面向不平衡高光谱遥感分类的SMOTE和旋转森林动态集成算法

童莹萍[1,2]，冯伟[1,2]，宋怡佳[1,2]，全英汇[1,2]，黄文江[3]，高连如[3]，朱文涛[1,2]，邢孟道[4]

1. 西安电子科技大学 电子工程学院, 西安 710071;
2. 西安电子科技大学 先进遥感技术研究院, 西安 710071;
3. 中国科学院空天信息创新研究院 数字地球重点实验室, 北京 100094;
4. 西安电子科技大学 前沿交叉研究院, 西安 710071

摘 要： 旋转森林RoF（Rotation Forest）是一种功能强大的集成分类器，它在高光谱图像分类中已经获得了很多成功的应用。然而，现实数据经常存在类别不平衡的问题，这使得传统的RoF算法侧重识别多数类别的样本，而忽略了少数类样本的分类精度。SMOTE（Synthetic Minority Oversampling Technique）算法通过模拟生成新样本的方式来增加少数类别样本的数量，进而达到平衡数据集类别的效果；但是SMOTE算法目前主要被用于数据预处理阶段，并且在处理多类问题时具有增加人工噪声的风险。为了解决高光谱数据学习中的多类不平衡问题，本文提出了一个新的SMOTE和RoF动态集成算法；该算法利用动态采样因子技术，将类别分布优化和基分类器训练过程进行融合。本实验利用Indian Pines、Salinas以及Pavia University这3个公开的高光谱数据对新的SMOTE和RoF动态集成算法的性能进行测试，同时选取4种对比算法，包括随机森林、传统的RoF以及通过随机过采样和SMOTE数据预处理后的RoF算法，并且采用总体分类精度、平均分类精度、F-measure、Gmean、最小召回率、集成分类器多样性、模型训练时间以及McNemar测试等为算法性能评价标准。实验结果表明本文方法具有明显的分类优势，可以保证在增加数据总体分类精度的基础上提高小类别样本的识别精度。

关键词： 集成学习，不平衡分类，旋转森林，SMOTE，动态采样

中图分类号： P2

引用格式： 童莹萍,冯伟,宋怡佳,全英汇,黄文江,高连如,朱文涛,邢孟道.2022.面向不平衡高光谱遥感分类的SMOTE和旋转森林动态集成算法.遥感学报,26(11): 2369-2381

Tong Y P, Feng W, Song Y J, Quan Y H, Huang W J, Gao L R, Zhu W T and Xing M D. 2022. Dynamic ensemble algorithm of SMOTE and rotation forest for imbalanced hyperspectral remote sensing classification. National Remote Sensing Bulletin, 26(11):2369-2381[DOI:10.11834/jrs.20210216]

1 引 言

高光谱图像不仅具有较高分辨率，而且包含非常丰富的光谱信息，已在军用和民用领域中获得了广泛应用（Tu等，2020；杜培军 等，2016）。但是，高光谱图像分类经常会遇到类别不平衡的问题（Rodriguez等，2006；韩竹 等，2020）。所谓的类别不平衡指的是数据中某些类别的样本数量明显少于其他类别（García等，2018）。通常情况下绝大多数分类器在对类别不平衡数据分类时都会出现一定程度的性能损失（Díez-Pastor等，2015；张永清 等，2020）。或者说分类器虽然获得了较高的总体分类精度OA（Overall Accuracy），却忽略了少数类样本的识别。

不平衡数据一般包括两种类型：二类和多类。类别间样本分布不均、多种类别重叠和样本噪声等问题，使得多类不平衡数据相对二类数据更加难以处理（Krawczyk，2016）。多类不平衡分类问

收稿日期： 2020-06-27；**预印本：** 2021-01-16

基金项目： 国家自然科学基金(编号:61772397,12005169,62201438)；陕西省自然科学基础研究计划(编号:2021JC-23)；榆林市科技局科技发展专项(编号:CXY-2020-094)；陕西林业科技创新重点专项(编号:SXLK2022-02-8)

第一作者简介： 童莹萍，研究方向为遥感图像处理。E-mail:yptong@stu.xidian.edu.cn

通信作者简介： 冯伟，研究方向为遥感图像处理、机器学习。E-mail:wfeng@xidian.edu.cn

题的理想处理结果是：一个分类器可以在不牺牲多数类别分类准确性的情况下，提高少数类的分类精度（Feng和Bao，2017；Feng等，2019a）。目前，研究学者针对以上问题已经提供了很多方法，例如，Jimenez-Castaño等（2020）优化了孪生支持向量机算法，使用高斯相似度并结合基于中心核比对的方法改善数据可分类性，用于处理不平衡数据；Arshad等（2019）提出了一种半监督深度模糊C均值聚类算法，提取新特征以控制冗余；Douzas等（2018）将K均值聚类算法与合成少数类过采样算法SMOTE（Synthetic Minority Oversampling Technique）相结合，克服其他过采样算法的缺点，消除类间不平衡和类内不平衡同时避免产生噪声样本，其中，SMOTE算法是对少数类别样本进行分析和模拟，并将人工模拟的新样本添加至数据集中，从而调节数据中类别的不平衡比率。但是大多数方法都集中于解决二类不平衡问题，不适用于多类情况（才子昕 等，2019）。

数据采样是常用的不平衡高光谱数据处理方法。典型的采样算法包括随机欠采样RUS（Random Undersampling）和随机过采样ROS（Random Oversampling）。RUS算法通过随机缩小多数类别的样本数量使样本达到平衡，但是RUS算法会导致重要信息的丢失；ROS算法的作用则与RUS完全相反，它是在少数类别中通过随机生成更多样本，直到所有类别达到平衡，然而ROS算法虽然可以克服RUS的缺陷，但是会引起算法的过拟合。

SMOTE是ROS的升级算法，它可以有效地避免过拟合现象，已被用于解决机器学习中的类别不平衡问题，例如，Bhagat和Patil（2015）首先使用SMOTE来平衡数据中各类别的样本数量，然后利用随机森林进行分类；Bandara等（2020）也是先通过SMOTE算法平衡数据集，然后构建滑坡敏感性模型，最后比较了随机森林与旋转森林算法在平衡数据情况下的性能表现。以上方法都可以在一定程度上提高分类器对小类别样本的分类精度，然而，这些技术仍侧重解决二类不平衡问题，在面临复杂的多类不平衡问题时难以拓展。在高光谱图像分类中，Cai和Zhang（2019）将SMOTE算法应用于Pavia University数据集上，有效解决高光谱数据不平衡问题，Zhou等（2020）通过SMOTE合成少数类别的人工样本，解决了多类情况下样本不平衡的问题。但是，大多数研究只将SMOTE算法应用于数据预处理阶段。由于SMOTE算法在训练样本时具有引入额外噪声的风险（Pan等，2020；Elreedy和Atiya，2019），一次性平衡样本可能导致产生的噪声无法修正，进而影响分类模型的训练效果。因此，如何改进SMOTE算法解决复杂的多类不平衡问题面临着巨大的挑战。

集成学习已经成功应用于高光谱图像分类（Mullick等，2020；Feng等，2019b）。旋转森林ROF（Rotation Forest）是一种功能强大的集成分类器（Tu等，2020；Rodriguez等，2006），引起了广泛关注。该算法通过结合多种随机因子来提升自身的学习性能，这些随机因子包括：数据随机选择、特征随机选择和特征空间旋转等。RoF算法的传统构造偏向于对多数类别进行分类，而忽略少数类样本的识别。目前虽然已有很多改进的RoF算法，但是仍然普遍适用于平衡数据集的学习。因此，改进RoF算法，使之适合不同比例的多类不平衡数据是一个非常有意义的研究方向。

本文提出了一个新的SMOTE和RoF的自适应内部集成算法。该算法利用动态采样因子技术，将类别分布优化和基分类器过程进行融合，并可以有效地解决高光谱数据中的多类不平衡问题。

2 RoF分类算法

RoF算法是一种随机森林算法衍生而来的集成分类器，它通过特征空间变换，可以更加有效地提高基分类器的多样性（Feng 等，2019a）。设$\{\boldsymbol{X},\ \boldsymbol{Y}\}=\{(\boldsymbol{x}_1,\ y_1),\ \cdots,\ (\boldsymbol{x}_N,\ y_N)\}$为训练样本集，其中$\boldsymbol{x}_i=[x_{i,\ 1},\ \cdots,\ x_{i,\ n}]\in\mathbb{R}^n$为特征为$\boldsymbol{F}$的训练样本，$\boldsymbol{Y}$为带有样本类别的向量，$y_j$为类别标签$\boldsymbol{\omega}=\{\omega_1,\ \cdots,\ \omega_L\}$的集合中的元素（$L$为样本的类别总数）。设$D_1,\ \cdots,\ D_T$为集成算法中的基分类器，其中$T$为分类器的数量。

为构造分类器D_i，首先将具有n个特征的$\boldsymbol{F}$随机分成H个互不相交的子集。假设H是n的因数，每个特征子集都包含$M=n/H$个特征，用$\boldsymbol{F}_{i,\ j}$表示用于训练分类器D_i的第j个特征子集，由$\boldsymbol{F}_{i,\ j}$组成的样本集合表示为$\boldsymbol{Z}$。其次，对$\boldsymbol{Z}$执行75%的样本随机采样，并对选中的样本集合进行主成分分析PCA（Principal Component Analysis）；然后保存主成分分析的系数$\boldsymbol{a}_{i,\ j}^{(1)},\ \cdots,\ \boldsymbol{a}_{i,\ j}^{(M_j)}$，每个系数的大小为$M\times 1$；最后将主成分分析的系数组成旋转矩阵$\boldsymbol{R}_i$。$\boldsymbol{R}_i$可以表示为

$$\boldsymbol{R}_i=\begin{bmatrix}\boldsymbol{a}_{i,1}^{(1)},\boldsymbol{a}_{i,1}^{(2)},\cdots,\boldsymbol{a}_{i,1}^{(M_1)} & [0] & \cdots & [0]\\ [0] & \boldsymbol{a}_{i,2}^{(1)},\boldsymbol{a}_{i,2}^{(2)},\cdots,\boldsymbol{a}_{i,2}^{(M_2)} & \cdots & [0]\\ \vdots & \vdots & \ddots & \vdots\\ [0] & [0] & \cdots & \boldsymbol{a}_{i,H}^{(1)},\boldsymbol{a}_{i,H}^{(2)},\cdots,\boldsymbol{a}_{i,H}^{(M_H)}\end{bmatrix} \tag{1}$$

式中，旋转矩阵的维度为$n\times\sum_j M_j$。RoF算法根据原始数据的特征$\boldsymbol{F}$对$\boldsymbol{R}_i$的列进行重新排序，获得新的旋转矩阵$\boldsymbol{R}_i^a$，其大小为$N\times n$。然后由$\boldsymbol{X}\boldsymbol{R}_i^a$生成一个新的训练样本集合，并训练一个分类器$D_i$。

3 SMOTE采样算法

SMOTE算法（Rodríguez 等，2020）是处理不平衡数据问题比较常用的算法。该算法步骤如图1所示，主要包括以下过程：对于少数类的样本$\boldsymbol{x}_i$，选择其K个最近邻的少数类样本$\bar{\boldsymbol{x}}$，在$\boldsymbol{x}_i$与$\bar{\boldsymbol{x}}$的连接线上插入新的合成少数类样本$\boldsymbol{x}_{\text{sym}}$，新的合成样本$\boldsymbol{x}_{\text{sym}}$可以表示为

$$\boldsymbol{x}_{\text{sym}}=\boldsymbol{x}_i+\alpha(\bar{\boldsymbol{x}}-\boldsymbol{x}_i) \tag{2}$$

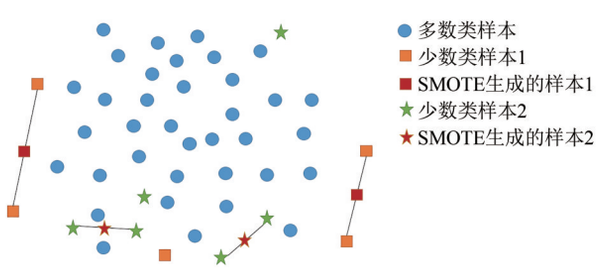


图1 SMOTE算法示意图

Fig.1 SMOTE algorithm diagram

式中，α是［0，1］范围内的随机变量。

为了研究SMOTE算法产生的人工样本与真实样本之间的区别，以下对这两种样本的均值和协方差矩阵进行比较分析。首先令$\Delta=\bar{\boldsymbol{x}}-\boldsymbol{x}_i$，新的合成样本$\boldsymbol{x}_{sym}=\boldsymbol{x}_i+\alpha\Delta$，其中$\alpha$是在$[0,\alpha^*]$中随机生成的，参数$\alpha^*$的值通常大于或等于1，这使得算法能够在连接样本$\boldsymbol{x}_i$和其随机选择的最近邻$\bar{\boldsymbol{x}}$的连线上进行外推或内插。

合成样本$\boldsymbol{x}_{\text{sym}}$的方差由以下公式给出：

$$E[\boldsymbol{x}_{\text{sym}}]\approx\mu_{\boldsymbol{x}_i}+\frac{C\alpha^*}{2}\int_{\boldsymbol{x}_i}p(\boldsymbol{x}_i)^{\frac{-2}{d}}\frac{\partial p(\boldsymbol{x}_i)}{\partial x}d\boldsymbol{x}_i \tag{3}$$

式中，$\left[\frac{\partial p(\boldsymbol{x})}{\partial x}\right]^{\mathrm{T}}=\left(\frac{\partial p(\boldsymbol{x})}{\partial x_1},\cdots,\frac{\partial p(\boldsymbol{x})}{\partial x_d}\right)$，$d$是向量的维度，$N$是所考虑的样本原始的数量，$p(\boldsymbol{x}_i)$是样本$\boldsymbol{x}_i$的条件密度函数，$\boldsymbol{I}$是单位矩阵，$C$的计算公式如下：

$$C=\frac{N!\Gamma(1+\frac{2}{d})^{\frac{2}{d}}\Gamma(K+\frac{2}{d}+1)}{\pi K!(d+2)\Gamma(N+\frac{2}{d}+1)} \tag{4}$$

式中，K为少数类样本的最近邻样本数，Γ为伽马函数。

合成样本$\boldsymbol{x}_{sym}$的协方差矩阵由以下公式给出：

$$\sum_{\boldsymbol{x}_{sym}}=\sum_{\boldsymbol{x}_i}+\frac{C\alpha^{*2}}{3}\int_{\boldsymbol{x}_i}\frac{p(\boldsymbol{x}_i)^{1-\frac{2}{d}}d\boldsymbol{x}_i\boldsymbol{I}}{-\frac{C^2\alpha^{*2}}{3}}\int_{\boldsymbol{x}_i}p(\boldsymbol{x}_i)^{\frac{-2}{d}}\frac{\partial p(\boldsymbol{x}_i)}{\partial x}d\boldsymbol{x}_i\int_{\boldsymbol{x}_i}p(\boldsymbol{x}_i)^{\frac{-2}{d}}\frac{\partial p(\boldsymbol{x}_i)^{\mathrm{T}}}{\partial x}d\boldsymbol{x}_i+$$
$$\frac{C\alpha^*}{2}\left(\int_{\boldsymbol{x}_i}p(\boldsymbol{x}_i)^{\frac{-2}{d}}\frac{\partial p(\boldsymbol{x}_i)}{\partial x}\left(\left(\boldsymbol{x}_i-\mu_{\boldsymbol{x}_i}\right)^{\mathrm{T}}\right)d\boldsymbol{x}_i+\int_{\boldsymbol{x}_i}p(\boldsymbol{x}_i)^{\frac{-2}{d}}\left(\boldsymbol{x}_i-\mu_{\boldsymbol{x}_i}\right)\frac{\partial p(\boldsymbol{x}_i)^{\mathrm{T}}}{\partial x}d\boldsymbol{x}_i\right) \tag{5}$$

式中，$\mu_{\boldsymbol{x}_i}$是真实样本的均值，$\sum_{\boldsymbol{x}_i}$是真实样本的协方差函数。若样本的概率密度为多元高斯分布，则式（3）与式（5）可以进一步简化为

$$E[\boldsymbol{x}_{sym}]\approx\mu_{\boldsymbol{x}_i} \tag{6}$$

$$\Sigma_{\boldsymbol{x}_{\text{sym}}}=\Sigma_{\boldsymbol{x}_i}+\left((2\pi)^{\frac{1-d}{2}}\frac{C\alpha^{*2}}{3}\det{}^{\frac{1-d}{2d}}(\Sigma_{\boldsymbol{x}_i})\left(\frac{d}{2d-1}\right)^{\frac{d}{2}}\right)\boldsymbol{I}+$$
$$\left(-2\pi C\alpha^*\det{}^{\frac{1}{d}}\left(\Sigma_{\boldsymbol{x}_i}\right)\left(\frac{d}{d-2}\right)^{\frac{d+2}{2}}\right)\boldsymbol{I} \tag{7}$$

式中，det表示行列式。式（7）对于维度$d>2$是有效的。对于任何$d>2$，$\frac{d}{d-2}$都大于1，因此生成的样本协方差矩阵的第三项将为负值。对于标准SMOTE算法，$\alpha^*=1$，式（7）第二项的值小于第三项。因此，由SMOTE生成的样本协方差矩阵将比原始少数类样本的协方差矩阵更为收缩。

上述证明了SMOTE生成的样本均值接近于真实样本的均值，两者的协方差矩阵之间存在一些差异。

4 SMOTE和RoF动态集成算法

本文提出了一种新的SMOTE和RoF动态集成算法。该算法使用样本权重和动态采样因子来降低SMOTE算法引入人工噪声的风险，并与RoF算法进行内部结合，最终用于多类不平衡高光谱数据学习问题。

4.1 样本重要性权重函数

本文提出一个新的样本重要性权重函数，来缓存SMOTE产生额外噪声的影响。该权重函数的概念源于集成间隔理论。在间隔理论中，对于给定的样本，它的间隔值越低，说明它包含的信息量越多，越容易获得更多分类器的关注（Zhou和Guo，2019）。

假设训练样本为$(\boldsymbol{x}, y)$，$h_t(\boldsymbol{x})$是由T个分类器预测的标签。$\theta \mapsto \delta(\theta)$是根据$\theta$是否为真将$\theta$映射为1或0的函数。$v(\boldsymbol{x}, c)=\sum_{t=1}^{T}\delta\left(h_t(\boldsymbol{x})=c\right)$为预测样本$\boldsymbol{x}$的标签为$c$的分类器个数。则样本$\boldsymbol{x}$的集成间隔值可以表示为：

$$\operatorname{margin}(\boldsymbol{x},y)=\frac{1}{T}\left(v(\boldsymbol{x},y)-\max_{c\neq y}v(\boldsymbol{x},c)\right) \tag{8}$$

样本的权重由以下公式定义：

$$W(\boldsymbol{x},y)=1-\left|\operatorname{margin}(\boldsymbol{x},y)\right| \tag{9}$$

对本文提出的样本重要性权重函数进行数学推导，式也可以写成：

$$W(\boldsymbol{x},y)=1-\frac{1}{T}\left(v\left(\boldsymbol{x},\tilde{h}_1(\boldsymbol{x})\right)-\min\left(v(\boldsymbol{x},y),v\left(\boldsymbol{x},\tilde{h}_2(\boldsymbol{x})\right)\right)\right) \tag{10}$$

式中，$\tilde{h}_1(\boldsymbol{x})$和$\tilde{h}_2(\boldsymbol{x})$是分类器数量最多的两个标签，$\tilde{h}_1(\boldsymbol{x})=\underset{c}{\operatorname{arcmax}}\, v(\boldsymbol{x}, c)$，$\tilde{h}_2(\boldsymbol{x})=\underset{c\neq\tilde{h}_1(\boldsymbol{x})}{\operatorname{arcmax}}\, v(\boldsymbol{x}, c)$。

统计预测样本$\boldsymbol{x}$获得每个类别的分类器数量，并对其进行降序排列，那么$W(\boldsymbol{x}, y)$可以看作是两个数量最多分类器的归一化差异。

分两步对式（9）进行证明。首先，假设样本真正的标签是由多个分类器投票选定，$\tilde{h}_1(\boldsymbol{x})=y$，那么$W(\boldsymbol{x}, y)=1-\left(v\left(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x})\right)-v\left(\boldsymbol{x}, \tilde{h}_2(\boldsymbol{x})\right)\right)$，由于$v(\boldsymbol{x}, \tilde{h}_2(\boldsymbol{x}))\leqslant v(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x}))=v(\boldsymbol{x}, y)$，式（9）中绝对值的符号没有改变。

若假设分类器对于样本$\boldsymbol{x}$的预测值与真实值不同，$\tilde{h}_1(\boldsymbol{x})\neq y$，那么$W(\boldsymbol{x}, y)=1-\left|v(\boldsymbol{x}, y)-v\left(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x})\right)\right|$。由于$v\left(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x})\right)\geqslant\max_{c} v(\boldsymbol{x}, c)$，样本的权重为$W(\boldsymbol{x}, y)=1-\left(v\left(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x})\right)-v(\boldsymbol{x}, y)\right)$。这意味着式（9）的margin值符号发生改变，即样本的权重为$W(\boldsymbol{x}, y)=1-\frac{1}{T}\left(v\left(\boldsymbol{x}, \tilde{h}_1(\boldsymbol{x})\right)-\min\left(v(\boldsymbol{x}, y), v\left(\boldsymbol{x}, \tilde{h}_2(\boldsymbol{x})\right)\right)\right)$。

综上，训练样本的自适应权重函数$W(\boldsymbol{x}_i)$可由以下公式计算：

$$W(\boldsymbol{x}_i)=1-\frac{1}{\sum_{c=1}^{L}(v_c)}\left|v(\boldsymbol{x}_i,y_i)-\max_{c\neq y_i}v(\boldsymbol{x}_i,c)\right|=1-\frac{1}{T}\left|v(\boldsymbol{x}_i,y_i)-\max_{c\neq y_i}v(\boldsymbol{x}_i,c)\right| \tag{11}$$

式中，$v(\boldsymbol{x}_i, y_i)$是真实类别y的投票数，$v(\boldsymbol{x}_i, c)$是其他类别c的投票数。样本具有较大的权重值$W(\boldsymbol{x}_i)$，表明这个样本越接近于当前的分类决策边界，并且具有较高的概率被选中参与SMOTE算法，为下一个分类器的训练提供新的类别平衡的训练样本集。因此，随着基分类器数量的增加，分布在决策边界的样本权重会迭代增加。

4.2 SMOTE和RoF动态集成算法

本文提出的SMOTE和RoF动态集成算法是一种数据采样与集成框架内部结合的算法，它不仅可以降低初始数据的不平衡率IR（Imbalance Ratio），还能保证训练集中样本的数量和质量。假设数据集$\boldsymbol{S}$中有L个类别，第i个类别的样本大小为N_i，将每个类别包含的样本集根据其样本数量N_i进行降序排列，则N_1为最大类的训练样本的数量，N_L为最小类别L的样本数。将$\boldsymbol{S}$的每个样本权重值W初始化为$1/N$。在每次迭代t中，以N_L为标准，对其他所有类别$c(c=2, \cdots, L)$利用重采样率$\omega\%$执行有放回随机采样，选取$\omega\%\cdot N_1$个样本，然后通过SMOTE算法生成剩下的$(1-\omega\%)\cdot N_1$个样本。最后通过将N_L、采样样本以及SMOTE产生的样本进行组合，获得各类别数量平衡的训练样本集$\boldsymbol{S}_t$。然后与传统RoF算法一样，对数据集$\boldsymbol{S}_t$进行特征空间旋转，以获得最终的训练数据集$\boldsymbol{S}_t'$，并训练一个分类器D_t。根据式（9）更新所有初始训练样本的权重值W，并更新重采样率$\omega\%$。重复以上步骤，直到t达到最大迭代次数。最终的预测结果通过利用最大投票规则将所有的分类器结果进行融合获得。本文算法的伪代码描述如下：

输入：训练样本集$\boldsymbol{S}$，特征集合$\boldsymbol{F}$，类别数L，N_i为第i个类别的样本数量，$\omega\%$为重采样率，H为随机分解的特征子集数量，D为基分类器，T为基分类器的数量，$E=\varnothing$为集成分类模型。

初始化：将所有$\boldsymbol{x}_i \in \boldsymbol{S}$的权重初始化为$W_1(\boldsymbol{x}_i) = 1/N$。

将$\boldsymbol{S}$中的L个类别按照包含的样本数量N_i进行降序排列。N_1为最大类的训练样本的数量，N_L为最小类别L的样本数。

1. for t=1：T

2. 定义$\boldsymbol{S}_{1,\,t}$为包含N_L个样本的集合

3. for c=2：T

4. 对类别c中的所有样本，根据W_t执行数量为$\omega\% \cdot N_1$的权重采样，获得样本集$\boldsymbol{S}_{c,\,t}$

5. 利用SMOTE算法，生成样本个数为$(1-\omega\%) \cdot N_1$的数据集$\boldsymbol{S}'_{c,\,t}$

6. end for

7. 将采样数据集$\boldsymbol{S}_{c,\,t}(c = 1, \cdots, L)$与人工生成的数据集$\boldsymbol{S}'_{c,\,t}(c = 2, \cdots, L)$进行结合，生成一个新的平衡数据集$\boldsymbol{S}_t$

8. 对数据集$\boldsymbol{S}_t$的特征集$\boldsymbol{F}$进行随机分配，得到H个不相交子集$\boldsymbol{F}_{t,\,h}$

9. for h=1：H

10. 将$\boldsymbol{F}_{t,\,h}$的特征组成$\boldsymbol{S}_{t,\,h}$，并将PCA应用于$\boldsymbol{S}_{t,\,h}$获得系数$c_{t,\,h}$

11. end for

12. $c_{t,\,h}$构成矩阵M_t，以原始数据特征$\boldsymbol{F}$为标准，对M_t的列进行重新排序，得到旋转矩阵M'_t，构建训练集$\boldsymbol{S}'_t = \left[\boldsymbol{S}_t \cdot M'_t,\ \boldsymbol{Y}_t\right]$

13. 用$\boldsymbol{S}'_t$训练一个分类器D_t

14. $\boldsymbol{E} \leftarrow \boldsymbol{E} \cup D_t$

15. 改变采样率$\omega\%$

16. for $\boldsymbol{x}_i \in \boldsymbol{S}$

17. 更新$W_t(\boldsymbol{x}_i)$

18. $W_{t+1}(\boldsymbol{x}_i) \leftarrow W_t(\boldsymbol{x}_i)$

19. end for

20. end for

21. 输出：集成模型$\boldsymbol{E}$

本文提出的算法利用动态采样为不同的基分类器提供具有多样性特征的训练样本集合。在使用SMOTE算法之前，以不同的采样率$\omega\%$对所有少数类样本进行随机采样。以下示例用于详细说明$\omega\%$的更新过程。

若设置采样$\omega\%$的范围为10%至100%，对于第一个分类器，$\omega\%$等于10%；第二个分类器$\omega\%$更新为20%，然后以此类推。当$\omega\%$等于100%时，对于类别$\omega\%$执行是完全的随机过采样；若集成模型尺寸$\omega\%$为30，那么每10个分类器，$\omega\%$执行3次从10%至100%的迭代更新。

5 实验结果

5.1 实验设计

由于过采样算法在解决不平衡问题时应用更为广泛，本实验主要选择ROS和SMOTE这两种过采样算法作为对比。为了验证本文算法在不平衡数据集上的分类优势，本文选用4种对比算法，分别是随机森林RF（Random Forest）、传统的RoF、ROS预处理与RoF结合的算法（记为ROSRoF）以及SMOTE预处理与RoF结合的算法（记为SMOTERoF）。

实验利用分类回归树CART（Classification and Regression Tree，）作为基分类器，所有集成模型由30个决策树组成。对于所有的RoF算法，参数H都设置为30；数据采样率$\omega\%$的范围设置为10%至100%；实验部分所有值均为将算法独立运行10次所得的结果均值。

5.2 评价指标

遥感图像有侧重于不同方面的评估指标（高连如 等，2007），本实验使用8种不同的比较方法对分类结果进行评估。利用平均分类精度AA（Average Accuracy）、总体分类精度OA（Overall Accuracy）、F-measure、Gmean和最小召回率（Minimum Recall）以及运算时间对算法进行测试。为了验证算法在集成框架提升方面的优势，集成分类器多样性（Feng等，2018）被用来作为集成算法的评估方法；另外，本实验使用McNemar非参数成对检验来比较本文算法与其他算法在统计意义上的性能差异；最后，为了直观地对比各个算法的分类表现，绘制各算法在高光谱图像上的分类结果，并将其与地面参考数据作对比。

5.3 实验数据

为了评估本文算法的有效性，本实验选用3个

公开的高光谱图像作为测试数据，它们分别是Indian Pines、Salinas以及Pavia University，数据集参数如表1所示。其中，Indian Pines图像具有20 m空间分辨率；Salinas图像空间分辨率为3.7 m，包含16个不同的类别；Pavia University图像波长范围为0.43—0.86 μm，空间分辨率高达1.3 m。

表1 实验数据集

Table 1 Experimental data set

数据集	Indian Pines	Salinas	Pavia University
像素	145×145	512×217	610×340
特征数	200	204	103
类别数	16	16	9

为了验证本文算法的分类表现，本实验采用了IR值不同的数据集。在Salinas图像数据以及Pavia University图像数据中，本实验采用无放回随机抽取方法，选择5%的原始数据构成训练样本，其余95%作为测试样本；而Indian Pines图像数据由于不平衡比率高，更有利于本实验设计，进而生成不同IR值的样本。

在Indian Pines图像数据中，构造4个数据集，方法是：对于多数类别样本，分别随机抽取原始数据的5%、10%、15%和20%以构造训练数据集，其它未被选择的样本作为测试样本；对于少数类别1、7、9、16，由于样本量过少，从每类中选取50%样本作为训练样本，剩下的作为测试样本。训练集的IR通过N_l/N_L计算得到；6个数据集的IR分别为12.20、17.50、36.80、49.10、12.51和19.83；本实验中所有训练样本及测试样本的具体信息由表2给出。

表2 训练样本与测试样本

Table 2 Training sets and test sets

类别	Indian Pines								Salinas（IR：12.51）		Pavia University（IR：19.83）	
	样本1（IR：12.20）		样本2（IR：17.50）		样本3（IR：36.80）		样本4（IR：41.90）					
	训练样本	测试样本	训练样本	测试样本	训练样本	测试样本	训练样本	测试样本	训练样本	测试样本	训练样本	测试样本
1	23	23	23	23	23	23	23	23	100	1909	331	6300
2	71	1357	142	1286	214	1214	285	1143	186	3540	932	17717
3	41	789	83	747	124	706	166	664	98	1878	104	1995
4	11	226	23	214	35	202	47	190	69	1325	153	2911
5	24	459	48	435	72	411	96	387	133	2545	67	1278
6	36	694	73	657	109	621	146	584	197	3762	251	4778
7	14	14	14	14	14	14	14	14	178	3401	66	1264
8	23	455	47	431	71	407	95	383	563	10708	184	3498
9	10	10	10	10	10	10	10	10	310	5893	47	900
10	48	924	97	875	145	827	194	778	163	3115	—	—
11	122	2333	245	2210	368	2087	491	1964	53	1015	—	—
12	29	564	59	534	88	505	118	475	96	1831	—	—
13	10	195	20	185	30	175	41	164	45	871	—	—
14	63	1202	126	1139	189	1076	253	1012	53	1017	—	—
15	19	367	38	348	57	329	77	309	363	6905	—	—
16	46	47	46	47	46	47	46	47	90	1717	—	—

5.4 结果分析

本文算法与对比算法在6个不平衡高光谱图像数据上的分类结果如表3和表4所示，F-measure、Gmean和最小召回率结果如表5所示。每组数据中性能最佳的结果已用粗体标出。

由表3和表4可以看出，与大多数现有文献的结论不同的是，RF的高光谱图像分类结果在统计意义上优于传统RoF的分类结果。传统的RoF侧重于多类样本的分类，而牺牲少数类样本的分类准确率（Breiman，2001）；尤其是当样本不平衡率增加时，传统RoF算法完全忽略了少数类样本的分类精度（Ghosh和Cabrera，2021）。此外，针对不平衡数据，本文算法与这两种未经数据预处理的分类算法相比始终可以达到最佳的分类结果。

表3 Salinas与Pavia University分类精度

Table 3 Salinas and Pavia University classification accuracy

/%

类别	Salinas					Pavia University				
	RF	RoF	ROSRoF	SMOTERoF	本文算法	RF	RoF	ROSRoF	SMOTERoF	本文算法
1	99.22	99.14	**99.64**	99.60	99.47	90.06	91.68	68.29	72.81	**94.10**
2	99.73	99.25	99.86	**99.93**	99.67	97.39	**97.95**	72.04	80.44	97.10
3	97.81	97.50	98.86	**99.18**	99.13	59.37	41.20	69.85	71.73	**74.96**
4	98.70	99.20	**99.55**	99.54	99.37	85.02	77.00	**97.28**	94.62	92.31
5	97.32	97.05	96.19	96.36	**98.55**	98.82	98.59	99.35	99.41	**99.76**
6	99.34	99.50	**99.89**	99.88	99.66	55.84	62.81	**88.82**	87.89	87.63
7	98.19	96.47	98.83	98.59	**99.44**	67.40	2.31	**91.30**	88.86	75.88
8	83.76	90.21	82.52	80.57	**91.34**	85.81	92.28	78.53	80.82	**88.36**
9	98.54	98.82	99.10	99.04	**99.18**	99.93	99.96	99.89	99.94	**99.99**
10	90.54	90.57	92.19	92.46	**96.94**	—	—	—	—	—
11	94.95	92.55	95.69	95.72	**97.22**	—	—	—	—	—
12	98.68	99.75	99.97	99.94	**99.98**	—	—	—	—	—
13	97.01	96.98	**98.15**	98.03	97.01	—	—	—	—	—
14	92.93	**94.08**	93.04	93.74	94.06	—	—	—	—	—
15	62.49	51.37	61.35	64.73	**65.28**	—	—	—	—	—
16	97.15	96.89	98.33	**98.60**	97.94	—	—	—	—	—
AA	94.15	93.71	94.57	94.74	**95.89**	82.18	73.75	85.04	86.28	**90.01**
OA	89.95	89.69	89.93	90.02	**92.74**	86.79	85.16	77.77	82.04	**92.83**

注：表格中加粗的数字表示每组数据中性能最佳的结果。

表4 Indian Pines分类精度

Table 4 Indian Pines classification accuracy

/%

类别	样本1(IR:12.20)					样本2(IR:17.50)				
	RF	RoF	ROS RoF	SMOTERoF	本文算法	RF	RoF	ROSRoF	SMOTERoF	本文算法
1	87.83	**91.30**	**91.30**	**91.30**	**91.30**	66.09	83.91	90.87	**94.35**	91.30
2	55.97	67.57	67.64	66.28	**74.12**	64.61	78.87	70.00	70.48	**81.66**
3	43.26	60.00	60.86	**69.01**	64.12	53.68	56.48	63.20	65.03	**69.65**
4	24.60	5.75	**80.00**	73.89	62.92	31.78	4.11	78.93	**80.89**	74.67
5	79.00	89.32	86.73	83.59	**90.15**	83.79	88.23	86.34	87.01	**90.32**
6	88.89	96.89	95.76	92.52	**96.96**	95.68	96.33	95.94	95.08	**96.86**
7	92.14	96.43	97.14	**97.86**	97.14	70.00	75.00	**88.57**	86.43	82.14
8	86.70	94.97	94.59	94.88	**96.13**	95.59	98.68	96.24	97.77	**99.21**
9	31.00	93.00	**94.00**	92.00	91.00	67.00	0.00	**100.00**	**100.00**	**100.00**
10	61.92	71.79	79.07	79.42	**82.73**	67.19	71.85	81.35	79.85	**85.23**
11	77.75	**86.64**	54.02	48.24	80.45	85.27	**91.19**	48.06	43.13	84.61
12	33.30	35.66	66.37	57.00	**66.44**	43.67	29.34	71.14	72.85	**81.40**
13	91.13	**97.74**	96.26	94.51	96.15	91.19	97.51	97.51	96.32	**99.19**
14	91.61	**97.56**	83.87	82.64	95.51	94.82	96.37	93.41	91.94	**96.67**
15	24.22	37.66	48.96	47.25	**49.51**	39.45	47.44	59.34	58.13	**64.20**
16	**98.72**	97.87	97.87	97.87	97.87	98.30	**100.00**	**100.00**	99.57	**100.00**
AA	66.75	76.26	80.90	79.27	**83.28**	71.76	69.71	82.56	82.43	**87.32**
OA	67.86	76.56	71.46	69.26	**80.76**	75.09	78.73	72.71	71.53	**85.58**

续表

类别	样本3(IR:36.80)					样本4(IR:49.10)				
	RF	RoF	ROS RoF	SMOTERoF	本文算法	RF	RoF	ROSRoF	SMOTERoF	本文算法
1	80.87	48.26	90.43	**91.30**	**91.30**	81.30	0.00	99.57	**100.00**	96.09
2	68.29	61.74	63.18	64.92	**81.89**	70.20	69.54	64.53	65.08	**83.65**
3	54.50	50.50	53.19	60.69	**76.59**	55.69	47.53	49.98	55.50	**78.33**
4	34.26	0.00	85.99	**89.70**	85.25	44.42	0.00	**90.74**	86.32	88.63
5	82.99	85.40	82.34	87.66	**92.87**	85.35	89.43	87.83	86.02	**93.59**
6	95.02	97.54	95.12	92.09	**97.10**	93.61	97.62	92.62	92.96	**97.89**
7	74.29	3.57	**92.86**	**92.86**	92.14	57.86	0.00	97.14	**99.29**	96.43
8	98.60	**99.93**	98.08	92.14	99.63	98.09	99.16	95.67	96.50	**99.56**
9	55.00	0.00	**100.00**	97.00	97.00	64.00	0.00	98.00	**100.00**	95.00
10	65.41	72.15	77.11	78.38	**81.46**	69.18	74.23	78.10	77.94	**86.09**
11	83.85	**86.99**	53.94	47.92	85.17	88.27	**88.52**	57.96	57.22	87.00
12	51.90	53.49	76.75	80.61	**82.97**	53.62	35.01	79.79	81.43	**88.61**
13	94.74	99.14	99.37	99.26	**99.43**	96.83	98.84	98.41	98.66	**99.39**
14	94.96	**96.38**	94.65	92.11	96.12	94.78	95.96	86.80	86.05	**95.46**
15	44.29	46.78	50.09	53.89	**64.13**	49.48	48.51	55.60	57.96	**71.65**
16	94.04	94.89	95.11	96.17	**97.45**	99.79	96.81	**100.00**	99.79	**100.00**
AA	73.31	62.30	81.76	82.29	**88.78**	75.16	69.71	83.30	83.79	**91.09**
OA	76.04	75.94	72.12	71.54	**86.39**	78.40	78.73	72.65	72.96	**88.40**

注：表格中加粗的数字表示每组数据中性能最佳的结果。

表5 不同算法的F-measure、Gmean和最小召回率

Table 5 F-measure, Gmean and minimum recall of each algorithm

样本		F-measure					Gmean					最小召回率				
		RF	RoF	ROS RoF	SMOT ERoF	本文算法	RF	RoF	ROS RoF	SMOT ERoF	本文算法	RF	RoF	ROS RoF	SMOT ERoF	本文算法
IndianPines	样本1	64.00	75.87	73.35	72.56	**79.92**	59.78	58.06	79.10	77.18	**81.71**	20.55	5.75	48.70	45.47	**49.51**
	样本2	70.85	74.30	75.28	75.30	**85.54**	68.12	0.00	80.85	80.52	**86.57**	31.78	0.00	48.06	43.13	**64.20**
	样本3	75.72	65.37	76.80	75.63	**87.77**	70.10	0.00	79.67	80.41	**88.21**	34.26	0.00	48.36	46.75	**64.13**
	样本4	75.50	74.30	76.11	76.68	**90.02**	72.05	0.00	81.35	82.08	**90.69**	39.86	0.00	49.98	54.95	**71.65**
Salinas		93.90	93.27	94.15	94.32	**95.99**	93.61	92.75	93.95	94.22	**95.45**	62.49	51.37	61.35	64.73	**65.28**
PaviaUniversity		84.75	77.50	80.46	82.54	**91.26**	80.45	18.15	84.12	85.69	**89.55**	55.84	2.31	68.18	71.60	**74.88**

注：表格中加粗的数字表示每组数据中性能最佳的结果。

通过ROS或SMOTE算法对数据预处理，能起到平衡数据集的效果，在不同样本上ROSRoF与SMOTERoF在AA方面较传统RoF都有明显的提升。由于本文算法将类别平衡融合于分类过程中，本文算法取得的分类结果与SMOTE算法作为预处理的分类结果相比更有优势。

本文算法不仅对少数类样本的分类有帮助，而且在很大程度上提高了总体分类精度。对于Indian Pines数据，在不平衡率较低的样本1中，本文所提出的算法与RF和传统RoF相比在AA方面分别提高了16.53%和7.02%，在OA方面分别提升了12.90%和4.20%，与ROSRoF和SMOTERoF相比在AA上分别提高了2.38%和4.01%，在OA上分别提高了9.3%和11.5%。

而在不平衡率较高的样本4中，本文算法与其他算法相比分类精度也有明显提升，说明本文算法在不平衡率改变的情况下同样具有很好的性能。另外对于F-measure、Gmean和最小召回率这3个评价指标，本文算法的分类结果明显优于其他分类算法。各算法的样本多样性和运行时间分别由表6和表7给出。本文算法利用动态采样技术，因此在集成分类器多样性上本文算法的性能显著提升。

表6 算法的样本多样性

Table 6 Sample diversity of each algorithm

算法	Indian Pines				Salinas	Pavia University
	样本1	样本2	样本3	样本4		
RF	0.1627	0.1640	0.1522	0.1562	0.0806	0.1069
RoF	0.1340	0.1317	0.1104	0.1079	0.0605	0.0705
ROSRoF	0.1596	0.1480	0.1401	0.1478	0.0622	0.0910
SMOTERoF	0.1781	0.1515	0.1530	0.1575	0.0627	0.1088
本文算法	0.1931	0.1828	0.1754	0.1751	0.0897	0.1172

表7 算法运行时间

Table 7 Run time of each algorithm

/s

算法	Indian Pines				Salinas	Pavia University
	样本1	样本2	样本3	样本4		
RF	0.40	0.64	0.90	1.22	2.02	1.53
RoF	37.70	60.23	91.59	117.32	269.63	74.20
ROSRoF	47.67	95.01	135.56	188.32	340.70	80.13
SMOTERoF	88.82	183.74	275.17	355.93	607.88	139.31
本文算法	65.35	86.08	155.00	173.01	389.89	230.90

在$p < 0.05$的情况下，McNemar的测试值超过1.96，就意味着两种算法之间存在显著差异性。McNemar的测试结果如表8所示。从表8中可以看出，本文算法与其他算法的McNemar测试结果均大于1.96，也就是说，与其他算法相比，本文算法的性能提升十分显著。

表8 McNemar的测试结果

Table 8 Results on Mcnemar's test

算法	Indian Pines				Salinas	Pavia University
	样本1	样本2	样本3	样本4		
本文算法与RF	25.300	24.457	23.973	21.998	23.038	36.938
本文算法与RoF	13.415	17.528	23.206	25.333	23.214	45.266
本文算法与ROSRoF	16.061	27.912	27.634	30.799	24.290	68.011
本文算法与SMOTERoF	23.855	25.889	31.429	30.767	24.083	58.636

不同算法的高光谱图像分类结果如图2—图4所示，其中图2显示的是Indian Pines数据在IR为49.1时的分类结果。不同数据上的分类结果表明本文提出的算法性能明显优于RF算法与传统RoF算法，在少数类别的样本分类精度高的同时，多数类样本同样能被更好地识别出来。

(a) 地面参考数据
(a) Ground truth

(b) RF分类结果
(b) RF classification result

(c) RoF分类结果
(c) RoF classification result

(d) ROSRoF分类结果
(d) ROSRoF classification result

(e) SMOTERoF分类结果
(e) SMOTERoF classification result

(f) 本文算法分类结果
(f) Proposed algorithm classification result

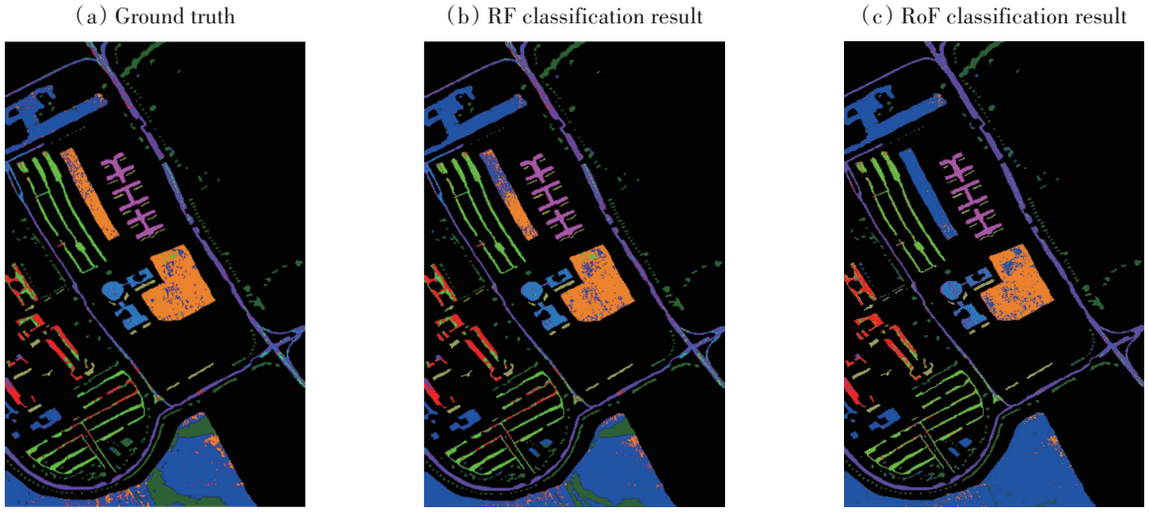
图2 Indian Pines分类结果

Fig. 2 Indian Pines classification results

(a) 地面参考数据
(a) Ground truth

(b) RF分类结果
(b) RF classification result

(c) RoF分类结果
(c) RoF classification result



(d) ROSRoF分类结果
(d) ROSRoF classification result

(e) SMOTERoF分类结果
(e) SMOTERoF classification result

(f) 本文算法分类结果
(f) Proposed algorithm classification result

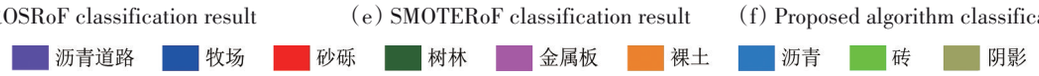
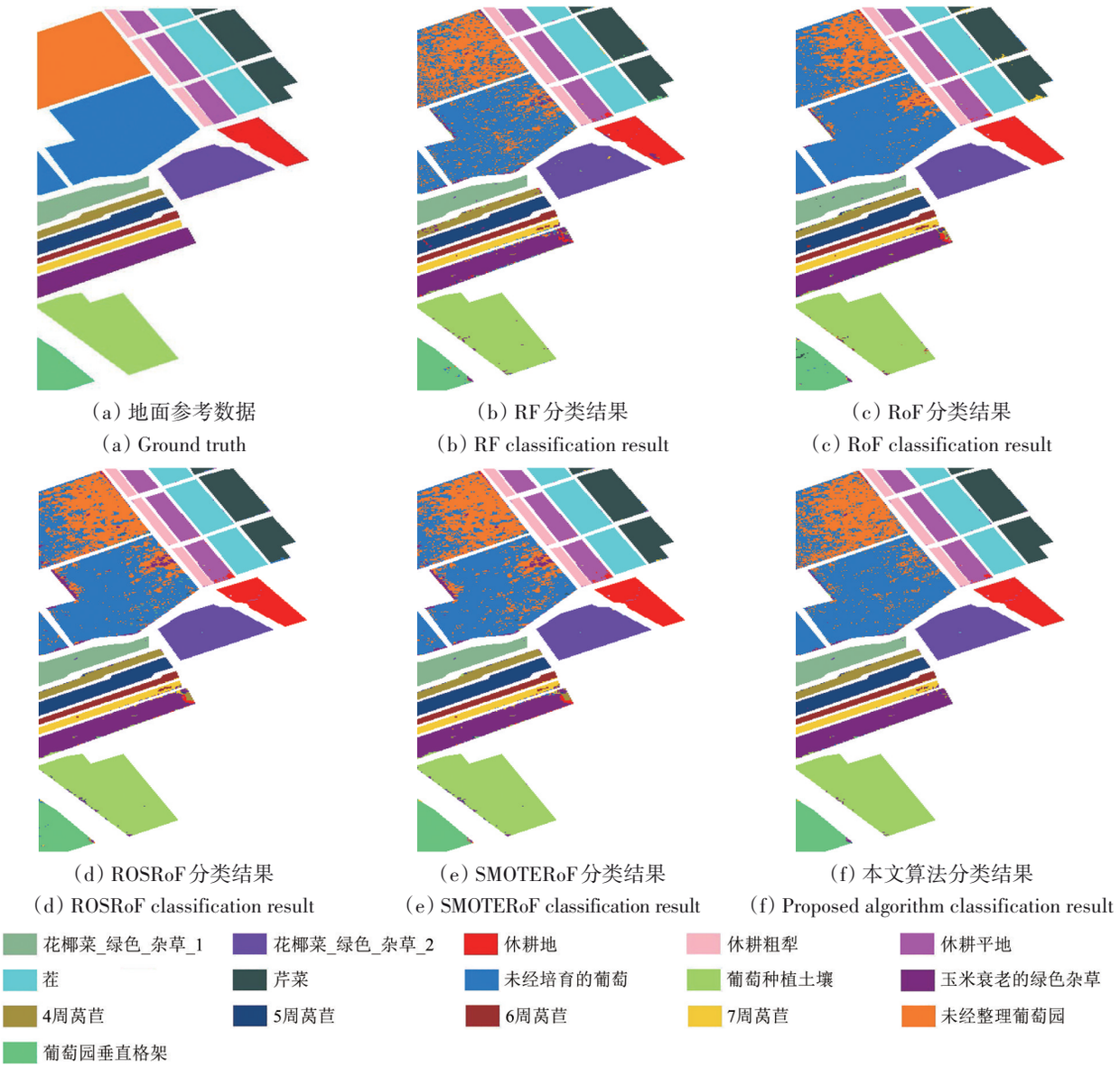


图3 Pavia University分类结果

Fig. 3 Pavia University classification results



（a）地面参考数据
（a）Ground truth

（b）RF分类结果
（b）RF classification result

（c）RoF分类结果
（c）RoF classification result

（d）ROSRoF分类结果
（d）ROSRoF classification result

（e）SMOTERoF分类结果
（e）SMOTERoF classification result

（f）本文算法分类结果
（f）Proposed algorithm classification result

图4 Salinas分类结果

Fig. 4 Salinas classification results

6 结 论

本文提出了一种基于SMOTE和RoF动态集成算法来解决多类高光谱数据不平衡分类问题。该算法是一种基于内部不平衡采样的集成方法，使用样本重要性函数确定样本的权重，以降低SMOTE算法引入人工噪声的风险，并利用动态采样来增加样本多样性。本文使用Indian Pines、Salinas和Pavia University高光谱数据对提出的算法进行多类不平衡的分类实验，选用4种对比算法，分别是RF、传统RoF算法，以及ROS和SMOTE预处理后再使用RoF分类的算法相比较，采用多种评价方法对算法的分类性能进行评价，包括总体分类精度、平均分类精度、F-measure、Gmean、最小召回率、样本多样性、模型训练时间以及McNemar测试。实验结果表明，与未经数据平衡处理的分类算法相比，本文算法具有优异的表现。

通过对比实验可以看出，在Salinas以及Pavia这类数据不平衡度较低的数据集中，本文算法在提高少数类样本分类精度的同时，仍旧在很大程度上提高了总体分类精度。而且，本文算法在不平衡率较高的Indian Pines数据集上的分类精度表明，本文算法在不平衡率该表的情况下同样拥有很好的性能。本文算法的评价参数例如McNemar的结果均大于1.96可以说明，与同类算法相比性能有十分显著的提升。同时，对比SMOTE只作为数据

预处理的方法，本文算法将数据采样与分类过程内部结合，不仅在多数类别上保持高分类精度，而且能很好地识别少数类样本。同时，本文算法采用动态采样技术，能有效提高样本的多样性。

在下一步工作中，可以将本文算法与一些特征提取算法相结合，例如奇异谱分析和稀疏表示，从而进一步提升样本多样性。

参考文献(References)

Arshad A, Riaz S and Jiao L C. 2019. Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification. IEEE Access, 7: 28100-28112 [DOI: 10.1109/ACCESS.2019.2901860]

Bandara A, Hettiarachchi Y, Hettiarachchi K, Munasinghe S, Wijesinghe I, Kusal H, Sidath M, Ishara W and Thayasivam U. 2020. A generalized ensemble machine learning approach for landslide susceptibility modeling//Sharma N, Chakrabarti A and Balas V E eds. Data Management, Analytics and Innovation. Singapore: Springer, 1016: 71-93 [DOI: 10.1007/978-981-13-9364-8_6]

Bhagat R C and Patil S S. 2015. Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest//2015 IEEE International Advance Computing Conference (IACC). Banglore: IEEE [DOI: 10.1109/IADCC.2015.7154739]

Breiman L. 2001. Random Forests. Machine Learning, 45: 5-32[DOI: 10.1023/A:1010933404324]

Cai L and Zhang G. 2019. Hyperspectral image classification with imbalanced data based on oversampling and convolutional neural network//Proceedings of SPIE 11342, AOPC 2019: AI in Optics and Photonics. Beijing: SPIE: 11342 [DOI: 10.1117/12.2543458]

Cai Z X, Wang X Y, Xu J and Jing L P. 2019. Sample adaptive classifier for imbalanced data. Computer Science, 46(1): 94-99 (才子昕, 王馨月, 徐剑, 景丽萍. 2019. 样本自适应的不平衡分类器. 计算机科学, 46(1): 94-99) [DOI: 10.11896/j.issn.1002-137X.2019.01.014]

Díez-Pastor J F, Rodríguez J J, García-Osorio C I and Kuncheva L I. 2015. Diversity techniques improve the performance of the best imbalance learning ensembles. Information Sciences, 325: 98-117 [DOI: 10.1016/j.ins.2015.07.025]

Douzas G, Bacao F and Last F. 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465: 1-20 [DOI: 10.1016/j.ins.2018.06.056]

Du P J, Xia J S, Xue Z H, Tan K, Su H J and Bao R. 2016. Review of hyperspectral remote sensing image classification. Journal of Remote Sensing, 20(2): 236-256 (杜培军, 夏俊士, 薛朝辉, 谭琨, 苏红军, 鲍蕊. 2016. 高光谱遥感影像分类研究进展. 遥感学报, 20(2): 236-256) [DOI: 10.11834/jrs.20165022]

Elreedy D and Atiya A F. 2019. A comprehensive analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences, 505: 32-64 [DOI: 10.1016/j.ins.2019.07.070]

Feng W and Bao W X. 2017. Weight-based rotation forest for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 14(11): 2167-2171 [DOI: 10.1109/LGRS.2017.2757043]

Feng W, Boukir S and Huang W. 2019a. Margin-based random forest for imbalanced land cover classification//2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE: 3085-3088 [DOI: 10.1109/IGARSS.2019.8898652]

Feng W, Dauphin G, Huang W J, Quan Y H and Liao W Z. 2019b. New margin-based subsampling iterative technique in modified random forests for classification. Knowledge-Based Systems, 182: 104845 [DOI: 10.1016/j.knosys.2019.07.016]

Feng W, Huang W J and Ren J C. 2018. Class imbalance ensemble learning based on the margin theory. Applied Sciences, 8(5): 815 [DOI: 10.3390/app8050815]

Gao L R, Zhang B, Zhang X and Shen X. 2007. Study on the method for estimating the noise in remote sensing images based on local standard deviations. Journal of Remote Sensing, 2007, 11, (2): 201-208 (高连如, 张兵, 张霞, 申茜. 2007. 基于局部标准差的遥感图像噪声评估方法研究. 遥感学报, (2): 201-208) [DOI: 10.11834/jrs.20070227]

García S, Zhang Z L, Altalhi A, Alshomrani S and Herrera F. 2018. Dynamic ensemble selection for multi-class imbalanced datasets. Information Sciences, 445-446: 22-37[DOI: 10.1016/j. ins. 2018.03.002]

Ghosh D and Cabrera J. Enriched random forest for high dimensional genomic data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1-1 [DOI: 10.1109/TCBB.2021.3089417]

Han Z, Gao L R, Zhang B, Sun X and Li Q T. 2020. Nonlinear hyperspectral unmixing algorithm based on deep autoencoder networks. Journal of Remote Sensing, 24(4): 388-400 (韩竹, 高连如, 张兵, 孙旭, 李庆亭. 2020. 高分五号高光谱图像自编码网络非线性解混. 遥感学报, 24(4): 388-400) [DOI: 10.11834/jrs.20209188]

Jimenez-Castaño C, Alvarez-Meza A and Orozco-Gutierrez A. 2020. Enhanced automatic twin support vector machine for imbalanced data classification. Pattern Recognition, 107: 107442 [DOI: 10.1016/j.patcog.2020.107442]

Krawczyk B. 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4): 221-232 [DOI: 10.1007/s13748-016-0094-0]

Mullick S S, Datta S, Dhekane S G and Das S. 2020. Appropriateness of performance indices for imbalanced data classification: an analysis. Pattern Recognition, 102: 107197 [DOI: 10.1016/j. patcog.2020.107197]

Pan T T, Zhao J H, Wu W and Yang J. 2020. Learning imbalanced datasets based on SMOTE and Gaussian distribution. Information Sciences, 512: 1214-1233 [DOI: 10.1016/j.ins.2019.10.048]

Rodríguez J J, Díez-Pastor J F, Arnaiz-González Á and Kuncheva L I. 2020. Random Balance ensembles for multiclass imbalance learning. Knowledge-Based Systems, 193: 105434[DOI: 10.1016/j.knosys.2019.105434]

Rodriguez J J, Kuncheva L I and Alonso C J. 2006. Rotation forest: a new classifier ensemble method. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 28(10): 1619-1630 [DOI: 10.1109/TPAMI.2006.211]

Tu X, Shen X B, Fu P, Wang T, Sun Q S and Ji Z X. 2020. Discriminant sub-dictionary learning with adaptive multiscale superpixel representation for hyperspectral image classification. Neurocomputing, 409: 131-145 [DOI: 10.1016/j.neucom.2020.05.082]

Zhang Y Q, Lu R Z, Qiao S J, Han N, Gutierrez L A and Zhou J L. 2020. A sampling method of imbalanced data based on sample space. Acta Automatica Sinica, 1-14 (张永清, 卢荣钊, 乔少杰, 韩楠, Gutierrez L A, 周激流. 2020. 一种基于样本空间的类别不平衡数据采样方法. 自动化学报, 1-14) [DOI: 10.16383/j.aas.c200034]

Zhou G and Guo F L. 2019. Research on sampling diversity method in ensemble learning base on margin//2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). Taiyuan: Shanxi University of Finance and Economics and hosted by AEIC Academic Exchange Center : 316-319[DOI: 10.1109/MLBDBI48998.2019.00071]

Zhou S, Sun L J, Xing W, Feng G J, Ji Y M, Yang J and Liu S C. 2020. Hyperspectral imaging of beet seed germination prediction. Infrared Physics & Technology, 108: 10336 [DOI: 10.1016/j.infrared.2020.103363]

Dynamic ensemble algorithm of SMOTE and rotation forest for imbalanced hyperspectral remote sensing classification

TONG Yingping[1,2], FENG Wei[1,2], SONG Yijia[1,2], QUAN Yinghui[1,2], HUANG Wenjiang[3], GAO Lianru[3], ZHU Wentao[1,2], XING Mengdao[4]

1.School of Electronic Engineering, Xidian University, Xi'an 710071, China;
2.Research Institute of Advanced Remote Sensing Technology, Xidian University, Xi'an 710071, China;
3.Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
4.Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China

Abstract: Rotation Forest (RoF), a powerful ensemble classifier, has obtained many successful applications in hyperspectral image classification. However, the data often has the problem of class imbalance. Consequently, the traditional RoF algorithm focuses on identifying the classes with majority samples, ignoring the accuracy of minority samples. The SMOTE (Synthetic Minority Oversampling Technique) algorithm increases the number of minority samples by simulating the way of generating new samples, thereby achieving the effect of balancing the categories of the data set. However, the SMOTE algorithm is mainly used in the data preprocessing stage and has the risk of increasing artificial noise when dealing with multi-class problems. Therefore, a novel dynamic ensemble algorithm based on SMOTE and RoF is proposed in this work to increase the classification accuracy of the multi-class imbalanced hyperspectral data. The proposed algorithm uses a dynamic sampling factor technology to merge the class distribution optimization with the base classifier. This algorithm not only realizes the adaptive generation of class balance data set but also reduces the influence of noise on the base classifier. In this experiment, three public hyperspectral images are used to test the performance of the algorithm, They are Indian Pines, Salinas and Pavia University. Four comparison algorithms are also selected, including random forest, traditional RoF, RoF algorithm with random oversampling, and SMOTE data preprocessing. The overall accuracy, average accuracy, F-measure, Gmean, minimum recall rate, ensemble classifier diversity, model training time, and McNemar test are the algorithm evaluation criteria. The experimental results demonstrate the effectiveness of the proposed method. The novel method not only obtains obvious classification advantages but also increases the recognition accuracy of minority samples while maintaining the overall classification accuracy of the data.

Key words: ensemble learning, imbalanced classification, rotation forest, SMOTE, dynamic sampling

Supported by National Natural Science Foundation of China (No. 61772397, 12005169, 62201438); Basic Research Program of Natural Science of Shaanxi Province (No. 2021JC-23); Science and Technology Development of Yulin Science and Technology Bureau (No. CXY-2020-094); Shaanxi Forestry Science and Technology Innovation Key Project (No. SXLK2022-02-8)