文章编号:1004-9037(2014)06-0981-05

图 Laplacian 和自训练用于高光谱数据半监督波段选择

黄 睿 吕智强

(上海大学通信与信息工程学院,上海,200072)

摘要:波段选择是数据降维的有效手段,但有限的标记样本影响了监督波段选择的性能。提出一种利用图 Laplacian 和自训练策略实现半监督波段选择的方法。该方法首先定义基于图的半监督特征评分准则以产生初始波段子集,接着在该子集基础上进行分类,采用自训练策略将部分可信度较高的非标记样本扩展至标记样本集合,再用特征评分准则对波段子集进行更新。重复该过程,获得最终波段子集。高光谱波段选择与分类实验比较了多种非监督、监督和半监督方法,实验结果表明所提算法能选择出更好的波段子集。

关键词:高光谱数据分类;波段选择;半监督学习;图 Laplacian;自训练

中图分类号:TP751.1 **文献标志码**:A

Semi-supervised Band Selection Based on Graph Laplacian and Self-training for Hyperspectral Data Classification

Huang Rui, *Lü Zhiqiang*

(School of Communication and Information Engineering, Shanghai University, Shanghai, 200072, China)

Abstract: Band selection is an effective method for dimensionality reduction. However, the information from the small size of labeled samples usually misleads the supervised band selection. A semi-supervised band selection method based on graph Laplacian and self-training idea is proposed. The method first puts forward the graph-based semi-supervised criterion for feature ranking to generate the initial band subset. The graph Laplacian used in the criterion is refined with aid of the label information. Then, the supervised classifications are carried out based on the band subset and some unlabeled samples with higher confidence values are added into the labeled sample set. Afterwards the band subset is updated according to the feature ranking based on the newly generated labeled and unlabeled data, and is used for classification. The process repeats to obtain the final subset. The experiments on hyperspectral data sets are carried out compare several unsupervised, supervised and semi-supervised band selection methods. Results show that the proposed method can produce the band subset with better performance.

Key words: hyperspectral data classification; band selection; semi-supervised learning; graph laplacian; self-training

引 言

高光谱遥感可提供观测对象从可见光到近红光谱范围内细致的光谱曲线,有利于地物的分类识别。但另一方面,数据维数的提高也给地物分类带来困扰。Hughes 现象[1]指出,在标记样本有限时数据维数的不断增长会导致监督分类器性能下降。由于标记样本的采集受客观条件限制,数量非常有限,因此先验知识的匮乏极大地制约了高维数据监督分类算法的性能。通过波段(特征)选择进行数据降维是缓和 Hughes 现象的有效手段[2]。然而,在主流的监督式波段选择方法中,标记样本有限同样会影响子集评价准则的有效性。发掘隐藏在大

基金项目:国家自然科学基金(61001162)资助项目。

收稿日期:2013-01-08;**修订日期**:2013-03-07

量非标记样本中的类别信息，并与有限的先验知识相结合，实现半监督的波段选择和分类，已成为领域内重要研究课题之一。

近年来，基于谱图理论的半监督特征选择日益引起关注。Zhao 等[3]较早提出半监督特征选择概念，通过正则化最小割聚类获得特征重要度权值，形成了适用于两类分类问题的半监督特征选择算法。局部敏感判别特征方法（Locality sensitive discriminant feature，LSDF）[4]受监督特征选择 Fisher score 和非监督特征选择 Laplacian score[5]启发，引入了类内图和类间图定义，对每个特征计算两种图的比值以实现特征排序。文献[6]在 LSDF 基础上通过广义特征值求解确定投影变换矩阵，并对变换矩阵进行系数分析，实现特征加权算法（Graph-based semi-supervised feature weighting，GSFW）。Zhong 等[7]提出了一种基于标签传递的 wrapper 式迭代特征选择方法，取得了较好的实验结果，但因特征选择过程是学习算法相关的，计算量较大。Chen 等[8]提出基于图的局部加权判别投影方法，实现高分辨率遥感影像波段选择，并将选择方法扩展到核空间。

本文提出了基于图 laplacian 和自训练策略的高光谱数据半监督波段选择方法（Graph-based self-training feature selection，GST_FS）。该方法首先定义适用于迭代的半监督特征评分准则以产生初始波段子集，对 LSDF 的特征评价准则进行改进，以适应自训练策略的迭代过程。在产生的初始子集基础上进行分类，采用自训练方法将部分可信度较高的非标记样本扩展至标记样本集合，再用特征评分准则对波段子集进行更新。重复该过程，获得最终波段子集。

1 算法描述

1.1 初始波段子集的生成

设数据集 $\boldsymbol{X}$ 可分为两个子集：标记样本集 $\boldsymbol{X}_L=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l]$ 及其对应标记 $\boldsymbol{Y}_L=[y_1,y_2,\cdots,y_l]$，未标记样本集 $\boldsymbol{X}_U=[\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}]$，其中 $\boldsymbol{x}_i\in\mathbf{R}^D$。令 $\boldsymbol{X}$ 的特征集合 $\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_D]^{\mathrm{T}}$ 记录样本的各特征值，其中 $\boldsymbol{f}_d\in\mathbf{R}^{l+u}(1\leqslant d\leqslant D)$。半监督波段选择就是要基于 $\boldsymbol{X}_L$ 和 $\boldsymbol{X}_U$ 选择出与目标任务最相关的波段子集。基于数据 $\boldsymbol{X}$ 可构造无向图 $G(\boldsymbol{V},\boldsymbol{E})$，其中 $\boldsymbol{V}$ 为图的节点集合，与样本对应，$\boldsymbol{E}$ 为图的边集合，表征节点间的连接关系。

文献[4]指出，好的特征（波段）应能保持数据的局部结构并具有强的类别辨识能力。构造类间图和类内图，其 Laplacian 表示分别为 $\boldsymbol{L}_b$ 和 $\boldsymbol{L}_w$，有

$$\boldsymbol{L}_w=\boldsymbol{D}_w-\boldsymbol{S}_w,\boldsymbol{D}_w=\mathrm{diag}(\boldsymbol{S}_w\boldsymbol{e}) \tag{1}$$

$$\boldsymbol{L}_b=\boldsymbol{D}_b-\boldsymbol{S}_b,\boldsymbol{D}_b=\mathrm{diag}(\boldsymbol{S}_b\boldsymbol{e}) \tag{2}$$

$$\boldsymbol{e}=(1,1,\cdots,1)^{\mathrm{T}} \tag{3}$$

其中，两个连接矩阵分别定义为

$$\boldsymbol{S}_{w,ij}=\begin{cases}rp(i)p(j) & y_i=y_j\\ 1 & (\boldsymbol{x}_i\in\boldsymbol{X}_U\text{且}\boldsymbol{x}_i\in\mathrm{KNN}(\boldsymbol{x}_j))\\ & \text{或}(\boldsymbol{x}_j\in\boldsymbol{X}_U\text{且}\boldsymbol{x}_j\in\mathrm{KNN}(\boldsymbol{x}_i))\\ 0 & \text{其他}\end{cases} \tag{4}$$

$$\boldsymbol{S}_{b,ij}=\begin{cases}1 & y_i\neq y_j\\ 0 & \text{其他}\end{cases} \tag{5}$$

式中：$\boldsymbol{x}_i\in\mathrm{KNN}(\boldsymbol{x}_j)$ 表示样本 $\boldsymbol{x}_i$ 是 $\boldsymbol{x}_j$ 的 K 最近邻；r 为常量，本文取为 100。$p(i)$ 为样本 $\boldsymbol{x}_i$ 具有标记 y_i 的可信度（当样本被学习算法赋予某类别标签时，样本属于该类别的概率）。当 $\boldsymbol{x}_i\in\boldsymbol{X}_L$ 时有 $p(i)=1$；当 $\boldsymbol{x}_i\in\boldsymbol{X}_U$ 时 $p(i)$ 由学习算法对样本归属类别概率的判定获得。

在此基础上，定义波段 d 的重要度评分准则为（$1\leqslant d\leqslant D$）

$$J(\boldsymbol{f}_d)=\frac{\boldsymbol{f}_d^{\mathrm{T}}\boldsymbol{L}_b\boldsymbol{f}_d}{\boldsymbol{f}_d^{\mathrm{T}}\boldsymbol{L}_w\boldsymbol{f}_d} \tag{6}$$

权值 $J(\boldsymbol{f}_d)$ 越大，表明波段 d 越重要。考虑到高光谱数据谱间具有高相关性，可采用去冗余算法选择前 D' 个波段构成子集[6]。

1.2 算法具体实现

算法首先利用半监督特征评分准则和谱间去冗余函数 FeaSelect 获得初始波段子集，子集规模为 D'。接着通过监督学习算法 Learn 对非标记样本集 $\boldsymbol{X}_U$ 分类，从未标记样本可信度集合 $\boldsymbol{P}=[p(1),p(2),\cdots,p(u)]$ 中选择至少 $s\%$ 可信度较高的样本加入到标记样本集 $\boldsymbol{X}_L$ 中，更新标记与非标记样本，并利用 FeaSelect 生成新的波段子集。该过程迭代 I_{ter} 次，产生最终子集 $\boldsymbol{F}'$。函数 ceil 表示向下取整。具体流程如表 1 所示。其中，学习算法 Learn 可采用极大似然分类器、K 最近邻分类器、支持向量机等实现。

由分类误差和训练样本数的关系[9-10]可知，加入 $\boldsymbol{X}_L$ 的未标记样本数量必须在第 t 次和 $t+1$ 次迭代过程中满足

$$|\boldsymbol{X}_L\cup\boldsymbol{X}_E^t|\left(1-2\frac{e_L|\boldsymbol{X}_L|+e^t|\boldsymbol{X}_E^t|}{|\boldsymbol{X}_L\cup\boldsymbol{X}_E^t|}\right)>$$

$$|\boldsymbol{X}_L \cup \boldsymbol{X}_E^{t+1}|\left(1-2\frac{e_L|\boldsymbol{X}_L|+e^{t+1}|\boldsymbol{X}_E^{t+1}|}{|\boldsymbol{X}_L \cup \boldsymbol{X}_E^{t+1}|}\right) \quad (7)$$

式中：$|\cdot|$表示集合规模，$\boldsymbol{X}_E^t$ 为第 t 次迭代中加入 $\boldsymbol{X}_L$ 的未标记样本集合，e_L 为原始标记样本分类误差，e^t 和 e^{t+1} 分别为未标记样本在两次迭代中的分类误差。由于该分类误差无法准确获得，采用如下方式近似

$$e^t = 1-\frac{1}{u}\sum_{i=1}^{u} p^t(i) \quad (8)$$

式(8)进一步可写为

$$\frac{|\boldsymbol{X}_E^{t+1}|}{|\boldsymbol{X}_E^t|} > \frac{1-2e^t}{1-2e^{t+1}} \quad (9)$$

当式(9)成立时，表明未标记样本的加入可提高学习算法性能。算法实现伪代码如下。

输入：$\boldsymbol{X}_L$，$\boldsymbol{X}_U$，D'，$s\%$，I_{ter}
输出：$\boldsymbol{F}'$
过程：
$\boldsymbol{F}' = \text{FeaSelect}(\boldsymbol{X}_L, \boldsymbol{X}_U, D')$
$t=1$；$u_0=0$；$e^0=0.5$
while $t \leqslant I_{\text{ter}}$ do
　$(\boldsymbol{Y}_U, \boldsymbol{P}^t) = \text{Learn}(\boldsymbol{X}_L, \boldsymbol{X}_U, \boldsymbol{F}')$ // 对 $\boldsymbol{X}_U$ 进行识别，返回类别标记和识别可信度
　IND=Sort($\boldsymbol{P}$,'decend') // 按可信度降序排列，返回位置索引 IND
　$\boldsymbol{X}'_U = \text{Rearrange}(\boldsymbol{X}_U, \text{IND})$ // 将 $\boldsymbol{X}_U$ 按 IND 重排
　$e^t = \text{CalError}(\boldsymbol{P}^t)$ // 按式(8)计算
　$u' = \max\left(\text{ceil}\left(\frac{1-2e^0}{1-2e^t}u_0\right), \text{ceil}\left(\frac{su}{100}\right)\right)$ // 保证式(9)成立
　$\boldsymbol{X}_E^t = [\boldsymbol{x}'_{U1}, \boldsymbol{x}'_{U2}, \cdots, \boldsymbol{x}'_{Uu'}]$
　$\boldsymbol{X}'_L = \boldsymbol{X}_L \cup \boldsymbol{X}_E^t$
　$\boldsymbol{X}'_U = \boldsymbol{X}_U - \boldsymbol{X}_E^t$
　$e^0 = e^t$；$u_0 = u'$；$t=t+1$
　$\boldsymbol{F}' = \text{FeaSelect}(\boldsymbol{X}'_L, \boldsymbol{X}'_U, D')$
end

2　高光谱数据波段选择实验

通过两个高光谱数据集的波段选择和分类实验，对所提半监督波段选择方法 GST_FS 与 5 种非监督、监督和半监督方法(Laplacian score、Fisher score、最小化误分典型相关性分析(Minimum misclassification canonical analysis，MMCA)[11]、LSDF 和 GSFW)进行性能比较。

实验数据 1 为 AVIRIS 在美国 Indian Pine 北部一块农业区获取的 220 波段高光谱影像部分，大小为 145 像素×145 像素；实验数据 2 为 DAIS 7915 在意大利城市 Pavia 的市中心获取的影像部分，大小为 400 像素×400 像素。考虑到水体和噪声的影响，分别选择 200 和 61 个波段进行实验。图 1，2 和表 1 分别为两个数据集的影像和实验数据描述。

图 1　AVIRIS 数据影像
Fig. 1　AVIRIS image

图 2　DAIS 数据影像
Fig. 2　DAIS image

表 1　实验数据描述
Table 1　Data description

类别	AVIRIS 数据	类别	DAIS 数据
Corn-min	834	Water	4 281
Corn	234	Tree	2 424
Grass/Pasture	497	Asphalt	1 251
Grass/Tree	747	Parking lot	1 475
Hay-windrowed	489	Bitumen	1 704
Soybeans-notill	968	Brick roof	287
Soybeans-min	2 468	Meadow	2 237
Soybeans-clean	614	Soil	685
Bldg-Grass-Tree-Drive	380	Shadow	241
总计	7 231		14 585

实验考察标记样本有限情况下的波段选择算法性能。每类标记样本数为各类样本总数的 10%，随机生成。出于计算量考虑，对于 AVIRIS 数据，随机生成的训练样本数占样本总量百分比为 3%；对于 DAIS 数据，随机生成的训练样本数占样

本总量百分比为 1%。两种数据中标记与未标记样本数之比为 1∶4。基于图的 Laplacian score，LSDF，GSFW 和 GST_FS 均以此标记样本和未标记样本构建 5NN 图。最终的实验结果为 5 次运行的平均值。学习算法为 KNN 分类器($K=5$)，未标记样本 $\boldsymbol{x}_i$ 的可信度可按式(10)计算

$$p(i)=\frac{\max_{c=1}^{C} k_c}{K} \tag{10}$$

式中：C 为数据类别数，k_c 表示样本 x_i 的 K 个标记样本最近邻中属于类别 c 的样本个数。

3 实验结果分析

3.1 不同波段选择方法的性能比较

图 3 给出了 6 种方法对 AVIRIS 数据选择 20 个波段，DAIS 数据选择 5 个波段时的测试样本分类精度比较。其中，LS 和 FS 分别为 Laplacian score 和 Fisher score 的缩写。可以看到，非监督的 Laplacian score 完全根据数据的局部结构信息选择波段，性能较差。MMCA 和 Fisher score 利用了标记样本类别信息，选择的波段子集性能总体好于 Laplacian score，但因受到标记样本不足的影响，性能起伏较大(Fisher score 在 AVIRIS 数据中的表现尤为明显)。3 种半监督波段选择方法由于同时利用了标记样本的类别信息和大量非标记样本中的局部结构信息，选择出的波段子集性能较好。其中，GSFW 综合考察了波段子集的整体性能，因此优于 LSDF；GST_FS 将 LSDF 的半监督特征排序算法与自训练策略结合起来，取得了最好的性能。

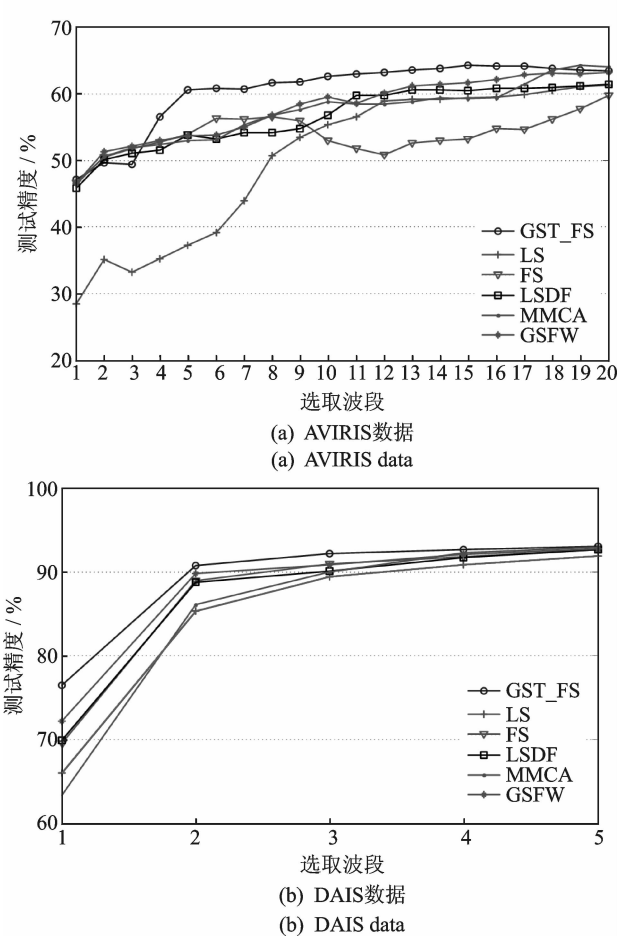

图 3 6 种波段选择方法的性能比较

Fig. 3 Performance comparisons among six band selection methods

3.2 标记样本数对波段子集的影响

实验考察了标记样本数变化对监督、半监督波段选择方法 Fisher score、MMCA、LSDF、GSFW 和 GST_FS 的影响。表 2 为标记样本数与未标记样本之比为 1∶4，2∶4 和 3∶4 变化时，5 种方法产生波段子集的平均分类精度比较。AVIRIS 数据的平均分类精度指选择 1～20 波段时的测试精度平均值；DAIS 数据的平均分类精度指选择 1～5 波段时的测试精度平均值。可以看出，随着标记样本的增多，测试精度有所提高。这说明标记样本对于提升波段子集质量有重要的指导作用，符合对标记样本的认知。

表 2 标记样本数变化对波段选择方法的影响

Table 2 Classification accuracies of band selection methods when number of labeled samples changes

类别		标记样本数		
		10%	20%	30%
AVIRIS 数据	GST_FS	60.37	62.13	62.11
	FS	53.86	55.32	55.56
	LSDF	56.56	57.50	58.54
	GSFW	57.84	59.26	60.39
	MMCA	57.12	61.07	60.90
DAIS 数据	GST_FS	89.01	88.96	90.18
	FS	86.76	87.64	87.84
	LSDF	86.62	87.66	88.36
	GSFW	87.51	88.18	89.12
	MMCA	84.93	87.04	87.60

3.3 未标记样本数对波段子集的影响

实验考察了未标记样本数目变化对基于图的波段选择方法 GST_FS，Laplacian score，GSFW 和 LSDF 的影响。表 3 为标记样本数与未标记样本之比为 1∶2，1∶3，1∶4 和 1∶5 变化时，4 种方法产生的波段子集的平均分类精度比较。可以看

出，随着未标记样本数的增加，Laplacian score、GSFW 和 LSDF 的精度变化较小；GST_FS 起伏稍大，但与未标记样本数并非正相关。这是由于未标记样本的引入在提供隐含类别信息的同时，也带来了不确定性。另一方面，过多的未标记样本不但无助于提高方法性能，也使得计算量难以负担。如何确定有效地未标记样本评价准则，选择适当数量的有代表性的未标记样本，值得进一步研究。

表 3　未标记样本数变化对波段选择方法的影响

Table 3　Classification accuracies of band selection methods when number of unlabeled samples changes

类别		未标记样本数			
		20%	30%	40%	50%
AVIRIS 数据	GST_FS	58.26	57.59	60.32	59.41
	LS	50.46	50.75	50.69	50.75
	LSDF	56.08	56.35	56.57	56.35
	GSFW	57.15	57.23	57.30	57.48
DAIS 数据	GST_FS	87.98	88.88	88.96	88.28
	LS	84.66	84.64	84.58	84.74
	LSDF	86.74	86.68	86.56	86.68
	GSFW	87.12	87.26	87.19	87.34

4　结束语

本文基于 LSDF 算法，提出了基于图 laplacian 和自训练策略的高光谱数据半监督波段选择方法 GST_FS。该方法首先定义适用于迭代的半监督特征评分准则以产生初始波段子集，接着在该子集基础上进行监督分类，采用自训练策略将部分可信度较高的非标记样本扩展至标记样本集合，再用特征评分准则对波段子集进行更新。重复该过程，获得最终波段子集。基于有限标记样本的高光谱数据波段选择和分类实验表明，算法利用了标记样本的类别信息和大量非标记样本中的局部结构信息，同时自训练策略的引入进一步提高了波段子集的质量，性能优于多种非监督、监督和半监督波段选择方法。

致谢　感谢美国 Purdue 大学 David A. Landgrebe 教授提供 AVIRIS 高光谱数据；意大利 Pavia 大学 HySens 项目提供 DAIS 高光谱数据。

参考文献：

[1] Hughes G F. On the mean accuracy of statistical pattern recognizers [J]. IEEE Transactions on Information Theory, 1968, 14(1): 55-63.

[2] 黄睿，何文勇. 基于粒子群算法和序贯搜索的高光谱波段选择[J]. 数据采集与处理，2012，27(4)：469-473.
Huang Rui, He Wenyong. Hyperspectral band selection based on particle swarm optimization and sequential search [J]. Journal of Data Acquisition and Processing, 2012, 27(4): 469-473.

[3] Zhao H, Liu H. Semi-supervised feature selection via spectral analysis [C] // 2007 SIAM International Conference on Data Mining. Minneapolis, Minnesota: [s. n.], 2007: 641-646.

[4] Zhao J, Lu K, He X. Locality sensitive semi-supervised feature selection [J]. Neurocomputing, 2008, 71: 1842-1849.

[5] He X, Cai D, Niyogil P. Laplacian score for feature selection [J]. Advances in Neural Information Processing Systems, 2006, 18: 507-514.

[6] 黄睿，陈玲. 图 Laplacian 半监督特征加权用于高光谱波段选择[J]. 应用科学学报，2011，29(6)：626-630.
Huang Rui, Chen Ling. Semi-supervised feature weighting using graph laplacian for hyperspectral band selection [J]. Journal of Applied Sciences, 2011, 29(6): 626-630.

[7] Zhong Z, Xie S, Fan W, et al. Graph-based iterative hybrid feature selection [C] // 8th International Conference on Data Mining. Pisa, Italy: [s. n.], 2008: 1133-1138.

[8] Chen X, Fang T, Huo H, et al. Graph-based feature selection for object-oriented classification in VHR airborne imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(1): 353-365.

[9] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data [C] // 17th International Conference on Machine Learning. Stanford, USA: CA, 2000: 327-334.

[10] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Trans Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[11] Chang C I, Du Q, Sun T, et al. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 1999, 37(6): 2631-2641.

作者简介：黄睿(1976-)，女，博士，副教授，研究方向：模式识别、图像处理、遥感信息智能处理，E-mail：huangr@shu.edu.cn；吕智强(1988-)，男，硕士研究生，研究方向：遥感影像信息提取与分类。