Pure Mathematics 理论数学, 2023, 13(10), 2755-2763
Published Online October 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.1310283

Hans汉斯

分组式OIF在高光谱图像分类的应用

周安旭

成都理工大学数理学院，四川 成都

收稿日期：2023年8月30日；录用日期：2023年9月30日；发布日期：2023年10月7日

摘　要

高光谱图像具有高维度和高相关性，导致“维度灾难”和计算成本高。本文利用波段的标准差和相关系数，选择信息量大且相关性小的波段作为特征波段。为克服原始OIF难以在高光谱图像中采用的困境，提出分组式OIF (Grouping OIF, G-OIF)将高光谱图像分为若干子集，分别计算每个子集的最佳波段组合，然后并集得到整个图像的最佳波段组合。使用Indian Pines数据集，采用随机森林和支持向量机作为分类器，比较不同的分组和波段数对分类效果的影响。最后发现使用G-OIF时分组越多，波段数越多，分类效果越好。G-OIF能够在保证精度的同时实现降维，并缓解“维度灾难”。

关键词

高光谱图像，波段选择，维度灾难，图像分类

Application of Grouped OIF in Hyperspectral Image Classification

Anxu Zhou

College of Mathematics and Physics, Chengdu University of Technology, Chengdu Sichuan

Received: Aug. 30th, 2023; accepted: Sep. 30th, 2023; published: Oct. 7th, 2023

Abstract

Hyperspectral images are highly dimensional and highly correlated, leading to the “curse of dimensionality” and high computational cost. In this paper, the standard deviation and correlation coefficient of the bands are used to select the bands with large amount of information and low correlation as the characteristic bands. In order to overcome the dilemma that the original OIF is difficult to use in hyperspectral images, a grouping OIF (Grouping OIF, G-OIF) is proposed to divide hyperspectral images into several subsets, calculate the best band combination for each subset,

文章引用: 周安旭. 分组式 OIF 在高光谱图像分类的应用[J]. 理论数学, 2023, 13(10): 2755-2763.
DOI: 10.12677/pm.2023.1310283

and then combine Get the best band combination for the entire image. Using the Indian Pines dataset, random forest and support vector machine are used as classifiers to compare the effects of different groups and band numbers on classification performance. Finally, it is found that when using G-OIF, there are more groups, more bands and better classification effect. G-OIF can achieve dimensionality reduction while ensuring accuracy, and alleviate the "curse of dimensionality".

Keywords

Hyperspectral Image, Band Selection, Curse of Dimensionality, Image Classification

1. 引言

高光谱遥感图像(Hyperspectral Image, HSI)是由数百个连续窄带组成的图像，具有很高的光谱相关性，HIS 的高维度数带来了“维度灾难”问题，即在固定的少量样本情况下，当维度降低时，HIS 的分类精度会随着维度增加而下降[1] [2] [3]。在遥感地物分类的应用中，大量的训练样本是昂贵且耗时的，甚至不可能。因此一般考虑利用其中具有代表性的波段作为“特征波段”。“特征波段”的产生一般包括两种技术方法：波段提取和波段选择。波段提取是利用线性或非线性的方式对原始高维波段进行变换，从而实现数据降维，其“特征波段”中的信息是原始所有数据信息的综合，而波段选择，是在原始的高维波段中，通过某些准则或方式在原始波段中选择出若干个波段，强调的是在原始波段中通过什么方式进行选择。

虽然他们都是降低数据维度的技术，但波段提取与波段选择相比，波段选择具有两个个优势：1) 从原始数据选择出的波段子集，没有进行相关的其他处理，依旧保持波段代表的物理意义[4]；2) 一般而言，不同的物质在光谱上会有不同的表现，但也可能由于光谱分辨率及光谱范围的限制表现为异物同谱。在数百个波段中，往往只有少数波段在地物彼此区分中起到了关键作用。通过波段选择，可以找到这些特定波段，从而提高对物体光谱性质的认识。

在 HIS 中，相邻的波段具有高度的相关性，但可能不携带有用的鉴别信息，导致所谓的 Hughes 现象(即“维度灾难”)和处理中的高计算成本，为了避免这些问题，波段选择取得了很好的效果，它可以去除冗余的波段。根据其是否使用除影像本身外的其他先验性信息，分为非监督波段选择和监督波段选择[5]。与监督波段选择相比，非监督波段选择技术不要求特定的应用，拥有更加灵活的使用场景。对于非监督波段选择技术，根据是否考虑了相关性可以将其分为两类：1) 不考相关性的方法常通过某种单一波段的指标来实现，例如信息熵、信噪比和信息散度等。2) 最佳指数因子(Optimal Index Factor, OIF)，计算波段的标准差和相关系数来衡量频带的重要性，并从中选择最优组合方案，这是一种考虑了多波段的相关性的方法。

OIF 在多光谱被广泛使用，2016 年赵庆展等人[6]，基于无人机多光谱遥感数据结合 OIF 方法、植被和水体指数等特征，提出了一种综合空－谱信息的最佳波段组合选择方法；2019 年郭力娜等人[7]，研究了基于 Landsat 8 OLI 影像的城市土地利用 OIF 选择方法，并验证了其有效性；2022 年王芳等人[8]，基于“高分二号”和“北京二号”卫星影像数据，考虑 OIF 来选择典型地物信息提取的最佳波段组合。这

些学者采用 OIF 的场景有个典型的特点，即都是多光谱数据，原因在于 OIF 需要计算所有波段的方差和多个波段间的相关性系数，这对于高达数百波段的高光谱遥感数据来说几乎是不可能完成的。因此，本文将提出一种分组式 OIF (Grouping OIF, G-OIF)的方式来选取高光谱遥感数据的“特征波段”，从而克服传统 OIF 在高光谱数据无法使用的困难。

2. 实验数据

本文选取高光谱数据集 Indian Pines，这是一个高光谱图像分割数据集，包括了美国印第安纳州的一个单一景观上的高光谱波段，像素为 145 × 145。对于每个像素，数据集包含 220 个光谱反射波段，代表了电磁光谱中不同部分的波长范围 0.4~2.5 μm。Indian Pines 由机载可视红外成像光谱仪(AVIRIS)于 1992 年对美国印第安纳州一块印度松树进行成像，然后截取尺寸为 145 × 145 的大小进行标注作为高光谱图像分类测试用途。Indian Pines 的细节如下表 1：

Table 1. Details of Indian Pines
表 1. Indian Pines 的细节

索引	类名	像素数
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-nottill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Indian Pines 数据集包括了 16 个有具体意义的类别，除此之外的所有像素类别均为索引为 0 的背景。Indian Pines 的彩色图像和真实地面(以灰度形式显示)如下图 1：

(a) 波段 29，19，9 的合成图像

(b) 真实地面

Figure 1. Visualization of the Indian Pines dataset
图 1. Indian Pines 数据集的可视化

3. 最佳指数因子的波段选取

3.1. 最佳指数因子的计算方式

OIF 方法主要考虑了三个方面：1) 选取的特征波段信息量尽可能大；2) 特征波段间的相关性要小，因为相关性越大代表蕴含不同地物间的信息差异越小，这不利于像素级的分类；3) 特征波段对不同地物的光谱差异越大越好。本文选取的 OIF 计算方式如下所示：

$$\mathrm{OIF}=\frac{\sum_{i=1}^{M}S_i}{\sum_{i=1}^{M}\left|R_{ij}\right|}, \tag{1}$$

其中 S_i 代表第 i 波段的标准差，R_{ij} 表示第 i 波段和第 j 波段之间的相关系数，M 表示的波段数，且 M 需要预先设置。OIF 的原理十分简单，通过计算考虑所有波段的标准差和多波段的相关系数，标准差越大和相关性越小，则选取的波段信息量越大。对于 S_i 来说，计算出所有波段的标准差是不难的。但是在面对数百的波段的高光谱数据，在计算 R_{ij} 时，设置的 M 越大波段间的组合越多，会对计算机内存造成极大压力，甚至出现内存不足的情况。

3.2. 分组式 OIF (G-OIF)的计算方式

为了缓解面对高光谱数据时计算 OIF 造成的内存压力，本文受分段信息熵[9]的启发，采用分组式的 OIF 来选择特征波段。如果将 Indian Pines 数据看作是一个向量空间，则将每一个像元可以看作一个 220 维的向量，整个数据集的维度为(145, 145, 220)。Indian Pines 数据集记作 $\mathbb{R}^{H\times W\times K}$，其中 $H\times W$ 代表影像尺寸，K 等于波段数。

以 K 这个维度为基准，先确定将 $\mathbb{R}^{H\times W\times K}$ 分为 n 组，记为 $\left\{\mathbb{R}^{H\times W\times K_k}\middle|\mathbb{R}^{H\times W\times K_1},\mathbb{R}^{H\times W\times K_2},\cdots,\mathbb{R}^{H\times W\times K_n},k=1,2,\cdots,n\right\}$。每组都可以看作是原数据集的一个子集，且每个子集间是互斥的，即：

$$\mathbb{R}^{H\times W\times K_1}\oplus\mathbb{R}^{H\times W\times K_2}\cdots\oplus\mathbb{R}^{H\times W\times K_n}=\mathbb{R}^{H\times W\times K}, \tag{2}$$

$\oplus$ 意思是直和。

在每个子集中根据公式(1)确定每个子集的最佳波段组合(Optimum Band Combinations, OBC)，记为 $\left\{OBC_k \middle| OBC_1, OBC_2, \cdots, OBC_n, k=1,2,\cdots,n\right\}$。最后确定的整个数据集的最佳波段组合为：

$$OBC = OBC_1 \bigcup OBC_2 \bigcup \cdots \bigcup OBC_n . \tag{3}$$

$\bigcup$表示并集，完整的计算过程如下：

算法 G-OIF 计算流程

1. 初始化超参数 M，n；
2. 初始化 $\left\{\mathbb{R}^{H\times W\times K_k} \middle| \mathbb{R}^{H\times W\times K_1}, \mathbb{R}^{H\times W\times K_2}, \cdots, \mathbb{R}^{H\times W\times K_n}, k=1,2,\cdots,n\right\}$；
3. 计算所有波段的标准差 S_i；
4. 在子集中计算所有可能的波段组合： $C_{K/n}^{M}$；
5. 根据波段组合计算 R_{ij}；
6. 在每个子集中计算所有波段组合的 OIF；
7. argmax{OIF}，得到 $\left\{OBC_k \middle| OBC_1, OBC_2, \cdots, OBC_n, k=1,2,\cdots,n\right\}$；
8. $OBC = OBC_1 \bigcup OBC_2 \bigcup \cdots \bigcup OBC_n$。

4. 实验与结果评价

4.1. 超参数初始化

G-OIF 要求首先对 $\mathbb{R}^{H\times W\times K}$ 做分组处理，其中要求两个超参数 M 和 n 合理。在本文中 n 要求为 $n|K$ ("|"是整除符号)，同时要求 $M \geq 3$，则在一个子集 OBC_k 中，一共有 $C_{220/n}^{M}$ 种组合。除此之外，若 M，n 设置不合理，计算组合时可能仍然会耗费大量时间和内存资源，在实验时不考虑这些不合理的情况。

4.2. 实验方法

随机森林(Random Forest, RF)和支持向量机(Support Vector Machine, SVM)是两种常用的用于高光谱图像分类的机器学习算法。RF 是一种集成学习方法，它构建多个决策树并输出类，即单个树的类的模式。它在高光谱图像分类中取得了巨大的成功。另一方面，SVM 是一种监督学习模型，它在高维空间中构建一个超平面或一组超平面，可用于分类或回归。SVM 已广泛应用于高光谱图像分类。

4.3. 结果分析比较

本文根据不同的超参数 M 和 n，利用 G-OIF 确定最佳组合波段后，并通过 RF 和 SVM 进行分类，将不同的波段组合进行比较。n 设置为 4、5、10、22、44 和 55，过滤过于耗费时间的 M，对应的 M 分别为 3-6、3-6、3-9、3-9、3-4 和 3，并且根据不同 M 时的分类结果进行分析。采取的评价指标为 Kappa 系数、平均精确度(Average Accuracy, AA)和总体精确度(Overall Accuracy, OA)，RF 和 SVM 的分类结果具体如下表 2 和表 3：

Table 2. Comparison of RF classification result indicators
表 2. RF 的分类结果指标对比

M	*n*	Kappa	AA	OA
3	4	67.3247	77.6123	71.8021
4		68.4170	77.5847	72.6803
5		68.1271	78.2510	72.3876
6		68.7185	80.8450	73.0120
3	5	68.7115	77.6395	72.9144
4		70.2122	80.9015	74.2316
5		70.1429	80.2262	74.2121
6		69.9163	82.8912	74.0560
3	10	71.2336	81.0328	75.1683
4		71.7284	81.3808	75.5098
5		71.9521	82.1119	75.6854
6		71.3556	80.3441	75.1390
7		71.9858	81.3446	75.7245
8		73.4633	82.8261	77.0221
9		73.0475	83.4640	76.6904
3	22	73.4636	81.1006	76.9929
4		73.4256	82.3830	77.0221
5		73.9945	83.2901	77.4612
6		74.1301	83.0920	77.6076
7		74.3419	84.0569	77.7442
8		74.5525	84.7625	77.9491
9		74.9015	84.9730	78.1027
3	44	74.6288	82.2749	77.9783
4		74.7381	83.8796	78.1247
3	55	74.4550	82.5032	77.7832

Table 3. Comparison of classification result indicators of SVM
表 3. SVM 的分类结果指标对比

M	*n*	Kappa	AA	OA
3	4	49.7161	45.5965	55.7811
4		50.4999	50.9068	56.7275
5		51.8914	49.4437	57.2251
6		51.4064	51.1167	57.6056
3	5	51.8307	49.4995	57.9374
4		53.0955	52.0180	58.9814
5		52.3624	53.8306	58.3472
6		55.2043	55.7733	60.7571
3	10	58.2282	55.6840	63.4111
4		58.9386	58.6005	64.1721
5		61.1730	61.7345	66.1333
6		61.5150	61.7708	66.4260
7		62.0356	63.8225	66.8260
8		62.6231	63.7218	67.4115
9		63.2430	64.2105	67.8798
3	22	65.3204	65.0986	69.6263
4		66.2821	67.3347	70.5630
5		68.3932	69.6027	72.3583
6		68.7811	70.0629	72.7583
7		69.3755	70.6976	73.2657
8		69.9564	72.2032	73.7243
9		70.7780	72.8834	74.5536
3	44	68.2118	68.0083	72.2314
4		69.1560	72.7610	73.1486
3	55	71.2660	73.1525	74.8463

首先讨论 n 对 G-OIF 的影响，为了容易观察，计算出每个子集 $\mathbb{R}^{H\times W\times K_k}$ 的各个指标的均值，然后可视化，如图 2。能够明显的看出随着 n 的增大，RF 和 SVM 的 Kappa、AA 和 OA 均呈现上升趋势。这说明随着 M，n 的变化，特征波段的数量也在变化，但其蕴含的信息量会越来越大，说明 G-OIF 的方法能够在实现减少计算成本的同时尽最大可能选出那些对分类有益的波段。

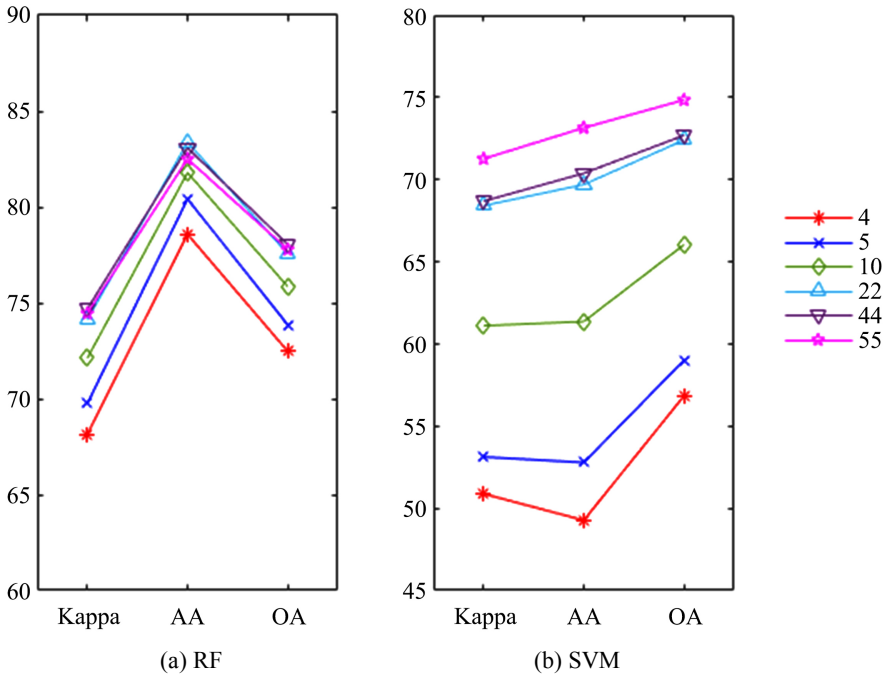

Figure 2. Mean values of various indicators of RF and SVM when n = 4, 5, 10, 22, 44, 45
图 2. n = 4, 5, 10, 22, 44, 45 时 RF 和 SVM 的各指标均值

接下来讨论当 n 固定时，即每个子集 $\mathbb{R}^{H\times W\times K_k}$ 中，M 变化时对 RF 和 SVM 产生的影响。从表 2 和表 3 能够很明显的看到，在每个子集中，选定的特征波段数量越多，其 Kappa、AA 和 OA 越大，说明分类效果与特征波段数量程正相关。

最后一个很明显的现象是：分组越多，同时设定的最佳波段组合的波段数越多，分类效果是越好的。原因是高光谱数据集的相邻波段相关性是很高的，分组越多时且选定的波段越多时更容易将那些蕴含更多差异性的波段选择出来。

5. 结论

原始的 OIF 计算方法在面对高光谱影像时，计算过多波段的组合会造成巨大的时间和内存压力，这甚至是无法完成的。本文受分段式信息熵的启发提出的 G-OIF 能够缓解这个问题，G-OIF 通过对高光谱数据进行分组，在每个子集中获得该子集的最佳组合波段，最后并集获得完整的最佳组合波段。G-OIF 在保证精度的同时实现了降维，并缓解了“维度灾难”。在未来，自适应的分组和波段数是一个具有潜力的研究方向。

参考文献

[1] Fauvel, M., Tarabalka, Y., Benediktsson, J.A., *et al.* (2012) Advances in Spectral-Spatial Classification of Hyperspectral Images. *Proceedings of the IEEE*, **101**, 652-75. https://doi.org/10.1109/JPROC.2012.2197589

[2] Landgrebe, D. (2002) Hyperspectral Image Data Analysis. *IEEE Signal Processing Magazine*, **19**, 17-28. https://doi.org/10.1109/79.974718

[3] Li, S., Song, W., Fang, L., *et al*. (2019) Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Transactions on Geoscience*, **57**, 6690-6709. https://doi.org/10.1109/TGRS.2019.2907932

[4] Varshney, P.K. and Arora, M.K. (2004) Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data. Springer, Berlin. https://doi.org/10.1007/978-3-662-05605-9

[5] Zhang, T., Li, P., Ding, Y., *et al*. (2022) Band Selection of Hyperspectral Images Based on Markov Clustering and Spectral Difference Measurement for Object Extraction. *ISPRS-International Archives of the Photogrammetry*, *Remote Sensing Spatial Information Sciences*, **43**, 449-455. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-449-2022

[6] 赵庆展, 刘伟, 尹小君, 张天毅. 基于无人机多光谱影像特征的最佳波段组合研究[J]. 农业机械学报, 2016, 47(3): 242-248+291.

[7] 郭力娜, 李帅, 张梦华, 牛振国, 李孟倩. 最佳波段组合的城市土地利用类型提取[J]. 测绘科学, 2019, 44(8): 161-167.

[8] 王芳. 最佳波段组合的典型地物信息提取[J]. 航天返回与遥感, 2022, 43(2): 82-91.

[9] 郑肇葆, 郑宏. 基于图像信息熵的高光谱图像分类[J]. 测绘地理信息, 2019, 44(5): 8-10.