第42卷，第4期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 4，pp1243-1249
2022年4月 Spectroscopy and Spectral Analysis April，2022

傅里叶变换中红外光谱的牛奶品质无损检测分级

肖仕杰[1]，王巧华[1,2*]，李春芳[3,4]，杜 超[3]，周增坡[4]，梁生超[4]，张淑君[3*]

1. 华中农业大学工学院，湖北 武汉 430070
2. 农业部长江中下游农业装备重点实验室，湖北 武汉 430070
3. 华中农业大学动物遗传育种与繁殖教育部实验室，湖北 武汉 430070
4. 河北省畜牧业协会，河北 石家庄 050000

摘 要 市场上普遍存在“高蛋白”，“高乳脂”等特色牛奶。为了实现对特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的无损快速分级，收集了河北省10个牧场不同月份(1月、3月—10月)的5 121份牛奶样本并采集中红外光谱数据，分别测定牛奶中的乳蛋白、乳脂和体细胞数，构建了牛奶品质分级模型。首先，分析牛奶光谱并去除冗余波段，最终选择925～1 597和1 712～3 024 cm^{-1}的敏感波段组合作为全光谱用于建立模型。为了提高模型的性能，采用标准正态变量变换(SNV)，多元散射校正(MSC)，一阶导数，二阶导数，一阶差分和二阶差分6种算法对光谱进行预处理并建立朴素贝叶斯模型(NB)和随机森林模型(RF)，确定二阶差分为最佳预处理方法，其测试集准确率分别为92.11%和96.87%。为了简化模型，利用无信息变量消除法(UVE)、竞争性自适应重加权算法(CARS)与稳定性竞争性自适应重加权采样算法(SCARS)以及UVE-CARS算法和UVE-SCARS算法对二阶差分后的光谱数据提取特征变量。然后，分别基于全光谱和所选特征变量数据，建立NB模型和RF模型。结果表明，SCARS算法为NB模型的最佳特征提取算法，模型的训练集准确率与测试集准确率分别为94.45%，93.94%；UVE-SCARS算法为RF模型的最佳特征提取算法，模型的训练集准确率与测试集准确率分别为99.86%，96.48%。综上，基于傅里叶变换中红外光谱技术建立的二阶差分-UVE-SCARS-RF模型，可以实现特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的无损快速分级，通过建立中红外光谱模型，首次将乳蛋白、乳脂含量和体细胞数直接结合进行分级鉴定，这是以往未曾有过的。模型应用方便，只需将获得的牛奶红外光谱数据输入模型即可输出预测类别，在牛奶产业中具有实际应用价值。

关键词 中红外光谱；牛奶；品质分级；无损检测；特征变量
中图分类号：S24 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)04-1243-07

引 言

牛奶富含蛋白质和脂肪。乳蛋白中含多种人体必需的氨基酸。乳脂能够提供能量和营养。牛奶的品质决定牛奶的口感[1]和价格[2]，直接关系到乳企的利润和发展。相关数据表明，2014年—2019年，我国每年原料奶产量均在3 000万吨以上[3]。随着生活水平的提高，消费者更加注重牛奶品质，因此市场上普遍出现“高蛋白”，“高乳脂”等特色牛奶。此外，研究表明，牛奶中体细胞数的变化会直接影响乳蛋白和乳脂的含量[4]。乳企在收购原料奶时会将其作为评价指标。

乳蛋白和乳脂含量，体细胞数的测定需要分开进行，使用不同的方法和仪器。传统的化学分析方法技术成熟、准确率高，但是耗时长且污染环境。若能找到一种方法同时对乳蛋白、乳脂含量和体细胞数直接进行检测并快速分级，将大大提高乳企的生产效率，节约生产成本。利用中红外光谱法检测牛奶操作简单且快速无损，在国外被用于牛奶成分(如蛋白成分和脂肪酸)[5-7]的含量预测和奶牛营养、健康与生殖状况监控[8]。在国内，中红外光谱在牛奶方面主要用于三聚氰胺和尿素等的掺假研究[9-10]。牛奶体细胞的无损研究方面，崔传金等和吴海云等[11-12]利用电参数和化学计量学方法进行了含量预测和分类研究。但是，关于牛奶体细胞的光谱

收稿日期：2021-04-08，**修订日期**：2021-05-30
基金项目：欧盟FP7构架项目(FP7-KBBE-2013-7-613689)资助
作者简介：肖仕杰，1993年生，华中农业大学工学院硕士研究生 e-mail：1175760869@qq.com
*通讯作者 e-mail：wqh@mail.hzau.edu.cn；sjxiaozhang@mail.hzau.edu.cn

无损检测鲜有报道。

利用傅里叶变换中红外光谱针对乳蛋白、乳脂和体细胞对牛奶进行分级研究。通过对特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的光谱差异进行分析，利用无信息变量消除法(uninformative variable elimination，UVE)、竞争性自适应重加权算法(competitive adaptive reweighed sampling，CARS)与稳定性竞争性自适应重加权采样算法(stability competitive adaptive reweighted sampling，SCARS)筛选出能代表 4 种牛奶品质差异的特征变量，并基于朴素贝叶斯(NB)和随机森林(RF)模型构建了牛奶检测分级模型。

1 实验部分

1.1 样品

牛奶于 2020 年 1 月、3 月至 10 月期间从河北省 10 个牧场获得，所有奶牛品种均为中国荷斯坦牛。牛奶采集利用全自动挤奶设备，每份牛奶采集 40 mL，分装到河北省奶牛生产性能测定(DHI)中心配置的全新专用取样瓶里并依次编号，为防止牛奶腐败变质，每个采样瓶里加入专用防腐剂布罗波尔 3.2～3.4 μL 并使其与牛奶充分混匀，及时放入专用冰箱冷藏保存。

1.2 仪器与设备

试验仪器与设备主要包括乳成分分析仪 MilkoScanTM FT＋(傅里叶变换中红外光谱仪，丹麦 FOSS 公司)；体细胞检测仪 FossomaticTM7(丹麦 FOSS 公司)，电热恒温水浴锅。

1.3 方法

1.3.1 光谱采集、乳蛋白和乳脂含量及体细胞数检测

将牛奶分批放入 42 ℃电热恒温水浴锅内预热 15～20 min 后摇晃均匀，使用 MilkoScanTM FT＋进行光谱采集以及蛋白质和脂肪的含量测定。FossomaticTM7 可视为自动荧光显微镜，物镜位于转盘上方，连续的牛奶液膜涂布在转盘周边，暴露在紫外光下，经吖啶橙染色的牛奶细胞荧光信号由光电倍增管检测并馈入放大系统，测得的脉冲被计数，每个脉冲等于 1 000 个细胞·mL^{-1}。

根据欧盟标准，脂肪的正常含量范围为 1.5%～9%，蛋白质的正常含量范围为 1%～7%，共筛选出5 121份牛奶。各牧场的样本分布如表 1 所示。

1.3.2 分级标准

参考 GB19301—2010《食品安全国家标准生乳》和 TTDSTIA001—2019《生乳用途分级技术规范》对牛奶进行分级，分级标准如表 2 所示。

1.4 数据处理

1.4.1 光谱预处理方法

牛奶本身作为胶体，当光束穿过时，会产生丁达尔效应，即光的散射，仪器在运行过程中也会产生随机噪声，基线漂移等，对中红外光谱产生影响[8]。本文采用 6 种算法对光谱进行预处理，包括标准正态变量变换(standard normal variable，SNV)，多元散射校正(multiplicative scatter correction，MSC)，一阶导数，二阶导数，一阶差分和二阶差分。

表 1 各牧场的样本分布统计

Table 1 Sample distribution statistics of each pasture

牧场标记	样本数量			
	特优优质奶	高蛋白特色奶	高乳脂特色奶	普通奶
1	118	183	50	119
2	102	56	151	92
3	144	106	338	249
4	40	22	112	44
5	258	132	167	143
6	187	131	130	141
7	63	66	41	62
8	169	149	81	147
9	139	221	57	176
10	122	89	165	159

表 2 分级标准

Table 2 Standard of classification

牛奶标记	级别	脂肪/%	蛋白质/%	体细胞/(10^4 个·mL^{-1})
A	特优优质奶	≥4 且≤9	≥3.7 且≤7	≤20
B	高蛋白特色奶	≥1.5 且≤3.4	≥3.7 且≤7	≤50
C	高乳脂特色奶	≥4 且≤9	≥1 且≤3.1	≤50
D	普通奶	3.1～4	2.8～3.4	≤100

1.4.2 特征变量选择

牛奶的原始光谱共有 1060 个波长，波长不同包含的信息不同，对模型的贡献率大小也不同，部分无信息变量对牛奶分级的中红外判别分析没有价值，反而容易降低模型的预测精度。UVE，CARS 和 SCARS 均以降低无信息变量为出发点，提取出能够代表 4 种牛奶差异的特征变量组合。

1.4.3 模型建立与性能评估

朴素贝叶斯(NB)[13]是一种以概率统计中的贝叶斯定理为理论基础的学习算法。已知先验概率，并计算给定的待分级牛奶属于特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的条件概率，再计算后验概率，选择后验概率最高的类别作为牛奶的预测类别。

随机森林(RF)[14]的本质是一个多决策树(随机方法形成)的分类器。当测试集中 4 种牛奶样本进入分类器时，实际上是由每棵决策树进行分类，选择分类结果最多的类别作为最终结果。

利用准确率作为模型的评价指标。训练集准确率与测试集准确率越高并且两者越接近，表明模型的精度高，可靠性好。

全部数据处理均在 MATLAB 2014b 中进行。

2 结果与讨论

2.1 光谱分析

在中红外范围内对牛奶样品的采集区域为 925～4 000 cm^{-1}，由于 3 680～4 000 cm^{-1} 区域对模型贡献率较低，因此，选择 925～3 680 cm^{-1} 的光谱进行分析。图 1 所示为特优

优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的平均光谱，从图中可以看出，特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的平均光谱吸收曲线紧密重合，每条曲线的变化趋势相似，表明特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的成分大致相同，但同时它们的光谱吸光度也存在差异，这表明 4 种牛奶的化学成分含量存在差异，这就为我们建立牛奶品质分级模型提供了理论依据。

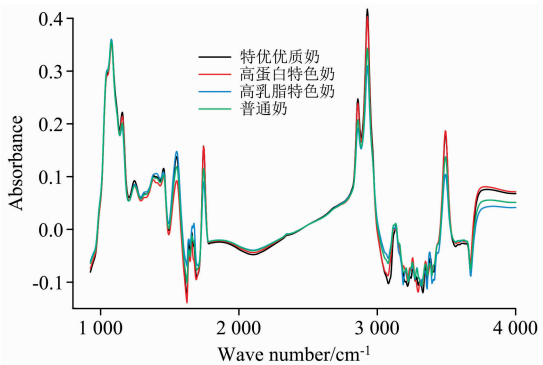

图 1　特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的平均光谱

Fig. 1　Mean spectra of special quality milk, high protein characteristic milk, high fat characteristic milk and ordinary milk

光谱中 1 250，1 550 和 1 650 cm^{-1} 峰与蛋白质的吸收有关[15]，1 774，2 856 和 2 928 cm^{-1} 峰与脂肪的吸收有关[10]。此外，水的吸收峰与牛奶相似，会对牛奶的中红外吸收造成干扰。1 597～1 712 和 3 024～3 680 cm^{-1} 区域由于水的吸收导致很低的信噪比[15-17]，建模前先将这些区域去除。最终取 925～1 597 和 1 712～3 024 cm^{-1} 的敏感波段组合用于后续模型的建立。

2.2　样本集划分

样本总数为 5 121，其中 A 级牛奶的样本数量为 1 342，B 级牛奶的样本数量为 1 155，C 级牛奶的样本数量为 1 292，D 级牛奶的样本数量为 1 332，利用随机法 RS 按照 7∶3 的原则划分样本集。划分后的训练集样本数量为 3 587，其中，A 级牛奶的样本数量为 940，B 级牛奶的样本数量为 809，C 级牛奶的样本数量为 905，D 级牛奶的样本数量为 933；测试集中样本总数为 1 534，其中，A 级牛奶的样本数量为 402，B 级牛奶的样本数量为 346，C 级牛奶的样本数量为 387，D 级牛奶的样本数量为 399。

2.3　光谱预处理

基于全光谱和预处理后的光谱数据，分别建立 NB 和 RF 模型，比较不同预处理对模型精度的影响，结果如表 3。对于 NB 模型，全光谱模型的训练集准确率与测试集准确率仅为 84.50％和 84.22％，与全光谱相比，所有预处理后的光谱数据建立的 NB 模型的训练集准确率与测试集准确率都有明显提升。其中，二阶差分处理后的光谱建立的 NB 模型精度最佳，训练集准确率与测试集准确率为 94.31％和 92.11％。对于 RF 模型，SNV 和 MSC 的模型准确率低于全光谱模型，其余 4 种预处理方法建立的 RF 模型准确率得到提高。二阶差分预处理后的光谱数据建立的 RF 模型精度最佳，训练集准确率和测试集准确率为 99.86％和 96.87％。因此，无论是 NB 模型还是 RF 模型，均选择二阶差分预处理作为最佳的预处理方法，并用于后续的建模分析。

表 3　采用不同预处理方法的全光谱预测模型

Table 3　Full spectrum prediction model using different pre-processing methods

模型	预处理	训练集准确率/％					测试集准确率/％				
		Total	A	B	C	D	Total	A	B	C	D
NB	全光谱	84.50	87.77	81.09	85.86	82.85	84.22	89.80	82.66	83.46	80.70
	SNV	90.72	89.47	94.56	92.71	86.71	88.46	87.06	92.20	89.66	85.46
	MSC	90.97	89.89	94.93	92.27	87.35	88.98	89.05	92.20	89.92	85.21
	一阶导数	90.69	92.34	88.63	91.27	90.25	88.85	94.03	84.68	89.15	86.97
	二阶导数	93.76	95.74	94.68	93.26	91.43	92.05	96.77	93.06	90.70	87.72
	一阶差分	90.38	92.34	87.64	90.94	90.25	88.85	93.78	84.39	89.41	87.22
	二阶差分	94.31	96.28	95.92	93.59	91.64	92.11	96.52	94.51	90.44	87.22
RF	全光谱	99.86	100	100	99.56	99.89	95.83	97.26	96.53	95.09	94.49
	SNV	99.86	100	100	99.56	99.89	94.52	97.01	95.66	95.09	90.48
	MSC	99.86	100	100	99.56	99.89	94.33	97.01	95.66	95.09	89.72
	一阶导数	99.86	100	100	99.67	99.79	96.15	98.01	97.69	96.12	92.98
	二阶导数	99.86	100	100	99.56	99.89	96.41	98.01	97.69	96.12	93.99
	一阶差分	99.86	100	100	99.56	99.89	96.09	98.01	97.69	95.87	92.98
	二阶差分	99.86	100	100	99.56	99.89	96.87	98.76	97.69	96.90	94.24

2.4　特征变量提取

2.4.1　UVE 算法提取特征变量

UVE 算法[18]的变量选择过程如图 2 所示，将阈值参数设为 0.9，主成分数取 20，建立 PLS 模型选择变量。图中左侧曲线为牛奶的光谱变量矩阵，右侧为添加的与牛奶光谱变量数相同的随机噪声矩阵，两条水平虚线处的值分别为＋95.57 和－95.57，代表随机噪声的最大阈值，两线之间为被剔除的无用变量，水平线之外则为建模的牛奶特征变量。

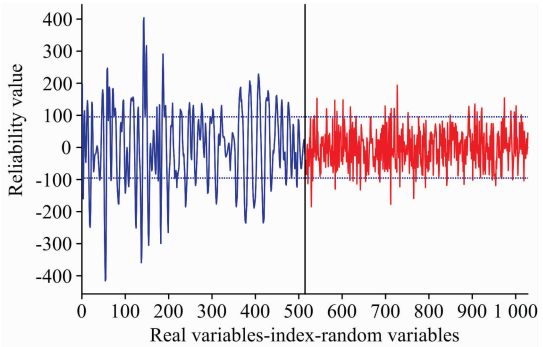

图 2 UVE 消除算法筛选特征波长

Fig. 2 Screening characteristic wavelengths by UVE

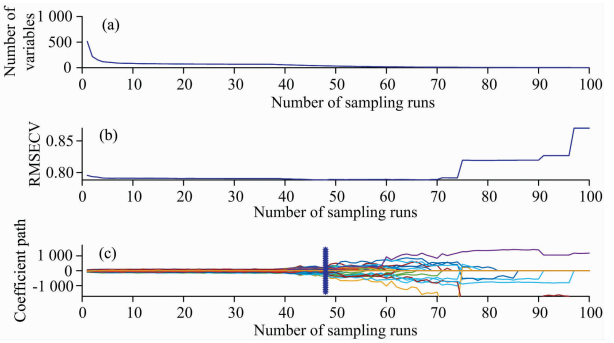

图 3 (a)采样变量数；(b)RMSECV；(c)回归系数路径

Fig. 3 (a) Number of sampling variables; (b) RMSECV; (c) Regression coefficient path

2.4.2 CARS 与 SCARS 算法提取特征变量

CARS 算法基于“优胜劣汰”准则剔除不适应的波长变量。SCARS 算法延续了 CARS 的提取过程[19]。由于两者的变量选择过程相似，仅以 CARS 为例对变量提取的过程进行分析。如图 3 所示，将采样次数设为 100，利用 5 折交叉验证，重采样率为 0.8。图 3(a)表明，迭代次数增加的过程，被选取的特征变量数量在逐步减少。此过程又可分为两个阶段，第一个阶段特征变量数呈指数衰减趋势，称为“粗选阶段”，第二个阶段特征变量数缓慢减少并趋于稳定，为“精选阶段”。图 3(b)为 RMSECV 的变化趋势。当采样次数小于 48，RMSECV 变化不明显，大于 48 时，RMSECV 缓慢增加，表明特征变量中可能包含了无用信息。图 3(c)中的竖线处对应迭代 48 次，可以取得最佳变量组合。

2.5 模型建立与比较

分别以 UVE，CARS 和 SCARS 提取的变量组合为自变量，以牛奶级别 A，B，C，D（在模型中分别记作 0，1，2，3）作为因变量建立 NB 模型和 RF 模型，结果如表 4。

对比 NB 模型可知，全光谱 NB 模型训练集准确率与测试集准确率分别为 94.31%，92.11%，预测性能较好。UVE，CARS 和 SCARS 提取特征变量后建立的模型均优于全光谱模型，表明 UVE，CARS 和 SCARS 算法适用于牛奶的品质分级，可以简化模型，提高模型精度。SCARS-NB 模型的精度优于 CARS-NB 模型和 UVE-NB 模型，训练集准确率与测试集准确率为 94.45%，93.94%。CARS，SCARS 提取的变量较少，为 37，20，仅占全光谱变量的 7.2%，3.9%。UVE 提取的变量数高达 229 个，占比达到 44.6%，变量数远大于 CARS，SCARS，导致模型运行速度慢，因此在 UVE 的基础上利用 CARS，SCARS 进行二次变量提取。UVE-CARS 和 UVE-SCARS 提取的变量数分别为 30 和 37，仅占 UVE 变量数的 13.1% 和 20.5%，变量数大大减少。从 UVE-CARS-NB 与 UVE-SCARS-NB 的预测结果来看，两种二次特征变量结合方法均对 UVE-NB 进行了优化，且 UVE-SCARS-NB 要优于 UVE-CARS-NB，训练集准确率与测试集准确率为 94.68%，93.61%。综合考虑，选择 SCARS-NB 模型作为牛奶品质分级的最优 NB 模型。

表 4 NB 模型和 RF 模型的预测结果

Table 4 Prediction results by NB and RF models

模型	筛选方法	变量数	训练集准确率/%					测试集准确率/%				
			Total	A	B	C	D	Total	A	B	C	D
NB	全光谱	514	94.31	96.28	95.92	93.59	91.64	92.11	96.52	94.51	90.44	87.22
	UVE	229	94.17	96.06	95.92	93.26	91.64	92.70	97.76	92.49	90.96	89.47
	CARS	37	93.73	95.21	95.43	94.14	90.35	92.50	96.77	93.06	92.25	87.97
	SCARS	20	94.45	95.96	95.30	94.92	91.75	93.94	97.26	93.93	93.02	91.48
	UVE-CARS	30	93.95	95.11	95.30	93.59	91.96	93.42	97.01	93.35	92.51	90.73
	UVE-SCARS	47	94.68	95.96	95.92	94.59	92.39	93.61	97.26	93.35	93.28	90.48
RF	全光谱	514	99.86	100	100	99.56	99.89	96.87	98.76	97.69	96.90	94.24
	UVE	229	99.86	100	100	99.56	99.89	96.74	98.51	97.69	96.38	94.49
	CARS	37	99.86	100	100	99.56	99.89	95.76	97.76	95.95	95.87	93.48
	SCARS	20	99.86	100	100	99.56	99.89	95.57	98.51	95.95	94.83	92.98
	UVE-CARS	30	99.86	100	100	99.67	99.79	95.83	97.26	97.40	96.12	92.73
	UVE-SCARS	47	99.86	100	100	99.56	99.89	96.48	98.26	97.40	95.87	94.49

对比 RF 模型可知，全光谱 RF 测试集准确率为 96.87%，模型的预测性能良好。UVE，CARS 和 SCARS 提取特征变量后建立的模型精度较全光谱模型均有不同程度的下降，但模型的测试集准确率均大于 95.5%，表明基于特征

变量的 RF 模型还是可行的，具有良好的精度。其中 UVE-RF 的精度优于 CARS-RF 和 SCARS-RF，测试集准确率为 96.74%，与全光谱 RF 接近。同样将 UVE 分别与 CARS 和 SCARS 相结合，进行二次特征变量提取并建立 RF 模型，但两种结合方法的模型精度较 UVE-RF 模型有所下降，这可能是因为 CARS 和 SCARS 在进一步剔除无用信息的同时将部分有用信息也剔除了。其中，UVE-SCARS-RF 的测试集准确率为 96.48%，与全光谱 RF 较接近。

进一步对比全光谱 RF，UVE-RF 和 UVE-SCARS-RF 模型的预测性能。与全光谱 RF 模型的测试集准确率相比，UVE-RF 模型精度下降 0.13%，UVE-SCARS-RF 模型精度下降 0.39%；对测试集的 1 534 份牛奶判别结果表明，UVE-RF 仅比全光谱 RF 模型多误判 2 个，UVE-SCARS-RF 比全光谱 RF 模型多误判 6 个。但在运行时间上，对测试集的 1 534份牛奶判别，全光谱 RF 模型的运行时间为 59.28 s；UVE 提取的特征变量数为全光谱变量的 44.55%，运行时间为全光谱 RF 模型的 44.74%；UVE-SCARS 提取的特征变量数为全光谱的 9.14%，运行时间仅为全光谱 RF 模型的 10.22%。综合考虑，最终选择 UVE-SCARS-RF 模型作为牛奶品质分级的最优 RF 模型。

2.6　最优模型的确定

对于 NB 模型，二阶差分-SCARS-NB 模型取得最优效果，训练集准确率与测试集准确率分别为 94.45% 和 93.94%，测试集中特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的预测准确率分别为 97.26%，93.93%，93.02%和 91.48%。对于 RF 模型，二阶差分-UVE-SCARS-RF 模型取得了最优效果，训练集准确率和测试集准确率为 99.86%，96.48%，测试集中特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶的预测准确率分别为 98.26%，97.40%，95.87% 和 94.49%。二阶差分-UVE-SCARS-RF 模型的训练集准确率与测试集准确率均高于 SCARS-NB 模型。综合考虑精度和效率，最终选择二阶差分-UVE-SCARS-RF 模型作为牛奶品质分级的最佳模型。

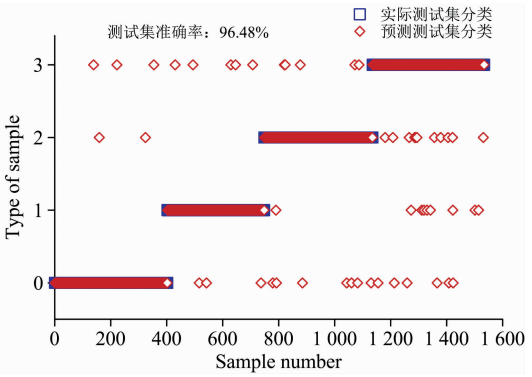

图 4　基于二阶差分-UVE-SCARS-RF 的分类模型

Fig. 4　Hierarchical model based on the second order difference-UVE-SCARS-RF

3　结　论

针对特优优质奶、高蛋白特色奶、高乳脂特色奶和普通奶建立了无损快速检测分级模型。选择来自 10 个牧场的 5 121份牛奶样本，保证了模型的通用性和可靠性。主要结论如下：

(1)探讨了牛奶品质分级的最佳预处理算法，结果表明无论是 NB 模型还是 RF 模型，二阶差分均为最佳预处理方法，并将其用于后续的建模分析。

(2)探讨了 UVE，CARS，SCARS，UVE-CARS 和 UVE-SCARS 5 种特征提取算法对 NB 模型和 RF 模型性能的影响。结果表明对于 NB 模型，SCARS 为最佳特征提取算法，对于 RF 模型，最佳的特征提取算法为 UVE-SCARS，但 RF 模型的精度优于 NB 模型。

(3)在实际生产中，效率也十分重要。在测试集中，二阶差分-SCARS-NB 模型的运行时间为 5.53 s，二阶差分-UVE-SCARS-RF 模型的运行时间为 6.06 s。综合考虑精度和效率，最终选择二阶差分-UVE-SCARS-RF 模型作为牛奶品质分级的最佳模型。

References

[1] GUO Li-ya, WU Jian-xin, ZHANG Xiao-jian, et al(郭利亚，吴建新，张晓建，等). China Dairy(中国乳业), 2020, (9): 66.

[2] CHEN Yan-sen, RUAN Jian, XIONG Jia-jun, et al(陈焱森，阮　健，熊家军，等). China Dairy Cattle(中国奶牛), 2018, (10): 66.

[3] CHEN Mei-jing, WANG Tong-tong, MENG Qing-yong(陈美静，王铜铜，孟庆勇). China Dairy(中国乳业), 2020, (7): 9.

[4] LI Chun-yan, QIN Bao-liang(李春艳，秦保亮). Graziery Veterinary Sciences · Electronic Version(畜牧兽医科学 · 电子版), 2019, (3): 16.

[5] De Marchi M, Bonfatti V, Cecchinato A, et al. Italian Journal of Animal Science, 2009, 8(2s): 399.

[6] Fleming A, Schenkel F S, Chen J, et al. Journal of Dairy Science, 2017, 100(6): 5073.

[7] Soyeurt H, Dehareng F, Gengler N, et al. Journal of Dairy Science, 2011, 94(4): 1657.

[8] DONG Li-feng, YAN Tian-hai, TU-yan, et al(董利锋，YAN Tianhai，屠　焰，等). Chinese Journal of Animal Nutrition(动物营养学报), 2016, 28(2): 326.

[9] YANG Yan-rong, YANG Ren-jie, DONG Gui-mei, et al(杨延荣，杨仁杰，董桂梅，等). The Journal of Light Scattering(光散射学报), 2014, 26(2): 203.

[10] YANG Ren-jie, LIU Rong, XU Ke-xin(杨仁杰，刘　蓉，徐可欣). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2011, 31(9): 2383.

[11] CUI Chuan-jin, GU Shao-peng, ZUO Yue-ming(崔传金, 古少鹏, 左月明). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2011, 42(1): 193.
[12] WU Hai-yun, ZUO Yue-ming, CUI Chuan-jin, et al(吴海云, 左月明, 崔传金, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2012, 43(8): 164.
[13] SHAO Le, YU Hong, LIU Xi-jing, et al(邵 乐, 于 红, 刘溪婧, 等). Journal of Dalian Ocean University(大连水产学院学报), 2010, 25(1): 45.
[14] LIU Meng, SHEN Si, WANG Nan(刘 猛, 申 思, 王 楠). Chinese Journal of Luminescence(发光学报), 2017, 38(5): 663.
[15] Bonfatti V, Di Martino G, Carnier P. J. Journal of Dairy Science, 2011, 94(12): 5776.
[16] Niero G, Penasa M, Gottardo P, et al. Journal of Dairy Science, 2016, 99(3): 1853.
[17] Etzion Y, Linker R, Cogan U, et al. Journal of Dairy Science, 2004, 87(9): 2779.
[18] FU Dan-dan, WANG Qiao-hua, GAO Sheng, et al(付丹丹, 王巧华, 高 升, 等). Chinese Journal of Analytical Chemistry(分析化学), 2020, 48(2): 289.
[19] GAO Sheng, WANG Qiao-hua, LI Qing-xu, et al(高 升, 王巧华, 李庆旭, 等). Chinese Journal of Analytical Chemistry(分析化学), 2019, 47(6): 941.

Nondestructive Testing and Grading of Milk Quality Based on Fourier Transform Mid-Infrared Spectroscopy

XIAO Shi-jie[1], WANG Qiao-hua[1, 2*], LI Chun-fang[3, 4], DU Chao[3], ZHOU Zeng-po[4], LIANG Sheng-chao[4], ZHANG Shu-jun[3*]

1. College of Engineering, Huazhong Agricultural University, Wuhan 430070, China
2. Key Laboratory of Agricultural Equipment in Mid-Lower Yangtze River; Ministry of Agriculture and Rural Affairs, Wuhan 430070, China
3. Key Laboratory of Animal Breeding and Reproduction of Minstry of Education, Huazhong Agricultural University, Wuhan 430070, China
4. Hebei Animal Husbandry Association, Shijiazhuang 050000, China

Abstract There are "high protein", "high fat", and other characteristics of milk in the market. In order to realize the nondestructive and rapid grading of super quality milk, high-protein characteristic milk, high-fat characteristic milk and ordinary milk, 5 121 milk samples from 10 pastures in Hebei Province in different months (January, March to October) were collected. Then the mid-infrared spectroscopy data were collected, the protein and fat content in milk were measured, the somatic cell number was measured, and the mid-infrared spectrum model of milk quality grading was established. Firstly, milk spectral analysis was carried out, and redundant bands were removed. Finally, the sensitive band combinations of 9 925～1 597 and 1 712～3 024 cm^{-1} were selected as the full spectrum to establish the model. In order to improve the prediction accuracy and efficiency of the model, six spectral pre-processing methods were used to improve the signal-to-noise ratio of the original spectrum, including Standard normal variable transform (SNV), multiple scattering correction (MSC), the first derivative and second derivative, first difference and second-order difference. Comparing the effects of different pretreatment methods by establishing naive Bayes model (NB) and random forest model (RF), the second-order difference obtained the best prediction accuracy. The testing set accuracy was 92.11% and 96.87%, respectively. So second-order difference was identified as the best pretreatment method for further analysis. In order to simplify the models, UVE (Uninformative variable elimination), CARS (Competitive adaptive reweighed sampling), SCARS (Stability Competitive adaptive reweighted sampling) were utilized to extract the characteristic variables from the pre-processed spectrum by second-order difference method. Then, the NB and RF models were established based on the full spectral data and the selected characteristic variable data. The results showed that SCARS was the best feature extraction algorithm for the NB model, and the accuracy rates of the training set and the testing set were 94.45% and 93.94%, respectively. UVE-SCARS is the best feature extraction algorithm of the RF model, and the accuracy of the training set and test set are 99.86% and 96.48%, respectively. In conclusion, the second-order difference-UVE-CARS-RF model established based on Fourier transform the mid-infrared spectroscopy technology can realize the rapid and non-destructive prediction of classification of 4 kinds of milk. Through the establishment of mid-infrared spectrum model, the combination of milk protein, milk fat content and somatic cell number is the first time for direct classification and identification, which is

unprecedented in previous studies. In applying the model, we only need to input the obtained milk mid-infrared spectral data into the model to output the prediction category, which has practical application value in the milk industry.

Keywords Mid-infrared spectrum; Milk; Quality grading; Nondestructive testing; Characteristics of the variable

(Received Apr. 8, 2021; accepted May 30, 2021)

* Corresponding authors

敬告读者——《光谱学与光谱分析》已全文上网

从 2008 年第 7 期开始在《光谱学与光谱分析》网站(www.gpxygpfx.com)“在线期刊”栏内发布《光谱学与光谱分析》期刊全文，读者可方便地免费下载摘要和 PDF 全文，欢迎浏览、检索本刊当期的全部内容；并陆续刊出自 2004 年以后出版的各期摘要和 PDF 全文内容。2009 年起《光谱学与光谱分析》每期出版日期改为每月 1 日。

《光谱学与光谱分析》期刊社