者波谷处，所选波段能代替全波段反映牛奶脂肪含量的光谱信息。

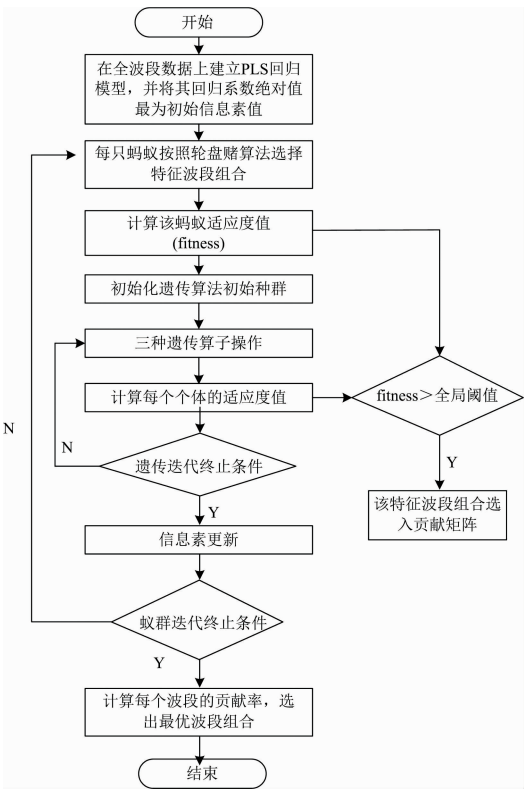

图 2　PLS-ACO-GA 算法流程图

Fig. 2　PLS-ACO-GA algorithm flowchart

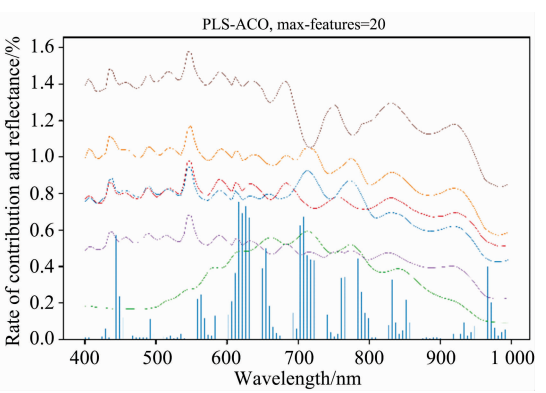

图 3　PLS-ACO 波段贡献率

Fig. 3　PLS-ACO band contributionrate

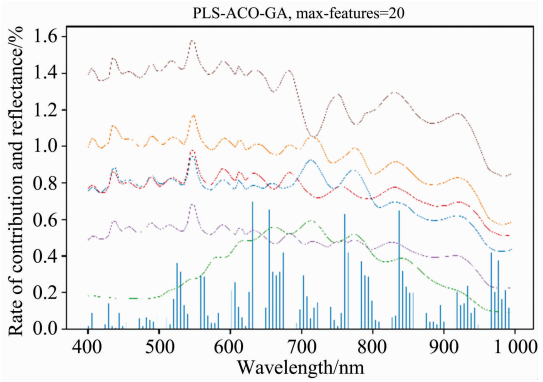

图 4 PLS-ACO-GA 波段贡献率

Fig. 4 PLS-ACO-GA band contributionrate

表 2 和表 3 分别是指设置不同 max-features 时，PLS-ACO 和 PLS-ACO-GA 所选特征波段组合建立 PLS 回归模型时得到的结果。

表 2 PLS-ACO，max-features 分别设置为 10，20，30 的 PLS 回归模型结果

Table 2 PLS regression model results with PLS-ACO and max-features set to 10，20 and 30 respectively

max-features	波长数	R_p^2	RMSEP	主成分数
Full	125	0.993 4	0.103 8	24
10	21	0.996 2	0.077 8	10
20	18	0.997 0	0.069 2	16
30	41	0.997 8	0.058 5	12

表 3 PLS-ACO-GA，max-features 分别设置为 10，20，30 的 PLS 回归模型结果

Table 3 PLS regression model results with PLS-ACO-GA and max-features set to 10，20 and 30 respectively

max-features	波长数	R_p^2	RMSEP	主成分数
Full	125	0.934 4	0.103 8	24
10	16	0.996 8	0.075 2	8
20	16	0.997 6	0.062 2	14
30	32	0.998 0	0.056 5	12

从实验结果可知，当 max-features 大小设置相同时，PLS-ACO 和 PLS-ACO-GA 相比全波段 PLS 回归建模结果，在提高了模型精度的同时，只需要更少的特征波段。PLS-ACO-GA 方法相比 PLS-ACO 具有有更加优异的表现，分析认为加入遗传算法后，种群将变得更加的丰富，为贡献矩阵提供了更多优秀的特征组合，最后按照波段贡献率来选择最终的特征波段组合，很多优秀随机特征波段组合共同构成了一个稳定的特征波段组合。加入遗传算法后，在一定的程度上避免了蚁群算法陷入局部最优；另一方面，选出的波段组合也更加具有代表性。从表 3 数据可知，在 PLS-ACO 中，虽然 R_p^2 没有显著提高，但 RMSEP 却明显降低了。max-features 分别为 10，20 和 30 时，RMSEP 分别降低了 25%，33%和 43%，在 PLS-ACO-GA 中，R_p^2 也没有得到显著的提高，但 RMSEP 得到了更显著的降低，分别降低了 27%，40%和 45%，并且所需要的特征波段数目更少。随着 max-features 的提高，模型精度也得到了相应的提高，但是当 max-features 等于 30 时，模型精度虽然得到了略微的提高，但是需要的特征数目却提高了两倍左右。最后选定 max-features 值为 20 时得到特征波段组合为 PLS-ACO 和 PLS-ACO-GA 的最优结果。

为了证明 PLS-ACO 和 PLS-ACO-GA 算法的优越性，用 CARS 算法，遗传算法[6]和传统蚁群算法对牛奶光谱数据进行特征波段选择来进行对比，对最终选择的波段组合建立 PLS 回归模型，以 R_p^2，RMSEP，选取特征数量大小作为特征选择方法好坏的评价标准。表 4 是建立 PLS 回归模型后的结果，从实验结果可以看出，PLS-ACO，和 PLS-ACO-GA 这两种算法，无论是模型精度，还是在选择更少的波段数量上，都有着很明显的优势。

表 4 不同特征选择方法下的 PLS 回归模型结果

Table 4 PLS regression model result sunder different feature selection methods

方法	波长数	R_p^2	RMSEP	主成分数
Full	125	0.934 4	0.103 8	24
CARS	40	0.995 1	0.088 7	12
GA	43	0.997 1	0.068 5	14
ACO	22	0.996 2	0.072 8	12
PLS-ACO	18	0.997 0	0.077 8	16
PLS-ACO-GA	16	0.997 6	0.062 2	14

在使用 PLS-ACO 和 PLS-ACO-GA 算法选择特征波段组合的过程中，一直选用 PLS 回归模型的精度作为评价特征波段组合的好坏标准，很容易陷入此特征波段选择方法只针对 PLS 回归模型有效果的情况。用多元线性回归(multiple linear regression，MLR)模型和随机森林回归(random forest regression，RFR)模型再次对 PLS-ACO 和 PLS-ACO-GA 这两种算法的适用性进行验证。PLS-ACO 和 PLS-ACO-GA 算法选取特征波段组合的数据进行 MLR 和 RFR 模型结果如表 5 和表 6 所示。实验数据表明，PLS-ACO 和 PLS-ACO-GA 算法在其他模型上也取得了不错的效果。在 MLR 模型中，对比全波段光谱数据，虽然 R_p^2 没有明显改变，但是 RMSEP 明显降低了，其中，PLS-ACO 降低了 15%，PLS-ACO-GA 降低了 47%。在 RFR 模型中，R_p^2 虽然也没有明显改变，但 RMSEP 的改善更加明显，其中，PLS-ACO 降低了 41%，PLS-ACO-GA 降低了 60%。通过对比可以发现，在 MLR、

表 5 PLS-ACO 和 PLS-ACO-GA 多元线性回归模型结果

Table 5 PLS-ACO and PLS-ACO-GA MLR results

方法	波长数	R_p^2	RMSEP
Full-MLR	125	0.991 4	0.118 0
PLS-ACO-MLR	18	0.997 0	0.069 2
PLS-ACO-GA-MLR	16	0.997 6	0.062 3

RFR 和 PLS 回归模型中，RFR 模型表现最佳，R_p^2 高达 0.999 994，RMSEP 仅为 0.003 078。

表 6　PLS-ACO 和 PLS-ACO-GA 随机森林回归模型结果

Table 6　PLS-ACO and PLS-ACO-GA RFR results

方法	波长数	R_p^2	RMSEP
Full-RFR	125	0.999 963	0.007 6
PLS-ACO-RFR	18	0.999 974	0.006 4
PLS-ACO-GA-RFR	16	0.999 994	0.003 0

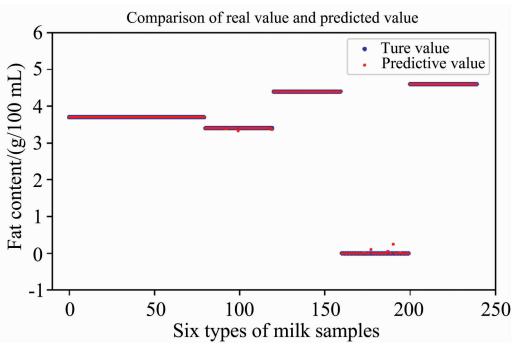

图 5　MSC-(PLS-ACO-GA)-RF 预测结果

Fig. 5　MSC-(PLS-ACO-GA)-RF prediction results

通过对比不同的预处理方法，特征波段选择方法，和不同的回归模型，选择 MSC 作为预处理的方法，用 PLS-ACO-GA 来进行特征波段选择，最后建立 RFR 模型得到的牛奶脂肪含量预测最为精准。将通过 MSC-(PLS-ACO-GA)-RFR 建模后得到的脂肪含量预测值与真实值做图对比，以便更加直观的感觉到预测效果，结果如图 5 所示。

从图 5 中可以看出，只有个别样本真实值和预测值偏差较大，大多数样本真实值和预测值基本重合，预测结果表现优异。

3　结　论

利用高光谱技术进行牛奶脂肪含量检测研究，使用不同预处理方法处理原始光谱数据，结果 MSC 取得了最优异的结果；进行特征波段选择，提出了 PLS-ACO 和 PLS-ACO-GA 这两种方法，通过与 CARS 算法，遗传算法和基本蚁群算法进行对比，在选取数目更少的基础上，进一步提高了模型的精度，最后比较不同预测模型方法，在 MLR，RFR 和 PLS 回归模型中，RFR 预测模型取得了最好的效果，R_p^2 为 0.999 994，RMSEP 为 0.003 078。研究表明，PLS-ACO 和 PLS-ACO-GA 这两种方法可以实现光谱数据特征波段选择，高光谱技术可以实现牛奶脂肪含量的精准预测，可为实现快速无损牛奶脂肪含量预测提供理论基础。

References

[1] Wu Chuyan, Song Tangshi, Xiang Mashi, et al. Optics Express, 2018, 26(8): 10119.

[2] LIANG Kun, DU Ying-ying, LU Wei, et al(梁　琨, 杜莹莹, 卢　伟, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2016, 2: 41.

[3] Li Hongdong, Liang Yizeng, Xu Qingsong. Analytica Chimica Acta, 2009, 648(1): 77.

[4] ZHANG Xiao-ming, MAO Zhi-kang, LI Shao-wen, et al(张小鸣, 冒智康, 李绍稳, 等). Jiangsu Agricultural Sciences(江苏农业科学), 2019, 47(19): 227.

[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] Wang Jingzhe, Ding Jianli, Abulimiti Aerzuna, et al. Peer J., 2018.

[7] WANG Si-han, WAN You-chuan, WANG Ming-wei, et al(王偲晗, 万幼川, 王明威, 等). Computer Engineering and Applications(计算机工程与应用), 2018, 54(1): 196.

[8] LIU Xin, MAO Zhi-kang, ZHANG Xiao-ming, et al(刘　鑫, 冒智康, 张小鸣, 等). Journal of Jiangsu University(江苏大学学报), 2020, 3: 12.

Spectral Selection Method Based on Ant Colony-Genetic Algorithm

HUANG Qing[1], XUE He-ru[1*], LIU Jiang-ping[1*], LIU Mei-chen[1], HU Peng-wei[1], SUN De-gang[2]
1. College of Computer and Information Engineering, Inner Mongolia Agricultural University, Huhhot 010000, China
2. College of Information Engineering, Shangdong HuaYu University of Technology, Dezhou 253000, China

Abstract As an important nutritional component in milk, fat is an important index to evaluate milk quality. Hyperspectral image technology can provide tens to thousands of bands of data and can reflect the subtle spectral differences of different components in milk. On the other hand, there is often a strong correlation between adjacent bands, which increases the amount of calculation and easily causes problems such as dimension disaster. Therefore, it is very important to select bands for hyperspectral data. This paper proposes a PLS-ACO feature band selection method combined with a genetic algorithm to form a new feature band selection method of PLS-ACO-GA. The two methods proposed in this paper are based on ant colony optimization. The absolute value of the regression coefficient of the PLS regression model is the main basis for evaluating the importance of wavelength, which is used as the heuristic information of ant colony optimization. Ant colony optimization is used for intelligent search, combined with genetic algorithm to produce more excellent characteristic band combinations. To avoid that pls-aco algorithm only obtains the optimal local solution. The optimal band combination can better reflect the information of fat composition in milk. By calculating the wavelength contribution rate, the optimal band combination is selected and compared with the spectral feature selection methods of genetic algorithm, cars algorithm and basic ant colony optimization. Finally, the prediction effects of the PLS regression model under different feature selection methods are compared. PLS-ACO, PLS-ACO-GA, CARS, GA and ACO screened 18, 16, 40, 43 and 42 characteristic bands in the spectrum of milk samples, respectively. The PLS prediction model after the PLS-GA-ACO screening band has the best effect. The prediction sets R_P^2 and RMSEP are 0.997 6 and 0.062 2 respectively, followed by PLS-ACO, and the prediction sets R_P^2 and RMSEP are 0.997 0 and 0.077 8 respectively. PLS-ACO and PLS-ACO-GA reduce the number of characteristic bands and improve the accuracy of the model. MLR, RFR and PLS regression prediction models are established based on PLS-ACO-GA data after characteristic band selection. The R_P^2 and RMSEP of the MLR prediction model are 0.997 6 and 0.062 3 respectively. R_P^2 and RMSEP of the RFR regression model were 0.999 9 and 0.003 0 respectively, and R_P^2 and RMSEP of the PLS regression model were 0.997 6 and 0.062 2 respectively. RFR model performs best among the three regression prediction models. The results show that hyperspectral technology can detect the fat content in milk, which provides a new, rapid and non-destructive method for the detection of fat content in milk.

Keywords Hyperspectral; Milk fat; Genetic algorithm; Ant colony algorithm; Characteritic band; Partial least squares

(Received Jul. 29, 2021; accepted Oct. 27, 2021)

* Corresponding authors