基于贝叶斯优化的SVM玉米品种鉴别研究

冯瑞杰[1]，陈争光[1,2*]，衣淑娟[3]

1. 黑龙江八一农垦大学信息与电气工程学院，黑龙江 大庆 163319
2. 黑龙江省现代农业物联网技术创新中心，黑龙江 大庆 163319
3. 黑龙江省水稻生态育秧装置及全程机械化工程技术研究中心，黑龙江 大庆 163319

摘 要 为了快速检测玉米品种类型，基于支持向量机(SVM)和近红外光谱联合建立玉米品种的分类模型。以郑单958、先玉335、京科968、登海605和德美亚等五个品种共计293个样本为研究对象，对采集的近红外光谱进行标准正态变量变换(SNV)处理后使用主成分分析法(PCA)对光谱数据进行降维处理。按照6∶1比例，随机选取251个样本为训练集，42个样本作为测试集，探讨贝叶斯优化算法(BO)对SVM模型性能的影响。分别使用网格搜索(GS)、遗传算法(GA)和BO算法等三种方法对SVM模型的两个重要参数惩罚因子C和径向基核函数参数γ进行寻优。选择各模型十折交叉验证识别准确率最高时对应的惩罚因子和核参数作为建模参数，建立SVM分类模型。将使用BO算法建立的SVM分类模型与使用GS和GA进行参数寻优后建立的模型性能进行比对。实验发现，使用BO优化的SVM分类模型相比于其他两种优化算法得到的SVM模型性能具有显著优势，测试集的识别准确率可达到100%。说明使用BO算法寻优的SVM模型参数是全局最优参数，其他两种优化算法寻优的参数可能陷入了局部最优，从而导致模型性能表现不佳。在进行PCA降维前后的光谱数据上分别建立BO-SVM模型，结果表明，BO算法对于高维数据优化效果不佳，更适用于低维数据。对于不同样本类别间数量不均衡导致模型性能表现不佳的问题，通过剔除郑丹958和先玉335两类数量较少的样本，使用剩余三个类别，共计248个样本重新建立SVM模型，实验发现，剔除两类小样本之后，各个模型在测试集上的性能均有提升，说明对于类间样本数量不均衡问题，某类样本数量越多，对于模型参数的修正就越细腻，模型对该类的拟合效果就越好。研究结果可用于玉米品种的快速鉴别，也可为基于近红外光谱的其他农产品分类和产地鉴别提供参考。

关键词 近红外光谱；玉米；贝叶斯优化；主成分分析；支持向量机
中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)06-1698-06

引 言

作为世界的三大作物之一，玉米对于我国经济发展和社会稳定具有非常重要的战略意义。玉米品种繁多，同一地区种植的部分玉米品种外观极其相似，很难通过肉眼区分，给农民的采购和市场的监管带来了一定的困难。因此，需要一种快速检测技术对玉米品种进行识别。

随着化学计量学和仪器测量技术的飞速发展，光谱分析已经被广泛应用于农业[1-2]、食品[3-4]、医药[5]等领域。近红外光谱分析具有分析速度快、分析效率高、分析成本低、对样品无损害、便于实现在线分析等优点。近年来，近红外光谱在农产品品种鉴别和产地溯源等方面得到广大科研工作者的重视。李杰等[6]利用近红外光谱结合无监督的主成分分析和有监督的线性判别分析方法分别构建茶叶品种鉴别模型，采用标准正态变量变换结合一阶导数预处理方式并结合无监督的主成分分析法实现绿茶样品种类鉴别分析，准确率达到75%，采用有监督的线性判别分析方法处理原始光谱数据，准确率可达到100%。高慧宇等[7]应用近红外光谱结合偏最小二乘判别分析建立转基因大豆的快速鉴别模型，通过选择样品形态、波长范围和光谱预处理方法对鉴别模型进行优化，提高模型鉴别正确率。有研究探索了近红外光谱结合BP神经网络建立北方粳稻种子快速鉴别模型，通过小波变换对全谱进行数据降维，分类准确率可达100%。

收稿日期：2021-05-10，**修订日期**：2021-07-05
基金项目：国家重点研发计划项目(2016YFD0701300)，黑龙江省省属高校基本科研业务费科研项目(ZRCPY201913)资助
作者简介：冯瑞杰，1994年生，黑龙江八一农垦大学信息与电气工程学院硕士研究生 e-mail：fengruijie11@163.com
* 通讯作者 e-mail：ruzee@sina.com

基于高维数据的分类方法很多，其中采用二分类的支持向量机由于其优越的表现得到广泛的应用。支持向量机(support vector machine，SVM)是机器学习中分析数据的监督式学习算法，被广泛应用于农业[8]、医疗[9]、工业设备故障检测[10]及图像分类[11]等领域。SVM 的核心思想是将低维空间中不可分的数据点映射到更高维的空间维度中，在高维空间中进行分离。为了简化计算过程，引入核函数定义从低维到高维空间的映射，以确保原始空间的变量可以很容易地计算内积。在 SVM 中，惩罚因子 C 和径向基核函数(radial basis function，RBF)参数 γ 两个参数决定 SVM 模型性能，因此参数寻优对 SVM 模型性能的表现至关重要。常用的参数寻优方法如网格搜索(grid search，GS)、遗传算法(genetic algorithm，GA)等普遍存在寻优时间长，针对非凸问题易陷入局部最优等不足。本研究采用贝叶斯优化(Bayesian optimization，BO)对 SVM 模型的惩罚因子 C 和 RBF 核参数 γ 进行寻优，以 5 种玉米种子作为研究对象，选择模型十折交叉验证识别准确率最高时对应的参数建立 SVM 玉米品种鉴别模型，为农产品的快速分类提供一种参考方法。

1　实验部分

1.1　样本与仪器

试验所用玉米种子购买于种子市场，包括郑单 958、先玉 335、京科 968、登海 605 和德美亚五个品种。每个品种取 200 粒作为一个样本，5 个品种分别有 22，23，63，85 和 100 个样本，共计 293 个样本，去除有破损、瘪粒的样本。将玉米样本放置于近红外光谱实验室 24 h 之后进行光谱扫描。

光谱采集设备是德国 Bruker 公司生产的 TANGO 品牌的近红外光谱仪，测量波长范围为11 520～4 000 cm^{-1}，测量样本的方式为漫反射和透射，分辨率为 8 cm^{-1}，每个样本扫描 32 次取平均值作为样本的光谱数据。将每类样本按照 6∶1 的比例随机划分训练集和测试集，全部 293 个样本最终划分为 251 个训练集样本和 42 个测试集样本。

1.2　建模方法及评价指标

1.2.1　支持向量机

SVM 的基本思想是结构风险最小化，通过核函数将数据从原始特征空间映射到高维特征空间，使线性内积运算非线性化，然后在高维特征空间建立使分类间隔最大化的最优超平面。惩罚因子 C 和 RBF 核函数参数 γ 是 SVM 中两个重要的参数。惩罚因子 $C>0$，C 越大对错误分类的惩罚越大，但容易出现过拟合；C 越小则对错误分类的惩罚减小，模型的复杂度降低，容易出现欠拟合。γ 决定数据映射到新特征空间后的分布，γ 越小，支持向量越多，模型平滑效应增大，容易欠拟合；γ 越大，支持向量越少，对未知样本分类效果很差，模型容易过拟合。支持向量的个数影响模型训练与预测的速度，因此在使用 SVM 建立判别模型时，惩罚参数 C 和核函数参数 γ 的选择至关重要。

1.2.2　贝叶斯优化

SVM 模型参数 C 和 γ 与模型性能之间呈现黑箱特点，即模型的性能与参数 C 和 γ 之间无法使用表达式描述，只能根据通过遍历离散的自变量取值得到最优 SVM 模型。贝叶斯优化[12]是一种十分高效的全局优化算法，主要用于机器学习调参，贝叶斯优化是一种不需要计算导数的系统化调优算法，采用高斯过程建立概率代理模型，考虑之前的参数信息，不断更新先验，使用采集函数来确定下一个评估点，可以在较短的时间内确定最佳参数。概率代理模型和采集函数是贝叶斯优化算法的两个核心组件。高斯过程是随机变量的集合，用以代替目标优化函数。在本研究中，高斯过程用于优化的 SVM 的参数组合，高斯过程的表达式如式(1)

$$f(x) \sim GP[m(x), k(x, x')] \tag{1}$$

式(1)中，均值函数 $m(x)=E(f(x))$，代表样本 $f(x)$的数学期望。协方差函数 $k(x, x')=E\{[f(x)-m(x)][f(x')-m(x')]\}$，高斯过程根据已经搜索的点估计其他点处目标函数的均值和方差，通过均值和方差构造采集函数，用于决定下次迭代时的采样点位置。

常见的超参数优化算法包括网格搜索、遗传算法，这些算法除了非常耗时之外，在遍历下一个离散参数时不考虑之前的参数信息，针对非凸问题容易陷入局部最优。而贝叶斯优化侧重于减少评估代价，迭代次数少，速度快，而且考虑之前的参数信息，针对非凸问题不易陷入局部最优。本研究选择贝叶斯优化作为 SVM 模型的参数寻优算法。

贝叶斯优化算法的过程如下：

(1) 在 SVM 模型的 C 和 γ 的设定搜索范围中随机选取 n_0 个采样点，以十折交叉验证的平均测试准确率为目标函数 f，模型的不同参数组合作为自变量 x，构成代理模型框架，得到目标函数的初始分布和采样点集 D；

(2) 通过最大化采集函数选择下一个采样点 x_t，得到采样点函数值 $f(x_t)$；

(3) 将新的采样点$[x_t, f(x_t)]$添加到采样点集 D 中，更新高斯过程代理模型，使得代理模型更加贴合目标函数的分布；

(4) 设定一个最大迭代次数，当迭代次数达到最大次数时，停止算法迭代，输出最优采样点以及对应的目标函数最优值，即 SVM 模型的最优参数 C 和 γ。

1.2.3　评价指标

本研究基于混淆矩阵，引入 f_1 评价指标作为模型的评价标准。f_1 评价指标的计算公式如式(2)

$$f_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{2}$$

式(2)中，recall 和 precision 分别叫做查全率和查准率，其定义如式(3)和式(4)

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{4}$$

式(3)和式(4)中，TP 为将正类预测为正类的个数；FN 为将正类预测为负类的个数；FP 为将负类预测为正类的个数。查全率(recall)越高，说明模型对正样本的识别能力越强；查准率(precision)越高，说明模型对负样本的区别能力越强。f_1 是两者的综合，f_1 越高，说明所建立的分类模型越稳健。recall 和 precision 任何一个数值减小，f_1 的值都会减小。

本研究还选择识别准确率作为玉米品种判别模型的评价指标。识别准确率是指正确预测的样本数占总预测样本数的比率，不考虑预测的样本是正类还是负类。

2 结果与讨论

2.1 数据预处理

为了降低光谱数据中谱带重叠、噪声信号对建模的干扰，在建模前需要对采集的原始光谱进行预处理。采用标准正态变量变换(standard normal variate transformation，SNV)对光谱数据进行预处理。SNV 主要用来消除固体颗粒大小、表面散射以及光程变化等因素对光谱数据的影响。图 1(a，b)分别为样本的原始光谱曲线和 SNV 处理后光谱曲线为全部玉米样本光谱数据经过 SNV 变换前后的光谱曲线。

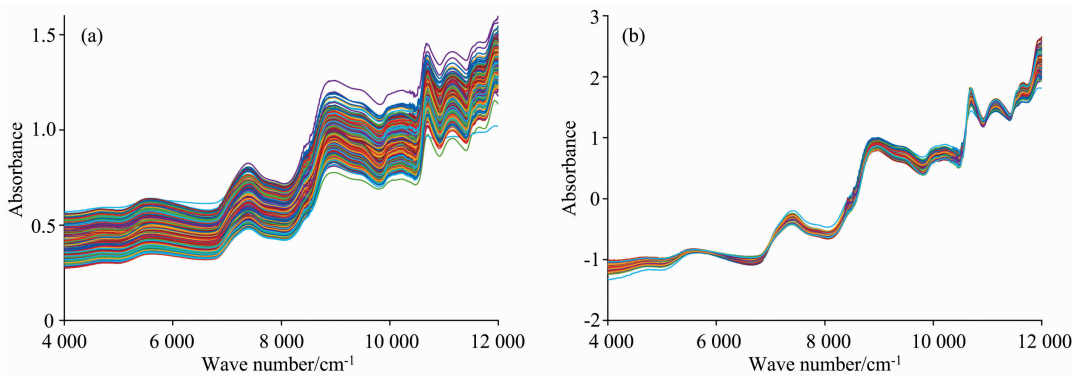

图 1 玉米种子原始近红外光谱图(a)及 SNV 处理后的光谱图(b)

Fig. 1 Original near-infrared spectra of corn seeds (a) and the spectra after SNV treatment (b)

2.2 不同优化方法下模型性能对比

在 SNV 预处理后的数据基础上，使用 10 折交叉验证分别建立 GS-SVM，GA-SVM 和 BO-SVM 模型，三种模型的参数以及性能指标如表 1 所示。表 1 的结果表明，BO 算法对 SVM 参数调优表现相比于 GS 和 GA 算法表现不佳，分析认为贝叶斯优化依赖于高斯过程建立概率代理模型，高斯过程作为一种概率分布，是事件最终结果的分布。高斯过程中的协方差函数 $k(x, x')$控制采样点的探索程度，对应于全局搜索，$k(x, x')$的计算依赖于已有样本的协方差矩阵。在高维数据的情形下，要使样本点布满整个搜索空间，需要大量的样本，有限的样本点在高维空间中的距离都会比较远，数据样本稀疏，会导致 $k(x, x')$近乎为无效函数。因此贝叶斯优化在高维数据中失去了其通过协方差函数进行探索的意义，近乎于完全随机搜索，算法不能通过采集函数进行高效的探索，有可能导致 SVM 模型陷入局部最优，模型表现不佳。说明在高维数据寻优方面，BO 算法并不是一个好的选择。

表 1 不同优化算法下的 SVM 模型性能对比

Table 1 Performance comparison of SVM models under different optimization algorithms

模型	惩罚因子 C	核参数 γ	训练集性能指标/%		测试集性能指标/%	
			准确率	f_1	准确率	f_1
GS-SVM	100.0	0.01	98.97	98.97	97.73	97.92
GA-SVM	256.0	0.003 9	98.40	98.13	97.73	97.92
BO-SVM	753.53	0.017 2	96.37	95.96	95.45	95.65

2.3 PCA 降维对贝叶斯优化及模型性能的影响

光谱数据经过 PCA 处理后，消除了数据特征间的共线性，去除了数据中不重要的特征，使得各个维度之间的数据相互正交，降低了数据的复杂性，并且大幅降低算法的计算开销。为了验证 PCA 降维对贝叶斯优化算法的影响，将高维度玉米近红外光谱数据利用 PCA 降维处理后保留 10 个主成分，前 10 个主成分的累计贡献率达到了 99.9%，在此基础上使用贝叶斯优化，对 SVM 模型参数 C 和 γ 进行优选并建立 PCA-BO-SVM 模型。采用十折交叉验证，计算模型的平均测试准确率，得到 SVM 模型的全局最优参数。同时在 PCA 降维的基础上建立 PCA-GS-SVM 和 PCA-GA-SVM 两种模型，三种模型性能参数如表 2 所示。由表 2 可知，对光谱数据使用 PCA 降维处理后，使用 GS 寻优得到的 SVM 模型核参数 γ 相比于 GA 以及 BO 算法寻得的核参数 γ 较大，模型出现轻微的过拟合，导致在测试集上表现不佳。对于 SVM 模型这样的连续型参数，GS 算法无法通过遍历所有 C 与 γ 可能参数组合去验证 SVM 参数空间中的所有参数，为了得到较优的参数组合，GS 算法必须加大网格搜索的密度，加之 GS 算法需要进行的交叉验证次数十分惊人，因此 GS 搜索方法耗费的时间成本巨大。

GA 算法的本质是随机性搜索，其调参的效果依赖于采样次数，采样次数越多，越有可能找到模型的全局最优参数，但随机采样点不容易落到最优参数组合上，并且 GA 算法无法利用之前采样点的评估效果进行主动寻优，寻优效率较低[13]，寻得的参数不一定是全局最优参数。BO 算法可以在很短的时间内寻得 SVM 的全局最优参数，这是因为 BO 算法使用采集函数，通过采集函数，在探索不确定区域和关注已知具有较优目标值的区域之间进行权衡，来确定下一个评估点。使用采集函数，可以使模型避开许多无用采样点的评估，准确描述出目标函数的分布，从而高效找到模型的最优参数组合。与 PCA-GS-SVM 和 PCA-GA-SVM 模型相比，PCA-BO-SVM 模型在测试集上的准确率和 f_1 值均达到

100%，说明经 BO 算法寻优后的 SVM 模型惩罚因子 C 和核参数 γ 均为全局最优参数，模型性能优于其他两种模型。

2.4　样本数量对训练模型的影响

四种模型在测试集上分类结果的混淆图如图 2，由混淆图可以看到，图 2(a)PCA-GS-SVM，图 2(b)PCA-GA-SVM 和图 2(c)BO-SVM 三种模型的识别错误率均与郑丹 958 有关，图 2(d)中 PCA-BO-SVM 模型在测试集中均可以正确识别各类玉米样本，识别效果优于其他三种模型。BO-SVM 模型的识别错误率与先玉 335 有关，这可能是由于郑丹 958 和先玉 335 样本数量较少导致模型对该样本的训练不够，在测试集上表现不佳所致。

为了验证这一猜想，剔除数据集中样本数量较少的郑丹 958(22 个样本)和先玉 335(23 个样本)两类样本，将剩余的 248 个玉米近红外光谱样本仍然按照 6∶1 的比例随机划分为训练集和测试集，使用 PCA-GS-SVM，PCA-GA-SVM，BO-SVM 和 PCA-BO-SVM 四种模型在训练集上建模，在测试集上进行玉米种类的识别，得到四种模型在三类玉米样本上的分类性能指标(表 3)。

表 2　降维后三种模型性能指标对比

Table 2　Comparison of performance indicators of the three models after dimensionality reduction

模型	惩罚因子 C	核参数 γ	训练集性能指标/%		测试集性能指标/%	
			准确率	f_1	准确率	f_1
PCA-GS-SVM	80.00	0.01	98.28	98.27	95.45	95.57
PCA-GA-SVM	254.67	0.0039	97.98	97.71	97.73	97.92
PCA-BO-SVM	503.05	0.0053	99.60	99.57	100	100

表 3　去除小样本后模型性能指标对比

Table 3　Comparison of model performance indicators after removing small samples

模型	惩罚因子 C	核参数 γ	训练集性能指标/%		测试集性能指标/%	
			准确率	f_1	准确率	f_1
PCA-GS-SVM	10.00	1.00	99.20	99.19	100.0	100.0
PCA-GA-SVM	101.92	2.02	100.0	100.0	100.0	100.0
BO-SVM	577.87	1.58	100.0	100.0	100.0	100.0
PCA-BO-SVM	387.27	1.90	100.0	100.0	100.0	100.0

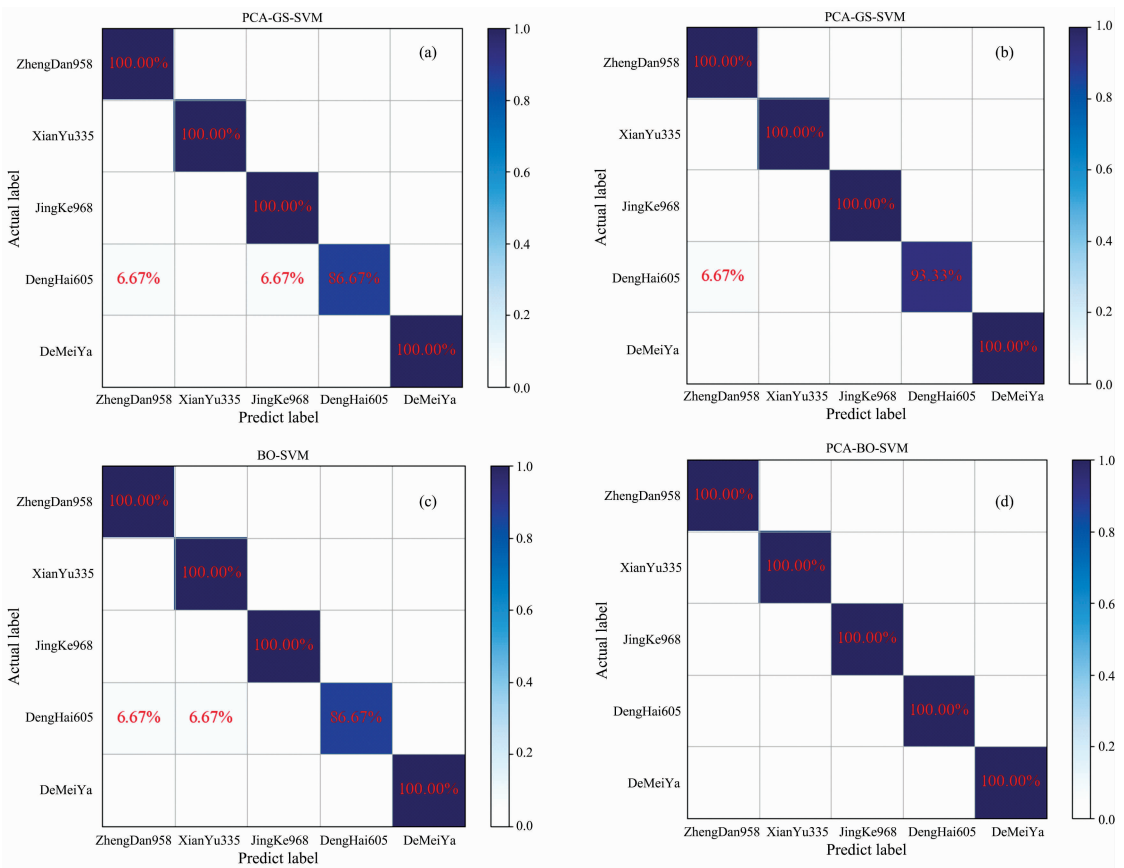

图 2　模型在光谱测试集上的混淆图

(a)：PCA-GS-SVM；(b)：PCA-GA-SVM；(c)：BO-SVM；(d)：PCA-BO-SVM

Fig. 2　Confusion map of the model on the spectral test set

(a)：PCA-GS-SVM；(b)：PCA-GA-SVM；(c)：BO-SVM；(d)：PCA-BO-SVM

由表 3 可以得出，在去掉郑丹 958 和先玉 335 两类小样本之后四种模型的训练集和测试集上的识别准确率均有显著提高，在测试集上的识别准确率均达到 100%。说明在类间数据量不平衡的模型训练过程中，模型对样本数据量较多的

类别拟合的更好，对该类的分类准确率较高[14]，但模型的泛化性能较弱。某种类别数据量越多，对模型参数的修正就越细腻，使模型更能刻画该类别的分布，对该类别数据的分类效果越好。

3 结 论

利用贝叶斯优化算法对 SVM 模型的两个超参数 C 和 γ 进行优化，结果表明，针对非凸优化问题，相较于网格搜索和遗传算法寻优，贝叶斯优化通过概率代理模型和采集函数来达到寻找模型全局最优参数的目的，充分利用完整的历史信息，避免不必要的参数评估，实现参数的高效优化，从而提高 SVM 模型的性能，基于贝叶斯优化的 SVM 模型的性能达到最优。由于贝叶斯优化适用于低维数据的模型参数优化，SVM 适合于小样本分类和回归，因此，数据降维能显著提高 SVM 模型的性能。此外，某类样本数量偏少会影响 SVM 模型的分类效果，导致模型的泛化性能减弱。本文利用 PCA，BO 和 SVM 构建了玉米品种的判别模型，为玉米品种的快速鉴别提供了一种新的方法。

References

[1] GUO Wen-chuan, ZHU De-kuan, ZHANG Qian, et al(郭文川, 朱德宽, 张 乾, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2020, 51(9): 350.

[2] CEN Zhong-yong, LEI Shun-xin, LEI Lei, et al(岑忠用, 雷顺新, 雷 蕾, 等). Journal of Huazhong Agricutural University(华中农业大学学报), 2021, 40(3): 1.

[3] WANG Qiao-hua, MEI Lu, GAO Sheng, et al(王巧华, 梅 璐, 高 升, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2019, 35(24): 314.

[4] YANG Chen-yu, YUAN Hong-fei, MA Hui-ling, et al(杨晨昱, 袁鸿飞, 马惠玲, 等). Food and Fermentation Industries(食品与发酵工业), 2021, 47(7): 211.

[5] LEI Xiao-qing, WANG Xiu-li, LI Geng, et al(雷晓晴, 王秀丽, 李 耿, 等). Chinese Traditional and Herbal Drugs(中草药), 2019, 50(16): 3947.

[6] LI Jie, LI Shang-ke, JIANG Li-wen, et al(李 杰, 李尚科, 蒋立文, 等). Journal of Instrumental Analysis(分析测试学报), 2020, 39(11): 1344.

[7] GAO Hui-yu, WANG Zhu, ZHANG Xue-song, et al(高慧宇, 王 竹, 张雪松, 等). Chinese Journal of Food Hygiene(中国食品卫生杂志), 2020, 32(3): 244.

[8] JIA Yin-jiang, JIANG Tao, SU Zhong-bin, et al(贾银江, 姜 涛, 苏中滨, 等). Journal of Northeast Agricultural University(东北农业大学学报), 2020, 51(7): 77.

[9] Madina H, Fatema S. International Journal of Electrical and Computer Engineering, 2017, 7(5): 2555.

[10] HAN Wei-yu, CHENG Long-sheng(韩卫宇, 程龙生). Computer Engineering and Applications (计算机工程与应用), 2020, 57(6): 239.

[11] LI Yu, GONG Xue-liang, ZHAO Quan-hua(李 玉, 宫学亮, 赵泉华). Chinese Journal of Scientific Instrument(仪器仪表学报), 2020, 41(12): 253.

[12] Shahriari B, Swersky K, Wang Z, et al. Proceedings of the IEEE, 2015, 104(1): 148.

[13] SANG He-cheng, SONG Shuan-jun, XING Xu-peng, et al(桑和成, 宋栓军, 邢旭朋, 等). Journal of Xi'an Polytechnic University(西安工程大学学报), 2021, 35(1): 44.

[14] XU Jian, WANG Xin-yue, CAI Zi-xin, et al(徐 剑, 王馨月, 才子昕, 等). Journal of Frontiers of Computer Science and Technology(计算机科学与探索), 2020, 14(3): 401.

Identification of Corn Varieties Based on Bayesian Optimization SVM

FENG Rui-jie[1], CHEN Zheng-guang[1,2*], YI Shu-juan[3]

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China
2. Technology Innovation Center for Heilongjiang Modern Agricultural Internet of Things, Daqing 163319, China
3. Heilongjiang Engineering Technology Research Center for Rice Ecological Seedings Device and Whole Process Mechanization, Daqing 163319, China

Abstract In order to detect corn varieties quickly, a classification model of corn varieties was established based on the combination of support vector machine (SVM) and near-infrared spectroscopy. 293 samples from five varieties, including Zhengdan 958, Xianyu 335, Jingke 968, Denghai 605 and Demeiya, were collected as research objects. After performing standard normal variable transformation (SNV) processing on the collected near-infrared spectra, the principal component analysis (PCA) method is used to reduce the dimensionality of the spectral data. According to the ratio of 6∶1, 251 samples were randomly selected as the training set and 42 samples as the test set to explore the influence of the Bayesian optimization (BO) algorithm on the performance of the SVM model. Three methods, including grid search(GS), genetic algorithm(GA) and BO algorithm, were used to optimize the two important parameters of the SVM model, namely, the penalty factor C and the radial basis kernel function parameter γ. The C and γ, corresponding to the highest recognition accuracy based on ten-fold cross-validation of each model, were used as modeling parameters, and the SVM classification model based on the three optimization algorithm methods were established. The SVM classification model based on BO is compared with the model based on GS and GA. The experimental results show that the performance of the SVM classification model optimized by BO is superior to that of the other two optimization algorithms, and the recognition accuracy on the test set can reach 100%. This shows that the parameters of the SVM model optimized by BO are the optimal global parameters, and the parameters optimized by the other two optimization algorithms may fall into the local optimal, resulting in poor performance of the model. BO-SVM models were established on the spectral data before and after PCA dimensionality reduction. The results show that BO is not good for high-dimensional data optimization, and it is more suitable for low dimensional data. For the problem of poor performance of the model caused by the imbalance of the number of different sample categories, the SVM models were re-established by removing the two small samples, Zheng Dan 958 and Xianyu 335, and using the remaining three categories, a total of 248 corn samples. The experimental results show that the performance of each model on the test set is improved after removing the two types of small samples, which indicates that for the problem of unbalanced sample number between classes, the more samples of a certain class, the more delicate the correction of model parameters, and the better the fitting effect of the model on this class. The results of this study can be used for rapid identification of corn varieties and can also provide references for the classification and origin identification of other agricultural products based on near-infrared spectroscopy.

Keywords Near infrared spectroscopy; Corn; Bayesian optimization; Principal component analysis; Support vector machine

(Received May 10, 2021; accepted Jul. 5, 2021)

* Corresponding author