doi:10.6041/j.issn.1000-1298.2017.04.040

基于高光谱成像的苹果品种快速鉴别

马惠玲[1]　王若琳[1]　蔡　骋[2]　王　栋[1]

(1.西北农林科技大学生命科学学院,陕西杨凌 712100;2.西北农林科技大学信息工程学院,陕西杨凌 712100)

摘要:以“乔纳金”苹果,“红富士”苹果和“秦冠”苹果共 90 个试验样本为试材分别采集 865 ~ 1 711 nm 的近红外波段高光谱图像,选取苹果图像感兴趣区域(ROI),以分辨率 2.8 nm 提取其平均反射光谱数据,分别利用 K 近邻法(KNN)和径向基核函数支持向量机(RBF - SVM)进行品种判别,5 折交叉检验。结果表明,3 种苹果的近红外高光谱图像均在波长 941 ~ 1 602 nm 之间变得清晰,该区域 200 个波段下的平均反射光谱数据经 KNN 法中的 10 种距离算法评判,当 K 取值 3 和 5 时,切比雪夫距离、欧几里得距离和明可夫斯基距离 3 种距离算法的识别正确率均达到 100%;SVM - RBF 核函数模型中,γ 取值为 2^{-8} ~ 1 的范围内识别正确率均在 92% 以上,当 γ 取值 2^{-5},C 取值为 16 和 32 时,识别正确率最高,为 96.67%。故利用近红外高光谱图像技术结合 KNN 计算对苹果品种进行快速鉴别是优异和可靠的方案。

关键词:苹果;品种鉴别;高光谱成像;K 近邻法;支持向量机

中图分类号:S661.1　　**文献标识码**:A　　**文章编号**:1000-1298(2017)04-0305-08

Rapid Identification of Apple Varieties Based on Hyperspectral Imaging

MA Huiling[1]　WANG Ruolin[1]　CAI Cheng[2]　WANG Dong[1]

(1. *College of Life Science*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*
2. *College of Information Engineering*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*)

Abstract: In order to achieve rapid non-destructive identification of apple varieties, the methodology of near-infrared hyperspectral imaging on identification of apple varieties was investigated. Near infrared hyperspectral images with wavelength from 865 ~ 1 711 nm of total 90 sample fruits were collected from three different varieties ("Jonagold", "Fuji" and "Qinguan" apples), and hyperspectral image area of the apple was selected as a region of interest (ROI). Reflection intensity data of the average reflex spectrum were extracted with resolution rate of 2.8 nm, then they were calculated with K-nearest neighbor (KNN) and the support vector machine (SVM) methods, respectively, which were checked with 5-fold cross-validation method. The results showed that the hyperspectral images of three varieties of apples all became clear within wavelength of 941 ~ 1 602 nm. Among ten distance-types' judgment of KNN with average reflection intensity at 200 wavelength-points, the identification accuracy of Chebychev, Euclidean and Minkowski reached the highest of 100% when the parameter K was set at 3 or 5. While using the support vector machine-radial basis function (SVM - RBF) model, the accuracy rate reached above 92% when the value of γ fell within 2^{-8} ~ 1. The highest recognition rate of this model reached 96.67% when γ was set at 2^{-5} and C took the value of 16 amd 32 at the same time. The results demonstrated that near-infrared hyperspectral imaging in combination with KNN was excellent and reliable for the rapid identification of apple varieties. This method could provide reference for identifying apple varieties in production.

Key words: apple; variety identification; hyperspectral image; K-nearest neighbor method; support vector machine

引言

苹果在世界果品市场上占据着重要地位,有些苹果品种成熟期、外形、色泽、风味差异较大,感观易于辨别;有些则成熟期、大小、着色、口感均比较接近,非专业人员品评辨识准确率不高,更难以在采后

收稿日期:2016 - 08 - 12　修回日期:2016 - 09 - 21

基金项目:陕西省农业科技创新与攻关项目(2015NY023)和农业部现代苹果产业技术体系项目(CARS - 28)

作者简介:马惠玲(1965—),女,教授,博士生导师,主要从事果实采后生理与技术研究,E-mail: ma_huiling65@ hotmail.com

处理流水线上快速鉴别。不同品种的田间栽培条件、采后贮藏特性、价格均差别较大,急需一种简便、无损的苹果品种快速鉴别方法。

目前,国内外学者在鉴别物质品种的研究方面主要采用近红外光谱技术(NIRS),根据全谱段或多波长下的光谱数据对待测物进行定性或定量分析,从而反映待测物的类型或有机物的含量[1]。有学者利用近红外光谱技术对葡萄[2]、梨[3]、苹果[4]、草莓[5]、鲜枣[6]和杨梅[7]等水果品种进行鉴别。此外,还有对水稻[8]、玉米[9-10]和小麦[11]等经济作物品种鉴别的研究报道。然而,近红外光谱技术只提供对检验客体某一较小区域的光谱信息,无法得到样本的空间信息,因而存在检测范围小和信息获取量少的弊端。

高光谱成像技术结合了传统成像技术和现代光谱技术,能够在更大范围内同时获取样本的光谱信息和空间信息,对于生物材料的光谱学特征差别识别敏感度高,因此能全面地反映被测物的种类及品质[12-13]。张初等[14]采用高光谱成像技术对西瓜种子品种进行鉴别,发现经过 SG 平滑算法对光谱进行预处理后,基于特征波长选择所建立的极限学习机(ELM)模型取得了最佳效果,建模集和预测集的识别准确率均达到 100%;WANG 等[15]利用高光谱成像对稻米品种进行鉴别,发现基于数据融合的逆反馈人工神经网络(BPNN)对稻米品种的识别率最高,为 94.45%,优于单独基于光谱数据(89.91%)或图像数据(88.09%)的结果。LIU 等[16]利用高光谱成像对 3 个荔枝品种进行分类,得出支持向量机(SVM)分类模型对校正集和预测集的识别率分别为 100% 和 87.81%。孙俊等[17]利用高光谱成像技术对 3 个地区的红豆样本进行分类,发现引入遗传算法(GA)优化的概率神经网络(PNN)模型识别效果最佳,识别正确率达到了 97.5%。

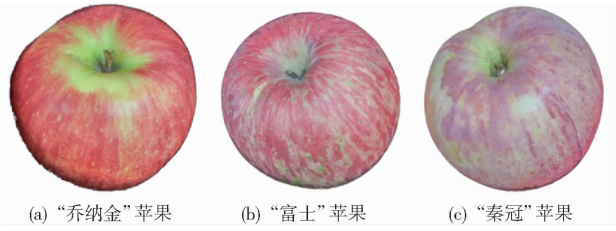
本文采用上市期有交叉,并均着红色的“乔纳金”、“红富士”和“秦冠”3 个品种的苹果为试材,逐一采用近红外高光谱成像技术获得各苹果的光谱数据,提取并分析各个苹果感兴趣区域(ROI)的光谱反射平均值,结合 K 近邻算法和支持向量机的建模方法,实现对苹果品种的分类鉴别,以期为近红外高光谱成像技术鉴别苹果品种的实际应用提供参考。

1 材料与方法

1.1 试验材料、设备及软件

1.1.1 试验材料

试验用苹果分 3 个品种(图 1),分别于商品成熟期采收:“乔纳金”苹果于 2015 年 9 月 20 日采于陕西省白水县西北农林科技大学苹果试验示范站果园;“红富士”苹果于 2015 年 10 月 14 日也采自该示范站果园;“秦冠”苹果于 2015 年 10 月 20 日采于陕西省宝鸡市扶风县段家湾镇青龙村农家果园。采收后各品种均选取大小相近、着色均匀、无机械损伤的果实 30 个,当天运回西北农林科技大学实验室,置于室温(20℃)下 24 h。逐一编号,1~30 号代表“乔纳金”苹果,种类代号为 1;31~60 代表“红富士”苹果,种类代号为 2;61~90 号代表“秦冠”苹果,种类代号为 3。此编号为后续划分训练样本和测试样本做准备。



(a) “乔纳金”苹果 (b) “富士”苹果 (c) “秦冠”苹果

图 1 苹果样本

Fig. 1 Apple samples

1.1.2 试验设备

高光谱图像采集系统结构如图 2 所示。该系统由 1 台 ImSpector N17E 型近红外光谱仪(芬兰 Spectral Imaging Ltd. 公司),1 台像素为 320×256 的 XEVA2616 型面阵 CCD 相机(比利时 XenICs Ltd. 公司),1 套高稳定性的卤钨灯白光光源,1 台高精度的电控平移台装置和计算机等组成。近红外高光谱成像系统采集光谱范围为 865~1 711 nm,光谱分辨率 2.8 nm,入射光狭缝宽度为 30 μm。

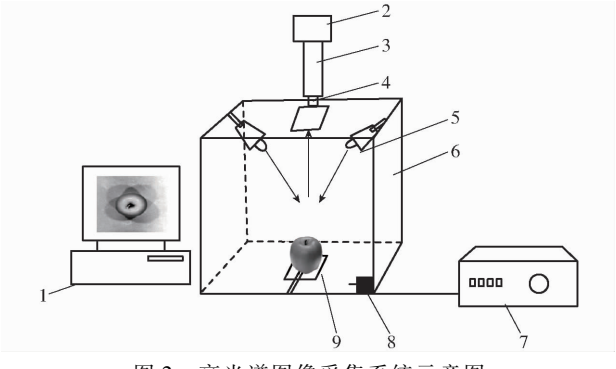


图 2 高光谱图像采集系统示意图

Fig. 2 Schematic of hyperspectral imaging system

1. 计算机 2. CCD 相机 3. 光谱仪 4. 镜头 5. 光源 6. 暗箱 7. 光源控制器 8. 电动机 9. 电控移动载物台

1.1.3 软件

高光谱图像采集由 Spectral SENS-V17E 软件(英国 Gilden Photonics Ltd. 公司)完成,后续的图像数据处理采用 ENVI 4.7 软件(美国 Research System 公司),光谱数据分析软件为 Matlab 7.4(美国 MathWorks 公司)。

1.2　试验方法

1.2.1　高光谱图像的采集

在高光谱图像数据采集前需要对镜头焦距进行调整，保证平台移动速度、相机曝光时间和光源相互匹配以确保采集图像清晰不失真[18]。经过多次调整及参数优化，最终确定高光谱摄像机的各参数为：平台移动速度 20 mm/s，相机的曝光时间 10 ms，物距 200 mm，采集到的高光谱图像块尺寸为 320 × 256 × 255（像素 × 波段数，即每个样品采集 255 个波段下像素为 320 × 256 的二维图像）。采集光谱时，每次将 1 个苹果样本放于载物台上，保持果柄-花萼轴与载物台垂直，使载物台中心对准相机。当平台移动时，近红外高光谱仪从上往下扫描样本，每次扫描得到 1 行图像的光谱信息，平台带动样本运动的过程中，获取其他位置直至整个样本的光谱信息。为了消除果实果柄面和果顶面形状差异对光谱采集的影响，对每个苹果样品均进行 2 次光谱采集，第 1 次采集果柄面，第 2 次采集果顶面，将 2 次光谱反射平均值作为该样品的光谱数据。以同样的方法依次采集 90 个苹果样品的近红外高光谱图像。这样，就得到了每个果实的物理和化学因素综合影响下的高光谱谱图（相当于各品种的高光谱指纹图谱）。

1.2.2　光谱校正

为了消除因苹果形状差异、光源强度在各波段下分布不均以及摄像头中暗电流存在而产生的噪声影响，需对获得的高光谱图像进行黑白标定[19]。首先对反射率为 99% 的标准白色校正板进行图像采集，得到全白的标定图像 W，然后关闭光源拧上镜头盖，采集全黑标定图像 D。图像校正公式为

$$R = \frac{R_0 - D}{W - D} \tag{1}$$

式中　R_0——原始的高光谱图像

R——校正后的图像

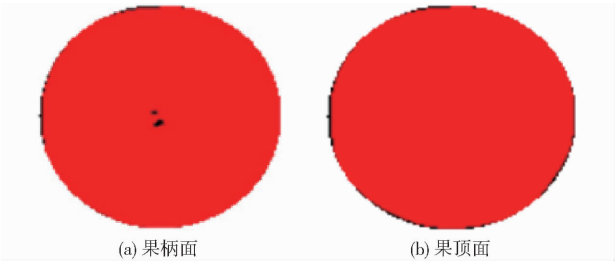
1.2.3　感兴趣区域的选取

利用 ENVI 4.7 软件，对图像经过滤波和形态学算法处理后，选取苹果的果柄面和果顶面分别作为感兴趣区域（Range of interest，ROI），如图 3 所示，并求 ROI 的反射光谱曲线。将两面的反射光谱曲线进行数据收集并取反射光谱平均值作为数据集合。

1.2.4　苹果理化特征的测定

（1）果实硬度

采用 GY-3 型果蔬硬度计（意大利）。各果实沿赤道线选取 2 个等分点，各点削去约 1.5 cm^2 的果皮，匀速插入直径 0.8 cm 的端头至刻度线（1 cm），记录最大穿透力，2 个点的值取平均后记为每个果的硬度。



(a) 果柄面　　(b) 果顶面

图 3　苹果果实 ROI 的选取

Fig. 3　ROI region selection of apple fruit

（2）可溶性固形物

采用 WY032T 型折光仪测量待测样液的折光率，在折光仪上直接读出可溶性固形物质量分数并重复 3 次。

（3）可滴定酸质量分数

称取 100 g 果肉榨成苹果汁并测量体积，每次取 5 mL 的苹果汁，定容至 100 mL 容量瓶中，取 25 mL 滤液加入酚酞指示剂，用 0.1 mol/L NaOH 溶液滴定至溶液颜色呈淡红色，且维持 1 min 不褪色，记录消耗的 NaOH 溶液体积。可滴定酸质量分数计算公式为

$$T = \frac{4NV_2VA}{WV_1} \times 100\%$$

式中　N——NaOH 溶液浓度，mol/L

W——样品鲜质量，g

V——样品液制成总体积，mL

V_1——吸取样品滤液体积，mL

V_2——滴定时消耗 NaOH 溶液用量，mL

A——与 1.00 mL NaOH 标准溶液相当的试样主体酸质量，g

1.2.5　不同品种的区分方法

K 近邻（K-nearest neighbor，KNN）算法的基本思想很直观，即对若干类样本，按照高维空间的距离度量，搜索出最相似的 K 个近邻，然后对近邻样本所属的类别标签进行分析，判断待测样本的类别[20]。由于不需要先验训练用于区分任意 2 个类别间的判别子分类器，并且支持增量学习等优良特性，KNN 算法具有很强的适应性。在采样数据较为充分的情况下，KNN 算法能够取得优秀的分类性能，具体表现在分类结果取决于近邻样本的类别标签，因此不会受到少量噪声的干扰，这大大提高了算法对噪声的鲁棒性和抗变换性。

在已知待测样本 K 个近邻的类别时，待测样本的类别有不同的判断方法。最直观的方法是找出这 K 个近邻中出现次数最多的类别来作为待测样本的类别。更完善的方法是将近邻与待测样本间的距离作为权重来计算这 K 个近邻所属类别的重要程度，加权统计得到最重要的类别来作为待测样本的类

别。

样本在特征空间中的相似程度由距离度量的定义来决定,着重点在特征空间中距离时,可以选用欧氏距离和城市街区距离等;着重于矢量的夹角时,则选用余弦相似度量更为合理。对于苹果光谱这样内部作用机理复杂的信号而言,难以先验地获知哪一种距离度量更为合理,常见的做法是对于可能的距离度量方案通过实证进行筛选,以获取能够与模型最吻合的距离度量。对于 n 维空间中的 2 个样本 $\boldsymbol{X}(X_1, X_2, \cdots, X_n)$ 和 $\boldsymbol{Y}(Y_1, Y_2, \cdots, Y_n)$,本文采用 Matlab 函数中内置的 10 种距离度量方法,其计算方法如下:

城市街区距离

$$d_{\text{cityblock}}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{n} |X_i - Y_i| \tag{2}$$

切比雪夫距离

$$d_{\text{Chebychev}}(\boldsymbol{X},\boldsymbol{Y}) = \max|X_i - Y_i| \tag{3}$$

相关距离

$$d_{\text{correlation}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{C_{\text{OV}}(\boldsymbol{X},\boldsymbol{Y})}{V_{\text{AR}}(\boldsymbol{X})V_{\text{AR}}(\boldsymbol{Y})} \tag{4}$$

式中 C_{OV}——协方差 V_{AR}——方差

余弦相似度

$$d_{\text{cosine}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{5}$$

欧几里得距离

$$d_{\text{Euclidean}}(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{6}$$

汉明距离

$$d_{\text{Hamming}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\#(X_i \neq Y_i)}{n} \tag{7}$$

式中,#代表将点转换为二进制数据,下同。

Jaccard 相似系数

$$d_{\text{Jaccard}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\#[(X_i \neq Y_i)] \cap ((X_i \neq 0) \cup (Y_i \neq 0))}{\#((X_i \neq 0) \cup (Y_i \neq 0))} \tag{8}$$

明可夫斯基距离

$$d_{\text{Minkowski}}(\boldsymbol{X},\boldsymbol{Y}) = \sqrt[p]{\sum_{i=1}^{n} (X_i - Y_i)^p} \tag{9}$$

由于式(9)代表对多个距离度量公式的概括,故 p 为某一个变量。

标准化欧氏距离

$$d_{\text{seuclidean}}(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{V_{\text{AR}}(X)} - \frac{Y_i - \overline{Y}}{V_{\text{AR}}(Y)}\right)^2} \tag{10}$$

斯皮尔曼距离(Spearman distance)

$$d_{\text{Spearman}}(\boldsymbol{X},\boldsymbol{Y}) = 1 - \frac{\sum_{i=1}^{n} \left(\text{rank}(X_i) - \frac{n+1}{2}\right)\left(\text{rank}(Y_i) - \frac{n+1}{2}\right)}{\sum_{i=1}^{n} \left(\text{rank}(X_i) - \frac{n+1}{2}\right)^2 \sum_{i=1}^{n} \left(\text{rank}(Y_i) - \frac{n+1}{2}\right)^2} \tag{11}$$

因为 KNN 算法采用了待测样本与其近邻间的距离来衡量二者间的相似度,所以该算法对数据的数值范围比较敏感,如果对特征值进行距离计算,则取值范围较大的特征值将对结果产生绝对的影响,而较小的特征值几乎没有作用,这是不合理的。因此需要先对样本数据进行归一化后再进行距离的计算和判断。在实际应用中,如果某种特征的确比其他特征对分类起到更重要作用,则可以根据实际情况增大其数值范围。

虽然 KNN 算法不需要消耗时间来对训练样本进行训练,而是直接拿训练样本来对测试样本进行类别判断,但是该算法仍然需要花费较长时间。因为该算法需要计算每个待测样本和全部训练样本间的距离,这在样本个数较多时或者样本维度较高时是很耗时的工作。因此有必要再探讨其他分类方法对苹果品种识别的可行性。

1.2.6 基于径向基核函数的支持向量机算法

支持向量机(Support vector machine,SVM)算法的基本思想是求出 2 个类别的决策面的表达式,进而可以对 2 个类别进行分类,多类别分类需要在二分类的基础上采取一对一或者一对多策略进行计算。

在分类过程中当 2 个类别线性不可分时,需要借助核函数将 2 个类别的数据映射到可分的高维空间中,再计算决策面的表达式。应用较多的核函数有线性核、多项式核、径向基核、Sigmoid 核 4 种,而无论是低维、高维、小样本、大样本等情况径向基核函数均适用,具有较宽的收敛域,是较为理想的分类依据函数[21]。本实验采取了 C 支持向量分类(C-support vector classification)的方法和径向基核函数(Radial basis function,RBF)。利用 SVM 进行模式分类时需要确定 2 个参数:惩罚因子 C 和 RBF 核函数中的半径参数 γ,参数 C 理解为调节优化方向中 2 个指标(间隔、分类准确度)偏好的权重。对于一个基于 RBF 的 SVM,其性能由参数(C, γ)决定,选取不同的 C 和 γ 就会得到不同的 SVM。其中 C 选择的越大,表示对错误率惩罚程度越大,C 较小时推广错误率的估计值比较高,C 较大时急剧降低,即性能得到迅速的提高。通过对参数(C, γ)的变

化，可选取最优参数组合来得到最优值，即此时的错误率最低[21]。所以本文中采用不同的 C 值和 γ 值进行交叉验证来计算分类结果的正确率。

径向基核函数公式为

$$K(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{\phi}(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{Y}) = \mathrm{e}^{-\gamma\|\boldsymbol{X}-\boldsymbol{Y}\|^2} \quad (\gamma>0) \tag{12}$$

式中，K 表示特征空间距离，$\boldsymbol{\phi}(\boldsymbol{X})$ 表示通过将输入空间内线性不可分的数据映射到一个高维的特征空间内使得数据在特征空间内是可分的。其中，虽然每个样本 $\boldsymbol{X}$ 被函数 $\boldsymbol{\phi}(\boldsymbol{X})$ 映射到无限维空间，但核是很容易计算的，从而可以方便地计算决策面的表达式。

1.2.7　交叉验证

交叉验证是指在给定的建模样本中，取大多数样本进行建模，留一小部分样本对刚建立的模型进行预测，并计算出这小部分样本的预测误差，记录它们的平方和。一直重复进行这个过程，直到所有的样本都被预测了一次而且仅被预报一次。j 折交叉验证是指在机器学习中，将数据集 A 分为训练集 B 和测试集 C，在样本量不充足的情况下，为了充分利用数据集对算法效果进行测试，将数据集 A 随机分为 j 个包，每次将其中 1 个包作为测试集，剩余 $j-1$ 个包作为训练集进行训练。这个方法能够同时重复运用同一批样本进行训练和验证，每次的结果验证一次，增加了验证的可信度，本文中采用了 5 折交叉检验的方法对 2 种分类模型进行预测。

2　结果与分析

2.1　光谱信息的提取与分析

根据苹果果实和背景不同的光谱特征，选择 1 069 nm 和 1 446 nm 的图像进行波段比处理，发现苹果果实区域的波段比大于 2，而背景噪声的值小于 1，故将阈值选为 2，进行降噪处理，然后将降噪后的二值图像进行滤波和形态学算法处理后作为掩膜来提取高光谱图像中苹果果实区域作为感兴趣区域，然后计算出 ROI 的平均光谱。

由于高光谱曲线在 865.11 ~ 941.46 nm 以及 1 602.15 ~ 1 711.71 nm 范围中有很大的噪声，导致该部分高光谱图像信息十分模糊，所以截取了 3 种苹果的近红外高光谱图像在波长为 941 ~ 1 602 nm 之间共 200 个波段的数据，即每个样本的维度为 200，即 KNN 算法中的各种距离公式中 $n = 200$，分别对应 200 个波段取值。各品种随机取 10 个果实为例，其 200 个波段内部分波长下反射强度及其随波长的变化趋势如图 4 所示。由图 4 可见，各个品种在选定波长范围内各个波长下的反射强度非常一致，“乔纳金”、“富士”、“秦冠”苹果各 30 个果实的观测值变异系数均低于 5%，依次为 3.2%、3.4%、4.9%。各波长下“乔纳金”苹果的取值与另外 2 种苹果有明显的分离；“富士”与“秦冠”苹果的取值有所交叉，显示了“富士”与“秦冠”苹果果实在各波长下的反射强度取值的相似性和难以区分性。故还需要借助谱图中数据信息，采用分类器计算后来确定二者区分的可能性。

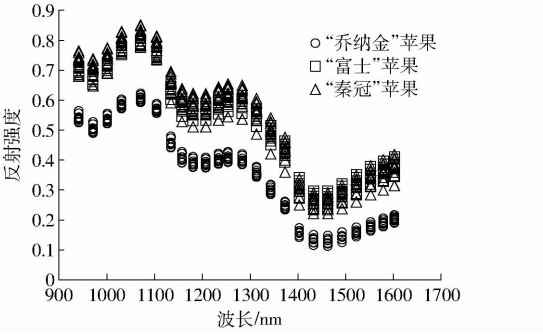


图 4　3 个品种苹果感兴趣区域的部分波长下反射强度

Fig. 4　Reflection intensity of three varieties of apples at part wavelengths in ROI region

2.2　建立 KNN 模型计算苹果品种的识别正确率

KNN 方法在类别决策时，只与极少量的相邻样本有关。由于 KNN 方法主要靠周围有限的邻近样本，而不是靠判别类域的方法来确定所属类别，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。

在 KNN 算法中利用不同的距离方法，在不同的 K 取值条件下，将各个品种的平均反射强度分为训练样本和校正样本并进行计算，苹果品种识别正确率如表 1 所示（K 取值分别为 1、3、5、7、9）。通过表 1 可知，K 选取 1、3、5 时，正确识别率普遍大于 K 取 7、9 的结果。这是因为当 K 值选取过大时，在特征空间中覆盖的局部区域过大，会降低分类器的拟合性能。对不同距离度量的实证中可以发现，K 取 1、3、5 时，基于空间中点之间度量的 5 种距离（城市街区距离、切比雪夫距离、欧几里得距离、明可夫斯基距离、标准欧氏距离），其识别正确率最高，均达到 98% 以上；基于余弦相似度、相关距离和斯皮尔曼距离的识别正确率其次，为 92% ~ 99%；基于集合运算的汉明距离和 Jaccard 相似系数度量性能不佳，识别正确率均低于 40%。说明基于点距离的度量方案可以更好地在光谱特征空间中表征数据样本，刻画出更为精确的苹果类别边界，这是因为苹果果实的高光谱信号值域较大，因此以浮点数来呈现数值时，仅当 2 个浮点数完全一样时才能够在求交集时获得非零值，出于测量误差的缘故，微小的噪声

都会造成样本相似度量的失效,因此基于集合运算的汉明距离和 Jaccard 相似系数不适用于苹果品种 KNN 判别分类。选取 K 值为 3 和 5 时,能够在拟合和泛化性能间获得较好的均衡,距离度量选取切比雪夫距离、欧几里得距离和明可夫斯基距离都达到了正确率 100% 的优异分类性能。

KNN 分类器在判别时采用近邻样本进行分析,因此是原生的多类别分类器,在分析其性能时,通常不再研讨两两类别之间的区分和混叠性能,而是考量多类别总体分类性能。而且在参数寻优后,获得了 100% 的品种判别性能,表明品种之间均具有优异的可分性。

表 1 不同 K 取值与各种距离的识别正确率

Tab. 1 Resolution of apple varieties at different K values and distance types %

K	城市街区距离	切比雪夫距离	相关距离	余弦相似度	欧几里得距离	汉明距离	Jaccard 相似系数	明可夫斯基距离	标准化欧氏距离	斯皮尔曼距离
1	99 ± 0.02	100	99 ± 0.02	93 ± 0.06	99 ± 0.02	40 ± 0.02	40 ± 0.02	99 ± 0.02	100	96 ± 0.05
3	99 ± 0.02	100	94 ± 0.04	94 ± 0.04	100	33	33	100	99 ± 0.02	94 ± 0.07
5	98 ± 0.03	100	96 ± 0.05	92 ± 0.06	100	33	33	100	99 ± 0.02	94 ± 0.07
7	98 ± 0.03	99 ± 0.02	93 ± 0.05	91 ± 0.07	98 ± 0.03	33	33	98 ± 0.03	99 ± 0.02	91 ± 0.07
9	94 ± 0.06	99 ± 0.02	94 ± 0.06	90 ± 0.07	98 ± 0.05	33	33	98 ± 0.05	99 ± 0.02	90 ± 0.05

2.3 建立支持向量机模型计算苹果品种的识别正确率

支持向量机基于核函数机制,能够将非线性的数据映射至高维线性可分的空间中,由于无法根据数据样本先验获得核函数的参数,通常在训练支持向量机中对可能选取的参数空间进行采样,遍历实证,选取最优的参数组合。本试验中选取拟合性能优越的径向基核函数,对惩罚因子 C 和径向基半径 γ 这 2 个参数进行实证寻优,苹果品种识别正确率如表 2 所示。在 γ 取值为 2^{-8} ~ 1 的条件下,在交叉验证中均可以取得 92% 以上的正确率,而随着 γ 取值的增加,拟合性能下降,在交叉验证中的识别正确率降低,故不必再增加 γ 取值来做验证。而其中当 γ 取 2^{-5},C 取 16 和 32 时,在交叉验证中的品种识别正确率最高,可以达到 96.67%,表明该间隔取值适当。

表 2 SVM 模型下不同 C 和 γ 值下的识别正确率

Tab. 2 Resolution of apple varieties at different C and γ values with SVM model %

lbγ	lb$C$						
	-1	0	1	2	3	4	5
-8	93.33	93.33	95.56	94.44	93.33	93.33	93.33
-7	93.33	95.56	94.44	93.33	93.33	93.33	95.56
-6	95.56	96.67	93.33	93.33	93.33	94.44	95.56
-5	95.56	94.44	94.44	93.33	94.44	96.67	96.67
-4	94.44	94.44	94.44	94.44	95.56	95.56	95.56
-3	94.44	94.44	94.44	95.56	95.56	95.56	95.56
-2	93.33	94.44	94.44	94.44	94.44	94.44	94.44
-1	93.33	93.33	93.33	93.33	93.33	93.33	93.33
0	92.22	93.33	93.33	93.33	93.33	93.33	93.33
1	81.11	91.11	92.22	92.22	92.22	92.22	92.22
2	66.67	81.11	81.11	81.11	81.11	81.11	81.11
3	45.56	61.11	63.33	63.33	63.33	63.33	63.33

对于支持向量机分类器,训练时考量的是结构代价函数和分类误差代价函数,因此未能对特定类别之间的区分性能进行呈现和分析,而且由于支持向量机分类性能不及 KNN 分类器,对于实际系统中建议选用 KNN 分类器。既然高光谱判别苹果品种类别实验方案的可行性得以肯定,此处不再对支持向量机中的特定类别间的分类性能进行研讨。

由于近红外图像是由测定目标与背景间的红外线强度差形成的,目标的形状和理化性质均对图像或反射光谱有影响。通过前人的研究,“乔纳金”苹果果形指数为 0.83[22],“富士”苹果为 0.82[23],“秦冠”苹果为 0.84[24];3 种苹果果形指数均在 0.83 左右,表明它们的外观差异并不明显,不是造成三者间近红外反射光谱差异的主要因素。3 个品种的果实硬度(表 3)从小到大虽然依次为“富士”苹果、“乔纳金”苹果、“秦冠”苹果,但是只有“秦冠”苹果与前二者的差异显著,与近红外高光谱中“乔纳金”苹果的反射强度与另外 2 个品种明显可分的差异性也不一致,因此,以硬度反映的果实质地和密度特性也不是影响近红外反射强度的唯一因素。

化合物分子中的含氢基团(OH、NH、CH)在近

表 3 各品种果实理化品质比较

Tab. 3 Comparison of physical-chemical indices for apple fruit of each variety

品种	可溶性固形物质量分数/%	可滴定酸质量分数/%	硬度/(kg·cm^{-2})
乔纳金	12.62 ± 0.29^{a}	0.29 ± 0.03^{a}	8.06 ± 0.87^{a}
富士	14.31 ± 0.95^{b}	0.39 ± 0.04^{b}	7.57 ± 0.76^{a}
秦冠	13.27 ± 1.21^{a}	0.28 ± 0.06^{a}	10.85 ± 1.00^{b}

注:观测值为平均值 ± 标准差,同列中不同字母表示差异显著($p \leqslant 0.05$)。

红外光谱区具有振动合频以及各级倍频的吸收[25]，故碳水化合物、酚类物质对近红外光有吸收。3 个品种果实的可溶性固形物和可滴定酸均为“富士”苹果显著高于其他 2 个品种(表 3)，与 3 个品种果实的近红外高光谱所表现的“富士”苹果与“秦冠”苹果交叉，“乔纳金”苹果明显较低的规律也不一致。丁秀玲[26]研究了 13 个品种苹果的理化特征。其中“乔纳金”、“富士”、“秦冠”苹果果实总酚含量(质量比)依次为 68.63、112.92、40.98 mg/(100 g)，3 个品种果实总酚含量与它在各波长的近红外反射强度的顺序也不同。由以上结果可以推测，本研究测定之外的化学组分也参与了近红外吸收而导致了近红外高光谱反射强度的品种间差异，利用反射强度差异建立的分类模型能够对不同品种成功识别所依据的是能够综合反映其理化特性的各品种近红外高光谱图，即分类数据来源充分体现了品种特异性，故该分类方法更加灵敏和实用。

3　讨论

不同种类苹果在高光谱信号特征空间中具有稳定和紧凑的局部性，并呈现非线性的表征，因此基于已有数据样本近邻度量的 KNN 分类器能够获得优异的分类性能，而且由于高光谱的数据描述为动态区间较大的浮点数，描述点之间关系的距离度量切比雪夫距离、欧几里得距离和明可夫斯基距离取得了更好的性能，而描述集合关系的距离度量汉明距离和 Jaccard 相似系数则性能欠佳。支持向量机分类器的试验表明，参数寻优后，分类正确率能够达到 96.67%，但是径向基参数 γ 取值(2^{-4}和2^{-5})相对较小，训练所获得的模型可能存在过拟合的风险，未来的研究还需要采集更充分的数据加以验证，这也表明不同品种的苹果在高光谱数据的特征空间中，存在着复杂的非线性穿插，使用径向基核函数映射后也无法完全实现线性可分，此时基于先验样本分布的 KNN 分类器由于不依赖模型假设，即使不同类别的苹果在高光谱特征空间中呈现为复杂的非线性分布，依然获得稳健、优越的分类性能。

本试验并没有使用单纯的化学计量算法进行分类，而是使用了高光谱无损测定结合机器学习的方法，可以利用更少的数据量得到更加准确的结果，使过程更加简便，结果更加准确。

以上结果表明生产上可以利用近红外高光谱成像技术在 865 ~ 1 711 nm 波段下的吸收值结合 KNN 机器学习的方法来设计检测机器来快速准确地鉴别苹果的品种，并有望建立用于苹果品种鉴别的高光谱指纹图谱。

4　结束语

采用近红外高光谱成像技术，通过一系列的图像处理，得到 3 个品种苹果果实在近红外 200 个波段的平均反射光谱，然后通过 Matlab 软件训练了 KNN 算法和 SVM(径向基核函数)的 2 个分类器。其中当 K 取值 3 和 5 时，KNN 分类器中切比雪夫距离、欧几里得距离和明可夫斯基距离 3 种距离算法的识别正确率均达到 100% 的优异性能，可以选定为最简单和高效的分析方法。SVM - RBF 核函数模型中当径向基参数 γ 取 2^{-5}，惩罚因子 C 取 16 和 32 时，品种识别正确率虽然高达 96.67%，可是其训练模型中存在过拟合风险，需要进一步论证。

参　考　文　献

1　闫润，王新忠，邱白晶，等. 基于特征光谱的草莓品种快速鉴别[J/OL]. 农业机械学报，2013，44(9)：182 - 186. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20130932&flag = 1. DOI：10.6041/j.issn.1000-1298.2013.09.032.
YAN Run，WANG Xinzhong，QIU Baijing，et al. Discrimination of strawberries varieties based on characteristic spectrum[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2013，44(9)：182 - 186. (in Chinese)

2　曹芳，吴迪，何勇，等. 基于可见-近红外反射光谱技术的葡萄品种鉴别方法的研究[J]. 光学学报，2009，29(2)：537 - 540.
CAO Fang，WU Di，HE Yong，et al. Variety discrimination of grapes based on visible-near reflection infrared spectroscopy[J]. Acta Optica Sinica，2009，29(2)：537 - 540. (in Chinese)

3　李江波，赵春江，陈立平，等. 基于可见/近红外光谱谱区有效波长的梨品种鉴别[J/OL]. 农业机械学报，2013，44(3)：153 - 157. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20130328&flag = 1. DOI：10.6041/j.issn.1000-1298.2013.03.028.
LI Jiangbo，ZHAO Chunjiang，CHEN Liping，et al. Variety identification of pears based on effective wavelengths in visible/near infrared region[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2013，44(3)：153 - 157. (in Chinese)

4　HE Y，LI X，SHAO Y. Quantitative analysis of the varieties of apple using near infrared spectroscopy by principal component analysis and BP model[C]// AI 2005：Advances in Artificial Intelligence，2005：1053 - 1056.

5　SÁNCHEZ M，JOSÉ De la Haba M，BENÍTEZ-LÓPEZ M，et al. Non-destructive characterization and quality control of intact strawberries based on NIR spectral data[J]. Journal of Food Engineering，2012，110(1)：102 - 108.

6 胡耀华, 刘聪, 何勇. 近红外光谱鉴别鲜枣品种及裂果研究[J]. 光谱学与光谱分析, 2013,33(12):3231 - 3234.
HU Yaohua, LIU Cong, HE Yong. Discrimination of the fresh jujube varieties and dehiscent fruit by NIR spectroscopy[J]. Spectroscopy and Spectral Analysis,2013,33(12):3231 - 3234. (in Chinese)

7 何勇, 李晓丽. 用近红外光谱鉴别杨梅品种的研究[J]. 红外与毫米波学报, 2006, 25(3):192 - 194.
HE Yong,LI Xiaoli. Didcriminating varieties of waxberry using near infrared spectra[J]. Journal of Infrared and Millimeter Waves, 2006, 25(3):192 - 194. (in Chinese)

8 李晓丽, 唐月明, 何勇,等. 基于可见/近红外光谱的水稻品种快速鉴别研究[J]. 光谱学与光谱分析, 2008, 28(3):578 - 581.
LI Xiaoli, TANG Yueming, HE Yong, et al. Discrimination of varieties of paddy based on Vis/NIR spectroscopy combined with chemometrics[J]. Spectroscopy and Spectral Analysis,2008, 28(3):578 - 581. (in Chinese)

9 邬文锦, 王红武, 陈绍江,等. 基于近红外光谱的商品玉米品种快速鉴别方法[J]. 光谱学与光谱分析, 2010, 30(5):1248 - 1251.
WU Wenjin, WANG Hongwu, CHEN Shaojiang, et al. Fast discrimination of commerical corn varieties based on near infrared spectra[J]. Spectroscopy and Spectral Analysis, 2010, 30(5):1248 - 1251. (in Chinese)

10 卢洋, 梁先扬, 李卫军,等. 基于近红外光谱短波段的玉米品种鉴别研究[J]. 河南大学学报:自然科学版, 2012, 42(3):239 - 243.
LU Yang, LIANG Xianyang, LI Weijun, et al. Study on corn varieties discrimination based on short wave band of NIR spectra [J]. Journal of Henan University:Natural Science, 2012, 42(3):239 - 243. (in Chinese)

11 赵花荣, 王晓燕, 陈冠华,等. 利用傅里叶变换红外光谱法鉴定小麦品种[J]. 光谱学与光谱分析, 2004, 24(11):1338 - 1341.
ZHAO Huarong, WANG Xiaoyan, CHEN Guanhua, et al. Identification of wheat varieties by FTIR spectrum [J]. Spectroscopy and Spectral Analysis, 2004, 24(11):1338 - 1341. (in Chinese)

12 周竹,李小昱,陶海龙,等. 基于高光谱成像技术的马铃薯外部缺陷检测[J]. 农业工程学报,2012, 28(21):221 - 228.
ZHOU Zhu, LI Xiaoyu,TAO Hailong, et al. Detection of potato external defects based on hyperspectral imaging technology[J]. Transactions of the CSAE,2012, 28(21):221 - 228. (in Chinese)

13 HUANG M, WAN X M, ZHANG M,et al. Detection of insect-damaged vegetable soybeans using hyperspectral transmittance image[J]. Journal of Food Engineering, 2013, 116(1): 45 - 49.

14 张初, 刘飞, 孔汶汶,等. 利用近红外高光谱图像技术快速鉴别西瓜种子品种[J]. 农业工程学报, 2013, 29(20):270 - 277.
ZHANG Chu, LIU Fei, KONG Wenwen, et al. Fast identification of watermelon seed variety using near infrared hyperspectral imaging technology[J]. Transactions of the CSAE,2013, 29(20):270 - 277. (in Chinese)

15 WANG L, LIU D, PU H, et al. Use of hyperspectral imaging to discriminate the variety and quality of rice[J]. Food Analytical Methods, 2014, 8(2):515 - 523.

16 LIU D, WANG L, SUN D, et al. Lychee variety discrimination by hyperspectral imaging coupled with multivariate classification [J]. Food Analytical Methods, 2014, 7(9):1848 - 1857.

17 孙俊, 路心资, 张晓东,等. 基于高光谱图像的红豆品种 GA - PNN 神经网络鉴别[J/OL]. 农业机械学报, 2016, 47(6):215 - 221. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20160628&flag=1. DOI:10.6041/j.issn.1000-1298.2016.06.028.
SUN Jun, LU Xinzi, ZHANG Xiaodong, et al. Identification of red bean varity with probabilistic GA - PNN based on hyperspectral imaging[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2016, 47(6):215 - 221. (in Chinese)

18 王思玲, 蔡骋, 马惠玲,等. 基于高光谱成像的苹果水心病无损检测[J]. 北方园艺, 2015(8):124 - 130.

19 邹小波,陈正伟,石吉勇,等.基于近红外高光谱图像的黄瓜叶片色素含量快速检测[J/OL].农业机械学报,2012,43(5):152 - 156. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20120526&flag=1. DOI:10.6041/j.issn.1000-1298.2012.05.026.
ZOU Xiaobo,CHEN Zhengwei, SHI Jiyong, et al. Rapid detection of cucumber leaves pigments based on near infrared hyperspectral image technology[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2012,43(5):152 - 156. (in Chinese)

20 CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems & Technology, 2011,2(3): Article No.27.

21 林升梁, 刘志. 基于 RBF 核函数的支持向量机参数选择[J]. 浙江工业大学学报, 2007, 35(2):163 - 167.
LIN Shengliang,LIU Zhi. Parameter selection in SVM with RBF kernel function[J]. Journal of Zhejiang University of Technology, 2007, 35(2):163 - 167. (in Chinese)

22 姜林. 苹果新品种"红乔纳金"的引种观察[J]. 落叶果树, 1995(2): 31.

23 贾希友, 马贵印, 王立刚,等. 高桩素可显著提高红富士苹果的果形指数[J]. 山西果树, 2000(1):7 - 9.

24 王田利, 李强, 曹红霞,等. 苹果新品种——粉红秦冠[J]. 山西果树, 2010(2):44 - 45.

25 李慧,曾铭,要磊,等.近红外光谱技术在食品检测中的应用研究进展[J].食品工业,2012(12):85 - 88.

26 丁秀玲. 不同品种苹果理化特性研究及其在曲奇饼干中的应用[D]. 杨凌:西北农林科技大学, 2011.
DING Xiuling. Physicochemical properties of different cultivars of apples and the use in cookies[D]. Yangling:Northwest A&F University, 2011. (in Chinese)