CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 41 卷第 11 期 2019 年 11 月
Vol. 41, No. 11, Nov. 2019

文章编号:1007-130X(2019)11-2063-08

一种改进的基于欧氏距离的 SDRSMOTE 算法*

李克文[1],林亚林[1],杨耀忠[2]

(1.中国石油大学(华东)计算机与通信工程学院,山东 青岛 266580;
2.中国石化胜利油田分公司信息化管理中心,山东 东营 257022)

摘　要:SMOTE 算法可以扩充少数类样本,提高不平衡数据集中少数类的分类能力,但是它在扩充少数类样本时对于边界样本的选择以及随机数的取值具有盲目性。针对此问题,将传统的 SMOTE 过采样算法进行改进,改进后的过采样算法定义为 SDRSMOTE,该算法综合考虑不平衡数据集中全部样本的分布状况,通过融合支持度 *sd* 和影响因素 *posFac* 来指导少数类样本的合成。在 WEKA 平台上分别使用 SMOTE、SDRSMOTE 算法对所选用的 6 个不平衡数据集进行过采样数据预处理,然后使用决策树、AdaBoost、Bagging 和朴素贝叶斯分类器对预处理后的数据集进行预测,选择 *F-value*、*G-mean* 和 *AUC* 作为分类性能的评价指标,实验表明 SDRSMOTE 算法预处理的不平衡数据集的分类效果更好,证明了该算法的有效性。

关键词:不平衡数据集;分类;边界样本;支持度;影响因素;欧氏距离;SMOTE

中图分类号:TP311　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2019.11.022

An improved SDRSMOTE algorithm based on Euclidean distance

LI Ke-wen[1], LIN Ya-lin[1], YANG Yao-zhong[2]

(1. College of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580;
2. Control Center of Informatization Sinopec Shengli Oil Field, Dongying 257022, China)

Abstract: The SMOTE algorithm can extend the minority samples and improve the classification ability of a few classes in the unbalanced data set. However, it blindly chooses boundary samples and the value of random numbers when extending the minority samples. This paper improves the traditional SMOTE oversampling algorithm, called SDRSMOTE. It takes into account all the unbalanced data sets. The distribution of all the samples, through the introduction of support degree *sd* and the influencing factor *posFac* to guide the synthesis of the minority samples. On the WEKA platform, the SMOTE and SDRSMOTE algorithms are used to preprocess the selected six unbalanced data sets and use the decision tree, AdaBoost, Bagging and Naive Bayes classifiers to predict the preprocessed datasets. The data set is classified, and *F-value*, *G-mean* and *AUC* are selected as evaluation indexes. The experiment shows that the unbalanced datasets preprocessed by the improved SDRSMOTE algorithm have better classification effect, which proves the effectiveness of the algorithm.

Key words: unbalanced data set; classification; boundary sample; support degree; influencing factor; Euclidean distance; SMOTE

* **收稿日期**:2018-12-21;**修回日期**:2019-04-23
通信作者:林亚林(664103127@qq.com)
通信地址:266580 山东省青岛市中国石油大学(华东)计算机与通信工程学院
Address: College of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580, Shandong, P. R. China

1 引言

近年来,机器学习得到了前所未有的发展,已经有决策树、朴素贝叶斯等许多相对成熟的分类算法。但是,这些传统的分类算法往往假设数据集是平衡的,即各个类别的样本数量基本相同。而在许多实际应用中,如软件缺陷预测、医疗诊断[1,2],少数类样本的数量远远少于其他类别的样本,这些场景中的数据集被称作不平衡数据集。在不平衡数据集中,本文定义样本数量少的类为少数类,样本数量多的类为多数类。传统的分类器对不平衡数据集进行分类时,由于多数类样本更容易学习,导致分类结果偏向于多数类,但人们最感兴趣的往往是少数类。例如软件缺陷预测中,几乎所有的数据集都是不平衡的,有缺陷的样本属于少数类,无缺陷的样本属于多数类,在实际应用中为有缺陷样本预测错误所付出的代价是惨痛的。因此,分类不平衡问题逐渐成为机器学习领域的研究热点,尤其是正确识别其中的少数类。

目前,研究人员主要从数据和算法2个层面解决分类不平衡问题。(1)数据层面通过采样法来增加少数类样本或减少多数类样本,从而使数据集中不同类别的样本数量趋于平衡。例如,Chawla等[3]提出了最经典的数据集重构算法合成少数类过采样技术SMOTE(Synthetic Minority Over-sampling TEchnique),该算法通过在同类近邻样本间线性插值来增加样本数量。Giraldo-Forero等[4]提出了基于距离度量的SMOTE类算法。古平等[5]提出了基于错分的过采样算法。Souto等[6]提出了将Tomek links算法与卷积神经网络CNN(Convolutional Neural Networks)算法结合起来的单边选择的欠采样算法。Laurikkala[7]提出了邻域清理NCL(Neighborhood Cleaning Rule)的欠采样算法。Nguyen等[8]提出了双判别器生成对抗网络D2GAN(Dual Discriminator Generative Adversarial Nets)扩充数据样本。(2)算法层面主要通过代价敏感学习和集成学习来提高分类模型的性能,从而改善分类不平衡问题。例如,徐丽丽等[9]通过为各类别分配不同的权重将加权支持向量机模型WSVM(Weighted Support Vector Machine)与模糊聚类结合提出一种新的不平衡数据加权集成学习算法。Nghe等[10]将采样技术与使用支持向量机的代价敏感学习算法进行结合来降低不平衡数据集误分类的成本。

基于此,本文提出一种改进的SMOTE算法(基于样本分布改进的SMOTE算法SDRSMOTE(Support Degree Random Synthetic Minority Over-sampling TEchnique))。在确定边界样本过程中,该算法以少数类样本为中心,距离为半径圈定区域,并统计该区域内的多数类样本个数作为该少数类样本的支持度 sd,根据支持度 sd 对少数类样本进行排序,使得支持度 sd 大的少数类样本优先被选择作为边界样本;在新样本生成过程中,考虑多数类样本分布对少数类样本生成的影响,引入影响因素 $posFac$ 对新样本生成过程中的随机数进行约束。该算法的主要贡献如下:(1)通过引入支持度 sd 确定边界样本的选择顺序,可以缓解传统SMOTE过采样算法中随机选择边界样本的盲目性;(2)通过综合总体样本的分布状况引入影响因素 $posFac$ 进行新样本的合成,使得新样本合成过程中的随机数取值更有针对性;(3)更加合理地扩展少数类样本,使得新数据集是趋于平衡的,提高分类器对于少数类样本的分类能力。

2 相关知识

2.1 不平衡数据分类方法

目前,处理不平衡数据的方法有很多,可以概括为2类,一类是从数据层面进行处理,另一类是从算法层面提高不平衡数据的分类性能。本文是从数据层面对不平衡数据进行处理。

数据层面主要是通过采样算法改变数据集中的样本分布状况,使得多数类与少数类的样本分布大致平衡。采样算法包括欠采样和过采样2种类型,这里仅介绍本文采用的过采样算法。

过采样技术就是使用复制或合成方法增加少数类样本的数量,该算法可以平衡不平衡数据集中的样本分布。经典的过采样算法包括随机过采样ROS(Random-Over-Sampling)和SMOTE。随机过采样是一种直接复制原有少数类样本实现平衡数据集样本分布的过采样算法,该算法容易造成分类模型过拟合[11]。

SMOTE是基于随机过采样的一种改进算法[12],它通过相邻2个少数类样本之间的线性插值来合成新的少数类样本。SMOTE具体算法流程:对于每个少数类样本 $\boldsymbol{x}_i$,计算 $\boldsymbol{x}_i$ 到其他所有少数类样本的欧氏距离,然后搜索 $\boldsymbol{x}_i$ 的 K 个最近邻样本,根据采样倍率 N 从 $\boldsymbol{x}_i$ 的 K 个最近邻样本随机选

择若干个样本(被选中的样本记为 $\boldsymbol{x}_j$)与 $\boldsymbol{x}_i$ 进行合成得到新样本 $\boldsymbol{x}_{new}$,$\boldsymbol{x}_{new}=\boldsymbol{x}_i+rand(0,1)(\boldsymbol{x}_j-\boldsymbol{x}_i)$,其中 $rand(0,1)$ 表示(0,1)内的随机数。例如,少数类样本点 $\boldsymbol{x}_i$ 为(4,7),被选中的 K 近邻样本点 $\boldsymbol{x}_j$ 为(3,5),合成新的少数类样本 $\boldsymbol{x}_{new}=(4,7)+rand(0,1)\times(-1,-2)$,假设随机数 $rand(0,1)$ 取值 0.3,则新样本 $\boldsymbol{x}_{new}$ 为(3.7,6.4)。SMOTE 算法避免了随机过采样算法中造成分类器模型过拟合的问题,但是对于少数类边界样本的选择以及新样本合成中随机数的取值范围没有进行精细的控制,从而导致合成的新样本质量较差。

2.2 不平衡数据分类算法评价指标

在分类问题中,传统的预测模型通常选用准确率(Accuracy)作为分类的评价指标,但是对于不平衡数据集而言,由于多数类样本数量远远多于少数类样本数量,所以用分类正确的样本数占样本总数的比例去衡量分类器的性能是不合适的。因此,本文采用不平衡分类[13]中常用的评价指标 *F-value*、*G-mean* 和 *AUC* 进行评估。首先介绍基本的分类评价指标。

在分类过程中,多分类问题也可以转化为二分类问题[14],因此本文主要讨论不平衡数据集的二分类问题。传统二分类过程中常用的混淆矩阵如表 1 所示,通常我们称少数类样本为正类样本(+),多数类样本为负类样本(−)。

Table 1 Confusion matrix for the two-class problem

表 1 二分类的混淆矩阵

分类	+(预测类)	−(预测类)
+(实际类)	*TP*	*FN*
−(实际类)	*FP*	*TN*

其中,*TP* (True Positive)表示真正例的样本数,*FN* (False Negative)表示假负例的样本数,*FP* (False Positive)表示假正例的样本数,*TN* (True Negative)表示真负例的样本数。

(1) 精确率(*precision*)的计算公式为:

$$precision=\frac{TP}{TP+FP}$$

(2) 召回率(*recall*)的计算公式为:

$$recall=\frac{TP}{TP+FN}$$

(3) *F-value* 值是精确率(*precision*)和召回率(*recall*)的调和均值,计算公式为:

$$F\text{-}value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision}$$

其中参数 β 常取 1,只有当精确率(*precision*)和召回率(*recall*)都高时,*F-value* 值才会高。因此,它可以作为不平衡缺陷数据集分类问题的有效评估标准。

(4) *G-mean* 是正类样本召回率和负类样本召回率的几何均值,计算公式为:

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

其值越大分类性能越好,只有当正类样本和负类样本的分类效果都比较好时,*G-mean* 值才会高。因此,它可以作为不平衡缺陷数据集分类问题的有效评估标准。

(5) ROC(Receiver Operating Characteristic)曲线是描述分类模型真正例率和假正例率关系的二维曲线,该曲线可以比较不同分类器的性能,但不能定量评估分类器的性能。*AUC*[15,16] 是 ROC 曲线与坐标轴所围成的面积,它的取值是 0～1,*AUC* 值越大,预测模型的性能越好。

3 SDRSMOTE 算法

3.1 边界样本的支持度

本文中边界样本支持度 *sd* 定义[17]:如图 1 所示,假设样本空间是二维的,我们以少数类样本为圆心,距离 S_{ave} 为半径圈定一个圆形区域,并统计该圆型区域内的多数类样本个数作为该少数类样本的支持度 *sd*,其中 S_{ave} 的值为所有少数样本与多数类样本的平均欧氏距离。

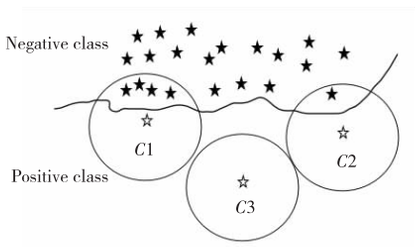

Figure 1 An example of boundary sample support degree

图 1 边界样本支持度示意图

首先,我们计算少数类样本 $\boldsymbol{x}_i$ 和所有多数类样本之间的欧氏距离 S_i :

$$S_i=\sum_{j=1}^{n}\sqrt{\|\boldsymbol{x}_i-\boldsymbol{y}_j\|^2},\quad i=1,2,\cdots,m;j=1,2,\cdots,n \tag{1}$$

其中,少数类样本的数量为 m ,多数类样本的数量为 n ,$\boldsymbol{x}_i$ 表示当前的少数类样本,$\boldsymbol{y}_j$ 表示当前的多数类样本。

然后,计算所有少数类样本和所有多数类样本

之间的距离 S：

$$S=\sum_{i=1}^{m}S_i \tag{2}$$

最后，计算少数类样本和多数类样本的平均欧氏距离 S_{ave}。

$$S_{ave}=\frac{S}{m\times n} \tag{3}$$

以 $\boldsymbol{x}_i$ 为中心，S_{ave} 为半径圈定区域，统计该区域内的多数类样本数量作为少数类样本 $\boldsymbol{x}_i$ 的支持度 sd。

图 1 中实心五角星表示多数类样本，空心五角星表示少数类样本，圆 $C1$ 内包含 4 个多数类样本，即圆 $C1$ 的支持度 sd 是 4，同理可得圆 $C2$ 的支持度 sd 是 1，圆 $C3$ 的支持度 sd 是 0。某一少数类样本的支持度越大，则意味着该样本被确定为边界样本的概率越高；相反，某一少数类样本的支持度越小，则意味着该样本被确定为边界样本的概率越低。通过引入支持度 sd 避免传统 SMOTE 过采样算法随机选择边界样本的盲目性。

3.2 新样本合成中随机数的确定(影响因素 *posFac*)

为了缓解 SMOTE 算法合成少数类样本时随机数取值的局限性，本文引入影响因素 $posFac$ 对随机数进行精细控制[12]，具体计算步骤如下所示：

(1) 计算少数类样本 $\boldsymbol{x}_i$ 与其 K 个同类近邻的平均欧氏距离 d_{pave-i}：

$$d_{pave-i}=\frac{\sum_{j=1}^{K}\sqrt{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}}{K} \tag{4}$$

其中，$\boldsymbol{x}_i$ 表示当前的少数类样本，$\boldsymbol{x}_j$ 表示样本 $\boldsymbol{x}_i$ 的第 j 个少数类近邻。

(2) 计算所有少数类样本与其 K 个同类近邻的平均欧氏距离 $d_{\text{pave-sum}}$：

$$d_{\text{pave-sum}}=\sum_{i=1}^{m}d_{pave-i} \tag{5}$$

其中，m 表示少数类样本的数量。

(3) 计算少数类样本集间的平均欧氏距离 d_{pave}：

$$d_{\text{pave}}=\frac{d_{\text{pave-sum}}}{m} \tag{6}$$

(4) 计算少数类样本与其 K 个多数类近邻的平均欧氏距离 d_{nave-i}：

$$d_{nave-i}=\frac{\sum_{j=1}^{K}\sqrt{\|\boldsymbol{x}_i-\boldsymbol{y}_j\|^2}}{K} \tag{7}$$

其中，$\boldsymbol{x}_i$ 表示当前的少数类样本，$\boldsymbol{y}_j$ 表示样本 $\boldsymbol{x}_i$ 的第 j 个多数类近邻。

(5) 计算所有少数类样本与其 K 个多数类近邻的平均欧氏距离 $d_{\text{nave-sum}}$：

$$d_{\text{nave-sum}}=\sum_{i=1}^{m}d_{nave-i} \tag{8}$$

其中，m 表示少数类样本的数量。

(6) 计算少数类样本与多数类样本间的平均欧氏距离 d_{nave}：

$$d_{\text{nave}}=\frac{d_{\text{nave-sum}}}{m} \tag{9}$$

(7) 在合成样本的过程中，计算当前被选中的边界样本与其 K 个同类近邻的平均欧氏距离 d_1。

(8) 在合成样本的过程中，计算当前被选中的边界样本与其 K 个多数类近邻的平均欧氏距离 d_2。

(9) 计算相对距离比 u：

$$u=\frac{d_1/d_{\text{pave}}}{d_2/d_{\text{nave}}} \tag{10}$$

(10) 根据相对距离比 u，计算影响因素 $posFac$ 的取值范围：

$$posFac=\begin{cases}rand(0,1), u<1\\ 0.5+0.5\times rand(0,1), 1\leqslant u\leqslant 2\\ 0.8+0.2\times rand(0,1), u>2\end{cases} \tag{11}$$

其中，$rand(0,1)$ 表示 $(0,1)$ 内的随机数，使用 $posFac$ 作为样本合成的随机数的取值，该影响因素考虑了周围多数类样本的影响，可以更加合理地合成少数类样本。

3.3 基于样本分布的 SDRSMOTE 算法

基于样本分布的 SDRSMOTE 算法，首先根据少数类样本的支持度 sd 选择边界样本 $\boldsymbol{x}_i$，然后使用 SMOTE 算法选择另一个与 $\boldsymbol{x}_i$ 相邻的少数类样本 $\boldsymbol{x}_j$。新样本的合成公式为：

$$\boldsymbol{x}_{\text{new}}=\boldsymbol{x}_i+posFac\times(\boldsymbol{x}_j-\boldsymbol{x}_i) \tag{12}$$

这种数据合成算法增加了边界上的少数类样本的数量，使得新样本生成过程中的随机数取值更有针对性，有效地缓解了传统 SMOTE 算法的盲目性，提高了少数类样本的分类性能。

SDRSMOTE 的基本步骤如下：

Step 1 将应合成的少数类样本数记为 Num；

Step 2 假定少数类样本的数量为 m，多数类样本的数量为 n。随机选择一个少数类样本 $\boldsymbol{x}_i$，并根据式(1)计算 $\boldsymbol{x}_i$ 与每个多数类样本之间的欧氏距离之和 S_i；

Step 3 根据式(2)计算所有 S_i 的累加和 S；

Step 4 根据式(3)计算少数类样本和多数类样本之间的平均欧氏距离 S_{ave};

Step 5 以 S_{ave} 为半径,依次选择每个少数类样本作为中心圈定区域,然后统计该区域中多数类样本的数量作为当前少数类样本的支持度 sd;

Step 6 根据支持度 sd 选择少数类样本进行新样本的合成,对所选样本使用 SMOTE 算法,根据式(11)引入影响因素 $posFac$,然后根据式(12)进行合成新样本;

Step 7 将新合成的少数类样本添加到数据集以参与训练和测试。

SDRSMOTE 算法在样本合成之前通过引入支持度 sd 确定边界样本的选择顺序,可以缓解传统 SMOTE 过采样算法随机选择边界样本的盲目性;在样本合成过程中通过综合总体样本的分布状况引入影响因素 $posFac$ 进行新样本的合成,使得新样本合成过程中的随机数取值更有针对性;该算法可以更加合理地扩展少数类样本,使得新数据集是趋于平衡的,提高分类器对于少数类样本的分类能力。

4 实验与分析

4.1 实验对象

为了评估 SDRSMOTE 在不平衡数据分类中的有效性,实验选取了 6 个分类不平衡数据集。blood、haberman、iris、breast 数据集来自 UCI[18] 机器学习库,PC3、JM1 来自 NASA 标准数据集,基本信息如表 2 所示。第 1、2 列分别表示序号、数据集名称;第 3 列表示特征维度,即属性个数;第 4 列表示样本总数,描述了数据集的规模;第 5、6 列分别表示少数类样本数、多数类样本数;第 7 列表示少数类比率即少数类样本数与总样本数的比值;第 8 列表示数据集的不平衡率,即多数类样本数与少数类样本数的比值,该值越大说明数据集的分类越不平衡[14]。

4.2 实验设计

实验使用开源数据挖掘工具 WEAK3.6,本文研究二分类问题,因此将多分类所用的数据集转化为二分类数据集,所用的数据集如表 2 所示。分别使用 SMOTE 算和 SDRSMOTE 算法在选取的数据集上进行过采样,其中邻域值 K 取 5,少数类样本插值倍频 N 取 1.2。在新数据集上分别使用具有代表性的经典分类算法决策树(J48)、朴素贝叶斯以及集成分类算法 AdaBoost、Bagging,采用十折交叉验证进行预测。对不平衡数据集分类效果的评价指标使用 $F\text{-}value$、$G\text{-}mean$ 和 AUC,验证对比传统 SMOTE 和 SDRSMOTE 算法生成少数类样本的合理性。

4.3 实验结果与分析

使用改进前的 SMOTE 和改进后的 SDRSMOTE 算法对数据集进行过采样处理后,再分别使用决策树(J48)、AdaBoost、Bagging、朴素贝叶斯 NB(Naive Bayes)4 种算法进行分类,得到分类后的 $F\text{-}value$、$G\text{-}mean$ 和 AUC 值,结果如表 3~表 5 所示,为了便于比较,采用“加粗”来标记在同一种分类模型下经过 SDRSMOTE 过采样后的性能优于 SMOTE 的数值。

从表 3 可以看出,采用 AdaBoost 对 6 个不平衡数据集分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 $F\text{-}value$ 的均值提高了0.014,平均提高 1.76%,效果最好;采用决策树分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 $F\text{-}value$ 的均值提高了 0.009,平均提高 1.1%;采用 Bagging 分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 $F\text{-}value$ 的均值提高了 0.005,平均提高 0.59%;采用朴素贝叶斯分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 $F\text{-}value$ 的均值提高了 0.004,平均提高 0.57%。使用 SDRSMOTE 算法结合决策树、AdaBoost、Bagging、朴素贝叶斯分类器上对

Table 2 Basic information of the data set

表 2 数据集基本信息

数据集	名称	特征数	样本总数	少数类样本数	多数类样本数	少数类比率/%	不平衡率/%
0	PC3	41	1 563	160	1 403	10.2	8.8
1	JM1	22	7 782	1 672	6 110	21.5	3.7
2	blood	5	748	178	570	23.8	3.2
3	haberman	4	306	81	225	26.5	2.8
4	iris	5	150	50	100	33.3	2.0
5	breast	10	699	241	458	34.5	1.9

Table 3 *F-value* of each prediction model on 6 data sets

表 3 各个预测模型在 6 个数据集上的 *F-value* 值

数据集	分类器							
	决策树		AdaBoost		Bagging		NB	
	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE
0	0.857	**0.869**	0.831	**0.846**	0.871	**0.885**	0.516	0.512
1	0.770	**0.775**	0.672	**0.699**	0.797	**0.803**	0.595	**0.609**
2	0.738	0.726	0.695	**0.706**	0.748	0.737	0.608	**0.614**
3	0.696	**0.735**	0.663	**0.685**	0.701	**0.714**	0.611	0.601
4	0.947	**0.952**	0.957	**0.962**	0.962	**0.966**	0.914	**0.928**
5	0.952	**0.957**	0.958	**0.962**	0.964	**0.968**	0.963	**0.964**
均值	0.827	**0.836**	0.796	**0.810**	0.841	**0.846**	0.701	**0.705**
平均提高	**0.009**		**0.014**		**0.005**		**0.004**	
平均提高率/%	**1.10**		**1.76**		**0.59**		**0.57**	

Table 4 *G-mean* of each prediction model on 6 data sets

表 4 各个预测模型在 6 个数据集上的 *G-mean* 值

数据集	分类器							
	决策树		AdaBoost		Bagging		NB	
	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE
0	0.759	**0.774**	0.717	**0.721**	0.743	**0.784**	0.618	**0.623**
1	0.740	**0.747**	0.630	**0.664**	0.767	**0.775**	0.558	**0.570**
2	0.719	0.707	0.681	**0.691**	0.731	0.720	0.588	**0.593**
3	0.690	**0.731**	0.654	**0.680**	0.694	**0.710**	0.614	0.606
4	0.947	**0.951**	0.956	**0.961**	0.961	**0.966**	0.914	**0.928**
5	0.951	**0.956**	0.958	**0.961**	0.963	**0.967**	0.962	**0.963**
均值	0.801	**0.811**	0.766	**0.780**	0.810	**0.821**	0.709	**0.714**
平均提高	**0.010**		**0.014**		**0.011**		**0.005**	
平均提高率/%	**1.25**		**1.83**		**1.36**		**0.71**	

Table 5 *AUC* of each prediction model on 6 data sets

表 5 各个预测模型在 6 个数据集上的 *AUC* 值

数据集	分类器							
	决策树		AdaBoost		Bagging		NB	
	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE	SMOTE	SDRSMOTE
0	0.786	**0.793**	0.851	**0.876**	0.911	**0.924**	0.787	**0.792**
1	0.788	0.787	0.722	**0.739**	0.847	**0.852**	0.626	**0.661**
2	0.757	**0.769**	0.759	**0.774**	0.806	**0.809**	0.724	**0.736**
3	0.704	**0.717**	0.689	**0.702**	0.781	0.779	0.628	**0.656**
4	0.954	**0.955**	0.988	0.987	0.990	**0.997**	0.986	0.985
5	0.958	**0.972**	0.993	0.992	0.986	**0.991**	0.986	0.985
均值	0.825	**0.832**	0.834	**0.845**	0.887	**0.892**	0.790	**0.798**
平均提高	**0.007**		**0.011**		**0.005**		**0.080**	
平均提高率/%	**0.85**		**1.32**		**0.56**		**1.01**	

不平衡数据集 PC3、JM1、blood、haberman、iris、breast 进行预测，*F-value* 均值明显提高，其中在 AdaBoost 分类器上效果最为显著。

从表 4 可以看出，采用 AdaBoost 对 6 个不平衡数据集分类时，SDRSMOTE 算法与 SMOTE 算法相比，在 6 个数据集上 *G-mean* 的均值提高了

0.014,平均提高 1.83%,效果最好;采用决策树分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *G-mean* 的均值提高了 0.01,平均提高 1.25%;采用 Bagging 分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *G-mean* 的均值提高了 0.011,平均提高 1.36%;采用朴素贝叶斯分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *G-mean* 的均值提高了 0.005,平均提高 0.71%。使用 SDRSMOTE 算法结合决策树、AdaBoost、Bagging、朴素贝叶斯分类器上对不平衡数据集 PC3、JM1、blood、haberman、iris、breast 进行预测,*G-mean* 均值明显提高,其中在 AdaBoost 分类器上效果最为显著。

从表 5 可以看出,采用 AdaBoost 对 6 个不平衡数据集分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *AUC* 的均值提高了 0.011,平均提高 1.32%,效果最好;采用决策树分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *AUC* 的均值提高了 0.007,平均提高 0.85%;采用 Bagging 分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *AUC* 的均值提高了 0.005,平均提高 0.56%;采用朴素贝叶斯分类时,SDRSMOTE 算法与 SMOTE 算法相比,在 6 个数据集上 *AUC* 的均值提高了 0.008,平均提高 1.01%。使用 SDRSMOTE 算法结合决策树、AdaBoost、Bagging、朴素贝叶斯分类器上对不平衡数据集 PC3、JM1、blood、haberman、iris、breast 进行预测,*AUC* 均值明显提高,其中在 AdaBoost 分类器上效果最为显著。

综合表 3～表 5 的 *F-value*、*G-mean* 和 *AUC* 值发现,SDRSMOTE 算法与 AdaBoost 分类器结合后的分类性能提升最为明显。

采用 AdaBoost 分类器对经过 SMOTE 和 SDRSMOTE 过采样算法处理的数据集进行分类,图 2～图 4 分别列出了评价指标 *F-value*、*G-mean* 和 *AUC* 对比结果。图 2 显示这些数据集经 SDRSMOTE 过采样处理后,AdaBoost 分类器的 *F-value* 值普遍高于经 SMOTE 处理的 *F-value* 值,在数据集 JM1 上尤其明显。图 3 的柱状图显示,这些数据集经 SDRSMOTE 过采样处理后,AdaBoost 分类器的 *G-mean* 值普遍高于经 SMOTE 处理的 *G-mean* 值,在数据集 JM1、haberman 上非常明显。从图 4 中可以清楚地看到,当数据集是 PC3、JM1、blood、haberman 时,SDRSMOTE 与 AdaBoost 结合的 *AUC* 值明显高于 SMOTE 与 AdaBoost 结合的 *AUC* 值;而在数据集 iris、breast 上 2 种模型的 *AUC* 值不相上下。综上所述,将 AdaBoost 作为不平衡数据集的预测模型时,SDRSMOTE 的实验结果明显优于 SMOTE 的。

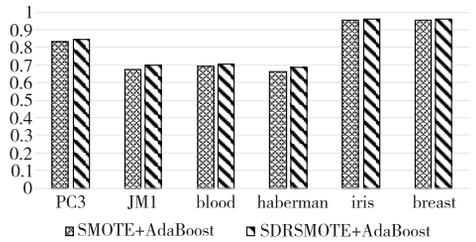

Figure 2 *F-value* for SMOTE+AdaBoost and SDRSMOTE+AdaBoost

图 2 SMOTE+AdaBoost 和 SDRSMOTE+AdaBoost 的 *F-value* 值

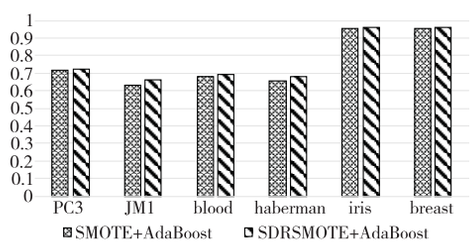

Figure 3 *G-mean* for SMOTE+AdaBoost and SDRSMOTE+AdaBoost

图 3 SMOTE+AdaBoost 和 SDRSMOTE+AdaBoost 的 *G-mean* 值

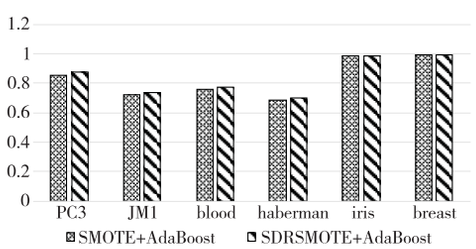

Figure 4 *AUC* for SMOTE+AdaBoost and SDRSMOTE+AdaBoost

图 4 SMOTE+AdaBoost 和 SDRSMOTE+AdaBoost 的 *AUC* 值

综合评价指标 *F-value*、*G-mean*、*AUC* 来看,SDRSMOTE 算法优于 SMOTE 算法,在 AdaBoost 分类器上性能提高尤为显著,即 SDRSMOTE 算法与该分类器结合使用效果最好。

5 结束语

本文提出的 SDRSMOTE 算法,在样本合成之前通过引入支持度 *sd* 确定边界样本的选择顺序,在样本合成时通过引入影响因素 *posFac* 对随机数

进行精细控制。它不仅可以避免传统 SMOTE 过采样算法选择边界样本的盲目性,而且还综合考虑了总体样本的分布状况。实验结果表明,该算法相比传统 SMOTE 过采样算法,在合成少数类样本时更加合理,有利于提高少数类的分类性能。但提出的 SDRSMOTE 算法增加了算法的时间复杂度,因此算法还需要进一步优化,以提高运行效率。

参考文献:

[1] Ye Feng,Ding Feng. Research and application of unbalanced data classification[J]. Computer Applications and Software, 2018,35(1):132-136. (in Chinese)

[2] Liu Dong-qi. Support vector machine based classification algorithms research for imbalanced data[D]. Hangzhou: Zhejiang University,2017. (in Chinese)

[3] 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.

[4] Giraldo-Forero A F,Jaramillo-Garzón J A,Ruiz-Muñoz J F, et al. Managing imbalanced data sets in multi-label problems: A case study with the SMOTE algorithm[C]// Proc of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Application, 2013:334-342.

[5] Gu Ping, Ouyang Yuan-you. Classification research for unbalanced data based on mixed-sampling[J]. Application Research of Computers,2015,32(2):379-381. (in Chinese)

[6] Souto M C P D,Bittencourt V G,Costa J A F. An empirical analysis of under-sampling techniques to balance a protein structural class dataset[J]. Lecture Notes in Computer Science,2006,4234:21-29.

[7] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[C]// Proc of Conference on AI in Medicine in Europe:Artificial Intelligence Medicine, 2001:63-66.

[8] Nguyen T D,Le T,Vu H,et al. Dual Discriminator Generative Adversarial Nets[C]// Proc of the 31th Conference of Neural Information Processing Systems, 2017:2671-2681.

[9] Xu Li-li Yan De-qin. Weighted ensemble learning algorithm for imbalanced data sets[J]. Microcomputer & its Applications,2015,34(23):7-10. (in Chinese)

[10] Nghe N T,Janecek P,Haddawy P. A comparative analysis of techniques for predicting academic performance[C]// Proc of the 37th IEEE Frontiers in Education Conference-global Engineering: Knowledge Without Borders, 2007: 7-12.

[11] Feng Hua-min,Li Ming-wei,Hou Xiao-lian,et al. Study of network intrusion detection method based on SMOTE and GBDT[J]. Application Research of Computers, 2017, 34(12):3745-3748. (in Chinese)

[12] Wei Hao,Li Hong,Liu Xiao-yu. An improved SMOTE algorithm[J]. Henan Science,2018,36(7):1009-1013. (in Chinese)

[13] Li Ke-wen, Yang Lei, Liu Wen-ying, et al. Classification method of imbalanced data based on RSBoost[J]. Computer Science,2015,42(9):249-252. (in Chinese)

[14] Yu Qiao,Jiang Shu-juan,Zhang Yan-mei,et al. The impact study of class imbalance on the performance of software defect prediction models[J]. Chinese Journal of Computers, 2018,41(4):809-824. (in Chinese)

[15] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997,30(7):1145-1159.

[16] Huang J,Ling C X. Using AUC and accuracy in evaluating learning algorithms[J]. IEEE Transactions on Knowledge & Data Engineering,2005,17(3):299-310.

[17] Li K,Zhang W,Lu Q,et al. An improved SMOTE imbalanced data classification method based on support degree[C]// Proc of International Conference on Identification,2014: 34-38.

[18] Center for Machine Learning and Intelligent Systems. UCI machine learning repository[DB/OL]. [2018-04-09]. http://archive.ics.uci.edu/ml/datasets.html.

附中文参考文献:

[1] 叶枫,丁锋. 不平衡数据分类研究及其应用[J]. 计算机应用与软件,2018,35(1):132-136.

[2] 刘东启. 基于支持向量机的不平衡数据分类算法研究[D]. 杭州:浙江大学,2017.

[5] 古平,欧阳源遊. 基于混合采样的非平衡数据集分类研究[J]. 计算机应用研究,2015,32(2):379-381.

[9] 徐丽丽,闫德勤. 不平衡数据加权集成学习算法[J]. 微型机与应用,2015,34(23):7-10.

[11] 封化民,李明伟,侯晓莲,等. 基于 SMOTE 和 GBDT 的网络入侵检测方法研究[J]. 计算机应用研究,2017,34(12):3745-3748.

[12] 魏浩,李红,刘小豫. 一种改进的 SMOTE 算法[J]. 河南科学,2018,36(7):1009-1013.

[13] 李克文,杨磊,刘文英,等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学,2015,42(9):249-252.

[14] 于巧,姜淑娟,张艳梅,等. 分类不平衡对软件缺陷预测模型性能的影响研究[J]. 计算机学报,2018,41(4):809-824.

作者简介:

李克文(1969-),男,黑龙江齐齐哈尔人,博士,教授,博士生导师,CCF 会员(E200014144M),研究方向为软件工程和人工智能。**E-mail**:likw@upc.edu.cn

LI Ke-wen,born in 1969,PhD,professor,PhD supervisor,CCF member(E200014144M),his research interests include software engineering,and artificial intelligence.