doi:10.3969/j.issn.1672-4348.2021.06.011

基于 PCA-KNN 的金线莲种类识别

柯程扬[1,2],刘丽桑[1,2],林赫[3],张荣升[1,2]

(1.福建工程学院 电子电气与物理学院,福建 福州 350118;
2.工业自动化福建省高校工程研究中心,福建 福州 350118;
3.国网福建省供电有限公司霞浦县供电公司,福建 宁德 355100)

摘要:为了能对金线莲品系进行方便准确地识别,提出基于 PCA-KNN 的金线莲叶片识别方法。通过图像预处理,获得特征较为明显的叶片区域,再提取纹理和颜色特征,进行特征融合,然后采用 PCA 降低特征维度,提高识别精度,最后通过训练 KNN 分类器完成分类。以 3 个品系的金线莲为例进行鉴别试验,结果表明,提出的识别方法与其它方法相比,正确识别率更高,达到 98.4%,能准确识别不同种类的金线莲。

关键词:金线莲叶片;特征提取;PCA 降维;KNN 算法;分类
中图分类号:TP391.41　　**文献标志码**:A　　**文章编号**:1672-4348(2021)06-0568-06

Species identification of anoectochilus roxburghii based on PCA-KNN

KE Chengyang[1,2], LIU Lisang[1,2], LIN He[3], ZHANG Rongsheng[1,2]

(1. School of Electronics, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou 350118, China;
2. Engineering Research Center of Industrial Automation in Colleges and Universities of Fujian Province, Fuzhou 350118, China;
3. Xiapu Power Supply Company, State Grid Fujian Power Supply Co., Ltd., Ningde 355100, China)

Abstract: In order to facilitate the accurate identification of the strain of anoectochilus roxburghii, a PCA-KNN based identification method was proposed. Through image preprocessing, the blade regions with more obvious features were obtained, then the texture and color features were extracted for feature fusion, and then PCA was used to reduce the feature dimension and improve the recognition accuracy. Finally, the classification was completed by training the KNN classifier. Identification tests of 3 strains of anoectochilus roxburghii were carried out, and results show that compared with other methods, the proposed method can effectively improve the recognition rate up to 98.4%, and it can accurately identify different categories of anoectochilus roxburghii.

Keywords: anoectochilus roxburghii; feature extraction; PCA dimension reduction; KNN algorithm; classification

金线莲享有“药王”的美称,但对生长环境要求严格,自然状态下繁殖率较低。不同种类的金线莲叶片外形十分相像,不法商家以次充好、不良掺假,严重影响了金线莲的临床疗效。金线莲的品系鉴定通常依赖于化学分析方法,主要包括显微鉴定法、高效液相色谱法、DNA(deoxyribonucleic acid)分子鉴定法和近红外光谱检测技术等,然而这些方法需要人工鉴别,存在一定的主观性而且效率低下[1-2]。因此实现不同品系的金线莲高效、准确鉴别是亟待解决的问题。

随着计算机和图像处理技术的发展,近年来,机器视觉技术已被引入到中药材鉴别的各个环节。文献[3]在传统经验鉴别的基础上,采用宽长比、RGB(red-green-blue)值、图像处理软件等现

收稿日期:2021-10-31
基金项目:福建省科技厅面上项目(2019J01773)
第一作者简介:柯程扬(1998—),男,福建厦门人,硕士研究生,研究方向:图像处理、深度学习、视觉 SLAM。
通信作者:刘丽桑(1984—),女,福建莆田人,副教授,博士,研究方向:人工智能、算法预测。

代处理方法与技术,实现了对枸杞子规格与等级的快速划分。文献[4]则利用了药材横切面显微图像的灰度信息,构建灰度匹配模板,实现与尺度及方位无关的中药材样品图像的自动识别。文献[5]利用自适应图像分割算法实现了对中药有效区域内不同部位的分割。然而,当前直接基于叶片图像对中药材进行识别的研究很少,文献[6]基于机器视觉对金线莲品系进行了研究,将提取的特征直接用于分类模型构建,但其构建的集成分类模型参数多,模型训练过程复杂。

鉴于此,本课题基于图像特征对金线莲品系进行识别分析,通过特征的有效融合降低分类模型设计复杂度,利用基于主成分分析(principal component analysis, PCA)的 K 最近邻(K nearest neighbor, KNN)[7-9]算法和支持向量机(support vector machines, SVM)[10-12]算法进行分类。

1　叶片的图像预处理

本研究以红霞金线莲、尖叶金线莲、台湾金线莲为例分析 3 类金线莲的品系鉴别问题,算法的整体流程如图 1 所示。

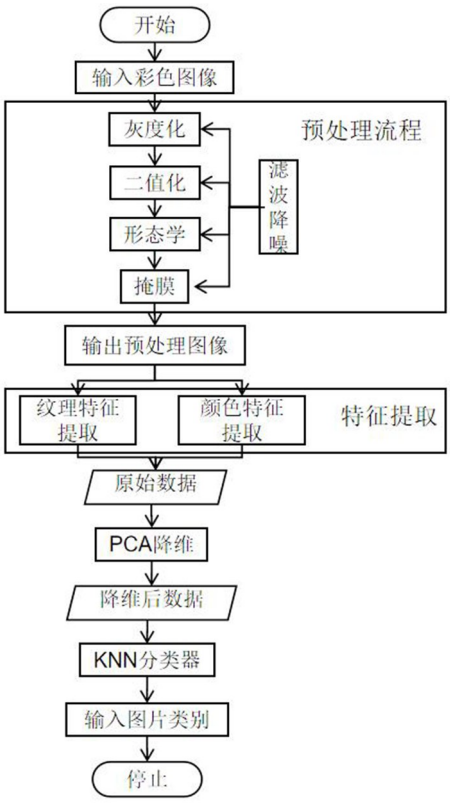

图 1　叶片图像识别整体流程图

Fig.1　Overall flow chart of blade image recognition

由于在自然环境条件下拍摄的干扰信息较多,为改善关键部分的显示效果以便于图像后续处理,首先对图像进行预处理。以红霞金线莲为例,具体步骤如下:

步骤一:将彩色图像缩小至分辨率为 800×800 像素。使用均值滤波和灰度化处理,将彩色图像(RGB 图像)转换为灰度图像,克服纹理提取时噪音及颜色的影响,如图 2(a)所示。

步骤二:采用高斯滤波和二值化处理,即采用大津法(Otsu)来选取滤波阀值 T, 凸显轮廓,对图像空间频率进行强化,提高目标的对比度,减少图像的数据量,如图 2(b)所示。

步骤三:对图像进行形态学处理,其中,闭运算消除叶片的内部孔洞,开运算去除叶柄影响,如图 2(c)所示。

步骤四:对图像进行掩膜,使用滤光片、胶片等物体或多值图像遮挡图像,以提取结构特征,统计屏蔽区,获取感兴趣区的行为,如图 2(d)所示。

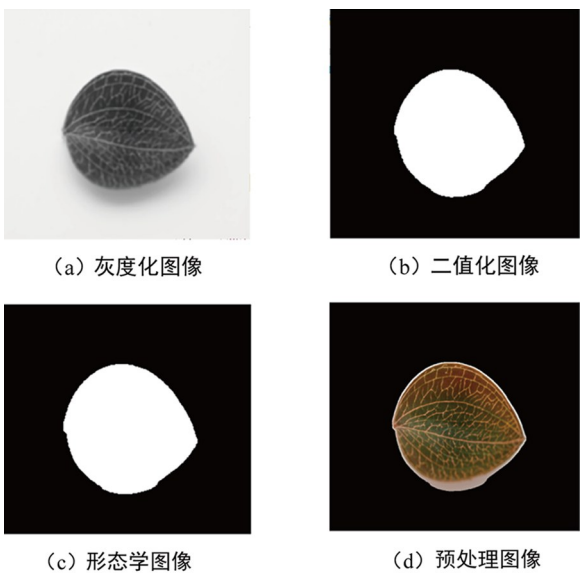

图 2　叶片预处理结果示例

Fig.2　Examples of blade preprocessing results

其中,对图像滤波降噪,在步骤一至步骤二均有采用,其作用是在尽可能保留图像细节特征的条件下,对目标图像进行噪声抑制,这是图像预处理中不可或缺的操作。通过滤波技术淡化部分无用信息,从而改善实物目标与领域或背景之间的灰度反差。

2 叶片的图像特征提取与融合

2.1 纹理特征提取

纹理特征是图像的重要底层特征之一,本试验采用灰度共生矩阵(gray level co-occurrence matrix,GLCM)在统计基础上完成提取。如果两个像素灰度在空间位置上表现出了联合分布的特征,则将其视为 GLCM 矩阵中的元素。根据相邻像素点之间距离参数 D 的不同可以得到不同距离的灰度共生矩阵,通过计算水平方向 $[0,D]$,右上角 45°方向 $[-D,D]$,竖直方向 $[-D,0]$,左上角 135°方向 $[-D,-D]$ 获得灰度共生矩阵。当图像中的各像素灰度值比较接近时,GLCM 存在较大的元素值。常用的 GLCM 矩阵特征指标包括:熵值、对比度、能量、自相关、逆差距等。本次试验提取 16 维数据,分别从水平方向、垂直方向、45°方向、135°方向等 4 个方向提取了关于能量 Asm、对比度 Con、逆差分矩 Idm 和熵值 Ent 数据。

$$\mathrm{Asm}=\sum_{i}\sum_{j}P^{2}(i,j) \tag{1}$$

$$\mathrm{Con}=\sum_{i}\sum_{j}|i-j|^{2}P(i,j) \tag{2}$$

$$\mathrm{Idm}=\sum_{i}\sum_{j}\frac{P(i,j)}{|i-j|^{2}} \tag{3}$$

$$\mathrm{Ent}=\sum_{i}\sum_{j}P(i,j)\log_{2}P(i,j) \tag{4}$$

式中,$P(i,j)$ 为归一化后的灰度共生矩阵的第 i 行第 j 列的值。

利用 GLCM,可以求得起点像素灰度值是 i ,并在离开某固定位置的点时灰度值为 j 的概率。GLCM 生成流程如下:

步骤一:假设点 (x,y) 位于某图像上,其灰度值为 (c_1,c_2) ;$(x+a,y+b)$ 为与该点存在距离或偏离的点。移动点 (x,y) 后,灰度值也会发生相应的变化。

步骤二:若将 k 视为灰度值级数,则存在 k^2 种 (c_1,c_2) 组合;在图像范围内,对每种 (c_1,c_2) 出现的次数进行统计,并以方阵形式排列。

步骤三:归一化处理出现的总数,得出 $P(c_1,c_2)$ 。在截取数值组合时,如果将差分值设为 (a,b) ,那么联合概率矩阵也会表现出众多差异。

通过对纹理特征进行提取后,得到 GLCM 数据,为 120 行 16 列矩阵。

2.2 颜色特征提取

HSV(hue-saturation-value)为使用者提供了直观的颜色描述方案,即颜色亮度、深度等颜色问题。其中,H 为色调;S 为饱和度;V 为明度、色彩的亮度。HSV 的诞生为颜色识别及机器视觉技术的发展提供了重要支持。HSV 空间模型可以对色彩明度、饱和度、色调进行直观的描述,比 RGB 空间更具有表达力。

在 HSV 模型里,V 分量与色彩无关,所以在提取颜色模型时,仅提取 R、G、B、H、S 共 5 个颜色分量。最后统计得到的颜色特征有 15 个(含一阶矩、二阶矩、三阶矩)。具体的计算公式如下:

$$\boldsymbol{C}_{i1}=\frac{1}{N}\sum_{j=1}P_{ij} \tag{5}$$

$$\boldsymbol{C}_{i2}=\left(\frac{1}{N}\sum_{j=1}(P_{ij}-\boldsymbol{C}_{i1})^{2}\right)^{\frac{1}{2}} \tag{6}$$

$$\boldsymbol{C}_{i3}=\left(\frac{1}{N}\sum_{j=1}(P_{ij}-\boldsymbol{C}_{i1})^{3}\right)^{\frac{1}{3}} \tag{7}$$

式中,$\boldsymbol{C}_{i1}$、$\boldsymbol{C}_{i2}$、$\boldsymbol{C}_{i3}$ 分别表示第 i 个颜色分量的一阶矩、二阶矩、三阶矩,P_{ij} 代表第 i 个颜色分量灰度值为 j 的像素点出现的概率。

通过对颜色特征进行提取后,得到 HSV 数据,为 120 行 15 列矩阵。

2.3 PCA 特征降维

PCA 是图像处理中经常用到的降维方法,它通过计算数据矩阵之间的协方差矩阵,得到协方差矩阵的特征向量;再选择特征值最大(协方差最大)的 S 个特征所对应的特征向量组成的矩阵;然后将数据矩阵转换到新的多维空间中,实现数据的降维操作,提高识别精度。

假设分析数据为 $\boldsymbol{N}$(n 个特征),则 PCA 整体计算流程为:

(1)对向量 $\boldsymbol{N}$ 进行中心化;

(2)计算 $\boldsymbol{N}$ 的协方差矩阵;

(3)计算协方差矩阵的特征值和特征向量;

(4)将原始特征数据投影到选取的特征向量上,得到降维后的 S 维特征,S 为经验值。

图 3 为数据降维后的坐标图,由于整体数据量较大,所以只从不同类别金线莲中分别随机选取 3 组数据进行绘图。图 3 中,横坐标为降维后维度,纵坐标为在不同维度下、不同类别的金线莲

数据量大小,1 为红霞金线莲,2 为尖叶金线莲,3 为台湾金线莲。从图 3 可见,前 3 维数据的区别较大,其余 5 维数据重叠较严重,不便用于区分,从而不利于金线莲的识别。

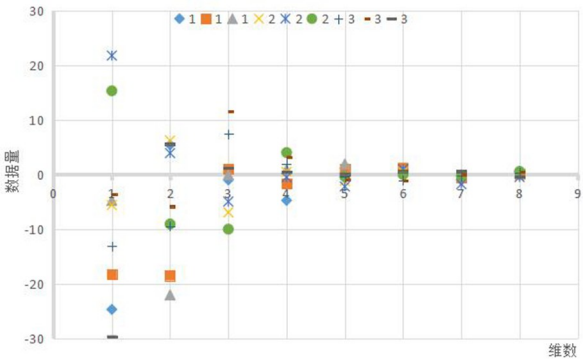

图 3　降维后坐标图

Fig.3　Coordinate diagram after dimension reduction

通过表 1 我们可以清楚地得知,前 5 维数据已经占整体的 99.61%。因此可以通过保留前 5 维数据,处理掉部分占比较小的干扰信息,这就能保留叶片提取特征的绝大部分信息,并且去除相似度极高的特征,从而很好地起到降低特征维度,提高结果的精确度的作用。

表 1　降维后各主成分贡献度

Tab.1　Contribution of each principal component after dimension reduction

维度	占比/%
第一维	73.08
第二维	22.62
第三维	3.14
第四维	0.52
第五维	0.25
第六维	0.18
第七维	0.13
第八维	0.05

3　基于 PCA-KNN 的金线莲识别算法

KNN 算法核心思想:如果一个样本在特征空间的最相似(最近临)的 N 个样本大部分属于某个类别,则该样本也属于这个类别。

对于 KNN 算法,首先构建一个训练样本集合 X,并设定 K 的初值(先确定一个初始值,再进行调整,实现最优参数);分类时,在训练样本集中选出与特征 Z 最接近的 k 个样本。

假设样本点 x 属于 n 维向量空间 R,则采用欧式距离计算样本点之间的近邻关系。对于样本集合 X, 设第 i 个样本属于 R。则样本 x_i 与样本 y_i 之间的欧式距离为:

$$\mathrm{dist}(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{8}$$

若给定一个待分类样本 x_s ,其中 $x_1,x_2,\ldots,x_k$ 表示与 x_s 距离最近的 k 个样本,假设离散目标函数为 f , v 表示类别标签。

$$\bar{f}(x_s)=\mathrm{argmax}\sum_{i=1}^{k}\delta(v,f(x_i)) \tag{9}$$

其中, $\bar{f}(x_s)$ 表示对 $f(x_s)$ 的估计,即待测样本 x_s 的类别。通过以上整体过程则可实现对未知样本的准确归类,分类匹配示意图如图 4 所示。

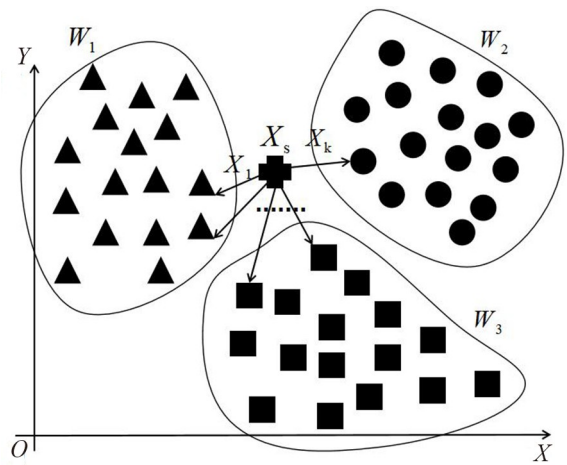

图 4　按距离匹配归类

Fig.4　Classification by distance matching

综上,基于 PCA 的 KNN 算法基本步骤为:

步骤一:获取 PCA 降维后的数据,共 9 列数据,假设调用指令提取 1 到 8 列的特征数据,存储于 jxl_data 中,再次调用指令提取第 0 列的标签,存储于 jxl_target 中。

步骤二:获取金线莲的特征值、目标值。将上一步获取的特征数据 jxl_data 作为特征值,jxl_target 作为目标值。

步骤三:将获取的特征值中 52%(经验值)的数据作为测试集,其余作为训练集。

步骤四:建立特征方程,即对特征值进行标准化处理。其特点为:由于每个特征的大小和取值范围等不一样,从而导致每个特征的权重不一样,而实际上是一样的。通过对原始数据进行变换,把数据变换到均值为 0,方差为 1 的范围内。这样每个特征值的权重都会变得一样,以便于计算机处理。

步骤五:将测试集送入算法训练,获取预测结果。

4 实验结果及分析

实验中,设定测试集大小为 0.52(test_size = 0.52),即 52%的数据用作测试集,剩余数据用于训练,根据经验选取 KNN 算法的 K 值为 1(n_neighbors = 1),比较不同主成分数下模型的效果。实验中用到的部分金线莲叶子图像如图 5 所示。

由表 2 可知,经过 PCA 降维后,选取主成分数为 5 作为 KNN 模型的输入时,该模型的识别率最高,达到 98.4%。

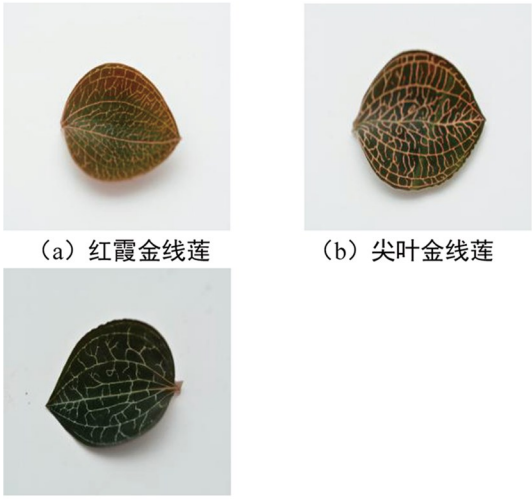

图 5 金线莲叶片图像

Fig.5 Blade images of anoectochilus roxburghii

为验证所提算法的有效性,本研究将其与 KNN、SVM、PCA-SVM 等算法对 63 幅待识别的金线莲叶子图进行识别,并将准确率作为分类结果评价指标。识别结果对比如表 3 所示。

表 2 不同主成分数下 KNN 模型下的识别率

Tab.2 Recognition rate of KNN model under different principal component numbers

主成分数	红霞金线莲识别率/%	尖叶金线莲识别率/%	台湾金线莲识别率/%	总识别率/%
1	56.5	70.0	75.0	66.7
2	69.0	77.8	75.0	73.0
3	83.3	77.3	88.2	82.5
4	95.2	87.0	78.9	87.3
5	100.0	95.5	100.0	98.4
6	95.5	95.5	94.7	95.2
7	91.7	95.0	100.0	95.2
8	95.7	95.2	100.0	96.8

表 3 各方法识别结果

Tab.3 Identification results of each method

方法	红霞金线莲准确识别数/幅	尖叶金线莲准确识别数/幅	台湾金线莲准确识别数/幅	准确率/%	错分率/%
sk-learn KNN	22	17	18	90.5	9.5
sk-learn SVM	22	19	18	93.7	6.3
PCAsk-learn KNN	22	21	19	98.4	1.6
PCAsk-learn SVM	20	22	19	96.8	3.2

由表 3 数据可知,对于未降维的数据,KNN 算法识别准确率为 90.5%,这说明 KNN 算法对于庞大数据的处理能力欠佳,对于小量数据处理效果较好。SVM 算法识别准确率为 93.7%,经过 PCA 降维后的 PCA-SVM 算法识别准确率为 96.8%,均低于本研究提出的 PCA-KNN 算法。

品系识别方面,利用 PCA 降维后的数据集并通过 KNN 算法实现了对金线莲的分类,即采用基于 PCA 的 KNN 算法进行金线莲品系识别更加精准,错分率为 1.6%,准确率达 98.4%。

5　结语

由于金线莲种类繁多且长相相近,人工识别耗时费力,而现有的方法对实验设备的条件要求高且具有局限性,为了能对品系进行方便准确地识别,本研究提出基于 PCA-KNN 的金线莲叶片识别方法。试验采集 3 种品系金线莲图像数据,对其进行预处理,进而提取得到 31 维金线莲的纹理、颜色特征数据;通过 PCA 进行特征降维,共提取累积贡献度达 99.61% 的 5 个主元,通过训练 KNN 分类器完成分类。实验表明,所提出的 PCA-KNN 算法能有效提高金线莲叶片识别准确率,较 KNN、SVM、PCA-SVM 等方法分别提高了 7.9%、4.7%、1.6%。因此,本方法已能对金线莲叶片做出准确的识别,下一步工作将深度学习用于提取特征以及分类识别,进一步优化算法。

参考文献:

[1] 陈莹,任丽,严桂杰,等. 不同来源金线莲的 HPLC 指纹图谱[J]. 沈阳药科大学学报,2019,36(9):794-804.

[2] LV T,TENG R D,SHAO Q S,et al. DNA barcodes for the identification of Anoectochilus roxburghii and its adulterants[J]. Planta,2015,242(5):1167-1174.

[3] 王丹,张久旭,范晶,等. 基于图像处理技术的枸杞子商品规格等级评价方法研究[J]. 世界科学技术-中医药现代化,2020,22(8):2817-2823.

[4] 王凤梅,卢文彪,陈仕妍. 基于灰度匹配模板的中药材显微图像识别[J]. 中国实验方剂学杂志,2019,25(11):167-172.

[5] 王琳, 胡翠英, 庞其昌, 等. 基于自适应图像分割的中药光谱图像检测[J]. 激光与光电子学进展, 2013, 50(12): 70-76.

[6] 谢文涌,柴琴琴,甘勇辉,等. 基于多特征提取和 Stacking 集成学习的金线莲品系分类[J]. 农业工程学报,2020,36(14):203-210.

[7] SYAHRORINI S,SYAMSUDIN D,SAPUTRA D H R,et al. K-nearest neighbor algorithm to identify cucumber maturity with extraction of one-order statistical features and gray-level co-occurrence[J]. IOP Conference Series:Earth and Environmental Science,2021,819(1):012010.

[8] 熊亚军,廖晓农,李梓铭,等. KNN 数据挖掘算法在北京地区霾等级预报中的应用[J]. 气象,2015,41(1):98-104.

[9] 耿丽娟,李星毅. 用于大数据分类的 KNN 算法研究[J]. 计算机应用研究,2014,31(5):1342-1344,1373.

[10] 楚松峰,赵凤霞,方双,等. 基于 PCA-SVM 的红枣缺陷识别方法[J]. 食品与机械,2021,37(1):156-160,198.

[11] 刘君君,王博亮,谢杰镇,等. SVM-KNN 分类器在赤潮生物图像识别中的应用[J]. 心智与计算,2009,3(1):31-36.

[12] 马娜,李艳文,徐苗. 基于改进 SVM 算法的植物叶片分类研究[J]. 山西农业大学学报(自然科学版),2018,38(11):33-38.

(责任编辑: 方素华)