第42卷，第4期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 4，pp1173-1178
2022年4月 Spectroscopy and Spectral Analysis April，2022

SVM自助重加权采样的蚕茧雌雄特征波长选择

陈楚汉[1]，钟杨生[2]，王先燕[3]，赵懿琨[1]，代 芬[1*]

1. 华南农业大学电子工程学院，广东 广州 510642
2. 华南农业大学动物科学学院，广东 广州 510642
3. 广东省蚕业技术推广中心，广东 广州 510640

摘 要 使用近红外光谱鉴别蚕茧雌雄设备成本较高，挑选有用特征可以减少成本。雌雄蚕茧的近红外光谱存在着共线性的关系，因此提出了一种包裹式的特征选择方法，基于支持向量机的自助重加权采样(BRS-SVM)的特征选择方法。使用NirQuest512近红外光谱仪采集了蚕茧的漫透射近红外光谱。用试验集的全波段建模得到特征重要度热图，并通过热图得到重要特征波段的范围。然后在重要特征波段范围内，分别用BRS-SVM、基于SVM的特征排序方法(MBR-SVM)、基于逻辑回归的特征排序方法(MBR-LR)、递归特征消除法(RFE)、连续投影算法(SPA)和遗传算法(GA)挑选单波段特征和连续波段面积特征，再分别用支持向量机(SVM)和逻辑回归(LR)建立雌雄分类模型。通过特征重要性热力图发现，蚕茧雌雄分类重要区域在900～1 399 nm内，用此波段范围建立SVM模型，试验集准确率为99.40%。用BRS-SVM挑选5个单波段特征，然后再用SVM建模，验证集准确率为93.88%，高出其他特征选择方法5%～12%，测试集准确率为89.56%，测试集准确率高出其他特征选择方法2%～4%。用BRS-SVM挑选27个单波段特征，建立SVM雌雄分类模型测试集准确率为94.97%，准确率达到生产条件要求。用BRS-SVM挑选的14个连续波段面积特征，再用SVM建模，测试集准确率为94.43%。在挑选少量特征情况下，我们提出的BRS-SVM要优于其他方法。用BRS-SVM挑选少量的特征，可以建立性能良好的蚕茧雌雄分类模型，有效减少了成本，具有重要的现实意义。

关键词 蚕茧；近红外光谱；特征选择
中图分类号：G307 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)04-1173-06

引 言

蚕茧雌雄鉴别是蚕茧杂交育种的重要一步[1]。从熟蚕上蔟到蚕蛹化蛾共约14 d，蚕种场一般在第8天进行削茧鉴蛹辨别雌雄，削茧鉴蛹时间只有4～5 d，在短时间内，完成削茧鉴蛹需要大量人工，劳动成本高。使用近红外光谱对蚕茧进行雌雄鉴别，成本比较高，使用较少的近红外波段可以节约成本。

目前关于蚕茧性别自动鉴定的方法大多都是有损的，需要人工削茧，这些方法有荧光蚕茧辨性[2]、磁共振成像、X射线成像技术，高光谱成像技术[3]，计算机视觉方法和近红外光谱分析[4-5]等。目前还没有结合化学计量学和近红外光谱的蚕茧性别自动鉴别的研究[6]。使用全波段光谱进行分析，仪器成本较高，无法大规模应用在实际生产中。

数据提取是把之前维度的特征映射到一个更低维度的空间[7]，但数据提取的方法无法减少使用的近红外光谱波段。在近红外光谱分析中，用特征选择方法挑选单波段特征[8]，然后用挑选出来特征波长对应的单波发光二极管(LED)或激光光源代替近红外光谱仪[9]，能节约设备成本。

根据上述需求，提出了一种基于统计学的包裹式方法，基于SVM的自助重加权采样(bootstraping re-weighted sampling support vector machines，BRS-SVM)的特征选择方法。近红外光谱分析依靠不同样品光谱间的微小变化进行分析[10]，连续波段面积能很好反映出不同样本光谱间的微小差异。用BRS-SVM分别挑选单波段特征和连续波段面积特

收稿日期：2021-07-13，**修订日期**：2021-11-03
基金项目：国家自然科学基金项目(61675003)，广州市科技计划项目(201707010346)，广东省现代农业产业技术体系创新团队项目(2019KJ124)，广州市科技计划项目(202103000090)资助
作者简介：陈楚汉，1997年生，华南农业大学电子工程学院硕士研究生 e-mail：597748426@qq.com
*通讯作者 e-mail：sunflower@scau.edu.cn

征，再用支持向量机(support vector machines，SVM)和逻辑回归(logistic regression，LR)建立雌雄分类模型，以挑选相同特征个数时模型的准确率对特征选择方法评估，并和其他特征选择方法比较，分析实验结果，以期选择合适数量的窄LED灯代替近红外光谱仪。

1 理 论

1.1 基于学习模型的特征排序

基于学习模型的特征排序方法是基于学习器，通过衡量学习器特征的权重大小，给特征重要性排序，去除不重要的特征。其优势是可以快速去除大量不重要特征，但是不适合挑选较少特征。本工作使用基于SVM的特征排序方法(model based ranking support vector machines，MBR-SVM)和逻辑回归LR的特征排序方法(model based ranking logistic regression，MBR-LR)。

1.2 递归特征消除

特征选择的方法分为过滤试，包裹式和嵌入式。包裹式特征选择法的特征选择过程与学习器相关，使用学习器的性能作为特征选择的评价准则，选择最有利于学习器性能的特征子集[11]。递归特征消除(RFE)是一种包裹式特征选择的方法，该方法类似使用了多次基于学习模型的特征排序方法，每次迭代消除少量特征。以SVM-REF为例，在每一轮训练过程中，会选择所有特征来进行训练，继而得到了分类的超平面，SVM-REF会消除较小的权重，本工作每次迭代消除两个特征。

1.3 连续投影算法

连续投影算法(successive projections algorithm，SPA)是前向特征变量选择方法。SPA利用向量的投影分析，通过将波长投影到其他波长上，比较投影向量大小，以投影向量最大的波长为待选波长，然后基于矫正模型选择最终的特征波长。SPA选择的是含有最少冗余信息及最小共线性的变量组合。

1.4 遗传算法

遗传算法(genetic algorithm，GA)是模拟达尔文进化论的自然选择和遗传学机理的生物进化过程的计算模型，是一种模拟自然进化过程搜索最优解的方法，利用选择，交叉和突变等进化因子使得种群的适应度不断增强，从而达到优胜劣汰的目的。本工作利用SVM给个体适应度评分。

1.5 基于SVM的自助重加权采样(BRS-SVM)

BRS-SVM是一种包裹式法，该方法通过统计学的方式，评价不同组合的特征子集的得分，逐步选取最优的特征子集，子集搜索策略是启发式搜索策略，这种搜索策略效率要远优于全局最优搜索；自助法是一种启发式搜索策略，在光谱特征选择中有着较好的效果[12]。BRS-SVM能够快速有效的寻找最优的特征组合。BRS-SVM大致可以分为子集搜索和子集评价部分，首先初始化每个特征的权重 u 和抽取特征的数量，其中每个特征的初始权重 u 相等且和为1，抽取的特征个数等于样本特征个数。子集搜索部分：(1)首先初始化 n 个样本空间，即重复 n 次将数据随机分成80%的训练集和20%验证集，样本空间个数 n 越大，统计次数就越多；(2)在 n 个样本空间下，每个样本空间按权重为 u 进行随机重复抽样，抽取出 m 个特征。子集评价部分：(1)根据自助法，排除重复的特征，剩下约0.632 m个不重复的特征；(2)每个样本空间分别用SVM建模，然后用验证集准确率评价抽取的特征子集；(3)得分前10%的特征子集有利于学习器的性能，以得分前10%的特征抽取频率更新特征的权重 u；(4)以所有样本空间抽取不重复特征个数的评价值更新抽取个数 m。重复子集搜索和子集评价部分，直到抽取个数 m 满足需求，算法流程图如图1所示。设置BRS-SVM的样本空间大小为200。

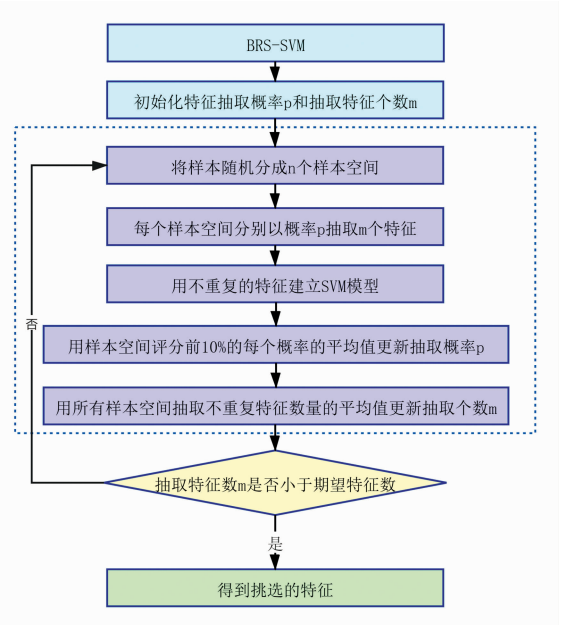

图1 BRS-SVM算法流程图

Fig. 1 BRS-SVM algorithm flow chart

1.6 计算环境

所有实验都重复计算50次，再求平均值，其中准确率的定义如式(1)所示

$$\text{accuracy} = \frac{T}{T+F} \times 100\% \tag{1}$$

式(1)中，T 为数据集分类正确的数量，F 为数据集分类错误的数量。

所有的运算都是在个人计算机上(Intel Core i5-4200，2.8 GHz CPU和12GB内存)用Pycharm(Python版本3.6.5，Tensorflow版本1.14.0，Keras版本2.3.1)进行的。

2 实验部分

2.1 仪器

样本的漫透射光谱采集使用课题组自行研制的种茧自动分选样机完成，光谱仪为海洋公司的NirQuest512型便携式光纤光谱仪，检测范围：900～1 699 nm。光谱仪设置积分时间为200 ms，平均次数为4以提高数据的稳定性，平滑宽度

为 4 以匹配系统的分辨率，样机如图 1 所示。样机工作步骤如下：

(1)将未剥壳的蚕茧放入左边进料口中，机械臂会抓取蚕茧到转盘中。

(2)转盘再将蚕茧转到光源(100 W 的卤灯泡)处，光源从上向下照射蚕茧，积分球在蚕茧下面采集蚕茧的漫透射光，通过 600 μm 光纤连接光谱仪。

(3)通过 USB 线将光谱仪采集的光谱数据传输给电脑，保存数据。

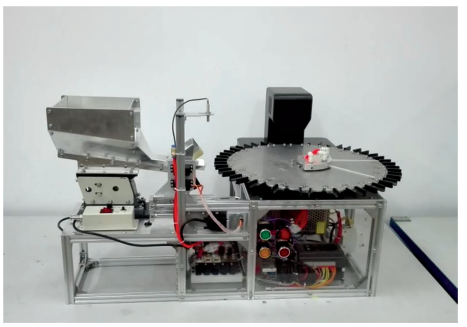

图 2　种茧自动分选样机

Fig. 2　Automatic silkworm sorting machine

2.2　样本

试验用的家蚕种茧样本来自于广东省蚕业推广中心和广东化州种茧场。将 2019 年 4 月至 2020 年 10 月采集的 4517 个近红外光谱样本作为试验的数据集，2021 年 6 月采集的 1 695个样本作为测试集，其中数据集信息如表 1 所示。9 芙×7 湘是 9 芙和 7 湘的第一代杂交品质，它们体型大小十分接近。试验集和测试集数据的采集时间不同，但他们品种接近，用测试集数据能很好验证试验的有效性。将茧壳削开，通过观察蚕茧尾部花纹来判断蚕蛹雌雄。

表 1　试验数据集的详细信息

Table 1　Details of the data sets

数据集名称	样本品种	样本数量/个	雌样本数/个	雄样本个数/个
试验集	9 芙	4 517	2 062	2 455
测试集	9 芙×7 湘	1 695	813	882

2.3　光谱数据

图 3 为 9 芙和 9 芙×7 湘通过 NirQuest512 型便携式光纤光谱仪采集到的雌雄蚕茧平均光谱，采集范围为 900～1 699 nm。由图 3 可以看出，两种品种的蚕茧雌雄光谱有 5 个相同的谱峰，峰值波长分别为 918，970，1 084，1 186和 1 269 nm。两种品种雌雄蚕茧的平均近红外光谱的谱峰差别不大，且它们谱峰都较宽。通常，雌蚕蛹的个体要比雄蚕蛹的大，所以相同品种情况下，雌蚕茧的平均近红外漫透射率要低于雄蚕茧的。雌雄蚕茧的漫透射近红外光谱存在交叉，但其交叉规律较为复杂，很难观察出雌雄蚕茧光谱差异较大的波长，因此需要使用相关算法挑选出相应的特征波长。

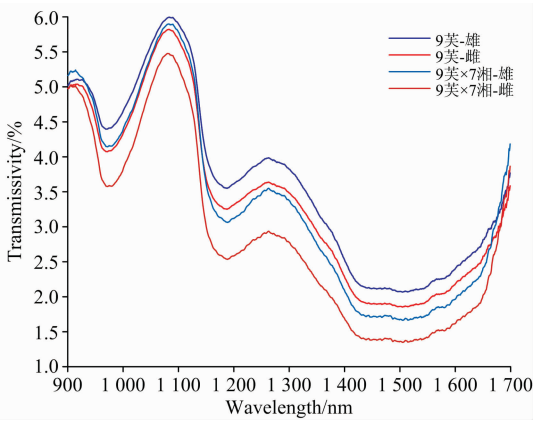

图 3　蚕茧平均近红外光谱

Fig. 3　Mean near infrared spectra of cocoon

3　结果与讨论

3.1　去除无信息波段

将试验集随机分为 80%的训练集和 20%的验证集。使用训练集的全波段光谱数据建立 SVM 模型，验证集准确率为 99.16%，以该 SVM 模型的权重大小为评判标准，权重越大特征越重要，将 900～1 699 nm 波段特征的重要性排序，并根据排序将重要程度缩放到 0～1，其中重要程度的计算如式(2)所示

$$\text{importance} = \frac{800 - S + 1}{800} \tag{2}$$

式(2)中，S 为特征重要性的排序。得到全波段特征重要性热力图，如图 4 所示，辞雄分类的重要特征都集中在 900～1 399 nm，使用该波段范围的训练集建立 SVM 模型，验证集准确率为 99.40%，所以我们认为雌雄分类信息大部分在 900～1 399 nm 波段内。

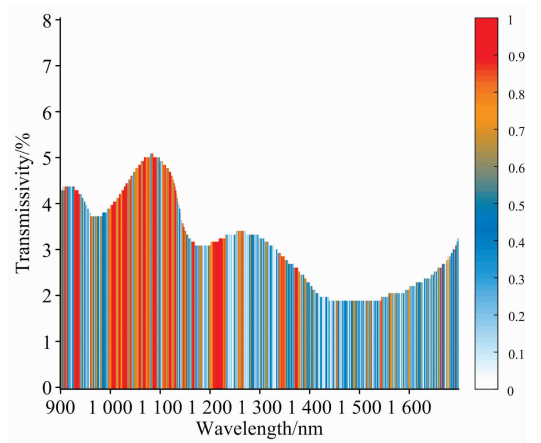

图 4　蚕茧近红外光谱特征重要性热力图

Fig. 4　The importance heatmap of near infrared spectral characteristics of cocoon

3.2　挑选单波段特征

在 900～1 399 nm 波段内挑选蚕茧雌雄分类的有用信息，分别使用 MBR-SVM，MBR-LR，REF-SVM 和 SPA 挑

选 5，10，20，30，40 和 50 个特征，GA 和 BRS-SVM 无法抽取固定的特征个数。将试验集随机分为 80%训练集和 20%验证集，使用挑选出来的特征训练 SVM 和 LR 雌雄分类模型，计算验证集准确率，重复上述 50 次，得到平均验证集准确率如图 5 所示，其中 MBR-SVM-SVM 表示使用 MBR-SVM 挑选特征，再使用 SVM 建模，同理可得其他图例含意。使用同种特征选择的方法挑选特征，再使用 SVM 模型建模的准确率比 LR 模型准确率高。挑选 5 个特征，BRS-SVM-SVM 验证集准确率为 93.88%，GA-SVM 验证集准确率为 89.24%，而其他特征选择方法只有 80%～82%。BRS-SVM 的性能要优于 GA-SVM，而 GA-SVM 的性能要优于其他算法。

用特征选择方法在试验集中挑选特征，得到的特征再用测试集建立分类模型，测试集准确率如图 5 所示。用测试集 900～1 399 nm 波段建立 SVM 雌雄分类模型准确率为 95.70%，建立 LR 雌雄分类模型准确率为 95.54%。用 BRS-SVM 挑选 5 个特征使用 SVM 建模准确率为 89.56%，其余准确率大多在 86%～87%，SVM 建模的准确率比 LR 的高，当挑选大于 9 个特征个数时，RFE-SVM，GA-SVM 和 BRS-SVM 性能接近，用 BRS-SVM 挑选 27 个特征 SVM 建模准确率为 94.97%，和使用 900～1 399 nm 波段建模准确率接近。通过上述实验，证明挑选单波段特征时我们的方法要优于其他方法，尤其是挑选特征数量较少的情况下。

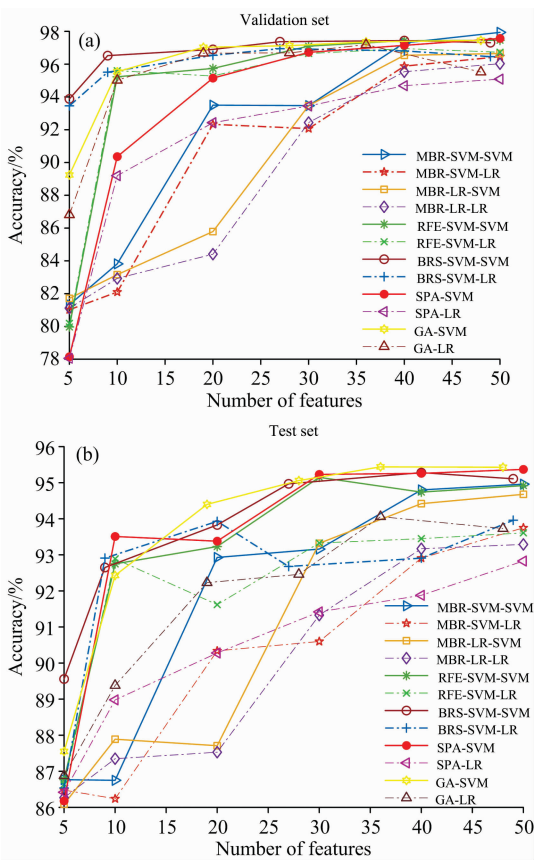

图 5 挑选的单波段特征的准确率图

Fig. 5 Accuracies of models using selected single-band features

3.3 挑选连续波段特征

计算试验集 900～1 399 nm 波段内的面积特征，如 900 nm 需要计算 900，900～901 和 900～902 nm 等 15 个连续波段的面积，1 385～1 399 nm 范围向 1 400 nm 后面的波段计算，共获取 7 500 个新的特征，再使用 MBR * SVM，MBR-LR，REF-SVM，SPA，GA 和 BRS-SVM 挑选连续波段的面积特征，其中 MBR-SVM，MBR-LR，REF-SVM 和 SPA 分别挑选 5，10，20，30，40 和 50 个特征，验证集准确率如图 6 所示，测试集准确率如图 6 所示。用 BRS-SVM 挑选 5 个特征再用 SVM 建模，验证集准确率为 94.17%，测试集准确率为 91.95%。用 REF-SVM 挑选 5 个特征再用 SVM 建模，验证集准确率为 86.30%，测试集准确率为 85.91%，用 GA 挑选 5 个特征再用 SVM 建模，验证集准确率为 89.30%，测试集准确率为 86.66%，在总特征数量较多且挑选少量特征的情况下，我们提出的 MBR-SVM 要优于 REF-SVM 和 GA，在挑选特征数量大于等于 20 个时，REF-SVM 的性能和 MBR-SVM，GA 相同。

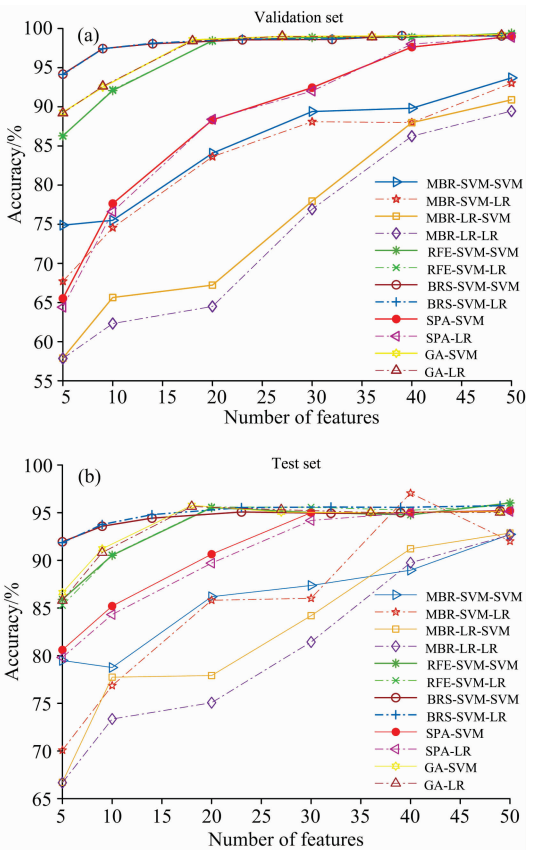

图 6 挑选的连续波段面积特征的准确率图

Fig. 6 Modeling accuracies of selected band area features

3.4 蚕茧近红外光谱的特征分析

图 7(a)为用 BRS-SVM 挑选的 27 个单波段特征，用这些特征建立 SVM 雌雄分类模型测试集准确率为 94.97%。图 7(b)为用 BRS-SVM 挑选的 14 个连续波段面积特征，用 SVM 建模测试集准确率为 94.43%，可用 13 个 LED 灯替代近红外光谱。可以根据实际生产需求选择合适的特征，成本

较低准确率要求不高，可选择挑选连续波段面积的特征，如用 BRS-SVM 挑选的 5 个连续波段面积特征，再用 SVM 建模测试集准确率为 91.95%，可用 5 个 LED 灯替代近红外光谱。

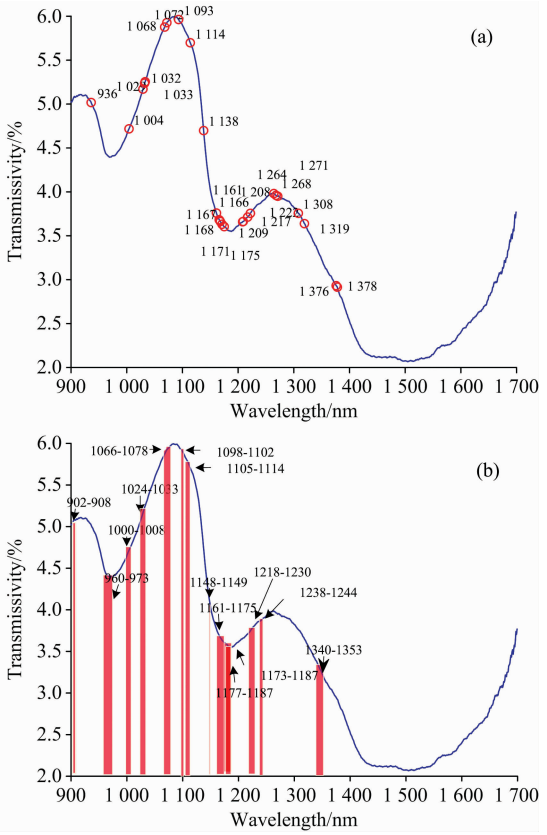

图 7　(a)BRS-SVM 挑选的 27 个单波段特征；(b)BRS-SVM 挑选的 14 个连续波段面积特征

Fig. 7　27 (a) single-band features selected by BRS-SVM; (b) 14 band area features selected by BRS-SVM

3.5　特征泛化能力分析

为了进一步验证挑选的特征的有效性，我们用 SW2540 型便携式光纤光谱仪采集 112 个 932 品种蚕茧的漫透射光谱和 77 个 7xia 品种蚕茧的漫透射光谱。用 BRS-SVM 挑选的 27 个单波段特征和 14 个连续波段面积特征建立 SVM 雌雄分类模型，准确率如表 2 所示。932 品种的分类模型效果差些，这是因为不同光谱仪或者不同品种的蚕茧采集的近红外光谱存在着差异。

表 2　932 和 7xia 品种蚕茧的 SVM 雌雄分类模型准确率

Table 2　Accuracy of SVM sex classification model for silkworm cocoons of 932 and 7xia

品种	单波段特征/%	连续波段面积特征/%
932	93.91	90.61
7xia	94.75	94.75

4　结　论

提出了一种包裹式的特征选择方法，基于支持向量机的自助重加权采样(BRS-SVM)的特征选择方法，分别对蚕茧近红外光谱单波段特征和连续波段特征进行选择，建立有效的雌雄分类模型。BRS-SVM 与其他特征选择方法相比性能均有一定优化，特别是在挑选少量特征时模型精度最高。在需求为低成本和低精度的情况下，挑选 5 个单波段特征，测试集准确率为 89.56%，在需求为高精度的情况下，挑选 14 个连续波段面积特征，测试集准确率为 94.97%。首次结合化学计量法分析蚕茧的近红外光谱，为蚕茧的雌雄检测应用提供一种实用的解决方案。

References

[1] FENG Wei-song(封槐松). China Sericultur(中国蚕业), 2018, (1): 1.

[2] ZHAO Ming-yan, JIANG Xin-yu, NIU Bao-long, et al(赵明岩, 蒋昕余, 牛宝龙, 等). Science of Sericulture(蚕业科学), 2018, 44(5): 711.

[3] Tao D, Wang Z, Li G, et al. Spectroscopy Letters, 2018, 51(8): 446.

[4] YAN Hui, LIANG Meng-xing, GUO Cheng, et al(颜　辉, 梁梦醒, 郭　成, 等). Science of Sericulture(蚕业科学), 2018, 44(2): 283.

[5] Zhu Z, Yuan H, Song C, et al. Sensors and Actuators B: Chemical, 2018, 268: 299.

[6] DAI Fen, CHE Xin-xin, PENG Si-ran, et al(代　芬, 车欣欣, 彭斯冉, 等). Journal of South China Agricultural University(华南农业大学学报), 2018, 39(2): 103.

[7] HONG Bin, DENG Bo, PENG Fu-yang, et al(洪　斌, 邓　波, 彭甫阳, 等). Chinese Journal of Computer(计算机科学), 2016, 43(8): 19.

[8] LI Yu-qiang, PAN Tian-hong, LI Hao-ran, et al(李鱼强, 潘天红, 李浩然, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(12): 3809.

[9] JIAO Lei-zi, DONG Da-ming, ZHAO Xian-de, et al(矫雷子, 董大明, 赵贤德, 等). Smart Agriculture(智慧农业), 2020, 2(2): 59.

[10] WANG Ling, LI Ding-ming, QIAN Hong-juan, et al(王　玲, 李定明, 钱红娟, 等). Chinese Journal of Analysis Laboratory(分析试验室), 2016, 35(10): 1203.

[11] Shardlow M. An Analysis of Feature Selection Selection Techniques, Mathematics, Computer Sciences, 2011.

[12] Zhang J, Xiong Y, Min S. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 223: 117110.

Feature Selection Algorithm for Identification of Male and Female Cocoons Based on SVM Bootstrapping Re-Weighted Sampling

CHEN Chu-han[1], ZHONG Yang-sheng[2], WANG Xian-yan[3], ZHAO Yi-kun[1], DAI Fen[1]*

1. College of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China
2. College of Animal Science, South China Agricultural University, Guangzhou 510642, China
3. Guangdong Sericulture Technology Promotion Center, Guangzhou 510640, China

Abstract The cost of identifying male and female cocoons by NIR is high, and the cost can be reduced by selecting useful features. Since there is a nonlinear relationship between the NIR spectra of female and male cocoons, a wrapper feature selection method, Bootstrapping Re-weighted Sampling Support Vector Machines (BRS-SVM), was proposed. The diffuse transmission NIR spectra of silkworm cocoons were collected by NirQuest512 NIR spectrometer. The heat map of characteristic importance was obtained by modeling the whole band of the test set, and the heat map obtained the range of important characteristic bands. Then, in the range of important characteristic bands, the single band features and continuous band area features were selected by BRS-SVM, Model-based ranking support vector machines (MBR-SVM), Model-based ranking Logistic Regression feature sorting method (MBR-LR), Recursive feature elimination (RFE), successive projections algorithm(SPA), Genetic Algorithm (GA), and then the support vector machines (SVM) and Logistic Regression (LR) sex classification models were established respectively. According to the characteristic importance heat map, it is found that the important area of male and female classification of silkworm cocoon was within 900～1 399 nm. We used this band to build the SVM model, and achieved 99.40% accuracy. BRS-SVM was used to select 5 single-band features. The accuracy of the test set is 89.56%, which is 2%～4% higher than other feature selection methods. RS-SVM was used to select 27 single-band features, and the accuracy of the test set of the SVM gender classification model was 94.97%, which reached the requirements of production conditions. The accuracy of modeling test set by BRS-SVM was 94.43% for 14 continuous band features. In the case of selecting a small number of features, our proposed BRS-SVM is superior to other methods. Using BRS-SVM to select a small number of features, we can establish a good performance of the female and male cocoon classification model, effectively reduce the cost, has important practical significance.

Keywords Cocoons; Near infrared spectrum; Feature selection

(Received Jul. 13, 2021; accepted Nov. 3, 2021)

* Corresponding author