Journal of Computer Applications
计算机应用,2020,40(6):1633 - 1637
ISSN 1001-9081
CODEN JYIIDU
2020-06-10
http://www.joca.cn

文章编号:1001-9081(2020)06-1633-05
DOI:10.11772/j.issn.1001-9081.2019101878

基于梯度分布调节策略的Xgboost算法优化

李　浩,朱　焱*
(西南交通大学 信息科学与技术学院,成都 611756)
(*通信作者电子邮箱yzhu@swjtu.edu.cn)

摘　要:为了解决集成学习模型Xgboost在二分类问题中少数类检出率低的问题,提出了基于梯度分布调节策略的改进的Xgboost算法——LCGHA-Xgboost。首先,通过定义损失贡献(LC)来模拟Xgboost算法中样本个体的损失量;而后,通过定义损失贡献密度(LCD)来衡量Xgboost算法中样本被正确分类的难易程度;最后,提出了梯度分布调节算法LCGHA,依据LCD动态调整样本个体的一阶梯度分布,间接地增大难分样本(主要存在于少数类中)的损失量,减小易分样本(主要存在于多数类中)的损失量,使Xgboost算法偏向对难分样本的学习。实验结果表明,与Xgboost、GBDT、随机森林(Random_Forest)这三大集成学习算法相比,LCGHA-Xgboost算法在多个UCI数据集上的召回率(Recall)值有5.4%~16.7%的提高,AUC值有0.94%~7.41%的提高;在垃圾网页数据集WebSpam-UK2007和DC2010数据集上所提算法的Recall值更是有44.4%~383.3%的提高,AUC值有5.8%~35.6%的提高。LCGHA-Xgboost算法可以有效提高对少数类的分类检出能力,减小少数类的分类错误率。

关键词:不平衡分类;Xgboost;梯度分布;损失贡献;损失贡献密度
中图分类号:TP181　　**文献标志码:**A

Xgboost algorithm optimization based on gradient distribution harmonized strategy

LI Hao, ZHU Yan*
(*School of Information Science and Technology*, *Southwest Jiaotong University*, *Chengdu Sichuan* 611756, *China*)

Abstract: In order to solve the problem of low detection rate of minority class by ensemble learning model eXtreme gradient boosting (Xgboost) in the binary classification problem, an improved Xgboost algorithm based on gradient distribution harmonized strategy called Loss Contribution Gradient Harmonized Algorithm (LCGHA)-Xgboost was proposed. Firstly, Loss Contribution (LC) was defined to simulate the losses of the samples in Xgboost algorithm. Secondly, by defining Loss Contribution Density (LCD), the difficulty of samples being correctly classified in Xgboost algorithm was measured. Finally, a gradient distribution harmonized algorithm called LCGHA was proposed to dynamically adjust the one order gradient distribution of samples according to the LCD. In the algorithm, the losses of hard samples (mainly in minority class) were indirectly increased, and the losses of easy samples (mainly in majority class) were indirectly reduced, making Xgboost algorithm tend to learn the hard samples. The experimental results show that compared with three ensemble learning algorithms Xgboost, GBDT (Gradient Boosting Decision Tree) and Random_Forest, LCGHA-Xgboost has the recall increased by 5.4%-16.7%, and Area Under the Curve (AUC) improved by 0.94%-7.41% on multiple UCI datasets, and the Recall increased by 44.4%-383.3%, and AUC improved by 5.8%-35.6% on WebSpam-UK2007 and DC2010 datasets. LCGHA-Xgboost can effectively improve the classification and detection ability for minority class, and reduce the classification error rate of minority class.

Key words: imbalanced classification; Xgboost; gradient distribution; loss contribution; loss contribution density

0　引言

在机器学习、模式识别等领域中,不平衡二分类问题一直是非常重要的研究课题。在现实生活中,不平衡数据存在于方方面面,如病毒检测、垃圾网页检测等。其特点是多数类的数据量一般远远多于少数类的数据量,数据呈现不平衡的分布状态。但是在许多不平衡分类问题中,少数类通常是更重要的。例如在垃圾检测领域中,垃圾网页的数据量远远少于正常网页的数据量。但相对正常网页,垃圾网页的漏检、误检往往会给社会带来更多的危害,带来更大的损失。因此,如何提高对少数类的分类准确率,减小少数类分类错误带来的损失,是数据挖掘中迫切需要解决的问题。Xgboost(eXtreme gradient boosting)算法是近年来兴起的一种高效集成学习算法,但是它在不平衡二分类问题上的表现却不令人满意。这是因为当多数类样本量远远多于少数类时,多数类的损失量占比将会远大于少数类,这导致Xgboost在建模时会偏重对多数类的学习,忽视少数类,从而降低Xgboost分类性能。

为了解决Xgboost在二分类问题中少数类检出率低的问题,许多改进的算法被提出。Luo等[1]提出以Xgboost作为基

收稿日期:2019-11-04;**修回日期:**2019-12-16;**录用日期:**2019-12-18。　**基金项目:**四川省科技计划项目(2019YFSY0032)。
作者简介:李浩(1992—),男,安徽定远人,硕士研究生,主要研究方向:不平衡数据分类;　朱焱(1965—),女,广西桂林人,教授,博士,主要研究方向:数据挖掘、Web异常模式发现、大数据管理与智能分析。

分类器，结合欠采样和Tomek-Link技术构造集成学习分类器。但是该方法是局限在样本集上的改进，对Xgboost算法本身并未改进；而且欠采样和Tomek-Link技术会使得数据集有信息损失，从而可能导致最终的分类结果不理想。Shi等[2]基于带权重的Xgboost构建了二层分类检测模型；但是该模型中针对Xgboost的权重计算方式过于简单，仅考虑了同类别样本数量的占比，且权重大小恒定不变，没有很好地平衡不同类别样本的损失量。Chen等[3]提出了基于IEEM（Interval Error Evaluation Method）对于不同的样本使用分段划等级的方式给予不同的重要性，从而提高了算法对少数类的关注；但是人为指定样本等级过于主观，容易造成较大的误差。Li等[5]针对少数类样本提出了基于梯度密度的调节策略，在神经网络使用梯度下降法学习数据的过程中，调节样本的梯度分布，增大少数类样本的梯度，减小多数类样本的梯度，从而提高神经网络对少数类样本的检出能力。正是受文献[5]的启发，本文提出了LCGHA（Loss Contribution Gradient Harmonized Algorithm）-Xgboost算法，LCGHA-Xgboost算法流程如图1所示，该算法不仅提高了Xgboost的性能，而且实现更加自动化和智能化。

LCGHA-Xgboost算法关注数据集中的难分样本，这些样本绝大部分存在于因为数量过少而导致难以被正确分类的少数类中，还有部分存在于多数类中；因此本文方法对少数类和多数类的学习能力都有较大提升，对算法性能的提升效果更加出色。LCGHA梯度调节算法定义损失贡献密度来衡量样本被正确分类的难易程度。依据损失贡献密度，动态调整样本一阶梯度分布，间接提高难分样本损失量在总损失量中的占比，使Xgboost偏重对难分样本的学习，从而提高Xgboost的分类能力。

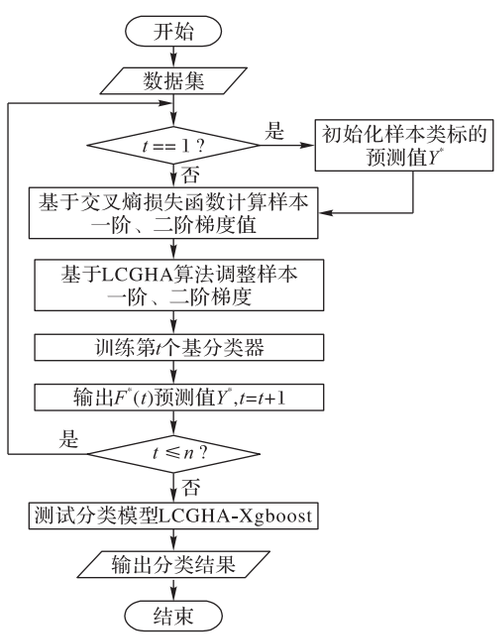

图1 LCGHA-Xgboost算法流程

Fig. 1 Flowchart of LCGHA-Xgboost algorithm

本文算法LCGHA-Xgboost与文献[2]算法和文献[3]算法的不同之处在于本文算法关注数据集中难分样本，算法作用范围更加广泛；并且本文算法对不同样本的关注度随模型训练的过程而变化，可以更好地拟合数据集样本，提高Xgboost分类能力。本文与文献[5]同样研究了梯度不平衡调节策略，但本文的损失贡献密度可以更好地衡量样本被正确分类的难易程度，更准确地识别难分样本和易分样本，从而进行更加合理的梯度调整。

通过在UCI数据集、WebSpam-UK2007和DC2010等多个数据集上进行实验验证得出，LCGHA梯度调节算法有效地提高了Xgboost对少数类样本的检出能力，增强了Xgboost算法性能。

1 相关技术

1.1 Xgboost算法原理及模型介绍

Xgboost是由Chen等[9]提出的一种集成学习算法。它在GBDT（Gradient Boosting Decision Tree）的基础上，通过改进模型拟合目标，在目标函数中添加正则项，进一步优化了性能。因为GBDT算法在优化损失函数时只使用了一阶导数信息，而Xgboost算法则是对损失函数进行了二阶泰勒展开，利用了一阶导数和二阶导数信息，从而能更好地拟合损失函数，减小优化过程中存在的误差。具体定义如下所示。

已知一个训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，$x_i\in X\subseteq \mathbf{R}^m$，$y_i\in Y\subseteq \mathbf{R}$，$X$为输入空间，$Y$为输出空间。Xgboost可以表示成加法模型，即：

$$\tilde{y}_i=\sum_{k=1}^{K}f_k(x_i) \tag{1}$$

式中：$\tilde{y}_i$表示训练模型的预测值；f_k表示第k个子模型；x_i表示第i个输入样本。Xgboost算法的优化目标包括损失函数和正则项两个部分，因此最终的优化目标为：

$$L(t)=\sum_{i=1}^{n}l(y_i,\tilde{y}_i(t-1)+f_t(x_i))+H(f_t) \tag{2}$$

式中：$L(t)$表示第t次迭代时的目标函数；y_i表示原始样本类标；$\tilde{y}_i(t-1)$表示样本在$t-1$次模型迭代时，模型的预测值；$f_t(x_i)$表示样本在第t次模型迭代时，模型的预测值；$H(f_t)$是目标函数的正则项。

对式(2)泰勒展开，可得：

$$\tilde{L}(t)\cong\sum_{j=1}^{r}\left[\omega_j\sum_{i\in I_j}g_i+\frac{1}{2}\omega_j^2\left(\sum_{i\in I_j}h_i+\lambda\right)\right]+\gamma T \tag{3}$$

式中：g_i表示样本x_i的一阶梯度；h_i表示样本x_i的二阶梯度；ω_j表示第j个节点的输出值；λ和γ是正则项系数，防止模型过拟合；I_j是第j个叶节点中的样本子集。

Xgboost模型的训练过程就是求解式(3)，找到最佳的ω_j^*及其对应的目标函数最优解，即：

$$\omega_j^*=-\sum_{i\in I_j}g_i\Big/\left(\sum_{i\in I_j}h_i+\lambda\right) \tag{4}$$

$$\tilde{L}(t)=-\frac{1}{2}\sum_{j=1}^{r}\left(\left(\sum_{i\in I_j}g_i\right)^2\Big/\left(\sum_{i\in I_j}h_i+\lambda\right)\right)+\gamma T \tag{5}$$

式(5)用来衡量一棵树的结构好坏，其值越低，代表树的结构越好。因此当树分裂节点时，可以得到式(6)：

$$Gain=L_{\text{left}}+L_{\text{right}}-L_{\text{father}} \tag{6}$$

如果$Gain$值大于零，则继续分裂节点；否则停止分裂节点。

1.2 梯度分布定义

在使用梯度下降法或牛顿法建模时,需要计算样本一阶梯度或二阶梯度。但是在同一个损失函数下,不同样本的梯度分布往往是不同的,因此有了梯度分布的概念。

定义 1 梯度分布。已知样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, $x_i\in X\subseteq \mathbf{R}^m$, $y_i\in Y\subseteq \mathbf{R}$, X 为输入空间,Y 为输出空间。样本集 D 的一阶梯度集合 $G=\{g_1,g_2,\cdots,g_n\}$,g_i 为样本 x_i 的梯度。

$$\chi(g_i,g_j)=\begin{cases}1, & g_i-\varepsilon\leqslant g_j\leqslant g_i+\varepsilon\\ 0, & \text{其他}\end{cases} \tag{7}$$

$$\delta(x_i)=\sum_{j=1}^{N}\chi(g_i,g_j) \tag{8}$$

其中:g_i 代表样本 x_i 的梯度;$g_j\in G$;$\chi(g_i,g_j)$ 表示样本 x 的梯度值是否落在以 g_i 为中心、ε 为半径的区域内;$\delta(x_i)$ 表示样本集 D 中落在以 g_i 为中心、ε 为半径的样本数。则 $\Phi=\{\delta(x_1),\delta(x_2),\cdots,\delta(x_n)\}$ 表示样本 D 内每个样本的梯度分布。由定义 1 可以得出样本集总体的梯度分布情况。

2 基于梯度分布调节策略的改进Xgboost算法

Xgboost在许多分类问题上表现都十分出色,但是在垃圾网页、故障检测等数据不平衡的领域中,对难分样本的检出能力较差,分类性能较低。为了提高Xgboost对难分样本的检出能力,本文定义了损失贡献和损失贡献密度,提出LCGHA-Xgboost算法。以损失贡献模拟样本损失量,以损失贡献密度衡量样本被正确分类的难易程度。依据损失贡献密度调节样本一阶梯度 g_i 的分布情况,达到间接提高难分样本损失量的目的,从而增强算法对难分样本的关注和学习,提高Xgboost的分类性能。

样本损失贡献(Loss Contribution, LC)的具体定义如下:

$$LC(x_i)=g_i^2/h_i \tag{9}$$

即损失贡献表现为以样本 x_i 的二阶梯度的平方除以一阶度。损失贡献分布定义如下。

定义 2 损失贡献分布。已知样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, $x_i\in X\subseteq \mathbf{R}^m$, $y_i\in Y\subseteq \mathbf{R}$, X 为输入空间,Y 为输出空间。样本集 D 的损失贡献集合 $LC\_Set=\{LC_1,LC_2,\cdots,LC_n\}$,$LC_i$ 表示样本 x_i 的损失贡献。

$$\chi(LC_i,LC_j)=\begin{cases}1, & LC_i-\varepsilon\leqslant LC_j\leqslant LC_i+\varepsilon\\ 0, & \text{其他}\end{cases} \tag{10}$$

$$\delta(x_i)=\sum_{j=1}^{N}\chi(LC_i,LC_j) \tag{11}$$

其中:$\chi(LC_i,LC_j)$ 表示样本 x_i 的损失贡献是否落在以 LC_i 为中心、ε 为半径的区域内;$\delta(x_i)$ 表示样本集 D 中落在以 LC_i 为中心、ε 为半径的样本数。则 $\Omega=\{\delta(x_1),\delta(x_2),\cdots,\delta(x_n)\}$ 表示样本集 D 内样本的损失贡献分布。相较定义 1,定义 2 是为了统计样本集总体的损失贡献分布情况。

损失贡献密度(Loss Contribution Density, LCD)定义如下。

定义 3 损失贡献密度。已知样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, $x_i\in X\subseteq \mathbf{R}^m$, $y_i\in Y\subseteq \mathbf{R}$, X 为输入空间,Y 为输出空间。N 代表样本 D 的样本数量;$LCD(x_i)$ 表示样本 x_i 的损失贡献密度。样本集 D 的损失贡献分布 $\Omega=\{\delta(x_1),\delta(x_2),\cdots,\delta(x_n)\}$,$\delta(x_i)$ 表示样本 x_i 损失贡献分布。

$$LCD(x_i)=\delta(x_i)/N \tag{12}$$

样本 x_i 损失贡献密度是样本的 $\delta(x_i)$ 值除以总样本数的商,依据定义 3 可以得到个体样本损失贡献分布区域的稀疏情况。样本 x_i 的损失贡献密度越大,代表在样本 x_i 附近与其损失贡献相近的样本就越多。这部分样本损失量占比就会越高,就会越发受到算法的关注和学习,因此越容易被算法正确检出。反之,样本 x_i 的损失贡献密度越小,就越难被算法正确检出。

使用损失贡献密度可以很好地衡量样本被正确分类的难易程度。想要提高对难分样本的检出能力,只需要通过调整样本的一阶梯度分布的方式,来间接地调整样本在式(5)中的损失量,具体定义如下。

$$g_i^*=g_i/LCD(x_i) \tag{13}$$

难分样本由于其LCD值远小于易分样本的LCD值,经过式(13)调整后,其一阶梯度的增幅会大于易分样本,从而间接导致式(5)中难分样本损失量增幅会大于易分样本的增幅,最终使得Xgboost偏向关注对难分样本的学习和分类。具体算法过程如下所示。

算法 1　LCGHA。

输入　数据集 D 类标 Y,数据集 D 预测值 Y^*;

输出　梯度分布调节后的样本一阶梯度集合 G^* 和原二阶梯度集合 H。

```
Begin
1)  根据输入的数据集D类标Y和Y*,计算样本的一阶梯度集合G和二阶梯度集合H
2)  根据式(9)计算样本的损失贡献集合LC_Set
3)  初始化样本损失贡献分布集合Ω={}
4)  for D中的每个样本x_i:
5)    初始化变量δ(x_i)=0
6)    for D中的每个样本x_j:
7)      根据式(10)和(11)计算样本x_i梯度分布
8)      若LC_i-ε≤LC_j≤LC_i+ε:
9)      则δ(x_i)=δ(x_i)+1
10)   end for
11)   Ω.append(δ(x_i))
12) end for
13) 根据式(12)计算每个样本的损失贡献密度LCD
14) 根据式(13)调节所有样本原一阶梯度集合G,得到G*
15) 输出数据集D的一节梯度集合G*和原二阶梯度H
End
```

算法 1 会在每个基分类器训练完成后,动态调整样本的一阶梯度分布。训练下一棵树时,算法就会着重关注难分样本。LCGHA-Xgboost算法的具体过程如下:

算法 2　LCGHA-Xgboost。

输入　训练集 D,测试集 T;

输出　数据集预测类标 Y^0。

```
Begin:
1)  输入训练集D,并开始训练模型
2)  for t←1 to n:
3)    若t==1:
4)      则初始化样本类标预测值Y*
5)    调用算法1,并输入Y和Y*,调整样本一阶梯度分布
6)    根据步骤5)返回的G*、H,训练第t棵回归树
7)    第t棵回归树与前t-1棵回归树组成强学习器F*(t)
```

8） 使用$F^*(t)$预测数据集D中样本，并得到数据集类标预测值Y^*和原数据集类标Y

9） end for

10） 将训练完成的$F^*(t)$作为最终的分类器$F(t)$

11） 输入测试集T，使用$F(t)$预测，得到预测类标Y^0

End

3 LCGHA-Xgboost性能验证与分析

3.1 实验数据集

为了检验算法性能优劣，本文选择了多个UCI数据集，它们有不同的不平衡比例。这些数据集来自于UCI数据库，是加州大学欧文分校（University of California，Irvine）提出的用于机器学习的数据库，而UCI数据集是一个常用的标准测试数据集（网址为：https://archive. ics. uci. edu/ml/index. php）。数据集数据分布如表1所示。

表1 UCI数据集数据分布

Tab. 1 Data distribution of UCI datasets

序号	数据集	样本数	特征数	不平衡比率
01	Ecoli	336	8	6. 636
02	Glass	214	10	3. 196
03	Yeast	1 484	8	8. 104
04	Climate	540	20	10. 730
05	Leaf	340	15	5. 538
06	Ionosphere	351	34	1. 786

此外本文还选择了样本更多、不平衡率更大的WebSpam-UK2007[11]和DC2010[12]作为本次的实验数据集。WebSpam-UK2007是由垃圾网页检测挑战赛和对抗性信息检索Web讨论组（AIRWEB）于2007年收集的公开资料集（下载网址：https://chato. cl/webspam/datasets/uk2007/），主要用于2008年垃圾网页检测挑战赛。DC2010数据集是由匈牙利科学院（the Hungarian Academy of Sciences）于2010年收集的公开数据集（下载网址为：https://dms. sztaki. hu/en/download/web-spam-resources），主要用于2010年ECML/PKDD发现挑战赛。其中WebSpam-UK2007数据集曾经过信息增益率优选特征处理，数据集具体数据分布如表2所示。

表2 垃圾网页数据集数据分布

Tab. 2 Data distribution of Web spam datasets

序号	数据集	样本数量	特征数量	不平衡比率
07	WebSpam-UK2007	5 798	75	17. 059
08	DC2010	1 412	96	25. 148

3.2 评价指标

解决不平衡分类问题，仅使用准确率、精确率等指标是不全面的。本文加入了AUC（Area Under the Curve）来共同评价模型的性能，AUC综合考虑少数类和多数类的分类准确性。混淆矩阵可以很好地表示出各种分类情况。

表3 混淆矩阵

Tab. 3 Confusion matrix

真实类别	预测类别	
	正类	负类
正类	*TP*	*FN*
负类	*FP*	*TN*

AUC是接受者操作特性曲线（Receiver Operating Characteristic curve，ROC）下的面积，取值在0到1之间。精确率（Precision）、召回率（Recall）、F1值和AUC等指标计算式如下：

$$Precision = TP/(TP + FP) \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

$$AUC = \frac{\sum_{i \in PositiveClass} rank_i - NP(NP + 1)/2}{NP \times NN} \tag{17}$$

其中：NP表示正类样本（少数类）总数；NN表示负类样本（多数类）总数；i表示正类样本；$rank_i$表示正类样本的置信度排序。

3.3 实验设计与结果分析

基于上述数据集和评价指标，本文设计分类检测实验，将LCGHA-Xgboost算法与当前较为流行的集成学习算法进行对比，这些算法包括随机森林（Random_Forest）、GBDT和Xgboost算法，实验对比结果如表4所示。

表4 不同数据集上不同算法的分类性能对比

Tab. 4 Classification performance comparison of different algorithms on different datasets

算法	Ecoli			Glass			Yeast			Climate		
	Recall	F1	AUC	Recall	F1	AUC	Recall	F1	AUC	Recall	F1	AUC
Random_Forest	0. 800 0	0. 888 9	0. 900 0	1. 000 0	0. 971 4	0. 990 7	0. 777 8	0. 770 6	0. 874 0	0. 750 0	0. 774 2	0. 865 9
GBDT	0. 866 7	0. 928 6	0. 933 3	0. 941 2	0. 969 7	0. 970 6	0. 777 8	0. 800 0	0. 878 6	0. 750 0	0. 774 2	0. 865 9
Xgboost	0. 933 3	0. 965 5	0. 966 7	0. 941 2	0. 969 7	0. 970 6	0. 759 3	0. 781 0	0. 868 2	0. 750 0	0. 750 0	0. 862 8
LCGHA-Xgboost	**0. 933 3**	0. 965 5	**0. 966 7**	**1. 000 0**	1. 000 0	**1. 000 0**	**0. 833 3**	0. 762 7	**0. 894 9**	**0. 812 5**	0. 896 6	**0. 906 2**
算法	Leaf			Ionosphere			WebSpam-UK2007			DC2010		
	Recall	F1	AUC	Recall	F1	AUC	Recall	F1	AUC	Recall	F1	AUC
Random_Forest	0. 666 7	0. 800 0	0. 784 7	0. 881 0	0. 913 6	0. 927 1	0. 159 3	0. 260 9	0. 577 7	0. 444 4	0. 516 1	0. 716 8
GBDT	0. 666 7	0. 800 0	**0. 833 3**	0. 881 0	0. 913 6	0. 927 1	0. 389 4	0. 377 7	0. 673 9	0. 500 0	0. 428 6	0. 733 6
Xgboost	0. 611 1	0. 666 7	0. 784 7	0. 881 0	0. 925 0	0. 933 8	0. 327 4	0. 418 1	0. 656 4	0. 500 0	0. 620 7	0. 747 8
LCGHA-Xgboost	**0. 666 7**	0. 727 3	0. 817 7	**0. 928 6**	0. 939 8	**0. 951 0**	**0. 769 9**	0. 304 2	**0. 783 6**	**0. 722 2**	0. 273 7	**0. 791 1**

从表4中可以看出，本文提出的LCGHA-Xgboost模型有效地提高了绝大多数数据集样本的分类性能，其中，在Glass数据集上，其AUC值达到了100%。本文算法通过调整样本一阶梯度增大难分样本损失量的策略有效提高了算法对少数类的关注和多数类中部分难分样本的关注，使得Xgboost对少数类和多数类的分类能力都得到了提升，从而做到在Glass数据集上完全分类正确。在Ecoli、Yeast、Climate、Ionosphere和DC2010数据集上，LCGHA-Xgboost分类性能最优，其AUC值

相较对比算法有0.94%~7.41%的提升。这是因为本文算法定义的损失贡献和损失贡献密度可以准确地模拟样本实际损失量和识别出难分样本,从而更加合理地提高难分样本的损失量占比,增强算法对少数类、多数类难分样本的检出能力。另外,在数据集WebSpam-UK2007上的实验结果显示,LCGHA-Xgboost算法AUC值相较传统Xgboost提高了19.3%,相较Random_Forest提高了近35.6%,主要原因是本文算法Recall提高了97.7%~383.3%,少数类样本被有效检出,所以算法分类性能得到较大提高。在Leaf数据集上,LCGHA-Xgboost算法性能略低于GBDT算法,高于RF和Xgboost算法。综合所有数据集上的表现可知,本文提出的LCGHA-Xgboost算法可以有效增强Xgboost对少数类的检出能力,提高算法性能。

4　结语

本文针对不平衡二分类问题中少数类误分错误率高的问题,提出了LCGHA-Xgboost算法。LCGHA-Xgboost算法定义了损失贡献和损失贡献密度,针对Xgboost算法的建模特性进行改进优化,提出了一阶梯度的调节策略LCGHA算法,以达到调整样本损失量的目的。实验结果表明,该算法在多数数据集上具有更高的AUC值,有效提高了Xgboost算法的分类性能。在接下来的工作中,将研究用误分类总代价替换传统的损失函数,结合优秀的进化算法,以达到更好的分类性能。

参考文献 (References)

[1] LUO R, DIAN S, WANG C, et al. Bagging of Xgboost classifiers with random under-sampling and tomek-link for noisy label-imbalanced data [J]. IOP Conference Series Materials Science and Engineering, 2018, 428: Article No. 012004.

[2] SHI H, WANG H, HUANG Y, et al. A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification [J]. Computer Methods and Programs in Biomedicine, 2019, 171:1-10.

[3] CHEN S, LIU X, LI B. A cost-sensitive loss function for machine learning [C]// Proceedings of the 2018 International Conference on Database Systems for Advanced Applications, LNCS 10829. Cham: Springer, 2018: 255-268.

[4] ALA'RAJ M, ABBOD M F. Classifiers consensus system approach for credit scoring [J]. Knowledge-Based Systems, 2016, 104: 89-105.

[5] LI B, LIU Y, WANG X. Gradient harmonized single-stage detector [C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8577-8584.

[6] OGUNLEYE A, WANG Q G. Enhanced XGBoost based automatic diagnosis system for chronic kidney disease [C]// Proceedings of the IEEE 14th International Conference on Control and Automation. Piscataway: IEEE, 2018: 805-810.

[7] BAHNSEN A C, AOUADA D, OTTERSTEN B, et al. Example-dependent cost-sensitive decision trees [J]. Expert Systems with Applications, 2015, 42(19): 6609-6619.

[8] ZHANG B, YU Y, LI J. Network intrusion detection based on stacked sparse autoencoder and binary tree ensemble method [C]// Proceedings of the 2018 IEEE International Conference on Communications Workshops . Piscataway: IEEE, 2018: 1 -6.

[9] 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. New York: ACM, 2016: 785-794.

[10] LIU H, ZHOU M, LU X S. Weighted Gini index feature selection method for imbalanced data [C]// Proceedings of the IEEE 15th International Conference on Networking, Sensing and Control. Piscataway: IEEE, 2018: 1-6.

[11] CASTILLO C, DONATO D, BECCHETTI L, et al. A reference collection for Web spam [J]. ACM SIGIR Forum, 2006, 40(2): 11-24.

[12] SIKLÓSI D, DARÓCZY B, BENCZÚR A A. Content-based trust and bias classification via biclustering [C]// Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. New York: ACM, 2012: 41-47.

[13] WANG J, SUN Z, BAO B, et al. Malicious synchrophasor detection based on highly imbalanced historical operational data [J]. CESS Journal of Power and Energy Systems, 2019, 5(1): 11-20.

[14] KANG Q, CHEN X, LI S, et al. A noise-filtered under-sampling scheme for imbalanced classification [J]. IEEE Transactions on Cybernetics, 2017, 47(12): 4263-4274.

[15] TANG Y, WANG X, JIA X, et al. Fusing multiple deep features for face anti-spoofing [C]// Proceeding of the 2018 Chinese Conference on Biometric Recognition, LNCS 10996. Cham: Springer, 2018: 321-330.

[16] IBRAHIM O A S, LANDA-SILVA D. Term frequency with average term occurrences for textual information retrieval [J]. Soft Computing, 2016, 20(8): 3045-3061.

[17] CHEN K, ZHANG Z, LONG J, et al. Turning from TF-IDF to TF-IGM for term weighting in text classification [J]. Expert Systems with Applications, 2016, 66: 245-260.

This work is partially supported by the Science and Technology Plan of Sichuan Province (2019YFSY0032).

LI Hao, born in 1992, M. S. candidate. His research interests include unbalanced data classification.

ZHU Yan, born in 1965, Ph. D., professor. Her research interests include data mining, Web anomaly detection, big data management and intelligent analysis.