基于加权正则化协同表示的非均衡分类方法

李艳婷[①] 王 帅[①] 金军委*[②] 马江涛[①] 陈雪艳[①] 陈俊龙[③]

[①](郑州轻工业大学计算机与通信工程学院 郑州 450001)

[②](河南工业大学人工智能与大数据学院 郑州 450001)

[③](华南理工大学计算机科学与工程学院 广州 510641)

摘 要：协同表示分类器及其变种在模式识别领域展现出优越的识别性能。然而，其成功很大程度上依赖于类别的平衡分布，高度非均衡的类别分布可能会严重影响其有效性。为弥补这一不足，该文把补子空间诱导的正则项引入到协同表示模型框架，使得改进后的正则化模型更具判别性。进一步，为提高非均衡数据集上少数类的识别准确率，根据每类训练样本的表示能力提出一种基于最近子空间的类权学习算法。该算法根据原始数据的先验信息自适应地获取每类的权重并且能够赋予少数类更大的权重，使得最终的分类结果对少数类更加公平。所提模型具有闭式解，这展示了该方法的计算效率。在权威公开的两类和多类非均衡数据集上的实验结果表明所提方法显著优于其他主流非均衡分类算法。

关键词：非均衡分类；自适应权重；补子空间；协同表示

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2023)07-2571-09

DOI: 10.11999/JEIT220753

Imbalanced Classification Based on Weighted Regularization Collaborative Representation

LI Yanting[①] WANG Shuai[①] JIN Junwei[②] MA Jiangtao[①]

CHEN Xueyan[①] CHEN Junlong[③]

[①](*College of Computer and Communication Engineering* , *Zhengzhou University of Light Industry*, *Zhengzhou* 450001, *China*)

[②](*College of Artifificial Intelligence and Big Data*, *Henan University of Technology*, *Zhengzhou* 450001, *China*)

[③](*School of Computer Science and Engineering*, *South China University of Technology*, *Guangzhou* 510641, *China*)

Abstract: Collaborative representation based classifier and its variants exhibit superior recognition performance in the field of pattern recognition. However, their success relies greatly on the balanced distribution of classes, and a highly imbalanced class distribution may seriously affect their effectiveness. To make up for this defect, this paper introduces the regularization term induced by the complemented subspace into the framework of collaborative representation model, which makes the improved regularization model more discriminative. Furthermore, in order to improve the recognition accuracy of the minority classes on imbalanced datasets, a class weight learning algorithm based on the nearest subspace is proposed according to the representation ability of each class of training samples. The algorithm obtains adaptively the weight of each class and can assign greater weights to the minority classes, so that the final classification results are more fair to the minority classes. The proposed model has a closed-form solution, which demonstrates its computational efficiency. Experimental results on authoritative public binary-class and multi-class imbalanced datasets show that the proposed method outperforms significantly other mainstream imbalanced classification algorithms.

Key words: Imbalanced classification; Adaptive weight; Complemented subspace; Collaborative representation

收稿日期：2022-06-27；改回日期：2023-03-30；网络出版：2023-03-31

*通信作者： 金军委 jinjunwei24@163.com

基金项目：国家自然科学基金(62106233, 62106068)，河南省科技攻关项目(222102210058, 222102210027, 202102210122)

Foundation Items: The National Natural Science Foundation of China (62106233, 62106068), The Science and Technology Research Project of Henan Province (222102210058, 222102210027, 202102210122)

1 引言

由于现实世界中非均衡数据集的普遍存在性，非均衡分类算法的研究已成为机器学习和模式识别领域的热点问题。非均衡分类旨在实现整个数据集特别是少数类的精准预测，其应用场景包括故障检测、疾病诊断、信息安全等[1]。大多数传统分类算法是在数据集均衡分布的假设下运行的，当它们处理非均衡问题时，其分类结果通常偏向于占主导地位的多数类，而对少数类的识别准确率偏低[2,3]。然而在实践中，少数类通常比多数类包含更重要且更有价值的信息，一旦误判可能会产生严重后果。例如，将入侵判断为正常行为可能引发重大网络安全事故；将癌症患者误诊为健康人会延误最佳治疗时间并威胁患者生命。因此，设计出有效的非均衡分类方法十分必要且迫切[4]。

现有的非均衡分类方法大致分为基于数据层面和基于算法层面两大类型[1]。前者的核心思想是通过采样技术均衡化类别分布来提高分类性能；而后者主要是优化现有的分类算法使其适应非均衡分类问题[5]。在各种采样技术中，最具代表性的是随机欠采样 (Random Under-Sampling, RUS) 和随机过采样 (Random Over-Sampling, ROS)。它们分别从多数类中随机删除样本和从少数类中随机复制实例以缩小少数类和多数类间的差距。但是RUS可能会删除一些重要的样本，而 ROS 可能会添加冗余信息。因此，开发合适的采样技术对于非均衡学习至关重要。到目前为止，科研人员已在欠采样中引入了数据清理和集成学习等一系列技术以提高欠采样性能[6]。有结果表明，对于不平衡率大于10的严重非平衡数据集，过采样通常比欠采样更有效。因此，越来越多的科研人员集中于研究过采样方法，其中一个典型算法是少数类合成过采样 (Synthetic Minority Over-sampling TEchnique, SMOTE)[7]。它随机选取少数类样本作为采样种子通过线性插值来添加少数类实例，但其有可能合成无效的少数类数据。随后，SMOTE 的很多变种比如自适应合成采样(ADAptive SYNthetic sampling, ADASYN) [8]、多数类加权少数类采样 (Majority Weighted Minority Oversampling TEchnique, MWMOTE)[9]、SMOTEENN[10] 等方法被提出以生成高质量的少数类实例。目前，SMOTE的改进算法主要集中在如何选取信息丰富的少数类样本作为采样种子和如何选择合适的插值方式来合成新的数据。Douzas等人[11]提出在每个选定的少数类样本的适当几何区域内合成新的样本以提升SMOTE的合成机制。Wang等人[12]提出一种基于局部分布的少数类过采样方法来选择信息丰富的少数类样本作为种子。Chen等人[13]基于K-means聚类自适应地选取采样种子来合成少数类数据。然而这些方法都是通过线性插值来合成少数类实例，Xie等人[14]基于非线性的Gauss分布模型进行插值从而进一步提高了SMOTE的性能。但不论过采样还是欠采样都改变了原始数据的分布，破坏了原始数据间的关系，使得最终的识别精度无法得到保证。本文致力于探究基于算法层面的方法。

在基于算法层面中，最具代表性的类型是基于代价敏感的非均衡学习方法[15]。该类型方法通过在决策阶段对错误分类的少数类样本赋予更高的惩罚来提高分类准确率。文献 [16] 在极限学习机的基础上基于代价敏感学习提出加权极限学习机 (Weighted Extreme Learning Machine, WELM) 算法。进一步，文献 [17] 提出方差约束的加权极限学习机 (Variances-constrained Weighted Extreme Learning Machine, VW-ELM) 来解决高度非均衡分类问题。合理代价矩阵的构建对于此类方法至关重要。在实践中，如何确定合适的代价矩阵仍然是项艰巨的任务。另一种类型的方法通过修改传统分类方法的目标函数使分类器对少数类更公平，其中稀疏表示的变种——基于稀疏监督表示的分类器 (Spare Supervised Representation-based Classifier, SSRC)[1] 表现出极大的优势。它通过引入标签信息和权重提高了少数类的分类准确率，但极高的计算复杂度限制其进一步发展和应用。受此启发，本文选用高效且复杂度低的协同表示分类器 (Collaborative Representation-based Classifier, CRC)[18]作为基础模型，在延续它已有优势的基础上着重弥补其在少数类上分类的不足。本文提出一种加权正则化协同表示 (Weighted Regularization Collaborative Representation, WRCR) 分类方法。首先分析现有的CRC方法及其变种的缺陷，然后从集合论的角度引入基于补子空间的正则项。进而根据每类训练样本对测试样本的表示能力赋予少数类更大的权重以提高整体特别是少数类的分类准确率。

本文的其余部分结构如下。第2节提出基于补子空间的加权正则化协同表示模型WRCR并给出它的优化求解过程。第3节给出实验对比结果和分析。第4节对全文工作进行总结。

2 加权正则化协同表示模型

本节详细介绍所提的WRCR模型。首先在CRC中引入由补子空间诱导的正则项，然后提出一种自适应类权学习方法，最后给出WRCR的优

化求解和分类准则。这里先给出本文常用的符号。$\boldsymbol{D}=[\boldsymbol{D}_1,\cdots,\boldsymbol{D}_n,\cdots,\boldsymbol{D}_N]$表示整个训练样本集，$\boldsymbol{D}_n\in R^{d\times M_n}$表示第$n$类的训练样本集，其中$N$表示总类别，$d$表示训练样本的特征维度，$M_n$表示第$n$类训练样本的个数。$M=M_1+M_2+\cdots+M_N$表示所有训练样本的个数。$\boldsymbol{x}\in R^d$表示一个测试样本，$\boldsymbol{c}=[\boldsymbol{c}_1;\cdots;\boldsymbol{c}_n;\cdots;\boldsymbol{c}_N]$是$\boldsymbol{x}$在$\boldsymbol{D}$上的表示系数向量，$\boldsymbol{c}^*=[\boldsymbol{c}_1^*;\cdots;\boldsymbol{c}_n^*;\cdots;\boldsymbol{c}_N^*]$是$\boldsymbol{x}$在$\boldsymbol{D}$上的最优表示系数向量。$\boldsymbol{c}_n$是$\boldsymbol{x}$在$\boldsymbol{D}_n$上的表示系数向量，$\boldsymbol{c}_n^*$是$\boldsymbol{x}$在$\boldsymbol{D}_n$上最优表示系数向量。$\boldsymbol{I}$是单位矩阵，$\gamma$和$\alpha$表示正则化参数，$\beta_n$表示第$n$类的权重。

2.1 WRCR模型

CRC模型由于简单、便于操作、识别准确率高已成为均衡分类问题的研究热点。然而，该方法对于解决非均衡数据分类问题并没有明显优势。图1的两个混淆矩阵显示CRC在两类以及多类非均衡数据上表现不佳。主要原因是其模型把所有的训练样本看成一个整体来分类，没有考虑少数类与多数类间样本个数的差异性。作为CRC的变种，竞争协同表示分类(Competitive-Collaborative Representation based Classification, CCRC)[19]模型虽然在每类样本间引入了竞争机制，但该方法仍具有一定缺陷。假设测试样本$\boldsymbol{x}$的真实标签是k，CCRC希望类内损失$||\boldsymbol{x}-\boldsymbol{D}_k\boldsymbol{c}_k||_2^2$尽可能小，类间损失$\left\{\|\boldsymbol{x}-\boldsymbol{D}_k\boldsymbol{c}_k\|_2^2\right\}_{n=1,n\neq k}^{N}$尽可能大。由于真实标签未知，该模型最小化了所有的损失之和$\sum_{n=1}^{N}||\boldsymbol{x}-\boldsymbol{D}_n\boldsymbol{c}_n||_2^2$。这似乎是合理的，但事实并非如此。由CCRC的分类准则可知，类间损失$\left\{\|\boldsymbol{x}-\boldsymbol{D}_k\boldsymbol{c}_k\|_2^2\right\}_{n=1,n\neq k}^{N}$越大越有利于分类，这与CCRC的最小化正则项$\sum_{n=1}^{N}||\boldsymbol{x}-\boldsymbol{D}_n\boldsymbol{c}_n||_2^2$相违背。本文从集合论的角度提出一种更具判别性的基于补子空间的CRC模型。由线性表示理论的假设可知，$E=\text{span}\{\boldsymbol{D}\}$和$E_n=\text{span}\{\boldsymbol{D}_n\}$分别表示整个训练样本集张成的全空间和第$n$类训练样本集张成的子空间。首先定义它们的和

$$E_m+E_n=\{p+q:p\in E_m,q\in E_n\}\tag{1}$$

则

$$E_m+E_n=\text{span}\{\boldsymbol{D}_m\cup\boldsymbol{D}_n\}\tag{2}$$

因此，全空间E可表示为所有子空间之和

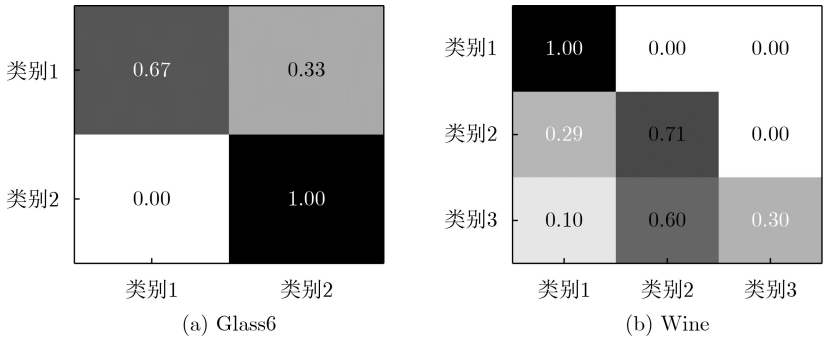
$$E=E_1+E_2+\cdots+E_N\tag{3}$$

我们将$E-E_n$定义为E_n的补子空间，则$E-E_n=\text{span}\{\boldsymbol{D}_{-n}\}$，其中$\boldsymbol{D}_{-n}$是指剔除第$n$类训练样本后剩余的训练样本集。那么$\left\{\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2\right\}_{n=1,n\neq k}^{N}$反映了补子空间$\{E-E_n\}_{n=1,n\neq k}^{N}$对测试样$\boldsymbol{x}$的表示能力。由于当$n\neq k$时，$E_k\subseteq E-E_n$，所以损失项$\left\{\|\boldsymbol{x}-\boldsymbol{D}_{-n}c_{-n}\|_2^2\right\}_{n=1,n\neq k}^{N}$应该越小越好，这里$\boldsymbol{c}_{-n}$是$\boldsymbol{D}_{-n}$对$x$的表示系数。因此将$\sum_{n=1,n\neq k}^{N}\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2$最小化是合理的。另外，由于测试样本$\boldsymbol{x}$真实标签未知，则需要加上$\|\boldsymbol{x}-\boldsymbol{D}_{-k}\boldsymbol{c}_{-k}\|_2^2$的最小化。目标函数初步确定为

$$\begin{aligned}c^*=&\arg\min_c\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\gamma\|\boldsymbol{c}\|_2^2\\&+\alpha\sum_{n=1}^{N}\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2\end{aligned}\tag{4}$$

尽管上述基于补子空间的协同表示模型继承了CRC简单高效的特性并提高了判别性，但其在处理非均衡分类问题时没有充分考虑类别分布信息，这会导致多数类的表示能力远超过少数类，从而使得最终的分类结果倾向于多数类。特别对于严重非均衡数据集，少数类由于训练样本个数太少，对测试样本的表示能力极低，从而进一步加大了少数类的重构误差，不利于最终的分类。利用这些先验信息，我们基于最近子空间分类 (Nearest Subspace Classification, NSC)[20]方法赋予不同类别不同的权重，并且增大少数类的权重，使得最终的分类结果对少数类更加公平。目标函数表示为

$$\begin{aligned}\boldsymbol{c}^*=&\arg\min_c\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\gamma\|\boldsymbol{c}\|_2^2\\&+\alpha\sum_{n=1}^{N}\beta_n\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2\end{aligned}\tag{5}$$



(a) Glass6　　(b) Wine

图 1 CRC在两个非均衡数据集上的混淆矩阵

其中，$\beta_n(n=1,2,\cdots,N)$是要学习的类权。我们把该模型称为加权正则化协同表示(WRCR)模型。

2.2 基于NSC的类权学习

我们利用NSC中每类训练样本集对测试样本的重构误差$\{\|\boldsymbol{x}-\boldsymbol{D}_n\boldsymbol{c}_n\|_2\}_{n=1}^N$来学习类权。由NSC模型可知，第$n$类训练样本对测试样本$\boldsymbol{x}$的最优表示系数为

$$\hat{\boldsymbol{c}}_n=\left(\boldsymbol{D}_n^{\mathrm{T}}\boldsymbol{D}_n\right)^{-1}\boldsymbol{D}_n^{\mathrm{T}}\boldsymbol{x} \tag{6}$$

第n类训练样本集$\boldsymbol{D}_n$对$\boldsymbol{x}$的重构误差为

$$r_n=\|\boldsymbol{x}-\boldsymbol{D}_n\hat{\boldsymbol{c}}_n\|_2 \tag{7}$$

我们定义最小重构误差

$$r_{\min}=\min\{r_n\} \tag{8}$$

显然，r_n越小，$\boldsymbol{D}_n$对$\boldsymbol{x}$的表示能力越强，$\boldsymbol{x}$属于第n类训练样本空间的概率就越大。针对非均衡分类问题，我们进一步发现少数类对测试样本的表示能力一般弱于多数类。这里，用图2展示的实验结果来解释这一现象。具体地，我们选用2个两类数据集Glass6，Newthyroid1和2个3类数据集Wine，Newthyroid作为基准数据集。首先计算每个测试样本在各类训练集中的重构误差占总体重构误差的比重，然后取所有测试样本的重构误差所占比重的平均值。如图2所示，对于两类数据集，多数类的重构误差远小于少数类的。对于3类数据集，我们分别标注各类为少数类、中间类、多数类，则可看出中间类的重构误差小于少数类的，多数类的重构误差小于中间类的。这可说明不论是两分类还是多分类问题，少数类的重构误差一般大于多数类的。也即测试样本属于少数类的概率一般小于属于多数类的概率。基于此，定义各类的类别权重β_n为

$$\beta_n=\exp\left(\frac{r_{\min}-r_n}{\delta}\right) \tag{9}$$

其中，$\delta>0$是调节类权的伸缩参数。显然该方法

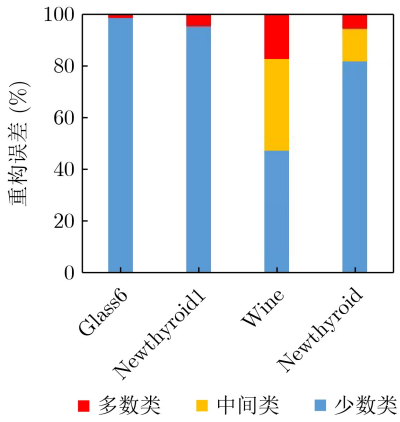


图 2 测试样本在各类训练集中的重构误差占总体重构误差的比重

能够赋予不同类不同的权重。不仅如此，它还能够使少数类获得更大的权重。这里，以两分类和多分类的情况分别来说明。首先，对于两分类，我们假设第1类是少数类。则由前面分析可知，重构误差$r_1>r_2$。那么$r_{\min}=r_2$。所以$\beta_1=\exp\left(\frac{r_2-r_1}{\delta}\right)<1$，$\beta_2=1$。又因为$\boldsymbol{D}_{-1}=\boldsymbol{D}_2$且$\boldsymbol{D}_{-2}=\boldsymbol{D}_1$，WRCR中引入的正则项$\sum_{n=1}^{2}\beta_n\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2$可写为$\beta_1\|\boldsymbol{x}-\boldsymbol{D}_2\boldsymbol{c}_2\|_2^2+\beta_2\|\boldsymbol{x}-\boldsymbol{D}_1\boldsymbol{c}_1\|_2^2$。由于$\beta_2>\beta_1$且第1类训练样本集$\boldsymbol{D}_1$的类权为$\beta_2$，我们得到基于NSC的类权学习算法能够赋予少数类更大的权重。对于多分类，假设总类别数为N且各类的训练样本个数$M_1<M_2<\cdots<M_N$，其中M_n表示第n类训练样本的个数。则我们得到重构误差$r_1>r_2>\cdots>r_N$。那么$r_{\min}=r_N$，$\beta_n=\exp\left(\frac{r_N-r_n}{\delta}\right)$。因此权重$\beta_1<\beta_2<\cdots<\beta_N$。由于$\boldsymbol{D}_{-n}$表示的是剔除了第$n$类训练样本后剩余的训练样本集，则$\boldsymbol{D}_{-n}$的样本个数为$M-M_n$，其中$M$表示训练样本总数。那么$M-M_1>M-M_2>\cdots>M-M_N$。因此WRCR中引入的正则项$\sum_{n=1}^{N}\beta_n\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2$使得训练样本集$\boldsymbol{D}_{-n}$的样本个数越小，它所对应的权重$\beta_n$越大。

2.3 WRCR的优化求解和分类准则

为解决WRCR模型中目标函数的最小化问题，首先定义一个新的矩阵$\bar{\boldsymbol{D}}_{-n}=[\boldsymbol{D}_1,\cdots,\boldsymbol{D}_{n-1},0,\boldsymbol{D}_{n+1},\cdots,\boldsymbol{D}_N]$。下面的定理保证了该模型具有闭式解。

定理1 给出类权β_n，则求解WRCR模型

$$\boldsymbol{c}^*=\left(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\alpha\sum_{n=1}^{N}\beta_n\bar{\boldsymbol{D}}_{-n}^{\mathrm{T}}\bar{\boldsymbol{D}}_{-n}+\gamma\boldsymbol{I}\right)^{-1}\cdot\left(\boldsymbol{D}+\alpha\sum_{n=1}^{N}\beta_n\bar{\boldsymbol{D}}_{-n}\right)^{\mathrm{T}}x \tag{10}$$

证明 我们看到WRCR的目标函数是关于c的凸可微函数，所以对其求导获得的极值点即为最小值点。为了便于计算我们将目标函数定义为ϑ

$$\vartheta=\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{c}\|_2^2+\gamma\|\boldsymbol{c}\|_2^2+\alpha\sum_{n=1}^{N}\beta_n\|\boldsymbol{x}-\boldsymbol{D}_{-n}\boldsymbol{c}_{-n}\|_2^2 \tag{11}$$

则对其求导并令导数为0得到

$$\begin{aligned}\frac{\partial\vartheta}{\partial c}=&-2\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{D}\boldsymbol{c})+2\gamma\boldsymbol{c}\\&+\alpha\sum_{n=1}^{N}\beta_n\left[-2\bar{\boldsymbol{D}}_{-n}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{D}}_{-n}\boldsymbol{c})\right]\\=&\,0\end{aligned}$$

所以

$$\boldsymbol{c}^{*}=\left(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\alpha\sum_{n=1}^{N}\beta_{n}\bar{\boldsymbol{D}}_{-n}^{\mathrm{T}}\bar{\boldsymbol{D}}_{-n}+\gamma I\right)^{-1}\cdot\left(\boldsymbol{D}+\alpha\sum_{n=1}^{N}\beta_{n}\bar{\boldsymbol{D}}_{-n}\right)^{\mathrm{T}}\boldsymbol{x}$$

证毕

求出最优表示系数$\boldsymbol{c}^{*}$后，我们计算每类的重构误差

$$r_{n}\left(\boldsymbol{x}\right)=\left\|\boldsymbol{x}-\boldsymbol{D}_{n}\boldsymbol{c}_{n}^{*}\right\|_{2},\ n=1,2,\cdots,N \tag{12}$$

通过最小重构误差准则来确定$\boldsymbol{x}$的类别

$$\mathrm{label}\left(\boldsymbol{x}\right)=\arg\min_{n}r_{n}\left(\boldsymbol{x}\right) \tag{13}$$

2.4 计算复杂度

由WRCR算法可知主要的计算复杂度取决于最优表示系数向量$\boldsymbol{c}^{*}$的求解。而矩阵的乘法和求逆运算占据$\boldsymbol{c}^{*}$的主要计算量。因此WRCR的计算复杂度为$O\left(dM^{2}+M^{3}\right)$，其中$d$表示训练样本的特征维度，$M$表示训练样本集的总个数。它和CRC具有同样低的计算复杂度，能够确保该方法的高效性。

3 实验结果

本文实验使用UCI[21]中的非均衡数据集进行评估，通过与基于CRC的分类方法和多个非均衡分类方法的对比来说明所提方法的有效性。

3.1 数据集

本文实验使用了UCI的9个两类和7个多类的非均衡数据集。这些数据集的详细特征信息如表1所述。类别分布表示每类样本的个数之比，不平衡率表示最多类的样本个数与最少类的样本个数之比。由表1看出，我们使用的数据库的不平衡率跨越范围较大，从1.10变动到71.51。不平衡率越高，准确分类的困难程度越大。

3.2 实验设置

在处理非均衡分类问题时，分类准确率不能有效评估非均衡分类算法的性能。这里，我们采用F-measure和G-mean来度量分类性能。不论两分类还是多分类，F-measure和G-mean越大，算法的分类性能越高。实验中我们使用5折交叉验证法，每个数据集被随机分为5个子集，选出1个子集作为测试集其余4个作为训练集。此方法被随机试验10次，取10次的平均值作为最终实验结果。在具体实验中，每个对比模型所涉及的参数都经过仔细调节使实验结果达到最优。对于所提的WRCR模型，γ, α, δ 3个参数对模型的性能评估至关重要。这里，我们设置γ, α的候选集为$\{10^{-1},10^{-2},\cdots,10^{-15}\}$, δ的候选集为$\{1,2,\cdots,10,10^{2},10^{3},10^{4},10^{5},10^{6}\}$。对这3个参数实施网格搜索算法来获取最优的实验结果。所有实验均以MATLAB为编程语言在CPU i5-8500和运行内存7.84 GB的笔记本端进行。

3.3 与基于CRC的分类方法的对比

由于WRCR方法是以CRC为基础模型通过借鉴CCRC的竞争机制和基于NSC赋予类权来解决非均衡分类问题，我们给出它和CRC, CCRC, NSC

表 1　16个非均衡数据集的详细信息

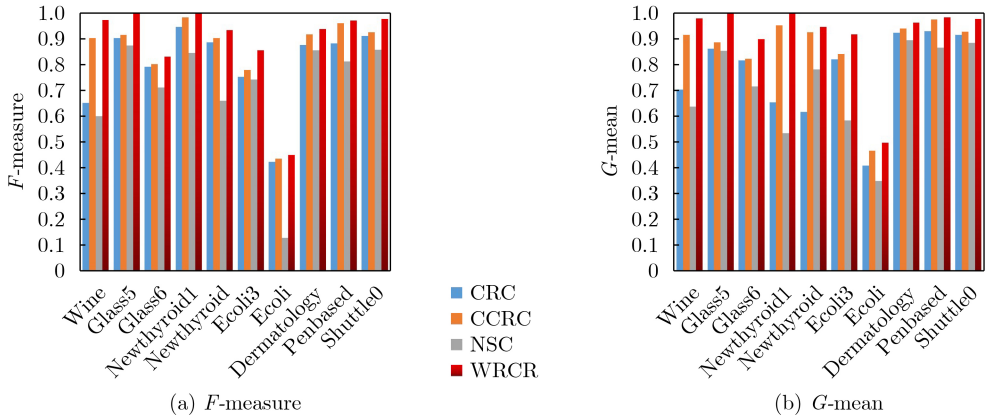
数据集	类别	样本总数	维度	类别分布	不平衡率
Wine	3	178	13	59: 71: 48	1.48
Glass5	2	214	9	9: 205	22.78
Glass6	2	214	9	29: 185	6.38
Newthyroid1	2	215	5	35: 180	5.14
Newthyroid	3	215	5	150: 35: 30	5.00
Ecoli3	2	336	7	35: 301	8.60
Ecoli	8	336	7	143: 77: 2: 2: 35: 20：5：52	71.51
Dermatology	6	366	33	111: 60: 71: 48: 48: 20	5.55
Penbased	10	1 100	16	115: 114: 114: 106: 114: 106: 105: 115: 105: 106	1.10
Shuttle0	2	1829	9	123: 1 706	13.87
Ecoli0vs1	2	220	7	77: 143	1.86
Balance-scale	3	625	4	49: 288: 288	5.88
ShuttleC0vsC4	2	1829	9	123: 1 706	13.86
Glass4	2	214	9	13: 201	15.46
Glass	3	163	4	70: 76: 17	4.47
Glass016vs2	2	192	9	17: 175	10.29

3个方法的实验对比结果以说明其有效性。图3以直方图的形式直观地展示了这4个基于CRC的方法在10个数据集上的F-measure和G-mean值。显然看出，NSC表现最差，CRC的分类结果略高于NSC，CCRC相比于CRC有明显提升，而所提的WRCR方法显著高于其它3种方法。为了展示WRCR在特定类别中的性能，图4给出了它在两类数据集 Glass6和多类数据集Wine上的各类识别结果。与图1的CRC方法相比，WRCR显著提高了少数类样本的识别准确率。WRCR的优越识别性能主要是由于以下两个方面： 一是在CRC模型中加入了更具判别性的基于补子空间的正则项；二是赋予了不同类别不同的权重使得分类结果对少数类更加公平。除了分类准确率，我们还测试了4种方法的运算效率。表2列出了Glass5数据集上一个测试样本的运行时间。可看到，WRCR运算速度很快并且和NSC, CRC, CCRC消耗了同等量级的运算时间，这也验证了WRCR算法的高效性。

3.4 与非均衡分类方法的对比

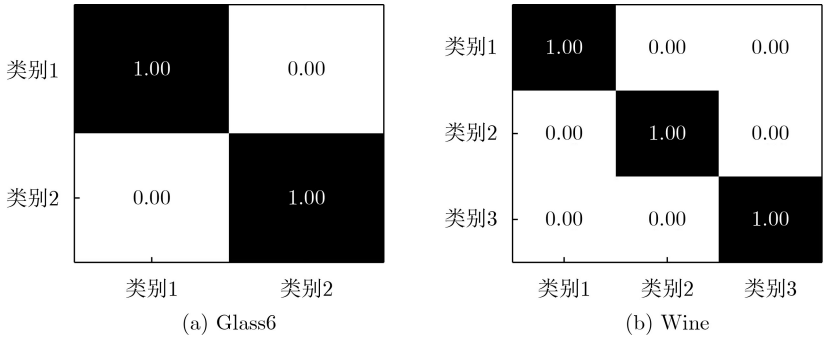
为了展示所提方法的有效性，本文将其与经典的非均衡分类方法RUS, ADASYN[8], SMOTE[7], MWMOTE[9], WELM[16], SMOTEENN[10], EasyEnsemble[22] 进行对比。表3和表4分别给出了WRCR与这些非均衡分类方法在F-measure和G-mean上的对比结果。最好的实验结果加粗显示，可清楚看到WRCR在其中14个数据集上的F-measure和G-mean值都高于对比方法。在另两个数据集Glass6和Ecoli上， 虽然它的F-measure稍低于对比方法，但G-mean远超其他方法。另外，WRCR在Wine, Glass5, Newthyroid1, Ecoli0vs1上的识别准确率能够达到100%。特别对于不平衡率高达22.78的严重非均衡数据集Glass5，WRCR不仅能够实现每类样本的精准识别，而且比其他方法的识别效果提升10%以上。

为了进一步展现WRCR方法的优越性能，我们将它和5种先进的非均衡分类方法GDO[14]，



(a) F-measure (b) G-mean

图 3 基于CRC的不同方法在10个非均衡数据集上的对比



(a) Glass6 (b) Wine

图 4 WRCR在两个数据集上的混淆矩阵

表 2 不同方法在Glass5数据集上的运行时间 (s)

	NSC	CRC	CCRC	WRCR
运行时间 (s)	3.19×10^{-3}	4.72×10^{-3}	7.35×10^{-3}	8.07×10^{-3}

VW-ELM [17], GEP[23], GMBSCL[5] 和GSE[6] 进行对比。值得一提的是这些对比方法都是近3年发表在国际Top期刊上并声称在我们使用的某些数据集上获得了最好的性能。这里，我们直接从这些方法的原论文中引用其实验结果。表5总结了不同方法在4个共有数据集Glass5, Newthyroid1, Ecoli0vs1, Ecoli3上的G-mean结果。可以发现，WRCR表现最好，这充分证实了所提方法可达到最先进的性能。

综上所述，所提方法WRCR能够提高非均衡数据集特别是少数类的分类准确率，从而有效解决非均衡分类问题。

4　结论

本文提出一种加权正则化协同表示的非均衡分类算法。它解决了CRC及其变种在非均衡数据集上分类效果不佳的问题。其关键是在CRC建模过

表 3　WRCR与经典非均衡算法在16个数据集上的F-measure (%) 值对比

数据集	ADASYN	SMOTEENN	WELM	RUS	SMOTE	MWMOTE	EasyEnsemble	WRCR
Wine	89.01	87.12	88.63	89.05	89.03	89.82	89.51	**100.00**
Glass5	77.44	77.86	64.31	87.15	68.72	79.22	88.42	**100.00**
Glass6	88.61	89.23	82.72	82.51	83.14	83.52	85.42	**90.04**
Newthyroid1	97.52	97.93	97.05	94.52	95.46	92.17	94.34	**100.00**
Newthyroid	92.55	92.61	90.44	93.26	91.72	92.81	93.22	**94.77**
Ecoli3	87.61	86.63	88.62	84.13	87.46	81.13	88.22	**98.35**
Ecoli	29.91	38.92	30.14	35.32	33.90	34.82	27.14	**53.10**
Dermatology	92.81	89.91	91.33	92.37	92.24	92.11	78.72	**96.25**
Penbased	95.63	97.52	97.85	97.31	98.40	95.82	90.52	**98.40**
Shuttle0	88.42	84.62	97.41	80.43	82.72	81.32	89.41	**97.87**
Ecoli0vs1	95.72	94.17	98.51	91.34	94.69	96.56	97.75	**100.00**
Balance-scale	54.26	52.47	51.38	47.59	50.58	54.63	55.76	**61.70**
ShuttleC0vsC4	93.96	89.35	96.47	91.25	85.19	93.42	81.38	**97.89**
Glass4	90.33	93.66	91.34	92.48	90.33	94.16	94.42	**96.18**
Glass	48.59	51.36	54.81	48.75	49.65	50.23	51.48	**56.06**
Glass016vs2	58.11	59.19	83.77	62.47	61.36	69.82	66.78	**84.09**

表 4　WRCR与经典非均衡算法在16个数据集上的G-mean (%) 值对比

数据集	ADASYN	SMOTEENN	WELM	RUS	SMOTE	MWMOTE	EasyEnsemble	WRCR
Wine	84.11	80.63	94.51	83.15	83.41	84.53	88.62	**100.00**
Glass5	88.13	90.52	88.92	91.24	87.52	89.74	88.64	**100.00**
Glass6	88.64	**89.22**	82.73	82.53	83.16	83.01	85.41	83.33
Newthyroid1	95.65	98.23	97.44	96.82	95.07	94.42	94.33	**100.00**
Newthyroid	90.53	90.42	89.91	87.23	91.74	92.43	89.14	**93.63**
Ecoli3	83.02	82.51	84.83	82.32	84.23	82.83	84.63	**87.49**
Ecoli	**62.31**	46.54	38.92	36.74	60.05	60.22	33.86	50.07
Dermatology	87.32	81.43	87.25	76.13	86.34	89.73	74.14	**93.98**
Penbased	91.83	95.52	95.36	91.51	94.34	93.15	87.92	**97.18**
Shuttle0	87.61	97.21	97.41	97.65	84.81	85.20	92.41	**97.87**
Ecoli0vs1	91.54	90.34	98.55	89.23	91.38	94.76	94.84	**100.00**
Balance-scale	52.83	54.37	50.65	48.98	54.76	52.42	55.78	**61.68**
ShuttleC0vsC4	92.51	86.76	92.36	90.18	83.57	91.83	87.38	**97.87**
Glass4	54.47	51.85	61.18	53.32	52.49	57.39	59.42	**66.66**
Glass	42.36	40.46	39.47	36.53	39.76	38.76	41.04	**44.01**
Glass016vs2	45.48	47.69	47.89	45.83	49.67	51.28	53.89	**66.66**

表 5 WRCR与先进非均衡算法的*G*-mean (%) 值对比

数据集	GDO	VW-ELM	GEP	GMBSCL	GSE	WRCR
Glass5	84.10	97.51	95.85	91.50	–	**100.00**
Newthyroid1	89.99	99.52	97.33	–	–	**100.00**
Ecoli0vs1	95.16	98.64	98.32	98.31	97.58	**100.00**
Ecoli3	88.67	91.20	92.57	–	88.53	**98.35**

程中引入由补子空间诱导的正则项。进一步为了提高少数类的识别准确率，本文提出了一种类权学习算法。该算法根据每类训练样本的表示能力自适应地学习每类的权重，从而赋予少数类更大的权重。所提模型能够以闭式解的形式有效解决。在不同非均衡数据集上的实验结果验证了所提方法的有效性。

参 考 文 献

[1] SHU Ting, ZHANG B, and TANG Yuanyan. Sparse supervised representation-based classifier for uncontrolled and imbalanced classification[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(8): 2847–2856. doi: 10.1109/TNNLS.2018.2884444.

[2] JIN Junwei, LI Yanting, and CHEN C L P. Pattern classification with corrupted labeling via robust broad learning system[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(10): 4959–4971. doi: 10.1109/TKDE.2021.3049540.

[3] JIN Junwei, LI Yanting, YANG Tiejun, *et al*. Discriminative group-sparsity constrained broad learning system for visual recognition[J]. *Information Sciences*, 2021, 576: 800–818. doi: 10.1016/j.ins.2021.06.008.

[4] JIN Junwei, QIN Zhenhao, YU Dengxiu, *et al*. Regularized discriminative broad learning system for image classification[J]. *Knowledge-Based Systems*, 2022, 251: 109306. doi: 10.1016/j.knosys.2022.109306.

[5] ZHU Zonghai, WANG Zhe, LI Dongdong, *et al*. Globalized multiple balanced subsets with collaborative learning for imbalanced data[J]. *IEEE Transactions on Cybernetics*, 2022, 52(4): 2407–2417. doi: 10.1109/TCYB.2020.3001158.

[6] ZHU Zonghai, WANG Zhe, LI Dongdong, *et al*. Geometric structural ensemble learning for imbalanced problems[J]. *IEEE Transactions on Cybernetics*, 2020, 50(4): 1617–1629. doi: 10.1109/TCYB.2018.2877663.

[7] 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: 321–357. doi: 10.1613/jair.953.

[8] HE Haibo, BAI Yang, GARCIA E A, *et al*. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. Proceedings of the International Joint Conference on Neural Networks, Hong Kong, China, 2008: 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[9] BARUA S, ISLAM M M, YAO Xin, *et al*. MWMOTE: Majority weighted minority oversampling technique for imbalanced data set learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(2): 405–425. doi: 10.1109/TKDE.2012.232.

[10] BATISTA G E A P A, PRATI R C, and MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. *ACM SIGKDD Explorations Newsletter*, 2004, 6(1): 20–29. doi: 10.1145/1007730.1007735.

[11] DOUZAS G and BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. *Information Sciences*, 2019, 501: 118–135. doi: 10.1016/j.ins.2019.06.007.

[12] WANG Xinyue, XU Jian, ZENG Tieyong, *et al*. Local distribution-based adaptive minority oversampling for imbalanced data classification[J]. *Neurocomputing*, 2021, 422: 200–213. doi: 10.1016/j.neucom.2020.05.030.

[13] CHEN Baiyun, XIA Shuyin, CHEN Zizhong, *et al*. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise[J]. *Information Sciences*, 2021, 553: 397–428. doi: 10.1016/j.ins.2020.10.013.

[14] XIE Yuxi, QIU Min, ZHANG Haibo, *et al*. Gaussian distribution based oversampling for imbalanced data classification[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(2): 667–679. doi: 10.1109/TKDE.2020.2985965.

[15] CAO Changjie, CUI Zongyong, WANG Liying, *et al*. Cost-sensitive awareness-based SAR automatic target recognition for imbalanced data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1–16. doi: 10.1109/TGRS.2021.3068447.

[16] ZONG Weiwei, HUANG Guangbin, and CHEN Yiqiang. Weighted extreme learning machine for imbalance learning[J]. *Neurocomputing*, 2013, 101: 229–242. doi: 10.1016/j.neucom.2012.08.010.

[17] LIU Zheng, JIN Wei, and MU Ying. Variances-constrained weighted extreme learning machine for imbalanced classification[J]. *Neurocomputing*, 2020, 403: 45–52. doi: 10.1016/j.neucom.2020.04.052.

[18] ZHANG Lei, YANG Meng, and FENG Xiangchu. Sparse

representation or collaborative representation: Which helps face recognition[C]. Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011: 471–478. doi: 10.1109/ICCV.2011.6126277.

[19] YUAN Haoliang, LI Xuecong, XU Fangyuan, *et al*. A collaborative-competitive representation based classifier model[J]. *Neurocomputing*, 2018, 275: 627–635. doi: 10.1016/j.neucom.2017.09.022.

[20] LI Yanting, JIN Junwei, ZHAO Liang, *et al*. A neighborhood prior constrained collaborative representation for classification[J]. *International Journal of Wavelets, Multiresolution and Information Processing*, 2021, 19(2): 2050073. doi: 10.1142/S0219691320500733.

[21] KHAN M M R, ARIF R B, SIDDIQUE M A B, *et al*. Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository[C]. Proceedings of the 4th International Conference on Electrical Engineering and Information & Communication Technology. Dhaka, Bangladesh, 2018: 124–129. doi: 10.1109/CEEICT.2018.8628041.

[22] LIU Xuying, WU Jianxin, and ZHOU Zhihua. Exploratory undersampling for class-imbalance learning[J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*), 2009, 39(2): 539–550. doi: 10.1109/TSMCB.2008.2007853.

[23] JEDRZEJOWICZ J and JEDRZEJOWICZ P. GEP-based classifier for mining imbalanced data[J]. *Expert Systems with Applications*, 2021, 164: 114058. doi: 10.1016/j.eswa.2020.114058.

李艳婷：女，博士，讲师，研究方向为模式识别、人工智能.
王　帅：男，硕士生，研究方向为模式识别、机器学习.
金军委：男，博士，讲师，研究方向为模式识别、人工智能.
马江涛：男，博士，副教授，研究方向为知识图谱、人工智能.
陈雪艳：女，博士，讲师，研究方向为通信工程、人工智能.
陈俊龙：男，教授，博士生导师，研究方向为宽度学习、人工智能等.

责任编辑：陈　倩