引 用 格 式：Yu Yao, Su Hongjun, Yao Wenjing. Boosting Ensemble Learning for Hyperspectral Image Classification Using Tangent Collaborative Representation[J]. Remote Sensing Technology and Application, 2020, 35(3): 634-644.[虞瑶, 苏红军, 姚文静. 基于Boosting的高光谱遥感切空间协同表示集成学习方法[J]. 遥感技术与应用, 2020, 35(3): 634-644.]
doi: 10.11873/j.issn.1004-0323.2020.3.0634

基于Boosting的高光谱遥感切空间协同表示集成学习方法

虞 瑶, 苏红军, 姚文静
(河海大学地球科学与工程学院, 江苏 南京 211100)

摘要：近年来, 协同表示分类(Collaborative Representation Classification, CRC)算法成为高光谱遥感影像分类的研究热点, 尤其是切空间协同表示分类(Tangent Space Collaborative Representation, TCRC)利用切平面估计测试样本的局部流形, 其分类精度得到了显著提高。为进一步提升高光谱遥感影像分类的准确性和可靠性, 提出了基于Boosting的高光谱遥感影像切空间协同表示分类算法(Boosting-based Tangent Space Collaborative Representation Classification, Boost TCRC)。Boost TCRC算法采用TCRC算法作为基分类器, 通过Boosting原理自适应地调整训练样本的权重, 增大错分样本的权重从而使得分类器专注于较难分类的训练样本, 然后在基于残差域融合时根据基分类器的分类表现赋予其权重, 最终采用最小重构误差的原则对测试样本进行分类。实验采用HyMap(Hyperspectral Mapper)和AVIRIS(Airborne Visible Infrared Imaging Spectrometer)等高光谱遥感影像数据对所提出算法的性能进行了综合评价, 结果表明：基于Boosting的集成方式可有效提升TCRC算法的分类效果。针对HyMap数据, Boost TCRC算法总体分类精度和Kappa系数分别为93.73%和0.920 8, 两种精度指标分别高于TCRC算法2.82%和0.032 3, 同时分别高于AdaBoost ELM算法1.81%和0.022 5。对于AVIRIS数据, Boost TCRC算法总体分类精度和kappa系数为84.11%和0.812 0, 两种精度指标分别高于TCRC算法3.97%和0.049 3, 同时分别高于AdaBoost ELM算法12.02%和0.143 6。

关 键 词：切空间协同表示；集成学习；Boosting；高光谱遥感分类
中图分类号：TP751 **文献标志码**：A **文章编号**：1004-0323(2020)03-0634-11

1 引 言

高光谱遥感能够获取数百个连续窄波段的光谱信息, 且所获取的信息包含丰富的空间和光谱信息, 成为地质制图、植被调查、城市规划、军事调查和环境监测等领域的有效技术手段[1-3]。其中实现地物目标分类是高光谱遥感数据应用研究的主要内容[4]。然而, 高光谱遥感影像分类面临着一些巨大的挑战, 主要是高光谱遥感影像数据量大而可利用的标签样本少。且当训练数据有限时, 随着波段数目地继续增加, 分类精度反而下降, 即所谓的“休斯现象”[6]。针对高光谱遥感影像分类面临的难题, 许多经典的机器学习和数据挖掘算法应用于高光谱遥感分类中, 并且取得了较好的效果, 例如：支持向量机(Support Vector Machine, SVM)[5]、极限学

收稿日期：2019-04-01；**修订日期**：2020-04-21
基金项目：国家自然科学基金项目“高光谱遥感影像多特征优化模型与协同表示分类”(41571325)、“高光谱遥感表示模型与分类器动态集成方法”(41871220)资助。
作者简介：虞 瑶(1995—), 女, 安徽安庆人, 硕士研究生, 主要从事高光谱遥感影像分类研究。E-mail: yuyaoyao_yy@163.com
通讯作者：苏红军(1985—), 男, 河南永城人, 教授, 博导, 主要从事高光谱遥感、资源环境遥感等研究。E-mail: hjsu@hhu.edu.cn

习机(Extreme Learning Machine,ELM)[7]和随机森林(Random Forest,RF)[8]等。近年来深度学习(Deep Learning)[9]和迁移学习(Transfer Learning)等算法成为高光谱遥感影像分类的研究热点。但任何一种分类算法都不是万能的,在取得较好的分类精度的同时也都有自身的缺陷。因此,除了发展性能更先进分类器外,利用集成学习综合各分类器的优点进行图像分类也成为热点方向[10]。

集成学习不是特指某种分类算法,而是集成多个基分类器共同决策的机器学习方法[11]。该方法通过选择简单的分类算法,获得多个不同的基分类器,然后采用某种集成方式组合成一个强分类器,从而显著提高分类系统的泛化能力和分类精度[12]。近年来,随着集成学习理论的提出与发展,基于集成学习的高光谱遥感影像分类算法引起了研究人员的广泛关注。随机森林(Random Forest,RF)是其中最具代表性的集成学习算法[14]。该算法基于决策树的集合,利用Bootstrape(自助法)采样方法生成训练子集,基于训练子集训练决策树,每颗决策树投票决策出最终分类结果[15]。针对高光谱遥感影像分类面临的高维灾难问题,引入子空间的概念,提出了基于随机子空间的极限学习机集成和基于旋转子空间的极限学习机等两种集成方式[16]。Boosting作为一种经典的集成学习方式,通过改变训练样本的权重分布来训练基分类器并将其预测结果组合成一个强分类器。Boosting集成策略中基分类器的选取至关重要,基于AdaBoost的神经网络集成算法利用神经网络算法作为基分类器取得了不错的分类表现[17]。由于ELM的高性能,Samat等[18]提出的基于AdaBoost的ELM集成算法分类精度显著优于基于AdaBoost的神经网络集成方法。但基分类器ELM随机化输入权重,导致该集成算法稳定性较差。多特征的利用可以进一步提升分类效果,Chen等[19]提出结合多特征和AdaBoost的算法,在集成过程中将各种类的特征赋予不同的基分类器,但是特征的种类对实验结果影响较大。Xia等[20]将扩展形态学属性剖面作为空间特征,采用随机森林作为基分类器,提出了基于Boosting的随机森林集成,极大地提高了分类精度。

近年来,协同表示(Collaborative Representation, CR)以模型简单、结构稳定在高光谱遥感分类中得到越来越多的关注。协同表示认为测试样本可以由训练样本集中同类的样本子集进行线性表示,分类依据为某类训练样本子集的表示估计值与测试样本的真实值最为接近。协同表示框架中的基于切空间的协同表示算法[21](Tangent Space Collaborative Representation,TCRC)利用了测试样本的简化切空间,分类精度进一步得到提升。在模式识别领域中,将多重协同表示与Boosting思想相结合的算法[22]已经取得了良好的识别性能,该算法采用CR作为基分类器,分类精度提升有限。因此考虑基于Boosting原理引入分类表现较优的TCRC作为基分类器,可进一步提升高光谱遥感影像分类的准确性和可靠性。

本文提出的Boost TCRC算法为同构集成,即利用相同的分类算法TCRC作为基分类器,结合Boosting算法思想,引入训练样本权重和分类器权重,在迭代训练过程中提高错误分类样本的权重,降低正确分类样本的权重,使得下一轮的分类器更加关注被错误分类的训练样本。并且根据基分类器的分类表现赋予其投票权重,错误率低的基分类器赋予高权重,错误率高的基分类器赋予低权重,最终实现各基分类在残差域有权重的融合。

为进一步提升切空间协同表示算法的分类精度,提出基于Boosting的切空间协同表示分类集成算法,探讨Boosting集成学习方式对高光谱遥感影像分类效果的影响。并且使用HyMap和AVIRIS两个不同的传感器的高光谱遥感影像数据进行实验验证。实验证明该算法是一种有效的同构集成学习方法。

2　切空间协同表示分类与Boosting

2.1　切空间协同表示

假设$\boldsymbol{X}\in\mathbf{R}^{d*M}$表示为$M$个训练样本的集合($d$表示波段数),其中包含$K$个类别。训练样本集构造相应的字典$\boldsymbol{D}=\left[\boldsymbol{D}_1,\boldsymbol{D}_2,...\boldsymbol{D}_K\right]$,其中第$m$类样本构成的字典表示为$\boldsymbol{D}_m=\{x_{mi}\}_{i=1}^{M_m}$,$m\in\{1,2,...,K\}$,$\sum_{m=1}^{K}M_m=M$。协同表示的目标函数表达式如下:

$$\boldsymbol{\alpha}=\arg\min_{\boldsymbol{\alpha}^*}\left(\left\|\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\alpha}^*\right\|_2^2\right)+\lambda\left\|\boldsymbol{\alpha}^*\right\|_2^2 \tag{1}$$

在高光谱遥感影像分类的问题中,总是假设同一类别的样本位于同一个低维流形中。根据该假设,测试样本的光谱空间及其可能的变化空间位于同一低维流形中。因此,转换表示的形式为:

$$T(\boldsymbol{y},\boldsymbol{v}):\boldsymbol{y}\in\mathrm{M}\xrightarrow{V}\boldsymbol{y}'\in\mathrm{M} \tag{2}$$

其中:$\boldsymbol{y}$和$\boldsymbol{y}'$分别代表原始和变换的光谱特征空间,并且$\boldsymbol{v}$反映光谱特征的各种变化。流形结构可有效

地解决非线性问题,从而提升分类表现。测试样本$\boldsymbol{y}$的局部流形结构可以由测试样本$\boldsymbol{y}$的切空间近似表示,将局部流形结构嵌入协同表示分类模型中,表达式如下所示:

$$T(\boldsymbol{y},\boldsymbol{v})=T(\boldsymbol{y},0)+\left.\frac{\partial T(\boldsymbol{y},\boldsymbol{v})}{\partial \boldsymbol{v}}\right|_{v=0}\boldsymbol{v}+o(\|\boldsymbol{v}\|^2)\approx \boldsymbol{y}+T(\boldsymbol{y})\boldsymbol{v} \tag{3}$$

$$(\boldsymbol{\alpha},\boldsymbol{v})=\arg\min_{\boldsymbol{\alpha},\boldsymbol{v}}\left(\left\|\boldsymbol{y}+T(\boldsymbol{y})\boldsymbol{v}-\boldsymbol{D}\boldsymbol{\alpha}\right\|_2^2+\lambda\|\boldsymbol{\alpha}\|_2^2\right) \tag{4}$$

其中:$T(\boldsymbol{y})=\left(\partial T(\boldsymbol{y},\boldsymbol{v})/\partial \boldsymbol{v}\right)\Big|_{v=0}$代表切空间的基。切空间距离比欧氏距离更能反映出样本点变换间的真实距离。切空间距离可以由测试样本$\boldsymbol{y}$与其邻域像元的光谱特征向量之差近似表示,即$\Delta\boldsymbol{y}=[\boldsymbol{y}_1'-\boldsymbol{y};\boldsymbol{y}_2'-\boldsymbol{y};\ldots\boldsymbol{y}_n'-\boldsymbol{y}]$。$n$表示为测试样本$\boldsymbol{y}$邻域像素的数量。当邻域足够大时,将会有:

$$\operatorname{span}(\Delta\boldsymbol{y})\cong\operatorname{span}(T(\boldsymbol{y})) \tag{5}$$

$$\forall\boldsymbol{v},\exists\boldsymbol{\beta}\Rightarrow T(\boldsymbol{y})\boldsymbol{v}=\Delta\boldsymbol{y}\boldsymbol{\beta} \tag{6}$$

$\Delta\boldsymbol{y}\boldsymbol{\beta}$具有自适应空间信息的局部流形结构,可以有效地提高样本间的可区分性。加入新的正则化项$\boldsymbol{\beta}$可使切空间协同表示模型更加稳定,切空间协同表示的目标函数变为:

$$(\boldsymbol{\alpha},\boldsymbol{\beta})=\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\left\|\boldsymbol{y}+\Delta\boldsymbol{y}\boldsymbol{\beta}-\boldsymbol{D}\boldsymbol{\alpha}\right\|_2^2+\lambda\|\boldsymbol{\alpha}\|_2^2+\eta\|\boldsymbol{\beta}\|_2^2\right) \tag{7}$$

其中:λ、η为正则化系数用于平衡惩罚项和误差项的大小。求出目标函数的最小值,求解得到$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$的解析解为:

$$\boldsymbol{\alpha}=(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\lambda\boldsymbol{I}-\boldsymbol{D}^{\mathrm{T}}\mathbf{P}\boldsymbol{D})^{-1}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{D}^{\mathrm{T}}\mathbf{P}\boldsymbol{y}) \tag{8}$$

$$\boldsymbol{\beta}=(\Delta\boldsymbol{y}^{\mathrm{T}}\Delta\boldsymbol{y}+\eta\boldsymbol{I})^{-1}(\Delta\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\alpha}-\Delta\boldsymbol{y}^{\mathrm{T}}\Delta\boldsymbol{y}) \tag{9}$$

其中:$\mathbf{P}=\Delta\boldsymbol{y}(\Delta\boldsymbol{y}^{\mathrm{T}}\Delta\boldsymbol{y}+\eta\boldsymbol{I})\Delta\boldsymbol{y}^{\mathrm{T}}$。如果测试样本$\boldsymbol{y}$属于第$m$类,则测试样本$\boldsymbol{y}$的最佳线性表示近似值为$\tilde{\boldsymbol{y}}_m=\boldsymbol{D}_m(\boldsymbol{D}_m^{\mathrm{T}}\boldsymbol{D}_m+\lambda\boldsymbol{I}-\boldsymbol{D}_m^{\mathrm{T}}\mathbf{P}\boldsymbol{D}_m)^{-1}(\boldsymbol{D}_m^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{D}_m^{\mathrm{T}}\mathbf{P}\boldsymbol{y})$。利用重构误差最小的原则对测试样本$\boldsymbol{y}$进行分类:

$$\operatorname{class}(\boldsymbol{y})=\arg\min_{m=1,2,\ldots K}r_m(\boldsymbol{y})=\arg\min_{m=1,2\ldots K}\left(\left\|\boldsymbol{y}+\Delta\boldsymbol{y}\boldsymbol{\beta}-\tilde{\boldsymbol{y}}_m\right\|_2^2\right) \tag{10}$$

2.2 Boosting

集成学习通过选择结构较为简单的学习算法作为基分类器,将多个基分类器的预测结果以某种结合策略集成,从而得到分类精度高且鲁棒性强的分类器。生成基分类器的方法一般分为两大类:①将不同学习算法应用于相同的训练样本集上,即异构集成;②将同一学习算法应用于不同的训练样本集上,可以通过对训练样本进行有放回采样或者改变输入特征,即同构集成。集成学习系统有效的关键在于能否产生具有差异性强和分类性能高的基分类器。差异性要求基分类器产生的泛化误差应尽可能不相关。为了达到预设的差异性,同构集成方式中经常使用3种策略:①基于不同训练样本的构造方式,如经典的集成方式Bagging和Boosting算法;②基于不同特征集的构造方式,如随机子空间算法和旋转森林等;③基于同一分类算法的不同参数组合,多数分类器中含有参数组合,利用不同的参数组合得到不同的分类结果。

集成学习通常也称为分类器集合或者多分类器系统,它认为不同的分类器具有不同的决策性能,组合不同的分类器一起使用,可以有效提高分类系统的分类精度和泛化能力。原因是多数的分类器取得局部解,而不同的分类器从不同出发点进行局部搜索,因此集成学习更容易逼近目标函数从而达到整体最优。而且集成学习还解决了单个分类器遇到的过适应问题,不易发生对训练数据过于精细刻画的现象。集成学习通常要求基分类器的分类精度稍微高于随机猜测。Bagging集成策略要求基分类器必须是不稳定的,也就是说分类器对样本或者参数越敏感,集成效果表现越好;而Boosting算法对稳定和不稳定分类器均适用。

Freund在Boosting的基础上提出了一种改进算法—AdaBoost(Adaptive Boosting)算法。该算法具体实现步骤为,首先令所有训练样本的初始权重值均为$1/N$,权重值代表被基分类器选为训练样本的概率。在之后的训练迭代过程中,若某训练样本被正确分类,权重值减小,则构造下一轮的训练样本集时,被选中的概率降低;若某训练样本被错误分类,权重值增大,则构造下一轮的训练样本集时,被选中的概率增加。因此,AdaBoost算法更关注那些较难分类的训练样本。权重更新过的训练样本集被用来训练下一个分类器。同时,在迭代训练过程中自适应地调整各基分类器的权重,决策投票时误差率大的基分类器获得低权重,而误差率小的基分类器获得高权重。最后各基分类器的预测值加权平均得到最终的集成分类结果。Bagging算法能明显减少分类的方差,而Boosting算法能同时减少分类的方差和偏差,因此大部分情况下Boosting算法要比Bagging算法准确性高,但Boosting算法对噪声十分敏感。并且当对训练数据产生过拟合现象时,Boosting算法可能会失效。AdaBoost算法的具体过程如下:

(1)选择基分类器 h。

(2)输入训练样本集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$N$ 表示训练样本总数，y_i 表示类别标签 $y_i\in\{-1,+1\}$，设置弱分类器数量 T。

(3)对训练集 S 的样本权值分布 D_i 进行初始化得到 D_1；$D_1(i)=1/N$, i=1, 2, ⋯, N。

(4)对 t=1：T 循环执行(a)～(c)。

(a)根据训练样本权值分布 D_t 随机可重复的获取第 t 个训练集 S_t，利用训练集合 S_t 和弱学习算法训练基分类器 $h_t(x)$，并计算基分类器 $h_t(x)$ 的加权错误率 ε_t：

$$\varepsilon_t=\sum_{i=1}^{m}D_t.*E,\text{其中 }E(i)=\begin{cases}0, h_t(x_i)=y_i\\1, h_t(x_i)\neq y_i\end{cases}$$

(b)根据错误率计算基分类器权值 α_t：

$$\alpha_t=\frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

(c)更新训练样本权值 $D_{t+1}=\frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$，其中 Z_t 为标准化因子，保证 D_{t+1} 是一个分布。

(5)输出强分类器 $H(x)=\text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$，其中 $H(x)$ 表示分类器组合模型，sign(.)为符号函数，取值为 1 或-1。

3　基于Boosting的切空间协同表示集成学习方法

训练样本的权重描述了训练样本对于测试样本分类的贡献率。每个训练样本重要性不同，提出的 Boost TCRC 算法可以重新平衡每个训练样本在每类中的重要性。在每次迭代训练过程中，根据分类器的当前错误率调整每个训练样本的权值，降低正确分类的训练数据的概率，增加错误分类的训练数据的概率。通过这种方式，Boost TCRC 算法专注于信息量大或分类难度较大的训练样本。同时该算法自适应地降低错误分类率较大的基分类器 TCRC 的权重和增加错误分类率较小的基分类器 TCRC 的权重。迭代结束后，根据 Boost TCRC 算法，测试样本 $\boldsymbol{y}$ 分类到具有最小加权残差的类别中：

$$\text{class}(\boldsymbol{y})=\arg\min_{m=1,2...K}r_m(\boldsymbol{y})\tag{11}$$

其中：$r_m(y)=\sum_{t=1}^{T}\partial_t\left\|\boldsymbol{y}+\Delta\boldsymbol{y}\boldsymbol{\beta}^t-\boldsymbol{D}_m^t\boldsymbol{\alpha}_m^t\right\|_2^2$。$T$ 个基分类器 TCRC 的权重系数为 $\boldsymbol{\partial}=[\partial_1,...,\partial_T]$ 且 $\sum_{t=1}^{T}\partial_t=1$。$\boldsymbol{\partial}$ 值越大，赋予基分类器的权重越大，而当 $\partial_t=0$ 时迭代停止。d_t 代表着测试样本与其真实类别的样本子集表示值之间的残差值减去其他类别的样本子集表示值的最小残差值。若 $d_t(i)<0$ 则表明第 i 个训练样本分类正确，$d_t(i)$ 值越小，说明第 i 个样本的可区分性更强。若 $d_t(i)>0$ 则表明第 i 个训练样本分类错误，$d_t(i)$ 值越大，说明第 i 个样本错分程度越明显。d_t 值越小，代表字典 X_t 的判别能力越强。针对切空间协同表示算法的分类特点，采用 ε_t 代替经典 AdaBoost 算法中错误率可以更加充分得知每一个训练样本的错分程度，从而更精确地调整每个训练样本的权重。可以很明显地看出每次迭代过程中 $|\varepsilon_t|<b_t$。Boost TCRC 算法的主要步骤如下：

(1)输入：基分类器 TCRC，训练集 $\boldsymbol{X}\in\mathbf{R}^{d*M}$，测试样本 $\boldsymbol{y}$，测试样本差值 $\Delta\boldsymbol{y}$，训练样本差值 $\Delta\boldsymbol{X}$，设置分类器的个数(集成次数)T，正则化参数 λ，η。

(2)对训练集 $\boldsymbol{X}$ 的样本权值分布 D_i 进行初始化得到 D_1：$D_1(i)=1/M$, i=1, 2, ..., M。

(3)对 t=1：T 循环执行(a)～(f)步骤。

(a)根据训练样本权值分布 D_t 获取第 t 个训练集 $\boldsymbol{X}_t$，利用训练集 $\boldsymbol{X}_t$ 构造字典 D_t；

(b)根据公式(8)和(9)计算 $d_t(i)=\left\|\boldsymbol{X}+\Delta\boldsymbol{X}\beta^t-\boldsymbol{D}_c^t\alpha_c{}^t\right\|_2^2-\min_{m\neq c}\left\|\boldsymbol{X}+\Delta\boldsymbol{X}\beta^t-\boldsymbol{D}_m^t\alpha_m{}^t\right\|_2^2$；

(c)计算 $\varepsilon_t=dot(D_t,d_t)$ 和 $b_t=\max|d_t|$；

(d)计算测试样本 $\boldsymbol{y}$ 的残差 $r_m(\boldsymbol{y})=\left\|\boldsymbol{y}+\Delta\boldsymbol{y}\boldsymbol{\beta}^t-\boldsymbol{D}_m^t\boldsymbol{\alpha}_m^t\right\|_2^2, m=1,2,\cdots,K$；

(e)计算基分类器权重 $\partial_t=\max\left\{\frac{1}{2b_t}\log\left(\frac{b_t-\varepsilon_t}{b_t+\varepsilon_t}\right),0\right\}$；

(f)更新训练样本的权重 $D_{t+1}(i)=\frac{D_t(i)}{Z_t}e^{\partial_i d_i}$，其中 Z_t 为标准化因子，保证是 D_{t+1} 一个分布；

(4) 输出基分类器的归一化后权重 $\{\partial_t\}_{t=1}^{T}$，采用如下的分类准则对测试样本 $\boldsymbol{y}$ 进行分类，最终得到测试样本的标签：

$$\text{class}(\boldsymbol{y})=\arg\min_{m=1,2...K}\sum_{t=1}^{T}\partial_t\left\|\boldsymbol{y}+\Delta\boldsymbol{y}\boldsymbol{\beta}^t-\boldsymbol{D}_m^t\boldsymbol{\alpha}_m^t\right\|_2^2$$

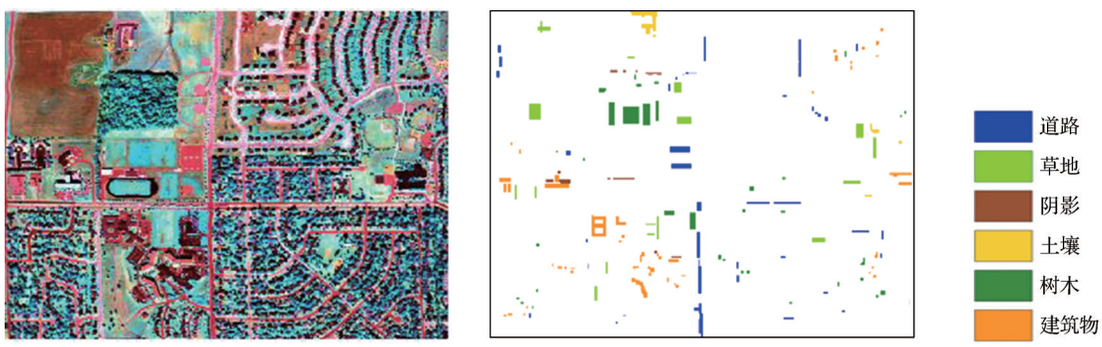
4　实验与分析

4.1　实验数据

实验数据一是普度大学(Purdue Campus)西拉斐特分校校区，该数据于 1999 年 9 月 30 日通过机载

高光谱制图仪(Hyperspectral Mapper,HYMAP)系统采集,在可见光和红外区域(400~2 400 nm)涵盖128条光谱波段。实验中剔除水吸收波段后保留126条波段,空间分辨率为3.5 m。该实验数据大小为377像素×512像素,共包含6类地物,分别为道路、草地、阴影、土壤、树木和建筑物。该高光谱数据假彩色图如图1(a)所示,样本分布如图1(b)所示。标记样本被随机分成训练样本和测试样本,共选取了90个训练样本,平均每类地物20个训练样本。

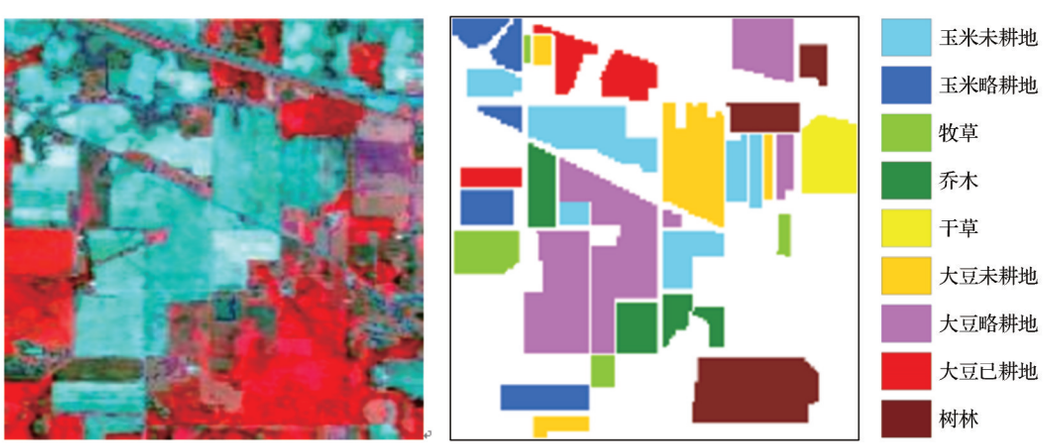
实验数据二是美国印第安纳州Indian Pines实验区,采用的是机载可见光红外成像光谱仪(Airborne Visible Infrared Imaging Spectrometer,AVIRIS)采集,包括从可见光到近红外(400~2 450 nm)的220个波段的光谱数据,剔除水吸收后保留200个光谱波段,空间分辨率约为20 m。该实验数据大小为145像素×145像素,共包括16类地物类型,为取得足够的训练样本去除7类地物,剩余的9类用于实验分析。该高光谱数据的假彩色图像如图2(a)所示,样本分布如图2(b)所示。实验根据样本分布图共随机选取466个像元作为训练样本,平均每类约占5%,其余像元作为测试样本进行精度评定。



(a) Purdue Campus数据假彩色图 (b) 地面真实值

图1 Purdue Campus 数据集

Fig.1 Purdue Campus data set



(a) Indian Pines数据假彩色图 (b) 地面真实值

图2 Purdue Campus数据集6种分类算法的分类效果图

Fig.2 Classification maps of Purdue Campus using six algorithms

4.2 实验设置

为验证Boost TCRC算法的有效性,采用Purdue Campus高光谱影像数据和Indian Pines高光谱影像数据进行实验。两组实验采用Boosting、随机森林(Random Forest)、ELM、AdaBoost ELM和TCRC分类器等作为对比算法,其中Boosting算法中基分类器为决策树,规模为20棵,随机森林算法中决策树规模为20棵。随机森林算法采用Bootstrap采样方法,对原始样本进行有放回的随机抽样,获得与原始训练样本集同等样本数量的训练样本子集。AdaBoost ELM算法为Samat等提出一种集成算法,该算法结合基分类器ELM算法和AdaBoost集成方式。两组数据中ELM算法和AdaBoost ELM算法均设置了最佳的隐含层神经元个

数。分类性能评价指标包括总体分类精度(Overall Accuracy,OA)、平均分类精度(Average Accuracy,AA)和Kappa系数。表1中详细列出了两个数据集中算法的最优参数,T代表集成次数,n的值表示相邻像素的数量,λ和η为Boost TCRC和TCRC算法中的正则化参数。六种方法重复10次实验取平均值,参数设置采用交叉验证的方法。

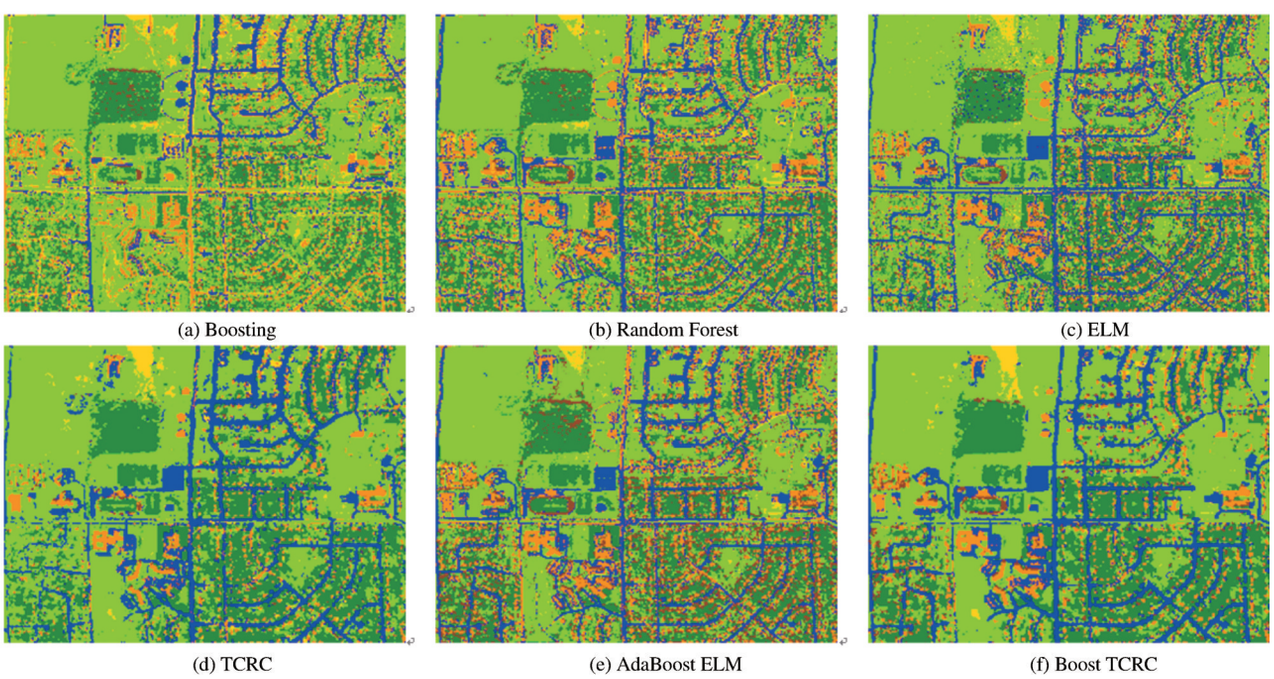
表1　实验设置的最佳参数

Table1　Optimal parameter setting in experiment

数据		Purdue Campus	Indian Pines
TCRC	λ	1e-6	1e-5
	η	1e-8	1e-8
	n	8	8
Boost TCRC	λ	1e-9	1e-9
	η	1e-8	1e-8
	n	8	8
	T	20	20

4.3　实验结果与分析

图3显示了第一组实验数据6种算法的分类结果图。不难发现,Boost TCRC分类器分类效果最好,AdaBoost ELM分类器次之,Boosting分类器效果最差。Boost TCRC算法中T设置为20次,AdaBoost ELM算法中T也设置为20次。表2中6种分类器(Boosting、RF、ELM、TCRC、AdaBoost ELM和Boots TCRC)的OA(%)值分别为86.09、87.31、89.36、90.91、91.92和93.73。可以看出,Boost TCRC算法总体分类精度、平均分类精度和Kappa



(a) Boosting　(b) Random Forest　(c) ELM

(d) TCRC　(e) AdaBoost ELM　(f) Boost TCRC

图3　Indian Pines数据集

Fig.3　Indian Pines data set

表2　实验一分类精度统计

Table2　Classification Accuracy Statistics of Experiment 1

类别	训练样本	测试样本	Boosting	RF	ELM	TCRC	AdaBoost ELM	Boost TCRC
道路	15	1 272	82.60	82.20	89.67	88.91	91.13	**92.44**
草地	15	1 099	89.78	90.93	89.12	85.54	**95.29**	89.49
阴影	15	204	71.03	94.02	66.31	90.46	81.87	88.43
土壤	15	364	73.24	94.45	80.66	86.82	86.85	91.19
树木	15	1 336	98.68	93.49	99.15	98.18	**99.33**	98.89
建筑物	15	1 270	83.72	79.67	89.60	94.63	87.32	**96.17**
总体分类精度/%			86.09	87.31	89.36	90.91	91.92	**93.73**
平均分类精度/%			83.18	89.13	85.75	90.76	90.30	**92.77**
Kappa系数			0.825 7	0.841 3	0.866 5	0.888 5	0.898 3	**0.920 8**
时间/s			1.25	0.10	0.14	18.40	1.30	126.95

系数比基分类器 TCRC 分别提高了 2.82%、2.01% 和 0.032，体现了 Boosting 集成方式的优越性和有效性。AdaBoost ELM 分类算法总体精度也比基分类器 ELM 算法提升了约 2.5%，但提升的幅度略低于 Boost TCRC 集成算法。RF 总体分类精度为 87.31%，比 AdaBoost ELM 算法和 Boost TCRC 算法总体分类精度分别低了 4.61% 和 6.42%。Boosting 算法总体分类精度最低，仅为 86.09%。Boost TCRC 分类器在 6 种算法中总体精度和平均精度均取得最佳的实验结果，其次是 AdaBoost ELM 分类器和 TCRC 分类器。

第二组实验数据 6 种分类算法的分类结果如图 4 所示。Boost TCRC 算法的分类结果比 TCRC 算法的分类结果更准确。由于利用邻域信息，Boost TCRC 算法的分类图比 AdaBoost ELM 算法的分类图更平滑。如表 3 所示，6 种分类器（Boosting、RF、ELM、TCRC、AdaBoost ELM 和 Boots TCRC）的 OA（%）值分别为 69.48、71.05、71.15、80.14、72.09 和 84.11。Boost TCRC 算法的分类性能均显著优于其他分类器，其中 Boost TCRC 算法总体分类精度、平均分类精度和 Kappa 系数相比较于 TCRC 算法分别提高了约 4%、0.9% 和 0.049 2，充分证明了集成学习的有效性。基分类 ELM 算法总体分类精度为 71.51%，所以 AdaBoost ELM 集成算法分类精度有限，约为 72.09%。随机森林总体分类精度为 71.05%，比 Boost TCRC 算法的总体分类精度低了约 13%。

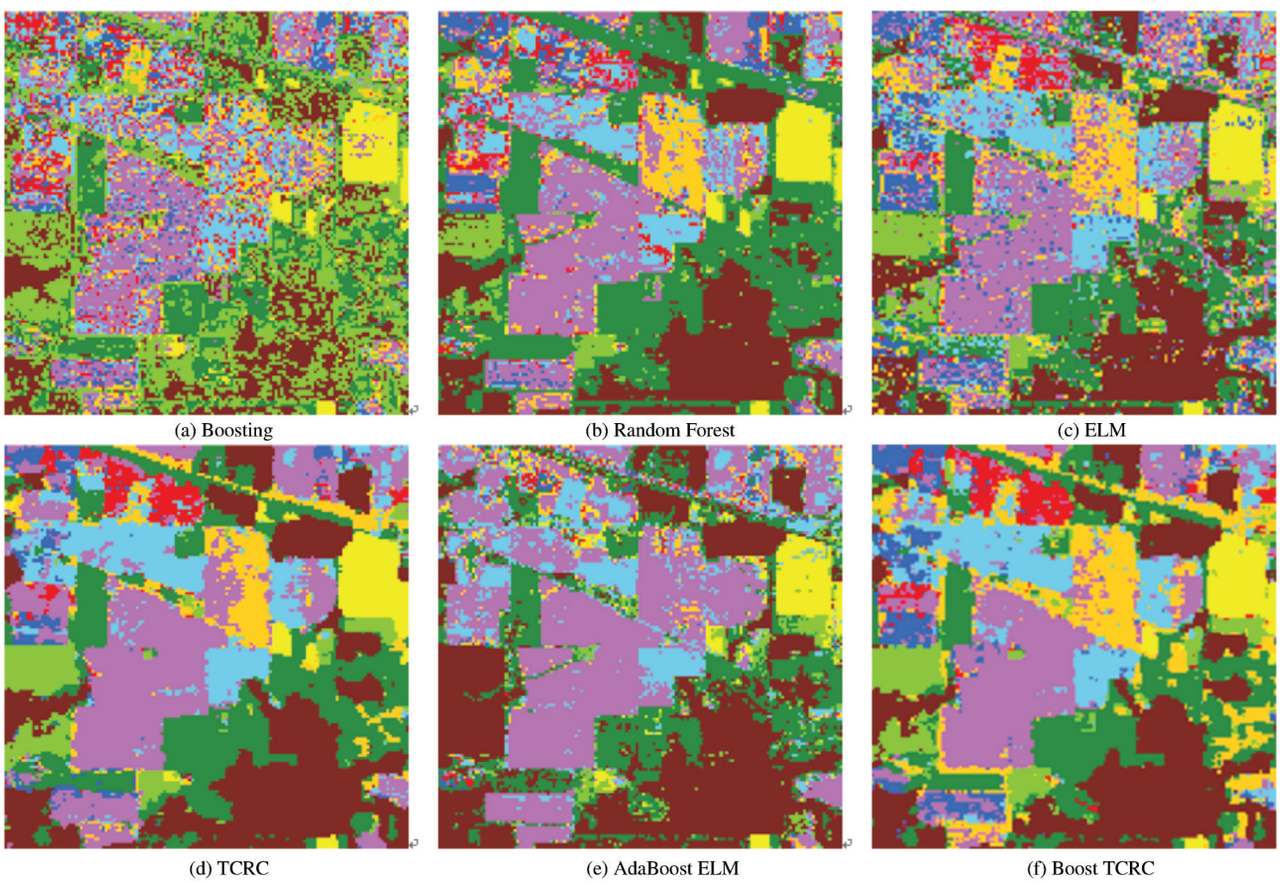


图 4 Indian Pines 数据集 6 种算法的分类效果图

Fig.4 Classification maps of Indian Pines using six algorithms

值得注意的是，Boost TCRC 分类器的总体精度高于 AdaBoost ELM 分类器约 12%，高于 Boosting 分类器约 14.63%，说明基分类器对集成的效果影响较大。

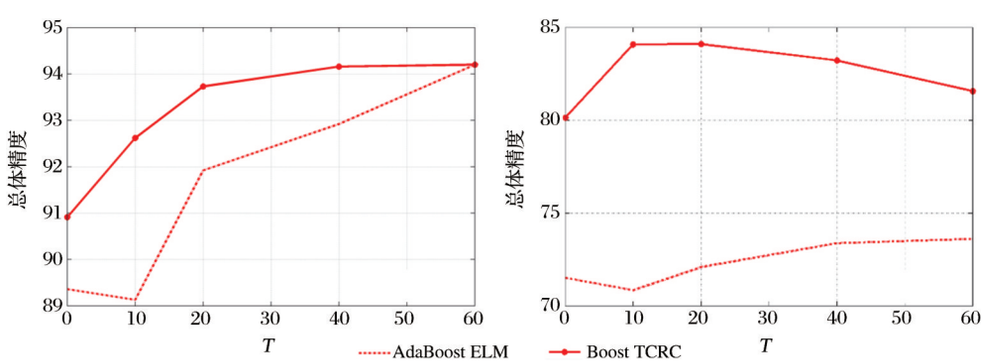
4.4 参数分析

图 5 对应 AdaBoost ELM 和 Boost TCRC 两种集成算法的总体分类精度值随集成次数 T 的变化。实验过程中 T 值分别设置为 10、20、40 和 60。图 5 中 X 坐标轴的零值代表分别两种集成算法的基分类器（TCRC，ELM）的总体分类精度值。从图 5(a) 中可发现，对于 HyMap 数据，Boost TCRC 分类器总体分类精度随 T 值增加而增大，当 T 值设置为 60 次时，总体分类精度达到约 94%。比 TCRC 算法总体分类精度提升了约 3%。值得注意的是，当 Indi-

表 3　实验二分类精度统计

Table 3　Classification Accuracy Statistics of Experiment 2

类别	训练样本	测试样本	Boosting	RF	ELM	TCRC	AdaBoost ELM	Boost TCRC
C1	72	1 356	55.41	57.69	62.81	77.25	64.70	**84.85**
C2	42	788	62.28	45.65	54.27	**82.03**	53.77	75.27
C3	25	458	73.51	73.44	87.20	**97.59**	90.84	95.83
C4	37	693	89.22	91.23	94.67	94.47	96.94	**97.29**
C5	24	454	94.05	97.05	99.04	99.52	99.80	**100**
C6	49	923	58.46	51.53	56.81	**80.01**	55.69	73.18
C7	123	2332	63.06	79.60	62.90	66.03	63.54	75.16
C8	30	563	52.91	35.99	62.93	81.43	71.67	**84.19**
C9	64	1 201	94.81	95.27	98.45	98.49	98.93	**99.14**
总体分类精度/%			69.48	71.05	71.51	80.14	72.09	**84.11**
平均分类精度/%			71.53	69.72	75.45	86.31	77.32	**87.21**
Kappa 系数			0.637 1	0.655 3	0.662 1	0.762 8	0.668 4	**0.812 0**
时间/s			12.21	0.35	0.13	164.30	2.30	3 651.34



(a) Purdue Campus数据不同T值下算法的分类结果　　(b) Indian Pines数据不同T值下算法的分类结果

图 5　两组数据不同 T 值下算法的分类结果

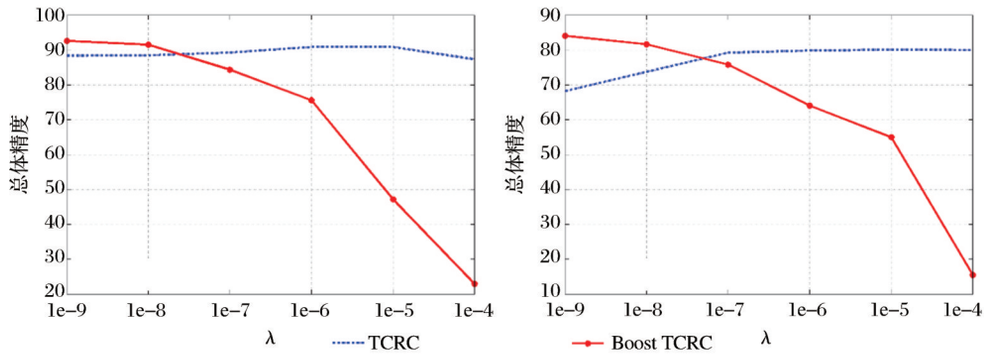
Fig.5　Overall accuracy of different algorithms for two data set with varying T

an Pines 数据集中 T 值设置大于 10 次时，Boost TCRC 算法的分类表现反而越来越差。可能的原因是 Indian Pines 数据分辨率不高，噪声较多，发生了 Boost TCRC 算法对训练数据过度拟合的现象。总体而言，对于两组数据，当 T 值为 10 次到 20 次左右时，Boost TCRC 算法总体精度得到明显提高，且算法分类表现优于 AdaBoost ELM 算法。

Boost TCRC 算法中受正则化参数影响较大，该算法存在两个正则化参数，分别为 λ 和 η。实验中参数 λ 设置的范围为 1e-9～1e-4。为了方便实施实验，T 设定为 10 次。AVIRIS 数据集中 Boost TCRC 和 TCRC 两种算法最佳参数设置均为 η=1e-8 和 n=8。参数 λ 变化趋势如图 6(b)所示，可以看出 Boost TCRC 算法对正则化参数 λ 的敏感性明显高于 TCRC 算法。当 λ=1e-4 时，Boost TCRC 算法的总体精度甚至下降到 17%。对于 Purdue Campus 数据集，Boost TCRC 和 TCRC 两算法最佳参数设置为 η=1e-8 和 n=8。参数 λ 的变化趋势如图 6(a)所示。和 AVIRIS 数据类似，Boost TCRC 算法也会随着 λ 的增加出现分类精度骤降的现象。

为了进一步比较不同算法的时间复杂度，在表 2 和表 3 中记录了两组实验数据中 6 种算法的运行时

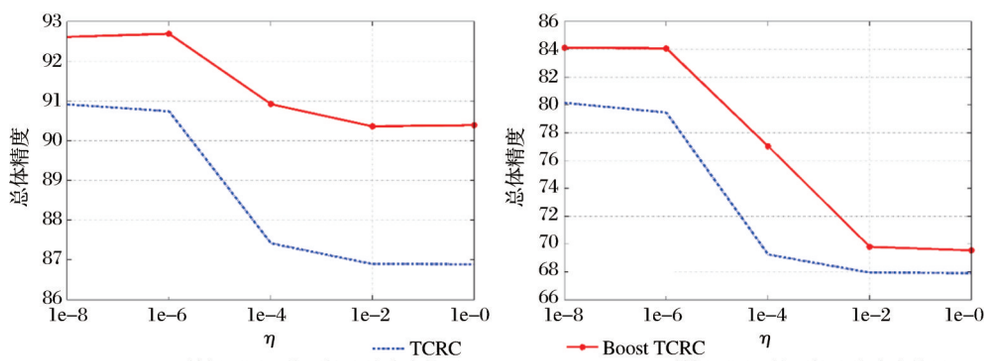
图 7 分析了两组高光谱数据中 Boost TCRC 和 TCRC 两种算法的总体分类精度值与正则化参数 η 之间的变化关系。实验中设置 η 范围为 1e-8～1e-0，T 也设定为 10 次。针对 HyMap 数据，Boost TCRC 算法最佳参数设置为 λ=1e-9，n=8，TCRC 算法最佳参数设置为 λ=1e-6，n=8。从图 7(a)中可得，Boost TCRC 算法的性能优于 TCRC 算法且两种算法均存在随着正则化参数 η 增加分类精度随之下降的趋势。AVIRIS 数据中 Boost TCRC 算法最佳参数设置为 λ=1e-9，n=8，TCRC 算法最佳参数设置为 λ=1e-4，n=8。与 HyMap 数据类似，Boost TCRC 算法分类精度高于 TCRC 算法，体现出了集成学习的优越性能。



(a) Purdue Campus数据不同λ值下算法的分类结果　(b) Indian Pines数据不同λ值下算法的分类结果

图 6　两组数据不同λ值下算法的分类结果

Fig.6　Overall accuracy of different algorithms for two data sets with varying λ



(a) Purdue Campus数据不同η值下算法的分类结果　(b) Indian Pines数据不同η值下算法的分类结果

图 7　两组数据不同η值下算法的分类结果

Fig.7　Overall accuracy of different algorithms for two data sets with varying η

间(仅分类过程,运行10次的平均时间)。实验基于2.8.GHz CPU和8G内存的计算机,所有实验均在Matlab软件中进行。本文对比的几种集成算法中,随机森林算法运算速率最快,AdaBoost ELM算法次之,Boosting算法运算速率最慢。与TCRC算法相比,由于串行生成基学习器,Boost TCRC算法导致了运算复杂度增加和运算效率的降低,耗时最多。但AdaBoost ELM集成算法采用运算速率快的基分类器ELM,所以较Boost TCRC算法耗时少。

5　结　语

为进一步提升高光谱遥感影像的分类效果,本文提出了基于Boosting的切空间协同表示高光谱遥感影像集成分类算法。该方法利用TCRC作为基分类器进行预测,自适应地学习基分类器TCRC和训练样本的权重,使得分类器专注于信息量较大或较难分类的训练样本,最终各基分类器在残差域实现有权重的融合。

采用两组高光谱影像数据进行实验验证,结果表明:①HyMap影像数据Boost TCRC算法的总体精度比TCRC算法提高了约3%,而在AVIRIS数据中总体精度提升了约4%。虽然两组数据Boost TCRC算法提升幅度不同,但均优于基分类器TCRC和AdaBoost ELM集成算法;②两组数据Boost TCRC算法的分类精度随着集成次数的增加呈现出不同趋势,但都能获得高于TCRC算法的分类精度。

本文算法虽在一定程度上提升了TCRC算法的分类效果,也存在一些不足之处,如Boost TCRC算法对正则化参数λ十分敏感,随着λ的增加会出现分类精度骤减的现象;而且迭代训练过程会不可避免地导致计算复杂度高和计算效率下降的问题。因此该算法如何减小计算复杂度有待进一步研究。

参考文献(References):

[1] Tong Qingxi, Zhang Bing, Zheng Lanfen. Hyperspectral Remote Sensing [M]. Beijing: Higher Education Press, 2006. [童庆禧,张兵,郑兰芬.高光谱遥感[M].北京:高等教育出版社,2006.]

[2] Landgrebe D. Hyperspectral Image Data Analysis[J]. IEEE Signal Processing Magazine, 2002, 19(1):17-28.

[3] Moon H, Ahn H, Kodell R L, *et al*. Ensemble Methods for

Classification of Patients for Personalized Medicine with High-dimensional Data [J]. Artificial Intelligence in Medicine, 2007, 41(3):197-207.

[4] Bioucas-Dias J M, Plaza A, Camps-Valls G, *et al*. Hyperspectral Remote Sensing Data Analysis and Future Challenges [J]. IEEE Geoscience and Remote Sensing Magazine, 2013, 1(2):6-36.

[5] Dozier J, Painter T H. Multispectral and Hyperspectral Remote Sensing of Alpine Snow Properties[J]. Annual Review Earth Planetary, Sciences, 2004, 32: 465-494.

[6] Braun A C, Weidner U, Hinz S. Classification in High-dimensional Feature Spaces-assessment Using SVM, IVM and RVM with Focus on Simulated EnMAP Data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2012, 5(2):436-443.

[7] Chen C, Li W, Su H, *et al*. Spectral-spatial Classification of Hyperspectral Image based on Kernel Extreme Learning Machine [J]. Remote Sensing, 2014, 6(6): 5795-5814. doi: 10.3390/rs6065795.

[8] Zhang Y, Cao G, Li X, *et al*. Cascaded Random Forest for Hyperspectral Image Classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018,11(4):1082-1094.

[9] Li S, Song W, Fang L, *et al*. Deep Learning for Hyperspectral Image Classification: An Overview[J].IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9):6690-6709.

[10] Zhang H, Li Y, Jiang Y, *et al*. Hyperspectral Classification based on Lightweight 3-D-CNN With Transfer Learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(8): 5813-5828.

[11] Chen Y, Wang Y, Gu Y, *et al*. Deep Learning Ensemble for Hyperspectral Image Classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(6): 1882-1897.

[12] Bao R, Xia J, Mura M D, *et al*. Combining Morphological Attribute Profiles via an Ensemble Method for Hyperspectral Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(3): 359-363.

[13] Benediktsson J A, Chanussot J, Fauvel M. Multiple Classifier Systems in Remote Sensing: from Basics to Recent Developments [C]//International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2007: 501-512.

[14] Chan J C W, Paelinckx D. Evaluation of Random Forest and Adaboost Tree-based Ensemble Classification and Spectral Band Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery[J]. Remote Sensing of Environment, 2008, 112(6): 2999-3011.doi: 10.1016/j.rse.2008.02.011.

[15] Gislason P O, Benediktsson J A, Sveinsson J R. Random Forests for Land Cover Classification[J]. Pattern Recognition Letters, 2006, 27(4): 294-300.

[16] Xia J, Dalla Mura M, Chanussot J, *et al*. Random Subspace Ensembles for Hyperspectral Image Classification with Extended Morphological Attribute Profiles [J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(9): 4768-4786.

[17] Haq Q S U, Tao L, Yang S. Neural Network based Adaboosting Approach for Hyperspectral Data Classification [C]// IEEE International Conference on Computer Science & Network Technology, 2012.

[18] Samat A, Du P, Liu S, *et al*. ELM^2: Ensemble Extreme Learning Machines for Hyperspectral Image Classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(4): 1060-1069.

[19] Chen Y, Zhao X, Lin Z. Joint Adaboost and Multifeature based Ensemble for Hyperspectral Image Classification[C]// 2014 IEEE Geoscience and Remote Sensing Symposium, 2014: 2874-2877.

[20] Xia J, Ghamisi P, Yokoya N, *et al*. Random Forest Ensembles and Extended Multiextinction Profiles for Hyperspectral Image Classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(1): 202-216.

[21] Su H, Bo Z, Qian D, *et al*. Tangent Distance-based Collaborative Representation for Hyperspectral Image Classification [J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(9):1236-1240.

[22] Chi Y, Porikli F. Classification and Boosting with Multiple Collaborative Representations[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(8):1519-1531.

Boosting Ensemble Learning for Hyperspectral Image Classification Using Tangent Collaborative Representation

Yu Yao, Su Hongjun, Yao Wenjing

(*School of Earth Sciences and Engineering, Hohai University, Nanjing* 211100, *China*)

Abstract: Recently, Collaborative Representation Classification (CRC) has attracted much attention in hyperspectral image analysis. Due to uses the tangent plane to estimate the local manifold of the test sample. Tangent Collaborative Representation Classification (TCRC) achieve better performance. Furthermore, in order to improve the classification accuracy and reliability of hyperspectral remote sensing images, a novel Boosting-based Tangent Collaborative Representation ensemble method (Boost TCRC) for hyperspectral image classification is proposed. In this algorithm, Boost TCRC algorithm choose TCRC as base classifier and adjust the weight of the training samples adaptively by using the principle of Boosting. Increasing the weight of the misclassified samples so that the classifier concentrates on the training samples that are difficult to classify. Then assigns the weights according to the classification performance of the base classifier based on the residual domain fusion. Finally, the principle of minimum reconstruction error is adopted to classify the test sample. The performance of the proposed algorithm was comprehensively evaluated by hyperspectral remote sensing image data such as HyMap (Hyperspectral Mapper) and AVIRIS (Airbone Visible Infrared Imaging Spectrometer). The Boosting method can effectively improve the classification effect of the TCRC algorithm. For HyMap data, the overall classification accuracy and kappa coefficient of Boost TCRC algorithm are 93.73% and 0.920 8 respectively. Two precision values are higher than TCRC algorithm by 2.82% and 0.032 3, and are higher than the AdaBoost ELM algorithm by 1.81% and 0.022 5. For AVIRIS data, the overall classification accuracy and kappa coefficient of Boost TCRC algorithm are 84.11% and 0.8120 respectively. Two precision values are higher than TCRC algorithm by 3.97% and 0.049 3, and are higher than AdaBoost ELM algorithm by 12.02% and 0.143 6.

Key words: Tangent collaborative representation; Ensemble learning; Boosting; Hyperspectral image classification