计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200138

改进的 XGBoost 在不平衡数据处理中的应用研究

宋玲玲[1]　王时绘[1,2]　杨　超[1,2,3]　盛　潇[1]
1 湖北大学计算机与信息工程学院　武汉 430062
2 湖北省教育信息化工程技术研究中心　武汉 430062
3 湖北大学数学与统计学学院应用数学湖北省重点实验室　武汉 430062
(ling2_song@stu.hubu.edu.cn)

摘　要　传统分类器在处理不平衡数据时,往往会倾向于保证多数类的准确率而牺牲少数类的准确率,导致少数类的误分率较高。针对这一问题,提出一种面向二分类不平衡数据的 XGBoost(eXtreme Gradient Boosting)改进方法。其主要思想是分别从数据、特征以及算法 3 个层面针对不平衡数据的特点进行改进。首先在数据层面,通过条件生成式对抗网络(Conditional Generative Adversarial Nets,CGAN)学习少数类样本的分布信息,训练生成器生成少数类补充样本,调节数据的不平衡性;其次在特征层面,先利用 XGBoost 进行特征组合生成新的特征,再通过最大相关最小冗余(minimal Redundancy-Maximal Relevance,mRMR)算法筛选出更适合不平衡数据分类的特征子集;最后在算法层面,引入针对不平衡数据分类问题的焦点损失函数(Focal Loss)来改进 XGBoost,改进后的 XGBoost 通过新的数据集训练得到最终模型。在实验阶段,选择 G-mean 和 AUC 作为评价指标,6 组 KEEL 数据集上的实验结果验证了所提改进方法的可行性;同时将该方法与现有的 4 种不平衡分类模型进行比较,实验结果表明所提改进方法具有较好的分类效果。

关键词:不平衡数据;XGBoost;Focal Loss;特征组合;mRMR;CGAN
中图法分类号　TP181

Application Research of Improved XGBoost in Imbalanced Data Processing

SONG Ling-ling[1],WANG Shi-hui[1,2],YANG Chao[1,2,3] and SHENG Xiao[1]
1 School of Computer and Information Engineering,Hubei University,Wuhan 430062,China
2 Hubei Provincial Education Information Engineering Technology Research Center,Wuhan 430062,China
3 Hubei Key Laboratory of Applied Mathematics,School of Mathematics and Statistics,Hubei University,Wuhan 430062,China

Abstract　When dealing with imbalanced data,traditional classifiers tend to guarantee the accuracy of the majority class and sacrifice the accuracy of the minority class,resulting in a higher error rate of the minority class. Aiming at this problem,an improved XGBoost method for binary imbalanced data is proposed. The main idea is to improve the characters of imbalanced data from three levels,data,features,and algorithms. Firstly,at the data level,Conditional Generative Adversarial Nets (CGAN) learns the distributive information of minority samples and then trains the generator to generate a few supple-mentary samples to adjust the imbalance of the data. Secondly,at the feature level,it uses XGBoost for feature combination to generate new features,and then uses the minimal Redundancy-Maximal Relevance (mRMR) algorithm to screen out a subset of features that are more suitable for imbalanced data classification. Finally,at the algorithm level,it introduces a Focal Loss function for imbalanced data classification to improve XGBoost. The improved XGBoost is trained on the new dataset to obtain the final model. In the experimental stage,G-mean and AUC are selected as the evaluation indicators. The experimental results on 6 sets of KEEL datasets verify the feasibility of the proposed improved method. At the same time,the method is compared with the existing four imbalanced classification models. The experimental results show that the proposed improved method has better classification effect.

Keywords　Imbalanced data,XGBoost,Focal Loss,Feature combination,mRMR,CGAN

到稿日期:2019-12-23　返修日期:2020-03-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61977021);应用数学湖北省重点实验室开放基金资助项目(HBAM201902)
This work was supported by the National Natural Science Foundation of China(61977021) and Open funded project of Hubei Key Laboratory of Applied Mathematics(HBAM201902).
通信作者:杨超(stevenyc@hubu.edu.cn)

1 引言

不平衡数据指在数据集中某类或者某些类的样本数量远大于其他类的样本数量，样本量较多的类被称为多数类（负类），样本量较少的类被称为少数类（正类）[1]。不平衡数据分类问题在医疗诊断[2-4]、机器故障检测[5-6]、信息安全[7-8]和计算机视觉[9-10]等实际应用中十分常见。在传统分类方法中，分类器为了保证训练模型的整体预测精度最大化，会优先考虑多数类样本的准确率，致使分类决策面偏向于少数类样本，从而导致模型对少数类样本的误分率较高。但在一些应用中，少数类的误分代价巨大。例如在医疗诊断中，将患者诊断为健康人的误分代价远远高于将健康人诊断为患者的误分代价[11]。因此，如何对不平衡数据进行正确分类成为机器学习和数据挖掘领域中的一个研究热点[12]。

不平衡数据分类问题的特点主要体现在数据本身分布不平衡以及传统分类算法在处理不平衡数据时存在局限性这两方面。相关研究人员多年来针对不平衡数据分类问题进行了深入的研究分析，相继提出了很多解决方法。这些解决方法按照不同类型可大概分为以下 3 个层面。

1）数据层面。通过对数据集进行数据分布调整以达到降低数据不平衡性的目的，主要方法是对数据进行重采样或数据分组[13]。重采样通常分为对数据集中的少数类样本进行添加的过采样和对多数类样本进行删除的欠采样。例如，Chawla 等[14]提出的经典过采样技术 SMOTE（Synthetic Minority Oversampling Technique）是通过在少数类样本之间线性生成新的样本来实现，但该算法容易产生样本分布边缘化现象；Cieslak 等[15]提出的欠采样技术是先对多数类进行聚类，然后从聚类后的每个簇中移除一定数量的样本。重采样技术在降低数据不平衡性的同时会引起一些问题，其中过采样人为添加样本会导致数据冗余，增加模型训练的复杂度；欠采样在移除多数类样本的过程中可能会丢失重要数据。数据分组是将不平衡数据集按照一定的划分规则分为多个平衡数据集，再将平衡数据集上训练出的多个分类器集成为一个分类器，以此改善数据类别不平衡问题。例如 Guo 等[16]使用 K-means 算法将多数类数据划分为多个簇，每个簇和少数类数据合并成一个新的相对平衡的数据集，再针对每个新的数据集进行模型训练。然而数据分组将数据进行了硬性分割，可能会破坏数据的整体分布信息。

2）特征层面。数据分布不平衡常常伴随着特征属性分布不均衡，导致在分类时对少数类的识别率不高[12]。高维数据集上的类别不平衡现象比较明显，但数据层面和算法层面的方法不能很好地解决高维不平衡数据分类问题[17]。数据类别不平衡问题中的特征选择指从全部特征集中挑选出更适合不平衡数据分类且更能反映不平衡数据特点的特征子集来构建模型。例如，Wang 等[18]提出了一种针对不平衡文本情感分类的特征选择算法 TSF（Two-Side Fisher），通过显式地组合正相关和负相关特征，减小了特征层面的不平衡性，但 TSF 算法的通用性有待考查。

3）算法层面。算法层面主要是针对传统分类算法对不平衡数据分类存在的局限性进行改进，以提高对少数类的识别率，其中代价敏感学习和集成学习方法尤为突出[19-20]。传统的分类算法通常假定所有样本的误分代价相同，而代价敏感学习更关注实际误分代价较高的类别的样本。在处理不平衡数据时，代价敏感学习通过给少数类赋予更高的误分代价，使得分类器在模型训练的过程中更关注于降低少数类的误分率，从而提高对少数类的分类效果。集成学习通过将多个分类器按照一定的方式集成在一起，来提高分类器的泛化能力，避免单个分类器对不平衡数据分类预测造成的偏差，从而提高分类效果[21]。目前基于集成学习的不平衡数据处理方法主要是将集成学习与其他不平衡数据分类处理方法结合，以全面提高分类效果。例如，Yuan 等[22]提出一种基于代价敏感学习的结构化支持向量机集成分类算法，增加了少数类样本的权重，提高了不平衡数据的分类精度。

基于以上的研究分析，本文提出一种针对不平衡数据二分类问题的 XGBoost 改进方法。该方法分别从数据、特征以及算法 3 个层面针对不平衡数据的特点进行改进。首先，在数据层面借助 CGAN 生成少数类补充样本，以降低数据的不平衡性；其次，在特征层面利用 XGBoost 建立决策树进行特征组合生成新的特征，再通过 mRMR 算法挑选出更适合不平衡数据分类的特征子集；最后，在算法层面引用目标检测中针对不平衡数据分类问题的 Focal Loss 来改进 XGBoost，改进后的 XGBoost 通过新的数据集训练得到最终模型。实验结果表明，本文改进的分类模型与其他 4 种具有代表性的不平衡分类模型（SMOTEBoost[23]，CUSBoost[24]，RUS-Boost[25]以及 KSMOTE-AdaBoost[26]）相比具有更好的分类效果，同时改进方法的阶段性对比实验也验证了改进的可行性。

2 相关算法简介

2.1 CGAN

条件生成式对抗网络（CGAN）是 2014 年提出的一种对生成式对抗网络（GAN）进行改进的深度学习网络，被广泛应用于生成样本[27]。GAN 主要包括两个部分：生成器 G（generator）和判别器 D（discriminator）。生成器 G 的目的是通过学习真实样本的分布使自身生成的样本更加接近真实样本，试图混淆判别器 D；判别器 D 的目的是识别区分原始数据集的真样本和生成器 G 生成的假样本。这个过程相当于两者的博弈，两者不断地进行对抗和相互迭代优化，最终达到一个动态平衡：生成器生成的样本接近于真实样本，判别器对于给定的样本预测为真的概率接近 0.5，即判别器识别不出真假样本。

GAN 的目标函数为：

$$\min_G \max_D V(D,G)=E_{x\sim p_{data}(x)}[\log D(x)]+E_{x\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

其中，x 为真实数据，$data$ 为真实数据的集合，p_{data} 为真实数据的分布，z 为噪音（输入数据），p_z 为原始噪音的分布。

从判别器 D 的角度考虑，希望尽可能区分真实样本和虚假样本，因此需要判别真实样本为真的概率 $D(x)$ 尽可能大，判别虚假样本为真的概率 $D(G(z))$ 尽可能小，即 $V(D,G)$ 尽可能大；从生成器 G 的角度考虑，希望生成的样本尽可能骗过判别器 D，因此 $D(G(z))$ 尽可能大，即 $V(D,G)$ 尽可能小。两个模型相互对抗，最终达到全局最优。

CGAN 在 GAN 的基础上增加了约束条件，即在生成器

G 和判别器 D 中均引入条件变量 y，y 可以是任意信息，如类别信息或者其他类型的数据。CGAN 通过使用额外信息 y 对模型增加条件，指导数据的生成过程。

CGAN 的目标函数为：

$$\min_G \max_D V(D,G)=E_{x\sim p_{data}(x)}[\log D(x|y)]+E_{z\sim p_z(z)}[\log(1-D(G(z|y)))] \quad (2)$$

本文改进方法中定义条件变量 y 是类别标签，将无监督的 GAN 变成有监督的条件 GAN。

2.2 mRMR 算法

最大相关最小冗余(mRMR)算法的主要思想是通过最大化特征与类别之间的相关性并最小化特征与特征之间的冗余度来选择最优特征子集[28]。特征与类别的相关性是通过特征与类别之间的所有互信息之和的平均值来计算的；特征与特征的冗余度是通过特征与特征之间的互信息之和再除以子集中特征个数的平方来计算的。

X,Y 之间的互信息 $I(X;Y)$ 定义为：

$$I(X;Y)=\iint p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)}\mathrm{d}X\mathrm{d}Y \quad (3)$$

其中，$p(X)$ 为变量 X 的概率，$p(Y)$ 为变量 Y 的概率，$p(X,Y)$ 为 X,Y 的联合概率。互信息表示两个变量之间的关系以及关系的强弱，可以理解为已知 Y 的值而造成对 X 值不确定性的减小，即 Y 中包含关于 X 的信息量。

特征与类别之间的相关性计算公式为：

$$D(S,c)=\frac{1}{|S|}\sum_{x_i\in s}I(x_i;c) \quad (4)$$

特征与特征之间的冗余度计算公式为：

$$R(S)=\frac{1}{|S|^2}\sum_{x_i,x_j\in s}I(x_i;x_j) \quad (5)$$

其中，x_i 为第 i 个特征，x_j 为第 j 个特征，c 为类别变量，S 为特征集合，$I(x_i,c)$ 为第 i 个特征和类别 c 之间的互信息，$I(x_i,x_j)$ 为第 i 个特征和第 j 个特征之间的互信息。

mRMR 结合了上述式(4)和式(5)两种约束条件：

$$mRMR=\max_S\left[\frac{1}{|S|}\sum_{x_i\in s}I(x_i;c)-\frac{1}{|S|^2}\sum_{x_i,x_j\in s}(x_i;x_j)\right] \quad (6)$$

2.3 FL_XGBoost 算法

2.3.1 Focal Loss

Focal Loss 是 Lin 等[29]于 2017 年在目标检测算法中专门为解决不平衡数据分类问题而提出的损失函数。它从两个方面来解决数据类别不平衡问题：1)损失函数更加倾向于关注少数类样本；2)避免易分类样本主导模型训练过程而导致模型性能降低。

Focal Loss 的核心思想为：假设给定不平衡数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$，$i=1,2,\cdots,n$，第 i 个样本的真实类别 $y_i\in\{0,1\}$，定义类别为 1 的样本数量远小于类别为 0 的样本数量。$\hat{y}_i$ 代表预测第 i 个样本类别为 1 的概率，$\hat{y}_i=p(y_i=1|x_i)$。预测第 i 个样本类别为 0 的概率为 $1-\hat{y}_i=p(y_i=0|x_i)$，则第 i 个样本被正确分类的概率为 $p(y_i|x_i)=\hat{y}_i^{y_i}(1-\hat{y}_i)^{1-y_i}$。单个交叉熵损失函数如下：

$$L_{CE}=\begin{cases}-\log\hat{y}_i, & y_i=1\\ -\log(1-\hat{y}_i), & y_i=0\end{cases} \quad (7)$$

1)引入系数 α 来调整不同类别的样本在损失函数中的权重大小，取 $\alpha\in(0.5,1)$ 来增加少数类样本的损失权重。

$$L_{CE\_\alpha}=\begin{cases}-\alpha\log\hat{y}_i, & y_i=1\\ -(1-\alpha)\log(1-\hat{y}_i), & y_i=0\end{cases} \quad (8)$$

2)引入聚焦系数 γ 来调整易分类样本和难分类样本的损失权重，$\gamma>0$。当 $y_i=1$ 时，$\hat{y}_i$ 越接近 1，代表该样本越容易被分类，则 $(1-\hat{y}_i)^\gamma$ 越小，达到了减少易分类样本损失权重的效果，使得算法更加关注于难分类样本。

$$L_{FL\_\alpha}=\begin{cases}-\alpha(1-\hat{y}_i)^\gamma\log\hat{y}_i, & y_i=1\\ -(1-\alpha)(\hat{y}_i^\gamma)\log(1-\hat{y}_i), & y_i=0\end{cases} \quad (9)$$

2.3.2 XGBoost 算法

极端梯度提升算法 XGBoost[30] 是 2014 年提出的基于 CART 回归树的一种 boosting 集成算法。它的目标是建立 K 棵回归树使得树群对样本的预测值尽可能接近样本的真实值，并且具有一定的泛化能力。其基本思想是对给定的训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$，$i=1,2,\cdots,n$，训练出 K 棵 CART 回归树 $F=\{f_1(x),f_2(x),\cdots,f_k(x),\cdots,f_K(x)\}$，$f_k(x_i)$ 代表样本 x_i 进入第 k 棵 CART 回归树后的预测结果输出，K 棵 CART 回归树集成的预测结果为 $\hat{y}_i=\sum_{k=1}^{K}f_k(x)$，$f_k(x)\in F$。XGBoost 的目标优化函数定义为：

$$Obj=\sum_{i=1}^{N}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \quad (10)$$

其中，$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$，$T$ 为第 k 棵树的叶子结点个数，ω_j 为该树第 j 个叶子结点的权重，γ 和 λ 为调节系数。第一部分 $\sum_{i=1}^{N}l(y_i,\hat{y}_i)$ 是模型的损失函数，第二部分 $\sum_{k=1}^{K}\Omega(f_k)$ 是为了控制模型的复杂度而添加的叶子结点权重和树深度的正则化项。目标函数最小化时，样本预测值越接近真实值，同时控制模型的复杂度，避免模型过拟合。

本文 $\sum_{i=1}^{N}l(y_i,\hat{y}_i)$ 模型的损失函数引用 Focal Loss，形成 FL_XGBoost 算法。

3 针对不平衡数据的 XGBoost 改进方法

针对传统分类器在不平衡数据上对少数类样本识别率较低这一问题，本文提出一种综合数据、特征以及算法 3 个层面针对不平衡数据进行改进的方法。

首先在数据层面，利用 CGAN 学习少数类样本分布信息，训练生成器生成逼真的少数类补充样本，以降低数据的不平衡性。与传统过采样技术不同的是，CGAN 可以有效解决样本重叠以及过拟合等问题，根据不同数据集的样本个数以及不平衡度决定生成少数类样本的个数。

其次在特征层面，利用 XGBoost 学习到的决策树对特征进行组合，生成新的特征。决策树实现特征组合的规则为：每个样本进入决策树后通过结点的划分最终落在某个叶子结点上，特征向量中该叶子结点对应的元素值赋为 1，这棵树的其他叶子结点对应的元素值为 0。所有的叶子结点对应的元素值构成的向量形成这棵树的特征向量，将模型中所有子树的特征向量连接起来形成最终的特征向量。新的特征向量的长

度是 XGBoost 模型中所有叶子结点数之和。组合的特征和原始特征构成新的特征集合。然后通过 mRMR 算法选择出更适合不平衡数据分类的特征子集。特征组合算法的主要步骤如算法 1 所示。

算法 1 特征组合算法

输入：样本集合 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$，S 中的第 i 个样本 $s_i=(x_i,y_i)$，$i=1,2,\cdots,n$，第 i 个样本的真实类别 $y_i\in\{0,1\}$，特征集合 W，分类器 XGBoost

输出：特征组合后新的特征集合 W′

1. for i=1,2,…,n do
2. s_i进入 XGBoost；
3. 得到 XGBoost 针对s_i优化出的组合特征v_i；
4. 将组合特征v_i加入特征集合 W 中；
5. end for

最后在算法层面，针对不平衡数据的特点，引入目标检测中针对数据类别不平衡问题的 Focal Loss 来改进 XGBoost 算法。通过实验研究分析发现，在本文实验中，α 取 0.75 且 γ 取 2 时效果较好。改进后的 XGBoost 通过特征选择后的数据集训练得到最终模型。

本文算法的训练过程如算法 2 所示。

算法 2 基于 Focal Loss 改进的 XGBoost 算法

输入：数据集 $D_1=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$，i=1，2，…，n，第 i 个样本的真实类别 $y_i\in\{0,1\}$

输出：分类模型

1. 利用 CGAN 补充少数类样本生成新的数据集D_2；
2. 调用算法 1 生成新的特征集合，构成数据集D_3；
3. 使用 mRMR 算法对数据集D_3进行特征选择，筛选出最优特征子集，构成数据集D_4；
4. 使用 Focal Loss 改进 XGBoost，形成 FL_XGBoost 算法；
5. FL_XGBoost 算法对数据集D_4进行训练得到最终模型。

4 实验与结果分析

4.1 实验数据介绍

为验证本文提出的改进方法针对不平衡数据分类问题具有较高的适用性，本文从 KEEL 机器学习数据库中选取 6 组二分类不平衡数据集，采用 10 次十折交叉验证后的平均值作为实验结果。数据集的不平衡度（多数类与少数类样本的比例）从 1.82 到 41.40 不等。具体信息如表 1 所列。

表 1 实验数据集信息

Table1 Information of experimental datasets

数据集	样本个数	特征维数	正样本数	负样本数	不平衡度
glass1	214	9	76	138	1.82
pima	768	8	268	500	1.87
ecoli2	336	7	52	284	5.46
glass6	214	9	29	185	6.38
yeast3	1484	8	163	1321	8.10
yeast6	1484	8	35	1449	41.40

4.2 性能评价指标

在传统的分类问题中，常采用精度 Accuracy 和错误率 Error Rate 作为评估指标，但这两者对于不平衡数据的分类问题并不适用[31-32]。当数据分布不平衡时，少数类样本对总体准确率的影响较小。即使分类器将全部样本都分类为多数类，仍然可以得到较高的准确率，因此对于不平衡数据分类问题需要选择专门的评价指标。本文选取比较典型的 G-mean[33-34] 和 AUC[35] 作为实验的评估指标。目前，针对不平衡数据分类问题的评估指标是在混淆矩阵的基础上得出的。混淆矩阵如表 2 所列。

表 2 混淆矩阵

Table 2 Confusion matrix

	实际正类	实际负类
预测正类	TP (True Positives)	FP (False Negatives)
预测负类	FN (False Positives)	TN (True Negatives)

根据表 2 可以得出如下指标的定义。

G-mean：正类中被正确分类的概率和负类中被正确分类的概率的综合指标。

$$G\text{-}mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}$$

分析可知，G-mean 结合了两类样本的分类准确率，相对于整体的准确率而言，G-mean 能够更好地衡量分类方法在不平衡数据集上的分类效果。

AUC：AUC 值就是 ROC 曲线下方所覆盖的面积，ROC 曲线是以假正率$\frac{FP}{TN+FP}$为横轴，以真正率$\frac{TP}{TP+FN}$为纵轴所绘制的曲线。ROC 曲线能够比较全面地反映分类器的性能，*AUC* 值越大，分类效果越好。

4.3 实验设计与结果

为验证本文所提方法的有效性，进行两组对比实验。第一组：分层次改进的方法与之前方法的实验对比，验证每阶段的改进是可行的。第二组：改进完全的方法与其他改进模型对比，验证本文提出的方法拥有较好的分类效果。

4.3.1 本文改进方法阶段性实验对比

实验为原始的 XGBoost、一阶段数据层面改进的方法、二阶段特征层面改进的方法以及改进完全的方法之间的相互对比。各方法在 6 个不同数据集上的 *G-mean* 和 *AUC* 对比结果如表 3 和表 4 所列，其中加粗项表示性能最好的分类效果。

表 3 阶段性 *G-mean* 的对比结果

Table 3 Phased *G-mean* comparison results

算法	数据集					
	glass1	pima	ecoli2	glass6	yeast3	yeast6
XGBoost	0.693	0.644	**0.917**	0.830	0.872	0.648
一阶段改进	0.665	0.731	0.880	0.832	0.895	0.869
二阶段改进	0.750	0.753	0.870	0.841	0.902	0.887
改进完全	**0.769**	**0.771**	0.908	**0.959**	**0.905**	**0.894**

表 4 阶段性 *AUC* 的对比结果

Table 4 Phased *AUC* comparison results

算法	数据集					
	glass1	pima	ecoli2	glass6	yeast3	yeast6
XGBoost	0.787	0.759	**0.989**	0.913	0.913	0.713
一阶段改进	0.787	0.770	0.954	0.941	0.962	0.933
二阶段改进	0.900	0.781	0.936	0.950	0.970	0.955
改进完全	**0.922**	**0.835**	0.988	**0.969**	**0.975**	**0.963**

从表 3 和表 4 可以看出，在数据集 glass1 中，一阶段改进的 *G-mean* 值略低于 XGBoost 算法，但改进完全的方法的 *G-mean* 值优于 XGBoost 算法；在数据集 ecoli2 中，本文提出

的改进方法的 *G-mean* 值和 *AUC* 值均低于 XGBoost 算法，原因可能在于数据的分布比较明显，传统分类器本身就能达到较好的分类效果，但补充少数类样本后，反而破坏了这种数据分布，导致分类效果降低。但本文提出的阶段性改进方法在所选的数据集 pima，glass6，yeast3 以及 yeast6 中，其 *G-mean* 值和 *AUC* 值均呈梯度提升，验证了改进方法的可行性。

4.3.2 本文改进方法与其他分类模型的实验对比

为了证明本文提出方法的有效性，选取了两个在不平衡数据分类领域较为经典的改进算法 SMOTEBoost 和 RUSBoost，以及近 3 年内较为流行的改进算法 CUSBoost 和 KSMOTE-AdaBoost，与本文提出的改进方法进行对比。各方法在 6 个不同数据集上的 *G-mean* 和 *AUC* 对比结果如表 5 和表 6 所列。

表 5 各算法的 *G-mean* 对比结果

Table 5 Comparison results of *G-mean* of each algorithm

算法	数据集					
	glass1	pima	ecoli2	glass6	yeast3	yeast6
SMOTE-Boost	0.623	0.351	0.876	0.917	0.902	0.761
CUS-Boost	0.778	0.611	0.880	0.889	0.874	0.747
RUS-Boost	0.253	0.271	0.573	0.754	0.685	0.615
KSMOTE-AdaBoost	**0.790**	0.707	0.882	0.890	0.859	0.704
本文算法	0.769	**0.771**	**0.908**	**0.959**	**0.905**	**0.894**

表 6 各算法的 *AUC* 对比结果

Table 6 Comparison results of *AUC* of each algorithm

算法	数据集					
	glass1	pima	ecoli2	glass6	yeast3	yeast6
SMOTE-Boost	0.779	0.655	0.931	0.930	0.958	0.892
CUS-Boost	0.857	0.672	0.921	0.895	0.963	0.848
RUS-Boost	0.624	0.572	0.840	0.928	0.927	0.860
KSMOTE-AdaBoost	0.889	0.759	0.941	0.954	0.959	0.932
本文算法	**0.922**	**0.835**	**0.989**	**0.969**	**0.975**	**0.963**

从表 5 和表 6 可以看出，本文提出的改进方法与其他改进的分类模型相比识别率更高，分类性能更优。在 5 组数据集 pima，ecoli2，glass6，yeast3 以及 yeast6 中，本文方法的 *G-mean* 值都高于其他改进模型，*G-mean* 值最高达到 95.9%；但在数据集 glass1 中，其 *G-mean* 值略低于对比模型。在 6 组数据集上，本文方法的 *AUC* 值均明显高于其他改进模型，最高达到 97.5%，说明本文提出的改进方法在整体上分类性能较好。

结束语 针对传统分类算法不能很好地解决不平衡类别问题，本文提出了从数据、特征以及算法 3 个层面解决二分类不平衡数据问题的综合改进方法。该方法首先在数据层面借助 CGAN 生成少数类补充样本，以降低数据的不平衡性；其次在特征层面利用 XGBoost 建立决策树对特征进行组合，再通过 mRMR 算法挑选出更适合不平衡数据分类的特征子集；最后在算法层面，引用专门针对不平衡数据分类问题的 Focal Loss 来改进 XGBoost，改进后的 XGBoost 利用新的数据集训练得到最终模型。两组对比实验的结果证明，本文的改进方法是可行的，且相比其他改进的不平衡分类模型分类效果更优，性能更好。但是该方法仅针对二分类数据，存在局限性。在接下来的工作中，我们将针对多分类问题展开研究。

参 考 文 献

[1] LIN W, TSAI C, HU Y, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409: 17-26.

[2] BHATTACHARYA S, RAJAN V, SHRIVASTAVA H. ICU mortality prediction: A classification algorithm for imbalanced datasets[C]// Proc of the 31st AAAI Conf on Artificial Intelligence. San Francisco: AAAI, 2017: 1288-1294.

[3] CHEN X, LIU P H, SUN Y Z, et al. Research on Disease Prediction Models Based on Imbalanced Medical Data Sets[J]. Chinese Journal of Computers, 2019, 42(3): 596-609.

[4] HU M M, CHEN X, SUN Y Z, et al. A Disease Prediction Model Based on Dynamic Sampling and Transfer Learning[J]. Chinese Journal of Computers, 2019, 42(10), 2339-2354.

[5] DUAN L, XIE M, BAI T, et al. A new support vector data description method for machinery fault diagnosis with unbalanced datasets[J]. Expert Systems with Applications. 2016, 64: 239-246.

[6] WANG F, XU T, TANG T, et al. Bilevel feature extraction-based text mining for fault diagnosis of railway systems[J]. IEEE Trans on Intelligent Transportation Systems, 2016, 18(1): 49-58.

[7] WANG S, YAO X. Using class imbalance learning for software defect prediction[J]. IEEE Trans on Reliability, 2013, 62(2): 434-443.

[8] XIONG W, LI B, HE L, et al. Collaborative web service QoS prediction on unbalanced data distribution[C]// IEEE Int Conf on Web Services. Anchorage: IEEE, 2014: 377-384.

[9] SHEN W, WANG X, WANG Y, et al. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection[C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3982-3991.

[10] POUYANFAR S, CHEN S C. Automatic video event detection for imbalance data using enhanced ensemble deep learning[J]. Int J of Semantic Computing, 2017, 11(1): 85-109.

[11] RAO R B. Data mining for improved cardiac care [J]. ACM SIGKDD Explorations Newsletter, 2006, 8(1): 3-10.

[12] LI Y X, CHAI Y, HU Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.

[13] GARCIA V, SANCHEZ J S, MOLLINEDAR A. On the eff ectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2011, 25(1): 13-21.

[14] 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.

[15] CIESLAK D A, CHAWLA N V, STRIEGELA. Combating imbalance in network intrusion datasets[C]// Proceedings of IEEE International Conference on Granular Computing. IEEE, 2006: 732-737.

[16] GUO H P, ZHOU J, WU C A, et al. K-nearest neighbor classifi-

cation method for class-imbalanced problem[J]. Journal of Computer Applications,2018,38(4):955-959,977.

[17] MEMBER M W,CHEN X W. Combating the Small Sample Class Imbalance Problem Using Feature Selection[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1388-1400.

[18] WANG J,LI D Y,WANG S G. Feature Selection Method for Imbalanced Text Sentiment Classification[J]. Computer Science,2016,43(10):206-210,224.

[19] ZHAO N,ZHANG X F,ZHANG L J. Overview of imbalanced data classification[J]. Computer Science,2018,45(S1):22-27

[20] WU Y X,WANG J L,YANG L,et al. Survey on Cost-sensitive Deep Learning Methods[J]. Computer Science, 2019,46(5): 1-12.

[21] CAO Y X,HUANG H Y. Imbalanced Data Classification Algorithm Based on Probability Sampling and Ensemble Learning [J]. Computer Science,2019,46(5):203-208.

[22] YUAN X M,YANG M,YANG Y. An Ensemble Classifier Based on Structural Support Vector Machine for Imbalanced Data.[J]. Pattern Recognition and Artificial Intelligence, 2013, 26(3):315-320.

[23] CHAWLA N V,LAZAREVIC A,HALL L O,et al. SMOTEBoost:improving prediction of the minority class in boosting [C]// Proceedings of the 2003 European Conference on Knowledge Discovery in Databases, LNCS 2838. Berlin: Springer, 2003:107-119.

[24] RAYHAN F,AHMED S,MAHBUB A,et al. CUSBoost:Cluster-based Under-sampling with Boosting for Imbalanced Classification[C]// 11th International Conference on Software Knowledge Information Management and Applications (SKIMA). 2017:1-6.

[25] SEIFFERT C,KHOSHGOFTAAR T M,VAN H J,et al. RUSBoost:a hybrid approach to alleviating class imbalance[J]. IEEE Transactions on Systems,Man and Cybernetics,Part A:Systems and Humans,2010,40(1):185-197.

[26] WANG Z Z,HUANG B,FAN Z J. et,al. Improved SMOTE unbalanced data integration classification algorithm[J]. Computer Application,2019,39(9):2591-2596.

[27] MIRZA M,OSINDER S. Conditional Generative Adversarial Nets[J]. arXiv:1411.1784,2014.

[28] PENG H,LONG F,DING C. Feature selection based on mutual information:Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2005,27(8):1226-1238.

[29] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern on Pattern Analysis & Machine Intelligence,2017,PP(99):2999-3007.

[30] CHEN T,GUESTRIN C. XGBoost:A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.

[31] CAI Z X ,WANG X Y,X J,et al. Sample Adaptive Classifier for Imbalanced Data[J]. Computer Science,2019,46(1):94-99.

[32] XIONG B Y,WANG G Y,DENG W B. Under-Sampling Method Based on Sample Weight for Imbalanced Data[J]. Journal of Computer Research and Development,2016,53(11):2613-2622.

[33] MATHEW J,PANG C K,LUO M,et al. Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines[J]. IEEE Transactions on Neural Networks and Learning Systems,2018,29(9):4065-4076.

[34] LI X F,L J,D Y F,et al. A New Learning Algorithm for Imbalanced Data-PCBoost[J]. Chinese Journal of Computers, 2012, 35(2):202-209.

[35] ZHANG N,CHEN Q. Ensemble learning training method based on AUC and Q statistics[J]. Journal of Computer Applications, 2019,39(4):935-939.

SONG Ling-ling, born in 1994, postgraduate. Her main research interests include machine learning and so on.

YANG Chao, born in 1982, Ph.D, associate professor, postgraduate supervisor, is a member of China Computer Federation. His main research interests include information security and computer immunology.