p 值统计量建模独立性的高光谱波段选择方法

张爱武[1,2], 康孝岩[1,2]

(1. 首都师范大学 三维信息获取与应用教育部重点实验室,北京 100048;
2. 首都师范大学 空间信息技术教育部工程研究中心,北京 100048)

摘 要: 近年来,*p* 值统计量的使用规范引起了统计学界的极大关注和集中讨论,广泛认为,*p* 值统计量可表达观测数据与备择假设之间的不相容程度。为探究高光谱图像波段的相关分析 *p* 值与其样本独立性的联系,进行了演绎推理和实例验证,研究表明,与相关系数 *r* 统计量相比,相关分析 *p* 统计量可直接表达波段样本的独立性,且 *p* 值矩阵具有高水平的自稀疏性,便于建模和计算。进而,对相关性 *p* 值矩阵进行直方图频数统计,提出一种基于 *p* 值的高光谱自适应波段选择方法 *p*SMBS。选取典型数据进行了监督分类实验,结果表明,在 Kappa 系数、总体精度(OA)和平均精度(AA)上,*p*SMBS 均优于同类方法 ABS、InfFS 和 LSFS。说明 *p*SMBS 在高光谱波段选择方面具有突出的有效性,这也佐证了相关性 *p* 值对波段独立性的强表征能力。

关键词: *p* 值统计量; 波段独立性; 自稀疏性; 非监督波段选择; 高光谱

中图分类号: TP753 **文献标志码:** A **DOI:** 10.3788/IRLA201847.0926005

Hyperspectral images band selection algorithm through *p*-value statistic modeling independence

Zhang Aiwu[1,2], Kang Xiaoyan[1,2]

(1.Key Laboratory of 3D Information Acquisition and Application, Ministry of Education, Capital Normal University, Beijing 100048, China; 2. Engineering Research Center of Spatial Information Technology, Ministry of Education, Capital Normal University, Beijing 100048, China)

Abstract: The usage specifications of *p*-value statistic were stimulated by highly visible discussions in the field of Statistics over the last few years. It is generally considered that a *p*-value can indicate how incompatible sample data are with the alternative hypothesis model. To explore the connection between the *p*-value of correlation analysis and spectral independence, the deductive reasoning and example verification were carried out. Compared with correlation coefficient (*r*-value statistic), results show that the band independence can be directly expressed by *p*-value statistic of correlation analysis. And *p*-value matrix has a kind of high-level self-sparsity, which can be used to model easily. And then an unsupervised band selection method (*p*-value sparsity matrix band selection, *p*SMBS) through *p*-value statistic modeling independence was proposed, based on the histogram frequency statistics of *p*-value matrix. Using two typical hyperspectral images (HSI) data, the experiments of supervised classification

收稿日期: 2018-04-07; **修订日期:** 2018-05-12

基金项目: 国家自然科学基金面上项目(41571369);国家重点研发计划项目(2016YFB0502500);青海省科技计划项目(2016-NK-138)

作者简介: 张爱武(1972-),女,教授,博士,主要从事空间信息获取与处理、计算机视觉与模式识别、图像处理等方面的研究。
Email: zhangaw98@163.com

were carried out. The results indicate that, on Kappa coefficient, overall accuracy (OA) and average accuracy (AA), *p*SMBS is superior to three kinds of methods, adaptive band selection (ABS), infinite feature selection (InfFS) and Laplacian score feature selection(LSFS). Therefore, the effectiveness and the practicability of *p*SMBS were verified on HSI band selection, and the characterization ability of *p*−value of correlation analysis on expressing band independence was evidenced.

Key words: *p*−value statistic; band independence; self−sparsity; unsupervised band selection; hyperspectral image

0 引 言

限于“大数据量、高冗余度”的特点,使得高光谱图像不易于被高效率和高精度地实现解混、分类、目标检测和物理量反演等典型应用,而降维是有效解决该问题的主要手段之一。作为降维的两种主要实现方式之一,波段选择以寻求“信息量大、独立性强”的特征波段来实现特征空间的简化[1−2]。按照侧重点不同,可将波段选择方法分为两类,一类侧重于提高算法精度,该类方法可以损失一定运算时间为代价,来换取精度性能的大幅提升,其建模的基础理论有遗传进化、稀疏表示和流形学习等;另一类则侧重于算法效率,在满足精度要求下,尽量缩短运行时间,其往往通过时间复杂度较低的经典统计量来实现。其中,后者为该研究所强调和关注,该类方法算法简单易行,可达到即时甚至实时的算法效果[3]。

在高光谱波段选择算法中,常用的参数统计量有均值、标准差、相关系数 r 和信息熵等。其中,谱间相关系数 r 常被用于间接表达波段的独立性[4−5],但少有文献关注波段间 r 为负值的情况及其处理方式;对相关系数检验的 p 值统计量的应用仅限于作为验证 r 值显著性水平的工具,而未对 p 值进一步地进行信息挖掘。近来一些学者针对 p 值统计量的含义和适用条件进行了梳理和讨论,美国统计学会发表了官方声明,阐述了使用 p 值的原则和使用条件[6−7],并提出 p 值大小可以表达样本数据与给定模型的相容/不相容程度(与原假设模型的相容程度、与备择假设模型的不相容程度)。基于此,该研究重点关注波段相关分析中的 p 值矩阵,挖掘其表达波段独立性的能力;然后结合直方图频率法,提出一种 p 值统计量的波段选择方法;最后,选用典型的同类方法,通过图像分类实验,对比验证该研究提出方法的有效性。

1 Pearson 线性相关系数的 p 值统计量

Pearson 线性相关系数 r 是由统计学家 Karl Pearson 所提出的,用于表达二元序列 $\{(x_i, y_i)\}_{i=1}^{N}$ 的线性相关程度的经典工具,其公式如下:

$$r=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \tag{1}$$

式中:N 为样本总数;$\bar{x}$和$\bar{y}$分别为 $\{x_i\}$ 和 $\{y_i\}$ 的均值。r 取值为[−1, 1],$|r|$越大表示变量之间的线性相关性越大;反之亦然。

p 值是由统计学家 Sir Ronald Aylmer Fisher (以下简称 Fisher) 所提出的,表达在原假设(Null Hypothesis)为真时,出现与当前观测结果相同或更极端情况出现的概率。在相关分析的假设检验中,原假设为无相关(No Correlation),此时,p 值是样本的一种统计量,其大小通过构造一个自由度为 v 的 t 统计量进行求解[8]:

$$p=1-A(t|v) \tag{2}$$

$$A(t|v)=\frac{1}{v^{1/2}B\left(\frac{1}{2},\frac{v}{2}\right)}\int_{-t}^{t}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}\mathrm{d}x \tag{3}$$

$$t=r\sqrt{\frac{N-2}{1-r^2}} \tag{4}$$

$$v=N-2 \tag{5}$$

式中:贝塔函数 $B\left(\frac{1}{2},\frac{v}{2}\right)=\int_0^1 x^{-1/2}\left(1-x\right)^{\frac{v}{2}-1}\mathrm{d}x$,另外,也可以通过 Γ 函数来间接求解 B 函数。并且规定 $A(0|v)=0$,$A(\infty|v)=1$。故而,p 值的值域为[0,1]。

在经典的原假设显著性检验中,除了原假设,往往还建立备择假设(有相关),并根据 p 值大小是否小于

显著性水平来判断 r 是否显著:当 p 足够小时,如小于0.05,则拒绝原假设,接受备择假设,表达 r 是显著的;而相反时,则接受原假设,表达 r 是不显著的。

2 谱间相关性 p 值矩阵及其性质

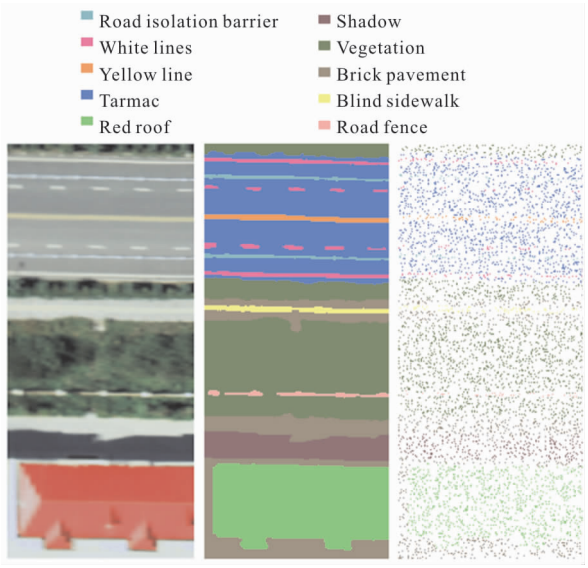
近年来,国内外统计学界的一些学者提出了 p 值统计量应用时的六个原则[6-7],明确了其含义和适用条件。经过大量文献研读,笔者研究发现在已有的波段选择方法中,甚至在整个光谱分析领域内,鲜见 p 值矩阵的信息挖掘及应用。鉴于此,文中首先将 p 值的使用原则引入到高光谱谱间相关性的分析中,从统计学原理上厘清谱间相关分析 p 值与波段样本独立性和波段本身独立性的表征关系;然后,选用样例来揭示和验证 p 值对波段样本独立性的表征能力以及 p 值矩阵的自稀疏性特点;最后,基于谱间相关性 p 值矩阵提出一种波段选择方法。

2.1 谱间相关性 p 值及对波段样本独立性的表达

参考文献[6-7]表明,p 值不能衡量某种假设为真的概率,但可以揭示样本数据与指定模型的相容/不相容程度;换句话说,p 值不解释假设本身,但可表达样本数据与假设之间的关系。诠释到谱间相关分析,以波段 M 和 N 的样本数据的相关分析为例,可以表述为,p 值的大小不解释波段之间是否独立或相关,但可以表达样本数据与零假设(假设 M 与 N 独立)之间的相容程度。简言之,p 值可以表达两波段的样本数据(而非两波段本身)之间的独立程度,即 p 值越大,样本之间的独立性越强;反之亦然。在此研究中,仅关注“p 值对波段的样本数据独立性的表达”;而对“p 值能否衡量波段本身的独立性”的讨论不在文中的研究之列。故为便于描述和理解,约定此研究中的波段独立性即为波段样本的独立性,波段之间的相关分析即为波段间的样本相关分析。

为了定量和直观地探讨 p 值的内涵,该研究选用两组高光谱影像作为实验样例:

(1) 样例一:数据来源于艇载可见/近红外短波高光谱相机探测的青海海晏县西海镇影像(图 1),截取大小为 900×400,波长范围为 0.4~1.0 μm,共800个波段。样例中包含了道路隔离护栏、道路白线、道路黄线、柏油路面、红色屋顶、阴影、植被、地砖路面、黄色盲道和道路围墙等 10 种覆盖目标。

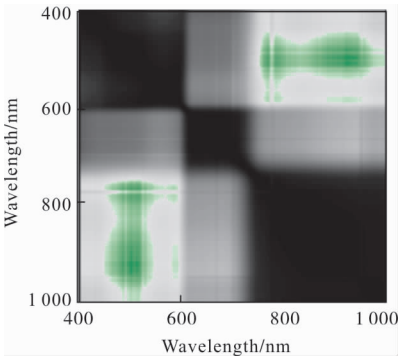


(a) 假彩色图像 (b) 地表参照 (c) 随机采样(训练样本)
(a) Pseudo color image sampling (b) Ground truth (c) Random sampling (training sample)

图 1 样例一

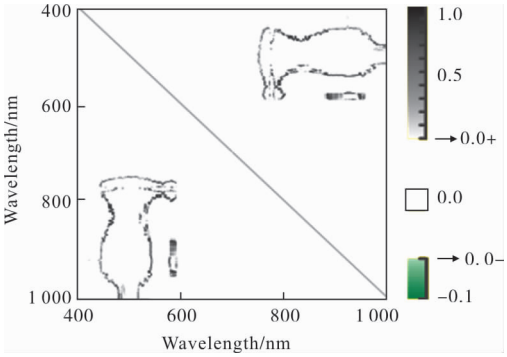
Fig.1 Sample 1

按照公式(1)~(5)对样例一进行两两波段之间的相关分析,以得到 r 矩阵和 p 矩阵,如图 2 所示,可视化矩阵中的任意一点表示两个波段之间的相关系数及显著性检验 p 值。研究发现,(1) 从图 2(a)可见,样例一波段之间存在负相关的情况(约占比 10%左右),这在前人研究中较少被提及;(2) 从图 2 (b)可见,p 值矩阵中的大部分元素为零值,而非零值则遍及(0,1)的值域范围;(3) 结合图 2(a)、(b)可见,非零值的 p 值主要对应于 r 矩阵中正负值相间位置,换句话说,r 绝对值大,则 p 值显示为零,而 r 趋近于零,则 p 由 0 逐渐趋近于 1。



(a) 相关性 r 矩阵

(a) Correlation matrix r



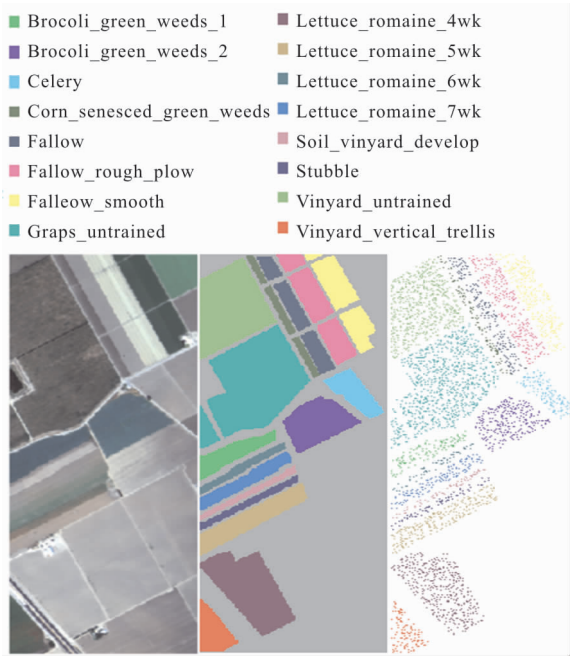
(b) 显著性 p 矩阵

(b) Significant level matrix p

图 2 样例一的相关分析

Fig.2 Correlation analysis of sample 1

(2) 样例二：数据来源于 AVIRIS 传感器探测的美国加州的 Salinas 山谷影像(图 3)，图像大小为 512 × 217，源数据共有 224 个波段，在除去水吸收噪声波段后余下 204 个波段，波谱范围为 0.4~2.5 μm，样例中覆盖有椰菜和绿草 1、椰菜和绿草 2、芹菜、成熟玉米和绿草、休耕地、粗糙的休耕地、平整的休耕地、葡萄藤、4 周的莴苣、5 周的莴苣、6 周的莴苣、7 周的莴苣、在开发的园壤土、残株、未结果实的葡萄园和葡萄园小路等 16 种目标。



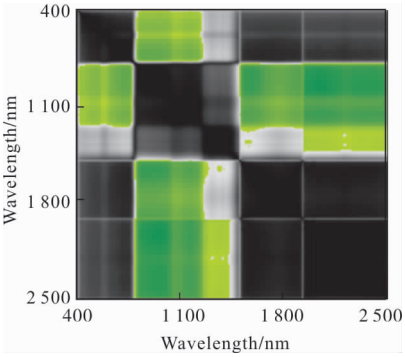
(a) 假彩色图像 (b) 地表参照 (c) 随机采样(训练样本)

(a) Pseudo color image (b) Ground truth (c) Random sampling (training sample)

图 3 样例二

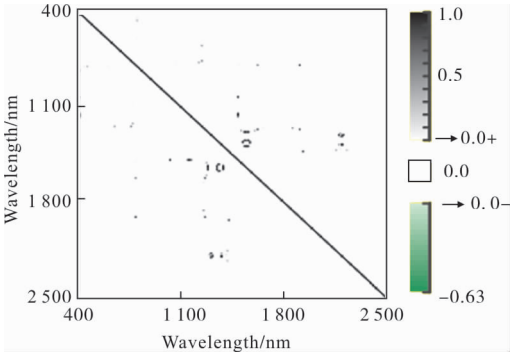
Fig.3 Sample 2

与样例一同理，样例二的 r 矩阵和 p 矩阵如图 4 所示。研究发现样例二表现出与样例一相似的规律：(1) 波段之间存在负相关的情况(约占比 30%左右)；(2) p 值矩阵中零值占了绝大部分，而非零值则遍及值域范围；(3) 非零值的 p 值主要对应于 r 矩阵中正负值相间位置，并随着$|r|$的增加而变小。



(a) 相关性 r 矩阵

(a) Correlation matrix r



(b) 显著性 p矩阵

(b) Significant level matrix p

图 4 样例二的相关分析

Fig.4 Correlation analysis of sample 2

两组不同传感器的高光谱样例表现出相同的结果：谱间相关分析中的 p 矩阵表现出与 r 矩阵强健的关系，p 值能一定程度上反映出$|r|$的大小，即 p 值越大，则$|r|$越小。谱间相关性与波段独立性的对应关系已是光谱分析领域的基本共识[3-5]，那么，结合统计学界对 p 值含义的研究前沿及上述研究结果，该研究认为谱间相关性 p 值可以表达波段独立性。

2.2 相关性 p 值矩阵的自稀疏性

在信号处理领域中，稀疏性是指原始信号进过变换后，大部分变换系数为 0 或接近于 0，而存在少量非零的变换系数的性质[9]；而矩阵的自稀疏性指的

是在未经稀疏变换时矩阵便呈现出较高的稀疏水平(较低的稀疏度)的性质。由此可见,该研究中两个样例的谱间相关性 p 值矩阵均具有典型的自稀疏性,如表 1 所示;并且,该研究认为 p 值矩阵的这种典型自稀疏性是高光谱谱间强相关性的必然结果,而非样例选择的偶然性所致。

表 1 样例影像的谱间相关分析结果统计

Tab.1 Results′ statistics of samples′ correlation analysis

Data set	Matrix r			Matrix p		
	Positive value	Negative value	Zero	Positive value	>0.05	Zero
Sample 1	88.60%	**11.40%**	0	23.60%	**1.46%**	**76.40%**
Sample 2	66.77%	**33.23%**	0	8.72%	**0.31%**	**91.28%**

对比可见,r 矩阵中没有零值,但有不小比例的负值;而 p 值矩阵中零值占了大部分,大于 0.05 的元素也占有不小的比重。这说明,r 值在表达波段独立性时,实质上是以|r|与独立性相互对应的,相比而言,p 值则可更直接地表征波段独立性;并且,未通过显著性检验的 r 理论上不应被使用,但在前人的研究中不显著的 r 并未被明确剔除;更为重要的一点,高水平的稀疏程度使得 p 值矩阵在同等条件下比 r 矩阵的处理效率高。

2.3 相关性 p 值建模独立性的波段选择算法

由上文可知,与 r 值矩阵相比,p 值在表达波段独立性时更为简便、直接,复杂度也更低。基于此,该研究利用相关性 p 值进行建模,来提出一种非监督的波段选择方法,其基本思路是选取出现独立性强(p 值大)且频率高的波段作为重要波段。利用 p 值稀疏矩阵提出的 pSMBS(p Value Sparse Matrix Band Selection)算法 1 的伪代码如下。

算法 1:相关性 p 值建模独立性的高光谱波段选择

输入:

含有 n 个波段的高光谱影像 $H=\{H_1, H_2, H_3, \cdots, H_n\}$;

欲选出的波段数量 k

输出:

选出的 k 个波段的集合

1:根据公式(1)计算高光谱影像的谱间相关系数 r 矩阵;

2:根据公式(2)~(5)计算谱间相关性 p 值矩阵 $M_{n\times n}$,去掉波段与其自身的 p 值后得到 $M'_{(n-1)\times n}$;

3:对 $M'_{(n-1)\times n}$ 各列降序排列,选择前 k 行组成矩阵 $MS_{k\times n}\in R^{k\times n}$,并将其对应的波段号组成矩阵 $B_{k\times n}$;

4:求 $B_{k\times n}$ 中相同波段号对应的 p 值之和并降序排序,返回前 k 个对应的波段号。

依据谱间相关性 p 值矩阵,首先,pSMBS 算法获取与任意单一波段 i (i = 1, 2, …, n) 独立性强(p 值大)的前 k 个波段的波段号及前 k 个 p 值,终可得到 $n\times k$ 个波段号(即 $B_{k\times n}$)及对应 p 值(即 $MS_{k\times n}$);然后,构造波段选择的参考值目标函数 $f_k(i)$:

$$f_k(i)=\sum_{u=1}^{k}\sum_{v=1}^{n}\mathrm{MS}(B_{uv}=i) \tag{6}$$

$f_k(i)$是 B 中的元素 i 在 MS 对应位置上的 p 值之和,表示在选择 k 个波段时,第 i 波段的选择参考值;最后对 $f_k(i)$(i=1, 2, …, n)进行降序排列,选择前 k 个值对应的波段。

3 波段选择实验

为了客观地探讨算法的有效性,文中选用了兼顾信息量和相关性的自适应波段选择方法 ABS (Adaptive Band Selection)[4,10]、基于拉普拉斯映射和局部保持投影的打分方法 LSFS(Laplacian Score Feature Selection)[11]和基于带权有向图的波段选择方法 InfFS(Infinite Feature Selection)[12]等同类算法中 3 种具有代表性的非监督方法进行对比实验。其中,ABS 可选出信息量高且相关性低的波段;LSFS 可以有效选出体现高光谱数据潜在流形结构的波段;而 InfFS 则从分类的角度,利用标准差和秩相关系数来揭示波段重要性。

采用上文两组数据为实验样例,以 5 为步长,对每种方法选择了 20 组子集(5–100);选取马氏距离(MDC, Mahalanobis Distance Classifier) 和随机森林(RFC, Random Forest Classifier)两种分类器对每个子集进行了监督分类 (随机选择小部分实况数据作为训练样本 (样例一选 10%(图 1 (c));样例二选 15%(图 3(c)),剩余部分作为测试样本),并分别统计了 Kappa 系数、总体分类精度 OA (Overall Accuracy) 和平均分类精度 AA (Average Accuracy),以对分类精度进行定量评价。

3.1 监督分类实验

3.1.1 样例一

监督分类的精度对比如图 5 所示。

样例一在两种分类方法下,均明显显示出文中研究所提方法 pSMBS 的优势,在 Kappa 系数、OA 和 AA 上,pSMBS 均高于 ABS、InfFS 和 LSBS 等三

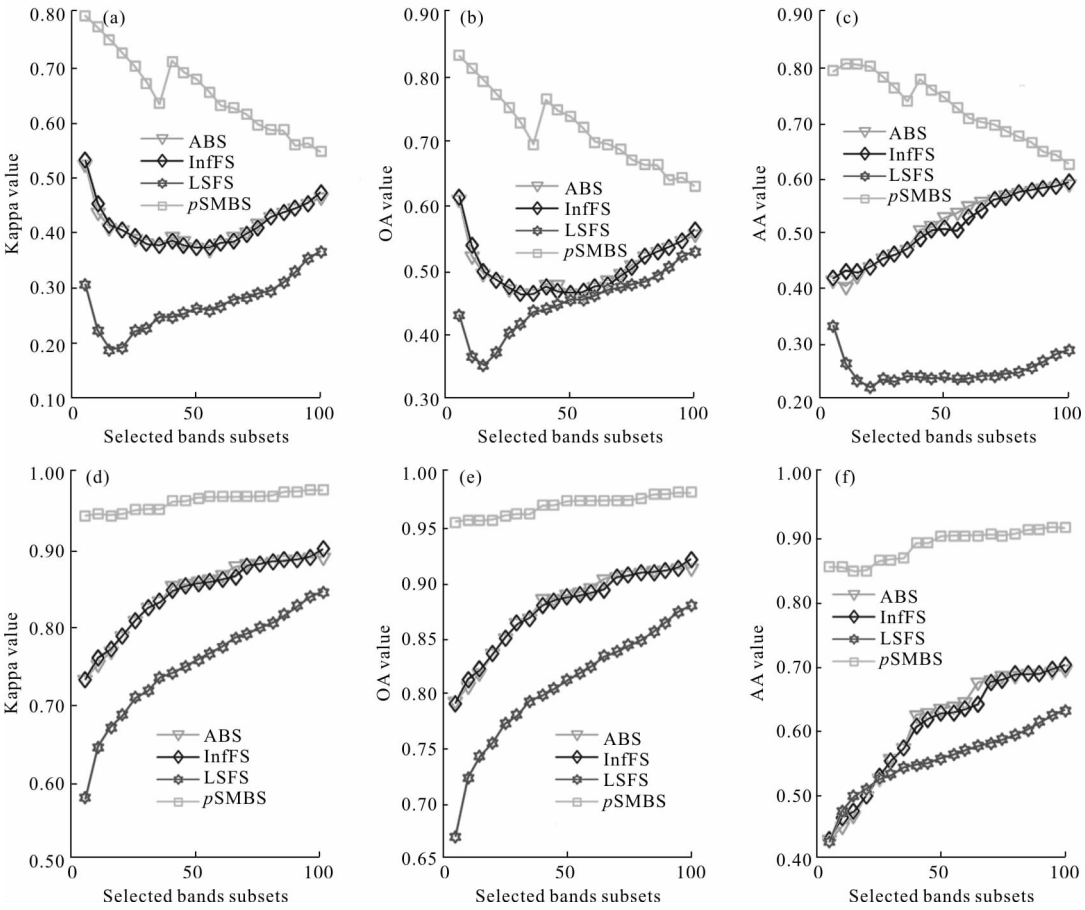


图 5 样例一的分类精度结果

Fig.5 Classification accuracy results of sample 1

种同类对比方法。在马氏距离分类器下:(1) *p*SMBS 在第 1 组子集(5 个波段)中便达到了高水平的精度(Kappa=0.790 3, OA=0.829 2 , AA=0.790 0),随后逐渐减小,但直到第 20 子集(100 个波段)仍高于其他三种方法;(2) ABS 和 InfFS 表现出相似的性能,Kappa 和 OA 均为先降低后升高,AA 则平稳增大;(3) LSFS 的性能较差。而随机森林分类器下, 四种算法的分类性能均大幅提升,但(1) *p*SMBS 性能优势仍较明显,且稳定性最高,在第 1 子集达到较高水平, 之后略有缓慢提高;(2) 三种对比方法中,ABS 和 InfFS 的性能仍趋同,均总体高于 LSFS。

表 2 进一步展示了各方法的 20 组的均值情况,可以看出两种分类器下, 较之三种对比方法, *p*SMBS 均表现出优异的性能(黑体表示最优,下划线表示次优, 下同):*p*SMBS 分别比次高高 0.240 5 (MDC Kappa)、0.215 5 (MDC OA)、0.211 3(MDC AA)、0.118 0 (RFC Kappa)、0.092 1 (RFC OA)、0.280 1(RFC AA)。

表 2 样例一 20 组波段子集的平均分类性能

Tab.2 Average classification performance of 20 subsets in sample 1

Method	MDC			RFC		
	Kappa	OA	AA	Kappa	OA	AA
ABS	0.410 8	0.498 7	<u>0.513 9</u>	<u>0.840 0</u>	<u>0.874 8</u>	<u>0.602 7</u>
InfFS	<u>0.411 2</u>	<u>0.499 3</u>	0.510 2	0.839 1	0.874 1	0.599 3
LSFS	0.268 0	0.446 0	0.251 5	0.747 5	0.803 4	0.549 7
*p*SMBS	**0.651 7**	**0.714 8**	**0.725 2**	**0.958 0**	**0.966 9**	**0.882 8**

3.1.2 样例二

监督分类结果如图 6 所示。

从图 6 中可以看出,*p*SMBS 与 ABS、InfFS 二法的性能相似,总体略高于后两者;而 LSFS 性能相对较差。从 MDC 看,(1) 前三者的 Kappa、OA 和 AA

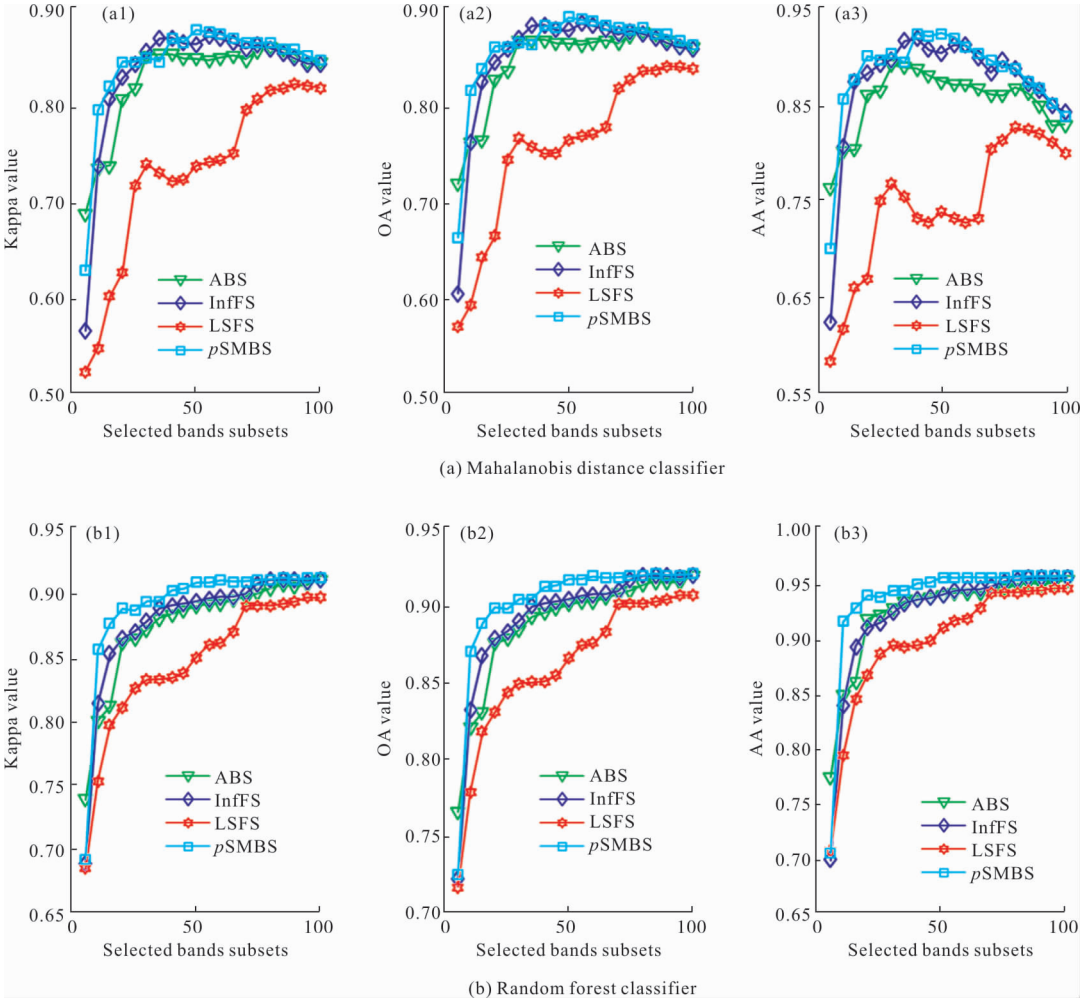


图 6 样例二的分类精度结果

Fig.6 Classification accuracy results of sample 2

三种指标均从起点快速达到较高水平，而后均略有下降；(2) ABS 仅在第 1 子集中高于 *p*SMBS 和 InfFS，在其他各子集均低于后两者；(3) *p*SMBS 仅在个别子集（第 5~7）上低于 InfFS，其他均高于 InfFS。从 RFC 看，(1) 四种方法的精度均随子集波段数量的增加而提高，总体而言，*p*SMBS 最优，ABS 和 InfFS 次之，LSFS 最弱；(2) 除第 1 子集上 *p*SMBS 为次优之外，其他各子集上，*p*SMBS 均为最优；(3)ABS 和 InfFS 的性能相当，在 Kappa 和 OA 上 InfFS 略占优势，而在 AA 上，ABS 略占优势。

从表 3 的平均分类性能上看，*p*SMBS 的精度性能有着相对优势，各指标均值为最优，在 MDC 下，分别比次优高 0.009 8(Kappa)、0.009 4(OA)、0.010 2(AA)；在 RFC 下，分别比次优高 0.011 2(Kappa)、0.010 0(OA)、0.013 1(AA)。

表 3 样例二20 组波段子集的平均分类性能

Tab.3 Average classification performance of 20 subsets in sample 2

Method	MDC			RFC		
	Kappa	OA	AA	Kappa	OA	AA
ABS	0.827 5	0.845 5	0.854 4	0.873 0	0.886 1	0.921 6
InfFS	0.833 6	0.850 3	0.872 0	0.877 1	0.889 8	0.919 5
LSFS	0.729 1	0.756 7	0.744 6	0.843 0	0.859 3	0.895 7
*p*SMBS	**0.843 4**	**0.859 7**	**0.882 2**	**0.888 3**	**0.899 8**	**0.934 7**

3.2 波段选择时间对比

表 4 为 *p*SMBS 与三种对比方法 ABS、InfFS 和 LSBS 的算法时间复杂度及两个样例的波段选择时间(表中，n 为波段总数；T 为像元点总数；k 为波段选择的数量)。每种算法在同一样例均运行 10 次，取时

间均值和标准差。运行硬件环境为戴尔 i7-6700 四核处理器,8 GB 内存;软件环境为 Windows 7 操作系统,MATLAB 2015a 平台。其中,LSFS 算法需获取样本(像元点)之间的亲和矩阵,但受限于硬件环境(内存),实践中只能选取部分样本点求解,故而参考文献[13]指出 LSFS 的算法复杂度是不明确的(*N*/*A*)。该研究随机选取 15 000 个样本点(此时仅亲和矩阵约需占 1.676 G 的内存) 进行 LSFS 的运算;而其他三种方法选用全体样本点, 三者的空间复杂度远低于 LSFS。

表 4 四种算法的时间复杂度及在样例中的计算时间对比

Tab.4 Contrast in computational complexity and computational time among the four algorithms on two samples

Method	Computational complexity	Computational time/s	
		Sample 1	Sample 2
ABS	$O(n^2T+nT+n^2)$	5.949 1±0.084 8	0.261 1±0.008 6
InfFS	$O(n^{2.37}(1+T))$[12]	1 192.862 8± 33.467 7	27.149 2±0.858 7
LSFS	N/A[13]	62.122 5±1.062 2	51.683 7±1.977 2
*p*SMBS	$O(n^2T+n^3+kn^2)$	**3.668 7**±0.090 3	**0.175 2**±0.012 6

一般的,$k<<n<<T$,因此从表中可看出,在时间复杂度上,*p*SMBS 与 ABS 相当, 两者均远小于 InfFS。从两个样例的波段选择时间上看,*p*SMBS 略小于 ABS, 且两者运算时间均不到 InfFS 用时的 1%,也均远小于 LSFS 的计算时间。

3.3 结果分析

两组样例高光谱影像数据,各 20 组子集的分类结果表明, 文中提出的 *p*SMBS 算法均优于 ABS、InfFS 和 LSFS 等三种对比方法, 验证了 *p*SMBS 的有效性。同时,*p*SMBS 的时间复杂度较低,可以满足高光谱波段选择的时效性。

选用经典的个体分类器 MDC 和组合分类器 RFC 等两种分类器进行对比, 表明了算法性能的健壮性;多组子集的实验表明,较之三种对比方法,在选择较少波段时,*p*SMBS 更易表现出最优的分类性能,且性能相对稳定。

波段相关性常被用于表达波段的独立性, 即相关性越小,则独立性越强[1-2,4-5,10,14];图 7(a1)和(b1)分别是两个样例各波段子集的谱间相关系数绝对值的均值, 可以明显看出,*p*SMBS 所选波段子集的谱间相关性均明显低于其他三种对比方法, 尤其是以兼顾信息量和相关性为目标的 ABS, 其谱间相关性均值也远高于 *p*SMBS。

同时,*p*SMBS 算法未考虑信息量的表达, 两组实验的结果(图 7(a2)和(b2))显示出 *p*SMBS 对信息量表达的不稳定性。

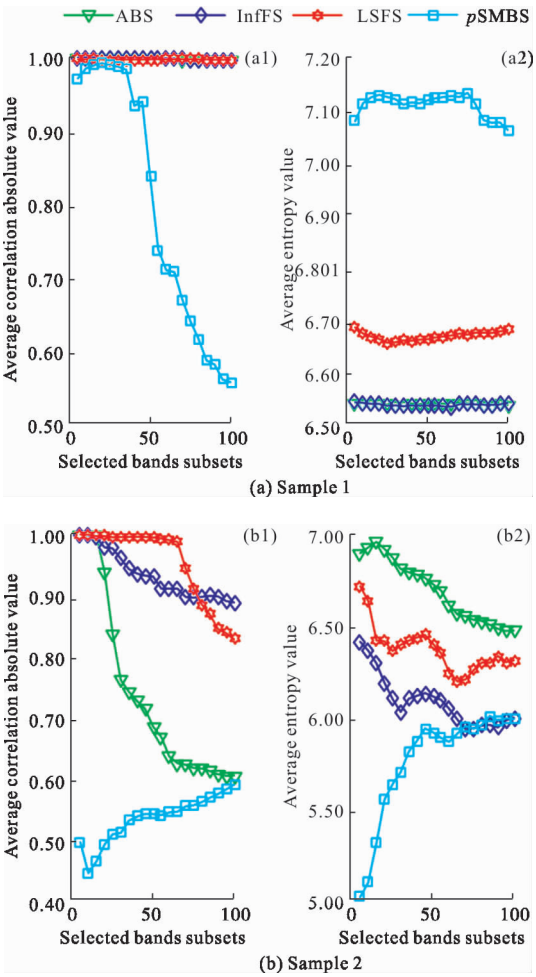


图 7 样例一的(a1)谱间相关系数均值和(a2)波段信息熵均值;样例二的(b1)谱间相关系数均值和(b2)波段信息熵均值

Fig.7 Average correlation coefficients (a1) and average entropy (a2) of the subsets of sample 1, average correlation coefficients (b1) and average entropy (b2) of the subsets of sample 2

4 结 论

p 值统计量的前沿理论认为其可以表达样本数

据与零假设的相容程度,本研究将其应用于高光谱谱间相关性分析中,原理推导和实证分析结果显示,(1)与谱间相关系数 r 矩阵相比,相关性 p 值矩阵更易表达波段的独立性,即 p 值越大,则波段独立性越强;(2) p 值矩阵具有高水平的自稀疏性,使得建模算法简单易算,算法复杂度低。进而,提出了一种 p 值建模独立性的波段选择方法 pSMBS,相对详实的监督分类实验表明,pSMBS 较好地度量了波段独立性,算法性能总体上高于三种同类算法,可行性较好。

此外,pSMBS 算法仅基于波段独立性建模,缺乏对波段信息量的考量,在下一步的工作中将结合信息量进一步深化非监督模型的研究。

参考文献:

[1] Qin Fangpu, Zhang Aiwu, Wang Shumin, et al. Hyperspectral band selection based on spectral clustering and inter-class separability factor [J]. ***Spectroscopy and Spectral Analysis***, 2015, 35(5): 1357-1364. (in Chinese)
秦方普, 张爱武, 王书民, 等. 基于谱聚类与类间可分性因子的高光谱波段选择 [J]. 光谱学与光谱分析, 2015, 35(5): 1357-1364.

[2] Sui C, Tian Y, Xu Y, et al. Unsupervised band selection by integrating the overall accuracy and redundancy [J]. ***IEEE Geoscience & Remote Sensing Letters***, 2015, 12(1): 185-189.

[3] Chang C I. Hyperspectral Data Processing: Algorithm Design and Analysis [M]. Hoboken, NJ: Wiley-Interscience, 2013.

[4] Zhang Aiwu, Du Nan, Kang Xiaoyan, et al. Hyperspectral adaptive band selection method through nonlinear transform and information adjacency correlation [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0538001. (in Chinese)
张爱武, 杜楠, 康孝岩, 等. 非线性变换和信息相邻相关的高光谱自适应波段选择[J]. 红外与激光工程, 2017, 46(5): 0538001.

[5] Gu Y, Zhang Y. Unsupervised subspace linear spectral mixture analysis for hyperspectral images [C]// International Conference on Image Processing, 2003. ICIP 2003. Proceedings. IEEE, 2003, 1: 801-804.

[6] Wasserstein R L, Lazar N A. The ASA's statement on p-values: context, process, and purpose [J]. ***American Statistician***, 2016, 70(2): 129-133.

[7] Nuzzo R. Statistical errors [J]. ***Nature***, 2014, 506(2): 150-152.

[8] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical Recipes in C: The Art of Scientific Computing [M]. 2nd ed. Cambridge: Cambridge University Press, 1992.

[9] Jiao Licheng, Zhao Jin, Yang Shuyuan, et al. Research advances on sparse cognitive learning, computing and recognition[J]. ***Chinese Journal of Computers***, 2016, 39(4): 835-852. (in Chinese)
焦李成, 赵进, 杨淑媛, 等. 稀疏认知学习、计算与识别的研究进展[J]. 计算机学报, 2016, 39(4): 835-852.

[10] Liu Chunhong, Zhao Chunhui, Zhang Lingyan. A new method of hyperspectral remote sensing image dimensional reduction [J]. ***Journal of Image and Graphics***, 2005, 10(2): 218-222. (in Chinese)
刘春红, 赵春晖, 张凌雁. 一种新的高光谱遥感图像降维方法[J]. 中国图象图形学报, 2005, 10(2): 218-222.

[11] He X, Cai D, Niyogi P. Laplacian score for feature selection [C]//International Conference on Neural Information Processing Systems, 2005: 507-514.

[12] Roffo G, Melzi S, Cristani M. Infinite Feature Selection [C]//IEEE International Conference on Computer Vision. IEEE Computer Society, 2015: 4202-4210.

[13] Roffo G, Melzi S. Ranking to Learn: Feature Ranking and Selection via Eigenvector Centrality [M]//New Frontiers in Mining Complex Patterns. Berlin Heidelberg: Springer, 2017: 19-35.

[14] Zhao Huijie, Li Mingkang, Li Na, et al. A band selection method based on improved subspace partition [J]. ***Infrared and Laser Engineering***, 2015, 44(10): 3155-3160. (in Chinese)
赵慧洁, 李明康, 李娜, 等. 一种基于改进子空间划分的波段选择方法[J]. 红外与激光工程, 2015, 44(10): 3155-3160.