文章编号：2095-2163(2023)07-0007-08 中图分类号：TP181 文献标志码：A

HSMOTE-AdaBoost：改进混合边界重采样集成分类算法

李 静，刘 姜，倪 枫，李笑语
（上海理工大学 管理学院，上海 200093）

摘 要：处理类不平衡问题时，已有的采样方法存在易受噪声影响和忽略边界样本的问题，尤其是忽略多数类样本的类内差异，位于边界的样本实例非常容易被错分，而这些样本对划分决策边界具有重要作用。将 SMOTE 过采样和 RUS 随机欠采样方法结合并进行改进，提出混合边界重采样算法（HSMOTE-AdaBoost）。HSMOTE-AdaBoost 算法首先对少数类运用 SMOTE 过采样，提高数据的平衡度；再使用 K 近邻算法清除噪声和采样方法产生的重叠实例；同时，基于与少数类样本的平均欧氏距离识别并保留边界多数类样本，然后对剩余的数据进行随机欠采样；最后，利用 AdaBoost 算法的优势，对平衡后的数据集进行多次迭代训练得到最终的分类模型。仿真实验结果表明，与传统的 SMOTE-Boost、RUS-Boost、PC-Boost 及改进后的算法 KSMOTE-AdaBoost 相比，该分类模型在不平衡数据集上的所有性能指标 $F-measure$，$G-mean$，AUC 值分别最高提升了 22.97%，13.88%和 10.03%，具有更优的分类效果。

关键词：类不平衡；SMOTE 过采样；AdaBoost 算法；噪声样本；边界样本

HSMOTE-AdaBoost：An integrated classification algorithm based on improved mixed boundary resampling

LI Jing，LIU Jiang，NI Feng，LI Xiaoyu
（Business School，University of Shanghai for Science and Technology，Shanghai 200093，China）

【Abstract】 When dealing with the class imbalance problem，existing sampling methods have the issues of being susceptible to noise and ignoring boundary samples，especially majority class boundary samples，making the boundary sample instances，which play an important role in determining the decision boundary，easily be misclassified. By improving the combination of SMOTE oversampling and random under sampling（RUS），a hybrid boundary resampling algorithm（HSMOTE-AdaBoost）is proposed. The HSMOTE-AdaBoost algorithm firstly performs SMOTE oversampling on the minority samples to improve data balance and effectiveness. Then，the paper uses the K nearest neighbor algorithm to remove noise and overlapping instances generated by the sampling method. Meanwhile，the paper recognizes and retains the boundary majority class samples based on the average Euclidean distance to the minority samples. After that，the remaining data is randomly undersampled. Finally，by making use of the advantages of AdaBoost algorithm ，the balanced dataset is trained for multiple iterations to obtain the final classification model. The experimental results show that，comparing with the traditional SMOTE-Boost，RUS-Boost，PC-Boost and the improved algorithm KSMOTE-AdaBoost，the increase of the $F-measure$，$G-mean$ and AUC values of HSMOTE-AdaBoost could reach 22.97%，13.88% and 10.03% respectively，implying a better performance of HSMOTE-AdaBoost.

【Key words】 class imbalance；SMOTE oversampling；AdaBoost algorithm；noise sample；boundary sample

0 引 言

在实际生活中，大部分的数据集都是不平衡的，即某些类会比其他类具有更多的实例，在这种情况下少数类的信息得不到充分的利用[1]。类不平衡问题给标准的分类学习算法带来了巨大的挑战，在不平衡的数据集上，大多数分类器倾向于对少数实例进行错误分类，而不是对多数实例进行错误分类[2-4]。但在现实生活中，少数类的错分代价会远高于多数类。类不平衡问题普遍存在于许多领域，

基金项目：国家自然科学基金（11701370）；上海市“系统科学”高峰学科建设项目。

作者简介：李 静（1998-），女，硕士研究生，主要研究方向：机器学习、数据挖掘；刘 姜（1983-），女，博士，副教授，硕士生导师，主要研究方向：符号计算、机器学习；倪 枫（1982-），男，博士，副教授，硕士生导师，主要研究方向：系统分析与集成；李笑语（2000-），女，本科生，主要研究方向：机器学习。

通讯作者：刘 姜 Email：jliu113@126.com

收稿日期：2022-07-27

如网络入侵检测[5]、情感分析[6-7]、欺诈检测[8-9]、医疗疾病诊断检测[10]和故障诊断[11]等领域。大多数标准分类算法往往表现出对多数类的偏倚,因此类的不平衡性往往会损害标准分类器的性能,尤其是对少数类的分类性能[12-13]。

目前,对于不平衡数据集分类问题的解决方案可以分为:数据级、算法级和集成层面[14-15]。其中,数据级解决方案包括许多不同形式的重采样技术,如随机过采样(ROS)和随机欠采样(RUS)[16-17]。其中,具有代表性的有 Chawla 等学者[18]提出了一新型过采样方法,SMOTE 在少数类样本与其 k 近邻的连线上随机生成新的样本。He 等学者[19]开发了一种自适应合成采样(ADASYN),使用密度分布来计算出每个少数样本需要合成的新样本数量。此外,还有几种混合的采样技术,其中一些方法就是将过采样与数据清理技术相结合,以减少过采样方法引入的重叠。典型的例子是 SMOTE-Tomek[20]和 SMOTE-ENN[21]。这些数据级方法虽然简单易用,但采样方法难以修改数据的分布和偏好相关联,一方面,欠采样会使数据集中一些有价值的数据信息未能得到利用;另一方面,过采样会生成多余数据,尤其是产生噪声数据。算法级的解决方案旨在开发新的算法或修改现有的算法,加强对少数类分类算法的学习。研究中,最受欢迎的分支是代价敏感学习,例如,Ting[22]提出了一种实例加权方法来裁剪代价敏感树以提高不平衡数据集的分类效果。Zhang 等学者[23]通过对目标函数中的多数类和少数类设置不同的代价,提出了一种代价敏感神经网络算法,使得少数类样本尽可能地被识别。其他基于原有算法的改进方案通常是对现有分类算法的改进,文献[24-26]通过引入权重参数来调整分类决策函数,使其向少数类样本偏倚。然而算法的应用具有特定的情形,大多数的分类模型一旦被确定,则不会动态地调整相应的模型结构及其参数,使得这些算法级改进方法无法动态地学习新增的样本。

集成层面解决方案就是将多个弱分类模型通过一定方式进行组合,得到一个新的、泛化性能更好的强分类器。Freund 等学者[27]提出的自适应增强(Adaptive Boosting, AdaBoost)算法是 Boosting 族中的代表算法。与其他分类器相比,AdaBoost 能够有效地避免过拟合。文献[28-29]分别将 AdaBoost 与过采样和欠采样结合,提出 SMOTEBoost 和 RUSBoost 算法。SMOTEBoost 运用 SMOTE 进行少数类样本的合成,经过多次迭代得到最终的强分类器。但 SMOTEBoost 算法在训练过程中生成的噪声数据会对分类性能产生影响。RUSBoost 则采用欠采样方法,先随机删除一些多数类样本,随后使用删除后的数据构造弱分类器。文献[30]提出一种新的学习算法 PCBoost,该算法考虑了属性特征,对少数类进行随机过采样,接着使用信息增益率来构造弱分类器,训练中错误分类的样本会被删除。

上述研究中,虽然有不少基于采样方法与 AdaBoost 进行结合,但在采样方法上还需做进一步探讨。由于集成学习在分类性能上的优越性以及过采样方法使用的普遍性,本文重点关注采样方法与集成学习结合构建分类器。现有的过采样方法容易引入影响分类性能的噪声样本,而欠采样则会丢失有用的多数类样本的特征信息,尤其是位于分类边界的多数类样本。

针对以上数据级改进方案中采样方法存在易受噪声影响和丢失多数类样本特征信息的问题,本文考虑了多数类样本的类内差异,并保留边界多数类样本;对少数类和多数类采取融合 SMOTE 过采样和 RUS 欠采样的混合采样策略,并结合集成技术 AdaBoost,提出了 HSMOTE-AdaBoost 算法。首先,针对少数类样本中数据缺少和噪声干扰的问题,引入经典的 SMOTE 采样算法和 K 近邻噪声清除算法。其次,由于现有的大多数算法尚未考虑到多数类样本之间的类内差异,本文基于平均欧式距离识别并保留边界多数类样本,再对剩下的样本进行随机欠采样。最后,运用以 J48 为基分类器的 AdaBoost 算法进行分类。结果表明,本文所提出的 HSMOTE-AdaBoost 算法与传统的 SMOTE-Boost, RUS-Boost, PC-Boost 算法及改进后的算法 KSMOTE-AdaBoost[31]相比,具有更好的分类效果。

1 理论基础

1.1 SMOTE 算法

Chawla 等学者[18]于 2002 年提出了少数类样本合成技术,即 SMOTE。该算法的流程步骤如下:

原始训练样本为 Tra,少数类样本是 P,多数类样本是 $NP = \{p_1, p_2, \cdots, p_{pnum}\}$,$N = \{n_1, n_2, \cdots, n_{nmum}\}$,这里 $pnum$,$nmum$ 分别表示少数类样本个数和多数类样本个数。

(1)计算少数类中的每个样本点 p_i 与所有的少数类样本 P 的欧式距离,得到样本点的 k 近邻。

(2)依据样本的采样倍率 U,在 k 近邻之间进行线性插值。

(3)合成新的少数类样本：

$$synthetic_j = p_i + r_j \times d_j \tag{1}$$

其中，d_j 是少数类样本点与其近邻的距离；r_j 是介于 0 到 1 之间的随机数。

(4)新合成的少数类样本与原始训练样本 Tra 进行合并，得到新的训练样本。

1.2 AdaBoost 算法

AdaBoost 算法是经典的 Boosting 算法，通过迭代的方式不断提高被误判样本的权值，从而进行样本权值的更新，分类器在下一次分类会把重心放在那些被误分的样本上，以此来达到正确分类所有样本的目的。算法的训练过程如下：

输入　$Tra = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ 为训练数据集，X_i 表示样本实例，Y_i 为类标签集合，$Y_i \in \{+1, -1\}, i = 1, 2, \cdots, n$；迭代循环次数为 M

输出　最终分类器 $G(X)$

1：训练数据的权值分布初始化：$D_1 = (w_{1,1}, \cdots, w_{1,i}, \cdots, w_{1,n})$，$w_{1,i} = \dfrac{1}{n}, i = 1, 2, \cdots, n$

2：for $m = 1$ to M：

2.1：使用具有权值分布 D_m 的训练数据集进行学习，得到基本分类器 $G_m(X)$

2.2：算出 $G_m(X)$ 在训练数据集上的分类误差率：

$$e_m = \sum_{i=1}^{n} w_{mi} I(G_m(X_i) \neq Y_i) \tag{2}$$

这里，I 为指示函数：

2.3：计算 $G_m(X)$ 的系数：

$$\alpha_m = \frac{1}{2} \ln \frac{1 - e_m}{e_m} \tag{3}$$

2.4：更新训练数据集的权值分布：

$$D_{m+1} = (w_{m+1,1}, \cdots, w_{m+1,i}, \cdots, w_{m+1,n}) \tag{4}$$

$$w_{m+1,i} = \frac{w_{m,i}}{Z_m} \exp(-\alpha_m Y_i G_m(X_i)) \tag{5}$$

其中，Z_m 为规范化因子，计算公式为：

$$Z_m = \sum_{i=1}^{n} w_{m,i} \exp(-\alpha_m Y_i G_m(X_i)) \tag{6}$$

3：构建基本分类器的线性组合：

$$f(X) = \sum_{m=1}^{M} \alpha_m G_m(X) \tag{7}$$

4：得到最终的分类器：

$$G(X) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(X)\right) \tag{8}$$

2 HSMOTE-AdaBoost 算法

针对不平衡数据的二分类问题，过采样会增加多余数据引入噪声样本，而欠采样会丢失一些有用信息，这 2 个问题极大地影响了算法的分类性能。HSMOTE-AdaBoost 算法考虑了边界多数类样本特征信息的价值，将 SMOTE 算法和 RUS 算法进行了结合，并针对 RUS 算法中存在随机删除边界多数类样本的问题进行了改进。位于 2 个类边界上的实例是该分类样本的必要实例，在确定决策边界时至关重要，将其删除会降低分类器的性能。因此，本文提出的 HSMOTE-AdaBoost 算法考虑了边界多数类样本，并试图保持必要的多数类实例，分类模型图如图 1 所示。首先，使用 SMOTE 过采样合成少数类实例，以平衡数据集；接着，删除原始数据中的一些噪声数据，提高合成样本的质量；然后，识别多数类的边界实例，并将其添加到输出数据集中。同时，对剩余的数据进行随机欠采样。最后，使用 AdaBoost 集成算法生成强分类器，实验结果表明该分类器的性能得到了有效的提升。

2.1 算法 HSMOTE-AdaBoost 的训练过程

(1)合成少数类样本。将合成少数类过采样(SMOTE)预处理技术应用到数据集 T_{imbal} 中，在少数类中引入人工实例，生成数据集 $\widehat{T}_{imbal}$。具体步骤详述如下。

① 原始训练样本是 T_{imbal}，少数类样本为 P，多数类是 N，$P = (p_1, p_2, \cdots, p_n)$，$N = (n_1, n_2, \cdots, n_m)$，$n, m$ 分别表示少数类样本个数及多数类样本个数。计算少数类样本点 p_j 与少数类 P 的欧式距离，得到该样本点的 k 近邻。

② 依据采样倍率 U 在 k 近邻中进行线性插值。

③ 合成新的少数类样本：

$$synthetic = p_j + r_i \times d_i \tag{9}$$

其中，d_i 表示少数类样本点与其近邻的距离，r_i 是介于 0 到 1 之间的随机数。

④ 新合成的少数类样本和原始训练样本进行合并，从而得到新的训练样本 $\widehat{T}_{imbal}$。

(2)识别和删除噪声。若少数类样本点在 k 近邻中有 k' 个多数类样本，显然 $0 \leqslant k' \leqslant k$，若 $k' \leqslant \dfrac{2}{3}k$，则为噪声。

(3)识别多数类的边界样本 N_{border}，并将其放入输出数据集 T_{bal} 中。具体步骤分述如下。

① 算出各多数类样本点 $n_i (i = 1, 2, \cdots\cdots, m)$ 到少数类样本点 $p_j (j = 1, 2, \cdots, n)$ 的距离之和为 $\sum_{j=1}^{n} d(n_i, p_j)$。

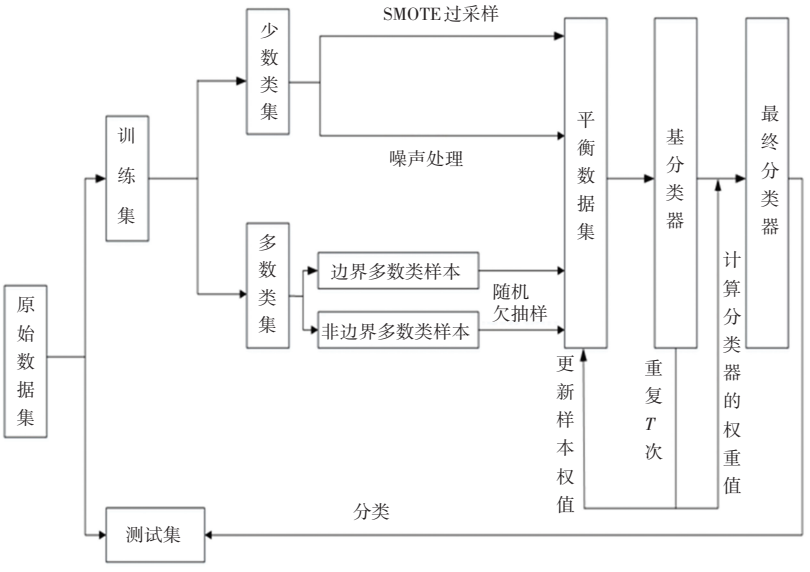

图 1 分类模型图

Fig. 1 Classification model diagram

② 求平均距离：

$$\bar{d} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(n_i, p_j)}{m} \tag{10}$$

③ 当多数类样本 n_i 满足 $\sum_{j=1}^{n} d(n_i, p_j) < \bar{d}$ 时，将其划分到 T_{bal} 中。

(4)对数据集 E 应用随机欠采样，将随机选取的样本推送到输出数据集 T_{bal}，并推得：

$$E = \widehat{T}_{imbal} - N_{border} \tag{11}$$

(5)将处理后的平衡数据集作为模型的输入，使用 AdaBoost 集成技术得到最终的分类结果。

2.2 算法 HSMOTE-AdaBoost 的详细步骤

输入 D 为数据集，K 为过采样算法中选择的最近邻个数，M 为迭代的次数

输出 集成分类器

1：将数据集 D 分成 10 份，其中 2 份为测试集 *TestData*，剩下即训练集 *TrainData*

2：运用 SMOTE 过采样算法生成新的少数类样本

3：用 k 近邻算法得到少数类样本的最近邻集合，依据判别条件对噪声样本进行识别及删除。判别条件为：若少数类样本中 k 近邻超过 2/3 的样本是多数类，则为噪声样本

4：基于所有多数类样本点对各个少数类样本点的平均欧式距离识别边界多数类样本，并将其单独放在 *NewTrainData* 中。识别条件为：若该多数类样本点到所有少数类样本点的欧式距离之和小于平均距离，则该样本点为边界多数类样本点

5：对剩余的数据进行随机欠抽样得到 *NewTrainData*

6：赋予 *NewTrainData* 中每个实例相同的权重。

7：for m = 1 to M

8：　　根据实例的权值有放回地抽样得到 D_m

9：　　由 D_m 得到基分类器 $G_m(X)$

10：　　使用 $G_m(X)$ 对 *NewTrainData* 中的实例进行分类，根据式(12)计算 $G_m(X)$ 的错误率 e_m：

$$e_m = \sum_{G_m(X_i) \neq Y_i} w_{m,i} \tag{12}$$

其中，$w_{m,i}$ 表示第 m 轮中第 i 个实例 X_i 的权值；$G_m(X_i) \neq Y_i$ 表示实例 X_i 被错分；Y_i 为类标签。

11：　　if $e_m > 0.5$ then 转步骤 7

12：　　end if

13：　　for 对正确分类的每个实例执行如下步骤：

14：　　更新权值：实例的权值乘以 $\frac{e_m}{(1 - e_m)}$

15：　　归一化所有实例的权值

16：　　end for

17：　　将每个类的权值赋予 0，对 *TestData* 的所有实例执行以下步骤

18：　　for i = 1 to M：

19：　　根据式(13)计算 $G_i(X)$ 对测试集中实例 X_i 的权值 W_i：

$$W_i = -\log\frac{e_i}{1 - e_i} \tag{13}$$

20：　　将 W_i 加到 $G_i(X)$ 所预测的类上，若 X_i 被正确分类，则其值为 0，否则为 1。

21：　　end for

22：　　返回权值最大的类，即为 X_i 的类别

23：end for

3　仿真实验

3.1　实验数据集

实验使用来自 KEEL 公开数据集的 6 组数据集，考虑了数据集的样本数、特征数及不平衡率。实验之前对数据集进行预处理，将训练集和测试集归一化到[0,1]区间，表 1 展示了实验数据集的基本信息。

表 1　实验数据集

Tab. 1　Experimental dataset

序号	数据集	样本总数	正类样本数	负类样本数	特征数	不平衡率
1	ecoli2	336	52	284	7	5.46
2	yeast1	1 484	429	1 055	8	2.46
3	glass1	214	76	138	9	1.82
4	glass6	214	29	185	9	6.38
5	ecoli3	336	35	301	7	8.60
6	ecoli1	336	77	259	7	3.36

3.2　性能评价指标

在对数据集进行二分类时，不能简单地采用整体精确度的高低来评价分类器性能的优劣。因为即使分类器对少数类样本的识别完全错误，总体的精确度也会比较高，所以仅靠这个单一评价指标并没有参考价值。为了全面体现分类性能，本文采用了综合评价指标 $F-measure$，$G-mean$ 和 AUC，$F-measure$，$G-mean$ 和 AUC 取值均在 0 到 1 之间，且值越大，说明分类器的分类性能越好。在介绍这些指标前，先引入混淆矩阵。混淆矩阵[32]可以将预测分类结果和实际分类结果以矩阵的形式直观地表示出来。混淆矩阵见表 2。

表 2　混淆矩阵

Tab. 2　Confusion matrix

	预测为正类	预测为负类
实际为正类	真正类(TP)	假负类(FN)
实际为负类	假正类(FP)	真负类(TN)

根据表 2 可得到以下评价指标。

(1)召回率 ($Recall$)。指真少数类占所有少数类的比例。可由如下公式来求值：

$$Recall = TP/(TP + FN) \tag{14}$$

(2) 精确率($Precision$)。指真少数类占所有被预测为少数类的比例。可由如下公式来求值：

$$Precision = TP/(TP + FP) \tag{15}$$

(3) $F-measure$[33]，又称 $F-Score$，兼顾精确度和召回率。可由如下公式来求值：

$$F-measure = \frac{(\alpha^2+1)\,Recall * Precision}{\alpha^2 Recall + Precision} \tag{16}$$

其中，α 是取值为 1 的比例系数。

(4) $G-mean$[34]。是一种有效衡量不平衡数据分类方法的指标，只有正类、负类两者召回率都高时，$G-mean$ 值才会高，表明分类器的分类性能较优。可由如下公式来求值：

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{17}$$

(5) AUC[35]。ROC 曲线以假正率 ($FPrate$) 和真正率($TPrate$) 为轴，以可视化的方式展现正确分类的收益和误分类的代价之间的关联。ROC 曲线下方的面积称为 AUC (Area Under Curve)，AUC 可以量化分类器的预测性能，AUC 的取值范围为 0 到 1，AUC 值越高，模型的预测性能越好。

3.3　实验设置

为了验证本文所提出的算法的优越性，将本文算法与传统的 SMOTE-Boost，RUS-Boost，PC-Boost 及改进后的 KSMOTE-AdaBoost 算法进行了比较。本文使用 Python 语言，Spyder 开发环境对各种算法进行仿真实验。与以往文献[36]一致，SMOTE 算法的 k 近邻数设为 5 时所合成的少数类样本质量较高。实验选用 J48 决策树作为基分类器，调用 Weka 软件中的 C4.5 函数包实现 J48 分类器。实验中将数据集的 20% 作为测试集，80% 作为训练集，采用 10 次五折交叉验证的平均值作为最终的评价结果。

3.4　实验结果分析

图 2~图 4 及表 3~表 5 列出了使用本文算法和其他 4 种不同采样方法与集成学习技术结合算法在 6 个不同数据集上的 $F-measure$、$G-mean$ 值及 AUC 的比较结果和具体数值，其中加粗部分为在该数据集上的最优结果。从图 2~图 4 可以看出，由本文算法所得到的 $F-measure$、$G-mean$ 及 AUC 值总体上优于其他 4 种对比算法。从表 3~表 5 可以得出本文的分类模型在 $F-measure$、$G-mean$ 及 AUC 值上均得到了提升，尤其是对比 RUS-Boost 算法。在数据集 ecoli2、yeast1、glass6、ecoli3 上，其分类评估指标均优于其他对比算法，$F-measure$ 的提升值高达 22.97%，$G-mean$ 的提升值为 13.88%，AUC 的最高提升值为 10.03%。从表 3 的 $F-measure$ 值可看出，HSMOTE-AdaBoost 算法除了在数据集 ecoli1 略低于 SMOTE-Boost 之外，在其他 5 个数据集上均有最佳表现。从表 4~表 5 得出，在数据集 glass1 上，$G-mean$ 值稍逊于 KSMOTE-AdaBoost 算法，AUC 指标略小于

SMOTE-Boost 算法,但其 *F - measure* 的提升值为 3.99%。原因在于其不平衡比率较小,边界样本的重要性未得到充分体现。虽然在部分数据集中的 *AUC* 值与其他 4 种算法相差不大,但是随着不平衡率的提升,性能提升百分比在不断增大,在不平衡比率最大的数据集 ecoli3 上,相对于其他 4 种算法,性能提升了 10.03%。此外,在部分数据集、如 yeast1,ecoli2 中,使用 HSMOTE-AdaBoost 算法优于 RUS-Boost 算法,这是因为在使用随机欠采样后有可能会删除一些潜在、有价值的多数类样本,而本文所提出的算法有效解决了这个问题,尽可能地保留了必要的实例从而提升了分类性能。

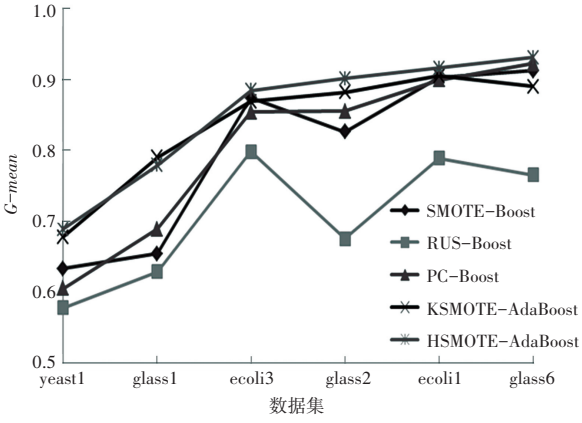

图 3 不同算法的 *G-mean* 对比图

Fig. 3 *G-mean* comparison of different algorithms

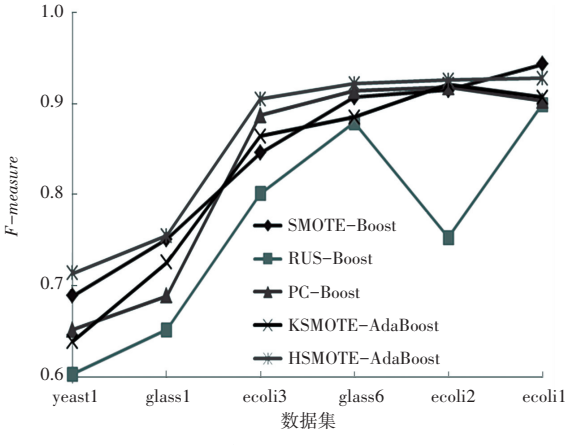

图 2 不同算法的 *F-measure* 对比图

Fig. 2 *F-measure* comparison of different algorithms

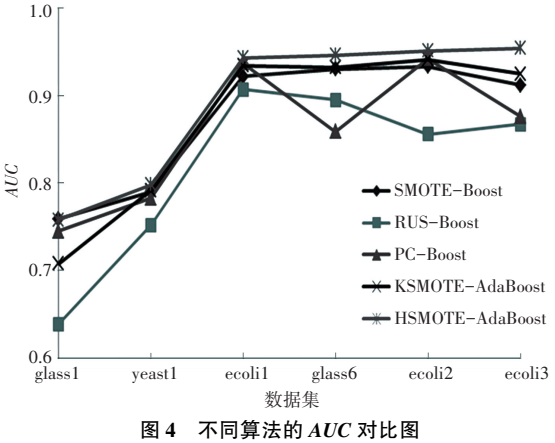

图 4 不同算法的 *AUC* 对比图

Fig. 4 *AUC* comparison of different algorithms

表 3 不同算法的 *F-measure*

Tab. 3 *F-measure* of different algorithms

数据集	SMOTE-Boost	RUS-Boost	PC-Boost	KSMOTE-AdaBoost	HSMOTE-AdaBoost
ecoli2	0.915	0.753	0.918	0.921	**0.926**
yeast1	0.689	0.603	0.652	0.639	**0.714**
glass1	0.751	0.652	0.689	0.726	**0.755**
glass6	0.907	0.879	0.914	0.885	**0.922**
ecoli3	0.846	0.801	0.887	0.864	**0.905**
ecoli1	**0.943**	0.899	0.903	0.907	0.928

表 4 不同算法的 *G-mean*

Tab. 4 *G-mean* of different algorithms

数据集	SMOTE-Boost	RUS-Boost	PC-Boost	KSMOTE-AdaBoost	HSMOTE-AdaBoost
ecoli2	0.826	0.675	0.855	0.882	**0.901**
yeast1	0.633	0.578	0.605	0.678	**0.689**
glass1	0.654	0.629	0.689	**0.790**	0.779
glass6	0.912	0.765	0.922	0.890	**0.931**
ecoli3	0.874	0.798	0.854	0.869	**0.884**
ecoli1	0.902	0.789	0.899	0.905	**0.916**

表 5　不同算法的 *AUC*

Tab. 5　*AUC* of different algorithms

数据集	SMOTE-Boost	RUS-Boost	PC-Boost	KSMOTE-AdaBoost	HSMOTE-AdaBoost
ecoli2	0.933	0.856	0.942	0.941	**0.951**
yeast1	0.790	0.752	0.783	0.792	**0.798**
glass1	**0.759**	0.639	0.745	0.708	0.758
glass6	0.930	0.895	0.859	0.932	**0.946**
ecoli3	0.912	0.867	0.876	0.925	**0.954**
ecoli1	0.922	0.907	0.937	0.934	**0.943**

4　结束语

针对二分类问题中不平衡数据集造成分类器分类性能低下的问题，本文考虑了多数类样本的类内差异，并保留了必要的边界多数类样本，将 SMOTE 过采样与随机欠采样组合运用及改进，并与 AdaBoost 相结合提出了一种解决类不平衡问题的 HSMOTE-AdaBoost 算法。与传统的 SMOTE-Boost、RUS-Boost、PC-Boost 及改进后的算法 KSMOTE-AdaBoost 相比，实验表明该算法训练的分类器能有效地处理类不平衡问题，分类性能更优。

此外，本文的改进算法也存在进一步研究的价值，接下来可以结合其他经典的集成算法研究其分类性能。其次，从实验结果可以发现，在类不平衡比率越大的数据集上的分类效果越好，亟需从理论上对该结果做进一步的探讨和验证。最后，本文所提出的算法对二分类问题的分类性能提升比较明显，然而在现实生活中的大多数数据都是多类别的，未来将着重研究如何提高多类别分类问题的分类性能。

参考文献

[1] WU Jianxin, BRUBAKER S C, MULLIN M D, et al. Fast asymmetric learning for cascade face detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(3): 369-382.

[2] GUO Haixiang, LI Yijing, JENNIFER S, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

[3] JOHNSON J M, KHOSHGOFTAAR T M. Survey on deep learning with class imbalance [J]. Journal of Big Data, 2019, 6 (1): 27.

[4] SÁEZ J A, LUENGO J, STEFANOWSKI J. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering [J]. Information Sciences: An International Journal, 2015, 291: 184-203.

[5] ZHOU Chenfeng, LECKIE C, KARUNASEKERA S. A survey of coordinated attacks and collaborative intrusion detection [J]. Computers & Security, 2009, 29(1): 124-140.

[6] NIR O, GILAD K, BRACHA S, et al. Sentiment analysis in transcribed utterances[M]// CAO T, LIM E, ZHOU Z H, Ho, et al. Advances in Knowledge Discovery and Data Mining. PAKDD 2015. Lecture Notes in Computer Science ( ). Cham: Springer, 2015, 9078: 27-38.

[7] GOPALAKRISHNAN V, RAMASWAMY C. Sentiment learning from imbalanced dataset: An ensemble based method [J]. International Journal of Artificial Intelligence, 2014, 12(2): 75-87.

[8] SIDNEY T, KOH Y S, GILLIAN D. Detecting online auction shilling frauds using supervised learning[J]. Expert Systems with Applications, 2014, 41(6): 3027-3040.

[9] WEI Wei, LI Jinjiu, CAO Longbing, et al. Effective detection of sophisticatedonline banking fraud on extremely imbalanced data [J]. World Wide Web, 2013,16(4): 449-475.

[10] AKRAM V, SAEED J. C-PUGP: A cluster-based positive unlabeled learning method for disease gene prediction and prioritization[J]. Computational Biology and Chemistry, 2018, 76: 23-31.

[11] KIM J H. Time frequency image and artificial neural network based classification of impact noise for machine fault diagnosis [J]. International Journal of Precision Engineering and Manufacturing, 2018, 19(6): 821-827.

[12] JL A, QZ A, QW A, et al. A novel oversampling technique for classimbalanced learning based on SMOTE and natural neighbors [J]. Information Sciences, 2021, 565: 438-455.

[13] YAN Yuanting, LIU Ruiqing, DING Zihan, et al. A Parameter-free Cleaning Method for SMOTE in Imbalanced Classification [J]. IEEE Access, 2019: 23537- 23548.

[14] PAULA B, LUIS T, RIBEIRO R. A survey of predictive modeling on imbalanced domains [J]. ACM Computing Surveys, 2016, 49 (2): 351-400.

[15] 李艳霞，柴毅，胡友强，等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34 (04): 673-688.

[16] 孟东霞，李玉鑑. 基于特征边界欠采样的不平衡数据处理方法 [J]. 统计与决策, 2021, 37 (11): 30-33.

[17] AREFEEN M A, NIMI S T, RAHMAN M S. Neural Network-Based Undersampling Techniques [J]. arXiv preprint arXiv: 1908.16487,2019.

[18] CHAWLA N V, BOWYER K W, HALL L O. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 1076-9757.

[19] HE Haibo, YANG Bai, GARCIA E A. ADASYN: adaptive synthetic sampling approach for imbalanced learning [J]. Proceedings of IEEE International Joint Conference on Neural

Networks, 2008, 12(5): 1322-1328.
[20] KANG Qi, CHEN Xiaoshuang, LI Sisi, et al. A Noise-filtered under-sampling scheme for imbalanced classification[J]. IEEE Transactions on Cybernetics, 2017, 47(12): 4263-4274.
[21] JIANG Kun, JING Lu, XIA Kuiliang. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE[J]. Arabian Journal for Science & Engineering, 2016, 41(8): 3255-3266.
[22] TING K M. An instance-weighting method to induce cost-sensitive trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 659-665.
[23] ZHANG Chong, TAN K C, LI Haizhou, et al. A cost-sensitive deep belief network for imbalanced classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 30(1): 109-122.
[24] IMAM T, TING K M, KAMRUZZAMAN J. Z-SVM: An svm for improved classification of imbalanced data[M]// SATTAR A, KANG B H. Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science( ). Berlin/Heidelberg: Springer, 2006, 4304: 264-273.
[25] 王伟,薛安荣,刘峰. 改进的 SVM 解决背景知识数据中的类不平衡[J]. 计算机应用研究,2011,28(08):2902-2904,2908.
[26] 杨扬,李善平. 基于实例重要性的 SVM 解不平衡数据分类[J]. 模式识别与人工智能,2009,22(06):913-918.
[27] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to Boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139.
[28] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: improving prediction of the minority class in boosting[J]. Lecture Notes in Computer Science, 2003, 2838(1): 107-119.
[29] CHRIS S, KHOAHFOTAAR T M, JASON V H, 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.
[30] LIU C L, HSIEH P Y. Model-based synthetic sampling for imbalanced data[J]. IEEE Transactions on Knowledge and Da-ta Engineering,2020,32(8) : 1543 - 1556.
[31] 王忠震,黄勃,方志军,等. 改进 SMOTE 的不平衡数据集成分类算法[J]. 计算机应用,2019, 39(09): 2591-2596.
[32] DENG Xinyang, LIU Qi, DENG Yong, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem[J]. Information Sciences, 2016, 340: 250-261.
[33] LIPTON Z C, ELKAN C, NARYANASWAMY B. Optimal thresholding of classifiers to maximize F1 measure[J]. Machine Learning Knowledge Discovery Databases, 2014, 8725: 225-239.
[34] KUBAT M, HOLTE R, MATWIN S. Learning when negative examples abound[J]. Lecture Notes in Computer Science, 1997, 1224(1): 146-153.
[35] FLACH P A. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics[C]// Proceedings of the Twentieth International Conference on International Conference on Machine Learning(ICML′03). Washington, DC USA :ACM, 2003:194-201.
[36] 石洪波,陈雨文,陈鑫. SMOTE 过采样及其改进算法研究综述[J]. 智能系统学报,2019, 14(06): 1073-1083.

(上接第 6 页)

参考文献

[1] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of t-he International Conference on Learning Representations. Scottsdale, USA: UMASS, 2013:1-12.
[2] PENNINGTON J, SOCHER R, MANNING C, et al. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference of Empirical Methods in NaturalLanguage Processing. Stroudsburg, PA: ACL, 2014: 1532-1543.
[3] 袁磊. 基于改进 CHI 特征选择的情感文本分类研究[J]. 传感器与微系统, 2017, 36(05): 47-51.
[4] 宋呈祥, 陈秀宏, 牛强. 文本分类中基于 CHI 改进的特征选择方法[J]. 传感器与微系统, 2019, 38(02): 37-40.
[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Minneapolis, Minnesota: ACL, 2019: 4171-4186.
[6] YANG Zhilin, DAI Zihang, YANG Yiming, et al. XLNet: Generalized autoregressive pretraining for language understanding[C]// Neural Information Processing Systems. Canada: NIPS Foundation,2019: 5754-5764.
[7] ADHIKARI A, RAM A, TANG R, et al. Rethinking complex neural network architectures for document classification[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: ACL, 2019: 4046-4051.
[8] 陈立潮, 秦杰, 陆望东, 等. 自注意力机制的短文本分类方法[J]. 计算机工程与设计, 2022, 43(03): 728-734.
[9] 杨青, 张亚文, 朱丽, 等. 基于注意力机制和 BiGRU 融合的文本情感分析[J]. 计算机科学, 2021, 48(11): 307-311.
[10] 孙红, 陈强越. 融合 BERT 词嵌入和注意力机制的中文文本分类[J]. 小型微型计算机系统,2022,43(01):22-26.
[11] 梁淑蓉, 谢晓兰, 陈基漓, 等. 基于 XLNet 的情感分析模型[J]. 科学技术与工程, 2021, 21(17): 7200-7207.
[12] 搜狗实验室. 搜狐新闻数据(SogouCS)[EB/OL]. [2020-01-13]. http://www.sogou.com/labs/resource/cs.php.
[13] CUI Yiming, CHE Wanxiang, LIU Ting, et al. Revisiting pre-trained models for chinese natural la-nguage processing[EB/OL]. [2020]. http://arxiv.org/abs/2004.13922.
[14] 陶志勇, 李小兵, 刘影, 等. 基于双向长短时记忆网络的改进注意力短文本分类方法[J]. 数据分析与知识发现, 2019, 3(12): 21-29.
[15] 滕金保, 孔韦韦, 田乔鑫, 等. 基于 CNN 和 LSTM 的多通道注意力机制文本分类模型[J]. 计算机工程与应用, 2021, 57(23): 154-162.