doi:10.6041/j.issn.1000-1298.2019.06.029

基于 SVC 和过采样的类别非均衡农业高光谱数据分类

袁培森[1]　翟肇裕[2]　任守纲[1,3]　顾兴健[1]　徐焕良[1,3]

(1.南京农业大学信息科学技术学院,南京 210095;2.马德里理工大学技术工程和电信系统高级学院,马德里 28040;
3.国家信息农业工程技术中心,南京 210095)

摘要:系统研究了农业高光谱数据中少数类的分类质量问题。为了提升少数类的分类质量,提出采用过采样 SMOTE 技术增加少数类新样本,同时研究了 SMOTE 技术中新样本生成策略和少数类采样倍率对高光谱数据中少数类分类结果的影响,以及不平衡数据集上分类器与模型的匹配度。在新的采样数据集上采用多类分类 SVC 技术对少数类分类,提升了非均衡高光谱数据集中少数类的分类质量。在真实数据集上进行了试验验证,并对不同的分类方法和系统参数进行了试验对比和分析,结果表明,本文方法能够显著地提高非均衡高光谱数据中少数类分类效果,平均分类精度不小于 0.82,平均召回率提升幅度为 11.11% ~26.15%,F_1 提升幅度为 5.81% ~40.85%。

关键词:高光谱数据分类;支持向量分类;过采样;非均衡数据;SMOTE

中图分类号:TP391;S24　　**文献标识码**:A　　**文章编号**:1000-1298(2019)06-0257-08

Classification of Unbalanced Agricultural Hyperspectral Data Based on SVC and Oversampling

YUAN Peisen[1]　ZHAI Zhaoyu[2]　REN Shougang[1,3]　GU Xingjian[1]　XU Huanliang[1,3]

(1. *College of Information Science and Technology*, *Nanjing Agricultural University*, *Nanjing* 210095, *China*
2. *Superior School of Technical Engineering and Telecommunication Systems*, *Technical University of Madrid*, *Madrid* 28040, *Spain*
3. *National Engineering and Technology Center for Agriculture*, *Nanjing* 210095, *China*)

Abstract: Hyperspectral technology is widely used in agricultural natural resources such as agro-ecological environment and land resource protection. Spectral imaging technology can effectively classify and identify ground objects. Therefore, the classification of hyperspectral data is one of the important contents of hyperspectral research. Category non-equilibrium problem is a common problem in agricultural hyperspectral data, and the classification quality of minority classes has great significance for the effective classification of hyperspectral data. However, the classification of minority classes is affected by the dominant majority classes. The general classification algorithm tends to the dominant majority classes classification, so that minority classes are usually submerged in the majority classes, bringing great challenge to classification accuracy and recall rate of the minority classes. The classification quality of the minority objects was studied in agricultural hyperspectral data. In order to improve the classification quality of minority classes, an oversampling technique SMOTE was proposed to add new samples for the minority classes. At the same time, the effects of new sample generation strategy and minority instance sampling rate on the classification results of minority samples in the agricultural hyperspectral data and the matching degree between the classifier and the model on the unbalanced data set were systematically studied. A multi-class classification SVC technique was used to classify minority classes on a new sampling data set, and it improved the classification accuracy of the minority classes in unbalanced agricultural hyperspectral dataset. The experimental verification was carried out on the real data set, and different classification methods and system parameters were tested and compared. The experimental results showed that the proposed method can greatly improve the effect of minority classification in

收稿日期:2018-11-26　修回日期:2018-12-19

基金项目:国家自然科学基金项目(61502236)和中央高校基本科研业务费专项资金项目(KYZ201752、KJQN201651)

作者简介:袁培森(1980—),男,讲师,博士,主要从事智能信息处理、海量数据处理与分析研究,E-mail:peiseny@njau.edu.cn

通信作者:徐焕良(1963—),男,教授,博士,主要从事农业信息化与大数据技术研究,E-mail:huanliangxu@njau.edu.cn

unbalanced agricultural hyperspectral data. The weight precision can reach above 0.82, the weight recall rate was obviously improved from 11.11% to 26.15%, and F_1 was increased from 5.81% to 40.85%. The method can provide a reference for the unbalanced agricultural hyperspectral data to improve the classification effect systematically.

Key words: hyperspectral data classification; SVC; oversampling; imbalanced data; SMOTE

0 引言

高光谱遥感技术通过采集不同目标反射或辐射出不同波长的电磁波，得到以像素为单位的遥感影像。高光谱图像(Hyperspectral image)拥有丰富的光谱信息,能提供准确、详细的土地覆盖材料描述,在农作物长势监测、作物养分监测、水分状况监测、作物生长参数估算、农作物高光谱遥感识别和分类等[1-6]方面具有广泛的应用。

高光谱图像波段多，光谱波段覆盖范围广，包含丰富的信息，并可结合空间信息和光谱信息,有效地反映目标的信息。高光谱图像分类是高光谱遥感对地观测研究的重要内容,其具体任务是对高光谱图像中的每一个像素所代表的目标进行有效的分类[7-8]。而非均衡问题是高光谱数据非常普遍的现象,当数据集中类别的比例严重失调时,传统的分类技术对于不平衡的数据集往往不能产生令人满意的结果[9]。因此非均衡高光谱数据的分类是一个很普遍和重要的问题[10-14]。

处理非均衡数据分类的主要技术包括数据采样[15-16]、基于代价的分类[17]等。其中,基于数据采样技术包括过采样和欠采样[18]。欠采样采用减少多类样本方案,但是由于高光谱图像训练样本一般较少,对采集到的高光谱图像进行标记代价较高。丢失的标类样本对分类模型具有较大的影响,因此对于不均衡农业高光谱数据导致的少数类分类质量问题,本文采用典型的合成少数类的过采样技术(Synthetic minority oversampling technique, SMOTE)[19]。

目前,在非均衡高光谱数据集上的分类研究,尤其是农业高光谱数据分析方面,许多学者进行了研究。ZHANG 等[10]提出,模型学习期间使用支持向量进行抽样,使得训练数据分布均衡。GARCÍA 等[20]使用随机采样和 PCA 技术对高光谱数据随机采样再降维,进而使用决策树来分类,分类精度虽有所提高,但是该方法试验结果仍有很大的提升空间。LI 等[14]提出基于正交补的数据子空间投影不平衡的高光谱图像分类技术。CHAO 等[13]针对复杂高光谱图像数据分类中少数类别分类精度低的问题,提出了一种基于 MK-LSSVM 的不平衡分类方法,该方法使用 K-means 聚类将多数类划分为不同的组,在聚类之后使用抽样技术平衡每个组和少数类,通过构建 MK-LSSVM 分类器对高光谱图像进行分类。GRAVES 等[21]通过成像光谱数据的分类不平衡数据集绘制物种图谱,这将有助于研究在越来越大的空间尺度上理解树种分布。

本文针对非均衡农业高光谱数据,采用过采样(Oversampling)技术处理少数类样本数据,通过对少数类采样个数的分析,研究少数类样本采样倍率对分类结果的影响。通过对非均衡高光谱数据采样,使少数类样本尽可能均衡,进而采用多类分类器——支持向量分类(Support vector classification, SVC)进行分类。采用 SMOTE[19]进行过采样少数类样本,研究 SMOTE 参数对分类精度的影响,研究非均衡高光谱数据分类器与模型的一致性对分类的影响。针对高光谱数据多类分类问题,采用 SVC[22]有效解决不平衡高光谱图像分类模型的集合学习过程。

1 非均衡数据及处理

1.1 非均衡数据

数据非均衡是数据挖掘等领域常见的问题,例如信用卡欺诈数据分析[16]、疾病诊断[23]、生物信息分析[24]、高光谱数据分析[21]等。BRANCO 等[25]总结了非均衡数据的问题,并对比了相关的方法和理论,同时得出,非均衡数据中,用户更加重视在目标变量域的子集上的预测性能,但是与用户更相关的样本在训练中的代表性较差,进而导致模型对少数类样本的错误估计。

非均衡数据问题的描述如下[25]:给定包含 n 个训练数据的数据集 $D=\{\boldsymbol{x}_i,y_i\}$,$i=1,2,\cdots,n$。向量 $\boldsymbol{x}_i\in\mathbf{R}^d$,$y_i\in\{-1,1\}$,定义用户关注度函数 φ 指定了用户关注的相关类,相关性阈值 τ,$D_r\in D$ 为训练样本中用户关注度超过阈值的子集,即 $D_r=\{(\boldsymbol{x}_i,y_i)\in D:\varphi(y_i)>\tau\}$,$D_n$ 为正常类,即 $D_n=\{(\boldsymbol{x}_i,y_i)\in D:\varphi(y_i)\leqslant\tau\}$,$D_n=D\backslash D_r$。非均衡比例因子 $I_m=\dfrac{|D_r|}{|D_n|}$为少数类与多数类的元素数量之比。在一般的非均衡数据集中,满足 $D_r\ll D_n$,即 I_m 通常非常低。

1.2　过采样

处理不均衡数据，有两个基本方法[16-17]：改变数据分布，在数据层面使得类别更均衡；改变分类算法，在传统分类算法的基础上对分类器采用加权方式，使得模型对少数类更加敏感。采样技术简单高效，常用的采样方法有：欠采样、过采样、欠采样和过采样综合采样[18]。过采样采用增加少数类样本，通过设置多数类和少数类的比例系数，在多数类样本不变情况下，生成指定数量的少数类样本，如图 1 所示，图中 η 为少数类的采样倍率。

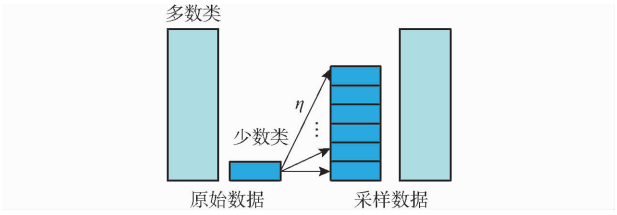

图 1　非均衡数据过采样示意图

Fig. 1　Illustration of unbalanced data oversampling processing

SMOTE[19]是典型的非均衡数据过采样技术，它通过在少数样本附近位置生成新样本达到类别平衡的目的，可以有效避免分类器过拟合。其处理基本过程如下：

(1)对少数类中每一个样本 $\boldsymbol{x}_i$，计算它到少数类样本集 D_r 中所有样本的欧氏距离，得到其中 k 个近邻。

(2)根据样本不平衡比例设置一个采样比例以确定采样倍率 η，对于每一个少数类样本 $\boldsymbol{x}_i$，从 k 个近邻中随机选择若干个样本，设选择的近邻为 $\boldsymbol{x}_i'$。

(3)对于每一个随机选出的近邻 $\boldsymbol{x}_i'$，分别与原样本构建新的样本 $\boldsymbol{x}_n$。

$$\boldsymbol{x}_n = \boldsymbol{x}_i + \mathrm{rand}(0,1) \mid \boldsymbol{x}_i - \boldsymbol{x}_i' \mid \tag{1}$$

式中　$\mathrm{rand}(\cdot)$——均匀分布函数

SMOTE 通过生成新的数据集来解决少数类分类不平衡的问题。假设初始数据集中少数类样本数为 $|D_r|$，多数类样本数为 $|D_n|$，首先增加 $(\eta-1)|D_r|$ 个少数类样本，并把最初的少数类样本和新增的少数类样本都放入新的数据集中。这样，新的数据集中少数类样本有 $\eta|D_r|$ 个，数据集共 $\eta|D_r|+|D_n|$ 个样本。

根据少数类选择策略的不同，SMOTE 包含 Regular、Borderline1、Borderline2 和 SVM 这 4 种策略[26]。

2　非均衡高光谱数据的多类分类处理

2.1　处理过程

非均衡农业高光谱数据的分类采用两阶段处理：数据预处理及过采样，生成新的采样数据集；采用多类分类器训练分类模型，对少数类分类质量进行评估和参数最优化选择。

采样预处理通过对少数类进行过采样使得各类实例大致平等。通过使用过采样，学习模型能够极大地克服由于多数类导致的模型先验偏差。

高光谱数据采用最小-最大规范化(Min - max normalization)[27]对原始数据进行线性变换，设样本属性 A 的值为 v，则该规范化为

$$v' = \frac{v - v_{\min}}{v_{\max} - v_{\min}}(m_{\mathrm{ax}} - m_{\mathrm{in}}) + m_{\mathrm{in}} \tag{2}$$

式中　$v_{\max}$——属性 A 最大值

$v_{\min}$——属性 A 最小值

m_{ax}——属性 A 的值域最大值

m_{in}——属性 A 的值域最小值

本文将原始数据变换到[0,1]区间，此时最大值和最小值分别为 0 和 1。

通过 SMOTE 对数据集 D 中指定的少数类和采样倍率 η 进行采样，新生成的数据集记为 D'，$D' = \eta|D_r| + |D_n|$。

2.2　高光谱多类分类

由于高光谱分类任务数据中通常包括多个类别。主要是通过组合多个二分类器来实现多分类器的构造[28]。一般的高光谱分类问题属于多类分类(Multi-class classification)，即将实例分类为 2 个类以上的分类问题。多类分类的假设是，每个样本有唯一的类标签。多类分类问题通过采用二类分类算法和一定的策略完成多类分类任务，采用的策略有 One *vs* all 和 One *vs* one[28]。

SVC[29]是基于支持向量机分类的一种技术。对于两类分类问题，SVC 问题可以归为以下问题：

给定两类问题的训练向量 $\boldsymbol{x}_i \in \mathbf{R}^d$，$i=1,2,\cdots,n$，类标号 $y_i \in \{-1,1\}$，SVC 问题的求解公式为

$$\begin{cases} J = \min\limits_{w,b,\zeta} \dfrac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + C\sum\limits_{i=1}^{n}\zeta_i \\ \text{s.t.} \begin{cases} y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geqslant 1 - \zeta_i \\ \zeta_i \geqslant 0 \end{cases} \end{cases} \tag{3}$$

式中　ζ_i——松弛变量　　b——截距

$\boldsymbol{w}$——权重向量　　J——目标函数

C——调和系数　　ϕ——核函数

式(3)的求解需要转换为对偶形式，其对偶形式为

$$\begin{cases} \min\limits_{\alpha} \dfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} - \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{s.t} \begin{cases} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \\ 0 \leqslant \alpha_i \leqslant C \quad (i=1,2,\cdots,n) \end{cases} \end{cases} \tag{4}$$

其中 $Q_{ij}=y_iy_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_j)$

式中 $\boldsymbol{e}$——单位向量 $\boldsymbol{y}$——类标号

$\boldsymbol{\alpha}$——拉格朗日乘子

α_i——拉格朗日乘子

$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$——核函数

$\boldsymbol{Q}$——$n\times n$ 的半正定矩阵

Q_{ij}——Q 的元素

式(4)通过核函数 $\boldsymbol{\phi}$ 使得训练向量映射到高维空间。常见的核函数选择为径向基函数(Radial basis function,RBF)[28]

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\exp(-\gamma\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2) \tag{5}$$

式中 γ——核的泛化能力参数,$\gamma\geqslant 0$

若 γ 越小则决策边界越精简,泛化能力越强。γ 越大表示决策边界越复杂,则泛化能力越弱。

对于样本 $\boldsymbol{x}$,其分类的决策函数可以表示为

$$f(\boldsymbol{x})=\mathrm{sgn}\left(\sum_{i=1}^{n}y_i\boldsymbol{\alpha}K(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{6}$$

对于多类问题,SVC 采用 One *vs* one[30] 策略。对于 m 个分类,训练 $M=\dfrac{m(m-1)}{2}$ 个分类器,每个分类器使用其中两类样本作为正负样本来训练;决策时使用投票原则,即将测试样本 $\boldsymbol{x}$ 输入到 M 个分类器,每个分类器输出一个判别结果,得票最多类作为类标号,如

$$y=\arg\max_{i\in\{1,2,\cdots,M\}}f_i(\boldsymbol{x}) \tag{7}$$

式中 f_i——分类器决策函数

3 试验结果与分析

3.1 试验数据集

本文测试的数据集为 Indian Pines,该数据集由 AVIRIS 传感器在印第安纳州西北部的印度松树林采集,由 145 像素 × 145 像素和 224 光谱反射波段组成,波长范围为 $4\times10^{-7}\sim2.5\times10^{-6}$ m。该数据集共包含了 16 类不同的农业对象,共标注 10 249 个像素类别,数据集真实类及分布比例具体如表 1 所示。本文算法采用 Python 3.6 实现。

本文所使用的数据集 Indian Pines 中类别分布如图 2 所示。从图 2 可以看出,该高光谱数据集少数类数量分布极不均衡。本文把类别中所占比例低于 3% 的类别作为少数类,其余作为多数类。因此,该数据集中少数类包括 6 个:Alfalfa、Corn、Grass - pasture - mowed、Oats、Wheat 和 Stone - Steel - Towers。

3.2 参数设置

本文参数设置如下:测试集和训练集的比例为 3∶7。径向基函数 RBF 参数 γ 为 0.125,C 为 1。SMOTE 参数 k 的范围为 3 ~ 7,默认为 6,新样本中生成策略默认为 SVM。少数类采样倍率默认为 5。

表 1 Indian Pines 数据集的 Groundtruth 类及其样本数

Tab. 1 Groundtruth class of Indian Pines dataset and its sample size

标号	类名	样本数	百分比/%
1	Alfalfa	46	0.45
2	Corn - notill	1428	13.93
3	Corn - mintill	830	8.10
4	Corn	237	2.31
5	Grass - pasture	483	4.71
6	Grass - trees	730	7.12
7	Grass - pasture - mowed	28	0.27
8	Hay - windrowed	478	4.66
9	Oats	20	0.20
10	Soybean - notill	972	9.48
11	Soybean - mintill	2455	23.95
12	Soybean - clean	593	5.79
13	Wheat	205	2.00
14	Woods	1265	12.34
15	Buildings - Grass - Trees - Drives	386	3.77
16	Stone - Steel - Towers	93	0.91

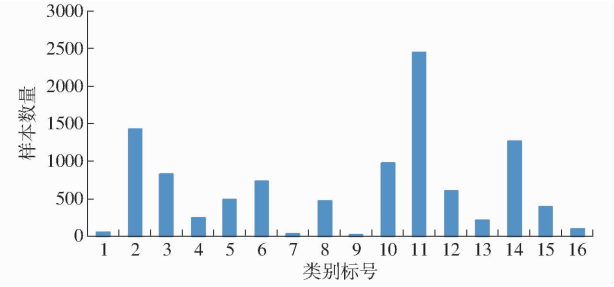

图 2 数据集样本类别分布

Fig. 2 Illustration samples distribution of Indian Pines datasets

3.3 分类精度指标

试验从半监督分类预测结果的质量分类效率进行了系统的研究和分析。定义 T_P(True positive)为正类并且也被预测成正类,F_P(False positive)为负类被预测成正类,T_N(True negative)为负类被预测成负类,F_N(False negative)为正类被预测成负类。

分类结果从 7 个方面进行度量:加权精度 p_w(Weight precision)、加权召回率 r_w(Weight recall)、加权 F_1 度量、分类准确率 A_c(Accuracy)、精度的几何平均值 G_M、平衡准确性指数(Index of balanced accuracy,IBA)和 Kappa 系数 K_a。定义分别为

$$p_w=\frac{\sum_{i=1}^{m}s_i\rho_i}{\sum_{i=1}^{m}s_i} \tag{8}$$

$$r_w = \frac{\sum_{i=1}^{m} s_i r_i}{\sum_{i=1}^{m} s_i} \tag{9}$$

$$F_1 = \frac{\sum_{i=1}^{m} s_i f_i}{\sum_{i=1}^{m} s_i} \tag{10}$$

$$A_c = \frac{T_P + T_N}{T_P + F_N + T_N + F_P} \tag{11}$$

其中　$\rho_i = \frac{T_P^i}{T_P^i + F_P^i}$　$r_i = \frac{T_P^i}{T_P^i + F_N^i}$　$f_i = \frac{2\rho_i r_i}{\rho_i + r_i}$

式中　s_i——第 i 类的支持度

i——高光谱数据类别，即每一类的真实类出现次数

T_P^i——第 i 类的 T_P 值

F_P^i——第 i 类的 F_P 值

F_N^i——第 i 类的 F_N 值

ρ_i——第 i 类的精度

r_i——第 i 类的召回率

f_i——第 i 类的 F_1 度量

G_M 度量用于评估不平衡分类应用的分类性能。两类精度的几何平均值为

$$G_M = \sqrt{\frac{T_P}{T_P + F_N} \frac{T_N}{T_N + F_P}} \tag{12}$$

G_M 在获得两类精度良好平衡的同时使其最大化。

IBA 量化了两类准确性平衡指数与选择的无偏估计总体准确性之间的权衡，计算式为

$$I_{BA_\theta} = [1 + \theta(T_P - T_N)] T_P T_N \tag{13}$$

I_{BA_θ}取决于用户定义的参数 θ，本文 θ 设置为 1。

Kappa 系数[31] K_a 准确度指标用于度量不平衡数据集的分类器与模型匹配的精细化程度，公式为

$$K_a = \frac{p_o - p_h}{1 - p_h} = 1 - \frac{1 - p_o}{1 - p_h} \tag{14}$$

式中　p_o——评估者之间相对观察到的一致意见的百分比

p_h——一致意见的预期次数

使用观察到的数据来计算每个观察者随机查看每个类别的概率 $K_a \in (0,1)$，K_a 在 0.61 ~ 0.80 之间表示模型具有较好的一致性[31]。

3.4　分类精度

在原始数据集和采样数据集上，对比了 SVC 和随机森林（Random forest，RF）[32] 在 Indian Pines 数据集的分类准确率。

图 3 是对 Indian Pines 原始数据集和采样数据集两个分类方法的分类准确率。采用 70% 数据集训练模型、30% 数据集作为测试集。在 Indian Pines 原始数据集上 SVC 在所有类上的分类准确率为 0.78，少数类的分类准确率为 0.65，RF 在所有类上分类准确率为 0.82，在少数类上的分类准确率为 0.38。结果表明 SVC 在原始数据集上的分类准确率比 RF 低 4.88%，但是对于少数类的分类准确率，SVC 比 RF 高 71.05%。因此，SVC 对少数类分类效果优于 RF。

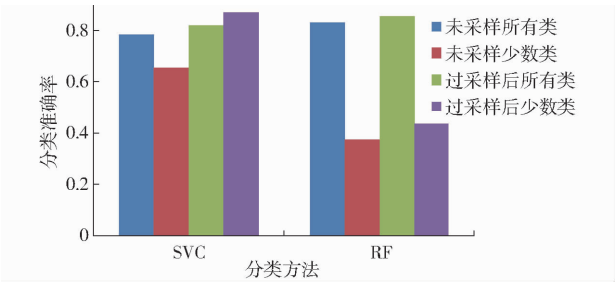

图 3　Indian Pines 数据集分类精度对比

Fig. 3　Classification accuracy comparison of Indian Pines dataset

对数据集中的少数类使用 SMOTE 采样之后，在采样数据集上，SVC 和 RF 的分类准确率有所提高，分别提高 32.93% 和 16.46%。结果表明，SCV 在少数类上的分类准确率和提高的比例优于 RF。因此，本文采用 SVC 对高光谱数据进行分类。

图 4 是在原始数据集上使用 SVC 分类的混淆矩阵，横轴为在 16 个类别上预测的类标号，纵轴为 16 个真实类标号。从图 4 可以看出，SVC 在未采样的高光谱数据集上的分类效果不理想，尤其是对少数类 1、4、7、9 这 4 个类分类精度比较低。

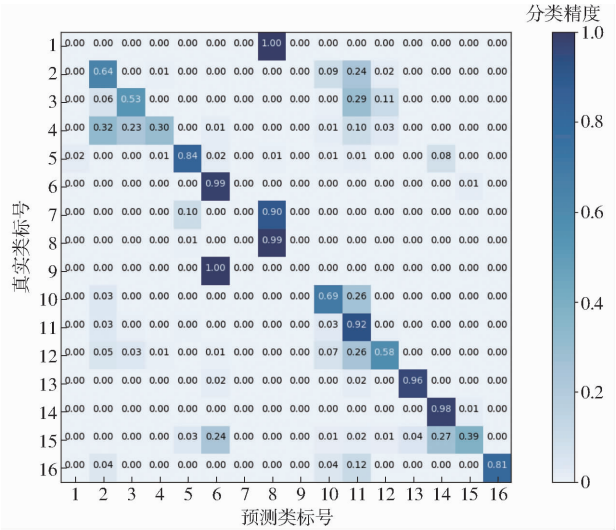

图 4　原始数据集使用 SVC 分类的混淆矩阵

Fig. 4　Confusion matrix of original imbalanced dataset with SVC

3.5　SMOTE 参数对精度的影响

3.5.1　新样本生成策略

SMOTE 的参数 k 设置为 6 时，数据集上少数类

识别准确率如表 2 所示。从表 2 可以看出,SVM 策略对少数类分类结果的加权召回率 r_w、F_1、G_M 和 IBA 结果较其它 3 个策略好。

表 2 SMOTE 的 4 种新样本生成策略对少数类的影响

Tab. 2 Influence of new instance generating strategy of SMOTE on minority classes

策略	p_w	r_w	F_1	G_M	IBA
Regular	0.96	0.86	0.90	0.92	0.84
Borderline1	0.96	0.85	0.90	0.91	0.83
Borderline2	0.95	0.83	0.87	0.89	0.80
SVM	0.95	0.87	0.91	0.92	0.85

为了进一步测试少数类上新样本生成策略对分类准确率的影响,在 k 为 6 时,测试了少数类分类准确率与 4 种新样本生成策略的关系,结果如图 5 所示。从图 5 可以看出,4 个策略中 SVM 的少数类分类准确率最高,为 0.873。

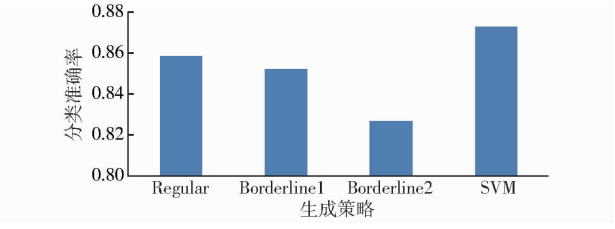

图 5 不同生成策略时少数类的分类准确率

Fig. 5 Classification accuracy of minority classes with new instance generating strategy of SMOTE

3.5.2 参数 k

SMOTE 新样本生成策略为 SVM 时,少数类结果度量与参数 k 关系如表 3 所示。从表 3 可以看出,k 为 6 时,少数类分类结果在加权召回率 r_w、F_1、G_M 和 IBA 指标上较好。

表 3 少数类结果度量与参数 k 的关系

Tab. 3 Classification performance of minority classes with parameters k of SMOTE

k	p_w	r_w	F_1	G_M	IBA
3	0.95	0.86	0.90	0.91	0.83
4	0.95	0.86	0.90	0.92	0.84
5	0.96	0.85	0.90	0.91	0.83
6	0.95	0.87	0.91	0.92	0.85
7	0.96	0.86	0.90	0.91	0.83

为了进一步测试少数类上参数 k 对分类准确率的影响,在生成策略为 SVM 时,测试了少数类的分类准确率与参数 k 的关系,结果如图 6 所示。参数 k 的取值范围为 3 ~ 7。少数类的分类准确率为 0.855 ~ 0.873,其中,k 为 6 时,分类准确率最高,为 0.873。

3.5.3 少数类采样倍率

高光谱数据中少数类采样倍率 η 与精度、召回率和 F_1 测试结果如图 7 所示。从图 7 可以看出,采用 SMOTE 对少数类采样的 p_w、r_w 和 F_1 影响非常显著;采样倍率 η 为 1 ~ 4 时,度量值提升幅度最大,p_w、r_w 和 F_1 分别提升了 8.67%、30.58% 和 25.81%;采样倍率为 5 ~ 15 时,p_w、r_w 和 F_1 变化不明显,结果比较稳定,具有较好的鲁棒性。

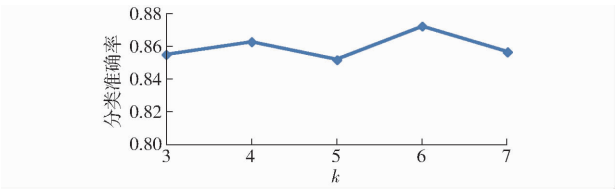

图 6 少数类的分类准确率与参数 k 的关系

Fig. 6 Classification accuracy of minority classes with parameters k of SMOTE

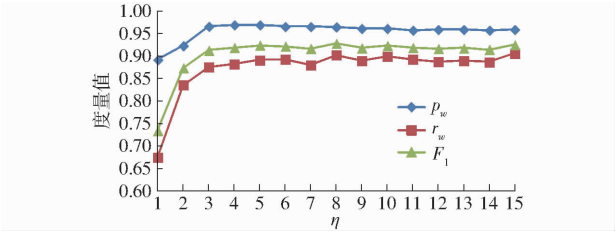

图 7 采样倍率 η 与 p_w、r_w 和 F_1 的关系

Fig. 7 Relationship of precison, recall rate and F_1 with sampling ratio η of SMOTE

3.6 Kappa 系数

参数 k 为 6 时,少数类上分类模型的 Kappa 系数 K_a 与 SMOTE 中新样本生成策略的关系如图 8 所示。4 个生成策略中少数类分类的 K_a 变化范围为 0.768 ~ 0.829,其中,SVM 策略的 Kappa 系数最高,为 0.829。

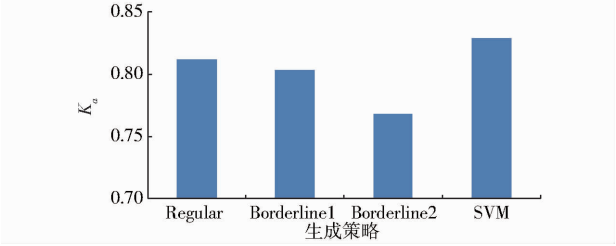

图 8 新样本生成策略与 K_a 的关系

Fig. 8 Relationship of K_a of model with new instance generating strategy of SMOTE

少数类上的分类模型的 K_a 与 SMOTE 中参数 k 关系如图 9 所示。参数 k 的取值范围为 3 ~ 7。少数类分类的 K_a 范围为 0.803 ~ 0.829,其中,k 为 6 时,Kappa 系数最高为 0.829。

从分类模型一致性度量 K_a 结果可知,参数 k 为 6 时,SVM 策略取得最优结果。

3.7 试验对比

表 4 是 SMOTE 参数 k 设置为 6,采用 SVM 策略时,在非采样数据集和采样数据集上的 p_w、r_w 和 F_1

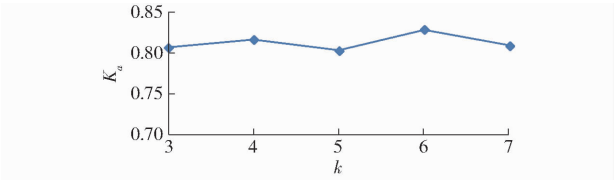

图 9　参数 k 与 K_a 的关系

Fig. 9　Relationship of K_a of model with parameters k of SMOTE

表 4　过采样数据集和原始数据集上少数类识别准确率对比

Tab. 4　Classification comparison over sampled and original unsampled dataset

标号	采样数据集			原始数据集		
	p_w	r_w	F_1	p_w	r_w	F_1
1	0.94	0.96	0.95	0	0	0
4	0.99	0.82	0.90	1	0.65	0.79
7	1	0.75	0.86	0	0	0
9	0.82	0.45	0.58	0	0	0
13	0.92	0.90	0.91	0.91	0.81	0.86
16	1	1	1.00	1	0.83	0.71

结果。SVC 分类器采用 RBF 核函数，各少数类采样的倍率 η 设置为 8。

表 4 中标号表示的类名与表 1 相同。表 4 结果表明，通过数据集 SMOTE 采样之后，少数类的 3 个分类指标均有大幅提升。Alfalfa、Grass－pasture－mowed、Oats 3 个类由于类数量在原始数据集上过于稀少，非采样集上 3 个指标都为 0。经过过采样之后，p_w、r_w 和 F_1 分别为 0.94、0.96、0.95、1.00、0.75、0.86 和 0.82、0.45、0.58。Corn 的 p_w 降低了 1%，但是 r_w 和 F_1 分别提升了 26.15% 和 13.92%。Wheat 的 p_w 提升了 1.1%，r_w 和 F_1 分别提升了 11.11% 和 5.81%，Stone－Steel－Towers 的 p_w 不变，但是 r_w 和 F_1 分别提升了 20.48% 和 40.85%。

上述结果表明，稀少的类在非采样数据集上，其分类的效果较差，往往淹没于多数类中，经过过采样，其分类的效果提升非常显著。Alfalfa、Oats 和 Grass－pasture－mowed 3 个类的 r_w 提升最显著。

表 5 是本文方法与 SVO[10] 及 SVM[33] 方法在平均分类精度 A_c 和 Kappa 系数 K_a 的试验对比结果。本文方法在参数 k 设置为 6、采用 SVM 策略时，分类的平均精度和 Kappa 系数相比于 SVO 分别提升了 6.72% 和 3.50%，相比于 SVM[33]，分别提升了 12.21% 和 3.62%。

表 5　与其他方法的准确率对比

Tab. 5　Classification accuracy comparison with other methods

参数	本文方法	SVO	SVM
A_c	0.873	0.818	0.778
K_a	0.829	0.801	0.800

图 10 是默认参数情况下，在采样数据集上使用 SVC 分类的混肴矩阵，横轴为 16 个预测类标号，纵轴为 16 个真实类标号。从图 10 可以看出，SVC 在采样高光谱数据集上的分类精度提升很大，对少数类 1、4、7、9、13 和 16 分类精度的分类效果提升显著。

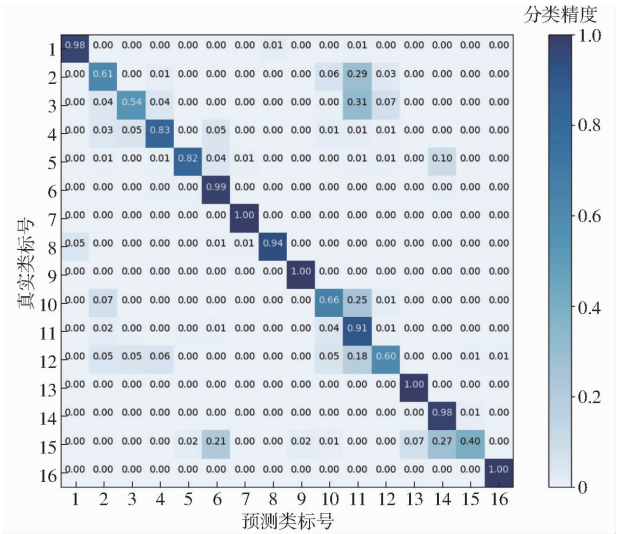

图 10　SMOTE 采样数据集上分类精度混肴矩阵

Fig. 10　Confusion matrix of oversampling dataset with SMOTE

4　结论

（1）针对农业高光谱数据的非均衡环境下少数类分类精度低的问题，研究了少数类的分类质量，利用过采样技术对数据进行处理，提升了少数类的分类质量。

（2）在高光谱数据集上进行了试验验证，对系统参数进行了试验和对比分析，试验结果表明，本文方法能够较好地提升高光谱数据集少数类分类精度，p_w 不小于 0.82，r_w 不小于 0.45，r_w 提升显著，提升幅度在 11.11% ~ 26.15% 之间。

参　考　文　献

[1]　王俊淑，江南，张国明，等. 高光谱遥感图像 DE－self－training 半监督分类算法[J/OL]. 农业机械学报，2015，46(5)：239－244.
WANG Junshu，JIANG Nan，ZHANG Guoming，et al. Semi-supervised classification algorithm for hyperspectral remote sensing image based on DE－self－training[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2015，46(5)：239－244. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20150534&flag=1. DOI：10.6041/j.issn.1000-1298.2015.05.034.(in Chinese)

[2]　高恒振. 高光谱遥感图像分类技术研究[D]. 长沙：国防科学技术大学，2011.

GAO Hengzhen. Research on classification technology of hyperspectral remote sensing images[D]. Changsha: National University of Defense Technology, 2011. (in Chinese)
[3] 孙俊,路心资,张晓东,等.基于高光谱图像的红豆品种 GA-PNN 神经网络鉴别[J/OL].农业机械学报,2016,47(6):215-221.
SUN Jun, LU Xinzi, ZHANG Xiaodong, et al. Identification of red bean variety with probabilistic GA-PNN based on hyperspectral imaging[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(6): 215-221. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20160628&flag=1. DOI: 10.6041/j.issn.1000-1298.2016.06.028. (in Chinese)
[4] CHEN X, LI S, PENG J. Hyperspectral imagery classification with multiple regularized collaborative representations[J]. IEEE Geoscience & Remote Sensing Letters, 2017, 14(7): 1121-1125.
[5] 姚付启,张振华,杨润亚,等. 基于红边参数的植被叶绿素含量高光谱估算模型[J]. 农业工程学报, 2009, 25(13):123-129.
YAO Fuqi, ZHANG Zhenhua, YANG Runya, et al. Hyperspectral models for estimating vegetation chlorophyll content based on red edge parameter[J]. Transactions of the CSAE, 2009, 25(13): 123-129. (in Chinese)
[6] 刘小丹,冯旭萍,刘飞,等. 基于近红外高光谱成像技术鉴别杂交稻品系[J]. 农业工程学报, 2017, 33(22): 189-194.
LIU Xiaodan, FENG Xuping, LIU Fei, et al. Identification of hybrid rice strain based on near-infrared hyperspectral imaging technology[J]. Transactions of the CSAE, 2017, 33(22): 189-194. (in Chinese)
[7] HAN M, ZHANG C. Spectral-spatial classification of hyperspectral image based on discriminant sparsity preserving embedding [J]. Neurocomputing, 2017, 243: 133-141.
[8] 张号逵,李映,姜晔楠. 深度学习在高光谱图像分类领域的研究现状与展望[J]. 自动化学报, 2018, 44(6): 961-977.
ZHANG Haokui, LI Ying, JIANG Yenan. Deep learning for hyperspectral imagery classification: the state of the art and prospects [J]. Acta Automatica Sinica, 2018, 44(6): 961-977. (in Chinese)
[9] KRAWCZYK B. Learning from imbalanced data: open challenges and future directions[J]. Progress in Artificial Intelligence, 2016, 5(4): 221-232.
[10] ZHANG X, SONG Q, ZHENG Y, et al. Classification of imbalanced hyperspectral imagery data using support vector sampling [C]// Geoscience and Remote Sensing Symposium. IEEE, 2014: 2870-2873.
[11] SUN T, JIAO L, FENG J, et al. Imbalanced hyperspectral image classification based on maximum margin[J]. IEEE Geoscience & Remote Sensing Letters, 2015, 12(3): 522-526.
[12] LI J, DU Q, LI W, et al. Representation-based hyperspectral image classification with imbalanced data[C]// Geoscience and Remote Sensing Symposium. IEEE, 2016: 3318-3321.
[13] CHAO S, CHU H. The imbalanced hyperspectral image classification based on sparse MK-LSSVM[J]. Urban Geotechnical Investigation & Surveying, 2016(2): 69-73.
[14] LI J, DU Q, LI Y, et al. Hyperspectral image classification with imbalanced data based on orthogonal complement subspace projection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(7): 3838-3851.
[15] HE H, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]// 2008 IEEE International Joint Conference on Neural Networks. IEEE, 2008: 1322-1328.
[16] ZHANG L, WANG W X. A re-sampling method for class imbalance learning with credit data[C]// International Conference of Information Technology, Computer Engineering and Management Sciences. IEEE Computer Society, 2011: 393-397.
[17] HE F, YANG H, MIAO Y, et al. A cost sensitive and class-imbalance classification method based on neural network for disease diagnosis [C]// International Conference on Information Technology in Medicine and Education. IEEE, 2017: 7-10.
[18] NGUYEN H M, COOPER E W, KAMEI K. A comparative study on sampling techniques for handling class imbalance in streaming data[C]// Joint, International Conference on Soft Computing and Intelligent Systems. IEEE, 2013: 1762-1767.
[19] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.
[20] GARCÍA V, SáNCHEZ J S, MOLLINEDA R A. Classification of high dimensional and imbalanced hyperspectral imagery data [C]// Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2011: 644-6511.
[21] GRAVES S, ASNER G, MARTIN R, et al. Tree species abundance predictions in a tropical agricultural landscape with a supervised classification model and imbalanced data[J]. Remote Sensing, 2016, 8(2): 1-22.
[22] GALL J, RAZAVI N, GOOL L V. An introduction to random forests for multi-class object detection[M]. Outdoor and Large-Scale Real-World Scene Analysis. Springer Berlin Heidelberg, 2012: 243-263.
[23] SCHUBACH M, RE M, ROBINSON P N, et al. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants [J]. Scientific Reports, 2017, 7(1): 2959.
[24] WANG Y, LI X, TAO B. Improving classification of mature microRNA by solving class imbalance problem[J]. Scientific Reports, 2016, 6: 25941.
[25] BRANCO P, RIBEIRO R P. A survey of predictive modeling on imbalanced domains[M]. ACM, 2016.
[26] SAIN H, PURNAMI S W. Combine sampling support vector machine for imbalanced data classification[J]. Procedia Computer Science, 2015, 72(1): 59-66.
[27] HAN J, KAMBER M. Data mining: concepts and techniques[M]. 3 rd. Morgan Kaufmann Publishers Inc., 2012: 113-115.
[28] BISHOP C M, NASRABADI N M. Pattern recognition and machine learning[M]. Academic Press, 2006: 461-462.
[29] ULLRICH C. Support vector classification[M]. Springer Berlin Heidelberg, 2009: 345-356.
[30] SHI Y, TIAN Y, KOU G, et al. Support vector machines for multi-class classification problems[M]. Springer London, 2011: 47-60.
[31] CHATFIELD, DOUGLAS G. Practical statistics for medical research[J]. Anz Journal of Surgery, 2010, 61(12): 963-964.
[32] CUTLER A, CUTLER D R, STEVENS J R. Random forests[J]. Machine Learning, 2004, 45(1): 157-176.
[33] MELGANI F, BRUZZONE L. Classification of hyperspectral remote sensing images with support vector machines[J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42(8): 1778-1790.