基于 a-BvSBEM 主动学习的高光谱图像分类

张 琳
(河海大学 南京 210000)

摘 要:在高光谱遥感图像分类中,需要大量的训练样本对分类器进行训练,然而对样本标记非常困难并且耗时、昂贵。针对样本标记困难的问题,提出了自适应的样本不确定性与代表性相结合的主动学习选择训练样本。样本的不确定性是利用最优标号与次优标号(best vs second-best,BvSB)的方法计算。用期望最大(expectation maximization,EM)聚类计算样本的代表性。然后将样本的不确定性与代表性通过自适应权重相结合,从而选出含信息量最大的未标注样本加入进行人工标注,并加入到训练样本。通过实验表明,此方法性能更加稳定,准确率也有一定的提高。

关键词:主动学习;样本不确定性与代表性;期望最大聚类;自适应;高光谱图像分类
中图分类号: P407.8 TN957.52 **文献标识码**:A **国家标准学科分类代码**: 520.6040

Hyperspectral image classification based on best vs second-Best active learning and expectation maximization cluster

Zhang Lin
(Hohai University, Nanjing 210000, China)

Abstract: In hyperspectral remote sensing image classification, needs a large number of training samples to trainclassifier, but labeled sanmpes is very difficult ,time-consuming and expensive. Therefor, we proposed aadaptive method combined representative samples with uncertainty samples to select samples. We use the active learning based on the best vs second-best(BvSB) for selecting training samples and take advantage of the expectation maximum (Expectation Maximization, EM) cluster to computerepresentativeness. Then uncertainty and representative of the samples combined with aadaptive weight to select most informative unlabeled samples for manual labeling, and join the training set for training classifier. Experiments show that our method is more stable performance and accuracy is also improved.

Keywords: active learning; representative and uncertainty of samples; expectation maximization; adaptive; hyperspectral image classification

1 引 言

近年来,高光谱图像分类成为研究的热点,这是由于高光谱图像含有丰富的光谱信息,能够更加准确的识别和区分地物。但在对高光谱遥感图像进行分类时,随着高光谱图像波段的增加,对训练样本数量的要求也急剧增加,而训练样本的数目非常有限,并不能满足不断增加的光谱带,这会导致"Hughes"现象。为解决分类过程中出现的"Hughes"现象,核方法被提出[1],由于核对维数不敏感,使它在高光谱图像分类中被广泛应用。将核函数应用于支持向量机(support vector machine, SVM)[2-6],可减少输入数据的维数,从而提高 SVM 的分类效果,传统的 SVM 分类器在高光谱图像分类中表现出了非常显著的效果。但传统的方法在选择训练样本时是随机选取的,这样选取的样本含有的信息量不够大,对分类器的训练不能达到很好的效果。近年来,用主动学习迭代地选择未标注样本进行标注并加入到训练样本集,这样避免了对信息量少的样本进行标注,从而节省了时间和人力[7]。在文献[7]中提出的基于最优标号和次优标号(best vs second-best, BvSB)主动学习能够适用于多分类的问题,并且选出的样本具有不确定性,包含更多的信息量,这样选出的样本能够训练出良好的分类器。但是 BvSB 选择出的样本仅考虑了不确定性,而忽略了样本的代表性。因此本文提出了一种自适应的不确定性与代表性相结合的主动学习方法来选择样本(a-BvSBEM)。

收稿日期:2017-01

2 基于 a-BvSBEM 的主动学习算法

2.1 BvSB 样本选择

在进行分类时，越靠近分类面的样本越具有不确定性，含有的信息量也越大，这样的样本作为训练样本，能够训练出更加准确的分类器。

设未标记样本集 $U=\{x_1,\cdots,x_n\}$，$Y=\{1,2,\cdots\}$ 为所有可能的类标号，未标记样本 x_i 属于各个类别的概率为 $P(y_i\mid x_i)$，基于 BvSB 的主动学习方法是找出未标记样本 x_i 的最优标号和次优标号。将 x_i 属于各个类别的最大概率 $P(y_{\text{Best}}\mid x_i)$ 的标号记为最优标号 y_{Best}，将 x_i 属于各个类别的次大概率 $P(y_{\text{Second-Best}}\mid x_i)$ 的标号记为次优标号 $y_{\text{Second-Best}}$。未标记样本 x_i 属于 y_{Best} 和 $y_{\text{Second-Best}}$ 的不确定性可以用 $P(y_{\text{Best}}\mid x_i)-P(y_{\text{Second-Best}}\mid x_i)$ 表示。当 $P(y_{\text{Best}}\mid x_i)-P(y_{\text{Second-Best}}\mid x_i)$ 取最小值时，x_i 属于 y_{Best} 和 $y_{\text{Second-Best}}$ 的不确定性越大，也就是 x_i 越靠近分类面，含有的信息量更适合作为训练样本。

这样 BvSB 准则[8]可以写为：

$$BvSB=\underset{x_i\subset U}{\operatorname{argmin}}(P(y_{\text{Best}}\mid x_i)-P(y_{\text{Second-Best}}\mid x_i)) \tag{1}$$

2.2 基于 BvSB 主动学习过程

1)将未标注样本加入到分类器中进行训练，并预测出未标记样本 x_i 属于各个类别的最大概率 $P(y_{\text{Best}}\mid x_i)$ 和次大概率 $P(y_{\text{Second-Best}}\mid x_i)$；

2)计算 BvSB 的值，并将其按照从小到大排列，每次迭代从中选出使 BvSB 最小的前 m 个样本记为未标记样本集 k_{mBvSB}，对 k_{mBvSB} 进行人工标注，并将其添加到训练样本中；

3)更新训练样本；

4)对分类器进行训练；

5)返回到 1)中进行下一次迭代，直到选出适当的样本。

2.3 EM 聚类

为了利用 EM 算法聚类高光谱图像[9-11]，假设从一个多元高斯分布中提取出的像素属于同一个聚类。这样，每个图像像素可以利用概率密度函数建模：

$$p(x)=\sum_{c=1}^{C}\omega_c\Phi_c(x;\mu_c,\Sigma_c) \tag{2}$$

式中：$\omega_c\in[0,1]$ 是聚类 c 的混合权重 $\sum_{c=1}^{C}\omega_c=1$；$\Phi(\mu,\Sigma)$ 是多元高斯密度，μ,Σ 分别为均值和协方差矩阵：

$$\Phi_c(x;\mu_c,\Sigma_c)=\frac{1}{(2\pi)^{\frac{B}{2}}}\frac{1}{|\Sigma_c|^{\frac{1}{2}}}\times$$

$$\exp\left\{-\frac{1}{2}(x-\mu_c)^{\mathrm{T}}\Sigma_c^{-1}(x-\mu_c)\right\} \tag{3}$$

参数 $\psi=\{C,\omega_c,\mu_c;c=1,2,\cdots,C\}$ 是通过迭代的方法估计的。EM 聚类算法概述如下：

令 $X=\{x_1,x_2,\cdots,x_n\}\in R^B$ 表示表示高光谱图像的 B 维光谱特征向量。$C_{\max}$ 表示聚类的最大个数。

1)初始化(迭代次数为 0)

(1)令 $C=C_{\max}$，确定第一次将 X 划分为 C 个聚类，表示为 Q_c^0，$c=1,2,\cdots,C$；从 X 中随机选择 C 个点作为聚类中心；

(2)根据欧氏距离最近，将剩余的点分别划分给 C 个聚类中心。

2)对于每次迭代 ite>0

(1)参数估计过程：$\mu_c^i,\Sigma_c^i,\omega_c^i$ 用最大似然估计来求，公式如下：

$$\mu_c^i=\frac{1}{m_c^{i-1}}\sum_{j=1}^{m_c^{i-1}}x_{j,c}^{i-1} \tag{4}$$

$$\Sigma_c^i=\frac{1}{m_c^{i-1}}\sum_{j=1}^{m_c^{i-1}}(x_{j,c}^{i-1}-\mu_c^i)(x_{j,c}^{i-1}-\mu_c^i)^{\mathrm{T}} \tag{5}$$

$$\omega_c^i=\frac{m_c^{i-1}}{n} \tag{6}$$

(2)聚类过程：

①根据最大后验概率准则计算给把每个特征向量划分给一个聚类，即对于 $xj\in Q_c^i$ 有：

$$\Pr(c\mid x_j)=\max_l\Pr(l\mid x_j) \tag{7}$$

$$\Pr(c\mid x_j)=\frac{\omega_c^i\Phi_c(x_j;\mu_c^i,\Sigma_c^i)}{\sum_{c=1}^{C}\omega_c^i\Phi_c(x_j;\mu_c^i,\Sigma_c^i)} \tag{8}$$

②对于估计聚类 $c(c=1,2,\cdots,C)$ 如果 m_c^i 小于特征向量的维数，将会将这个特征向量从这个聚类中移除，并分配给下一个聚类；

③如果不满足收敛准则，则返回参数估计过程，若满足收敛准则，则停止。

2.4 自适应过程

要选择的样本需要既具有较大不确定性又具有很好的代表性，因此用权重将样本的不确定性度量准则与代表性度量准则相结合，如下式：

$$F_a(x_i)=BvSB\,(x_i)^{\alpha}+d\,(x_i)^{\alpha} \tag{9}$$

式中：d 表示各个样本距离聚类中心的欧氏距离，$\alpha\in W=[0.1,0.2,\cdots,0.9,1]$ 是一个权重参数，表示样本不确定性和代表性的贡献程度。当 $\alpha>0.5$ 表示不确定性准则占主导地位，$\alpha<0.5$ 表示代表性占主导地位，$\alpha=0.5$ 时，表明代表性与不确定贡献程度相同。由于在主动学习和迭代过程中，两者的重要性在不断地变化，因此预先定义一个权重系数集合 $W=[0.1,0.2,\cdots,0.9,1]$，对于每个不同的 $\alpha\in W$，用式(9)计算出既具有不确定性又具有代表性的样本。

2.5 基于 a-BvSBEM 的主动学习算法

1)根据式(1)计算出 BvSB；

2)根据 EM 聚类算法计算出各个样本距离聚类中心的欧氏距离 d；根据式(9)计算出 $F_a(\mathrm{x_i})$；

3)从 $F_a(x_i)$ 中选出前 m 个样本，进行人工标注，并加入到训练样本中；

4)更新训练样本集,并对分类器进行训练;

5)返回1),直到迭代结束。

3 实验结果与分析

本次实验是在MATLAB R2013a平台下进行的,实验从总精度OA(overall accurary)、平均分类精度AA(average accuracy)和Kappa系数方面进行分析。OA、AA和Kappa值均采用10次实验结果的平均值。

3.1 实验数据

数据1:Indian Pines数据,是AVIRIS传感器在美国印第安纳州西北部的Indian Pines上空拍摄获取。AVIRIS传感器在0.2~2.3 μm的光谱范围内形成220个波段。在实验中将20个噪声波段移除,剩余的200个光谱波段用于实验。Indian Pines图像的光谱分辨率10 nm、空间分辨率分别为20 m、空间维度为145×145。真实的地物包含16类,共有10 366个标记像素。

数据2:University of Pavia数据,为ROSIS光学传感器在意大利南部的Pavia市的Pavia University上空拍摄的,数据大小为610×340,具有很高的空间分辨率(1.3 m/pixel)。除去噪声波段,图像包含有103个连续波段,该地区共包含9种地物,共42 776个样本。

3.2 实验数据设置

本文的核函数采用的是径向基核函数,期中g为0.04,惩罚函数C为100 000。

利用Indian Pines数据进行实验时,$C_{max}=17$,University of Pavia数据进行实验时,$C_{max}=10$。

3.3 实验结果分析

本文共做了3个实验,分别是本文提出的基于a-BvSBEM主动学习的方法选择样本、随机的选择样本(random selection samples,RS)、基于最优标号与次优标号(BvSB)。

1)Indian Pines数据:本实验首先随机选取80个样本(每类选取5个样本),然后在主动学习的过程中,每次迭代选取10个样本加入到训练样本中,共迭代40次,最终的训练样本为480个。实验的分类结果如表1所示。从表1中的比较结果可以看出a-BvSBEM和BvSB主动学习选择的样本训练出的分类器的分类正确率比RS的高5%左右。虽然a-BvSBEM方法的分类正确率只比BvSB方法的分类正确率高1%左右,但是从图1可以看出,此方法训练出的分类器性能更稳定一些。图2所示为Indian Pines的分类实验图。

表1 Indian Pines测试集上的分类精度

	a-BvSBEM	BvSB	RS
Overall accuracy	**72.97**	71.18	67.92
Average accuracy	**67.8**	65.71	52.06
kstatistic	**0.677 6**	0.664 3	0.626

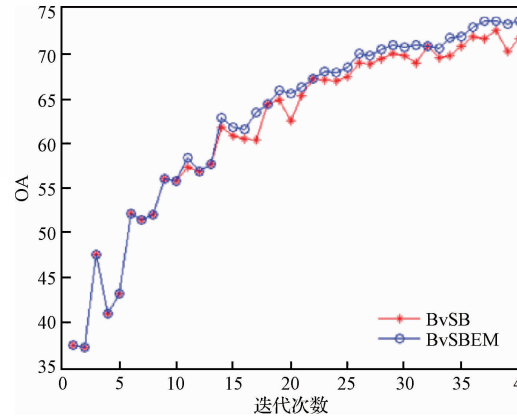

图1 Indian Pines分类器性能比较

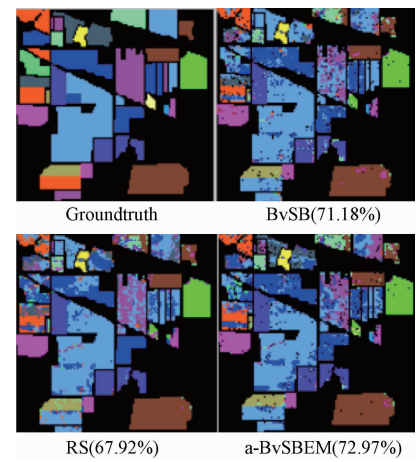

图2 Indian Pines的分类实验

2)University of Pavia数据:本实验首先随机选取45个样本(每类选取5个样本),然后在主动学习的过程中,每次迭代选取5个样本加入到训练样本中,共迭代30次。因此,训练样本的总数为195。实验的分类结果如表2所示。从表2及图3可以得出和Indian Pines数据一样的结果。图4所示为University of Pavia的分类实验图。

表2 University of Pavia测试集上的分类精度

	a-BvSBEM	BvSB	RS
overall accuracy	**83.4**	82.04	48
average accuracy	**75**	74.96	64.34
kstatistic	**0.769 5**	0.756 4	0.433 2

从Indian Pines数据和University of Pavia数据的分类结果可以看出,提出的基于a-BvSBEM的主动学习方法,在训练样本数量很少的情况下,就可以达到很好的效果,并且性能更加稳定。

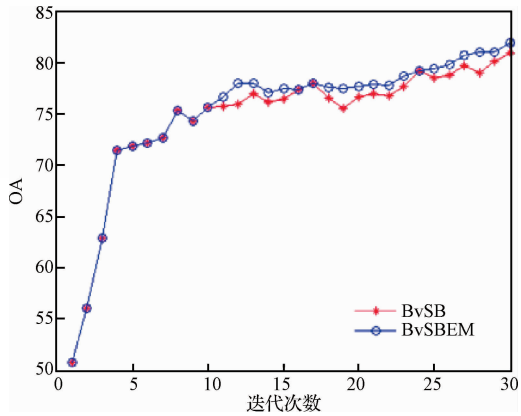

图 3　University of Pavia 分类器性能比较

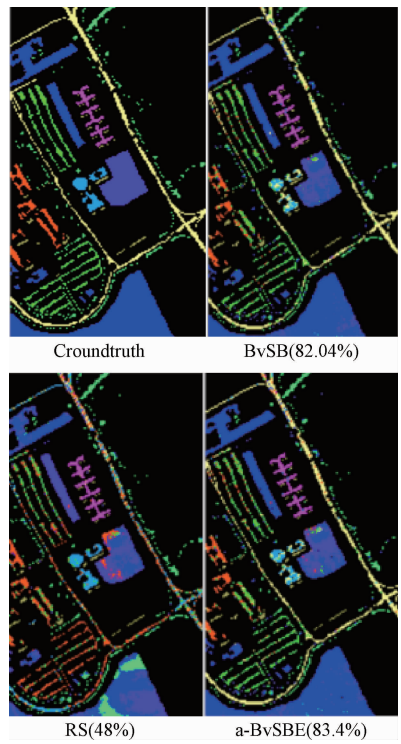

图 4　University of Pavia 的分类实验图

4　结　论

本文提出的基于 a-BvSBEM 的主动学习方法合的方法，结合了样本的不确定性和代表性，使选择的样本含有更多的信息量，能够训练出分类准确率高的分类器。由于本文结合了不确定性，在实验的大多数情况下，训练出的分类器的性能要比 BvSB 的性能更加稳定。并且本文用自适应的方法进行权重系数的选择，更加灵活。

参 考 文 献

[1] LI J，MARPU P R，PLAZA A，et al. Generalized composite kernel framework for hyperspectral image classification[J]. IEEE Transactions on Geoscience & Remote Sensing，2013，51(9)：4816-4829.

[2] 王建国，张鑫礼，张文兴. 核模糊 C 均值聚类粒度支持向量机方法研究[J]. 中国测试，2016，42(2)：96-99.

[3] 张国刚，徐向辉，Zhang G S，等. 基于加权纹理特征的 SAR 图像目标识别算法[J]. 国外电子测量技术，2015，34(9)：22-25.

[4] 宋倩，黄睿. 基于属性剖面和支持向量机的遥感图像检索[J]. 电子测量技术，2016，39(8)：96-99.

[5] 汪济洲，鲁昌华，蒋薇薇，等. 一种新的基于混合粒子的粒化支持向量机算法[J]. 电子测量与仪器学报，2015，29(4)：591-597.

[6] HU D M，LIU Q，NIU G，et al. Study on phase retardation characteristic of LCVR using dispersion analysis and SVM[J]. Instrumentation，2015，2(2)：11-17.

[7] DEMIR B，PERSELLO C，BRUZZONE L. Batch-mode active-learningmethods for the interactive classification of remote sensing images[J]. Transactions on Geoscience & Remote Sensing，2011，49(3)：1014-1031.

[8] 陈荣，曹永锋，孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报，2011，37(8)：954-962.

[9] JOSHI A J，PORIKLI F，PAPANIKOLOPOULOS N. Multi-class active learning for image classification[C]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition，2009：2372-2379.

[10] TARABALKA Y. Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques[J]. IEEE Transanctions on Geoscience & Remote Sensing，2009，47(8)：2973-2987.

[11] 袁永华，李玉，赵雪梅. 基于谱聚类的高分辨率全色遥感影像分割[J]. 仪器仪表学报，2016，37(7)：1656-1664.

作 者 简 介

张琳，1990 年出生，河海大学硕士研究生，研究方向为高光谱图像分类。
E-mail：719523698@qq. com