

doi: 10. 6046/gtzyyg. 2013. 02. 16

引用格式: 任广波,张杰,马毅,等. 基于半监督学习的遥感影像分类训练样本时空拓展方法[J]. 国土资源遥感,2013,25(2): 87-94. (Ren G B,Zhang J,Ma Y,et al. Extending method of remote sensing image training sample based on semi-supervised learning in both time and spatial domain[J]. Remote Sensing for Land and Resources,2013,25(2): 87-94.)

基于半监督学习的遥感影像分类 训练样本时空拓展方法

任广波,张 杰,马 毅,宋平舰
(国家海洋局第一海洋研究所,青岛 266061)

摘要: 针对无法直接获取训练样本的遥感影像分类问题,从满足条件的其他影像中选择替代训练样本是最直接的方法,但由于地物类型在不同影像中的辐射环境不同,导致替代训练样本对待分类影像的代表性较差,无法保证分类精度。以直推式支持向量机(transductive support vector machine,TSVM)分类为例,发展了一种基于半监督学习的遥感影像训练样本时空拓展方法。该方法采用非监督方法从待分类影像中选择大量未标记样本,挖掘各类地物在特征空间中的结构信息;以替代训练样本所拟合的分类面为初始面,通过自适应渐进式的优化,实现对待分类影像的高精度分类。该方法要求训练样本的来源影像与待分类影像具有相似的地物分布和相近的时相。以 SPOT5 和 QuickBird 影像分类为例,分别通过基于像元的和基于分割对象的分类实验证实,该文提出的方法可有效地实现训练样本的时空拓展应用。

关键词: 遥感分类;半监督学习;直推式支持向量机(TSVM);样本拓展应用

中图法分类号: TP 75 **文献标志码:** A **文章编号:** 1001-070X(2013)02-0087-08

0 引言

在遥感影像分类中,对某些无法实施现场踏勘和没有历史资料可参考的区域进行分类时,从其他类似影像中选择替代训练样本进行拓展应用是最直接的方法;面向大规模、重复性的遥感监测活动,该方法又可为样本的重复利用提供可行途径。由于地物类型在不同影像中的辐射环境不同,导致替代训练样本对待分类影像的代表性较差,以致所训练出的分类器产生偏差,无法保证获得好的分类精度^[1-4]。

训练样本的拓展应用是遥感影像自动化分类的重要途径之一,更是当前模式识别领域面临的亟待解决的技术问题^[5]。Du 等^[6]、张友水等^[7]、Koukal 等^[8]和 Ren 等^[9]在对影像进行相对辐射校正或绝对辐射校正的基础上,提出了遥感影像分类样本的拓展应用方法。然而,且不论一般的影像不可能完成绝对辐射校正^[2],仅相对辐射校正也因其操作流程繁琐复杂、对操作者主观依赖性强而较难实现。

遥感影像中,同类型样本会在特征空间中表现出明显聚类特征。通过引入大量的未标记样本,挖

掘属于不同类型的样本在特征空间中的分布特征;利用这些特征,即使所选的训练样本在对影像类型特征的代表性上存在偏差,也能通过有效利用未标记样本估计出一个能在待分类影像中有较好泛化能力(generalization ability,即机器学习算法对新鲜样本的适应能力)的分类面,这便是本文要引入的解决训练样本时空拓展问题的半监督学习方法。

半监督学习是数据挖掘领域中的一种方法,其同时利用已标记样本(即训练样本)和未标记样本,通过挖掘未标记样本中所隐含的各待分类类型在特征空间中的固有结构信息,对已标记样本因代表性不好而造成的拟合分类器有偏差情况进行矫正。该思想产生于 20 世纪 60 年代,Scudder^[10],Fralick^[11]以递归式自学习的方式最早发现了未标记样本的作用;直到 20 世纪 70 年代,可有效估计不完全数据中各生成成分参数的 EM 算法得到了普及,使得半监督学习思想逐渐得到了认可。半监督学习思想已在文本分类^[12-13]和 Web 挖掘^[14-15]等领域得到广泛应用。在遥感影像分类中,有关半监督学习的研究成果不多^[5,16-22]。上述研究的出发点是解决在精力或资源有限情况下的小样本分类问题,但训练

样本依然来自待分类影像自身,而基于其他影像替代训练样本的半监督学习分类方法的文献报道却不多。同时,半监督学习算法中应用最多的是 Joachims^[23]提出的训练算法,但该算法实现条件苛刻且效率低。陈毅松等^[24]、沈新宇等^[25]和廖东平等^[26]针对该问题提出了相应的优化算法,但都因未考虑遥感影像分类特点(数据量大、类别分布不均衡和先验知识难以获取等)导致在遥感影像分类中适用性不好。

针对以上问题,本文以直推式支持向量机(transductive support vector machine, TSVM)分类为例,开展基于半监督学习的遥感影像分类训练样本拓展应用方法研究。选择覆盖我国海岸带区域的 2 景 SPOT5 影像和 2 景覆盖海岛的 QuickBrid 影像,进行训练样本拓展分类应用实验(对 SPOT5 影像应

用基于像元的分类策略,对 QuickBird 影像应用基于分割对象的分类策略),以证实本文提出的方法在遥感影像分类训练样本时空拓展应用中的有效性。

1 数据与方法

1.1 数据和实验设计

1.1.1 遥感影像和覆盖区域

选择覆盖相似地物分布区域、获取时相相近的 SPOT5 和 QuickBird 遥感影像进行分类实验。选择这 2 种影像的原因是其分别为当前中、高空间分辨率遥感应用中的主流影像,也是 m 级和亚 m 级空间分辨率遥感影像的代表影像,开展基于这 2 种影像的分类研究对其他同类影像有参考意义。影像覆盖区域和相应的快视图如图 1 所示。

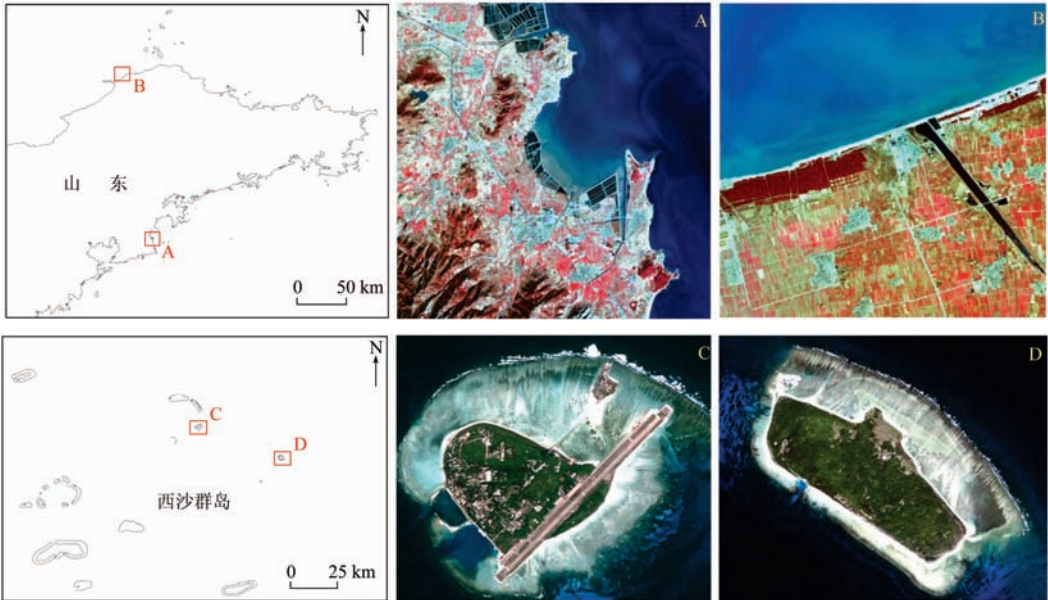


图 1 SPOT5 影像(A,B)和 QuickBrid 影像(C,D)覆盖区域和相应的快视图
Fig.1 SPOT5 (A,B) and QuickBird (C,D) images and their coverage

图 1 中,SPOT5 影像 A 覆盖青岛市崂山区东海岸的鳌山卫,影像 B 覆盖烟台市龙口市栾家口,由于该类影像没有蓝光波段,故其快视图为近红外波段(R)、红光波段(G)和绿光波段(B)的彩色合成影像; QuickBird 影像 C 和 D 分别覆盖我国三沙市西沙群岛的永兴岛和东岛,均为多光谱影像,其快视图为红光波段(R)、绿光波段(G)和蓝光波段(B)的彩色合成影像。2 组影像详细信息如表 1 所示。

对于 SOPT5 影像,2 景影像同处胶东半岛,所包含的主要地物类型相似(包括木本植被、草本植被、水体、人工建筑、沙滩和裸地),虽不是同一年获取的影像,但获取的季节相近,影像中地物所处辐射环境也较为相似;对于 QuickBird 影像,2 景影像虽获取时间相差半年,但由于同处热带,季节对地物的影

响较小,且 2 岛距离较近,岛体底质、礁盘和岛上地物覆盖类型均较相似(包括木本植被、草本植被、水体、珊瑚沙滩和人工建筑等 5 种类型)。

表 1 SPOT5 和 QuickBird 影像信息

Tab.1 Information of the SPOT5 and QuickBird images				
影像类型	获取日期	覆盖区域	空间分辨率/m	主要地物类型
SPOT5	2003-11-03	青岛鳌山卫	10.0	木本植被、草本植被、水体、人工建筑、沙滩、裸地
	2004-11-16	烟台栾家口		
QuickBird	2006-06-06	三沙永兴岛	2.4	木本植被、草本植被、水体、珊瑚沙滩、人工建筑
	2005-11-04	三沙东岛		

1.1.2 实验设计

分类实验中,分别选择影像 A 和影像 C 为替代训练样本来源影像;影像 B 和影像 D 为待分类影像。根据 SPOT5 和 QuickBird 影像不同空间分辨率的应用特点,拟对 SPOT5 影像采用基于像元的分类策略,对 QuickBird 影像采用基于分割对象的分类策略。相应的分类实验设计如下:

1)在基于像元的分类实验中,从 SPOT5 影像 A 中采用监督方法独立选择共 3 组训练样本集(分别为 24,48 和 96 个样本),样本数量在每个类别中平均分布;从影像 B 中以随机方式独立选择 4 组未标记样本集(分别为 100,200,500 和 1 000 个样本);训练样本集和未标记样本集逐一搭配进行分类实验(共 12 种搭配结果)。采用这种组合实验方式的目的是观察在遥感影像训练样本拓展应用中,训练样本和未标记样本之间的样本数量比例是否与分类精度存在关系。若有关系,如何配置 2 种样本的比例使分类结果达到最优,并探究随样本数量的增加和

比例的变化对分类结果的影响。

2)在基于分割对象的分类实验中,首先对 QuickBird 影像 C 和 D 进行图像分割处理,采用同样的分割尺度和参数,要求每景影像中分割斑块数量尽量少,且每个斑块中只能包含一种地物。由于分割结果影像中斑块的数量较为有限,故分类时将待分类影像中所有的分割斑块都作为未标记样本参与半监督学习;待分类器优化过程结束,分类即完成。而对于训练样本,分别独立选择 20,40 和 60 个样本,样本数量在每个类别中平均分布。

3)训练样本的选择和最终分类结果的精度验证依据为:①“我国近海海洋综合调查与评价专项”对上述 2 个区域的人机交互解译的最终提取成果;②分别于 2006 年和 2008 年开展的 2 次现场踏勘所获取的数据资料。

1.2 方法

本文提出的基于半监督学习的遥感影像训练样本时空拓展方法的技术流程如图 2 所示。

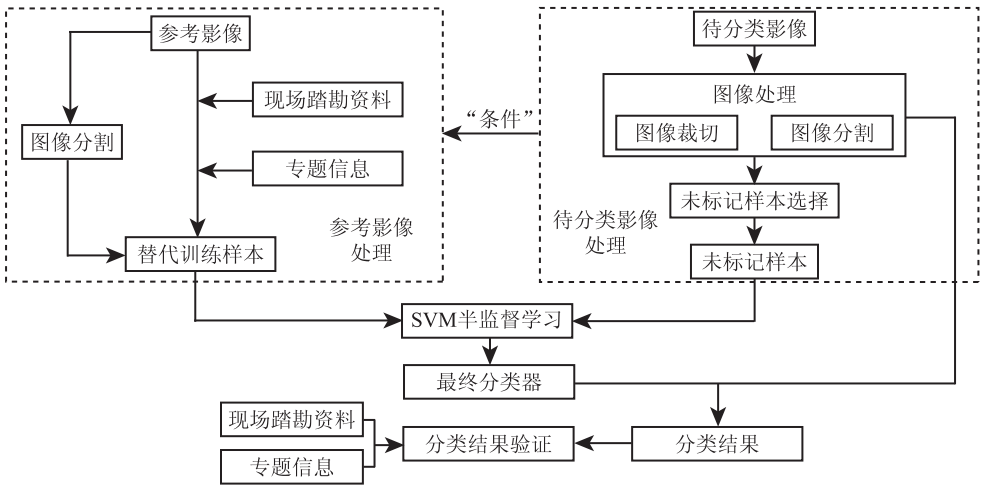


图 2 基于半监督学习的遥感影像分类样本拓展方法技术流程

Fig.2 Flow chat of the method of remote sensing images training sample extending based on semi-supervised learning

本文方法与一般遥感影像半监督学习分类方法的区别:①替代训练样本和未标记样本分别来源于不同的影像;②在基于分割对象的分类中,将待分类影像中所有的分割斑块作为未标记样本参与分类器的半监督学习与优化。图 2 的“条件”指分类中涉及的 2 景影像需要有相似的地物类型分布和相近的获取时相,以保证用于拓展分类的训练样本对待分类影像地物类型的代表性不会有太大偏差,使半监督学习过程中分类面能向正确的方向优化。

1.2.1 SVM 分类器半监督学习

当训练样本对待分类影像的代表性较差时,根据其结构风险最小化的原理,支持向量机(support

vector machine,SVM)优秀的分类和泛化能力反而成为其最大的弱点。针对这一问题,机器学习领域提出了半监督学习的 SVM 分类方法,即 TSVM 方法。该方法通过寻找参与分类器拟合的大量未标记样本所表现出的地物类型在特征空间所固有的分布和结构,使分类器在待分类影像中获得较好的分类泛化能力。

以 2 类分类问题为例,样本集 $X = \{x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+n}\}$ 的前 l 个样本为训练样本,对应 $Y = \{y_1, y_2, \dots, y_l\} \subset \{-1, +1\}$; 后 n 个样本为未标记样本,要求拟合一个分类超平面 $f(x_i) = wx_i + b = 0$, 对属于“+1”类的样本 $f(x_i) > 0$, 反之 $f(x_i) < 0$; w 和 b 分别为特征空间中分类超平面的法向量和偏移

量。对于软分类(训练过程中允许有少量的训练样本被错分)问题的最优分类面拟合,实际上是求解以下优化问题,使 $\min(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i + C^* \sum_{j=1}^n \zeta_j)$ 服从

$$\begin{aligned} y_i[wK(x_i) + b] &\geq 1 - \xi_i \quad (i=1, 2, \dots, l), \\ y_j[wK(x_j) + b] &\geq 1 - \zeta_j \quad (j=1, 2, \dots, n), \end{aligned} \quad (1)$$

式中: $K(x)$ 为将高维特征空间映射到低维空间的核函数; C 和 C^* 分别为训练样本和未标记样本的惩罚因子,描述在多大程度上可以容忍某些样本被错分,惩罚因子越大,越不允许错分; ξ 为松弛变量($\xi_i \geq 0, \zeta_j \geq 0$),表示被错分的样本点 x 的错分程度, $1 - \xi$ 为 x 点到分类面的距离。

针对上述分类面拟合最优化问题,Joachims^[23]给出了迄今应用最广的训练算法,但该算法需要事先确定属于某个类别未标记样本的具体数目,对于遥感影像中的地物分类(特别是进行样本拓展应用的分类问题),事先确定某个类别样本的具体数目或者比例是不可能的,一旦错误估计,对分类结果的影响将是灾难性的。陈毅松^[24-26]等对分类器训练中未标记样本的标注过程进行了改进,但若应用到遥感图像分类中,还存在重要问题,即对未标记样本的标注过程未考虑到各个类别在特征空间中的实际分布。上述工作或直接对等地标注一对样本,或按照某个固定阈值标注一个区域的样本,或考虑到阈值的差别而对区域标注阈值进行人为估测,但都无法适应遥感图像分类时的数据量大、类别分布不均衡和先验知识难以获取等情况,因而需要对训练学习算法进行改进。

1.2.2 改进的 TSVM 分类器半监督学习

面向分类样本拓展应用的遥感图像分类问题和特点,针对目前半监督学习算法面临的问题,提出一种改进的 TSVM 算法。

对于有 M 个类别的遥感影像样本时空拓展分类问题,算法流程如下:

假设影像 A 和影像 B 具有相似的地物类型和获取时相,将影像 A 作为训练样本的来源影像,影像 B 作为待分类影像。步骤如下:

1) 从影像 A 中监督选择 K 个训练样本,组成训练样本集 L ; 从影像 B 中非监督选择 N 个未标记样本,组成未标记样本集 U (其中, $K < N$)。

2) 给出惩罚因子 C , 应用样本集 L 训练初始的 SVM 分类器,用 $f^{(0)}[SVM(x)]$ 表示; 考虑第 t 次递归训练($t \geq 1$), 分类器用 $f^{(t-1)}[SVM(x)]$ 表示。

3) 用训练的分类器 $f^{(t-1)}[SVM(x)]$ 对未标记

样本集 $U^{(t-1)}$ 中的样本进行类别预测,得到属于每个类别的样本数目 $N_i (i=1, 2, \dots, M)$ 。设置判别函数阈值 $\sigma_i = 1 - \lambda N_i / N$ (λ 为调节参数,默认为 1), 对于分类问题中有特别倾向的,可将之变小。对于类别 i , 将满足 $||wK(x_i) + b| - 1| \leq \sigma_i$ 并可预测出类别的样本加入到训练样本集 L 中,将该部分样本从未标记样本集 U 中移除,新样本集分别用 $L^{(t)}$ 和 $U^{(t)}$ 表示。应用新样本集 $L^{(t)}$ 重新输入训练分类器,得到 $f^{(t)}[SVM(x)]$ 。

4) 当满足 $||wK(x_i) + b| - 1| \leq \sigma_i$ 的样本不存在时,停止递归,输出分类函数 $f^{(t)}[SVM(x)]$; 否则,返回步骤 3 继续训练。

与传统算法相比,本文提出的算法有 4 个特点: ①采用效率更高的区域标注法,即每次递归将满足阈值条件的未标记样本一次性标注; ②对未标记样本的标注过程考虑到了各类别在特征空间中的分布特征,即对每个类别都设置了与其先验概率估计值 N_i / N 相关的标注判别阈值,随着被标记的未标记样本数的增加, N_i / N 越接近于真实的先验概率; 但初始判别阈值因无法准确估计其先验概率而都设置为一个相同较小的值,因为对于每个类别初始的已标记样本数量 N_i 都是一样且较小的; ③标注过程对处于支持向量面 2 侧的未标记点都进行考虑,包括因已标记样本代表性问题导致的未标记样本点都落在初始分类面和初始支持向量面之间的情况; ④预留对有倾向类别的调节参数,对遥感图像分类中常见的、在图像中处于像元数量弱势但极其重要的类别,可适当调节 λ 来增加对其的关注度。

对基于支持向量机的半监督学习算法中区域标注法的收敛性已在陈毅松等^[24]的研究中给出了证明,不再赘述。

2 结果与分析

2.1 基于像元的样本拓展分类结果

根据实验设计,分别从 SPOT5 影像 A 中独立地监督选择 24, 48 和 96 个训练样本。由于待分类影像 B 中地物类型的先验分布未知,故样本在木本植被、草本植被、水体、人工建筑、沙滩和裸地等 6 种地物类型中平均分布。同时,从待分类影像中独立地随机选取 100, 200, 500 和 1 000 个未标记样本参与半监督学习。初始惩罚因子 C 的取值通过实现对已标记样本的最优分类来进行估计。

在相同样本组合下,分别进行了 SVM 监督分类、传统 TSVM 分类和本文算法的半监督学习分类(图 3),其分类精度见表 2。

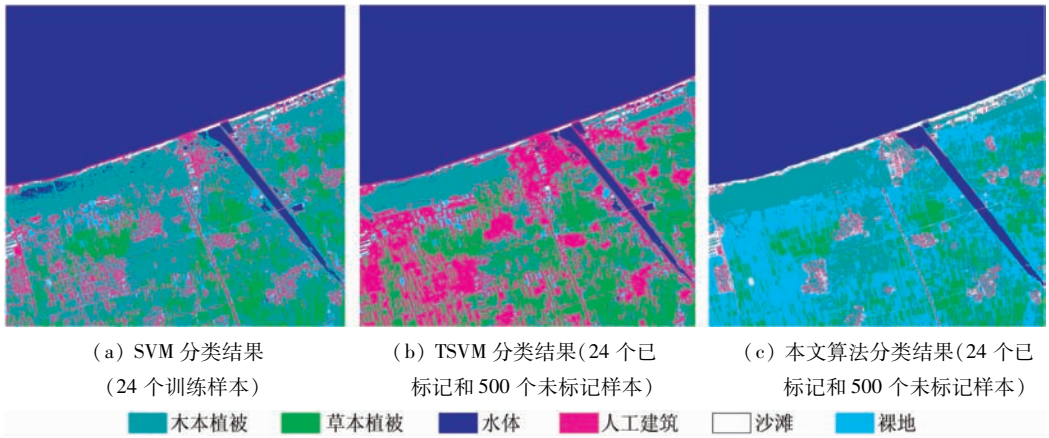


图 3 从影像 A 中选择训练样本在影像 B 中拓展应用分类结果对比

Fig. 3 Comparison of different classification results of image B with the training samples selected from image A

表 2 基于像元的不同分类结果精度比较

Tab. 2 Comparison of the accuracies of different classification results based on pixel							
训练样本数	未标记样本数	SVM		TSVM		本文算法	
		总体精度/%	Kappa	总体精度/%	Kappa	总体精度/%	Kappa
24	100	44.1	0.368	60.2	0.556	70.6	0.668
	200	44.1	0.368	65.4	0.581	77.9	0.753
	500	44.1	0.368	72.3	0.681	85.6	0.811
	1 000	44.1	0.368	73.5	0.678	82.1	0.780
48	100	42.2	0.327	52.2	0.483	63.3	0.573
	200	42.2	0.327	60.3	0.555	73.3	0.681
	500	42.2	0.327	70.1	0.654	81.0	0.772
	1 000	42.2	0.327	70.6	0.667	81.2	0.776
96	100	49.7	0.454	55.3	0.467	61.2	0.554
	200	49.7	0.454	60.5	0.539	68.6	0.608
	500	49.7	0.454	73.1	0.665	79.9	0.703
	1 000	49.7	0.454	75.4	0.676	81.2	0.745

分析图 3 和表 2 不难发现：

1) 半监督学习方法可有效完成训练样本的时空拓展应用。实验结果显示,在 24 个训练样本和 500 个未标记样本的组合试验中,与样本直接拓展应用的 SVM 监督分类方法相比,本文提出的方法使分类精度提高了约 41 个百分点,达到 85.6% ; 并取得了比传统 TSVM 方法更高的分类精度,且在每一个分类组合中都提高了约 10 个百分点。

本文提出的方法获得了比其他 2 种方法更好的分类效果。从原理上分析,基于统计学习理论的 SVM 方法在处理小样本问题上具有优势,但在处理样本有偏分布问题时,其优秀的泛化能力反而变成了弱点(如图 3(a)中由于裸地类型在该区域的光谱特征较为相似,导致该类型样本在特征空间中分布集中;而木本植被和人工建筑样本特征在特征空间中分布相对分散,致使最终分类结果中木本植被和人工建筑的过度分类);传统 TSVM 方法要求估计各类别在待分类影像中的比重,对于本文所面临的未知区域分类问题,估计先验比重将面临较大经验风险(如图 3(b)中因对裸地所占比重的估计过小,导致人工建筑和裸地都出现了较严重的分类错误);而本文方法根据未标记样本在特征空间中聚类的特点,在每一次递归中,渐进式地计算和调整分类面的变化方向和幅度。分类结果证实本文方法可有效完成分类训练样本的拓展应用(图 3(c))。

2) 当进行拓展应用的训练样本对待分类图像地物类型的空间特征结构描述不好时,过多使用训练样本会对分类结果产生负作用(如表 2 中,对于相同的未标记样本,24 个训练样本组合的分类精度要高于 48 和 96 个的分类精度)。

在半监督学习中,当未标记样本在特征空间中表现出待分类地物类型固有特征中心和训练样本的位置有明显差异时,少量具有较大偏差分布的训练样本集中的样本,可以被设置的惩罚因子 C 所容忍和忽略,而此时训练样本的作用是为 SVM 分类提供一个初始的有导向性作用的分类面;但当这样的训练样本增多到惩罚因子无法容忍时,其必将给分类结果造成影响。

3) 当训练样本的数目一定时,未标记样本应用得多并不一定都可提高分类精度。从表 2 看出,当

分类训练样本为 24 个时,未标记样本从 500 增加到 1 000 个,反而导致了分类精度的下降。

对式(1)分析表明,随着未标记样本数目的增多,特别是相对于训练样本占绝对优势时,目标函数值的走向在较大程度上受到未标记样本的支配,训练样本便失去了对其进行类别引导的作用。本文的分类实验结果显示,当未标记样本和训练样本数量差距保持在 10 ~ 20 个时,可取得较好的分类效果(如实验结果在 24,48 和 96 个已标记样本情况下,其最高分类精度出现在与 500 和 1 000 个未标记样本搭配分类的时候),这与任广波等^[19]在基于生成模型的半监督学习分类中得到的结论一致。

在分类效率方面,通过 1.2 节的方法描述,改进的 TSVM 方法在算法复杂度上与传统的 TSVM 是相同的,但由于可自适应地调整分类面变化的方向和幅度,相比每次递归只对固定数目的样本进行标注,

可有效地提高效率(在本文遥感图像分类实验中,递归 8 次以内便可达到收敛,而 TSVM 方法则平均需要 50 次以上)。

2.2 基于对象的样本拓展分类结果

不同于基于像元的分类,针对高分辨率遥感影像(如 QuickBird),基于分割斑块的分类已成为研究的热点之一。首先对影像 C 和影像 D 进行图像分割,由于 2 景影像分别覆盖的永兴岛和东岛上的地物类型、尺度等都较为相似,因此用相同的分割参数(分割尺度参数 40,形状参数 0.6,紧致度参数 0.5)进行分割,获得 2 232 个分割斑块(图 4(a))。

按照实验设计,以分割斑块为基本单元,选择分割后影像每个斑块分别在 4 个波段的像点均值为特征,从影像 C 中独立选择 3 组训练样本(分别有 20, 40 和 60 个样本),对影像 D 进行分类(图 4(b)(c))。

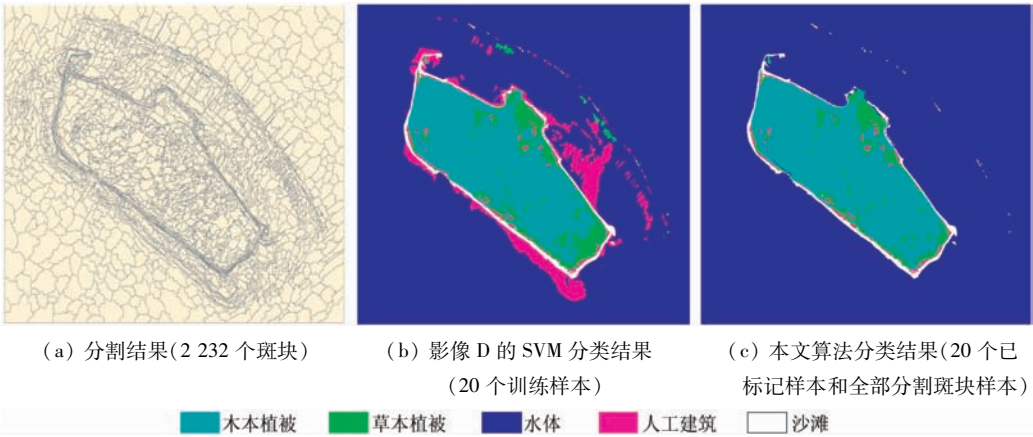


图 4 从影像 C 中选择的分类样本在影像 D 中拓展应用分类结果

Fig. 4 Comparison of different classification results of image D with training sample selected from image C

由于在基于像元的分类中已经证实本文提出的方法较传统的 TSVM 方法更有效,故本实验仅就使用本文提出的半监督学习方法前后的效果进行分析,分类精度见表 3。

表 3 基于分割对象的不同分类结果精度比较
Tab. 3 Comparison of the accuracies of different classification results based on object

训练样 本数	未标记样本	SVM		本文算法	
		总体精度/%	Kappa	总体精度/%	Kappa
20	全部分割斑块	65.3	0.577	86.4	0.827
40	全部分割斑块	62.5	0.552	86.1	0.796
60	全部分割斑块	58.9	0.502	78.9	0.689

基于分割对象的分类结果和基于像元的分类结果表现出相似的特点,不再重复论述。不同的是,基于分割对象的分类中将待分类影像中所有的分割斑块都作为未标记样本参与半监督学习,分类工作随着半监督学习过程的结束而完成。这在机器学习理

论中是真正意义上的“直推式学习”。同时,若将对影像的分割作为影像预处理而不作为分类过程的工作,则可获得比基于像元分类方法更好的分类效率。

另外,在最终分类结果中,仍有部分斑块被错误分类(如东岛东北部的浪花被错分为沙滩,环绕岛陆的部分沙滩被错分为人工建筑),这是由于地物的同谱异物问题造成的。可通过利用分割对象特有的形状参数和地物位置关系等信息对分类结果进行二次判定,以获得更高精度的分类结果。

3 讨论

1) 训练样本的拓展应用有望成为实现遥感图像自动化分类的有效途径。因本文所涉及分类问题是在无法从待分类图像中获取训练样本时的分类样本拓展应用问题,训练样本的获取源有着较大的不确定性;或者说只要保证训练样本对待分类图像

有一定的代表性,样本取自何处便不再重要。如此,不妨设想可建立一个较为全面的训练样本集,在涉及样本拓展应用分类问题时,根据待分类图像的类型、区域、待分类类别等条件,从样本集中自动选择训练样本、开展分类。这将很大程度地提高分类过程的自动化程度。

2) 基于分割对象的半监督学习应用潜力巨大。分割对象带来的一系列基于像元分类所不具备的特征维度(如形状指数和空间语义关系等),将有望帮助我们开辟一条更高效、高精度、高自动化处理无样本或样本代表性不好的遥感图像分类的途径。

3) 训练样本对待分类影像具有一定的代表性,是半监督学习方法取得较好分类效果的前提,但实际应用中这种代表性却不容易准确地控制。因此如何使所选取的训练样本对待分类影像的“代表性”足以让半监督学习取得好效果,是一个值得深入研究的问题。

4 结论

1) 半监督学习方法有效利用了未标记样本所表现出的地物类别在特征空间中的固有结构信息,成功实现了遥感影像训练样本的拓展应用;所提出的半监督学习算法能更好地解决样本拓展应用中待分类影像地物类别比例无法准确估计的问题,取得比传统 TSVM 方法更好的分类效果。

2) 过多地使用训练样本和未标记样本都可能会对训练样本拓展应用效果产生不利影响,其原因是由于半监督学习算法的限制,过多地使用上述 2 类样本都可能会对分类过程产生误导。

3) 基于分割对象的半监督学习分类方法是真正意义上的直推式学习方法,可获得比基于像元的分类方法更高的分类效率。

参考文献(References):

- [1] Jackson Q, Landgrebe D A. An adaptive classifier design for high - dimensional data analysis with a limited training data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2001, 39(12): 2664 - 2679.
- [2] Jensen J R. Introductory digital image processing: A remote sensing perspective[M]. 3rd ed. New Jersey: Prentice Hall, 2006: 195 - 209.
- [3] Kaufman Y J, Fraster R S. Atmospheric effect on classification of finite fields[J]. Remote Sensing of Environment, 1984, 15(2): 95 - 118.
- [4] Cracknell A P, Hayes L W. Atmospheric corrections to passive satellite remote sensing data[M]//Cracknell A P, Hayes L W B. Chapter 8 in Introduction to Remote Sensing. London: Taylor and Francis, 1993: 116 - 158.
- [5] Camps V G, Bandos T, Zhou D Y. Semi - supervised graph based hyperspectral image classification[J]. IEEE Transaction on Geoscience and Remote Sensing, 2007, 45(10): 3044 - 3054.
- [6] Du Y, Teillet P M, Cihlar J. Radiometric normalization of multitemporal high - resolution satellite images with quality control for land cover change detection[J]. Remote Sensing of Environment, 2002, 82(1): 123 - 134.
- [7] 张友水, 冯学智, 周成虎. 多时相 TM 影像相对辐射校正研究[J]. 测绘学报, 2006, 35(2): 122 - 127.
Zhang Y S, Feng X Z, Zhou C H. Relative radiometric correction for multitemporal TM images[J]. Acta Geodaetica et Cartographica Sinica, 2006, 35(2): 122 - 127.
- [8] Koukal T, Suppan F, Schneider W. The impact of relative radiometric calibration on the accuracy of KNN - predictions of forest attributes[J]. Remote Sensing of Environment, 2007, 110(4): 431 - 437.
- [9] Ren G B, Zhang J, Ma Y, et al. A method for classification training sample spatial - time expanding of remote sensing images[C]//International Conference on Space Information Technology, Beijing, Proceedings of the SPIE, 2009: 76510G1 - G7.
- [10] Scudder I I. Probability of error of some adaptive pattern - recognition machines [J]. IEEE Transactions on Information Theory, 1965, 11(3): 363 - 371.
- [11] Fralick S. Learning to recognize patterns without a teacher[J]. IEEE Transactions on Information Theory, 1967, 13(1): 57 - 64.
- [12] Nigam K, McCallum A, Mitchell T. Semi - supervised text classification using EM semi - supervised learning[M]. Cambridge MA: MIT Press, 2006: 33 - 55.
- [13] István N T, Richárd F, János C. On positive and unlabeled learning for text classification[C]//István N T, Richárd F, János C. Lecture Notes in Computer Science. London: Springer, 2011, 6836: 219 - 226.
- [14] Schenker A, Bunke H, Last M, et al. A graph - based framework for web document mining[M]//Schenker A, Bunke H, Last M, et al. Document Analysis Systems. London: Springer, 2004: 401 - 412.
- [15] Yang Y, Wu F, Nie F, et al. Web and personal image annotation by mining label correlation with relaxed visual graph embedding image processing[J]. IEEE Transactions, 2012, 21(3): 1339 - 1351.
- [16] Shahshahani B M, Landgrebe D A. 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 - 1092.
- [17] 骆剑承, 王钦敏, 马江洪, 等. 遥感图像最大似然分类方法的 EM 改进算法[J]. 测绘学报, 2002, 31(3): 234 - 239.
Luo J C, Wang Q M, Ma J H, et al. The EM - based maximum likelihood classifier for remotely sensed data[J]. Acta Geodaetica et Cartographica Sinica, 2002, 31(3): 234 - 239.
- [18] Tuia D, Camps V G. Semi - supervised remote sensing image classification with cluster kernels[J]. IEEE Geoscience and Remote Sensing Letters, 2009, 6(2): 224 - 228.
- [19] 任广波, 张杰, 马毅, 等. 生成模型学习的遥感影像半监督分类[J]. 遥感学报, 2010, 14(6): 1090 - 1104.
Ren G B, Zhang J, Ma Y, et al. Generative model based semi - supervised learning method of remote sensing image classification

- [J]. Journal of Remote Sensing, 2010, 14(6): 1090 – 1104.
- [20] Tuia D, Pasolli E, Emery W J. Using active learning to adapt remote sensing image classifiers [J]. Remote Sensing of Environment, 2011, 115(9): 2232 – 2242.
- [21] Galante N, Siqueia R, Sant' Anna S J, et al. Semi – supervised remote sensing image classification methods assessment [C]//IGARSS. Geoscience and Remote Sensing Symposium. Vancouver: IEEE International, 2011: 2939 – 2942.
- [22] Mishra N S, Ghosh S, Ghosh A. Semi – supervised fuzzy clustering algorithms for change detection in remote sensing images [C]//Mishra N S, Ghosh S, Ghosh A. Lecture Notes in Computer Science. London: Springer, 2012, 7143: 269 – 276.
- [23] Joachims T. Transductive inference for text classification using support vector machines [C]//Proceedings of the 16th International Conference on Machine Learning (ICML). San Francisco: Morgan Kaufmann Publishers, 1999: 200 – 209.
- [24] 陈毅松, 汪国平, 董士海. 基于支持向量机的渐进直推式分类学习算法 [J]. 软件学报, 2003, 14(3): 451 – 460.
- Chen Y S, Wang G P, Dong S H. A progressive transductive inference algorithm based on support vector machine [J]. Journal of Software, 2003, 14(3): 451 – 460.
- [25] 沈新宇, 许宏丽, 官腾飞. 基于直推式支持向量机的图像分类算法 [J]. 计算机应用, 2007, 27(6): 1463 – 1464.
- Shen X Y, Xu H L, Guan T F. Image classification based on transductive support vector machines [J]. Computer Application, 2007, 27(6): 1463 – 1464.
- [26] 廖东平, 魏玺章, 黎 湘, 等. 一种改进的渐进直推式支持向量机分类学习算法 [J]. 信号处理, 2008, 24(2): 213 – 218.
- Liao D P, Wei X Z, Li X, et al. An improved learning algorithm with progressive transductive support vector machine [J]. Signal Processing, 2008, 24(2): 213 – 218.

Extending method of remote sensing image training sample based on semi – supervised learning in both time and spatial domain

REN Guangbo, ZHANG Jie, MA Yi, SONG Pingjian

(First Institute of Oceanography, SOA, Qingdao 266061, China)

Abstract: In classification of remote sensing images without any training samples, the choice of training samples from other representative images might be the only direct way; nevertheless, due to the difference of radiometric environments, the classification training samples from one image could not be well representative of other images. It is known that labeled samples from one image may not be effective for classifying others with high accuracy. In view of the above problem, a novel semi – supervised transductive support vector machine (TSVM) method is proposed. The authors first chose a large quantities of unlabeled samples from the images which need to be classified in an unsupervised way, then extracted the inherent construction information of different classes in the feature space. Next, with the help of semi – supervised learning theory, the authors trained a classifier which was pre – trained by the labeled samples from another image in a recursive way, and at last an optimized classifier was obtained. It should be noted that two images involved in the method must have familiar land covers and acquired times. Classification experiments of SPOT5 and QuickBird remote sensing images were undertaken by the authors, and the classification results prove that the method proposed in this paper can effectively realize the sample extending application in both time and spatial domain.

Key words: remote sensing classification; semi – supervised learning; transductive support vector machine (TSVM); sample extending application

第一作者简介: 任广波 (1983 –), 男, 助理研究员, 主要研究方向为海岛海岸带遥感。E – mail: renguangbo@yahoo. com. cn。

(责任编辑: 邢 宇)