计算机科学
COMPUTER SCIENCE

基于代价敏感激活函数 XGBoost 的不平衡数据分类方法

李京泰，王晓丹

引用本文

李京泰，王晓丹. 基于代价敏感激活函数 XGBoost 的不平衡数据分类方法[J]. 计算机科学, 2022, 49(5): 135-143.

LI Jing-tai, WANG Xiao-dan. XGBoost for Imbalanced Data Based on Cost-sensitive Activation Function[J]. Computer Science, 2022, 49(5): 135-143.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400064

基于代价敏感激活函数 XGBoost 的不平衡数据分类方法

李京泰　王晓丹
空军工程大学防空反导学院　西安 710051
(afeulijingtai@163.com)

摘　要　为解决在数据不平衡条件下使用 XGBoost 框架处理二分类问题时算法对少数类样本的识别能力下降的问题,提出了基于代价敏感激活函数的 XGBoost 算法(Cost-sensitive Activation Function XGBoost,CSAF-XGBoost)。在 XGBoost 框架构建决策树时,数据不平衡会影响分裂点的选择,导致少数类样本被误分。通过引入代价敏感激活函数改变样本在不同预测结果下损失函数的梯度变化,来解决被误分的少数类样本因梯度变化小而无法在 XGBoost 迭代过程中被有效分类的问题。通过实验分析了激活函数的参数与数据不平衡度的关系,并对 CSAF-XGBoost 算法与 SMOTE-XGBoost,ADASYN-XGBoost,Focal loss-XGBoost,Weight-XGBoost 优化算法在 UCI 公共数据集上的分类性能进行了对比。结果表明,在 F1 值和 AUC 值相同或有提高的情况下,CSAF-XGBoost 算法对少数类样本的检出率比最优算法平均提高了 6.75%,最多提高了 15%,证明了 CSAF-XGBoost 算法对少数类样本有更高的识别能力,且具有广泛的适用性。

关键词:代价敏感;Logistic 回归;数据不平衡分类;XGBoost;激活函数
中图法分类号　TP391.4

XGBoost for Imbalanced Data Based on Cost-sensitive Activation Function

LI Jing-tai and WANG Xiao-dan
Air and Missile Defense College,Air Force Engineering University,Xi'an 710051,China

Abstract　For binary classification with category imbalance,a cost-sensitive activation function XGBoost algorithm(CSAF-XGBoost) is proposed to promote the ability of recognizing minority samples. When XGBoost algorithm constructs decision trees,unbalanced data will affect split point selection,which lead to misclassification of minority. By constructing cost-sensitive activation function (CSAF),samples in different estimation are under different gradient variations,which approach the problem that the gradient variation of misclassified minority sample is too small to make samples be recognized correctly in iterations. The experiments analyze the relation of imbalanced rate (IR) to parameters,and compare performance with SMOTE-XGBoost,ADASYN-XGBoost,Focal loss-XGBoost and Weight-XGBoost on UCI datasets. As for recall rate of minority,CSAF-XGBoost surpasses the best methods 6.75% in average and 15%in maximum with F1-score and AUC score in the same level. The results prove CSAF-XGBoost has better performance in recognizing minority class samples and wider applicability.

Keywords　Cost-sensitive,Logistic regression,Data imbalanced classification,XGBoost,Activation function

1　引言

传统分类器在设计时通常假设不同类别的样本数量相同,但在实际的分类任务中,不同类别的样本数量往往是不同的,这种情况被称为数据不平衡。例如,在入侵检测任务中,疑似入侵行为的样本远远少于正常行为样本,正确识别少数类样本将获得更大的价值。在数据不平衡情况下提高分类器性能主要有两种方式:数据预处理和算法改进。

数据预处理是通过对训练数据采用欠采样或过采样的策略来改善其分布的方法。Deng 等[1]提出一种基于超平面排序,分层抽样、多类样本重组的数据采样方法,以得到可用于机器学习的分类平衡数据集。Douzas 等[2]利用条件生成对抗网络(Conditional Generative Adversarial Nets,CGAN)生成少数类样本;Zhang 等[3]用高斯混合模型(Gaussian Mixture Model,GMM)拟合样本分布,结合少数类样本合成过采样技术(Synthetic Minority Over-sampling Technique,SMOTE)从而生成样本;Yi 等[4]为解决 SMOTE 算法在对少数类样本进行线性插值时未考虑不同样本分布的问题,提出了结合聚类的 SMOTE 算法;Tao 等[5]在分析大量过采样方法后,针对产生重复样本和噪声样本的问题,提出了基于实质否定选择的过采样方法,实验结果表明,相比其他过采样方法,该方法具有明显优势。

改进算法的方法主要有代价敏感学习方法和集成学习方法[6]。不同决策结果的代价不同,例如在软件漏洞检测中,未能发现漏洞的代价比错误判断为漏洞的代价更高。代价敏感学习根据定义的代价信息,如样本获取代价、样本误分类代价等[7],来改进算法性能。数据不平衡通常伴随样本误分类代价不平衡[8],例如在入侵检测中具有高误分代价的入侵行为

到稿日期:2021-04-07　返修日期:2021-09-07
通信作者:王晓丹(afeu_wang@163.com)

的样本数量远远少于低误分代价的正常行为的样本数量，因此可以利用误分类代价信息改进样本的分类过程[9]。Ping等[10]将代价敏感与决策树相结合，提出了基于聚类的弱平衡代价敏感随机森林算法，用于充分学习少数类样本。Jing等[11]通过随机分隔得到多个平衡子集，利用代价敏感多集学习的方法得到分类特征。Tao等[12]为提高集成支持向量机的分类性能，提出了自适应地确定误分类代价的方法。在分类器迭代时，通过自适应地调节不同少数类的贡献，使决策向远离少数类的方向倾斜，提高了分类器的性能。集成学习方法结合了多个分类器来给出样本的预测类别[13]。为解决不平衡数据问题，集成学习常与代价敏感和数据预处理等方法结合[6,13]。Garcia等[14]为解决多标签的不平衡分类问题，提出了动态集成选择方法，每个子分类器的价值由样本权重来评估，根据样本集的特征动态选择合适的分类器集来提高分类效果。Chen等[15]通过随机欠采样多份与少数类等量的多数类样本来构成多个训练子集，并用其训练多个差异化的Ext-GBDT模型，最后以所有子模型的预测概率的平均值为分类结果。Tao等[16]提出的基于最大软边界的密度敏感支持向量数据描述分类算法可以更好地利用训练集中的少数类样本，实验结果表明该方法在多个数据集上超越了传统方法。Zhang等[17]综合代价敏感和间隔理论设计了CMBoost算法，在客户信用数据集上取得了很好的效果。

XGBoost是基于梯度提升的集成决策树算法，在医疗检测[18]、地质探测[19]等领域被广泛运用。但XGBoost在处理信用卡风险预测、网络入侵检测、医疗检测等不平衡数据集时效果不理想，通常需要结合样本采样技术来提高性能[20-22]。Chen等[23]提出了w-XGBoost算法，通过设计权重函数来赋予每个样本不同的权重，使算法关注高权重样本；Zou等[24]提出了数据调节策略(Data Adjustment，DA)，对少数类样本采用SMOTE，对多数类样本采用随机欠采样(Random Under-sampling，RUS)，以减小数据的不平衡程度；Saner等[25]对比了随机过采样、SMOTE、ADASYN算法结合XGBoost分类模型的效果，根据正负样本交叉熵损失的比值来决定过采样策略。

本文通过总结数据不平衡条件下XGBoost进行二分类任务时，分裂点选择和样本预测值更新这两个过程的规律发现，样本被误分类后的梯度变化是影响分类器分类效果的重要因素。通过构建代价敏感激活函数，将Logistic损失函数的一阶梯度变化与正负类别的误分类代价关联，使样本在不同分类结果下的一阶梯度变化不同，提高了算法对少数类样本的重视程度。本文在4组公共数据集上进行了实验，相比其他4种方法，本文提出的代价敏感激活函数XGBoost算法在多个评价指标上获得了较好的结果。

2 XGBoost算法的原理

对于有 n 个样本、m 个特征的数据集 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}(|\mathcal{D}|=n,\boldsymbol{x}_i\in\mathbb{R}^m,y_i\in\mathbb{R})$，样本的预测值 $\phi(\boldsymbol{x}_i)$ 是各棵决策树的结果 $f_i(\boldsymbol{x}_i)$ 之和。假设生成了 K 棵决策树，则样本预测值为：

$$\phi(\boldsymbol{x}_i)=\sum_{i=1}^{K}f_i(\boldsymbol{x}_i) \tag{1}$$

下文介绍决策树的生成过程。算法的损失函数为：

$$\mathcal{L}(\phi)=\sum_i l(y_i,\hat{y}_i)+\sum_k\Omega(f_k) \tag{2}$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|W\|^2 \tag{3}$$

其中，$l(y_i,\hat{y}_i)$ 是样本损失，$\Omega(f_k)$ 是正则化函数，惩罚系数 γ 降低了树的叶节点数量 T，$\boldsymbol{W}=(w_1,w_2,w_3,\cdots,w_T)$ 是全部叶节点权重构成的向量，λ 系数则限制了叶节点权重的大小。

第 t 次迭代生成的决策树 f_t 降低了第 $t-1$ 次迭代后的损失，如式(4)所示：

$$\mathcal{L}^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(\boldsymbol{x}_i))+\sum_{i=1}^{t-1}\Omega(f_i)+\Omega(f_t) \tag{4}$$

对损失函数进行二阶泰勒展开，如式(5)、式(6)所示：

$$\mathcal{L}^{(t)}\simeq\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{(t-1)})+g_if_t(\boldsymbol{x}_i)+\frac{1}{2}h_if_t^2(\boldsymbol{x}_i)\right]+\sum_{i=1}^{t-1}\Omega(f_i)+\Omega(f_t) \tag{5}$$

$$g_i=\partial_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)}),h_i=\partial^2_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)}) \tag{6}$$

$\text{leaf}(\boldsymbol{x}_i)=j$ 表示样本 $\boldsymbol{x}_i$ 被决策树 f_t 分类至第 j 个叶节点，决策树的结果 $f_t(\boldsymbol{x}_i)$ 为该叶节点的权重 w_j。去掉常量 $l(y_i,\hat{y}_i^{(t-1)})$ 和 $\sum_{i=1}^{t-1}\Omega(f_i)$ 后，第 t 次迭代生成的决策树把样本分类至 T 个叶节点后的损失为：

$$\tilde{\mathcal{L}}^{(t)}=\sum_{j=1}^{T}\left[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2\right]+\gamma T \tag{7}$$

其中，$I_j=\{i|\text{leaf}(\boldsymbol{x}_i)=j\}$ 表示第 j 个叶节点中样本的索引。将叶节点 j 的损失关于 w_j 求偏导，求得使损失最小的叶节点权重 w_j^*：

$$w_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \tag{8}$$

式(8)被称作叶节点权重计算公式。将 w_j^* 代入叶节点损失函数，得到叶节点 j 的损失为：

$$\mathcal{L}_{\text{before}}=-\frac{1}{2}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma \tag{9}$$

将式(9)作为节点的结构损失，将节点中的样本根据第 i 个特征的值 σ 分为两个节点，即左节点 $S_L=\{\boldsymbol{x}_i|\sigma(\boldsymbol{x}_i)\leqslant\theta\}$ 和右节点 $S_R=\{\boldsymbol{x}_i|\sigma(\boldsymbol{x}_i)>\theta\}$，$\theta$ 值称为分裂点。样本索引分别为 I_L 和 I_R，分裂后的结构损失为：

$$\mathcal{L}_{\text{after}}=-\frac{1}{2}\left(\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}\right)+2\gamma \tag{10}$$

故叶节点 j 以第 i 个特征的 θ 为分裂点获得的结构增益为：

$$\begin{aligned}\mathcal{L}_{\text{split}}&=\mathcal{L}_{\text{before}}-\mathcal{L}_{\text{after}}\\&=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}\right]-\gamma\end{aligned} \tag{11}$$

式(11)被称为分裂增益公式[26]，用于衡量一个分裂节点的好坏。分裂系数 γ 用于减小结构的复杂度，防止决策树过拟合。XGBoost算法根据分裂增益公式在各个特征中寻找分裂点以进行分裂，决策树构建完成后计算叶节点权重以更新样本预测值，下一次迭代则以新的样本预测值构建一棵决策树，最终以多颗决策树的预测结果之和作为最终的预测值。

3 不平衡数据对XGBoost框架的影响

3.1 不同分类结果下损失函数的梯度变化

混淆矩阵可用于表示样本的分类结果。在二分类问题

中，通常定义少数类为正类、多数类为负类，分别用 1 和 0 标记，混淆矩阵如式(12)所示：

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \tag{12}$$

其中，TP(True Positive)表示预测为正类实际也为正类的样本数量，FP(False Positive)表示预测为正类但实际为负类的样本数量，FN(False Negative)表示预测为负类但实际为正类的样本数量，TN(True Negative)表示预测为负类实际也为负类的样本数量。每一列之和为样本某一类别的真实数量，每一行之和为预测样本为某一类别的数量。

对于二分类问题，XGBoost 一般采用逻辑回归：

$$L(P(y_i|\boldsymbol{x}_i))=-\log P(y_i|\boldsymbol{x}_i) \tag{13}$$

其中，$P(y_i|\boldsymbol{x}_i)$指样本$(\boldsymbol{x}_i,y_i)$属于真实类别 y_i 的概率，在二分类问题中类别标签 $Y=\{0,1\}$，$y_i\in Y$，0 和 1 分别表示负类和正类。$P(y_i|\boldsymbol{x}_i)$通过 Sigmoid 函数映射得到，函数形式如式(14)所示：

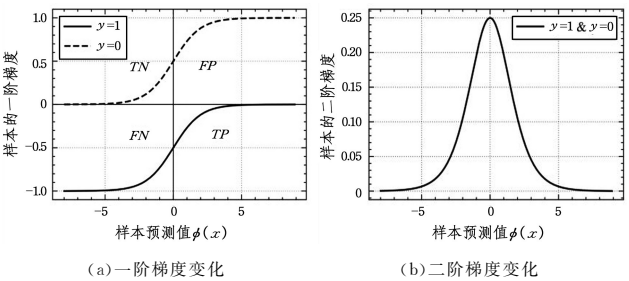
$$P(y_i|\boldsymbol{x}_i)=\begin{cases} S(\phi(\boldsymbol{x}_i))=\dfrac{1}{1+\mathrm{e}^{-\phi(\boldsymbol{x}_i)}}, & y_i=1 \\ 1-S(\phi(\boldsymbol{x}_i))=\dfrac{1}{1+\mathrm{e}^{\phi(\boldsymbol{x}_i)}}, & y_i=0 \end{cases} \tag{14}$$

将式(14)代入损失函数并对预测值 $\phi(\boldsymbol{x}_i)$进行求导，得到样本损失函数关于预测值的一阶梯度 g_i 和二阶梯度 h_i：

$$g_i=\partial L(P(y_i|\boldsymbol{x}_i))/\partial\phi(\boldsymbol{x}_i)=\begin{cases} -\dfrac{1}{1+\mathrm{e}^{\phi(\boldsymbol{x}_i)}}, & y_i=1 \\ \dfrac{1}{1+\mathrm{e}^{-\phi(\boldsymbol{x}_i)}}, & y_i=0 \end{cases} \tag{15}$$

$$h_i=\partial^2 L(P(y_i|\boldsymbol{x}_i))/\partial\phi(\boldsymbol{x}_i)^2=\begin{cases} \dfrac{\mathrm{e}^{\phi(\boldsymbol{x}_i)}}{(1+\mathrm{e}^{\phi(\boldsymbol{x}_i)})^2}, & y_i=1 \\ \dfrac{\mathrm{e}^{-\phi(\boldsymbol{x}_i)}}{(1+\mathrm{e}^{-\phi(\boldsymbol{x}_i)})^2}, & y_i=0 \end{cases} \tag{16}$$

样本被正确分类时一阶梯度的绝对值减小，当 TN 类样本预测值减小至 −5.0、TP 类样本的预测值增大至 5.0 时，其趋于 0。样本被错误分类时一阶梯度的绝对值增大，FN 类样本预测值减小至 −5、FP 类样本预测值增大至 5 时，其趋于 1，随后预测值的改变不会引起梯度值的较大变化。正负样本的二阶梯度相同，当样本预测值的绝对值由 0 增大时，样本的二阶梯度 h_i 由最大值 0.25 减小并逐渐趋于 0，如图 1 所示。



(a)一阶梯度变化　(b)二阶梯度变化

图 1　损失函数关于预测值的一阶梯度变化和二阶梯度变化

Fig. 1　First and second order gradient of loss function with respect to predicted value

3.2　样本数量不平衡对分裂点的影响

在选择分裂点时，将样本按某特征的值排序。分裂增益式(11)表明，由于被错误预测的正负类样本的 g 值符号相反，为使分裂后结构损失降低最多，需要找到某一特征的分裂点将节点中的 FP 和 FN 类样本分裂后置于不同节点中。

若 FP 类样本分布在区间(a,b)内，FN 类样本分布在区间(c,b)内，则两区间重合的区域称为混合区间。

若节点中的样本按特征 α 的值从小到大排列，则 FP 类样本分布在$(-\infty,A)$区间，FN 类样本分布在$(B,+\infty)$内，混合区间为(A,B)，则分裂点在混合区间内，如图 2 所示。

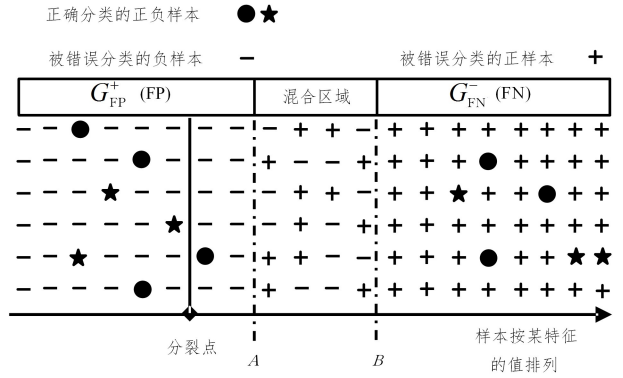


图 2　样本按某一特征值排序后的分布图

Fig. 2　Distribution of samples arranged by the value of an attribute

证明：分裂后左右节点的样本 h 值之和分别为 H_{L} 和 H_{R}。区间$(-\infty,A)$内样本的一阶梯度和为 G_{FP}^{+}，区间$(B,+\infty)$内样本的一阶梯度和为 G_{FN}^{-}，混合区间(A,B)内一阶梯度和为 G_{M}，则分裂点在 A 处的结构损失如式(17)所示：

$$\mathscr{L}_A=-\frac{1}{2}\left(\frac{(G_{\mathrm{FP}}^{+})^2}{H_{\mathrm{L}}+\lambda}+\frac{(G_{\mathrm{FN}}^{-}+G_{\mathrm{M}})^2}{H_{\mathrm{R}}+\lambda}\right)+2\gamma \tag{17}$$

当分裂点 P 小于 A 时，区间(P,A)内样本一阶梯度和为 ΔG，则左节点 $\sum_{i\in I_{\mathrm{L}}} g_i=G_{\mathrm{FP}}^{+}-\Delta G$，右节点 $\sum_{i\in I_{\mathrm{R}}} g_i=G_{\mathrm{FN}}^{-}+G_{\mathrm{M}}+\Delta G$。由于 ΔG 大于 0，因此在 P 处的结构损失大于在 A 处的结构损失。同理，B 点右侧处的结构损失大于在 B 处的结构损失，因此分裂点一定处于混合区间(A,B)内，结论成立。

分裂点在混合区间内从 A 点向右移动的过程中，若在当前位置 P_n 和下一位置 P_{n+1} 的区间(P_n,P_{n+1})内样本的一阶梯度和 G' 大于 0，则分裂点 P_{n+1} 优于 P_n。

证明：分析结构损失时主要关注 G 的变化，并且由于混合区域内样本占总样本的比例很小，当分 裂点在区间(A,B)移动时，忽略 H 的变化。假设 H_{L} 和 H_{R} 相等为 H。在混和区间(A,P_1)内样本的一阶梯度和为 $G_m^{(0)}$，在混和区间(P_1,B)内样本的一阶梯度和为 $G_r^{(0)}$，如图 3 所示。

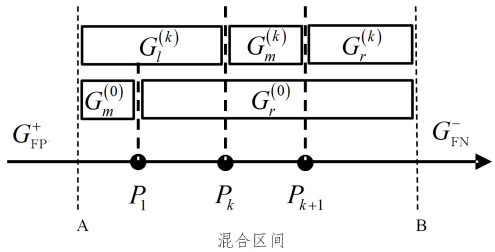


图 3　分裂候选点在样本混合区间的移动

Fig. 3　Split candidate moves in sample mixing region

当 $n=0$ 时，$P_0=A$。当分裂点由 A 向 P_1 移动时，分裂点

在 P_1 时的损失小于在 A 点的损失。

$$\mathscr{L}_A-\mathscr{L}_{P_1}=-\frac{2G_m^{(0)}(G_{\mathrm{FN}}^{-}-G_{\mathrm{FP}}^{+}+G_r^{(0)})}{H+\lambda}>0 \tag{18}$$

由于混合区域中的样本数量只占节点中样本的一部分，因此$|G_r^{(0)}|<|G_{\mathrm{FN}}^{-}-G_{\mathrm{FP}}^{+}|$，故 $G_{\mathrm{FN}}^{-}-G_{\mathrm{FP}}^{+}+G_r^{(0)}<0$，$G_m^{(0)}>0$。

当 $n=k$ 时分裂点由 P_k 向 P_{k+1} 移动（$P_k<P_{k+1}$），在混合区间(A,P_k)内样本的一阶梯度和为 $G_l^{(k)}$，在混合区间(P_k,P_{k+1})内样本的一阶梯度和为 $G_m^{(k)}$，在混合区间(P_{k+1},B)内样本的一阶梯度和为 $G_r^{(k)}$。当分裂点在 P_k 时，满足条件 $G_l^{(k)}>0$。

$$\mathscr{L}_{P_k}-\mathscr{L}_{P_{k+1}}=-\frac{2G_m^{(k)}(G_{\mathrm{FN}}^{-}-G_{\mathrm{FP}}^{+}+G_r^{(k)}-G_l^{(k)})}{H+\lambda}>0 \tag{19}$$

其中，$G_{\mathrm{FN}}^{-}-G_{\mathrm{FP}}^{+}+G_r^{(k)}-G_l^{(k)}<0$，故 $G_m^{(k)}>0$，结论成立。

在不平衡样本集中，多数类样本在混合区域的数量远高于少数类。分裂点向右移动时，若经过的区间内样本的一阶梯度和大于 0，且区间中部分少数类样本周围存在大量多数类样本，为使分裂增益更大，则将少数类样本分入多数类样本的节点中。

3.3 对样本预测值更新的影响

第 $t+1$ 轮迭代输入的样本预测值 $\hat{y}_i^{(t+1)}$ 为第 t 轮迭代输入的样本预测值 $\hat{y}_i^{(t)}$ 加上样本在第 t 轮迭代生成的决策树中所在叶节点的权重 ω_j^*。

$$\hat{y}_i^{(t+1)}=\hat{y}_i^{(t)}+\omega_j^* \tag{20}$$

样本的一阶梯度随预测值变化，样本预测值更新后，若 $\omega_j^*>0$，各类样本预测值增大，梯度增大；若 $\omega_j^*<0$，各类样本预测值减小，梯度减小。如图 4 所示。

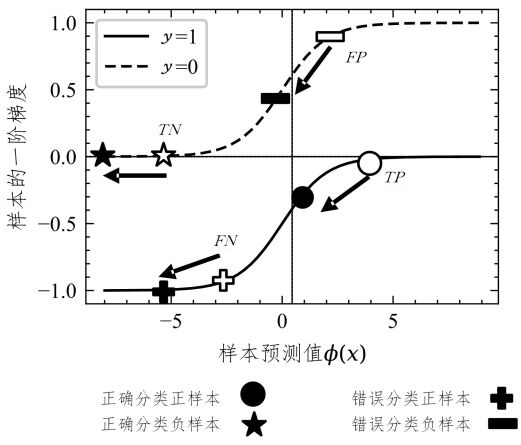


图 4 预测值更新后样本在不同分类结果下的梯度变化

Fig. 4 Gradient variations of samples after updating predictive value

将样本预测为真实类别的预测值称为期望值，正样本的期望值为$+\infty$，负样本的期望值为$-\infty$。样本的一阶梯度绝对值$|g_i|$的变化范围为(0,1)，当预测值偏离期望值时其趋于1，当预测值接近期望值时其趋于 0。当不同类别的样本数量差距较大时，由于样本的一阶梯度变化范围小，因此叶节点权重的大小和正负主要由多数类决定。例如，节点中有 10 个 $g=0.2$ 的FN类样本和2个 $g=-0.8$ 的FP类样本，最终，$\sum_{i\in I_j} g_i>0$，$\omega_j^*<0$，节点内的少数类样本偏离期望值。

综上所述，数据不平衡会对算法产生以下影响，这些影响共同导致少数类样本无法被有效预测。

(1)分裂点在混合区间移动时，少数类样本周围存在大量多数类样本，使部分少数类样本被分入多数类样本的节点中。

(2)样本预测值的更新对不同类别的样本的影响不同。对于有大量负样本的节点，更新后的负样本更接近期望值，同时使节点内的正样本偏离期望值而产生误分类正样本。被误分的正样本的一阶梯度变化较小，且绝对值最大为 1，在计算节点的梯度和时被数量较多的负样本的梯度抵消，无法根据节点的梯度值对正样本的预测值进行有效的更新。

为改善上述情况，高误分代价样本与低误分代价样本在被误分时应获得不同的梯度变化。高误分代价样本在被误分时一阶梯度应快速增加，并且一阶梯度的绝对值应比低代价样本高几倍，以防止多数类样本因数量优势而干扰决策。

4 代价敏感激活函数

对于二分类问题，误分类代价矩阵为：

$$Cost=\begin{pmatrix} c_{++} & c_{+-} \\ c_{-+} & c_{--} \end{pmatrix} \tag{21}$$

其中，c_{ab} 是将 a 类样本被预测为 b 类而造成的代价损失。根据贝叶斯最优决策理论，将期望代价最小的一类作为判别结果。若判断样本类别为正类，则相应的期望代价的计算和判别过程为：

$$c_{++}P_+(\boldsymbol{x}_i)+c_{-+}(1-P_+(\boldsymbol{x}_i))<c_{+-}P_+(\boldsymbol{x}_i)+c_{--}(1-P_+(\boldsymbol{x}_i)) \tag{22}$$

$$P_+(\boldsymbol{x}_i)>P_b=\frac{c_{-+}-c_{--}}{(c_{-+}-c_{--}+c_{+-}-c_{++})} \tag{23}$$

其中，$P_+(\boldsymbol{x}_i)$计算样本为正类的概率，P_b 为概率决策阈值。对于二分类问题，当样本为正类的概率大于 P_b 时，预测样本为正类，否则样本为负类。当误分类代价相同时，$P_b=0.5$，属于传统贝叶斯决策的无偏损失。

同样给出预测值决策阈值 E，当预测值大于 E 时，预测样本为正类。在 Sigmoid 激活函数中，预测值为 0 时预测为正类的概率为 0.5，预测值决策阈值 0 与概率决策阈值 P_b 对应。当误分类代价不同时，概率决策边界为：

$$\hat{y}_i=\begin{cases}1, & P_+(\boldsymbol{x}_i)>P_b\\ 0, & P_+(\boldsymbol{x}_i)<P_b\end{cases} \tag{24}$$

为使预测值决策阈值 0 与概率决策阈值 P_b 对应，需要对 Sigmoid 激活函数进行平移，在 Sigmoid 函数中添加参数 α。

$$P(y_i|\boldsymbol{x}_i)=\begin{cases}J(\phi(\boldsymbol{x}_i))=\dfrac{1}{1+\alpha e^{-\phi(\boldsymbol{x}_i)}}, & y_i=1\\ 1-J(\phi(\boldsymbol{x}_i))=\dfrac{\alpha}{\alpha+e^{\phi(\boldsymbol{x}_i)}}, & y_i=0\end{cases} \tag{25}$$

其中，α 参数使预测值增加 $-\log\alpha$。样本预测值为 0 时，被判断为正类的概率为 $1/(\alpha+1)$，等于概率决策边界 P_b，计算得到 α 与分类代价信息的关系：

$$\begin{cases}\dfrac{1}{1+\alpha e^{-\phi(\boldsymbol{x}_i)}}=P_b\\ \phi(\boldsymbol{x}_i)=0\end{cases} \tag{26}$$

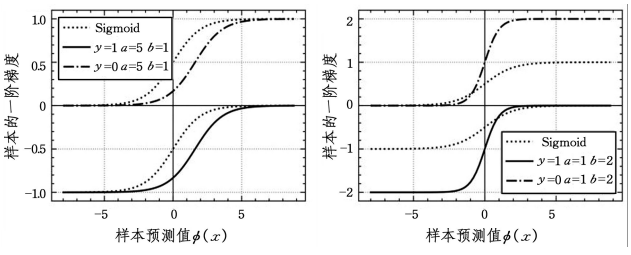
$$P_b=\frac{1}{1+\alpha} \tag{27}$$

$$\alpha=\frac{c_{+-}-c_{++}}{c_{-+}-c_{--}} \tag{28}$$

当正确分类的代价为 0 时，$\alpha=c_{+-}/c_{-+}$，称作样本的误分类代价比值。为提高少数类样本的梯度绝对值和变化速度，引入了 β 参数。

$$P(y_i|\boldsymbol{x}_i)=\begin{cases}J(\phi(\boldsymbol{x}_i))=\dfrac{1}{1+\alpha e^{-\beta\phi(\boldsymbol{x}_i)}}, & y_i=1\\ 1-J(\phi(\boldsymbol{x}_i))=\dfrac{\alpha}{\alpha+e^{\beta\phi(\boldsymbol{x}_i)}}, & y_i=0\end{cases}\tag{29}$$

其中，β 值影响激活函数中概率值随预测值变化的快慢，使得损失函数一阶梯度的绝对值最大为 β，二阶梯度的最大值等于 $\beta^2/4$。当正样本的误分代价比负样本的误分代价高时（$c_{+-}>c_{-+}$），$\alpha>1$，误分代价低的负样本的梯度绝对值降低，误分代价高的正样本的梯度绝对值增大，如图 5 所示。



(a)Change parameter a (b)Change parameter b

图 5 改变参数后样本的一阶梯度曲线

Fig. 5 Comparison of one order gradient of loss function after changing parameters

同一节点中的样本会获得相同的预测更新值 ω_j^*。为使具有高误分代价的少数类样本在决策树构建过程中更受重视，在预测值更新后偏离期望值时其梯度绝对值应快速增大，在接近期望值时其梯度绝对值应缓慢减小，使少数类样本被持续关注；相反，对于误分类代价小的多数类样本，在预测值更新后偏离期望值时其梯度绝对值应快速减小，在接近期望值时其梯度绝对值应缓慢增大，以减少算法对小代价样本的关注。少数类样本在本次迭代过程偏离期望值后，会依据误分代价获得足够大的一阶梯度，使其在下一次迭代过程中被分入少数类样本较多的节点而被正确预测。为此，构造分段激活函数：

$$P_+(\boldsymbol{x}_i)=\begin{cases}\dfrac{1}{1+\alpha e^{-\beta\phi(\boldsymbol{x}_i)}}, & P_+<P_b\\ \dfrac{1}{1+\alpha e^{-\phi(\boldsymbol{x}_i)}}, & P_+\geqslant P_b\end{cases}\tag{30}$$

由于 $P_+(\boldsymbol{x}_i)$ 函数的梯度在 $x=0$ 处不连续，因此定义函数在 0 处的 n 阶梯度等于预测值趋近 0^+ 和 0^- 时 n 阶梯度和的 1/2。

$$P_+(\boldsymbol{x}_i)=\begin{cases}\dfrac{1}{1+\alpha e^{-\beta\phi(\boldsymbol{x}_i)}}, & \phi(\boldsymbol{x}_i)<0\\ \dfrac{1}{1+\alpha e^{-\phi(\boldsymbol{x}_i)}}, & \phi(\boldsymbol{x}_i)\geqslant 0\end{cases}\tag{31}$$

$$P_+^{(n)}(0)=\frac{1}{2}(P_+^{(n)}(0^+)+P_+^{(n)}(0^-))\tag{32}$$

对比代价敏感激活函数与 Sigmoid 函数，当误分类代价比值 α 增大时，样本更容易被预测为正类，并且当预测值小于决策边界时，样本预测概率随预测值变化更快。将式(31)代入 Logistic 回归损失函数，绘制 $\alpha=4$，$\beta=2$ 时样本损失函数的一阶梯度变化图像，如图 6 所示。

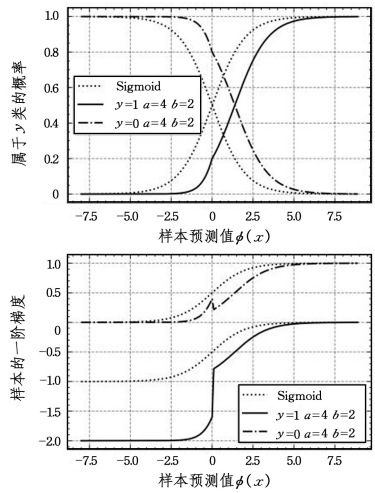


图 6 代价敏感激活函数及损失函数的一阶梯度变化

Fig. 6 Cost-sensitive activation function and first order gradient of loss function

相比 Sigmoid 函数，代价敏感激活函数有以下特性：

(1)高代价样本被误分时，一阶梯度有一个阶跃式的增长，梯度的绝对值最大为 β，且梯度更快逼近最大值；而被正确分类后，样本仍保持一定梯度，使少数类样本对算法的决策过程仍有较大影响。

(2)低代价样本在被正确分类后，一阶梯度快速降低，更快接近 0；被错误分类后，一阶梯度缓慢增加，并且相比 Sigmoid 函数，样本的一阶梯度更小。

(3)低代价样本即将被误分时，梯度会有较大的提升，使样本在下一轮中被正确分类，使一阶梯度接近 0，减小了低代价样本对决策的影响。

综上，当混合区域中低代价样本较多时，可以减少多数类节点中少数类样本的数量，提高算法对被误分少数类样本的关注，使少数类样本接近期望值，防止少数类样本被误分。

β 值的选择与代价比例 α 有关。当预测值为 0，使用代价敏感激活函数时多数类样本（负样本）Logistic 损失函数的一阶梯度小于使用 Sigmoid 函数时的一阶梯度 0.5，得到的关系式如下：

$$\left.\frac{\beta e^{\beta x}}{\alpha+e^{\beta x}}\right|_{x=0}\leqslant\frac{1}{2}\tag{33}$$

$$\beta\leqslant\frac{1}{2}(\alpha+1)\tag{34}$$

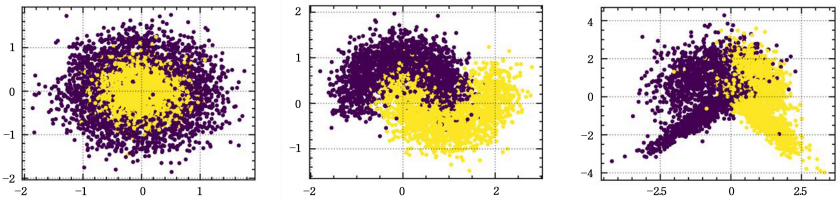
5 实验与结果分析

5.1 函数参数的选择

本文提出的代价敏感算法使用基于类别的误分类代价[27]，即同一类样本的误分类代价相同。代价矩阵的设置可由该领域专家根据经验给出，或采用参数学习的方法获得[28]。由于采用了不同领域的数据集，因此本文采用参数学习的方法来确认类别误分类代价。

使用统计模型分析参数的初始值。为得到不同样本不平衡比例的数据集，在 python 环境中使用 sklearn 包中的 dataset 工具生成特征维数为 2、不同分布的数据集 c_1，c_2，c_3（random_state=2）。实验数据集中，正负样本由随机

采样得到，3 种数据分布模拟在不同特征中复杂的样本分布情况。由于现实问题中少数类样本更难获得，因此每次随机抽取负样本2000 例，正样本按照正负样本比例 R（负样本数量除以正样本数量）进行欠采样。从 c_1，c_2，c_3 中随机抽取正负样本各 1000 例构成测试集，数据集分布如图 7 所示。

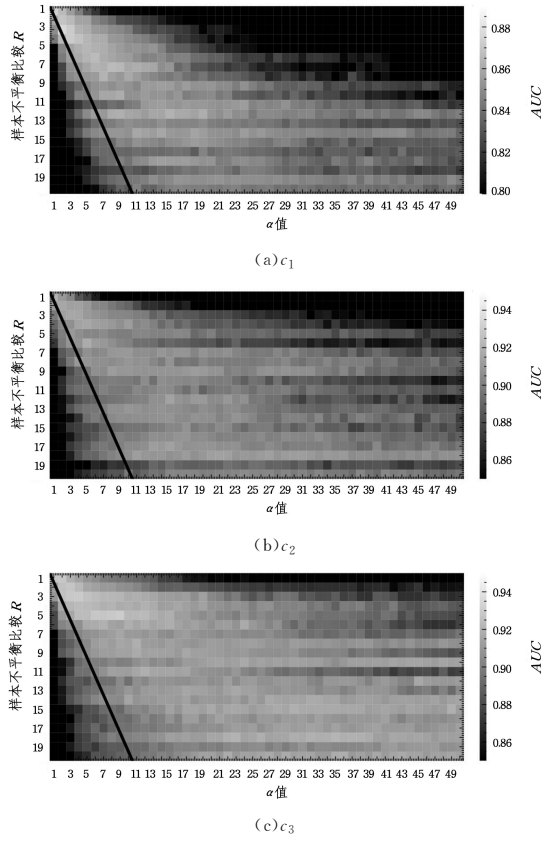


(a)c_1 (b)c_2 (c)c_3

图 7 不同数据集的分布

Fig. 7 Distributions of different datasets

设 $\beta=1$，观察不同 α 值下模型的 AUC 值，AUC 值表示在不同决策域或分界值下，以正例中被判为正例的概率为纵坐标，以被判定为正例却不是正例的概率为横坐标绘制的曲线下的面积。在不同样本不平衡比例 R 的数据集中，当 $1<\alpha<R/2$，$R>2$ 时，模型的 AUC 值有显著提升。这部分提升是由于增加代价参数后修正了样本不平衡导致的分类阈值偏移，使得测试集中的少数类样本能够被正确分类。当 $\alpha>R/2$，$R>2$ 时，增大 α 会使 AUC 值产生小幅波动，这是由于分类决策面在少数类和多数类样本混合区域内移动时，少数类样本的查全率（被预测为少数类且实际为少数类的样本数量除以实际为少数类的样本数量）提高而查准率（被预测为少数类且实际为少数类的样本数量除以预测为少数类的样本数量）降低，如图 8 所示。



(a)c_1

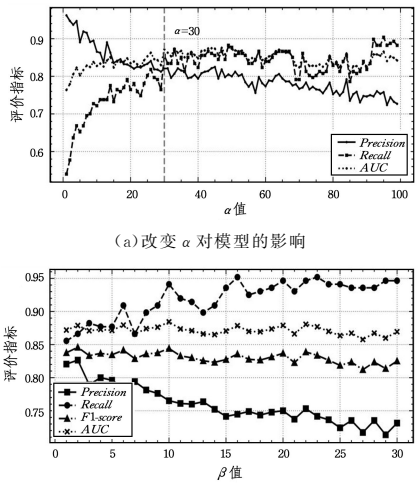
(b)c_2

(c)c_3

图 8 代价比值 α 对模型性能的影响

Fig. 8 Influence of cost ratio and on model performance

对于参数 β，使用不平衡比例较大的真实数据集进行实验。Wilt 数据集的不平衡比例 $R=57.6$，随着 α，β 的增加，CSAF-XGBoost 算法对少数类样本的查准率（Precision）下降，而查全率（Recall）提高，F_1 值（查准率和查全率的调和平均数）与 AUC 值则增长至一定程度后保持稳定。可以根据实际需求选择 CSAF-XGBoost 算法的 β 值。如果对少数类样本的查全率要求高，则可以适当提高 β 值；如果对少数类样本的查准率要求高，则可以适当降低 β 值，如图 9 所示。



(a)改变 α 对模型的影响

(b)改变 β 对模型的影响

图 9 CSAF-XGBoost 在不同 α，β 值下的性能

Fig. 9 Performance of CSAF-XGBoost with different α and β

5.2 不平衡比例变化下算法的稳健性分析

为观察不平衡数据对 XGBoost 框架分类造成的影响，以少数类样本的查全率（*Recall*）、F_1 值与 AUC 值为评价指标。通过实验发现，各性能指标随样本不平衡比例的增加呈下降趋势，当样本不平衡比例为 10 时，查全率由 0.9～0.95 下降至 0.55～0.75，*F-score* 和 AUC 值由 0.9～0.95 下降至 0.75～0.85，样本不平衡比例超过 10 后，各项指标继续下降。这说明数据不平衡情况使 XGBoost 对少数类样本的识别能力下降，并且样本不平衡比例越大，算法对少数类样本的识别能力就越弱。由于随机采样使各个数据集中少数类样本的分布不同，使曲线在下降的过程中产生波动，因此使用代价敏感激活函数后，各项指标均有所提高，接近样本平衡（不平衡比例为 1）时的识别效果。在样本不平衡比例达到 10 时，CSAF-XGBoost 算法对少数类样本的各项指标仍维持在 0.9 左右，若继续增大不平衡比例，则本文算法对少数类样样本的识别能力下降缓慢，如图 10 所示。

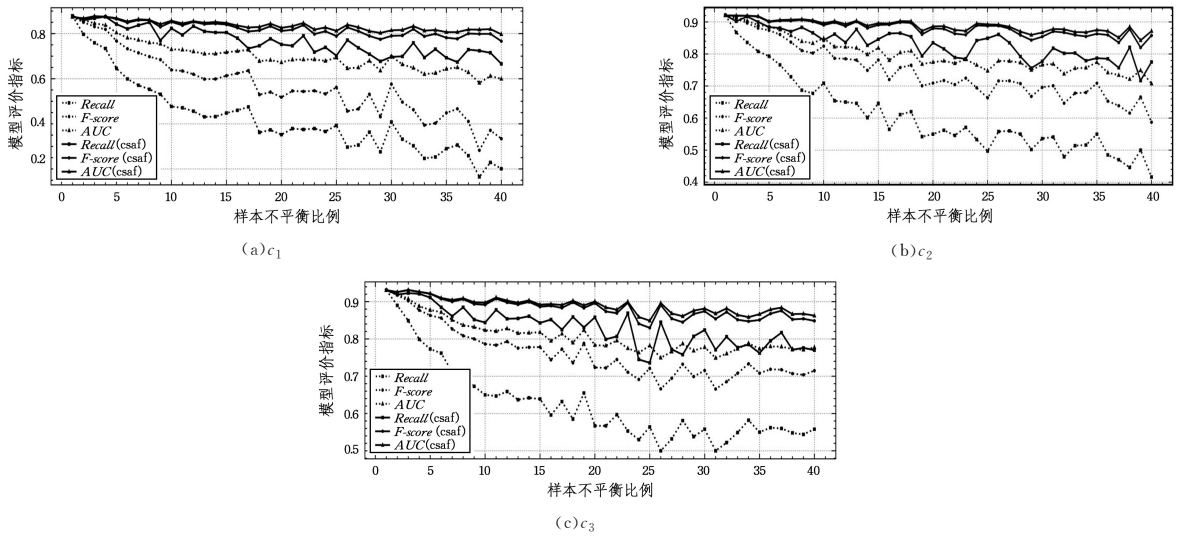


图 10 CSAF-XGBoost 在高度不平衡情况下对少数类样本的识别能力

Fig. 10 Recognition ability of CSAF-XGBoost to minority class samples with highly unbalanced dataset

5.3 不同算法的分类效果对比

使用 SMOTE-XGBoost 算法[21,29]、ADASYN-XGBoost 算法[25,29]、Weighted-XGBoost[23,30]、Focal-XGBoost[30-32] 以及本文的代价敏感激活函数这 5 种方法在医学、勘探、工业筛选等应用领域中的 4 种数据集上进行实验(来自 UCI 数据集)，数据集的统计特性如表 1 所列。

所有算法的超参数采用贝叶斯优化[33]，算法中非默认的超参数如表 2 所列。

实验结果表明，使用代价敏感激活函数后，算法对少数类样本的各个评价指标均有较大的提高。相比 Weighted-XGBoost 算法，少数类样本的查全率提高了 0.05～0.20，F_1 值和 *AUC* 值提高了 0.06～0.10。相比 ADASYN-XGBoost 算法，查全率提高了 0.09～0.11，F_1 值和 *AUC* 值提高了 0.01～0.08，如图 11 所示。

表 1 实验数据集的基本统计特性

Table 1 Statistical characteristics of datasets

数据集名称	特征数量	不平衡比例	样本数量	训练集			测试集		
				总计	正样本数量	负样本数量	总计	正样本数量	负样本数量
wilt	6	57.64	4839	4339	74	4265	500	187	313
glass1	9	1.66	214	149	56	93	65	20	45
yeast3	8	7.87	1484	1038	117	921	446	46	400
ecoli1	8	3.12	336	235	57	178	101	20	81

表 2 实验所用算法的超参数

Table 2 Hyper parameters of algorithms

算法名称	数据集名称			
	wilt	glass1	yeast3	ecoli1
XGBoost(origin)	'learning_rate':0.434, 'max_depth':5, 'min_child_weight':1.275, 'gamma':0.169, 'n_estimators':40	'learning_rate':0.5, 'max_depth':7, 'min_child_weight':1.578, 'gamma':0.1, 'n_estimators':42	'learning_rate':0.5, 'max_depth':7, 'min_child_weight':7.438, 'gamma':0.1, 'n_estimators':13	'learning_rate':0.5, 'max_depth':9, 'min_child_weight':3.346, 'gamma':0.449, 'n_estimators':35
Weight-XGBoost	imbalance_alpha=60	imbalance_alpha=5	imbalance_alpha=8.1	imbalance_alpha=3.4
Focal-XGBoost	gamma=2	gamma=2	gamma=2	gamma=3
Csaf-XGBoost	alpha=32,beta=7	alpha=3,beta=2	alpha=5,beta=3	alpha=3,beta=1.2
SMOTE-XGBoost	SMOTE: random_state=1, k_neighbors=20 XGBoost: 'gamma':1.0, 'learning_rate':0.5, 'max_depth':4.0, 'min_child_weight':2.79, 'n_estimators':39	SMOTE: random_state=1, k_neighbors=5 XGBoost: 'gamma':0.1, 'learning_rate':0.5, 'max_depth':6 'min_child_weight':0.1, 'n_estimators':33	SMOTE: random_state=1, k_neighbors=10 XGBoost: 'gamma':0.1, 'learning_rate':0.5, 'max_depth':7, 'min_child_weight':2.55,'n_estimators':45	SMOTE: random_state=1, k_neighbors=10 XGBoost: 'gamma':0.1, 'learning_rate':0.5, 'max_depth':8,'min_child_weight':1.77, 'n_estimators':35
ADASYN-XGBoost	ADASYN: random_state=1, n_neighbors=20 XGBoost: 'gamma':0.19, 'learning_rate':0.47, 'max_depth':4, 'min_child_weight':0.58, 'n_estimators':50	ADASYN: random_state=1, n_neighbors=5 XGBoost: 'gamma':0.1, 'learning_rate':0.5, 'max_depth':6, 'min_child_weight':0.1, 'n_estimators':27	ADASYN: random_state=1, n_neighbors=10 XGBoost: 'gamma':0.14, 'learning_rate':0.32, 'max_depth':6, 'min_child_weight':2.32,'n_estimators':34	ADASYN: random_state=1, n_neighbors=10 XGBoost: 'gamma':0.1, 'learning_rate':0.5, 'max_depth':6, 'min_child_weight':0.47, n_estimators':43

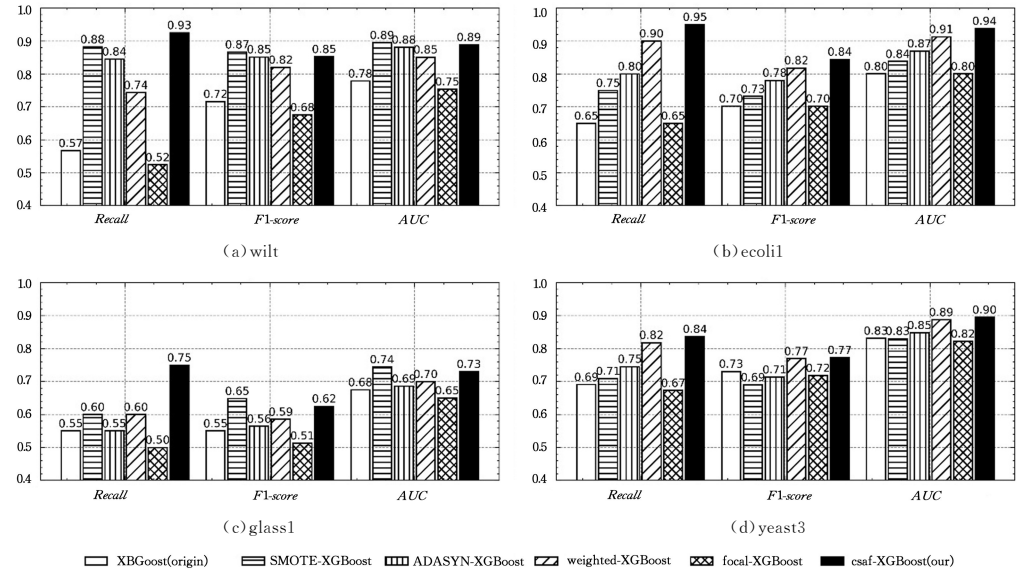
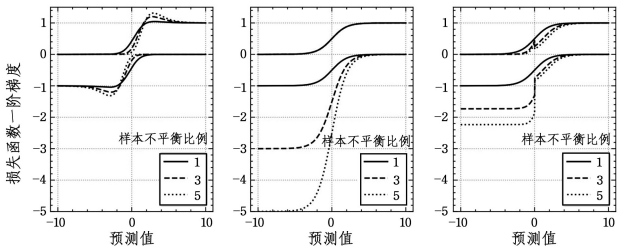


图 11 4 种方法在 4 组数据集下的性能对比

Fig. 11 Comparison of performance between four methods on four datasets

对比在数据不平衡比例 R 为 1,3,5 时不同损失函数的一阶梯度曲线发现,Focal Loss 损失函数的一阶梯度随 γ 值的变化改变很小,当 $\gamma=2.5$ 时,少数类样本梯度绝对值的最大值为 1.17,梯度变化不明显使 Focal-XGBoost 算法的识别能力的提升有限。而 Weighted Cross-entropy 损失函数的梯度变化明显,少数类样本的梯度随参数 α 成倍增加,梯度最大值为 α,对提升改进算法的性能具有显著作用,说明改变样本的梯度是提升算法分类性能的重要因素。代价敏感的激活函数不仅提高了少数类样本的梯度绝对值,还降低了多数类样本的梯度,不同分类结果的样本的梯度变化快慢也不同,使算法的分类效果优于使用 Weighted Cross-entropy 损失函数的 Weighted-XGBoost 算法,如图 12 所示。



(a)Focal loss 损失函数 (b)Weighted Cross-entropy 损失函数 (c)Cost-Logistic 损失函数

图 12 当数据不平衡比例为 1,3,5 时不同损失函数的一阶梯度

Fig. 12 One order gradient of different loss functions when imbalanced rate is 1,3,5

结束语 XGBoost 是一个性能优异的机器学习框架,但在数据不平衡条件下,该框架使用 Logistic 损失函数来解决二分类问题时性能下降显著。为提升 XGBoost 框架在不平衡数据条件下的分类性能,本文引入代价敏感激活函数来对其进行改进。

(1)分析不平衡数据对 XGBoost 框架进行二分类过程的影响,总结样本的一阶梯度在节点分裂点位置选择和样本预测值更新两个过程中的作用,发现少数类样本损失函数的一阶梯度绝对值小、变化不明显是导致其在算法迭代过程中无法被有效预测的重要原因。

(2)构造代价敏感激活函数,使不同误分类别的样本获得不同的梯度变化,高代价样本更受重视。对比 4 种方法在 4 组公共数据集上的分类效果,本文提出的代价敏感激活函数方法使 XGBoost 算法对少数类样本具有更强的识别能力。

本文提出的代价敏感激活函数基于类别的误分类代价,虽然简化了误分类代价的设置过程,但模型性能会有所下降。下一步可以将 CSAF-XGBoost 算法改进为基于样本的误分类代价。另外,根据人工数据集总结的参数初始化方法能否广泛用于其他领域,需要额外的大量实验进行验证。在进行参数优化时,本文使用网格搜索的方法,为进一步改进本文算法,可以结合更加高效的参数学习方法,以提升模型性能。后续还将对代价敏感激活函数的优化效果进行理论分析,为将该方法应用到其他分类器提供理论支撑。

参 考 文 献

[1] DENG M Y, GUO Y S, LIU T. Research on Imbalanced Data Sampling Method Based on Stratification and Recombination [J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35(8): 122-128.

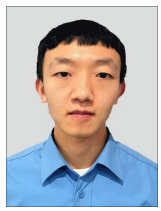
[2] GEORGIOS D, FERNADO B, et al. Effective data generation for imbalanced learning using conditional generative adversarial networks [J]. Expert Systems with Application, 2018, 91(1): 464-471.

[3] ZHANG H, HUANG L, WU C Q, et al. An Effective Convolutional Neural Network Based on SMOTE and Gaussian Mixture Model for Intrusion Detection in Imbalanced Dataset [J/OL]. Computer Networks, 2020, 177. https://www.sciencedirect.com/science/article/abs/pii/S1389128620300712.

[4] YI H K, JIANG Q C, YAN X F, et al. Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique with Wind Turbine Fault Detection Application [J]. IEEE Transactions on Industrial Informatics, 2021, 17(9): 5867-5875.

[5] TAO X M, LI Q, REN C, et al. Real-value negative selection

oversampling for imbalanced data set learning [J]. Expert Systems with Applications,2019,129:118-134.
[6] LI Y,LIU Z D,ZHANG H J. Review on ensemble algorithms for imbalanced data classification [J]. Application Research of Computers,2014,5:13-17.
[7] TURNEY P. Types of Cost in Inductive Concept Learning [J]. arXiv:0212034,2002.
[8] LI Y X,CHAI Y,HU Y Q,et al. Review of imbalanced data classification methods [J]. Control and Decision,2019,34(4):4-19.
[9] BADRAN M F,SAHAR N M,SARI S,et al. Intrusion-Detection System Based on Hybrid Models:Review Paper [C]// IOP Conference Series:Materials Science and Engineering. 2020.
[10] PING R,ZHOU S S,LI D. Cost sensitive random forest classification algorithm for highly unbalanced data [J]. Pattern Recognition and Artificial Intelligence,2020,201(3):62-70.
[11] JING X Y,ZHANG X Y,ZHU X K,et al. Multiset Feature Learning for Highly Imbalanced Data Classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,43(1):139-156.
[12] TAO X M,LI Q,GUO W,et al. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification [J]. Information Sciences,2019,487:31-56.
[13] GALAR M. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches [J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews,2012,42(4):463-484.
[14] GARCIA S,ZHANG Z L,ALTALHI A,et al. Dynamic ensemble selection for multi-class imbalanced datasets [J]. Information Sciences,2018,445:22-37.
[15] CHEN Q W,WANG W,MA D,et al. Class-imbalance credit scoring using Ext-GBDT ensemble [J]. Application Research of Computers,2018,35(2):421-427.
[16] TAO X M,CHEN W,LI X,et al. The ensemble of density-sensitive SVDD classifier based on maximum soft margin for imbalanced datasets [J/OL]. Knowledge-Based Systems, 2021, 219(7). https://www. sciencedirect. com/science/article/abs/pii/S095070512100160X.
[17] ZHANG Z,QIU J X,DAI W. A New Improved Boosting for Imbalanced Data Classification [C]// IOP Conference Series Materials Science and Engineering. 2019.
[18] SHI H T,WANG H R,HUANG Y X,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.
[19] DING H,LIU K,CHEN X Z,et al. Optimized Segmentation Based on the Weighted Aggregation Method for Loess Bank Gully Mapping [J]. Remote Sensing,2020,12(5):793-813.
[20] THABTAH F,HAMMOUD S,KAMALOV F,et al. Data imbalance in classification:Experimental evaluation [J]. Information Sciences,2020,513:429-441.
[21] ABAD Z S H,MASLOVE D M,LEE J. Predicting Discharge Destination of Critically Ill Patients Using Machine Learning [J]. IEEE Journal of Biomedical Health Informatics, 2021, 25(3):827-837.
[22] CHANG Y C,CHANG K H,WU G J. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions [J]. Applied Soft Computing,2018,73:914-920.
[23] CHEN W B,FU K,ZUO J W,et al. Radar emitter classification for large data set based on weighted-xgboost [J]. IET Radar Sonar and Navigation,2017,11(8):1203-1207.
[24] ZOU S H,SUN H Z,XU G S,et al. Ensemble Strategy for Insider Threat Detection from User Activity Logs [J]. CMC-Computers Materials & Continua,2020,65(2):1321-1334.
[25] SANER C B,KESICI M,YASLAN Y,et al. Improving the Performance of Transient Stability Prediction using Resampling Methods [C]// Proceedings of the 2019 11th International Conference on Electrical and Electronics Engineering (ICEEE). Bursa:IEEE,2019:146-150.
[26] CHEN T,GUESTRIN C. XGBoost: A Scalable Tree Boosting System [M]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco:Association for Computing Machinery,2016:785-794.
[27] ZHOU Z H,LIU X Y. On multi-class cost-sensitive learning [J]. Computational Intelligence,2010,26(3):232-257.
[28] WAN J W,YANG M. Survey on Cost - sensitive Learning Method [J]. Journal of Software,2020,31(1):113-136.
[29] NASARIAN E,ABDAR M,FAHAMI M A,et al. Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach [J]. Pattern Recognition Letters,2020,133:33-40.
[30] WANG C,DENG C Y,WANG S Z. Imbalance-XGBoost:leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost [J]. Pattern Recognition Letters,2020,136:190-197.
[31] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal Loss for Dense Object Detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(2):318-327.
[32] TRAN G S,NGHIEM T P,NGUYEN V T,et al. Improving Accuracy of Lung Nodule Classification Using Deep Learning with Focal Loss [J/OL]. Journal of Healthcare Engineering, 2019. https://www. hindawi. com/journals/jhe/2019/5156416/.
[33] BERGSTRA J,BENGIO Y. Random Search for Hyper-Parameter Optimization [J]. Journal of Machine Learning Research, 2012,13:281-305.



LI Jing-tai, born in 1998, postgraduate. His main research interests include machine learning and steganalysis.

WANG Xiao-dan, born in 1966, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include machine learning and intelligent information processing.

(责任编辑:喻藜)