

基于 BP 神经网络对薇甘菊预处理方法的选取

黄亦其¹, 李 婕^{1,2}, 赵建晔³, 杨 睿^{1,2}, 李岩舟¹, 孙中宇⁴, 乔 曦^{1,2*}

(1. 广西大学机械工程学院, 广西南宁 530004; 2. 中国农业科学院深圳农业基因组研究所, 广东深圳 518120; 3. 青岛中农万联科技有限公司, 山东青岛 266106; 4. 广东省科学院广州地理研究所, 广东广州 510070)

摘要 以入侵植物薇甘菊高光谱图像为研究对象, 基于 4 种预处理方法对薇甘菊高光谱图像进行降低噪声处理, 分别研究了基于主成分分析的特征提取方法和基于 BP 神经网络的分类模型, 筛选出薇甘菊高光谱识别的最优预处理方法, 以实现薇甘菊的快速准确识别。结果显示, 预处理方法为一阶、二阶微分的识别率分别为 81.2% 和 76.92%; 标准正态变量变换 (SNV) 和一阶微分+SG 平滑的识别率分别为 89.74% 和 87.18%。多次试验得到基于 SNV 预处理方法的识别率最稳定, 即得到最优预处理方法为 SNV。

关键词 高光谱技术; 薇甘菊目标识别; 特征集选取; BP 神经网络模型

中图分类号 S126 文献标识码 A

文章编号 0517-6611(2020)05-0246-04

doi:10.3969/j.issn.0517-6611.2020.05.069

开放科学(资源服务)标识码(OSID):



Selection of Pretreatment Method for *Mikania micrantha* Based on BP Neural Network

HUANG Yi-qi¹, LI Jie^{1,2}, ZHAO Jian-ye³ et al (1. College of Mechanical Engineering, Guangxi University, Nanning, Guangxi 530004; 2. Shenzhen Institute of Agricultural Genomics, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120; 3. Qingdao China-Agri IOT Science & Technology Co., Ltd., Qingdao, Shandong 266106)

Abstract With invasive plants *Mikania micrantha* Kunt hyperspectral image as the research object, based on four kinds of pretreatment method of *M. micrantha* hyperspectral image noise reduction processing, we studied the feature extraction method based on principal component analysis and the classification model based on BP neural network. The optimal pretreatment method for *M. micrantha* hyperspectral identification was screened, in order to realize fast and exact recognition of *M. micrantha*. The experimental results showed that the recognition rates of first and second orders were 81.2% and 76.92%, respectively. The recognition rates of standard normal variable transformation (SNV) and first order differential +SG smoothing were 89.74% and 87.18%, respectively. Multiple experiments showed that the recognition rate of SNV-based pretreatment method was stable, in other words, the optimal pretreatment method was SNV.

Key words Hyperspectral technology; *Mikania micrantha* Kunt target recognition; Feature set selection; BP neural network

薇甘菊 (*Mikania micrantha* Kunt) 与附主植物在可见光下视觉辨识度低, 分布无规律, 受复杂多变的野外环境限制, 一般检测方法很难实现对薇甘菊实时精准监测。人工踏查^[1]是比较普遍的针对薇甘菊的监测方法之一, 但其受地形影响, 无法对薇甘菊的分布及危害进行准确的定量评测。由于具有光谱分辨率高、波段多、信息量丰富等特点, 高光谱监测技术应用于薇甘菊入侵的监测上具有很大的优势。

目前, 监测方法主要有图像识别法、高光谱监测法、雷达数据辅助识别法和中低分辨率时序序列数据分析法等^[2]。Chance 等^[3]利用小型光谱成像仪 (CASI) 的图像, 在加拿大不列颠哥伦比亚省萨里的非森林植被环境中提取喜马拉雅黑莓 (*Rubus armeniacus*)、英国常春藤 (*Hedera helix*) 和其他物种的光谱, 对所有的光谱通过光谱通道选择算法处理, 结果表明对英国常春藤和喜马拉雅黑莓的检测准确率分别为 80.0% 和 76.4%。而 Narumalani 等^[4]对光谱图像进行去噪处理, 基于光谱角匹配法检测了某区域内 4 种入侵物种及其分布, 总体准确率为 74%。

高光谱数据降维方法主要分为 2 种^[5]: 一种是基于变换

的特征提取法, 即通过数据变换产生一个新的特征空间, 将全部波段的特征属性完成从高维空间到低维空间转换, 从而保留尽可能高的分类精度。特征提取中典型的算法有包括主成分分析 (principal components analysis, PCA)^[6]、奇异值分解 (singular value decomposition, SVD)^[7] 等。另一类是基于非变换的特征选择, 也称局部特征波段选取。常见的波段选择方法包括基于信息量的波段选择以及基于类间可分性的波段选择^[5]。鉴于此, 笔者使用主成分分析法完成特征提取, 以较低的维数达到较好的识别结果, 提高识别效率。

准确地识别外来入侵植物是对其开展监测工作的前提与基础, 任智伟等^[5]通过将基于用 PCA 法和信息量改进 PCA 法对高光谱图像数据降维的结果输入最小距离分类器比较其分类精度来对比分析 2 种降维方法的效果。喻俊等^[8]先利用 S-Golay 滤波剔除异常光谱数据, 再采用光谱微分法与植被指数法构建光谱特征最后应用人工神经网络法结合因子分析法对典型植被进行分类识别。而吴培强等^[9]以黄河口湿地为研究区分别建立支持向量机、人工神经网络和光谱角制图 3 种分类模型开展利用全波段的和所选择特征波段的分类对比试验。笔者通过成像光谱仪获取以薇甘菊为识别目标的高光谱图像, 人工筛选获取薇甘菊类和非薇甘菊类高光谱数据来用作识别薇甘菊目标的研究, 并基于主成分分析方法 (PAC) 对其做适当的降维处理。

1 薇甘菊数据的采集

高光谱图像的采集采用人工手持德国 Cubert 公司生产的 S185 高速成像光谱仪。共采集到薇甘菊高光谱图像 27

基金项目 国家自然科学基金青年科学基金项目 (31801804); 大鹏新区产业发展专项资金项目 (KY20180117); 南宁市重点研发计划 (20192065); 广东省科学院实施驱动发展能力建设专项 (2018GDASCX-0101)。

作者简介 黄亦其 (1973—), 男, 广西隆安人, 副教授, 博士, 硕士生导师, 从事农业信息获取与计算机视觉方面的研究。* 通信作者, 助理研究员, 博士, 从事农业信息获取与计算机视觉方面的研究。

收稿日期 2019-08-05; **修回日期** 2019-09-29

张,分辨率为 1 000×1 000。部分薇甘菊高光谱图像样本如图 1 所示。观察统计显示,样本中包括 3 类其他植物叶子以及其中 1 类植物的花,还有一些黑色背景。



图 1 薇甘菊高光谱图像样本

Fig. 1 Hyperspectral image of samples of *Mikania micrantha*

Cube Ware 是与 S185 配套的光谱分析软件。通过易于操作的用户界面来记录和分析高光谱图像。主要用到信号即图像处理模块,即从 Cube Ware 中获取的高光谱数据样本的预处理,数据降维、BP 神经网络模型的构建等都在 Matlab 2018a 中实现。

1.1 高光谱数据样本获取 通过 Cube Ware 软件打开样本集的薇甘菊高光谱图像,随机从 18 张图像中以像素点为单位采集高光谱数据,每个像素点对应 1 条高光谱曲线,每 1 条曲线所对应的光谱数据都能以 ASCII 码的格式保存(图 2)。共采集到高光谱数据样本 720 个,该研究主要对薇甘菊进行识别,因此图片中出现的其他类植物分别记为其他类植物 1,其他类植物 2,其他类植物 3 叶子和其他类植物 3 叶子花,具体的样本采集数目见表 1。

表 1 光谱数据样本采集数目比较

Table1 Comparison of the number of collected spectral data samples			
样本类别 Sample type	总样本数 Total samples//个	样本名称 Sample name	样本数目 Sample number//个
薇甘菊类 <i>Mikania micrantha</i>	360	薇甘菊花	180
		薇甘菊叶子	180
非薇甘菊类 Others	360	其他植物 1	90
		其他植物 2	90
		其他植物 3 叶子	45
		其他植物 3 花	45
		黑色背景	90

1.2 异常样本剔除 由于环境干扰或人为因素,在其采集与传输样本数据过程中可能造成某些数据不正常或丢失,这种数据被称为异常值^[10]。高光谱数据分析过程中,为了提

高建模的准确性与稳定性,对异常样本进行剔除处理,该研究将采用马氏距离(MD)剔除方法剔除异常样本^[11]。

马氏距离剔除方法基于马氏距离,通过计算高光谱数据马氏距离即数据的协方差距离来判定是否为异常光谱数据样本。马氏距离主要计算公式(1)为:

$$MD_k = \sqrt{(X_k - \mu)^T \Sigma^{-1} (X_k - \mu)}$$
 (1)

式中, $k=1, \cdots, n$,为样本编号, X_k 为每个样本的光谱值, μ 为样本集的光谱平均值, Σ^{-1} 为协方差矩阵。

分别对薇甘菊类和非薇甘菊类进行了异常光谱数据样本剔除,薇甘菊类剔除 9 个样本,非薇甘菊类剔除 9 个样本。剔除异常光谱数据后,薇甘菊类样本数目为 351,非薇甘菊类样本数目为 351。

2 高光谱数据预处理

为了减弱原始光谱数据中的噪声,需要对剔除样本的光谱数据样本进行光谱数据预处理,以便更有效地利用光谱中包含的有效信息,大而提升模型的预测能力以及稳定性。

2.1 平滑 平滑通常被用于消除光谱数据中高频噪声的干扰,提高光谱信噪比。S-G 平滑是一种在时域内基于局域多项式最小二乘法拟合的滤波方法。其运算公式(2)为:

$$y_i = \frac{1}{n} \sum_{j=-m}^m c_j y_{i+j}$$
 (2)

式中, y_i 为在点 i 过滤掉的值; y_{i+j} 为在点 i 未过滤掉的值; j 指定该点左侧平滑数据点数; m 为指定该点右侧平滑的数据点数; c_j 为卷积参数。

2.2 微分 微分指对光谱数据求取不同阶数导数的方法,通过求导可以有效降低噪声对光谱有用信息的干扰,从而达到抑制无用信息和突出感兴趣信息的目的^[12],在消除其他背景方面具有明显的优势。该研究使用的微分法主要有一阶微分(FDR)和二阶微分(SDR)。

2.3 标准正态变量变换(SNV) 标准正态变量变换与 MSC 对高光谱数据处理有相似的作用,即可消除因样本间颗粒大小差异、光谱变换及表面散射等原因引起的高光谱数据误差。计算公式如式(3):

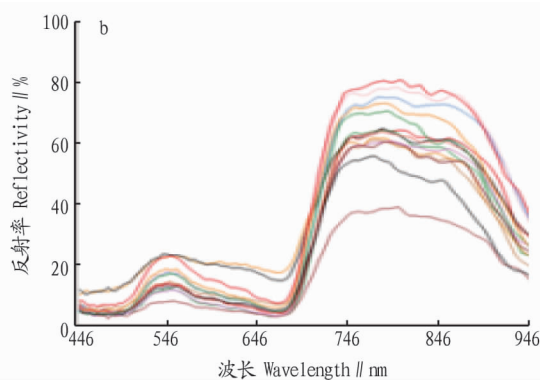
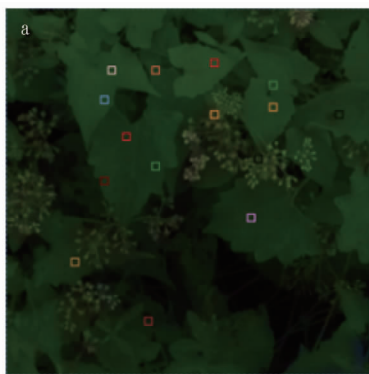
$$Z_i = \frac{x_i - \mu}{\sqrt{\frac{\sum_{k=1}^n (x_{ik} - \mu)^2}{n-1}}}$$
 (3)

式中, x_i 为第 i 个样本的原始光谱数据, x_{ik} 为第 i 个样本中第 k 波段处的光谱反射值, n 为波长点数, $k=1, 2, \cdots, n$ 。 μ 为构成光谱所有点的均值。

3 基于全波段的薇甘菊特征提取

高光谱数据因其波段窄、波段多等特点,能够更大程度上反映地物局部精细信息,在分析植被的反射光谱差异性方面具有重要的实用意义和研究价值^[13]。因此,能有效地适用高光谱数据的降维算法对薇甘菊识别的准确性至关重要。以下是对预处理后的高光谱数据基于主成分变换完成薇甘菊光谱特征提取。

主成分分析(principal component analysis, PCA)是由 Hotelling 于 1993 年提出的一种多元统计分析方法,目前仍作



注:a. 高光谱图像;b. 高光谱曲线

Note: a. Hyperspectral images; b. Hyperspectral curves

图2 高光谱样本采集

Fig. 2 Collection diagrams of hyperspectral samples

为数据挖掘工具广泛应用^[14]。主成分分析的核心是通过正交变换将 1 组存在相关性的高维数据转换为少数的线性不相关的变量,转换后的这组变量称为主成分。这些主成分是原来数据变量的线性组合^[15],能够反映原变量的数据特征。

该研究对 4 种预处理后的光谱数据集分别进行主成分分析计算,初步根据前 k 个主成分的贡献率和来确定要选择的主成分的个数,一般情况下主成分贡献率和需满足大于等于 85%。第 k 个主成分的方差贡献率计算公式如式(4),前 k 个主成分方差贡献率之和的计算公式如式(5):

$$\eta_i = \frac{\lambda_i}{\sum_{i=1}^m \lambda_i} \quad (4)$$

$$G(k) = \sum_{j=1}^k \eta_j \quad (5)$$

式中, λ_i 为特征值, η_i 为第 i 个方差贡献率, k 为主成分数。

观察图 3 光谱数据降维后的主成分贡献率图,基于 4 种不同预处理的主成分贡献率大小的变化均非常剧烈,从第一主成分到最后呈下滑趋势。第一主成分的方差贡献率最多即第一主成分包含有降维前高光谱数据的大部分信息。观察图 4 主成分贡献率可得,所有预处理方法的主成分贡献率和大概在 25 个主成分位置曲线趋于平滑,即前 25 个主成分贡献率之和基本在 99%~100%,完全能代表降维前的高光谱数据中包含的信息,故定最大主成分数目为 25,即尽可能用最少的主成分数目最大程度代表全部的光谱数据信息,提取出最优的识别薇甘菊的高光谱数据特征集。

在主成分分析的基础上还可以通过主成分得分图初步地对样本进行分析判别。由图 5 可知,薇甘菊与非薇甘菊类(图中的其他类)样本均以坐标原点为中心聚集分布,证明第一、二主成分对高光谱数据具有一定的聚类效果。比较发现,标准正态变量变换聚类效果较好。同时,各主成分得分图中非薇甘菊类样本比薇甘菊类聚集更紧密,聚类效果更明显,但 2 类样本间交叉重叠明显,还需要基于支持向量机分类器对其进行精确分类。

4 BP 神经网络建模

BP (back propagation) 神经网络是 1986 年由 Rumelhart 和 McClelland 为首的科学家提出的概念,是一种按照误差逆

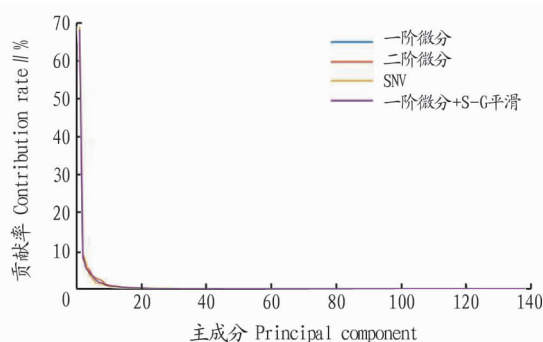


图3 主成分贡献率

Fig. 3 Contribution rate of principal components

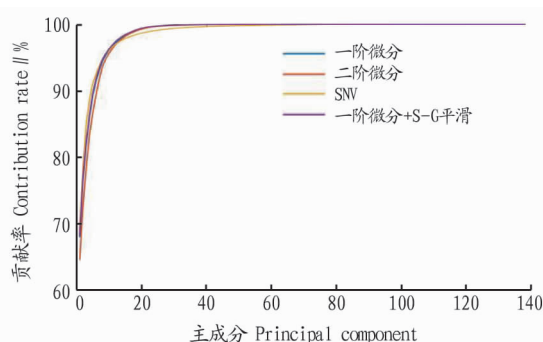


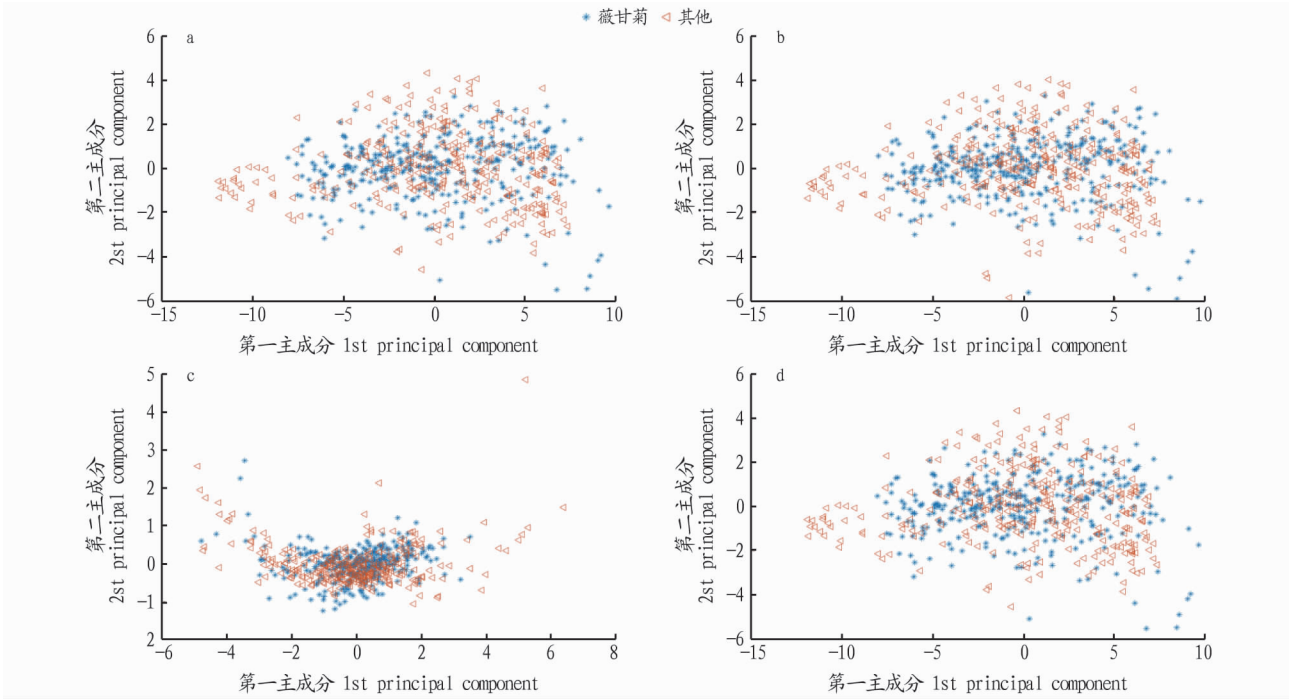
图4 前 k 个主成分贡献率曲线

Fig. 4 Cumulative contribution rate curve of the first k principal components

向传播算法训练的多层前馈神经网络,是目前应用最广泛的神经网络^[16]。BP 神经网络是一种多层前馈,并通过误差反传训练的一种网络,其主要方法是梯度下降法,使网络的实际输出值和期望输出值的误差均方差达到最小^[16]。该研究使用了 BP 算法信号的前向传播和误差的反向传播 2 个部分,计算误差输出时按从输入到输出的方向进行,而调整权重和阈值则从输出到输入的方向进行^[17]。BP 神经网络结构如图 6 所示。

在 BP 神经网络模型中使用函数 sigmoid 作为传递函数,计算公式(6)为

$$F(x) = \frac{1}{1 + e^{-x}} \quad (6)$$



注:a.一阶微分主成分分析;b.二阶微分主成分分析;c.一阶微分+SG 主成分分析;d. SNV 主成分分析
Note:a. First-order differential principal component analysis; b. Second-order differential principal component analysis; c. First-order differential +SG principal component analysis; d. SNV incipal component analysis

图 5 基于不同预处理方法的主成分得分

Fig.5 Scores of principal components based on different pretreatment methods

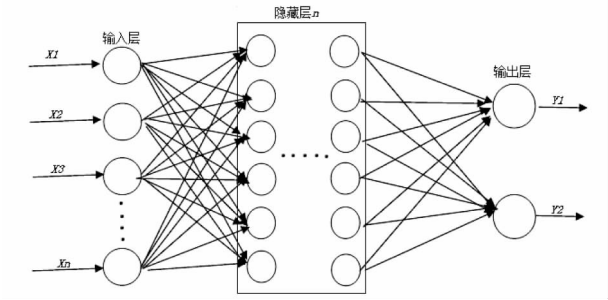


图 6 BP 网络结构图
Fig.6 Simulation model of the pneumatic device

式中, x 为参数。

将 4 种预处理后的主成分作为 BP 神经网络的输入建立神经网络模型,并相互对比,确定最优分类准确率的预处理方法。BP 神经网络建模后,进行参数调节,隐藏层均设置为 10,输出层神经元个数均设置为 2,而后调试隐藏层和输出层学习率、训练次数等重要参数。对神经网络的参数进行反复调整直到获得最佳分类准确率。基于 4 种预处理方法分别建立 4 个神经网络模型。建模时按 2:1 分为训练集和测试集。

从表 2 可以看出,基于 SNV 和一阶微分+SG 预处理方法进行分类得到的分类准确率较优,但是就维度而言,SNV 预处理方法的维度降到 9,比一阶微分+S-G 卷积平滑预处理方法降维程度要好。通过比较得到二阶微分对薇甘菊的识别效果最差。

从图 7 和 8 可以看出,通过 SNV 方法处理过得薇甘菊数据,得到的识别程度相比于 一阶微分+S-G 卷积平滑的更加

表 2 BP 神经网络模型分类准确率比较

Table 2 Comparison of the classification accuracy of BP neural network model				
模型 Model	预处理 Pretreatment	主成分个数 Number of principal component	累计可信度 Cumulative credibility %	分类准确率 Classification accuracy/%
1	一阶微分	15	98.328 1	81.20(190/234)
2	二阶微分	15	98.201 8	76.92(180/234)
3	SNV	9	95.564 1	89.74(210/234)
4	一阶微分+S-G 卷积平滑	15	98.447 3	87.18(204/234)

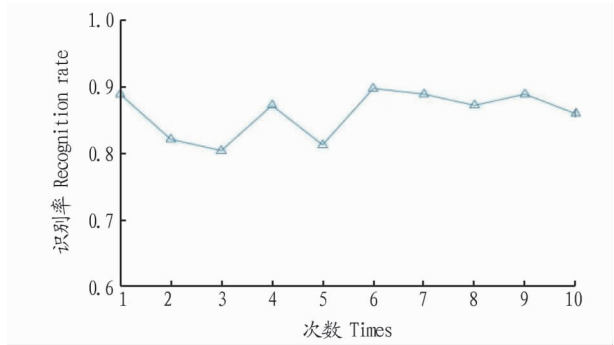


图 7 SNV 识别率变化曲线
Fig.7 SNV recognition rate curve

稳定。结合上述分析,得出基于 SNV 对薇甘菊的识别效果更好。并得到最优的预处理方法为 SNV,且主成分个数为 9 时,对薇甘菊的识别效果最佳。

好的教学效果。随着蓝墨云等移动课堂 APP 技术的发展,相信能为本科专业教学带来更多创新之处。

参考文献

- [1] 肖继坪,罗静波,于虹漫,等.提高基础生物化学课堂教学效果的途径探索[J].安徽农业科学,2018,46(16):234-236.
- [2] 赵玉强,张瑞娟.信息化背景下生物化学课程教学的调查研究及对策分析[J].安徽农业科学,2018,46(21):230-233.
- [3] 杜春梅,朱德旋,董锡文,等.高校生物化学课程翻转课堂的教学设计与实践[J].安徽农业科学,2018,46(34):228-229,234.
- [4] 岳红,张淑芳,田小海.高校生物化学教学现状分析与改革创新研究[J].创新创业理论与实践,2019(19):16-17.
- [5] 朱云霞,张红玲.“互联网+”背景下高校教学改革初探[J].黑龙江教育

- (高教研究与评估),2019(2):61-63.
- [6] 张莹.基于蓝墨云班课的移动混合式教学模式探析[J].现代商贸工业,2019,40(16):159-160.
- [7] 崔为体,谢彦杰,沈文飏.高校专业课精准教学的需求和可行性分析:以南京农业大学生物化学课程群教学为例[J].中国农业教育,2018(4):63-67,95.
- [8] 周威,胡梁斌,李红波,等.食品科学与工程专业生物化学课程教学探索与实践[J].科教导刊,2016(10):90-92.
- [9] 王裴.对高校课堂中互动式教学方法的探索与实践[J].才智,2019(10):31.
- [10] 毛孙忠,雷康福,张伟,等.医学生物化学课程试行“随堂章节考试+期末考”考核实践与探索[J].现代医药卫生,2019,35(12):1884-1886.

(上接第249页)

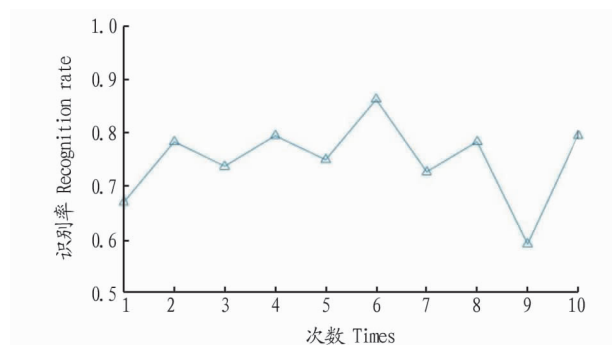


图8 一阶微分+SG 识别率变化曲线

Fig.8 First-order differential +SG identification rate curve

5 结论

该研究主要以外来入侵植物薇甘菊为研究对象,对野外环境下高光谱数据进行分析研究。研究了高光谱数据的异常数据样本剔除及预处理方法,基于高光谱数据降维特征提取方法,基于BP神经网络的模型建立方法。该研究得出了以下结论:

(1)采集野外薇甘菊高光谱图像过程中,应用马氏距离剔除法剔除异常样本,分别剔除了薇甘菊类9个异常样本,以及非薇甘菊类9个异常样本,得到用于后续研究的光谱数据样本共702组,薇甘菊类351组,非薇甘菊类351组。

(2)通过4种预处理方法对高光谱数据预处理,从而达到了降噪、消除基线漂移等预期的效果,但其对识别模型建立的影响还需要进一步验证分析。

(3)对上述所得4个特征集分别进行了基于主成分分析的高光谱数据降维,得到最大主成分为前25个主成分用于识别薇甘菊的特征集。比较分析显示,标准正态变量预处理方法使得高光谱数据聚类效果更好。

(4)对降维方法降维后特征集建立基于BP神经网络分类模型。在针对主成分分析降维后的4个特征集分别建立BP神经网络分类模型,确定了4个特征集中最终的主成分数目以达到特征集进一步降维的目的,并且得出所选主成分

数目对应的测试集识别准确率。其中SNV-BP对应的测试集不但识别准确率高且其所建立的BP分类模型最稳定,最终确定基于前9个主成分数目基础上测试集识别准确率最佳为89.74%,从而验证了BP神经网络能够对野外薇甘菊进行识别。

参考文献

- [1] 柳帅,林辉,孙华,等.基于Pleiades-1卫星数据薇甘菊信息提取[J].中南林业科技大学学报,2014,34(11):116-119.
- [2] 孙玉芳,姜丽华,李刚,等.外来植物入侵遥感监测预警研究进展[J].中国农业资源与区划,2016,37(8):223-229.
- [3] CHANCE C M, COOPS N C, CROSBY K, et al. Spectral wavelength selection and detection of two invasive plant species in an urban area[J]. Canadian journal of remote sensing, 2016, 42(1):27-40.
- [4] NARUMALANI S, MISHRA D R, WILSON R, et al. Detecting and mapping four invasive species along the floodplain of North Platte River, Nebraska[J]. Weed technology, 2009, 23(1):99-107.
- [5] 任智伟,吴玲达.基于信息量改进主成分分析的高光谱图像特征提取方法[J].兵器装备工程学报,2018,39(7):151-154.
- [6] 张兵.高光谱图像处理与信息提取前沿[J].遥感学报,2016,20(5):1062-1090.
- [7] VILLA A, CHANUSSOT J, JUTTEN C, et al. On the use of ICA for hyperspectral image analysis[C]//2009 IEEE international geoscience and remote sensing symposium. Cape Town, South Africa:IEEE, 2009.
- [8] 喻俊,李晓敏,张权,等.基于实测高光谱数据的太湖湖滨带典型植被分类[J].江苏农业科学,2017,45(5):240-244.
- [9] 吴培强,张杰,马毅,等.基于地物光谱可分性的CHRIS高光谱影像波段选择及其分类应用[J].海洋科学,2015,39(2):20-24.
- [10] 闵顺耕,李宁,张明祥.近红外光谱分析中异常值的判别与定量模型优化[J].光谱学与光谱分析,2004,24(10):1205-1209.
- [11] FILZMOSER P, HIRON K. Outlier detection for compositional data using robust methods[J]. Mathematical geosciences, 2008, 40(3):233-248.
- [12] 况润元,曾帅,赵哲,等.基于实测高光谱数据的鄱阳湖湿地植被光谱差异波段提取[J].湖泊科学,2017,29(6):1485-1490.
- [13] 陈彦兵,况润元,曾帅.基于高光谱数据的鄱阳湖湿地典型植被识别分析[J].人民长江,2018,49(20):19-23.
- [14] 顾绍红,王永生,王光霞.主成分分析模型在数据处理中的应用[J].测绘科学技术学报,2007,24(5):387-390.
- [15] 唐红,郑文斌,李宪霞.主成分分析在光全散射特征波长选择中的应用[J].光学精密工程,2010,18(8):1691-1698.
- [16] 邱晨,罗璟,赵朝文,等.基于BP神经网络的空气质量模型分类预测研究[J].软件,2019,40(2):129-132.
- [17] 来纯晓,武振国,金松林,等.基于BP神经网络的小麦抗寒性模型构建[J].河南科技学院学报(自然科学版),2019,47(3):72-78.