基于改进标签传播算法的高光谱图像半监督分类

崔宾阁,吴子宾,秦学川,马秀丹
(山东科技大学 计算机科学与工程学院,山东 青岛 266590)

摘 要:针对标签传播算法缺乏对新生成样本的评价进而影响分类精度的问题,本文提出一种利用阈值的标签传播算法来提高高光谱图像的分类精度。首先,用基于图像融合和递归滤波的特征提取方法对原始高光谱图像进行处理。然后,给出一个阈值并对标签传播算法新生成样本进行评价,保留一些可信度较高的样本。最后,保留的新样本和已标记样本之和作为训练样本,对图像进行分类。实验表明,基于改进标签传播算法优于其他的高光谱图像分类算法。

关键词:标签传播;高光谱图像分类;阈值法;递归滤波

中图分类号:TP75 **文献标志码**:A **文章编号**:1672-3767(2016)06-0101-07

Semi-supervised Classification of Hyperspectral Images Based on Modified Label Propagation Algorithm

CUI Binge, WU Zibin, QIN Xuechuan, MA Xiudan
(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: To solve the lack of evaluation of the label propagation algorithm for new samples which further affects the classification accuracy, this paper proposed a new label propagation algorithm about the threshold to improve the classification accuracy of hyperspectral images. First of all, the original hyperspectral images were processed with the method of feature extraction based on image fusion and recursive filtering. Then a threshold was given and the new samples produced by label propagation algorithm were evaluated. Some samples with higher credibility were kept. Finally, with the newly-kept samples and tagged samples as the training samples, the images were classified. Experimental results show that the modified label propagation algorithm is better than other hyperspectral image classification algorithms.

Key words: label propagation; hyperspectral image; threshold value method; recursive filtering

高光谱分辨率图像可以由高光谱卫星传感器获得,例如机载可见/红外成像光谱仪(Airborne Visible/Infrared Imaging Spectrometer)。高光谱图像提供了对应地物物理材质的详细光谱信息,因此高光谱图像能够区分不同地貌特征。

目前常用的高光谱图像的分类方法有监督分类方法[1-4]、半监督分类方法[5-7]、非监督分类方法。其中,半监督算法因其能够在标记样本稀少[8]的情况下提高分类精度而得到了越来越多的关注。

标签传播算法[9](Label Propagation)是由 Zhu 等于 2002 年提出的一种基于图的半监督学习方法[10],

收稿日期:2016-06-09
基金项目:国家自然科学基金青年基金项目(41406200);山东省自然科学基金青年基金项目(ZR2014DQ030)
作者简介:崔宾阁(1979—),男,山东烟台人,副教授,硕士生导师,主要从事机器学习、数据挖掘、人工智能、模式识别、遥感图像处理、大数据和云计算研究.
吴子宾(1990—),男,山东聊城人,硕士研究生,主要从事高光谱遥感图像分类的研究,本文通信作者.
E-mail:568690239@qq.com

因其不受数据分布形状的影响、算法简单、执行时间短且分类性能好的优点引起了国内外学者的关注，并被大量应用到图像分类领域中。但标签传播算法仍存在一些问题，如在信噪比较差的图像中分类效果较低，分类精度易受样本所属标签类概率的影响等。

针对标签算法存在的问题，本文提出一种改进的标签传播算法(modified label propagation，MLP)，其主要思想是：用基于图像融合和递归滤波的高光谱图像特征提取方法[11]提高图像的质量，然后给出一个阈值对标签传播算法产生的新样本进行评价，当生成样本的可信度低于阈值时，认为是错误标记样本并去除该样本，最后将保留的可信度较高的新样本与已标记样本之和作为训练样本，对图像进行分类。实验结果表明，本文提出的 MLP 算法能够显著提高高光谱图像分类精度，且在训练样本稀少的情况下，分类效果更佳。

1 标签传播算法思想及其局限性

1.1 标签传播算法及局限性分析

标签传播算法是一种基于图的半监督学习方法，图中的节点表示已标记和未标记的数据，图中的边表示两个节点的相似度；节点的标签通过边传播到相邻节点，为图中的所有节点定义该节点所属类别的概率分布表，图中的所有节点都根据它相邻节点的概率分布来更新自己的概率分布。该算法在传播过程中迭代执行直到节点的概率分布收敛。然后选出未标记样本对应概率最大的类，作为样本的标记信息。

标签传播过程如图 1 所示，黑色和灰色节点是不同类别的已标记数据，白色节点是未标记数据。以箭头上的概率从已标记数据到未标记数据传播标签。

图 1 中标签传播算法通过近邻点之间的标记传播来对节点进行分类，新产生样本的标签由样本所属标签类的概率决定，当概率很小时，该数据有可能是噪声数据或是其它类，由此产生错误标记，进而对图像分类造成干扰。为排除错误标记的样本产生的干扰，可以添加阈值来对新产生的样本进行评价，提高新生成标记样本的可信度。

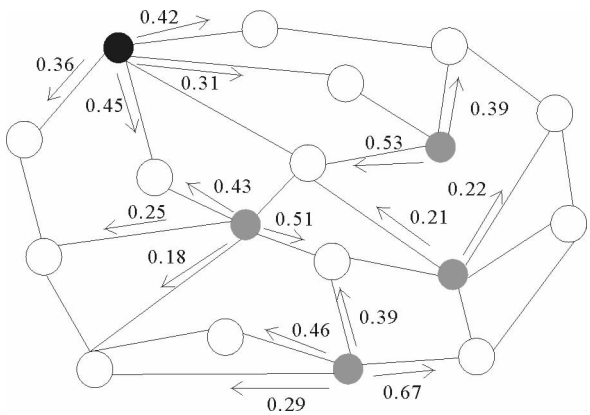

图 1 标签传播过程

Fig. 1 The procedure of label propagation

1.2 标签传播算法

算法 1：标签传播算法(LP)

输入：已标记样本及对应标签集合 $D_l=\{(x_1,y_1),\cdots,(x_l,y_l)\}\subset R^N$，未标记样本集合 $D_u=\{x_{l+1},\cdots,x_n\}\subset R^N$，标签集 $L=\{1,\cdots,c\}$

输出：未标记样本及对应标签集合 $D'_u=\{(x_{l+1},y_{l+1}),\cdots,(x_n,y_n)\}\subset R^N$

1) n 个样本集合$\{x_1,x_2,\cdots,x_l,x_{l+1},x_{l+2},\cdots,x_n\}$作为图中的所有节点，计算关联矩阵 $\boldsymbol{W}^{n\times n}$

$$w_{ij}=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right),\quad \forall i\neq j。\tag{1}$$

2) 根据已有的 W 计算传播概率矩阵 $\boldsymbol{S}^{n\times n}$

$$S_{ij}=\frac{W_{ij}}{\sum_{k=1}^{n}W_{ik}}。\tag{2}$$

其中 S_{ij} 表示从节点 i 到节点 j 的传播概率。

3) 初始化标记矩阵 $\boldsymbol{A}^{n\times c}$，初始化概率分布 P，

$$A_{ij}=\begin{cases}1, y_i=j, i\leqslant 1\\ 0, y_i\neq j, i\leqslant 1\\ 1/c, l<i\leqslant n\end{cases};\tag{3}$$

$$P_{ij}=A_{ij},1\leqslant i\leqslant n,1\leqslant j\leqslant c。\tag{4}$$

4）传播。每个节点根据传播概率 P 把它周围节点传播来的标记信息按权重相加，并更新自己的概率分布 F，F_{ij} 为第 i 个样本属于第 j 个类的概率

$$P_{ij}=\sum_{k=1}^{n}S_{ik}P_{kj},1\leqslant i\leqslant n,1\leqslant j\leqslant c。\tag{5}$$

5）限定已标记样本。把已标记样本的概率分布重新赋值为初始值

$$P_{ij}=A_{ij},1\leqslant i\leqslant l,1\leqslant j\leqslant c。\tag{6}$$

6）重复步骤 4，直到 P 收敛。

7）对未标记样本进行标记，

$$y_i=\arg\max_{j\leqslant c}P_{ij},1\leqslant i\leqslant n。\tag{7}$$

8）$D_l\bigcup D'_u$ 作为训练数据，用 SVM 进行分类。

2 改进的标签传播算法

2.1 改进的标签传播算法思想

改进的标签传播算法思想是：用基于图像融合和递归滤波的高光谱图像特征提取方法产生的图像作为标签算法的输入图像；再增设一个阈值来对标签传播算法产生的标记样本进行评价，如果产生的标记样本所属标签类的概率小于阈值，则将该样本从标记样本集中去掉。

2.2 改进的标签传播算法

算法 2：改进的标签传播算法（MLP）

输入：已标记样本及对应标签集合 $D_l=\{(x_1,y_1),\cdots,(x_l,y_l)\}\subset R^N$，标签传播算法输出的未标记样本及对应标签集合 $D'_u=\{(x_{l+1},y_{l+1}),\cdots,(x_n,y_n)\}\subset R^N$，阈值 t

输出：训练样本集 D，分类结果

1）对图像进行基于图像融合和递归滤波的特征提取操作

①图像融合：图像融合能有效地去除噪声的影响，将高光谱图像分割成相邻波段的 k 个子集，用均值法对每个子集中的波段进行融合。

$$Q^k=\frac{\sum_{n=1}^{N_k}P_n^k}{N_k}。\tag{8}$$

其中，P_n^k 表示第 k 个子集中的第 n 个波段，N_k 表示第 k 个子集中的波段数，Q^k 表示第 k 个子集融合后的图像。

②递归滤波：对融合后的图像进行递归滤波操作，递归滤波可以高效地利用图像的空间信息。

$$J[m]=(1-a^b)\cdot I[m]+a^b\cdot J[m-1]。\tag{9}$$

$J[m]$ 表示滤波结果图像，$a=\exp(-\frac{\sqrt{2}}{\sigma_s})\in[0,1]$ 表示反馈系数，$I[m]$ 表示输入的像素信息，b 表示像素之间的距离。

2）将传统标签传播算法产生的新增样本集 $D'_u=\{(x_{l+1},y_{l+1}),\cdots,(x_n,y_n)\}\subset R^N$ 作为候选集，初始化训练样本集 $D=D_l$。

3）对候选集中每个样本评价，当 t 小于等于 P_{ij} 时，将对应的第 i 个样本加入到训练样本集 D 中

$$D=D\bigcup\{(x_i,y_i)\},t\leqslant P_{ij},l<i\leqslant n。\tag{10}$$

4）训练样本集 D 作为训练样本用 SVM 对处理后的图像进行分类。

3 实验

3.1 实验数据

实验采用的是由机载可见/红外成像光谱仪传感器（AVIRIS）1992 年在印第安纳州西北部的 Indian

Pines 地区获得的数据。高光谱图像由 220 个光谱波段的 145×145 个像素组成，包括 16 个地物类型，由于噪声及吸水率的影响，去掉低信噪比、水汽吸收波段以及传感器故障波段的 20 个波段。

每个单独实验做 20 次，20 次实验取均值作为获得值；S 表示每个类的已标记样本数量；总体精度(overall accuracy，OA)可以用训练样本中分类正确的样本总数和训练样本总数的比值来表示；平均精度(average accuracy，AA)可以用每种地物分类正确的样本数和每种地物样本数比值的平均值来表示。

3.2 实验结果及分析

3.2.1 对于阈值的讨论

在实验中，用 Indian Pines 的 AVIRIS 数据集评估阈值对分类性能的影响。图 2 展示了不同阈值和不同样本数总体精度的变化曲线，t 和 S 分别取{0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.75，0.8，0.85，0.9，1}和{3，5，10}，可以看到在 $S=3$、$S=5$ 和 $S=10$ 时，OA 分别在 $t=0.85$、$t=0.8$ 和 $t=0.9$ 时达到最高，且在达到最高之前，OA 随着 t 的增加而增加，而在到达最高之后 OA 随着 t 的增加递减。

由此可以得出：

1) 随着 t 的增加，标记样本的可靠性越来越高，新增样本的数量越来越少。

2) $t=0$ 意味着增加的新样本都没有被删除，样本数最多，但样本的可靠性最低；$t=1$，相当于新样本全部被删除，样本数最少，样本的可靠性最高。

3) 寻找最优阈值，即找到产生的标记样本的可靠性与数量之间的最优线性组合。

3.2.2 图像特征提取效果的验证

在相同阈值条件下，讨论以下两种特征提取方法对图像分类精度的影响。每个类取 10 个样本。

方法 1 先用基于图像融合和递归滤波的特征提取方法对原始高光谱图像进行处理，然后用带有阈值的标签传播算法对处理后的图像进行分类。

方法 2 没有经过特征提取处理，直接用带有阈值的标签传播算法对原始图像进行分类。

实验结果由图 3 所示，可看到：

1) 由方法 1 和方法 2 得到的图像分类趋势基本吻合。

2) 由方法 1 得到的分类精度远远高于由方法 2 得到的图像分类精度。

实验结果说明先用特征提取方法提高图像质量，再对图像进行分类，可以显著提高图像分类精度。

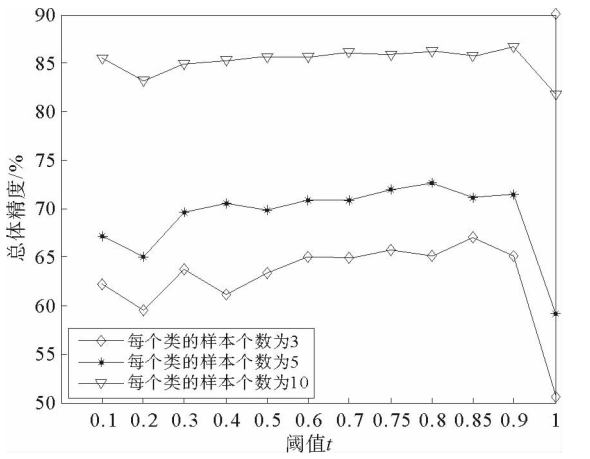

图 2 每个类的样本数不同时阈值 t 对于 Indian Pines 的 AVIRIS 数据分类性能的影响

Fig. 2 Influence of t on the performance for the AVIRIS data of Indian Pines

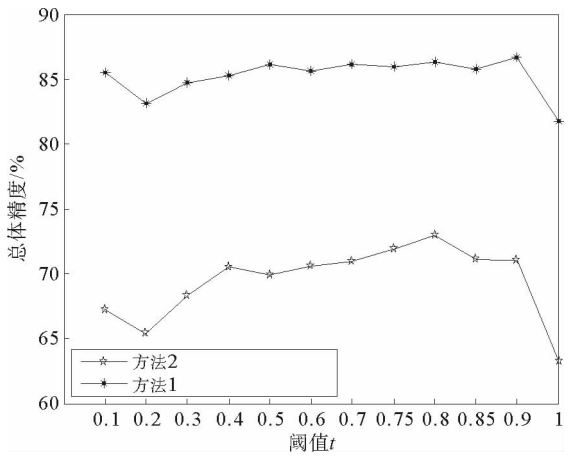

图 3 特征提取方法对于 Indian Pines 的 AVIRIS 数据分类性能的验证

Fig. 3 Influence of adding a feature extraction method on the performance for the AVIRIS data of Indian Pines

3.2.3 LP 与 MLP 的对比实验

将改进的标签传播算法用IndianPines的AVIRIS数据进行实验，并与标签传播算法进行比较。结果

如表 1 所示。

可看出:与原标签传播算法(LP)相比,改进的标签算法(MLP)在 $S=10$ 时,OA、AA 和 Kappa 均提高了 20%左右,分类效果提升明显。

3.2.4 几种常用图像分类算法的对比实验

将本文的 MLP 算法与四种常用的分类方法利用 Indian Pines 的 AVIRIS 数据进行比较,这些算法包括:Support Vector Machine(SVM)、Spectral-Spatial Hyperspectral Image Classification With Edge-Preserving Filtering(EPF)[1]、Feature Extraction of Hyperspectral Images With Image Fusion and Recursive Filtering(IFRF)[11]和 Intrinsic Image Decomposition for Feature Extraction of Hyperspectral Images(IID)[2]。SVM 算法采用高斯核函数,EPF 算法采用文献[1]中的默认参数,IFRF 算法采用 20 个特征值,IID 算法中将图像的波段分给成相邻的 4 个子集。

表 2 为 5 种图像分类算法对图像分类后的总体精度比较,表中加粗的为在 S 取不同的值时各种方法的最优值,可见在 S 取不同值时,本文的 MLP 均优于其他分类算法。

表 1 LP 与 MLP 对于 Indian Pines 的 AVIRIS 数据分类的对比实验

Tab. 1 classification experiments of LP and MLP for the AVIRIS data of Indian Pines

Algorithm	—	$S=3$	$S=5$	$S=10$	$S=15$
LP	OA	56.92	57.9	65.13	79.12
	AA	55.21	56.8	66.01	80.43
	Kappa	54.49	58.1	64.28	76.45
MLP	OA	65.14	71.61	86.72	93.45
	AA	66.14	71.02	85.72	92.25
	Kappa	65.38	70.19	86.65	93.13

表 2 5 种图像分类算法对于 Indian Pines 的 AVIRIS 数据分类的总体精度比较

Tab. 2 OA of the five image classified algorithms for the AVIRIS data of Indian Pines

Algorithm	$S=3$	$S=5$	$S=10$	$S=15$
SVM	20.21	35.1	56.38	63.49
EPF	23.33	39.42	66.96	79.06
IFRF	39.24	60.63	76.55	85.06
IID	47.48	69.4	83.46	89.31
MLP	**65.14**	**71.61**	**86.72**	**93.45**

表 3 4 种图像分类算法对于 Indian Pines 的 AVIRIS 数据的分类精度

Tab. 3 Classification accuracies of four image classified algorithms for the AVIRIS data of Indian Pines

	Train	Test	EPF	IFRF	IID	MLP
Alfalfa	15	46	79.800	80.69	**99.36**	83.28
Corn-no till	15	1 428	79.600	84.86	95.63	**96.36**
Corn-min till	15	830.	80.980	68.17	84.69	**95.27**
Corn	15	237	45.210	75.60	72.59	**76.96**
Grass/pasture	15	483	96.950	87.57	90.78	**98.22**
Grass/trees	15	730	90.410	91.79	97.94	**98.58**
Grass/pasture-mowed	15	28	**86.180**	54.66	69.15	75.48
Hay-windrowed	15	478	100.000	100.00	100.00	**100.00**
Oats	15	20	90.250	67.90	72.25	**90.44**
Soybeans-no till	15	972	66.660	79.41	80.68	**84.41**
Soybeans-min till	15	2 455	79.750	92.14	92.24	**94.94**
Soybeans-clean till	15	593	57.200	73.54	82.97	**92.92**
Wheat	15	205	**99.280**	96.12	98.34	98.47
Woods	15	1 265	96.460	98.33	98.01	**98.92**
Bldg-Grass-Tree-Drives	15	386	66.620	78.31	86.77	**99.50**
Stone-steel towers	15	93	94.560	97.72	**98.90**	92.25
OA	—	—	77.237	85.03	89.35	**93.45**
AA	—	—	81.870	82.93	88.77	**92.25**
Kappa	—	—	74.195	83.07	88.91	**93.13**

表3展示出五种算法对图像分类的平均精度(AA)。每个类取15个已标记样本。由表3可以得出,本文的MLP能够得到较高的AA,分别比EPF、IFRF和IID提高10.38%、9.32%和3.48%;而且MLP在许多地物的分类精度上也优于其他图像分类算法,尤其是在Corn-min till、Soybeans-clean till和Bldg-Grass-Tree-Drives等地物中分类精度更为突出。与目前图像分类效果最好的IID算法相比,MLP在Stone-steel towers、Alfalfa等地物中分类精度接近于IID算法;而在Corn-min till、Grass/pasture、Oats等地物中,MLP算法的分类效果要优于IID算法。

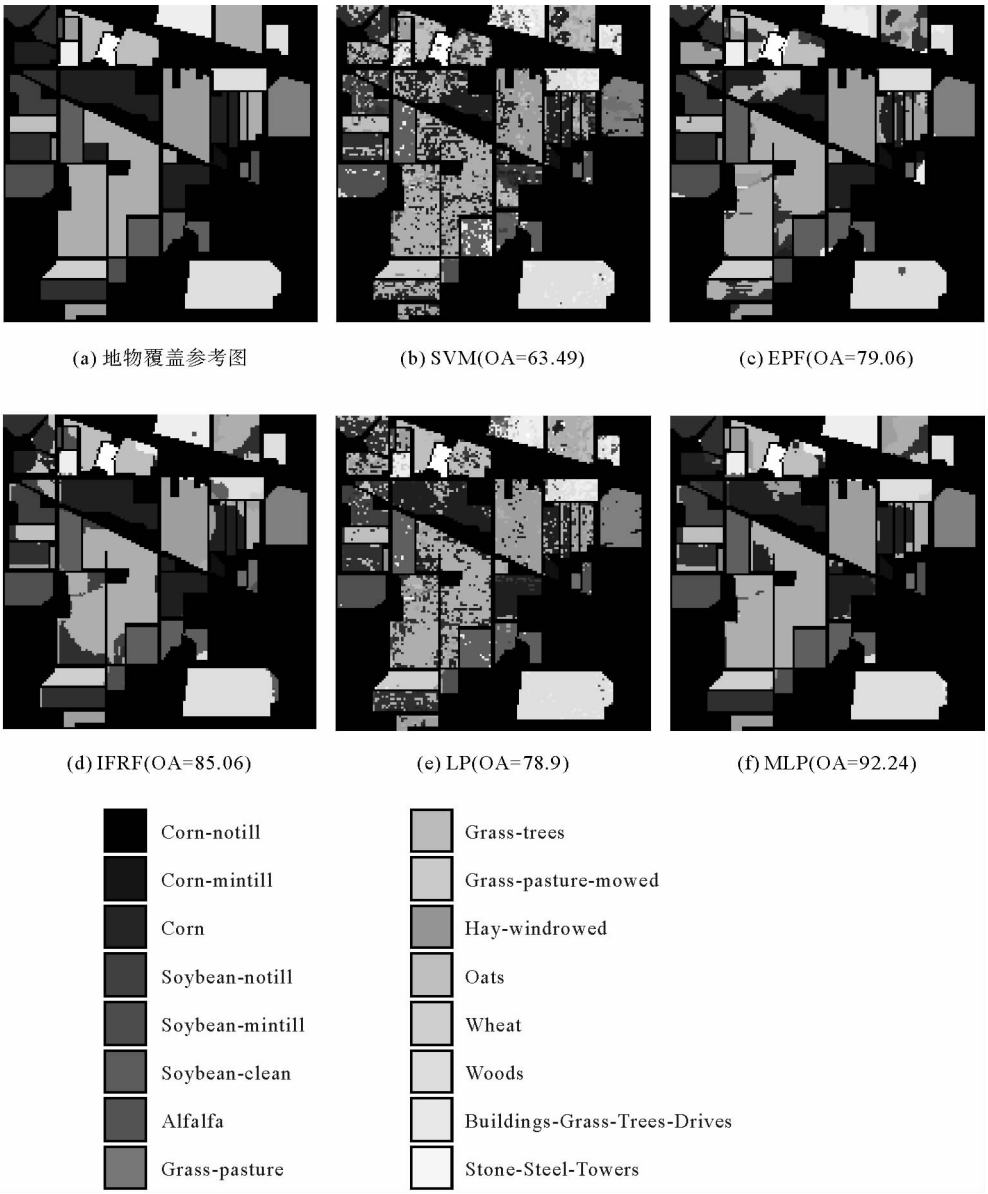

图4 不同图像分类算法对Indian Pines的AVIRIS数据分类性能的比较

Fig. 4 Comparison of the image classification algorithms for the AVIRIS data of Indian Pines

图4(a)是地物覆盖参考图,图4(b)-(f)分别是SVM、EPF、IFRF、LP和MLP的地物分类图。可以看出:图4(f)最接近地物覆盖参考图4(a),MLP分类效果最好。

4 结论

本文提出的改进的标签传播算法(MLP),通过对图像进行融合及迭代滤波提高图像质量,然后通过添加阈值筛选标签传播算法产生的未标记样本增加标记样本的可信度,最后用标记样本和已筛选的未标记样本训练 SVM 进行遥感图像分类。

为了评估 MLP 的分类性能,本文使用 Indian Pines 的 AVIRIS 数据集对 SVM,EPF,IFRF,IID 和 LP 的分类结果进行比较,由实验结果发现,MLP 能够受到不同阈值参数的影响,阈值的合理设置可以显著提高图像分类精度。在已标记样本数量不变的情况下能够明显提高高光谱图像的总体精度。

本文提出的一种改进的标签传播算法尽管能在稀少样本的情况下获得较高的分类精度,但对于阈值的设置是通过实验获得的,没有形式化的表示出来,阈值的设置问题将会是以后研究的重点,改进的标签传播算法是否能够应用到其他高光谱应用中还有待进行深入研究。

参考文献:

[1]KANG X D,LI S T,ATLI J. Spectral-spatial hyperspectral image classification with edge-preserving filtering[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,52(5):2666-2677.

[2]KANG X D,LI S T,FANG L Y,et al. Intrinsic image decomposition for feature extraction of hyperspectral images[J]. IEEE Transactions on Geoscience and Remote Sensing,2015,53(4):2241-2253.

[3]MELGANI F,BRUZZONE L. Classification of hyperspectral remote sensing images with support vector machines[J]. IEEE Transactions on Geoscience and Remote Sensing,2004,42(8):1778-1790.

[4]BOVOLO F,BRUZZONE L,CARLINE L. A novel technique for subpixel image classification based on support vection machine[J]. IEEE Transactions on Image Processing,2010,19(11):2983-2999.

[5]KANG X D,LI S T,FANG L Y,et al. Extended random walker-based classification of hyperspectral Images[J]. IEEE Transactions on Geoscience and Remote Sensing,2015,53(1):144-153.

[6]LI U J,DIAS B,PLAZA J M,et al. Semi-supervised hyperspectral image classification using soft sparse multinomial logistic regression[J]. IEEE Transactions on Geoscience and Remote Sensing,2013,10(2):318-322.

[7]YANG L,YANG S,JIN P,et al. Semi-supervised hyperspectral image classification using spatio-spectral laplacian support vector machine[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,11(3):651-655.

[8]SHAHSHAHANI,B. M. ,LANDGREBED. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing,1994,32(5):1087-1095.

[9]WANG L G,HAO S Y,WANG Q M,et al. Semi-supervised classification for hyperspectral imagery based on spatial-spectral label propagation[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2014,97(1):123-137.

[10]CAMPS-VALLS G,V. BANDOS T,ZHOU D Y. Semi-supervised graph-based hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2007,45(10):3044-3054.

[11]KANG X D,LI S T,ATLI J. Feature extraction of hyperspectral images with image fusion and recursive filtering[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,52(6):3742-3752.

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

[13]LI J,DIAS B,PLAZA A,et al. Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and markovrandom fields[J]. IEEE Transactions on Geoscience and Remote Sensing,2012,50(3):809-823.

(责任编辑:傅　游)