基于正样本和未标记样本的遥感图像分类方法

裔 阳,周绍光,赵鹏飞,胡屹群
YI Yang, ZHOU Shaoguang, ZHAO Pengfei, HU Yiqun

河海大学 地球科学与工程学院,南京 211100
School of Earth Science and Engineering, Hohai University, Nanjing 211100, China

YI Yang, ZHOU Shaoguang, ZHAO Pengfei, et al. Classification method of remote sensing image based on positive and unlabeled data. Computer Engineering and Applications, 2018, 54(4):160-166.

Abstract: Traditional classifier is made up of both positive and negative data. It is a common situation in remote sensing image classification: users are only interested in one specific land-cover type. However, labeling land-cover is a time consuming and labor intensive process, and unlabeled data are usually obtained easily and contain useful information. For this reason, a remote sensing image classification method based on Positive and Unlabeled data(PUL) is proposed. Firstly, according to the inherent characteristics of positive data and combined with support vector data description confident positive and negative samples can be extracted from unlabeled data, and those examples are eliminated from unlabeled set. Then it uses above extracted samples to train a SVM classifier and extract relative confident positive and negative sample from unlabeled set again. The extraction rule is based on the performance of unlabeled set in the SVM classifier. The last step is weighted SVM process. The weight of initial positive and negative samples is 1. The weight of samples extracted by SVM classifier is between 0 and 1. To verify the effectiveness of PUL method, it does classification experiment in remote sensing image and is compared with One-Class SVM(OC-SVM), Gauss Data Description(GDD), Support Vector Data Description(SVDD), Biased SVM and Multi-class SVM. The results show that PUL is helpful to the improvement of classification and better than above OC-SVM methods and Multi-class SVM.

Key words: Biased Support Vector Machine(SVM); Support Vector Data Description(SVDD); Gauss Data Description (GDD); One-Class SVM(OC-SVM); remote sensing image classification; Multi-class SVM

摘 要:传统分类器的构建需要正样本和负样本两类数据。在遥感影像分类中,常出现这样一类情形:感兴趣的地物只有一种。由于标记样本耗时耗力,未标记样本往往容易获取并且包含有用信息,鉴于此,提出了一种基于正样本和未标记样本的遥感图像分类方法(PUL)。首先,根据正样本固有特征并结合支持向量数据描述(SVDD)从未标记集筛选出可信正负样本,再将其从未标记集中剔除;接着将其带入SVM训练,根据未标记集在分类器中的表现设立阈值,再从未标记集中筛选出相对可靠的正负样本;最后是加权SVM(Weighted SVM)过程,初始正样本及提取出的可靠正负样本权重为1,SVM训练筛选出的样本权重范围0~1。为验证PUL的有效性,在遥感影像进行分类实验,并与单类支持向量机(OC-SVM)、高斯数据描述(GDD)、支持向量数据描述(SVDD)、有偏SVM(Biased SVM)以及多类SVM分类对比,实验结果表明PUL提高了分类效果,优于上述单类分类方法及多类SVM方法。

关键词:有偏SVM;支持向量数据描述;高斯数据描述;单类支持向量机;遥感图像分类;多类SVM
文献标志码:A **中图分类号**:TP751.1 doi:10.3778/j.issn.1002-8331.1609-0184

1 引言

遥感影像分类是遥感信息提取的重要手段,是目前遥感技术中的热点研究内容[1]。传统的遥感影像分类方法需要标记所有地物类别,但当感兴趣的地物只有一种时,其他地物并不在考虑范围之内,标记所有类别未免过于浪费[2]。比如进行温度反演之前需要将水体去除,

基金项目:国家自然科学基金(No.41271420/D010702)。
作者简介:裔阳(1991—),男,硕士生,主要研究方向:模式识别;周绍光(1966—),男,副教授,主要研究方向:遥感图像处理。
收稿日期:2016-09-13 **修回日期**:2016-11-01 **文章编号**:1002-8331(2018)04-0160-07
CNKI **网络优先出版**:2017-02-28, http://kns.cnki.net/kcms/detail/11.2127.TP.20170228.1845.028.html

则分类地物就是水体;从遥感影像中检索道路更新交通系统,则只要提取出道路,其他植被建筑等地物不需要考虑在内等,以上方法被称为单类分类(one class classification)[3]。单类分类的一个巨大优势在于仅依靠正样本就可以达到近似于传统分类方法的分类效果,意味着遥感影像的分类目标唯一,训练样本只要某类地物作为训练样本,剩余的都当作负样本,大大减少了人工标记代价。

单类分类是在负样本无法获取或者获取难度较大的情形下被提出的,相关学者已经做了大量研究并且取得了令人鼓舞的效果[4-8]。单类分类器的分类方法主要有基于密度的方法、基于神经网络的方法、基于聚类的方法、基于支持域的方法[9]。其中,常用并且效果较好的有单类支持向量机(OC-SVM)[10]、高斯数据描述(GDD)[11]、支持向量数据描述(SVDD)[12]。这些方法都是将感兴趣的类别作为目标类,其余的当作离群类,同时训练数据只需要目标类。比如支持向量数据描述将训练样本通过核函数映射到高维空间,寻找到一个体积尽可能小包含所有样本的超球体;单类支持向量机原理与支持向量数据描述相似,在高维空间最大化超平面与原点(视为负例)的间隔;高斯数据描述是假设目标数据服从高斯分布,并且可以从目标数据中估计得到。

相对于获取标记正样本的耗时耗力,未标记样本的获取要容易许多。虽然标记样本和未标记样本的获取难度有巨大的反差,但是未标记样本同样蕴含着对于构建分类器有用的信息[13],因此利用正样本和未标记样本数据学习(PU学习)分类器成为一种重要的分类问题[14-16]。现阶段,PU学习主要用于文本分类[17-18],已经有许多研究成果。PU学习的一个思路是从未标记样本中筛选出可信的负例样本,然后进行两类分类[19-20]。Liu等人将部分正样本集放入未标记集作为Spy带入I-SE训练,观察Spy在未标记集中的表现建立阈值函数,筛选出可信负样本集RN[20],同时也利用提出的l-DNF算法提取可信负样本。富震提出了改进的1-DNF方法筛选出可信负样本集RN的基础上,结合主动学习技术迭代运行SVM,得到了较好的分类结果[21];Ke等将正样本和未标记样本分别作为正样本集和负样本集赋以不同权重训练,迭代筛选出可信的正样本集,直至未标记集中可被认为是只有很少一部分正样本时迭代停止,最后赋以原始正样本、可信正样本、剩余的未标记样本不同权重获取最终的分类器[22]。

虽然以上PU学习都取得了不错的效果,但都是从未标记集中获取可信负集RN从而进行后续分类。当初始的正样本数量很有限时,从未标记集中提取过多的RN进行两类分类会导致样本数量严重不平衡,影响最终的分类精度,且上述算法大多使用用于文本分类的1-DNF算法获取可靠负样本集RN,在遥感图像中并不适用。Ke等不断从未标记集中提取出正样本,不会出现上述数据不平衡问题,但是需要迭代运行SVM算法,效率较低。鉴于此,结合正样本固有特征和支持向量数据描述提出了一种FCPN(Find Confident Positive and Negative samples)方法。从未标记样本集筛选出可信正负样本,将该方法的结果作为PUL的输入。为验证PUL的有效性,在高光谱遥感影像Pavia University中进行分类实验,并与GDD、SVDD、OC-SVM、Biased-SVM以及多类SVM方法进行对比实验。

2　PUL方法

2.1　支持向量数据描述(SVDD)

SVDD是Tax等人提出的一种基于支持域的单类分类算法[12]。假设训练数据集$\{x_i, i=1,2,\cdots,l\}$,SVDD的目的是在高维空间寻找到一个体积尽可能小的包含所有训练样本的超球体。SVDD的目标函数如下:

$$\min R^2+C\sum_{i=1}^{l}\delta_i$$
$$\text{s.t.}\ \ \|x_i-O\|\leqslant R^2+\delta_i,\delta_i\geqslant 0 \tag{1}$$

式(1)中R表示超球体的半径,C是惩罚因子,用来权衡训练样本被排除在超球体外的数目和超球体体积的大小,表明为了获取最优超球体,允许一部分样本被隔离在超球体的外部。δ_i是松弛标量,O是超球体的球心,引入拉格朗日乘子进行求解:

$$L(R,O,\delta_i,\alpha_i,\lambda_i)=R^2+C\sum_{i=1}^{n}\delta_i-\sum_{i=1}^{n}\alpha_i(R^2+\lambda_i-\|\phi(x_i)-O\|^2)-\sum_{i=1}^{n}\lambda_i\delta_i \tag{2}$$

超球体建立完成之后,对于一个新的未知点x,若满足$f(x)$小于0,则被认为是目标类:

$$f(x)=k(x,x)-2\sum_{i=1}^{n}\alpha_i k(x_i,x)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(x_i,x_j)-R^2 \tag{3}$$

2.2　有偏SVM(Biased SVM)

支持向量机[23]是由支持向量构成的两类分类器,适合小样本、高维数据的分类问题[24],但是PU学习只有少量正样本和大量的未标记样本,因此Liu等人提出Biased SVM方法[20]。Biased SVM认为未标记集是含有噪声数据(正样本)的负样本集,该分类器将标记样本作为正集和未标记样本U作为负集带入训练。假设m和n分别是标记正样本和未标记样本的数目,对于所有训练样本,下式问题需要解决:

$$\min_{w,b,\xi}\frac{1}{2}\|w^2\|+C_+\sum_{i=1}^{m}\xi_i+C_-\sum_{i=m+1}^{m+n}\xi_i$$
$$\text{s.t. } y_i(w\cdot x_i+b)\geqslant 1-\xi_i, i=1,2,\cdots,m+n;$$
$$\xi_i\geqslant 0, i=1,2,\cdots,m+n \quad (4)$$

C_+ 和 C_- 分别是正样本和负样本错分类的惩罚因子，ξ_i 是松弛变量，实践证明Biased SVM比大部分两步策略效果更好[20]。Biased SVM不合理之处是它赋以未标记集 U 中所有样本同一权值，而 U 中也包含部分正样本，当 U 中的正样本较多时，Biased SVM效果会有所降低。

2.3 FCPN

文本分类中的1-DNF算法是将正例集合中出现频率大于在未标识集合中出现频率的特征构造正例特征集合 PF，未标记样本中不含有任何 PF 中特征的被判定为可信负样本(Confident Negative Samples)。然而，与文本特征不同的是遥感图像正负样本之间的特征数目相同，不同的是特征之间的差异性，FCPN方法的一个重要假设就是不同类别的样本之间的特征差异性大于同类样本。FCPN主要有两部分构成：

(1)根据遥感影像固有特征获取初始可信正样本和可信负样本；

(2)利用训练样本建立多个超球体，从初始可信正负样本中筛选得到最终的可信正负样本。

2.3.1 获取初始可信正负样本

设正样本集中有 l 个正样本，$P=\{x_1,x_2,\cdots,x_l\}$，每个样本有 n 个特征 $x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$；设 $\boldsymbol{m}$ 为 $3\times n$ 的矩阵，其中 $i=1,2,\cdots,n$ 是矩阵 $\boldsymbol{m}$ 的列坐标，具体 $\boldsymbol{m}$ 矩阵如下：

$$\boldsymbol{m}(1,i)=(\sum_{j=1}^{l}x_{ji})/l \quad (5)$$

$$\boldsymbol{m}(2,i)=(\sum_{j=1}^{l}|x_{ji}-m(1,i)|/m(1,i))/l \quad (6)$$

$$\boldsymbol{m}(3,i)=(\sum_{j=1}^{l}\|x_{ji}-m(1,i)|/m(1,i)-m(2,i)|/m(2,i))/l \quad (7)$$

设未标记集中有 N 个未标记样本，$U=\{x_1,x_2,\cdots,x_N\}$，样本特征 $x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$；设 $\boldsymbol{SUMF}$ 为 $N\times 4$ 的矩阵，其中 $k=1,2,\cdots,N$ 是矩阵 $\boldsymbol{SUMF}$ 的行坐标，具体 $\boldsymbol{SUMF}$ 矩阵前两列如下：

$$\boldsymbol{SUMF}(k,1)=\sum_{i=1}^{n}|x_{ki}-\boldsymbol{m}(1,i)|/\boldsymbol{m}(1,i) \quad (8)$$

$$\boldsymbol{SUMF}(k,2)=\sum_{i=1}^{n}\|x_{ki}-\boldsymbol{m}(1,i)|/\boldsymbol{m}(1,i)-\boldsymbol{m}(2,j)|/\boldsymbol{m}(2,i) \quad (9)$$

算法流程：

(1)根据正样本集 P 得到上述大小为 $3\times n$ 的 $\boldsymbol{m}$ 矩阵。

(2)输入 N 个验证集样本进行类别判定，建立 $N\times 3$ 的矩阵 $\boldsymbol{SUMF}$，并得到 $\boldsymbol{SUMF}$ 的前两列，$\boldsymbol{SUMF}$ 的后两列初始值为0。

(3)当未标记集 U 的某个样本 x_k 的第 i 个特征同时满足以下3个条件时，$\boldsymbol{SUMF}(k,3)+1$。

条件1

$|x_{ki}-m(1,i)|/m(1,i)>\boldsymbol{S}_1*\boldsymbol{m}(2,i)$

条件2

$\boldsymbol{SUMF}(k,i)<\boldsymbol{S}_2*\boldsymbol{mean}(SUMF(:,2))$

条件3

$\|x_{ki}-\boldsymbol{m}(1,i)|/\boldsymbol{m}(1,i)-\boldsymbol{m}(2,i)|/\boldsymbol{m}(2,i)>\boldsymbol{S}_3*\boldsymbol{m}(3,i)$

$\boldsymbol{SUMF}$ 的第3列数值范围为0到 n，当样本 k 中满足上述3个条件的特征数目大于 $0.2n$ 时，该样本被当作可信负样本。其中，$\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$ 的取值由验证集调整，一般分别取1、1.2、3，训练不同类别样本时只需要微调，$\boldsymbol{S}_1$:[1,1.2]，$\boldsymbol{S}_2$:[1.1,1.3]，$\boldsymbol{S}_3$:[2.8,3.2]，步长均为0.1。

(4)当未标记集 U 的某个样本 x_k 的第 i 个特征同时满足以下3个条件时，$\boldsymbol{SUMF}(k,4)+1$。

条件1

$|x_{ki}-\boldsymbol{m}(1,i)|/\boldsymbol{m}(1,i)<T_1*\boldsymbol{m}(2,i)$

条件2

$\boldsymbol{SUMF}(k,i)>T_2*\boldsymbol{mean}(\boldsymbol{SUMF}(:,2))$

条件3

$\|x_{ki}-\boldsymbol{m}(1,i)|/\boldsymbol{m}(1,i)-\boldsymbol{m}(2,i)|/\boldsymbol{m}(2,i)<T_3*\boldsymbol{m}(3,i)$

$\boldsymbol{SUMF}$ 的第4列数值范围同样为0到 n，不难发现可信正样本与可信负样本求得的条件类似，当样本 k 中满足上述3个条件的特征数目大于 $0.3n$ 时，该样本被当作可信负样本。其中，T_1、T_2、T_3 取值由验证集调整，一般取0.8、1.1、0.9，不同类别训练样本可做微调，T_1:[0.7,0.9]，T_2:[1,1.2]，T_3:[0.8,1]，步长均为0.1。需要说明的是该方法的一个重要假设是不同类别样本之间的差异性大于同类样本，而上述过程的条件2与其不符，这是因为通过大量数据对比发现满足条件1和3而被判别错误的样本往往不满足条件2，通过添加条件2能够取得预期目的。

2.3.2 结合SVDD获取最终可信正负样本

SVDD将训练样本映射至高维数据并用体积最小的超球体将其包裹住，并能在遥感影像分类取得良好效果[25]。只依赖一个超球体能够达到理想的分类精度，但可能会将部分负样本错分入超球体中，由于上述步骤已经得到初始可信正负样本，为了筛选出的最终可信正负样本错误率降到最低，从正样本集中随机抽取10组相同数量的样本分别建立超球，当需要判定类别的样本同时在10个超球内时，该样本作为最终可信正样本。图1显示了一种简单情形：绿色和红色的点分别表示正样本和负样本，可以看出超球边界1和2里面都包含了大部分正样本，但是也存在少部分负样本，而两个边界重叠部分全部为正样本。虽然重叠部分也去除了部分正样本，但留下的都是可信的负样本。FCPN方法的第一步

就已经得到了初始可信的正负样本,再经过多个超球判定后留下的样本被认为是最终的可信正负样本。

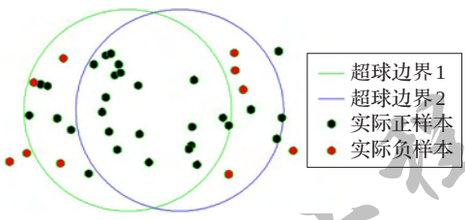

图1　多超球筛选可信正负样本

2.3.3　加权支持向量机(Weighted SVM)

类似于Biased SVM对正负样本赋以不同的权重,Weighted SVM针对不同样本的类别可信度大小赋以不同的权重。假设初始的正样本集 P ,PCPN筛选出部分可信正负样本,分别为 PO_1、NE_1。然后将 P 和 PO_1 作为正样本, NE_1 作为负样本带入SVM训练;由于SVM的类别判定阈值为0,即阈值大于0则为正,反之为负,为了从剩下的未标记集获取可信度较高的样本,求得未标记集所有判定为正的样本的阈值去平均,记为 $th1$,相应的负样本阈值为 th_2 ,提取出当阈值大于 $th1$ 的正样本、阈值小于 th_2 的负样本,假设此步骤得到的正负样本为 PO_2 、NE_2 ;最后将 P、PO_1、PO_2 和 NE_1 、NE_2 带入Weighted SVM,最小化下列公式:

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C_{ppn1}\sum_{i=1}^{m}\xi_i+C_{pn2}\sum_{i=m+1}^{m+n}\xi_i$$
$$\text{s.t.}\quad y_i(w\cdot x_i+b)\geqslant 1-\xi_i,\ i=1,2,\cdots,m+n;$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,m+n \qquad (10)$$

其中 C_{ppn1}、C_{pn2} 分别为 P、PO_1、NE_1 和 PO_2、NE_2 的惩罚因子,给不同训练样本错误赋以不同权重,在验证集中可调节 C_{ppn1}、C_{pn2} 获取最优分类器。根据样本可信度的差异, C_{ppn1} 取1, C_{pn2} 取值范围(0,1]。

2.3.4　PUL流程图

PUL方法流程图如图2所示。

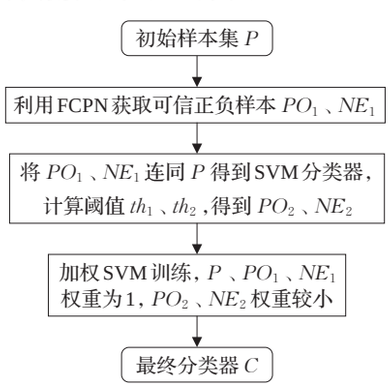

图2　PUL方法流程图

3　实验设计及结果

3.1　数据集描述

实验数据是高光谱遥感影像Pavia University,是用ROSIS高光谱传感器系统在意大利南部的Pavia市的Pavia University上空拍摄的,数据大小为610×340像素,除去噪声波段,图像包含有103个连续波段,空间分辨率为1.3 m,该地区共包含Asphalt、Meadows、Gravel等9种地物,一共标记42 776个点,如表1所示。为提高运算速率,采用Chang等的MVPCA算法[26]进行波段选择,最终取36、50、63、69、93这5个波段的数据,求得每个波段的均值、方差、同质性、对比度和二阶矩[2]。所有特征通过灰度共生矩阵以3乘3窗口大小求得,然后将数据归一致0-1,一共得到25个特征,每次分类将目标样本视为正样本,其余样本作为负类样本并建立测试集、验证集和未标记集。其中测试集test_set包含50%正样本和80%负样本,验证集va_set由30%正样本和16%的负样本构成,剩余的作为未标记样本集 *unlabeled_set* ,从未标记样本集选出50个正样本作为初始训练正集 P ,所有样本均为随机选取。

表1　标记地物类别及数目

类别	地物	样本数目
1	Asphalt	6 631
2	Meadows	18 649
3	Gravel	2 099
4	Trees	3 064
5	Painted metal sheets	1 345
6	Bare Soil	5 029
7	Bitumen	1 330
8	Self-Blocking Bricks	3 682
9	Shadows	947

3.2　实现FCPN

为了验证FCPN方法的有效性,每次训练都是随机选取其中50个目标类样本作为初始正样本集 P ,每类样本训练5次求平均值。具体结果如表2所示,第二列是用FCPN方法提取出的可信正样本的平均数量,括号内是初始可信正样本中错误样本的数目;第三列是用FCPN方法提取出的可信负样本的平均数量;第四列是用SVDD判别之后剩余的正样本数量,括号内都是错误样本数目。

不难看出,仅依据影像固有特征筛选出的初始可信正样本中约有5%致20%的错误,因此后续用多超球SVDD判别出最终的可信正样本很有必要,而初始可信负样本的错误率近乎为0,不需要多超球SVDD判定步骤。

多超球SVDD的参数选择流程如下:

(1)随机将50正样本分成10组,每组30个;

(2)惩罚因子 c 范围[0.01,0.2],步长0.01,核宽度 s 范围[0.02,20],步长0.02;

(3)求出每组样本相应参数在验证集的正确率;

(4)选取正确率最高的参数作为多超球SVDD的最终参数。

对比表2的第一列与第三列,SVDD判别后的正样本错误率大大降低,近乎为0。可见FCPN最终筛选出的可信正负样本错误率极低,为后续分类的正确性提供了保障。

表2　各类筛选的可信正负样本数目

地物	初始可信正	初始可信负	SVDD判别后
Asphalt	112.4(15.6)	216.4(0)	96.2(0)
Meadows	128.8(2.8)	296.2(0.2)	97.4(0)
Gravel	100(17.2)	446.2(0)	80.8(0)
Trees	181.2(10.4)	163(0)	144.4(0)
Painted metal sheets	28.0(5.4)	277.2(0)	19.8(0)
Bare Soil	134.2(22.0)	66.4(0)	54.6(0)
Bitumen	46.2(1.8)	543.0(0)	32.8(0)
Self-Blocking Bricks	123.2(11.0)	49.0(0)	61.0(0.2)
Shadows	34.0(3.2)	653.0(0)	15.2(0)

3.3　影像分类结果

为验证PUL方法的有效性,本文与OC-SVM、GDD、SVDD及Biased SVM进行对比。上述4种方法都已被证明有着很好的分类效果。除了Biased SVM需要未标记集数据,另外3种方法都是单类分类方法,即只需要目标样本作为训练数据。无论是PUL还是另外四种方法,都需要通过包含正样本和负样本的验证集 *va_set* 调节具体参数,更多的标记样本意味着更多的人工标记负担,所有实验方法的初始训练样本数目均为50个。其中PUL、OC-SVM和Biased SVM使用Libsvm工具箱实现,SVDD和GDD则使用Data Description Toolbox,所有操作均在Matlab中完成,图3是影像原图。

图3　影像原图

为了比较的公正性,5种方法均用在验证集 *va_set* 实验效果最好的参数。实验中SVDD和GDD均使用径向基(RBF)核函数,用以下方法调节两个参数:惩罚因子 c 和核宽度 s,惩罚因子 c 取值范围[0.01,0.2],核宽度 s 取值范围[0.2,200]。OC-SVM同样使用RBF核函数,训练样本拒绝率 *nu* 范围为[0.01,0.2],训练步长为0.01,核宽度 s 范围为[0.1,200],步长为0.1。Biased SVM的核宽度 s 范围为[0.1,200],步长为0.1,正样本权重为1,负样本权重为 $1,2^{-1},\cdots,2^{-7}$。PUL在获取 PO_2 和 NE_2 之后,赋以 P、PO_1、NE_1 的权重为 1,PO_2 和 NE_2 的权重同样为 $1,2^{-1},\cdots,2^{-7}$。

从表1可知不同类别之间的样本数量差距很大,Shadows类只有不到1 000个,Meadows类则数目接近20 000,无论是验证集还是测试集,负类样本占据了绝大部分。若某分类器假阴率(Flase Negative Rate)很高,即使大部分正样本都被判别错误,正确率依然很高,当然,该分类器实际是很糟糕的。因此使用 *Fscore* 作为分类器的判别指标[27], $Fscore=2\times r\times p/(r+p)$,$r$ 和 p 分别表示召回率(Recall Rate)和准确率(Precision Rate)。图3是影像原图,由于样本类别较多,图4只显示Asphalt、Gravel、Trees、Bare Soil、Bitumen的分类效果图,图4的列表示具体类别,行表示不同方法,图5是文中提及方法所有地物的Fscore柱状图,表3是相应的Fscore表格,每类的最高值加粗表示。总体来讲,PUL方法提取的相应类别地物与原图的相似度最高,GDD和SVDD召回率(Recall Rate)较低,过多正样本没有被识别出来,尤其是Trees和Bare Soil类,所以GDD和SVDD的Fscore值几乎都是最低的;OC-SVM效果相对而言要比GDD和SVDD好很多,每一类的Fscore值比较平稳,但是比PUL要低,只有Gravel类与PUL效果接近;Biased-SVM效果与PUL比较接近,在类Asphalt、Trees、Painted metal sheets、Shadows中还略高于PUL方法,但是Biased-SVM单纯地将所有未标记样本均作为负样本,导致未标记集中含有数量较多正样本时效果不佳,其Fscore值反差很大,这一点在类Meadows反应特别明显,其Fscore值只有0.023 1。PUL分类方法在大多数类别中的Fscore值都比其他方法更高,且相对稳定。

以上方法是用单类分类方法将9类地物分别提取出来,下面用多类SVM算法与PUL相比较。训练样本共450个,每类50个样本,测试样本每类200个,用Matlab自带的遗传算法参数寻优,核宽度 s 取值范围[0.1,100],惩罚因子 c 取值范围[0.1,50],步长均为0.1,采用“一对多”的方法进行多类分类。多类SVM分类的总体精度是89.06%,表4是分类混淆矩阵,矩阵内都是分类5次求平均的值,第一行是类别,第一列是该类的Fscore值,其中类Gravel、Bare Soil、Bitumen的Fscore值略高于PUL,其余的则是PUL方法更优。

综上所述,本文PUL方法可以很好地进行遥感图像分类,比其他单类分类方法更优,同时也优于多类SVM分类方法。

4　结语

本文基于有限正样本和未标记数据提出了PUL遥感图像分类方法。PUL的优势在于可以依靠有限正样本从未标记样本中提取可信的正负样本,然后通过SVM筛选出相对可信的正负样本,最后用加权SVM得

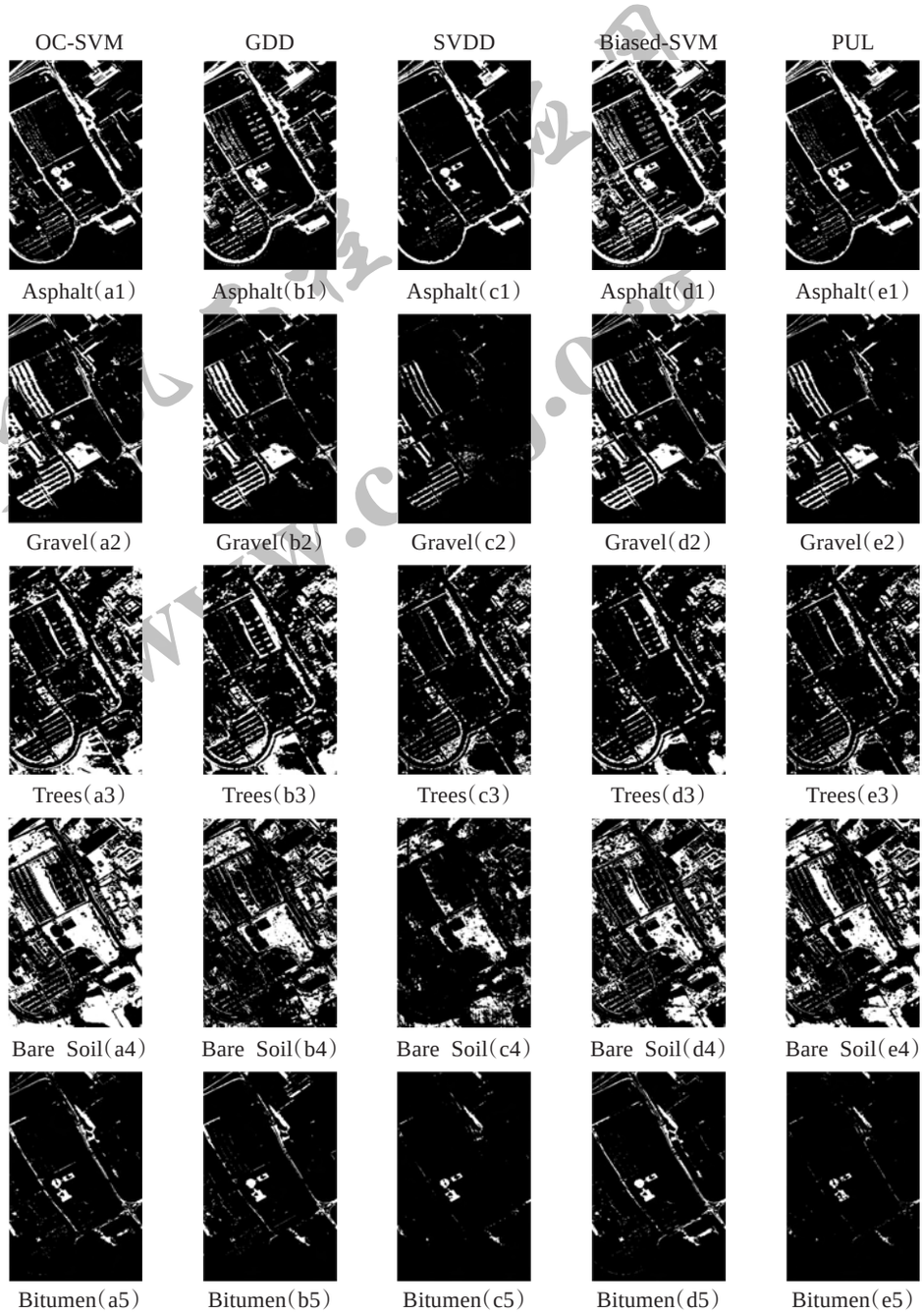

图4　几种方法分类效果图

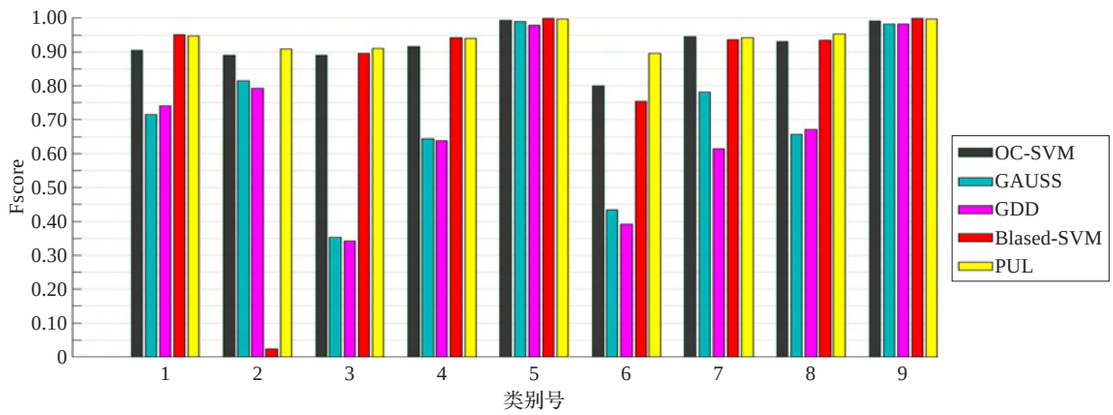

图5　几种方法 Fscore 对比图

表3 几种方法Fscore具体值

类别	1	2	3	4	5	6	7	8	9
OC-SVM	0.904 8	0.890 6	0.891 1	0.916 2	0.992 4	0.801 3	0.945 1	0.931 6	0.991 0
GDD	0.715 3	0.816 2	0.352 7	0.643 2	0.988 7	0.433 3	0.781 4	0.656 2	0.981 2
SVDD	0.742 1	0.792 4	0.341 2	0.637 6	0.979 2	0.391 8	0.614 6	0.672 3	0.982 3
B-SVM	**0.951 2**	0.023 1	0.896 6	**0.941 5**	**0.999 1**	0.754 3	0.936 2	0.934 2	**0.998 1**
PUL	0.948 0	**0.909 2**	**0.911 5**	0.940 1	0.996 1	**0.895 4**	**0.942 6**	**0.952 8**	0.997 1

表4 多类分类混淆矩阵

Fscore	1	2	3	4	5	6	7	8	9
0.914 9	172.4	0.0	1.0	0.0	1.0	0.0	9.8	1.0	0.0
0.921 8	0.2	178.4	0.0	33.0	0.0	18.8	0.0	0.0	0.6
0.930 2	10.4	0.2	182.4	0.0	0.0	0.0	1.2	52.0	0.0
0.892 6	0.0	0.0	0.0	164.8	0.0	6.2	0.0	0.0	0.6
0.993 6	0.8	0.0	0.0	0.0	198.0	1.6	0.0	0.0	0.0
0.905 4	0.2	21.4	0.0	2.2	0.0	170.4	0.0	1.0	0.2
0.964 6	10.0	0.0	1.0	0.0	0.0	0.0	188.8	0.0	0.0
0.825 6	5.8	0.0	15.6	0.0	1.0	3.0	0.2	146.0	0.2
0.995 4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	198.4

到最终的分类器,而不是直接将未标记样本作为负样本再加权处理。实验结果表明,PUL方法能够很好地完成遥感图像分类,并且有较高Fscore值。本文不足之处是PUL方法中间过程参数较多,需要调节,这将是进一步研究的内容。

参考文献:

[1] 贾坤,李强子,田亦陈,等.遥感影像分类方法研究进展[J].光谱学与光谱分析,2011,31(10):2618-2623.

[2] Li W,Guo Q,Elkan C.A positive and unlabeled learning algorithm for one-class classification of remote-sensing data[J].IEEE Transactions on Geoscience and Remote Sensing,2011,49(2):717-725.

[3] Tax D M.One-class classification[D].Delft University of Technology,2001.

[4] Kowalczyk A,Raskutti B.One class SVM for yeast regulation prediction[J].ACM SIGKDD Explorations Newsletter,2002,4(2):99-100.

[5] Munoz-Mari J,Camps-Valls G,Gomez-Chova L,et al.Combination of one-class remote sensing image classifiers[C]//Proceedings of Geoscience and Remote Sensing Symposium,2008:1509-1512.

[6] Sanchez-Hernandez C,Boyd D S,Foody G M.One-class classification for mapping a specific land-cover class:SVDD classification of fenland[J].IEEE Transactions on Geoscience and Remote Sensing,2007,45(4):1061-1073.

[7] Zeng Y,Lan J.Imitate geometric manifold coverage method for one-class classification of remote sensing data[J].Arabian Journal of Geosciences,2015,8(2):631-638.

[8] Guerbai Y,Chibani Y,Hadjadji B.The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters[J].Pattern Recognition,2015,48(1):103-113.

[9] 潘志松,陈斌,缪志敏,等.One-Class分类器研究[J].电子学报,2009,37(11):2496-2503.

[10] Chen Y,Zhou X S,Huang T S.One-class SVM for learning in image retrieval[C]//Proceedings of International Conference on Image Processing,2001:34-37.

[11] Tax D M J,Duin R P W.Uniform object generation for optimizing one—class classifiers[J].Journal of Machine Learning Research,2001,2(2):155-173.

[12] Tax D M,Duin R P.Support vector domain description[J].Pattern Recognition Letters,1999,20(11):1191-1199.

[13] Castelli V.The relative value of labeled and unlabeled samples in pattern recognition[D].Stanford University,1995:355-355.

[14] Elkan C,Noto K.Learning classifiers from only positive and unlabeled data[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Las Vegas,Nevada,USA,2008:213-220.

[15] Calvo B,Larrañaga P,Lozano J A.Learning Bayesian classifiers from positive and unlabeled examples[J].Pattern Recognition Letters,2007,28(16):2375-2384.

[16] Mordelet F,Vert J P.A bagging SVM to learn from positive and unlabeled examples[J].Pattern Recognition Letters,2014,37:201-209.

[17] Li X,Liu B.Learning to classify texts using positive and unlabeled data[C]//Proceedings of International Joint Conference on Artificial Intelligence,2003:587-592.

(下转230页)