大豆科学 Soybean Science
http://www.haasep.cn
2018,37(4):596-605
DOI:10.11861/j.issn.1000-9841.2018.04.0596

基于高光谱图像和邻域粗糙集理论的大豆品种识别算法及其综合性能评估

刘 瑶[1],李梓楠[1],吴 涛[1],刘 潋[2],孟祥丽[1]

(1.岭南师范学院 信息工程学院,广东 湛江 524048;2.哈尔滨工程大学 信息与通信工程学院,黑龙江 哈尔滨 150001)

摘 要:应用高光谱图像技术可以实现大豆品种的快速、高效、无损鉴别检测。高光谱图像数据量大、波段数量多,这导致数据传输、存储和处理有一定难度,故需应用波段选择方法进行数据降维。目前,面向品种识别的高光谱图像波段选择算法大多数都是以分类性能作为算法的评价标准,忽略了算法的稳定性。该文针对大豆品种识别问题,研究基于邻域粗糙集理论中的依赖度、一致性和信息熵等属性选择准则的高光谱波段选择算法,以 Jaccard 系数为稳定性度量指标,研究算法的稳定性随数据集扰动和子集大小的变化情况。针对波段选择算法的稳定性度量与分类模型之间是相互独立的这一问题,不能盲目追求高的稳定性而忽略特征子集的分类效果,提出当波段子集大小相同时采用 Pareto 最优解来评估算法综合性能;当波段子集大小不同时,采用兼顾分类性能、稳定性和子集大小的综合评价函数(PSN)评估算法性能。研究结果对获取综合性能最佳的波段子集有一定的理论及应用价值。

关键词:大豆;高光谱图像;邻域粗糙集;波段选择;综合性能

A Soybean Variety Identification Algorithm Based on Hyperspectral Image and Neighborhood Rough Set Theory and Its Comprehensive Performance Evaluation

LIU Yao[1], LI Zi-nan[1], WU Tao[1], LIU Lian[2], MENG Xiang-li[1]

(1. School of Information Engineering, Lingnan Normal University, Zhanjiang 524048, China; 2. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: The rapid, efficient and nondestructive identification for soybeans varieties can be realized by using hyperspectral image technology. Hyperspectral image data is large and contains hundreds of bands, therefore it is difficult for data transmission, storage and processing. It is necessary to use band selection method for dimension reduction. At present, band selection algorithms of hyperspectral image for variety identification mainly take classification performance as evaluation criteria, and the stability of the algorithms is ignored. To solve the problem of variety identification for soybeans, hyperspectral band selection algorithms based on the dependence, consistency and information entropy criteria in neighborhood rough set theory were studied in this paper. By introducing Jaccard index as a metric of stability, the changes of stability of the algorithm with the dataset perturbation and the subset size were explored. Since the stability measurement and classification model of band selection algorithm are independent of each other, it couldn't pursue high stability only and ignored the classification effect. To assess the comprehensive performance of algorithms, for the subsets with the same size, Pareto optimal solutions was proposed. For the subsets with different size, a comprehensive evaluation function of classification performance, stability and subset size was proposed. The results have certain theoretical and application value for obtaining the best band subset of comprehensive performance.

Keywords: Soybean; Hyperspectral image; Neighborhood rough set; Band selection; Comprehensive performance

大豆作为重要的粮食作物和油料作物,是优质植物蛋白和油脂的主要来源,在农业生产和食物消费系统中具有重要地位。种子的质量和纯度直接关系到大豆的产量与质量。如果大豆种子中混杂掺假,将极大损害国家和农民的利益。因此,大豆品种鉴别在作物育种、农业生产以及市场流通等领域都具有十分重要的意义,是保证品种优良遗传性状得以充分发挥,保护农业生产安全,促进农业生产持续优质、高产、稳产的有效手段。目前,种子品种鉴别的常用方法有形态学方法、荧光扫描分析法、化学鉴定法等[1]。在这些方法中,形态学法需要鉴定者有丰富的经验,但有一定的主观性、精度不高;荧光扫描法、化学鉴定法鉴别精度较高,但是过程繁琐、检测时间长,且对样本具有破坏性,不适

收稿日期:2018-03-05
基金项目:黑龙江省自然科学基金重点项目(ZD201303);湛江市科技攻关计划项目(2017B01143);岭南师范学院自然科学研究人才专项(ZL1902)。
第一作者简介:刘瑶(1982-),女,博士,讲师,主要从事高光谱图像处理技术研究。E-mail:liuyao@lingnan.edu.cn。
通讯作者:吴涛(1980-),男,博士,副教授,主要从事网络学习、智能信息处理研究。E-mail:wutao@whu.edu.cn。

宜对样品进行批量快速分析。

光谱分析技术能通过光谱信息反映大豆种子品种间内部物理结构和化学成分的差异,可实现快速、高效、无损的品种鉴别检测。朱大洲等[2]应用近红外光谱仪采集大豆的漫反射光谱,结合软独立建模分类方法建立大豆的定性分析模型。杨冬风等[3]应用主成分分析和离散多带小波变换提取大豆的光谱特征,建立基于 BP 神经网络的大豆品种识别模型。柴玉华等[4]利用主成分分析后的高光谱图像的纹理特征对大豆进行分级。Tan 等[5]应用 PCA 方法提取高光谱图像的能量、熵、惯性矩和相关性等特征变量,构建神经网络分类器。Liu Y 等[6]将基于粗糙集的高光谱波段选择技术引入到大豆品种检测领域,从而实现大豆分类特征波段信息的充分提取。可见,基于高光谱技术的品种识别研究大多数仅是以分类性能作为模型好坏的评价标准,忽略了算法的稳定性。稳定性定义为在具有相同分布的不同训练数据集中,特征选择算法所获得的特征子集表现出的鲁棒性,即选择结果的可重复性[7]。对于一个不稳定的特征选择算法,在训练集中减少或者添加一些样本后,选择结果不可重复;甚至当训练集样本没有发生任何变化的时候,选择结果仍然不同。算法不稳定的主要原因是在设计算法的时候,没有考虑稳定性问题。经典的特征选择方法目的是找到一个能够达到最高的分类准确率的最小特征子集,而未考虑算法的稳定性。追求特征子集最小的设计方案会忽略冗余特征中含有的重要信息。

在应用高光谱图像数据进行分类的过程中,波段选择是一个重要步骤,其任务是找出特定的波段子集来有效地描述不同类型样本之间的差异。在农业领域,专家们希望找出所有关键特征,并且通过这些特征值的变化对类别差异的内在机理进行解释和研究。如果波段选择算法的稳定性较差,当测试数据的样本发生微弱变化时,就会在相同的波段选择算法下导致所选择出的波段子集产生很大的变化,更有甚者,波段子集完全不同。专家无法从变化较大的波段子集中获取有规律的信息,这会给研究带来困扰,并且这种波段子集实际意义不大。因此,高光谱波段选择算法的稳定性研究在提高分类性能的可信度、找出关键特征、减小测量工作量、提高效率等方面非常重要,应在考虑所选波段子集分类精度基础上,综合考虑稳定性,寻求综合性能最佳的算法。

本文将粗糙集理论和高光谱技术相结合,并将其引入到大豆品种识别领域,研究基于依赖度、一致性和信息熵等属性选择准则的高光谱波段选择算法。除分类性能外,引入另一种衡量波段选择算法优劣的重要指标—稳定性度量指标,研究稳定性随样本扰动情况和样本子集大小的变化情况。并根据波段子集大小相同与否两种情况评估算法的综合性能,以获取分类准确率高且稳定的波段子集,为建立大豆高光谱无损检测系统,实现快速、准确、无损、稳定的检测提供基础。

1　材料与方法

1.1　高光谱数据的获取

本试验所用 3 种大豆样品由东北农业大学国家大豆研究中心提供,分别是东农 42、东农 51 和东农 61,在每个品种中选取 110 粒正常完好的籽粒,共计 330 个样本用于试验。试验所用仪器是美国 Headwall 公司的高光谱图像采集系统,该系统包括图像光谱仪(Hyperspec VNIR-A)、CCD 像头、镜头、光源为 150 W 可调功率光纤卤素灯、由步进电机控制的样本输送平台。高光谱仪的狭缝是 25 μm,有效光谱的范围 400 ~ 1 000 nm,光谱分辨率 2 ~ 3 nm,光谱采样间隔为 0.74 nm。

高光谱成像系统由于光源强度分布不均匀,存在一些噪声,如暗电流,这严重影响采集的高光谱图像质量,因此需要对每幅样本图像进行校正。图 1 为大豆样本东农 51 在 706.15 nm 波段下的校正后图像。

高光谱数据中除了包含样品的自身信息外,还包含其它无用的信息和噪声。为了消除这些因素的影响,在建模之前,对原始光谱数据进行预处理。光谱预处理的方法主要有去趋势、基线补偿、变量标准化、多元散射校正、一阶和二阶求导处理等,本试验采用多元散射校正和变量标准化相结合的预处理方法[8]。

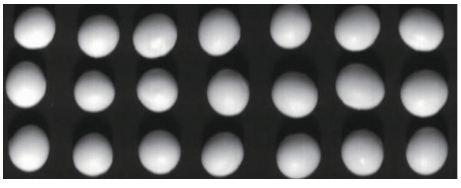


图 1　706.15 nm 波段下东农 51 的校正后图像

Fig. 1　Image of Dongnong 51 after correction in 706.15 nm

1.2　基于邻域粗糙集的大豆高光谱波段选择算法

1.2.1　邻域粗糙集　粗糙集理论由波兰数学家 Pawlak 提出,在机器学习、数据挖掘、模式识别等领域有比较成功的应用[9]。Pawlak 粗糙集能够有效地处理符号型数值,但对连续属性的数值处理能力

非常有限,需要先进行离散化处理。为了解决粗糙集离散化过程中的信息损失问题,本文在粗糙集理论中引入邻域系统概念,提出邻域粗糙集(neighborhood rough set,NRS)模型[10]。

定义 1:给定实数空间上的非空有限集合 $U=\{x_1,x_2,\cdots,x_n\}$,对于 U 上的任意样本 x_i,其邻域为 $\delta=\{x\mid x\in U,\Delta(x,x_i)\leqslant\delta\}$,其中 $\delta\geqslant 0$,Δ 是一个距离度量函数,用 P 范数表示为 $\Delta_p(x_1,x_2)=(\sum_{i=1}^{N}|f(x_1,a_i)-f(x_2,a_i)|^P)^{1/P}$,称 $\delta(x_i)$ 为由 x_i 生成的 δ 邻域信息粒子。

定义 2:已知样本集合 U,C 是条件属性,D 是决策属性,C 的邻域关系为 N,则称 $NDT=\langle U,N,D\rangle$ 为邻域决策系统。

样本之间的邻域关系可以根据所选择的邻域半径和集合的属性来确定,进而计算出样本集合的边界和近似区域。

定义 3:给定邻域决策系统 $NDT=\langle U,N,D\rangle$,D 将 U 划分为 N 个等价类 $X_1,X_2,\cdots,X_N$,$B\subseteq C$,B 生成的邻域关系为 N_B,那么决策 D 关于 B 的邻域下、上近似分别定义为:

$$\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i \tag{1}$$

$$\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i \tag{2}$$

其中,$\underline{N_B}X_i=\{x_i\mid\delta_B(x_i)\subseteq X,x_i\in U\}$,$\overline{N_B}X_i=\{x_i\mid\delta_B(x_i)\cap X\neq\varphi,x_i\in U\}$。

1.2.2 特征波段选择算法 对于高光谱数据,每个光谱波段的图像数据就是一个条件属性,设有 m 个波段,记为 $C=\{c_1,c_2,\cdots,c_m\}$,有 n 个样本,记为 $S=\{s_1,s_2,\cdots,s_n\}$,对应的高光谱波段信息矩阵可以表示为 $W=\{w_{ij}\mid i=1,2,\cdots,n;j=1,2,\cdots,m\}$,其中 w_{ij} 为第 $s_i(s_i\in S)$ 个样本在第 $c_j(c_j\in C)$ 个波段下的光谱值。在 NRS 中,已知的农作物品种可以看作是决策属性,记为 $D=\{d_i\mid i=1,2,\cdots,n\}$,$d_i$ 是样本 s_i 的品种类别标签,波段信息可以看作是条件属性 C,条件属性和决策属性共同构成一个决策属性表。

粗糙集理论的精髓是约简,即在保持信息系统分类能力不变的前提下,通过消除冗余属性,最终得到信息系统的分类或决策规则的方法。属性约简被证明是一个 NP-hard 问题,一个信息系统的约简通常不是唯一的。因此,粗糙集理论的主要研究方向之一是研究出快速的、有效的属性约简算法。本文针对大豆高光谱数据的分类问题,重点研究基于依赖度、一致性和信息熵等的启发式属性约简算法。

(1)基于依赖度的波段选择算法(dependency measure of the NRS,DMNRS)

定义 4:已知邻域决策系统 $NDT=\langle U,N,D\rangle$,$B\subseteq C$,决策属性 D 对条件属性 B 的依赖度定义为:

$$\gamma_B(D)=\frac{|\underline{N_B}D|}{|U|} \tag{3}$$

$\gamma_B(D)$ 是指在条件属性 B 上,决策属性 D 能够完全包含的样本数目与样本总体数目的比值,其取值在 0 和 1 之间。

(2)基于可变精度的波段选择算法(variable precision of the NRS,VPNRS)

可变精度粗糙集模型引入了变精度因子 β,即允许一定程度的分类错误率存在。它在一个给定的相对较小的错误率的条件下,把研究对象尽可能多地归入到同一个类别中,以此来增强粗糙集模型的数据分析能力[11]。

定义 5:设 U 为一个有限的非空论域,$\forall X,Y\subseteq U$,集合 X 关于集合 Y 的包含程度定义为:

$$I(X,Y)=\frac{|Y\cap X|}{|X|} \tag{4}$$

定义 6:设邻域决策系统 $NDT=\langle U,N,D\rangle$,对于 $X\subseteq U$,$0.5<\beta\leqslant 1$,则 X 的可变精度 β 下近似和 β 上近似分别为:

$$\underline{N^\beta}X=\{x_i\mid I(\delta(x_i),X)\geqslant\beta,x_i\in U\} \tag{5}$$

$$\overline{N^\beta}X=\{x_i\mid I(\delta(x_i),X)\geqslant 1-\beta,x_i\in U\} \tag{6}$$

满足以上两个表达式的粗糙集模型是 VPNRS。

(3)基于一致性的波段选择算法(consistency measure of the NRS,CMNRS)

定义 7:设邻域决策系统 $NDT=\langle U,N,D\rangle$,在邻域 $\delta(x_i)$ 中,类 ω_j 的概率分布为 $P(\omega_j\mid\delta(x_i))$,$P(\omega_j\mid\delta(x_i))=n_j/N$,其中,$n_j$ 为邻域 $\delta(x_i)$ 中属于第 j 类的样本数量,N 是邻域内样本的数量,$j=1,2,\cdots,c$,如果 $P(\omega_l\mid\delta(x_i))=\max_j P(\omega_j\mid\delta(x_i))$,则 x_i 的邻域决策(Neighborhood Decision,ND)函数为:

$$ND(x_i)=\omega_l \tag{7}$$

$ND(x_i)$ 是根据 x_i 的邻域中的各个类的概率分布来对 x_i 进行决策。

定义 8:0-1 错误分类损失函数定义为:

$$\lambda(\omega(x_i)\mid ND(x_i))=\begin{cases}0,\omega(x_i)=ND(x_i)\\1,\omega(x_i)\neq ND(x_i)\end{cases} \tag{8}$$

其中 $\omega(x_i)$ 是 x_i 的真实类别。

定义 9:邻域决策误差率定义如下:

$$NDER=\frac{1}{n}\sum_{i=1}^{n}\lambda(\omega(x_i)\mid ND(x_i)) \tag{9}$$

其中 n 表示样本的总量,称 1-NDER 为邻域识别率(neighborhood recognition rate,NRR)。

邻域决策误差率本质上是按照多数决策原则,

根据样本邻域内类的分布信息,给各样本重新分配决策类,然后统计重新分配的类别与实际类别之间的差异率[12]。

(4)基于邻域信息熵的波段选择算法(neighborhood mutual information,NMI)

把信息系统当作离散信息源,把邻域内各个对象看作是一组随机事件,则可以根据信息熵的变化来判断属性的重要度。在分类问题中,互信息可以用来表示特征与决策之间的相关性[6]。

定义 10:已知邻域决策系统 $NDT=\langle U,N,D\rangle$,$S,R\subseteq C$ 是样本的两个条件属性集,则 S 与 R 的邻域互信息定义为:

$$NMI_\delta(R;S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|\delta_R(x_i)|\cdot|\delta_S(x_i)|}{n|\delta_{S\cup R}(x_i)|} \tag{10}$$

邻域互信息可以看作属性 S 与 R 的关联程度的重要指标,反映了已知属性 S 后,属性 R 不确定性减小量。

$NMI_\delta^{x_i}(R;D)=-\log\frac{|\delta_R(x_i)|\cdot|D_{x_i}|}{n|\delta_{R\cup D}(x_i)|}$为属性集 R 中包含决策属性 D 的信息量。决策的邻域互信息值越大说明此属性集能够提供的分类信息越多,也意味着包含在此属性集中的特征预测能力越强。

(5)基于最大相关最小冗余的波段选择算法(maximal relevance minimal redundancy,MRMR)

应用最大相关最小冗余策略寻找最优的波段子集要考虑两方面重要因素,其一是波段与类别之间的相关性,其二是波段之间的冗余性。波段与类别之间的相关性越大,表示波段包含的分类信息就越多,波段之间的冗余性越小,表示波段之间的相似性越小、重复信息越少。

通过两种结合算子最大相关最小冗余商(MRMR quotient,MRMRQ)和最大相关最小冗余差(MRMR difference,MRMRD)[13]将最大相关性与最小冗余性结合起来,实现高光谱波段选择。

定义 11:两种结合算子 Φ_{MRMRD} 和 Φ_{MRMRQ} 的定义如下:

$$\Phi_{MRMRD}=\frac{1}{|S|}\sum_{b_i\in S}NMI_\delta(b_i;D)-\frac{1}{|S|^2}\sum_{b_i,b_j\in S}NMI_\delta(b_i;b_j) \tag{11}$$

$$\Phi_{MRMRQ}=\frac{1}{|S|}\sum_{b_i\in S}NMI_\delta(b_i;D)\Big/\frac{1}{|S|^2}\sum_{b_i,b_j\in S}NMI_\delta(b_i;b_j) \tag{12}$$

1.2.3　*前向贪心搜索算法*　基于启发式搜索机制的约简算法是设计一种属性选择准则作为选择当前最好属性的依据。本文设计一个由条件属性、相对属性子集和决策属性 3 个变量决定的属性重要度函数。任意一个条件属性的重要度可以表现为:在添加一个条件属性到属性子集之后,依赖度、邻域识别率或互信息的增加值。

定义 12:给定一个邻域决策系统 $NDT=\langle U,N,D\rangle$,$B\subset C$,$a\in C-B$,a 相对于 B 的重要度被定义为:

(1)对于 DMNRS 和 VPNRS 波段选择算法,

$$SIG(a,B,D)=\gamma_{B\cup a}(D)-\gamma_B(D) \tag{13}$$

(2)对于 CMNRS 波段选择算法,

$$SIG(a,B,D)=NRR_{B\cup a}(D)-NRR_B(D) \tag{14}$$

(3)对于 NMI 波段选择算法,

$$SIG(a,B,D)=NMI_\delta(B\cup\{a\};D)-NMI_\delta(B;D) \tag{15}$$

(4)对于 MRMR 波段选择算法,

$$SIG(a,B,D)=\Phi_{MRMRD}(B\cup\{a\};D)-\Phi_{MRMRD}(B;D) \tag{16}$$

$$SIG(a,B,D)=\Phi_{MRMRQ}(B\cup\{a\};D)-\Phi_{MRMRQ}(B;D) \tag{17}$$

根据属性重要度准则,构造以空集为起点的前向贪心式属性约简算法。每次迭代,都需要计算剩余的每个属性的重要度,按照从大到小排序,选择重要度值最大的属性加入约简集合中。如果剩余的每个属性的重要度都为零,则算法停止。系统的重要度不会因加入任何新的属性而发生变化,约简属性子集不再影响信息决策系统的分类能力,即属性子集的分类能力达到最大。图 2 为基于前向贪心搜索策略的波段选择算法流程。算法会因不同的邻域 δ 取值而产生很多不同的波段子集,各个子集的分类性能要由极限学习机(ELM)分类器来评价。具有较高的分类准确率,且维数较小的子集被选择作为最后约简结果。

1.3　稳定性测度

稳定性度量包含 4 个关键步骤[14]:(1)通过随机抽样或交叉验证的方式生成样本子集。(2)在每个样本子集上进行特征选择。(3)计算特征选择的结果之间的成对相似性。(4)最终的稳定性即为计算所有成对相似性的平均值。其中,步骤(3)是度量算法稳定性的核心。目前,国内外学者提出了多种度量准则,归纳起来有基于特征子集、特征排序和特征权重的稳定性度量方法[15]。本文采用基于特征子集的稳定性度量方法中的 Jaccard 系数,计算方法如下:

$$sim(f_i,f_j)=\frac{|f_i\cap f_j|}{|f_i\cup f_j|} \tag{18}$$

给定一个包括 n 个样本的数据集 D,从中随机

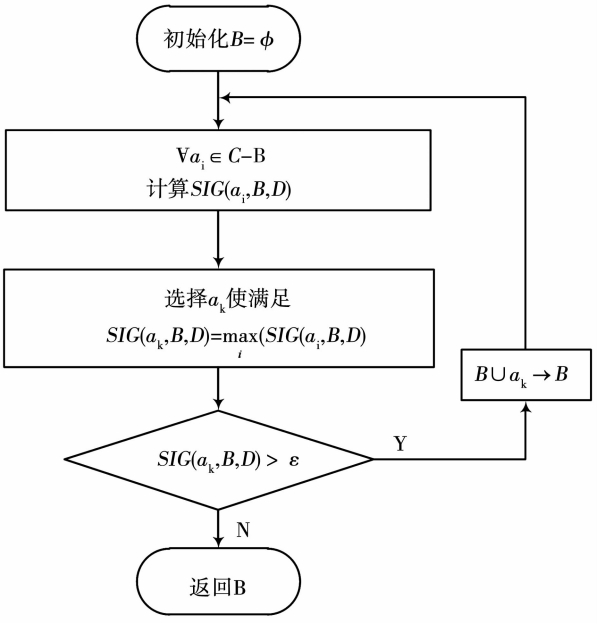


图 2 基于前向贪心搜索策略的波段选择算法流程图

Fig. 2 Flow chart for the band selection by forward greedy search strategy

抽取 m 个样本子集,大小都为 βn,参数 m 和 β 为可变参数。输出结果可以通过在 m 个样本子集上进行特征选择得到,根据上面介绍的相似性度量方法计算任意两个特征选择结果之间的相似性,所有 $m(m-1)/2$对的相似性均值即为整体的稳定性,特征选择算法 F 的稳定性计算公式如下:

$$sim_F = \frac{2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} sim(f_i, f_j)}{m(m-1)} \tag{19}$$

sim_F 越大表示算法 F 的稳定性越好。

1.4 综合评价方法

特征选择算法的稳定性度量与分类模型之间是相互独立的,所以这可能会出现某种特征选择方法很稳定,但却不能取得良好的分类效果的情况。因此,需要在保证一定的分类精度的前提下来研究特征选择算法的稳定性问题,不能盲目追求高的稳定性而忽略特征子集的分类效果。本文分别考虑当波段选择算法选择出的波段子集大小相同和不同这两种情况下的算法性能综合评估方法。

1.4.1 Pareto 最优解 当波段子集大小相同时,面向大豆品种识别的高光谱图像特征波段选择问题是要对两个目标即分类准确性和稳定性同时优化的问题,可作为一个多目标优化问题来考虑。

多目标优化问题就是在一定的约束条件下,求解多个目标函数的最优化问题,可以形式化表示为:

$$\min f(x) = [f_1(x), f_2(x), \cdots, f_p(x)]^T \tag{20}$$

其中 $x = (x_1, x_2, \cdots, x_n)^T$ 称为决策向量,$f_1(x)$,$f_2(x)$,…,$f_p(x)$ 称为目标函数,p 维向量 $[f_1(x), f_2(x), \cdots, f_p(x)]^T$ 所在的空间称为目标空间。

在多目标优化问题中,大多数情况下各子目标可能是相互冲突的,即某子目标的改善可能引起其它子目标性能的降低,即同时使多个子目标均达到最优往往是不可能的。解决多目标优化问题的最终手段是在各子目标之间进行协调权衡和折衷处理,使各子目标函数尽可能地达到最优。

定义 13:决策变量 $x_u \in X$ 称为 *Pareto* 最优解,当且仅当不存在决策变量 $x_v \in X$ 使得相应的目标向量 $v = f(x_v) = (v_1, v_2, \cdots, v_n)$ 优于 $u = f(x_u) = (u_1, u_2, \cdots, u_n)$。

由于 Pareto 最优解并不是唯一的,而是一个 Pareto 最优解集[16],因此研究者可以根据自己的偏好和对各目标函数的重视程度,从 Pareto 最优解集中选出最适合实际情况的满意解。

1.4.2 分类性能、稳定性和子集大小综合评价函数 Saeys 等[17]对一些特征选择算法的学习模型的稳定性和分类性能进行对比,并应用稳定性和分类性能的调和平均函数对模型的综合性能进行评价。设算法的稳定性为 stability,分类性能为 performance,调和平均函数计算如下:

$$\frac{(\beta^2+1)\,stability \times performance}{\beta^2\, stability + performance} \tag{21}$$

参数 β 用来调整调和平均函数中稳定性与分类性能两者的重要性比例,β 越大说明稳定性指标越重要,当 $\beta=1$ 表示稳定性和分类性能同等重要。公式(21)兼顾了稳定性和分类性能,但没有考虑波段子集大小对算法综合性能的影响。这里提出分类性能、稳定性和子集大小综合评价函数(PSN):

$$PSN = \frac{(\beta^2+1)\,stability \times performance}{\beta^2\, stability + performance + \alpha \times \frac{N}{203}} \tag{22}$$

式中,参数 α 用来调整子集大小这个参数的重要性,α 设置的越大,说明较小的子集比较受到青睐,$N/203$ 代表波段子集中的 N 个波段占整体 203 个波段的百分比。

1.5 稳定性实验设计

在度量特征选择方法的稳定性时,大多数研究者采用从原始数据集中随机删除一些样本,或者是利用交叉验证的方式生成多个扰动样本子集。现有的稳定性研究都是隐含地假设样本子集中的样本差异很小,但是在实际应用中,这个假设是不成立的。无论是交叉验证还是随机删除都不能保证所构成的样本子集的差异很小。当样本子集的差异比较大时,基于这些样本子集得到的关于稳定性

的结论将不再正确。当采用随机删除的样本扰动方法进行试验时,没有考虑稳定性受样本子集之间的相似性影响的情况,因此难以确定特征选择算法的稳定是因为数据集比较相似还是因为算法本身内在比较稳定。交叉验证方法虽然能够保证每两个样本子集之间有相同的重叠率,但产生的样本子集之间重叠率是固定不变的,不能清楚地看出子集之间重叠率对于算法稳定性的影响。因此本试验采用一种固定重叠率的样本子集构成方法,产生若干个大小相同,重叠率相同的样本子集。通过改变重叠率来控制样本子集的相似性,采用在不同的重叠率情况下得到的试验结果来说明样本扰动程度对稳定性的影响。根据稳定性的定义,应在数据集轻微扰动的情况下来度量特征选择算法的稳定性。因此本试验中,样本子集之间的重叠率较高,在90%以上。

在样本子集产生的过程中,还要注意一个问题,即样本子集中每类样本所占的比重应与原数据集中每类样本所占的比重相同,以避免数据扰动造成不同类别的样本分配不均衡的问题。从每一类样本里移除 l 个样本,剩余的样本构成样本子集。本试验中,选择 3 个不同程度的样本扰动,l 分别取 1,3 和 5,所对应的样本子集间重叠率分别为 98.15%(53/54),94.23%(49/52)和 90%(45/50)。图 3 是 $l=5$ 时样本子集的构造过程。

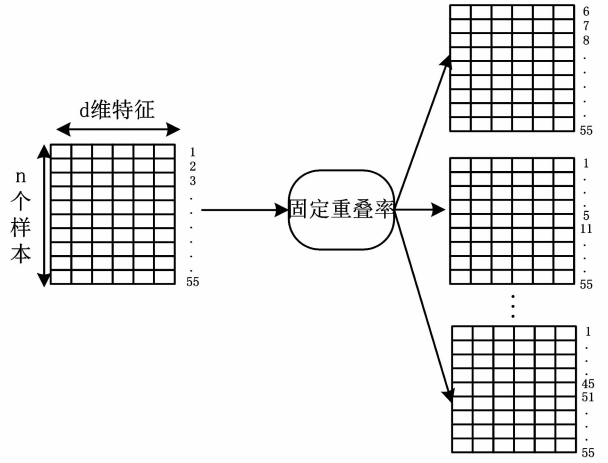


图 3　$l=5$ 时固定重叠率的样本子集构造

Fig. 3　The sample subset of fixed overlapping rate when $l=5$

2　结果与分析

2.1　分类准确率

在每个大豆品种样本中各随机选取 80 个样本,共计 240 个样本作为训练集,然后每组随机选取 30 个样本共 90 个样本用作测试,建立 ELM 分类模型。ELM 建模分析中选取“Sine”函数作为隐含层激励函数,隐含层神经元个数设定为 200[18]。图 4 给出了在 100 次随机选取训练样本和测试样本的试验中基于 NRS 的波段选择算法的最大分类准确率和平均分类准确率随邻域半径 δ 的变化情况,其中,VPNRS 波段选择算法根据前期的研究结果选择参数 β 的取值为 0.7[19]。

从图 4 中可以看出对于同一个高光谱数据集,很多波段子集(可以是同一种波段选择算法产生的,也可以是不同的波段选择算法产生的)有相同或者近似的分类性能。当 $\delta>0.12$ 时,多数情况下,基于 NRS 的各算法的最大分类准确率能达到 100%,平均分类准确率也都在 95% 以上。也就是说,最佳波段子集并不唯一。在评估波段选择算法性能优劣时,除分类性能外,稳定性也应作为一个重要的指标被考虑,进而达到综合评估算法性能的目的。

2.2　稳定性随子集大小变化

为了研究基于 NRS 的不同属性选择准则的波段选择算法的稳定性,我们在大豆高光谱数据集上进行试验。针对前面提到的 VPNRS、CMNRS、DMNRS、NMI、MRMRD 和 MRMRQ 这 6 种波段选择算法,本文选择 Jaccard 系数为稳定性度量指标,对不同波段选择算法的稳定性进行对比。在不同扰动程度下,6 种波段选择算法的稳定性随波段子集大小的变化情况如图 5 所示。从图 5 中可以看出,当样本数据扰动增加时,各个算法的稳定性都有所降低。不同算法之间的稳定性变化差异比较大,稳定性的差异在扰动较小的时候更为明显。

无论在何种程度样本扰动的情况下,MRMRD 方法都是最稳定的;VPNRS、DMNRS 和 MRMRQ 这 3 种算法稳定性居中,当扰动较小时,DMNRS 算法的稳定性略高于 VPNRS 和 MRMRQ 算法;当扰动较大时,MRMRQ 算法的稳定性略高;CMNRS 和 NMI 算法最不稳定。在图 5(a)中,各种算法的稳定性随子集大小变化较大,而在图 5(c)中,稳定性随子集大小变化幅度减小,稳定性曲线都较为平缓,即随着样本子集扰动的增加,子集大小对算法的稳定性的影响减小。

对于同一种算法,当波段子集大小不同时,一些算法的稳定性变化也很大。在图 5 中,随着波段子集大小的增加,VPNRS、DMNRS 和 CMNRS 算法的稳定性呈明显减小趋势,也就是说在 10 个扰动的样本子集中进行波段选择,这 3 种算法都具有选择出的前几个波段高度相似的特点,并且这个特点在 VPNRS 算法上体现的更为典型,因为在 3 种扰动情况下,VPNRS 算法的稳定性都保持随波段子集大小

的增加而减小的趋势,而 DMNRS 和 CMNRS 算法在另外两种扰动下这种变化趋势不太明显,尤其是 DMNRS 算法。与这 3 种算法不同的是 MRMRQ 算法,它的稳定性随着子集大小的增加而增加,也就是说波段子集中包含的波段越多,MRMRQ 算法越稳定。在扰动较大且波段子集较大时,MRMRQ 算法的稳定性最接近于 MRMRD 算法的稳定性。

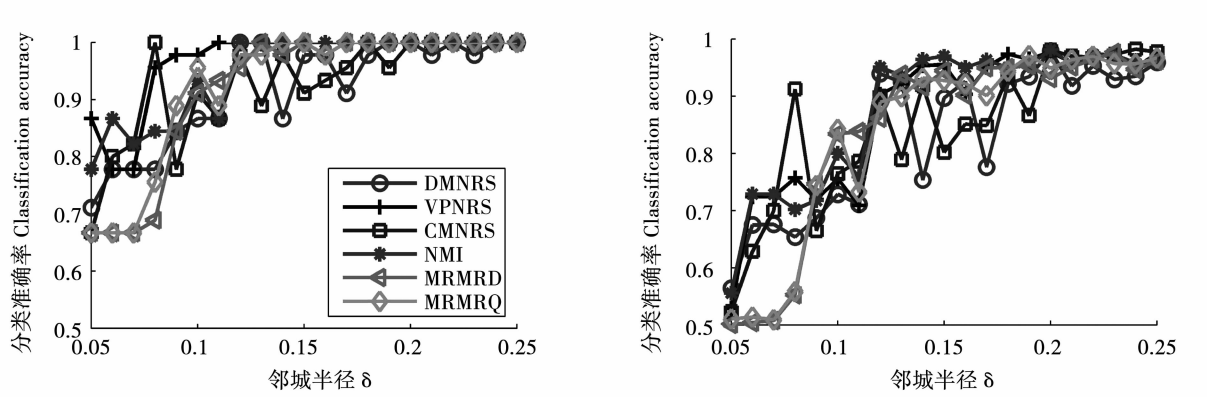
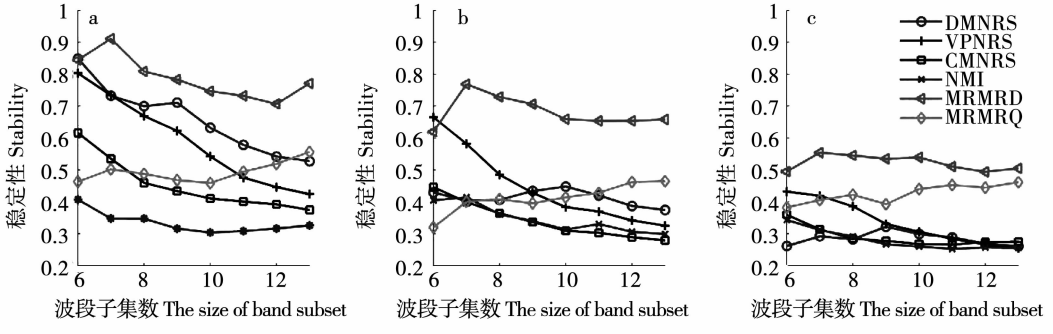


图 4 基于 ELM 分类器的 NRS 算法分类准确率最大准确率(a)平均准确率(b)

Fig. 4 Classification accuracy of the algorithms based on the NRS by the ELM classifier maximal accuracy(a) average accuracy(b)



a:$l=1$; b:$l=3$; c:$l=5$.

图 5 稳定性随波段子集大小的变化情况

Fig. 5 Variation of the stability with the change of subset size

2.3 子集大小相同情况下,算法综合性能评估

当选择出的特征波段子集大小相同时,为了综合评估基于 NRS 理论的各种不同测度的大豆品种检测算法的性能,我们采用 Pareto 最优解。图 6 为子集大小分别为 6,8,10 和 12 时,3 种不同扰动程度下(l 分别取 1,3 和 5),Pareto 最优解图。

在图 6 的每个子图中,右上角的点所代表的算法的性能优于左下角的点所代表的算法。在一部分子图中,可以得到唯一的 Pareto 最优解,如图 6(e),点(0.94,0.73)位于最右上方,即 MRMRD 算法在稳定性和分类性能上均为最优。然而,在大部分子图中,Pareto 最优解并不唯一,如图 6(b),CMNRS 算法的准确率为 0.89,优于 MRMRD 算法(准确率为 0.84),但 CMNRS 算法的稳定性为 0.45,远低于 MRMRD 算法(稳定性为 0.62)。在这种情况下,研究者可以根据自己的偏好和对各评价指标的重视程度,从 Pareto 最优解集中选出最适合实际情况的满意解。

2.4 子集大小不同情况下,算法综合性能评估

表 1 列出了 VPNRS、DMNRS、CMNRS、NMI、MRMRD 和 MRMRQ 这 6 种算法在两种不同的邻域下选择出的波段数量、平均分类准确率和稳定性。表中数据的选择原则是每种算法选择一个分类准确率最高的情况,另外再选择一个分类准确率相对比较高、但波段子集较小的情况。如果单独从分类性能、稳定性或者子集大小的角度,可以选择出最佳的算法,但是如果从三者综合的角度,很难从表中直接看出哪种算法最佳。本文引入 PSN 综合评价函数进行性能评估。

图 7 以 $l=3$ 为例分析 PSN 值与算法性能之间的关系,图中横坐标对应表 1 中的算法前面的序号。研究人员若权衡了稳定性和分类性能之间的关系,认为两者重要度相当,且子集大小对整体性能影响不大,则可令参数 $\beta=1$ 且 $\alpha=1$。从图 7(a)中可以看出,算法 10 性能最优,其次是算法 11 和算法 9。若在一些实际应用中,分类性能的重要度远大于稳

定性的重要度,则可以设定参数 $\beta=0.1$。此时,各种算法的 PSN 值如图 7(b)所示,综合性能最佳的算法仍为算法 10,但次最佳算法已由 $\beta=1$ 时的算法 11 变为算法 9。从表 1 中可以看出,算法 9 的稳定性为 0.60,略低于算法 11 的稳定性(0.63),但算法 9 的分类准确率为 94.47%,略高于算法 11 的分类准确率(93.18%)。在分类性能更重要的情况下,应用 PSN 综合评价函数,算法 9 优先于算法 11 被选出。

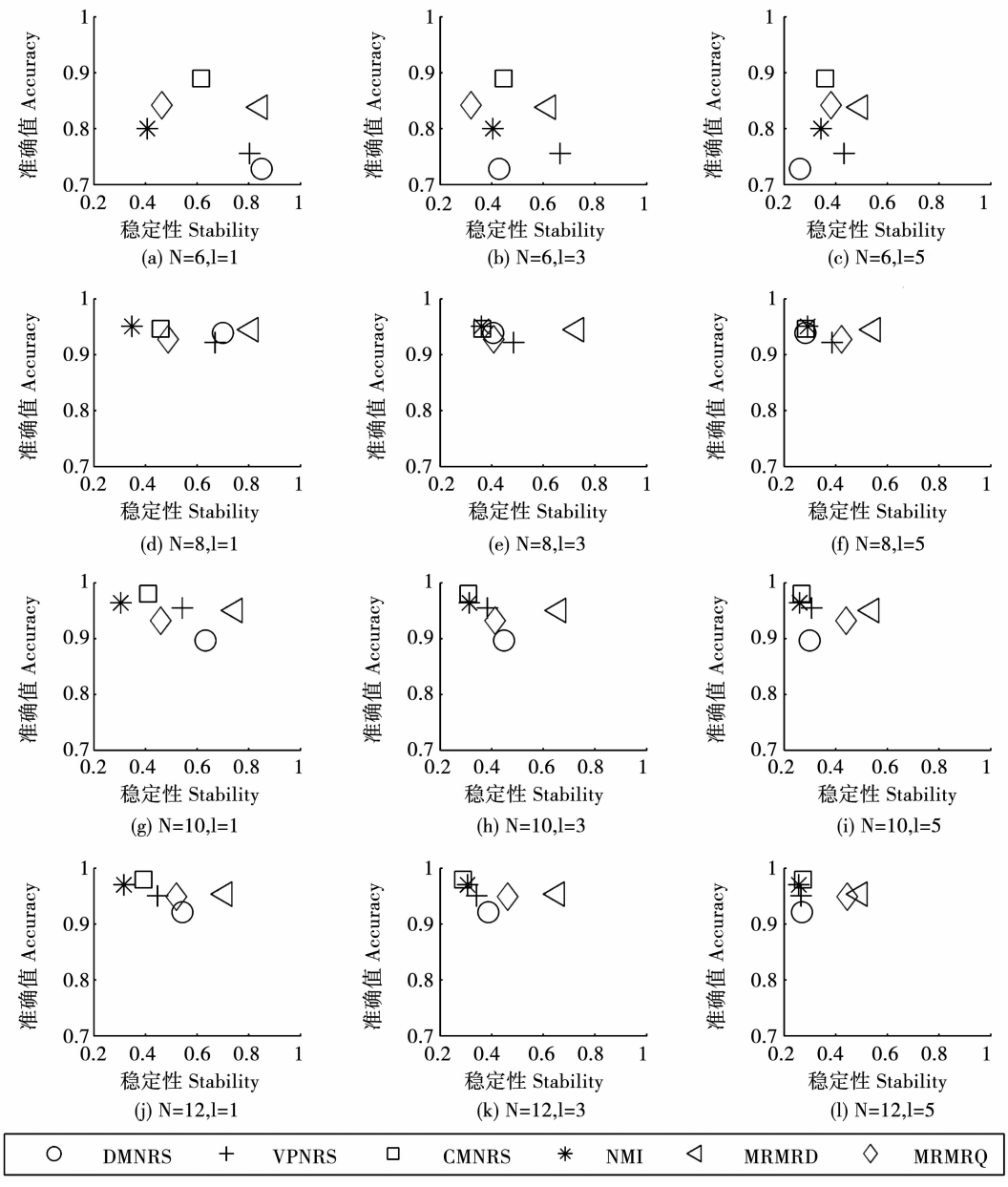


图 6　不同扰动程度下 Pareto 最优解图(子集大小为 6、8、10 和 12 时)

Fig. 6　Pareto optimal solution graph under different perturbation levels (the subset size is 6, 8, 10, and 12)

参数 α 主要用来调整子集大小的重要性,在实际应用中,不太需要考虑子集中波段数量时,α 可以取较小的值,如 $\alpha=1$,但当波段数量对整个算法性能影响较大时,α 需取较大的值。从图 7(c)可以看出,算法 11 和算法 9 的 PSN 值分别为 0.572 1 和 0.570 2,均略大于算法 10 的 PSN 值 0.564 8。当研究人员以得到一个较小的子集且该子集的分类性能和稳定性同等重要为目的时,算法 11 和算法 9 因具有子集维数较小的优势而被综合评价函数选出。图 7(d)所示的算法 9、10 和 11 这 3 个算法的 PSN 值之间的差异比图 7(c)图中的更大。综上,本文提出的分类性能、稳定性和子集大小综合评价函数可快速、有效地选择出综合性能最佳的算法。重要的是,研究人员可根据实际情况灵活地设置参数,达到权衡各个性能指标重要度的目的。

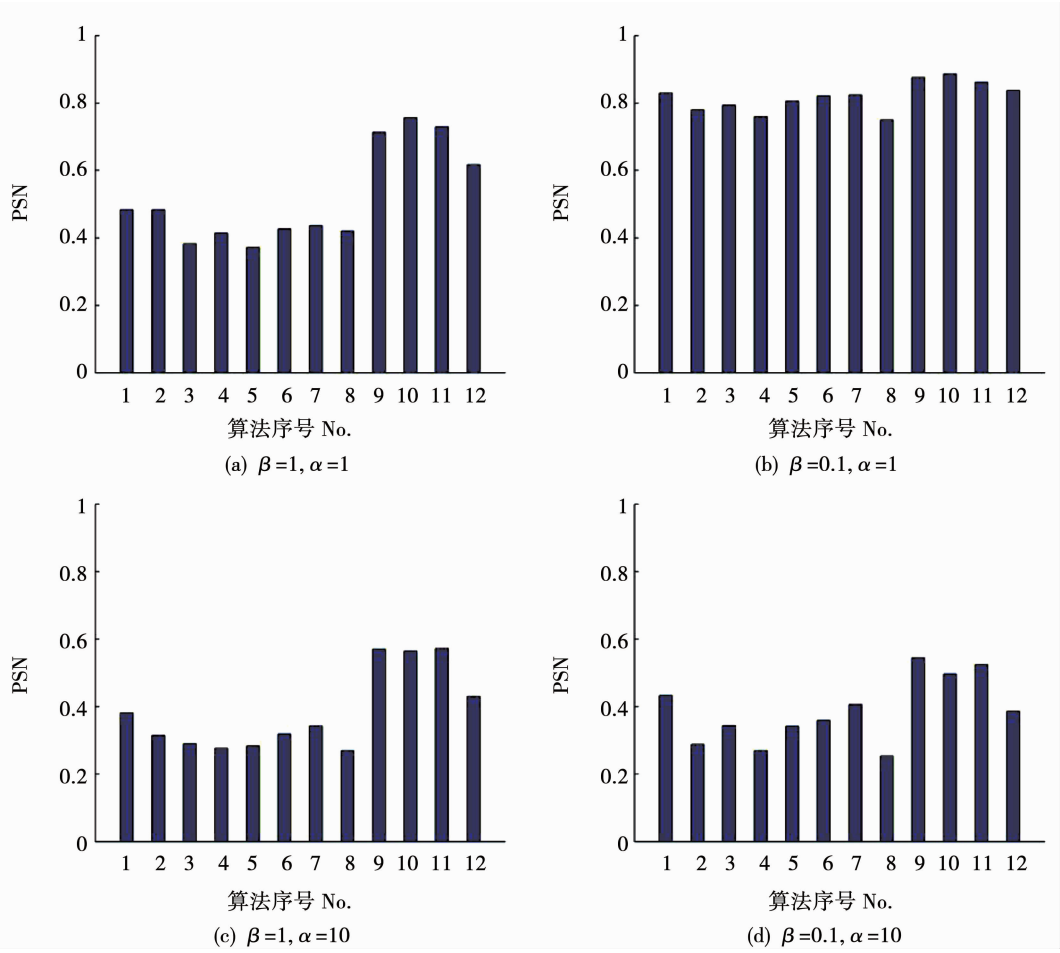


图 7 当 l 为 3 时,不同参数下算法的 PSN 值

Fig. 7 The PSN values of the algorithms with different parameters when $l = 3$

表 1 各种算法的波段数量、平均准确率和稳定性

Table 1 The number of bands, the average accuracy and the stability of the algorithms

序号 No.	算法 Algorithm	邻域 δ	子集大小 Size	平均准确率 Average accuracy	稳定性 Stability		
					$l=1$	$l=3$	$l=5$
1	DMNRS	0.12	8	93.91%	0.66	0.34	0.30
2	DMNRS	0.20	17	97.89%	0.51	0.35	0.26
3	VPNRS	0.15	9	95.49%	0.29	0.25	0.21
4	VPNRS	0.18	15	97.33%	0.30	0.28	0.26
5	CMNRS	0.24	9	97.76%	0.28	0.24	0.22
6	CMNRS	0.27	10	98.02%	0.32	0.29	0.26
7	NMI	0.12	8	95.07%	0.37	0.30	0.26
8	NMI	0.20	17	98.10%	0.33	0.29	0.27
9	MRMRD	0.15	9	94.47%	0.76	0.60	0.48
10	MRMRD	0.23	13	97.62%	0.77	0.66	0.51
11	MRMRQ	0.14	10	93.18%	0.68	0.63	0.56
12	MRMRQ	0.19	15	97.09%	0.57	0.49	0.41

3 结 论

为了解决同种物质的分类难题,本文应用高光谱图像技术实现大豆品种的鉴别检测。利用 NRS 属性约简思想提取特征波段,实现大豆分类特征信息的充分提取。在邻域半径参数选择适当的情况下,利用 ELM 模型进行大豆品种识别,VPNRS、DMNRS、CMNRS、NMI、MRMRD 和 MRMRQ 这 6 种算法

都能取得很好的分类效果,最大分类准确率达到100%,平均分类准确率达到95.60%以上。但在大豆品种分类识别中,除分类性能以外,稳定性也应该作为一个重要评价指标来考虑。在实际应用中,人们通常希望从有一定的物理含义的波段数据中发现一些规律。若波段选择算法的稳定性相对比较低,选择结果会产生很大的变化。对于研究者来说,无法从变化较大的波段子集中获取有规律的信息,会给研究带来困扰,这种波段子集实际意义不大。因此,需要寻求能保持较高的分类精度且波段选择结果稳定的算法。本文在以 Jaccard 系数为稳定性度量指标的基础上,针对波段子集大小相同和不同两种情况,分别提出了 Pareto 最优解和兼顾分类性能、稳定性和子集大小的综合评价函数两种方法来评估算法性能。这对获取综合性能最佳的波段子集有一定的理论及应用价值。

本研究丰富了现有的大豆品种识别技术和手段,以本研究作为基础,建立大豆高光谱无损检测系统,并应用于生产实践中,能实现快速、准确、无损、稳定的检测,对实现农产品检测的自动化和智能化有重要意义,有利于促进大豆种植业持续优质、高产、稳产,保护农民利益,保障国家粮食安全。

参考文献

[1] 魏利峰. 玉米种子高光谱图像品种检测方法研究[D]. 沈阳:沈阳农业大学, 2017. (Wei L F. Research on detection method of maize variety based on hyperspectral image[D]. Shenyang: Shenyang Agricultural University, 2017.)

[2] 朱大洲, 王坤, 周光华, 等. 单粒大豆的近红外光谱特征及品种鉴别研究[J]. 光谱学与光谱分析, 2010, 30(12): 3217-3221. (Zhu D Z, Wang K, Zhou G H, et al. The NIR spectra based variety discrimination for single soybean seed[J]. Spectroscopy and Spectral Analysis, 2010, 30(12): 3217-3221.)

[3] 杨冬风, 朱洪德. 基于近红外透射光谱分析和 BP 神经网络的大豆品种识别[J]. 大豆科学, 2013, 32(2): 249-253. (Yang D F, Zhu H D. Recognition of soybean varieties based on near infrared transmittance spectroscopy and BP neural network[J]. Soybean Science, 2013, 32(2): 249-253.)

[4] 柴玉华, 侯升飞, 彭长禄. 基于高光谱图像技术的大豆分级识别方法研究[J]. 东北农业大学学报, 2014, 45(4): 107-112. (Chai Y H, Hou S F, Peng C L. Identification of different soybean grades based on hyperspectral imagery[J]. Journal of Northeast Agricultural University, 2014, 45(4): 107-112.)

[5] Tan K, Chai Y, Song W, et al. Identification of soybean seed varieties based on hyperspectral image[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(9): 235-242.

[6] Liu Y, Xie H, Chen Y H, et al. Neighborhood mutual information and its application on hyperspectral band selection for classification[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 157(157): 140-151.

[7] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: A study on high- dimensional spaces[J]. Knowledge & Information Systems, 2007, 12(1): 95-116.

[8] 严衍禄, 陈斌, 朱大洲. 近红外光谱分析的原理、技术与应用[M]. 北京: 中国轻工业出版社, 2013. (Yan Y L, Chen B, Zhu D Z. Near infrared spectroscopy-principles, technologies and applications[M]. Beijing: China Light Industry Press, 2013.)

[9] Dong Y, Xiang B, Geng Y, et al. Rough set based wavelength selection in near-infrared spectral analysis[J]. Chemometrics and Intelligent Laboratory Systems, 2013, 126(126): 21-29.

[10] Chen Y, Xue Y, Ma Y, et al. Measures of uncertainty for neighborhood rough sets[J]. Knowledge-Based Systems, 2017, 120(C): 226-235.

[11] Pan X, Zhang S, Zhang H, et al. A variable precision rough set approach to the remote sensing land use/cover classification[J]. Computers & Geosciences, 2010, 36(12):1466-1473.

[12] Hu Q, Pedrycz W, Yu D, et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40(1): 137-150.

[13] Liu Y, Chen Y, Tan K, et al. Maximum relevance, minimum redundancy band selection based on neighborhood rough set for hyperspectral data classification[J]. Measurement Science and Technology, 2016, 27(12): 125501.

[14] 季金胜. 高分辨率遥感影像典型地物目标的特征选择及其稳定性研究[D]. 上海:上海交通大学, 2015. (Ji J S. Feature selection and its stability for typical geoobjects of the high-resolution remote sensing image[D]. Shanghai: Shanghai Jiao Tong University, 2015.)

[15] Kuncheva L I. A stability index for feature selection[C]//Artificial Intelligence and Applications, 2007: 421-427.

[16] 俞研, 黄皓. 面向入侵检测的基于多目标遗传算法的特征选择[J]. 计算机科学, 2007, 34(3):197-200. (Yu Y, Huang H. Feature selection using multi-objective genetic algorithms for intrusion detection[J]. Computer Science, 2007, 34(3):197-200.)

[17] Saeys Y, Abeel T, Peer Y V D. Robust feature selection using ensemble feature selection techniques[J]. Lecture Notes in Computer Science, 2008, 5212: 313-325.

[18] 刘瑶, 谭克竹, 陈月华, 等. 基于分段主成分分析和高光谱技术的大豆品种识别[J]. 大豆科学, 2016, 35(4): 672-678. (Liu Y, Tan K Z, Chen Y H, et al. Variety recognition of soybeans using segmented principal component analysis and hyperspectral technology[J]. Soybean Science, 2016, 35(4): 672-678.)

[19] Liu Y, Xie H, Wang L, et al. Hyperspectral band selection based on a variable precision neighborhood rough set[J]. Applied optics, 2016, 55(3): 462-472.