基于形态特征参数的茶叶精选方法

吴正敏，曹成茂※，王二锐，罗 坤，张金炎，孙 燕
（安徽农业大学工学院，合肥 230036）

摘 要：夏秋季节的梗与叶片的色泽差异小，采用传统色选机难以实现精选。该文提出依据茶叶形态特征的多特征向量分选法，以期实现茶叶精选算法快速建模，提高分选精度。采集动态下落过程中的茶叶图像，开发基于图像处理的特征提取程序自动提取多组茶叶样本形态特征参数，采用随机森林算法判定特征权重并进行特征选择，建立逻辑回归、决策树和支持向量机 3 种不同分类算法对样本进行分类，验证特征的可分性，并分析不同分类算法对复杂茶叶样本分类效果的影响。试验结果表明：1）形态特征参数圆形度 E 的重要性权重最大，为 0.467，最终将重要性阈值设定为 0.05，选择圆形度 E、矩形度 R、线性度 Len、周长 C 和紧凑度 J 5 种形态特征向量建立数据集；2）在测试数据集中，逻辑回归（logistic regression, LR）、决策树（decision tree, DT）和支持向量机（support vector machine, SVM）3 种分类算法的平均准确率为 0.924，说明所选特征具有明显的可分性；3）根据输出的混淆矩阵，3 种分类算法中支持向量机算法识别效果最好，准确率和调和平均数（$F1$）得分分别为 93.8%和 94.7%。该方法可快速应用于其他类型茶叶精选和茶叶实际生产过程，有效提高茶叶品质。

关键词：形态特征；决策树；支持向量机；逻辑回归；随机森林；茶叶
doi：10.11975/j.issn.1002-6819.2019.11.036
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1002-6819(2019)-11-0315-07

吴正敏，曹成茂，王二锐，罗 坤，张金炎，孙 燕. 基于形态特征参数的茶叶精选方法[J]. 农业工程学报，2019，35(11)：315－321. doi：10.11975/j.issn.1002-6819.2019.11.036 http://www.tcsae.org
Wu Zhengmin, Cao Chengmao, Wang Errui, Luo Kun, Zhang Jinyan, Sun Yan. Tea selection method based on morphology feature parameters[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(11): 315－321. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2019.11.036 http://www.tcsae.org

0 引 言

茶叶随着生长季节的变化，其成品茶色泽也在改变，春茶茶叶偏绿，梗、叶色泽差异性较大，色选分离效果较好，夏秋茶毛茶偏墨绿，梗、叶色泽差异较小[1]，色选机基于茶叶良品与不良品光学特性分选难度很大；茶叶有六大品种，多种加工工艺，其成品茶形态特征复杂，如何快速建立不同类型茶叶形选模型，有效提高良品与不良品的分离率，依然是茶叶精加工的关键问题。本文研究对象大红袍是乌龙茶，加工过程中进行了轻发酵，其梗叶色泽差异更小，分离难度更大；此外，夏秋茶在加工过程中多经过揉捻环节，毛茶含梗率较高，必须有效分离，以提高茶叶品质。

近年来机器视觉广泛应用于农产品分选[2-5]，在茶叶识别、品质鉴定和分选领域，杨福增等[6]针对清明期“午子仙毫”，提取茶叶 G 分量，综合茶叶嫩芽形状特征，最终的识别准确率为 94%，董春旺等[7]基于机器视觉和工艺参数对针芽形绿茶外形进行评价，宋彦团队[8]针对 7 个等级祁门红茶，建立了数字化等级鉴定方法，Borah 等[9-11]研究了基于纹理特征的茶叶分类方法，Cimpoiu 等[12]采用神经网络实现对茶叶分类。在茶叶精选环节，张春燕等[13]提出基于最小风险贝叶斯分类器的茶叶茶梗分类方法，高达睿等[14]建立了基于茶叶颜色和形状特征的茶叶分选系统，但并不适用于颜色差异小、形态特征复杂的茶叶样本。本文基于茶叶形态特征提出了一种多特征向量下茶叶良品与不良品分选快速建模的方法，建立多个复杂形态特征描绘子，自动判别特征向量权重，快速选择有效特征；在 Python3 环境中开发逻辑回归、决策树和支持向量机 3 种算法实现茶叶精选，采用网格搜索和 K 折交叉验证方法优化算法模型，进行特征向量和分选算法评估。

1 材料与方法

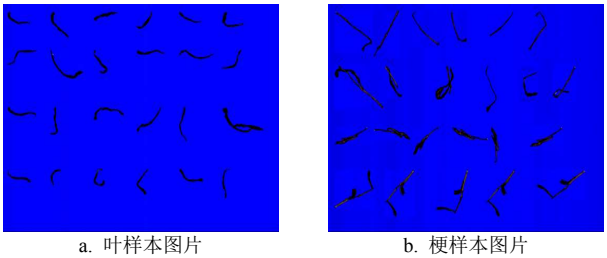
试验中选择武夷山大红袍作为样本，在单通道茶叶色选机中采集大红袍下落过程中的样本图像如图 1 所示，其中叶样本即良品如图 1a 所示，梗样本即不良品如图 1b 所示，梗样本由单梗、梗叶缠绕、一梗一叶及多叶等组成，其形态极不规律，部分梗与叶的形态相似度较高，梗叶分离难度非常大，为更好地建立特征描绘子，本文先对图像进行预处理，再提取特征参数。

1.1 图像预处理

批量加载样本图像，提取蓝色平面，再进行二值化，由于样本图像目标与背景区分度较大，选择 4 种较为简单的阈值分割方法对做过同样处理的样本图像进行阈值分割，全局阈值 Otsu 法阈值分割和双峰法分割图像的效

收稿日期：2018-12-14 修订日期：2019-04-18
基金项目：安徽省科技重大专项（18030701195）和安徽省高校自然科学研究项目（KJ2016A233）联合资助
作者简介：吴正敏，博士研究生，研究方向为茶叶智能化精加工。
Email：wuzhengmin@ahau.edu.com
※通信作者：曹成茂，教授，博士生导师，主要从事智能检测与控制技术、农业机械化工程研究。Email：caochengmao@sina.com

果更好，细节保留更为完整；迭代法全阈值分割后的图像和局部阈值分割后的图像细节丢失较多，为保证后期提取特征参数的准确性，迭代法全阈值分割和局部阈值分割图像的方法不宜采用，考虑到全局阈值 Otsu 法阈值分割较双峰法分割图像更为简单，处理更快，后期样本图像均采用全局阈值 Otsu 法阈值分割进行图像阈值分割。



a. 叶样本图片 b. 梗样本图片
a. Leaf samples image b. Stalk samples image

图 1 样本图片

Fig.1 Samples image

样本形态特征描绘子周长、圆形度、线性度等参数的提取需要图像的边缘信息，本文选择 Roberts、Sobel、Prewitt、LOG、Canny 这 5 种常用的边缘算子进行边缘提取，总体来看，Canny 算子提取的边缘细节最完整，不容易受噪声干扰，能够检测到真正的弱边缘。故后期样本图像边缘信息采用 Canny 算子提取。

1.2 建立形态特征描绘子

根据预处理后的图像信息提取基本形态特征面积 A、周长 C、长轴 L、短轴 S[15]。其中：面积表示茶叶图像边界线内包含的所有像素个数；周长 C 由茶叶样本图像边界像素点的总和来计算；长轴 L 为区域最小外接矩形的长；短轴 S 为区域最小外接矩形的宽。

大红袍样本基本形态特征向量受样本成像过程各因素影响较大，因此由基本形态特征建立复杂特征描绘子圆形度、直径、紧凑度、矩形度、细长度、对角线长度和线性度如表 1 所示。

表 1 复杂特征描绘子

Table 1 Feature description

特征类型 Feature types	定义 Definition	计算公式 Computational formulas
圆形度 Circularity	茶叶边界形状接近圆的程度，由面积和周长计算得到	$E=4\pi A/C^2$
直径 Diameter	茶叶面积相等的圆的直径	$D=2\sqrt{\frac{A}{\pi}}$
紧凑度 Compactness	茶叶直径与长轴之比	$J=D/L$
矩形度 Rectangularity	茶叶面积与区域最小外接矩形面积比	$R=A/(L\times S)$
对角线长 Diagonal length	茶叶图像最小外接矩对角线长度	$\mathrm{Dia}=\sqrt{L^2+S^2}$
细长度 Slightness	区域最小外接矩形长宽之比	$T=L/S$
线性度 Linearity	直线长度与最小外接矩对角线比	$\mathrm{Len}=\mathrm{Hough\_len}/\sqrt{L^2+S^2}$，其中，Hough_len 为样本图像基于霍夫变换得到的直线长

注：公式中 E、D、J、R、Dia、T、Len、A、C、L、S 分别为代表圆形度、直径、紧凑度、矩形度、细长度、对角线长度、线性度、面积、周长、长轴和短轴。

Note: E, D, J, R, Dia, T, Len, A, C, L and S represent the circularity,diameter, compactness, rectangularity, diagonal length, slightness, linearity,area, perimeter, long and short axis length.

根据描绘子基本特性，开发特征提取算法，从样本图像中提取出周长 C、长轴 L、短轴 S、圆形度 E、直径 D、紧凑度 J、矩形度 R、细长度 T、对角线长度 Dia、线性度 Len 10 个特征描绘子的参数值。

1.3 茶叶形态特征自动提取算法

根据图像预处理流程和特征描绘子描述方法，开发茶叶形态特征自动提取算法得到茶叶形态特征的原始数据集。基本步骤如下：

1）批量读取待处理图片×numm（图片总数）；

2）定义全局变量 Sum_num（样本总数）、Data（数据集）

3）对于每一幅读取的 RGB 图像，均提取图像蓝色平面，采用最大类间方差法分割图像，得到二值化图像，滤除干扰对象，基于 Canny 算子提取边缘，确定边缘连通域和区域连通域，统计每张图片上的样本总数 num；

4）计算每个样本的形态特征参数 Data=[Len T J E R D C L S Dig]和样本质心;

5）经过 num×numm 次循环，得到所有样本的特征数据。

1.4 随机森林算法与特征向量的快速选择

为实现茶叶分类算法的快速开发和优化，本文采用随机森林算法判断特征向量的重要性[16-18]，通过随机森林中所有决策树得到的平均不纯度衰减来度量特征的重要性[19]，基本计算方法如式（1）所示。

$$X_{\mathrm{importance}}=\sum(\mathrm{erroob2}-\mathrm{erroob1})/N \quad (1)$$

式中 erroob1 为袋外数据未加入干扰的数据误差，erroob2 为袋外数据所有样本的特征 X 加入噪声干扰后的误差，N 为随机森林中树的棵数。

计算出每个特征的重要性，设定一定的阈值，进行数据压缩，将提取的大红袍原始样本形态特征数据集进行归一化处理，基于 Python3.0 和 scikit-learn 库编写基于随机森林算法的特征重要性获取算法[20]，得出各特征数据重要性如图 2 所示。

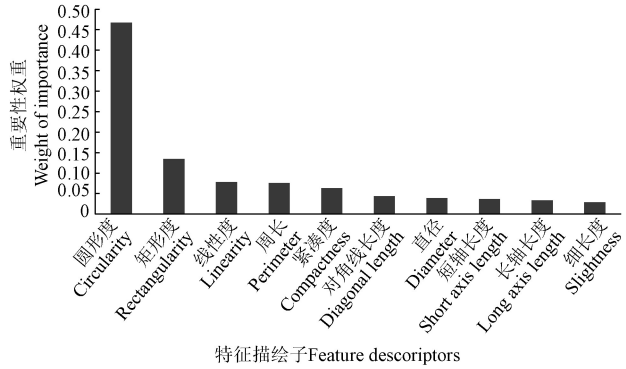


图 2 特征重要性

Fig.2 Feature importance

如图 2 所示，圆形度 E 权重最大，为 0.467，细长度 T 最小，为 0.029，可以看出圆形度这一特征在大红袍梗、叶分离中贡献率较大，该特征是有面积与周长的比值得到，从一定程度上消除了环境变化、光学等因素的影响，有较强的适应性，在其他类型茶叶梗、叶分离中可参考，

在一些茶叶分级、分类的文献中也有所运用，如高睿达[14]在六安瓜片的分级中便使用了该特征。为减少分类时的运算量，将权重阈值设为 0.05，最终选择圆形度 E、矩形度 R、线性度 Len、周长 C 和紧凑度 J 5 种特征向量，用于验证逻辑回归、决策树和支持向量机 3 种算法实现大红袍良品、不良品分离的效果。

1.5　机器学习系统基本结构

大红袍良品与不良品分类是典型的二分类问题，考虑到算法的易用性和准确率，文中选择了逻辑回归、决策树和支持向量机 3 种算法，逻辑回归是一种简单却又快速而强大的算法；决策树的优势在于它的模型可见性，能够清晰地看到它每一步是如何判定和执行；支持向量机一直在传统机器学习算法中占据重要地位，也是在目前众多实际运用如茶叶色选机优先选择的算法。为更贴近实际运用，本文最终选择这 3 种算法进行分类结果的验证。根据特征选择获得的特征向量，建立数据集，用于分类算法的训练、验证和测试。机器学习系统基本结构如图 3 所示。

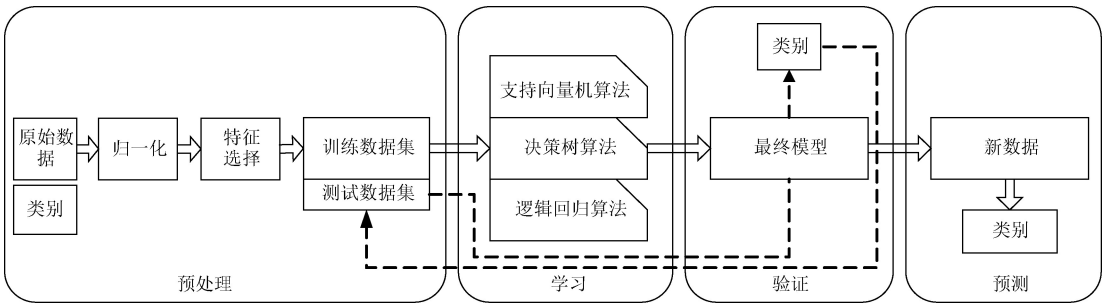


图 3　机器学习系统基本结构

Fig.3　Basic structure of machine learning system

1.6　试验设计

选择 480 个大红袍样本，采集样本动态下落过程中的图片，批量输入到茶叶形态特征自动提取算法程序中，建立样本形态特征数据集，首先将原始数据进行归一化处理，预处理后的特征向量进行随机分割，80%用于训练，20%用于测试，采用 10 折交叉验证选择分类模型最优参数，随机将训练数据集划分为 10 份，其中 9 份用于训练，剩下的 1 份用于验证。根据上述机器学习系统参数优化过程获得逻辑回归、决策树和支持向量机最优模型统计训练和测试数据集最终的评价结果。根据分类器输出的混淆矩阵真正（TP）、真负（TN）、假正（FP）及假负（FN）的样本数量，利用式（2）公式计算准确率 Accuracy、真正率 Precision、召回率 Recall 和调和平均数 F1 作为分类算法的评价指标[21]。

$$\begin{cases}\text{Accuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN})\\ \text{Precision}=(\text{TP})/(\text{TP}+\text{FP})\\ \text{Recall}=(\text{TP})/(\text{TP}+\text{FN})\\ \text{F1}=2(\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall})\end{cases}\tag{2}$$

2　机器学习系统设计与参数优化

2.1　逻辑回归算法

逻辑回归是经典的二分类算法，也可以实现多分类[22-23]。本文主要针对大红袍的梗叶进行分离，属于二分类问题，建立式（3）预测函数。

$$h_\theta(x)=g\left(\theta^{\mathrm{T}}x\right)=1/\left(1+e^{-\theta^{\mathrm{T}}x}\right)\tag{3}$$

式中 $\theta_0+\theta_1x_1+\ldots+\theta_nx_n=\sum_{i=1}^{n}\theta_ix_i=\theta^{\mathrm{T}}x$，$\theta$为权重；$x$ 为样本特征向量。

对于二分类任务（0,1），整合两种情况下的预测结果，得到（4）式

$$P(y/x;\theta)=\left(h_\theta(x)^y\left(1-h_\theta(x)^{1-y}\right)\right)\tag{4}$$

式中 y 为类别标签，利用似然函数求解，此时应用梯度上升求最大值，引入 $J(\theta)=-\frac{1}{m}l(\theta)$，$l(\theta)$ 表示对数似然函数，转换为梯度下降任务，求导得到损失函数（5）式。

$$\frac{\partial}{\partial\theta_j}J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x_i)-y_i\right)x_i^j\tag{5}$$

参数更新

$$\theta_j=\theta_j-\partial\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x_i)-y_i\right)x_i^j\tag{6}$$

式中∂表示学习率，m 为样本个数，i 表示第 i 个样本，j 表示第 i 个样本的第 j 个特征值。

文中采用网格搜索调优超参的方法确定最佳正则化惩罚系数，如图 4 输出学习曲线和验证曲线对优化过程进行观察。

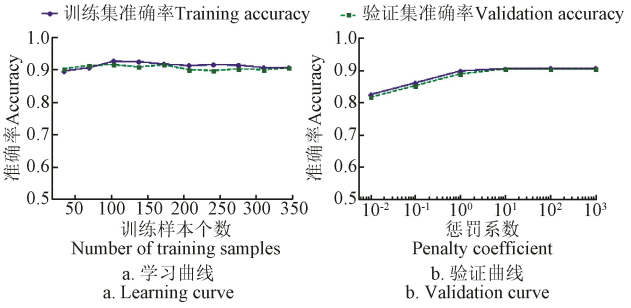


图 4　学习曲线和验证曲线

Fig.4　Learning curve and validation curve

从图 4a 可以看出训练数据准确率与验证集偏差较小，说明模型泛化能力较强，输入样本数据集，执行网格搜索程序后，从图 4b 验证曲线也可以看出惩罚系数的最优值在 10 左右，设置过低时，会导致分类准确率下降，

而大于 10 以后，模型基本稳定。

2.2 决策树算法

决策树是一种树型结构，其中每个内部节点表示在一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别[24-25]。建立决策树的关键是选择哪个属性作为分类依据，根据不同的目标函数，建立决策树主要有信息增益、信息增益率、Gini 系数 3 种算法[26]，其中：

信息增益：表示得知特征 X 的信息而使得类 Y 的信息的不确定性减少的程度，定义为训练数据集 D 的经验熵 $H(D)$与特征 X 给定条件下 D 的经验条件熵 $H(D|X)$之差，即

$$g(D,A)=H(D)-H(D/A) \tag{7}$$

信息增益率

$$g_r(D,X)/H(X) \tag{8}$$

Gini 系数

$$\mathrm{Gini}(p)=\sum_{k=1}^{k}p_k(1-p_k)=1-\sum_{k=1}^{k}p_k^2=1-\sum_{k=1}^{k}\left(\frac{|C_k|}{|D|}\right)^2 \tag{9}$$

式中 k 为类别，$|D|$表示样本个数，$|C_k|$为属于类 C_k 的样本个数。

上述 3 种分类依据，信息增益受数据样本自身熵影响很大；信息增益率考虑了自身熵的影响；Gini 系数在特征数据越纯时，值越低，应用更为广泛。

采用决策树算法对大红袍梗、叶样本进行分类试验，选择 Gini 系数作为分类依据，优化后的树模型参数最小叶子节点个数设为 4，最大深度设为 5。

2.3 支持向量机算法

SVM 作为传统机器学习的一个非常重要的分类算法，它是一种通用的前馈网络类型。根据核函数的不同可分为线性支持向量机和非线性支持向量机，支持向量机算法的主要影响因素是核函数的选择和相应参数的设置[26-30]。文中采用网格搜索调优超参的方法确定最佳正则化惩罚系数 C 和核函数类型及相应参数，基本算法实现步骤如下：

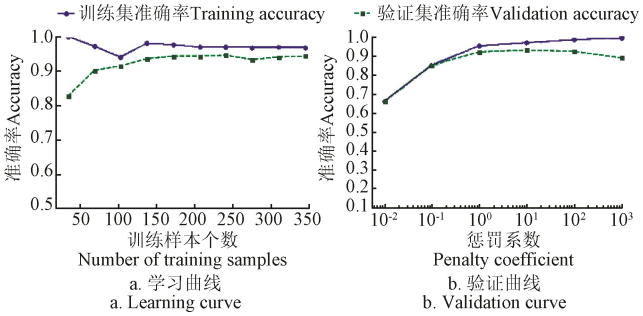
文中采用网格搜索调优超参的方法确定最佳正则化惩罚系数 C 和核函数类型及相应参数，基本算法实现步骤如下：

入样本数据

$$\mathrm{Data}=\{(x_{11},x_{12},x_{13},x_{14},x_{15},y_1),(x_{21},x_{22},x_{23},x_{24},x_{25},y_2),\ldots,(x_{i1},x_{i2},x_{i3},x_{i4},x_{i5},y_i)\} \tag{10}$$

其中特征向量个数为 5，样本数量 i，$y_i\in\{+1,-1\}$，x_i 为第 i 个茶叶样本实例，y_i 为 x_i 的类标记：当 y_i=+1，称 x_i 为良品；当 y_i=−1，称 x_i 为不良品。

2）首先将原始数据进行归一化处理，预处理后的特征向量进行随机分割，80%用于训练，20%用于测试，采用 10 折交叉验证，随机将训练数据集划分为 10 份，其中 9 份用于训练，剩下的 1 份用于验证。以准确率作为参数优化的评价参数，根据这些独立且不同的数据子集得到的模型性能评价结果，计算出平均性能，这样可以降低对数据的敏感性，提高模型的泛化能力，结果如图 5 所示。



a. 学习曲线 a. Learning curve　b. 验证曲线 b. Validation curve

图 5 学习曲线和验证曲线

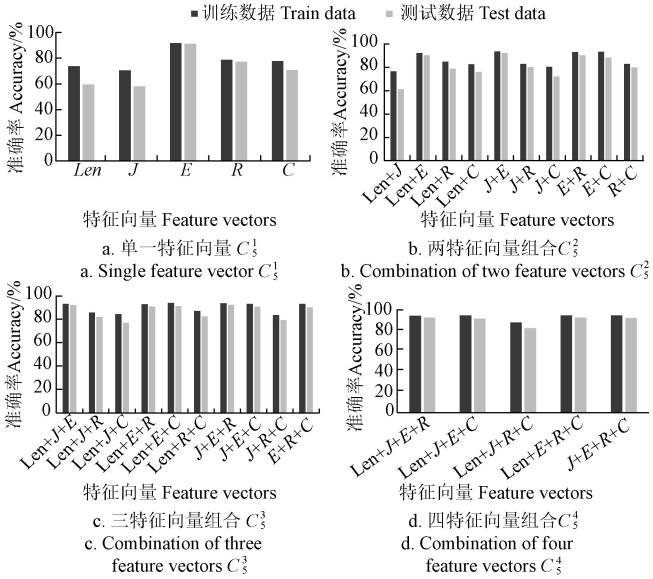
Fig.5 Learning curve and validation curve

从图 5a 学习曲线可以看出训练数据集和测试数据集的准确率之间有较小差距，存在轻微过拟合现象，输入样本数据集，执行网格搜索程序后，最终输出的最佳参数组合是：惩罚系数 C=10，核函数 kernel=径向基函数（radial basis function，rbf），核参数 gamma=0.1，从图 5b 验证曲线也可以看出惩罚系数 C 的最优值在 10 左右，设置过低时，会导致分类准确率下降，设置过高会导致过拟合现象更加严重。

3 试验与结果分析

3.1 特征向量组合试验结果与分析

通过参数优化得到逻辑回归、决策树和支持向量机 3 种算法的最优模型，根据随机森林判定的特征重要性，选择了圆形度 E、矩形度 R、线性度 Len、周长 C 和紧凑度 J 5 种特征向量组合建立了样本特征数据集，在逻辑回归、决策树和支持向量机 3 种算法模型中验证 C_5^1、C_5^2、C_5^3、C_5^4 4 种形式特征向量组合的分类结果，以 3 种算法准确率 Accuracy 平均值作为评价指标，各特征向量组合效果如图 6 所示



a. 单一特征向量 C_5^1　a. Single feature vector C_5^1

b. 两特征向量组合C_5^2　b. Combination of two feature vectors C_5^2

c. 三特征向量组合 C_5^3　c. Combination of three feature vectors C_5^3

d. 四特征向量组合C_5^4　d. Combination of four feature vectors C_5^4

注：C_5^1、C_5^2、C_5^3、C_5^4 分别表示 E、R、Len、C 和 J 5 个不同的特征向量中取出 1、2、3、4 个特征的组合。

Note: C_5^1, C_5^2, C_5^3, C_5^4 represent the combinations of one, two, three and four features from the five different eigenvectors of E, R, Len, C and J.

图 6 C_5^1、C_5^2、C_5^3、C_5^4 分类结果

Fig.6 Classification results of C_5^1, C_5^2, C_5^3, C_5^4

试验结果表明：1）从单一特征向量 C_5^1 组合测试集分类结果看，单个特征 3 种算法分类后的准确率均值大小与特征重要性基本呈正相关，圆形度 E 特征在训练数据集和测试数据集分类正确率均值最高，分别达到 91.6%和 91.1%，紧凑度 J 特征在训练数据集和测试数据集上分类正确率最低，分别为 70.5%和 58.1%；2）C_5^1、C_5^2、C_5^3、C_5^4 4 种形式特征向量组合在测试集上的最高准确率均值分别为：91.1%、92.02%、92.13%、92.27%，随着特征向量个数的增加，测试集上的最高准确率均值也在不断增加；3）特征选择对最终算法的评估有着较大的影响，不同算法对相同特征的表现情况会有所差异。

3.2　不同分类算法分选试验结果与分析

选择圆形度 E、矩形度 R、线性度 Len、周长 C 和紧凑度 J，5 种特征向量建立了样本特征数据集，在 3 种最优模型条件下得到完整训练集和测试集的评价指标得分如表 2 所示，3 种不同分类算法在训练集与测试集上分类结果偏差如图 7 所示。

表 2　3 种不同分类算法不同评价指标的得分表

Table 2　Score of three different classification algorithms with different evaluation indicator

项目 Item	训练数据集 Train dataset				测试数据集 Test dataset			
	准确率 Accuracy	真正率 Precision	召回率 Recall	调和平均数 Harmonic average $F1$	准确率 Accuracy	真正率 Precision	召回率 Recall	调和平均数 Harmonic average $F1$
逻辑回归 Logistic regression	0.909	0.923	0.941	0.932	0.917	0.914	0.946	0.929
决策树 Decision tree	0.970	0.977	0.977	0.977	0.917	0.926	0.946	0.935
支持向量机 Support vector machine	0.969	0.965	0.982	0.977	0.938	0.931	0.964	0.947
均值 Mean	0.949	0.955	0.967	0.962	0.924	0.927	0.952	0.937

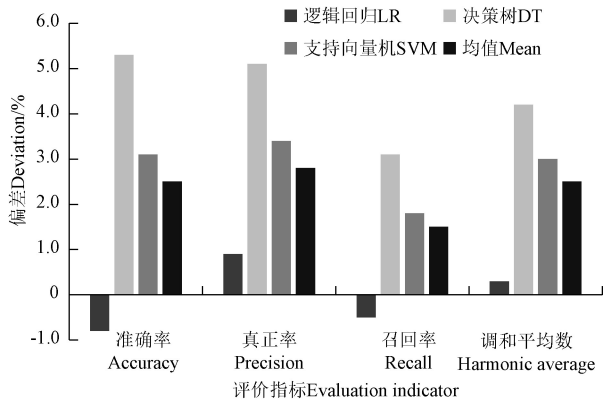


图 7　3 种不同分类算法在训练集与测试集上分类结果偏差

Fig.7　Deviation of classification results from three different classification algorithms on train and test dataset

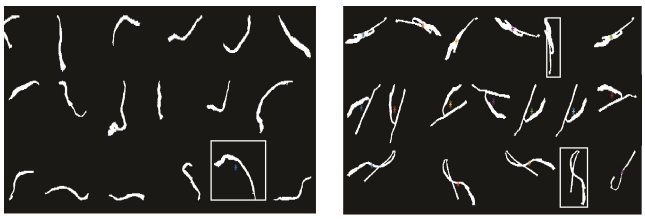
试验结果表明：1）如表 2 所示 3 种不同分类算法的训练数据集准确率、真正率、召回率和调和平均数 $F1$ 平均得分都达到了 0.95 左右，测试数据集准确率、真正率、召回率和 $F1$ 平均得分都超过了 0.92，说明建立的大红袍形态特征描绘子具有一定的可分性，效果较佳，从图 6、图 7 也可以看出，所选 5 种特征用于分类时最高准确率比单一特征向量分类时最高正确率、最低正确率分别提高了 1%和 36.2%，说明特征描绘子的数量及重要性会对分类结果产生重要影响,多特征向量条件下进行特征选择和模型比较可以实现模型快速筛选,有效减少算法开发时间；

2）从 3 种不同类型分类算法在样本形态特征数据集上的分类结果看，支持向量机算法的效果最好，测试试验结果中准确率和 $F1$ 达到了 93.8%和 94.7%，而且其在训练数据集、测试数据集准确率、真正率、召回率和 $F1$ 得分均超过了 0.965 和 0.93，但支持向量机算法在训练集和测试集上的 4 种评价指标得分偏差均大于 2%，准确率提高的过程也伴随着过拟合风险的增大；

3）从准确率和召回率这 2 个评价指标上看，逻辑回归和决策树在测试集得分均相同，但其他 2 个指标决策树算法都略高于逻辑回归算法，从这一结果我们可以看出多个评价指标更有利于我们选出最佳的分类算法。

4）从图 7 中 3 种不同分类算法不同评价指标训练集与测试集分类得分偏差我们可以看出，逻辑回归算法的泛化能力更强，决策树算法产生过拟合的风险更大，而从表 2 我们得出逻辑回归算法的得分最低，支持向量机的得分最高，所以在评价特征向量可分性时，可以选择多个多种算法评价结果均值作为最终的评判依据。

为更好地分析分类算法的效果，以及分析茶叶形态特征描绘子对分类的影响，算法开发过程中对样本图像做了可视化处理，对判定为不良品的样本进行形心点标记，这也是模拟动态检测过程，通过给定的形心点位置，可以有效剔除不良品。如图 8 所示，图 8a 叶样本有 1 个样本被标记，即被误识别为梗；图 8b 梗样本有 2 处未被标记的，即梗未被识别出来，由于样本形态太过复杂多样，在分类过程中还是存在少数样本被误分的情况。



a. 叶识别结果　a. Leaves'test results

b. 梗识别结果　b. Stalks' test results

图 8　测试结果

Fig.8　Test results

4 结 论

本文基于茶叶形态特征建立了一种多特征向量下茶叶良品与不良品分选快速建模的方法，验证了逻辑回归、决策树和支持向量机 3 种分类算法在大红袍样本数据集上准确率、真正率、召回率和 *F*1，4 种评价指标的评价效果，最终的试验结果表明：

1）采用随机森林算法进行特征重要性判定，在多特征向量下选择圆形度 *E*、矩形度 *R*、线性度 Len、周长 *C* 和紧凑度 *J*，5 种形态特征向量进行大红袍良品与不良品分选，效果明显；

2）在特征选择过程中，利用多种分类算法叠加验证结果更准确；

3）3 种算法中支持向量机算法分类效果最佳，但在试验中也发现，分类算法模型的参数选择对最终结果影响较大，采用交叉验证等参数优化方式可以有效提高模型的泛化能力和分类准确率；

4）该方法也可快速开发其他类型的茶叶精选算法模型，试验中选择的是大红袍动态下落过程中的图片，符合茶叶精选过程的实际工况，可推广到茶叶实际生产的精加工过程中。

[参 考 文 献]

[1] 刘跃云. 夏秋绿茶色泽提升技术研究[D]. 重庆：西南大学，2011.

[2] 彭江南，谢宗铭，杨丽明，等. 基于 Seed Identification 软件的棉籽机器视觉快速精选[J]. 农业工程学报，2013，29(23)：147－152.
Peng Jiangnan, Xie Zongming, Yang Liming, et al. Rapid selection of cottonseed machine vision based on seed identification software[J].Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE), 2013, 29(23): 147－152. (in Chinese with English abstract)

[3] Kurtulmus F, Alibas I, Kavdir I. Classification of pepper seeds using machine vision based on neural network[J]. International Journal of Agricultural & Biological Engineering, 2016, 9(1): 51－62.

[4] Wang Weilin, Li Changying. A multimodal machine vision system for quality inspection of onions[J]. Journal of Food Engineering, 2015, 166: 291－301.

[5] 王红军，熊俊涛，黎邹邹，等. 基于机器视觉图像特征参数的马铃薯质量和形状分级方法[J]. 农业工程学报，2016，32(8)：272－277.
Wang Hongjun, Xiong Juntao, Li Zouzou, et al. Potato grading method of weight and shape based on imaging characteristics parameters in machine vision system[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(8): 272－277. (in Chinese with English abstract)

[6] 杨福增，杨亮亮，田艳娜，等. 基于颜色和形状特征的茶叶嫩芽识别方法[J]. 农业机械学报，2009，40(增刊 1)：119－123.
Yang Fuzeng, Yang Liangliang, Tian Yanna, et al, Recognition of the tea sprout based on color and shape features[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(Supp.1): 119－123. (in Chinese with English abstract)

[7] 董春旺，朱宏凯，周小芬，等. 基于机器视觉和工艺参数的针芽形绿茶外形品质评价[J]. 农业机械学报，2017，48(9)：38－45.
Dong Chunwang, Zhu Hongkai, Zhou Xiaofen, et al. Quality evaluation for appearance of needle green tea based on machine vision and process parameters[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(9): 38－45. (in Chinese with English abstract)

[8] 宋彦，谢汉垒，宁井铭，等. 基于机器视觉形状特征参数的祁门红茶等级识别[J]. 农业工程学报，2018，34(23)：279－286.
Song Yan, Xie Hanlei, Ning Jingming, et al. Grading Keemun black tea based on shape feature parameters of machine vision[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(23): 279－286. (in Chinese with English abstract)

[9] Borah S, Hines E L, Bhuyan M. Wavelet transform based image texture analysis for size estimation applied to the sorting of tea granules[J]. Journal of Food Engineering, 2007, 79(2): 629－639.

[10] Laddi A, Sharma S, Kumar A, et al. Classification of tea grains based upon image texture feature analysis under different illumination conditions[J]. Journal of Food Engineering, 2013, 115(2): 226－231.

[11] Tang Zhe, Su Yuancheng, Er M J, et al. A local binary pattern based texture descriptors for classification of tea leaves[J]. Neurocomputing, 2015, 168(30): 1011－1023.

[12] Cimpoiu C, Cristea V M ,Hosu A, et al. Antioxidant activity prediction and classification of some teas using artificial neural networks[J]. Food Chemistry, 2011, 127(3): 1323－1328.

[13] 张春燕，陈笋，张俊峰，等. 基于最小风险贝叶斯分类器的茶叶茶梗分类[J]. 计算机工程与应用，2012，48(28)：187－192，239.
Zhang Chunyan, Chen Sun, Zhang Junfeng, et al. Classification of tea and stalk based on minimum risk Bayesian classifier[J]. Computer Engineering and Applications, 2012, 48(28): 187－192, 239. (in Chinese with English abstract)

[14] 高达睿. 基于颜色和形状特征的茶叶分选研究[D]. 合肥：中国科学技术大学，2016.
Gao Darui. Rsearch on the Tea Sorting Based on Characteristic of Color and Shape[D]. Hefei: University of Science and Technology of China, 2016. (in Chinese with English abstract)

[15] 刘希. 基于彩色线阵 CCD 的茶叶分选控制系统设计[D]. 南京：南京林业大学，2014.
Liu Xi. The Design of Tea Sorter Control System Based on Color Linear CCD[D]. Nanjing: Nanjing Forestry University, 2014. (in Chinese with English abstract)

[16] Sebastion Rasch. Python Machine Learning[M]. 高明等译. 北京：机械工业出版社，2017.

[17] Breimen L. Random Forests[J]. Machine Learning, 2001, 45(1)：5－32.

[18] 徐少成，李东喜. 基于随机森林的加权特征选择算法[J]. 统计与决策，2018，34(18)：25－28.
Xu Shaocheng, Li Dongxi. Weighted feature selection algorithm based on random forest [J]. Statistics & Decision, 2018, 34(18): 25－28. (in Chinese with English abstract)

[19] Strobl C, Boulesteix A L, Kneib T, et al. Conditional variable importance for random forests[J]. BMC Bioinformatics, 2008, 9(1): 1－11.

[20] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of newtests[J]. Pattern Recognition, 2014, 44(2): 330－349.

[21] Powers, David M W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation[J]. Journal of Machine Learning Technologies, 2011, 2(1): 37－63.

[22] 金志刚，苏菲. 基于 FSVM 与多类逻辑回归的两级入侵检测模型[J]. 南开大学学报：自然科学版，2018，51(3)：1－6.
Jin Zhigang, Su Fei. A two-stage model intrusion detection system based on SVM and multi-class logistic regression[J]. Acta Scientiarum Naturalium Universitatis Nankaiensis, 2018, 51(3): 1－6. (in Chinese with English abstract)
[23] 刘敏洁，许旺，王建华，等. 基于人工神经网络和二元逻辑回归的甜玉米种子生活力检测模型研究[J]. 中国农业大学学报，2018，23(7)：1－10.
Liu Minjie, Xu Xuan, Wang Jianhua, et al. Seed viability testing model of sweet corn based on artificial neural network and binary logisitic regression[J]. Journal of China Agricultural University, 2018, 23(7): 1－10. (in Chinese with English abstract)
[24] Chandra B, Kothari R, Paul P. A new node splitting measure for decision tree construction[J]. Pattern Recognition, 2010, 43(8): 2725－2731.
[25] Liu W, Chawla S, Cieslak D A, et al. A Robust decision tree algorithm for imbalanced data sets[C]//Proceedings of the SIAM International Conference on Data Mining. America: SIAM, 2010, 766－777.
[26] Umano M, Okamolo H, Hatono I, et al. Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis system[C]//Proceedings of the 3 IEEE International Conference on Fuzzy Systems. New York: IEEE Press, 1994, 3: 2113－2118.
[27] Ju Hongyun, Zhang Junben, Li Chaofeng et al. Remote sensing image based on *K*-means and SVM automatic classification method[J]. Application Research of computers, 2007, 24(11): 318－320.
[28] Ma Jiajun, Zhou Shuisheng, Li Chen, et al. A sparse robust model for large scale multi-class classification based on K-SVCR[J]. Pattern Recognition Letters, 2019, 117: 16－23.
[29] Zhang J, Zhang P, Li Z. Fuzzy support vector machine based on color modeling for facial complexion recognition in traditional chinese medicine[J]. Chinese Journal of Electronics, 2016, 25(3): 474－480.
[30] Nasiri J A, Charkari N M, Jalili S. Least squares twin multi-class classification support vector machine [J]. Pattern Recognition, 2015, 48(3): 984－992.

Tea selection method based on morphology feature parameters

Wu Zhengmin, Cao Chengmao※, Wang Errui, Luo Kun, Zhang Jinyan, Sun Yan

(*College of Engineering, Anhui Agricultural University*, *Hefei* 230036, *China*)

Abstract: The color between stalks and leaves of tea in summer and autumn is similar, which means the traditional color sorter is difficult to sort based on optical characteristics. To realize the rapid modeling of tea selection algorithm and improve the sorting accuracy, a method for sorting the fine and bad products of tea by multi-feature vectors based on the morphological characteristics was introduced in this paper. First, Wuyishan Dahongpao tea was selected as a test sample to collect images during the dynamic drop process. The blue element image was extracted, and single sample's binary image and edge were obtained by analysis of whole image connection area. Then, feature extraction program was developed based on image processing algorithm to extract morphological feature parameters of the tea samples automatically. Four simple shape descriptors-the sample perimeter, area, the length and width of minimum bounding rectangle were extracted. On this basis, eight complex shape descriptors-circularity, rectangularity, linearity, slightness, diameter, diagonal of minimum bounding rectangle, compactness and centroid were calculated. In addition, the random forest algorithm was used to determine the above features weight, the feature was selected according to weight threshold. Finally, logistic regression (LR), decision tree (DT) and support vector machine (SVM) that three different classification algorithms were established to classify the samples, verify the validity of the features and analyze the effects of different classification algorithms on the classification of tea. The original data were normalized and randomly segmented 80% used for training, 20% for testing. 10-fold cross-validation was used to select the optimal parameters of the classification model, and the training dataset was randomly divided into 10 parts, of which 9 parts were used for training, and the remaining 1 part was used for verification. According to the above machine learning system parameter optimization process to obtain the logical regression, decision tree and support vector machine optimal model, and statistical the final evaluation results on test dataset. The test results showed that: 1) The circularity weight was the highest, at 0.467, and five eigenvectors of circularity, rectangularity, linearity, perimeter and compactness were finally selected with the weight threshold value which was 0.05; 2) In the test dataset, the average accuracy *F*1 of the three classification algorithms was 0.924, suggesting that the established tea morphological feature descriptors has certain separability and better effect; 3)When testing test-dataset, the accuracy score was 91.7% and *F*1 score of logistic regression (LR) was 92.9%, the accuracy score was 91.7% and *F*1 score of support vector machine (SVM) was 94.7%.Support vector machine (SVM) algorithm was the best recognition effect in three classification algorithms; 4) From three different classification algorithms assessment score deviation, we can see that the generalization ability of the logic regression algorithm was stronger, the decision tree algorithm has a greater risk of over fitting. We get the lowest accuracy and *F*1 score of the logistic regression algorithm, while the support vector machine accuracy and *F*1 score were the highest, so in the evaluation of eigenvector comparability, multiple algorithms can be selected to evaluate the results of the average as the final basis for evaluation. In the experiment, we acquired dynamic image, which stay in line with the actual working conditions of the tea selection process, and can be extended to the actual processing of tea production.

Keywords: morphology; decision tree; support vector machine; logistic regression; random forest; tea