ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 727－740
DOI:10. 16337/j. 1004-9037. 2023. 03. 020

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于Bootstrap方法最大熵优化过采样算法

雷天纲，陈　刚

（大连海事大学理学院，大连 116026）

摘　要：随着数据时代的到来，非平衡数据的分类问题受到越来越多的关注。在非平衡数据的分类问题中，往往因为少数类样本与多数类样本比例失衡而导致分类结果错误。因此，提出了一种在最大熵原理下基于自助法（Bootstrap method）的过采样算法。首先，通过自助法获得数据样本的概率分布，并用最大熵原理对概率分布进行优化；其次，根据少数类生成新的少数类的能力不同，提出基于少数类样本分布的概率增强算法。该算法使数据随机性得到了充分体现，保证了少数类样本的概率密度在数据集平衡前后保持一致性，从而提高分类算法的有效性；最后，通过从UCI和KEEL数据库选取8组数据进行实验，实验结果表明所提出的新算法比现有的其他算法更有效。

关键词：非平衡数据；自助法；最大熵原理；概率增强；分类

中图分类号：TP373　　**文献标志码：**A

An Over-Sampling Algorithm for Maximum Entropy Optimization Based on Bootstrap Method

LEI Tiangang, CHEN Gang

(School of Science, Dalian Maritime University, Dalian 116026, China)

Abstract: With the advent of the data era, the classification of unbalanced data is receiving more and more attention. In the classification of unbalanced data, classification results are often incorrect due to an imbalance in the ratio of minority class samples to majority class ones. Therefore, we propose an oversampling algorithm based on the Bootstrap method under the maximum entropy principle. Firstly, the probability distribution of the data sample is obtaited through self-help method and optimized using the principle of maximum entropy. Secondly, a probability enhancement algorithm based on minority class sample distribution is proposed based on different abilities of minority classes to generate new minority classes. The algorithm allows the randomness of the data to be fully represented and ensures that the probability density of the minority class remains consistent before and after the data set is balanced, thus improving the effectiveness of the classification algorithm. Finally, experiments are conducted by selecting eight data sets from the UCI and KEEL databases, whose results show that the proposed algorithm is more effective than other algorithms.

Key words: unbalanced data; Bootstrap method; principle of maximum entropy; probability enhancement; classification

基金项目：国家自然科学基金（11571056）。
收稿日期：2022-01-20；**修订日期：**2023-04-20

引 言

随着大数据时代的到来,数据大量存在于人们的日常生活中。在这些数据中绝大多数是非平衡数据[1],而非平衡数据的分类[2]要比平衡数据的分类更加复杂且有意义。因此,非平衡数据的分类问题被广大研究者所关注,例如:癌症诊断[3]、网络入侵检验[4]、生物信息学[5]、信用卡还款欺诈[6]、语音识别[7]等。为了解决类不平衡问题,研究者们先后提出许多不同的解决方法[8-12]。这些方法可以大致分为两类,分别是数据层面的方法与算法层面的方法。

数据层面的方法主要是重采样算法,包括过采样算法[13-19]、欠采样算法[20-24]和过采样欠采样相结合[25]的算法。通常情况下,由于欠采样算法在采样过程中会发生信息损失[26-27],所以对过采样算法更加感兴趣。过采样算法中最常用的就是SMOTE(Synthetic minority over-sampling technique)[13],SMOTE算法本质是根据少数类的几何特征生成新的少数类,首先计算出每个少数类样本的K个近邻;其次,从K个近邻中随机挑选N个样本进行分析,根据分析人工构造新的少数类;最后,将合成的新样本与原始数据样本结合起来,使非平衡数据转换为平衡数据。随后,基于SMOTE算法的一系列改进算法[18, 28-32]被开发出来。Han等[28]提出了Borderline-SMOTE算法,该算法仅仅对边界线附近的少数样本进行过采样。Borderline-SMOTE又可分为Borderline-SMOTE1和Borderline-SMOTE2,Borderline-SMOTE1对边界点生成新样本时,在K近邻的少数类中选择一个样本生成新样本,Borderline-SMOTE2则是在K近邻中的任意一个样本生成新样本。ADASYN(Adaptive synthetic)[33]的基本思想是为更难学习的少数类样本使用加权分布生成更多合成样本。RSMOTE(Robust SMOTE)算法[32]是一种自适应鲁棒SMOTE算法。Kernel-ADASYN算法[34]提出了一种新的基于内核的自适应合成过采样方法。Bunkhumpornpat等[29]提出了Safe-Level SMOTE算法,该算法以划分的安全级别来过采样少数类。SPIDER2(Selective preprocessing of imbalanced data 2)[15]分为两个步骤对多数类样本和少数类样本进行预处理。首先,定义多数类样本的特征并检测噪声样本,根据新标记的选项重新分类为少数类。随后,将同样的操作运用在少数类样本上。陈刚等[35]提出一种基于GMM-EM(Gaussian mixed mode-expectation maximization)的非平衡数据的概率增强算法,该算法通过均值最大化算法获得少数类原始数据的统计特征,在保持统计特征不变的情况下进行过采样生成高质量的少数类新样本。

对于算法层面的方法,主要的策略是提出新算法或者对现有分类算法进行改进,以便提升分类算法的分类性能。Rivera等[36]对基于先验的OUPS(Over-sampling using propensity scores)和SLOUPS(Safe level OUPS)算法进行了改进。Dong等[37]采用深度学习方法进行非平衡数据分类,建立了基于批量增量少数类修正的非平衡深度学习模型。Tao等[38]提出了一种基于成本敏感的非平衡数据分类集成方法,是一种用于集成分类的基于自适应成本敏感的支持向量机。

上述介绍的方法虽然在一定程度上能够解决一些非平衡问题,但这些方法或多或少存在一些局限性。例如:SMOTE系列算法不能有效地扩展正样本的训练领域,导致生成的新样本缺乏多样性,不能准确地近似正样本的概率分布[39]。对于以Kernel-ADASYN算法为代表的核函数类算法虽然考虑了少数类的统计特征,但本质上是基于分类器的一种改进算法。而基于GMM-EM的概率增强算法在原始少数类样本概率密度的获取上有一定的局限性。首先,一些对高斯分布拟合差的样本数据会造成参数估计不准确,产生质量较差的少数类新样本,对数据的分布有一定的要求。其次,EM算法对初值敏感,聚类的结果往往随着初始值的不同而波动较大。最后,EM算法虽然一定会收敛但不一定能达到全局最优,只有优化的目标函数是一个凸函数,才能保证得到全局最优解。

在GMM-EM算法中,高斯混合模型只针对分布为高斯分布的数据表现出不错的拟合效果。然而

在实际应用中所获得的数据分布大多数都是不确定的,对于这些数据GMM-EM算法很难产生理想的分类效果。因此,能够提出一种在获得数据样本分布方面没有限制的方法十分重要。基于此,本文选用Bootstrap方法,它对原始样本的概率分布没有限制条件,能够获得最接近数据真实分布的概率,因而可以避免GMM-EM算法的不足。然而,Bootstrap方法在数据量较小时容易出现估计不准确的问题。因此,将数据转换为特征向量,并利用最大熵模型来选择熵最大的特征概率模型。最大熵模型可以处理多种类型的特征,并且可以很好地处理不完整的数据。将Bootstrap方法获得的特征期望值当做最大熵模型的约束条件,并利用优化算法求解最大熵模型的参数,得到一个概率分布。相较于单独使用Bootstrap方法或最大熵模型,本文方法在小规模数据上表现更好,同时用Bootstrap方法得到的概率分布充当最大熵模型的限制条件确保最大熵模型的输出概率分布与真实的概率分布尽可能接近,因此在实际的过程中要优于单独使用的两种算法。最后,利用最大熵模型来估计少数类样本的概率分布,并根据概率分布来决定过采样的权重,以平衡数据集中不同类别的样本数量,从而提高分类的效果。实验结果表明,相较于常用的过采样算法,提出的新算法具有更好的性能和优势。

1 准备知识

1.1 Bootstrap方法

Bootstrap方法是Efron[40]在20世纪70年代后期建立的。其主要内容是在原始数据的样本中作有放回的重复抽样,抽样得到的样本量为n,原始数据中每个样本每次被抽到的概率均为$1/n$,获得的样本称为Bootstrap样本。相继地、独立地抽取多个Bootstrap样本,利用这些样本对原始数据进行统计推断。这一方法可以用于对总体掌握较少的情况,它是近代统计中数据处理常用的方法。

为了得到Bootstrap方法,首先介绍经验分布函数,假设$X_1, X_2, \cdots, X_n$是总体分布F的一个样本,用$S(x)$表示$X_1, X_2, \cdots, X_n$中不大于x的随机变量的个数,其中$-\infty < x < \infty$,则经验分布函数可表示为

$$F_n(x)=\frac{1}{n}S(x) \quad -\infty<x<\infty \tag{1}$$

而经验分布函数成立的理论依据就是格里汶科定理。格里汶科定理中提到,对于任一实数x,当$n\to\infty$时,经验分布函数$F_n(x)$概率取值为1时一致收敛于分布$F(x)$,即

$$P\left\{\lim_{n\to\infty}\sup_{-\infty<x<\infty}\left|F_n(x)-F(x)\right|=0\right\}=1 \tag{2}$$

因此,对于任一实数x当n充分大时,经验分布函数$F_n(x)$可当作$F(x)$来使用。

下面介绍Bootstrap方法的具体内容。

首先,获得样本数据集$x=(x_1, x_2, \cdots, x_n)$,按放回抽样的方法抽得容量为$n$的样本$x^*=\left(x_1^*, x_2^*, \cdots, x_n^*\right)$。其次,相继地、独立地求出$B$个($B\geqslant 1\,000$)容量为$n$的Bootstrap样本,对于第$i$个Bootstrap样本,计算每个样本$x_j, j=1, 2, \cdots, n$,在Bootstrap样本中出现的频率$f_{x_j}^i=\frac{k_{x_j}^i}{n}$,其中$i=1, 2, \cdots, B$,$j=1, 2, \cdots, n$,$k_{x_j}^i$为样本$x_j$在第$i$个Bootstrap样本中出现的次数。最后,用频率估计概率计算每个样本$x_j, j=1, 2, \cdots, n$的概率

$$f_{x_j}=\sum_{i=1}^{B}\frac{f_{x_j}^i}{B} \quad i=1, 2, \cdots, B;\ j=1, 2, \cdots, n \tag{3}$$

这样就得到了样本数据分布的Bootstrap估计。

本文采用Bootstrap方法进行数据处理。具体而言,首先获得样本数据集,然后按放回抽样的方法

抽取容量为n的Bootstrap样本,并计算每个样本在Bootstrap样本中出现的频率;接着,利用频率估计概率计算每个样本的概率,从而得到样本数据分布的Bootstrap估计。该方法在统计学中被广泛应用,可用于对总体掌握较少的情况下进行推断。因此,Bootstrap方法是一种重要的数据处理方法,其应用范围广泛,对于数据分析和统计推断具有重要意义。需要注意的是,当样本量较小时,Bootstrap方法的效果并不理想,此时需要采用最大熵模型进行优化。

1.2 最大熵模型及优化

在信息熵与概率统计中,熵是表示随机变量不确定的度量,一个离散随机变量X的概率分布为$P(X=x_i)=p_i, i=1,2,\cdots,n$,则其熵为

$$H(X)=-\sum_{i=1}^{n} p_i \lg p_i \tag{4}$$

最大熵原理[41]作为概率模型学习的准则之一。通常认为最好的模型熵也是最大的,因此最大熵原理就是要在那些满足约束条件的模型集合中选出熵最大的模型。熵满足不等式$0\leqslant H(P)\leqslant \lg|X|$,其中$|X|$是$X$的取值个数,所以只有$X$是均匀分布时不等式右边等号成立,熵最大。将最大熵原理应用到分类问题中就得到最大熵模型。如果限制条件都满足的所有模型集合为$\mathcal{C}\equiv\left\{P\in P\middle| E_p(f_i)=E_{\hat{p}}(f_i), i=1,2,\cdots,n\right\}$,且定义在条件概率分布$P(Y|X)$上的条件熵为

$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\lg P(y|x) \tag{5}$$

则满足限制条件的模型集合$\mathcal{C}$中条件熵$H(P)$最大的模型就是最大熵模型。

为了便于求解,将最大值问题转化为等价的最小值问题

$$\min_{P\in\mathcal{C}} -H(P)=\sum_{x,y}\tilde{P}(x)P(y|x)\lg P(y|x) \tag{6}$$

之后求解约束最优化问题得出的解就是最大熵模型学习的解。本研究中对最大熵模型的学习采用了改进的迭代尺度法,改进的迭代尺度算法(Improved iterative scaling)[42]是一种常用的最大熵模型学习的最优化算法。

2 基于自助法的最大熵原理优化下过采样算法

由于人类面对现实世界的复杂性以及获取数据技术的限制,常常得到一类不完整、缺失的非平衡数据集。少数类的样本数量较少,导致分类器在训练和测试过程中对于少数类别的样本分类效果较差。根据以上讨论,本文利用自助法与最大熵原理,针对非平衡数据的分类问题提出一种基于自助法在最大熵原理优化下的过采样算法,算法的内容如下:

Step 1 基于自助法获得数据的概率分布。首先,从总体数据集中获取一个样本数据集,该样本数据集的大小与总体数据集相同。其次,为了估计总体数据集的概率分布,使用自助法对样本数据集进行随机放回抽样,得到多个自助样本数据集。对于每个自助样本数据集,使用每个样本在该自助样本数据集中出现的频率来估计总体数据集中该样本的概率,最终得到总体数据集的概率分布估计值。该方法可以有效地解决样本数据集过小导致的概率分布估计不准确的问题。

Step 2 利用自助法获得的概率分布作为最大熵模型的限制条件,从而优化概率分布。

(1) 根据最大熵原理构建最大熵模型。由于本文研究的非平衡数据分类问题为二分类问题,所以构建的最大熵模型为

$$\begin{cases}\max\limits_{P\in\mathcal{C}} H(P)=-\sum\limits_{x,y}\tilde{P}(x)P(y|x)\lg P(y|x)\\ \text{s.t.}\quad E_p(f_i)=E_{\hat{p}}(f_i)\quad i=1,2,\cdots,n\\ \qquad\sum\limits_y P(y|x)=1\end{cases}\tag{7}$$

为方便求解,将最大熵模型转化为最优化问题

$$\begin{cases}\min\limits_{P\in\mathcal{C}} -H(P)=\sum\limits_{x,y}\tilde{P}(x)P(y|x)\lg P(y|x)\\ \text{s.t.}\quad E_p(f_i)=E_{\hat{p}}(f_i)\quad i=1,2,\cdots,n\\ \qquad\sum\limits_y P(y|x)=1\end{cases}\tag{8}$$

(2) 为求解最优化问题,引入拉格函数作为辅助函数进行求解

$$L(P,w)\equiv -H(P)+w_0\left(1-\sum_y P(y|x)\right)+\sum_{i=1}^{n} w_i\left(E_{\tilde{P}}(f_i)-E_P(f_i)\right)\tag{9}$$

由于拉格朗日函数是凸函数,所以可以将最优化的原始问题 $\min\limits_{P\in C}\max\limits_{w} L(P,w)$ 转化为对偶问题 $\max\limits_{w}\min\limits_{P\in C} L(P,w)$。求解后得到最大熵模型

$$P_w(y|x)=\frac{1}{Z_w}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)\tag{10}$$

式中:$Z_w=\sum\limits_y\exp\left(\sum\limits_{i=1}^{n} w_i f_i(x,y)\right)$ 为规范化因子;$f_i(x,y)$ 为特征函数;w_i 为特征的权重。

(3) 通过改变最大熵模型的限制条件进行模型优化。将自助法得到的概率分布 $P(X,Y)$ 代替训练集的联合分布的经验分布 $\tilde{P}(X=x,Y=y)$,利用改进的迭代尺度法对最大熵模型进行求解得到模型最优参数。

(4) 将求得的最优解代入最大熵模型,得到优化后少数类的概率 f'_{x_s}。

Step 3 将少数类分布的概率权重用作过采样的权重。具体做法是用少数类样本的概率权重决定采样权重。通过上述方法得到了少数类样本的概率分布。因此每个少数类样本的概率权重为

$$\omega_{x_s}=\frac{f'_{x_s}}{\sum\limits_{s=1}^{n} f'_{x_s}}\quad s=1,2,\cdots,n\tag{11}$$

式中:f'_{x_s} 为少数类样本 x_s 的概率值;n 为少数类样本的数量。

Step 4 确定过采样的数量。由 Step 3 中得到的采样权重决定每个少数类样本的采样次数,总的过采样数量为

$$\Delta=N_{\text{maj}}-N_{\text{min}}\tag{12}$$

式中:N_{maj} 为多数类样本数;N_{min} 为少数类样本数。

Step 5 对少数类样本进行过采样产生新的少数类样本。为了防止采样前后样本重复,假定 x_i 是与少数类样本 x_s 欧式距离最近的少数类,即

$$x_i=\min_{s\neq i}^{n} D(x_i,x_s)\quad i,s=1,2,\cdots,n\tag{13}$$

则新生成的少数类样本 x_s^k 为

$$x_s^k=x_s\pm\frac{1}{2}rD(x_i,x_s)\quad k=1,2,\cdots,S_{x_s}\tag{14}$$

式中：r为(0,1)之间的随机数；$x_s^k=\left[x_s^{1(k)},x_s^{2(k)},\cdots,x_s^{p(k)}\right]$表示少数类第$s$个原始样本生成的第$k$个新样本。

Step 6 算法评估。本文旨在通过评估决策树分类器在分类问题中的有效性，来比较本文提出的新算法与其他经典算法（包括ORIGINAL、SMOTE、Borderline1-SMOTE、ADASYN和Borderline2-SMOTE）的性能。为了定量评估每种算法的性能，将使用多种评价指标，包括AUC、Acc、Sen、Spe、Pre和F_1值。

本文算法伪代码如下。

输入：包含N个训练样本的数据集$D=\left\{x_j,y_j\right\},j=1,2,\cdots,N$，其中$x_j$代表第$j^{\text{th}}$样本，$y_j$是$x_j$的类别标签；

输出：生成新的少数类。

(1) 确定样本数据集的经验分布。

(2) for D do sampling with replacement

(3) 得到B个Bootstrap样本。

(4) end for;

(5) 计算x_j在Bootstrap样本中出现的次数

(6) $\text{set}\ f_j'=\sum_{i=1}^{B}\dfrac{\text{count}\left(x_j\right)/N}{B},i=1,2,\cdots,B$

(7) 得到样本数据集的经验分布$\tilde{P}(X,Y)$。

(8) 利用最大熵模型进行模型优化。$f_i\left(x,y\right)$是特征函数，w_i是特征的权重。

(9) for each i do

(10) δ_i是方程$\sum_{x,y}\tilde{P}(x)P_w\left(y|x\right)f_i\left(x,y\right)\exp\left(\delta_i f^{\#}\left(x,y\right)\right)=E_{\tilde{P}}\left(f_i\right)$的解。

更新w_i值：$w_i\leftarrow w_i+\delta_i$直至收敛。

(11) end for;

(12) 产生新样本。

(13) $\Delta=N_{\text{maj}}-N_{\text{min}}$定义为新样本的生成数。

(14) n代表少数类样本数。

(15) for $s=1$ to n do

(16) 计算ω_{x_s}

(17) 计算每个少数类样本生成新样本的数量$S_{x_s}=\left[\omega_{x_s}\cdot\Delta\right]$，其中[ ]为取整函数。

(18) if $S_{x_s}=0$ then $S_{x_s}=1$。

(19) else if $\sum_{s=1}^{n}S_{x_s}>\Delta$ then

按少数类样本的概率降序取前Δ个生成的新的少数类样本。

(20) else if $\sum_{s=1}^{n}S_{x_s}<\Delta$ then

继续按少数类样本的概率降序生成新的少数类样本直至所有新的少数类样本数量为Δ。

(21) end if

(22) end for;

(23) while 数据非平衡 do
(24) for $s=1$ to n do
(25) 计算x_i;
(26) 计算x_s^k;
(27) end for
(28) end while

3 算例分析

为了评估本文提出的新算法,进行了实验研究。首先,介绍实验的数据集、分类算法,然后给出实验的结果并进行算法比较,最后对算法的时间复杂度进行分析。

3.1 数据集

在实验中,选择8组来自UCI和KEEL数据库中的非平衡数据集,如表1所示。这些数据在大小、功能、数量和不平衡比率(IR)上各不相同,这样保证了全面的性能评估。由于只考虑二分类问题,所以对原始数据进行了转换,使它们成为两个类别的数据。不平衡比率的计算公式为

$$\mathrm{IR}=\frac{\text{多数类样本数量}}{\text{少数类样本数量}} \tag{15}$$

表1 非平衡数据集的基本信息

Table 1 Basic information of unbalanced datasets

数据集	样本数量	属性个数	IR
Wine_1_vs_2	130	13	1.20
Data_banknote	1 372	4	1.25
Ecoli_cp_vs_im	220	7	1.86
Pima	768	8	1.87
Yeast1	1 484	8	2.46
Glass_2_vs_7	105	9	2.72
New_thyroid	180	5	5.00
yeast-0-5-6-7-9_vs_4	528	8	9.35

3.2 评估指标

现在考虑一个两类分类问题,其中结果被标记为正的或负的。给定一个由多数类样本和少数类样本组成的测试数据集,任何分类模型的任务都是为每个样本分配一个类标签。假定少数类为正多数类为负,则从一个二值分类器中可以获得4种可能的结果:真阳性TP、假阳性FP、真阴性TN和假阴性FN。如果预测的结果是正的,实际值也是正的,那么它被称为真阳性。同样地,当预测结果和实际值均为负时则为真阴性。当预测结果为负而实际值为正时,则为假阴性。如果预测结果为正而实际值为负时,则它被视为假阳性。由于准确率(Acc)不能单一地作为评估非平衡数据分类问题的指标。因此,采用AUC、Acc、Sen、Spe、Pre和F_1值来对算法的分类性能进行评估,它们通过表2的混淆矩阵来计算。具体计算公式为

表2 混淆矩阵

Table 2 Confusion matrix

真实值	预测值	
	少数类	多数类
少数类	TP	FN
多数类	FP	TN

$$\mathrm{AUC}=\frac{1+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}-\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}}{2} \tag{16}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{TN}} \tag{17}$$

$$\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{18}$$

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{19}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{20}$$

$$F_1=\frac{2\cdot\mathrm{Sen}\cdot\mathrm{Pre}}{\mathrm{Pre}+\mathrm{Sen}} \tag{21}$$

3.3 分类算法

所关注的是过采样算法的比较,因此没有对分类器进行比较。在实验中,选取了决策树分类器作为实验中的分类器,通常决策树主要有3种实现,分别是ID3算法、CART算法[43]和C4.5算法。其中CART算法称为分类回归树算法,它是决策树生成的一种有效算法。CART算法是一种二分递归分割技术,把当前样本划分为两个子样本,使得生成的每个非叶子结点都有两个分支,因此CART算法生成的决策树是结构简洁的二叉树。因为感兴趣的是每种过采样算法的相对性能,而不是它们的绝对性能,所以在实验中采取了CART算法的默认参数进行试验。

3.4 实验结果

在理想的情况下基于训练数据构建的分类模型应该表示概率分布$P(Y|X)$的"真实模型"。然而在实际的应用中,由于训练集与真实概率分布之间的偏差,很难得到真实模型。因此从训练数据集归纳出的分类模型通常是近似的"真实模型"。在本实验中,新算法对上述8组非平衡数据进行了过采样,而过采样的关键在于采样样本的选取,所以获取被采样样本的概率值至关重要。表3即为少数类样本在新算法优化后得出的概率分布的概率值$p(\mathrm{class1}|x_i)$,其中class1代表少数类,x_i为每个数据集的少数类样本。

表3 少数类概率分布概率值

Table 3 Minority class probability distribution probability value

数据集	概率值							
Wine_1_vs_2	0.751	0.784	0.718	0.612	…	0.720	0.738	0.678
Data_banknote	0.579	0.561	0.563	0.547	…	0.561	0.550	0.549
Ecoli_cp_vs_im	0.503	0.373	0.612	0.526	…	0.595	0.634	0.580
Pima	0.535	0.526	0.538	0.458	…	0.571	0.618	0.432
Yeast1	0.388	0.435	0.417	0.436	…	0.411	0.405	0.419
Glass_2_vs_7	0.483	0.545	0.636	0.617	…	0.642	0.696	0.603
New_thyroid	0.582	0.485	0.508	0.491	…	0.523	0.537	0.412
yeast-0-5-6-7-9_vs_4	0.317	0.333	0.338	0.313	…	0.283	0.270	0.289

3.5 算法比较

为了进一步评估新算法的性能,将本文提出的新算法与其他5种算法(ADASYN, SMOTE, Borderline1-SMOTE, Borderline2-SMOTE和ORIGINAL)在AUC、Acc、Sen、Spe、Pre和F_1值评价指标中进行比较,需要强调的是,本文中所有涉及到的算法都是通过python代码实现的。在比较算法中,ADASYN、SMOTE、Borderline1-SMOTE、Borderline2-SMOTE这4种算法都使用了python包imbalaned-learn 0.8.1版本的默认参数。ORIGINAL算法保持原始非平衡数据不变,直接对非平衡数据进行分类。图1直观地展示了各种算法评价指标比较情况。

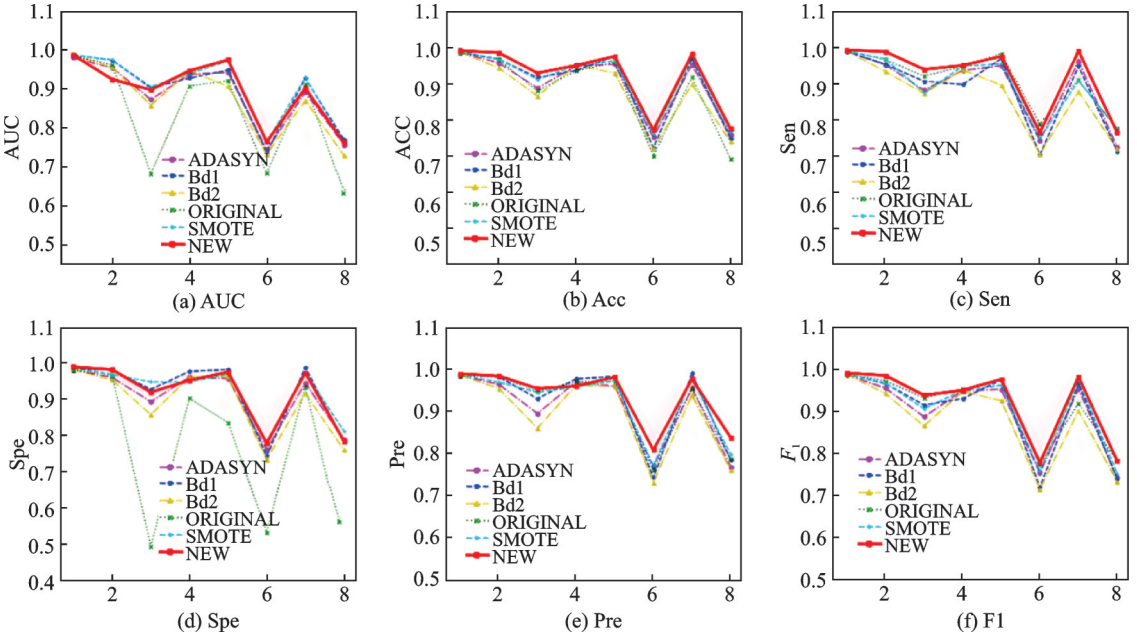

图1 概率增强算法评价指标曲线图

Fig.1 Evaluation index curves of probabilistic enhancement algorithms

AUC是一种常用的分类模型评价指标,用于衡量分类器的性能。AUC的取值范围在0到1之间,值越大表示分类器的性能越好。6种算法的AUC对比如表4所示,表中Data 1~Data 8分别对应数据Data_banknote、Ecoli_cp_vs_im、yeast-0-5-6-7-9_vs_4、Glass_2_vs_7、New_thyroid、Pima、Wine_1_vs_2、Yeast1。通过表4对比实验数据的AUC评价指标,证明了新算法在多个数据集上的有效性。实验结果表明,本文算法在Data 4、Data 5和Data 6数据集上的AUC值分别为0.945、0.973和0.762。虽然本文算法与SMOTE算法的获胜次数相同,但是两者相比较,本文算法要优于SMOTE算法。这表明本文算法能够有效地提高分类器的性能,在与之相比较的过采样算法中占优。

表4 基于决策树分类器的AUC比较结果

Table 4 AUC comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.978	0.953	0.871	0.936	0.941	0.741	0.890	0.754
Borderline1	**0.986**	0.972	0.901	0.926	0.947	0.735	0.926	**0.767**
Borderline2	0.985	0.954	0.856	**0.945**	0.905	0.730	0.868	0.727
ORIGINAL	0.984	0.960	0.680	0.906	0.919	0.681	0.909	0.631
SMOTE	0.985	**0.973**	**0.904**	0.938	0.970	0.761	**0.928**	0.759
本文算法	0.985	0.923	0.896	**0.945**	**0.973**	**0.762**	0.900	0.760

Acc指标是分类器分类准确率的一种度量方式,它表示分类器在所有样本中正确分类的比例。在本文中,Acc指标被用来评估不同算法在非平衡数据分类问题上的表现。6种算法的Acc对比如表5所示。从表5中可以看出,本文算法在大多数数据集上的Acc指标都比其他算法高,特别是在Data 2和Data 7上,本文算法的Acc指标分别达到了0.983和0.979,远高于其他方法。同时,在其他算法中,

表5 基于决策树分类器的Acc比较结果

Table 5 Acc comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.983	0.955	0.885	0.947	0.951	0.749	0.950	0.755
Borderline1	0.984	0.965	0.914	0.934	**0.973**	0.720	0.966	0.745
Borderline2	0.985	0.941	0.863	**0.948**	0.927	0.716	0.895	0.736
ORIGINAL	0.984	0.964	0.877	0.933	0.956	0.697	0.915	0.689
SMOTE	0.986	0.965	0.909	**0.948**	0.960	0.758	0.972	0.762
本文算法	**0.989**	**0.983**	**0.927**	**0.948**	**0.973**	**0.769**	**0.979**	**0.772**

SMOTE算法在大多数数据集上也表现出了较好的Acc指标。

Sen指标是分类器正确预测为正类的样本数占实际为正类的样本数的比例。表6比较了不同过采样算法和本文算法在8个数据集上分类器的敏感度表现。从表中可以看出,本文算法在大多数数据集上的Sen指标都比其他方法高,特别是在Data 1、Data 2和Data 7上,本文算法的Sen指标分别达到了0.991、0.986和0.988,远高于其他方法。这表明本文算法在处理非平衡数据分类问题上具有很好的效果,能够更好地识别出少数类样本。

表6 基于决策树分类器的Sen比较结果

Table 6 Sen comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.987	0.950	0.879	0.934	0.947	0.740	0.959	0.722
Borderline1	0.986	0.951	0.904	0.895	0.967	0.700	0.946	0.710
Borderline2	0.988	0.931	0.870	0.936	0.893	0.702	0.875	0.716
ORIGINAL	0.988	0.964	0.918	0.945	**0.980**	**0.786**	0.907	**0.773**
SMOTE	0.987	0.966	0.872	0.946	0.953	0.746	0.905	0.715
本文算法	**0.991**	**0.986**	**0.937**	**0.948**	0.973	0.762	**0.988**	0.762

Spe指标是指分类器在预测负样本时的准确率,即真实负样本中被正确预测为负样本的比例。表7列出了基于决策树分类器的不同算法在8个数据集上的Spe指标比较结果。从表中可以看出,本文算法在大多数数据集上都取得了最好的结果,特别是在Data 1和Data 2上,其Spe指标分别达到了0.987和0.980,远高于其他方法。这表明本文算法在处理非平衡数据分类问题时具有较好的性能。同时,可以发现在不同数据集上,不同的算法表现不同。例如,在Data 3和Data 8上,SMOTE方法表现最好,而在Data 6上,ADASYN方法表现最好。这说明在实际应用中,需要根据具体数据集的特点选择合适的数据处理方法。

Pre指标是指分类器在预测时正确分类的样本数占总样本数的比例。表8展示了基于决策树分类器的不同算法在预测准确率方面的比较结果。从表中可以看出,本文算法在大部分数据集上的Pre值均高于其他数据处理方法,表明本文算法在分类器预测方面具有更高的准确性。具体来说,本文算法在Data 1、Data 2、Data 3、Data 6和Data 8上的Pre值分别为0.987、0.982、0.952、0.806和0.836,均高于其他数据处理方法。而在Data 4、Data 5和Data 7上,本文算法的Pre值分别为0.958、0.979和0.975,略低于Borderline1方法。综上所述,基于决策树分类器的实验结果表明,本文算法在预测准确率方面具有较高的优势,可以有效地提高分类器的预测准确性。

表7 基于决策树分类器的Spe比较结果

Table 7 Spe comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.979	0.958	0.891	0.959	0.955	0.758	0.943	0.785
Borderline1	0.982	0.979	0.925	**0.975**	**0.980**	0.740	**0.986**	0.780
Borderline2	0.980	0.951	0.855	0.959	0.960	0.730	0.914	0.757
ORIGINAL	0.979	0.963	0.490	0.900	0.833	0.529	0.933	0.480
SMOTE	0.984	0.966	**0.946**	0.946	0.967	0.770	0.971	**0.810**
本文算法	**0.987**	**0.980**	0.917	0.950	0.973	**0.776**	0.971	0.782

表8 基于决策树分类器的Pre比较结果

Table 8 Pre comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.983	0.962	0.893	0.968	0.957	0.767	0.953	0.766
Borderline1	0.986	0.980	0.928	**0.975**	**0.981**	0.740	**0.988**	0.782
Borderline2	0.984	0.952	0.858	0.960	0.957	0.729	0.936	0.759
ORIGINAL	0.983	0.980	0.945	0.968	0.970	0.758	0.951	0.785
SMOTE	0.984	0.967	0.943	0.956	0.971	0.769	0.975	0.796
本文算法	**0.987**	**0.982**	**0.952**	0.958	0.979	**0.806**	0.975	**0.836**

F_1值是评估分类器性能的重要指标,它综合了分类器的精确度和召回率。表9展示了多种过采样算法的F_1值比较结果,比较了5种不同的算法和本文提出的算法在8个数据集上的F_1值。从表中可以看出,本文算法几乎在所有数据集上表现出色,其F_1值比其他方法高。特别是在Data 1、Data 2和Data 7数据集上,本文算法的F_1值分别为0.989、0.983和0.980,明显高于其他方法。这表明本文算法在这些数据集上具有更好的分类性能。此外,还可以看到,ADASYN和Borderline2方法在某些数据集上表现不佳,其F_1值较低。这可能是因为这些方法在生成合成样本时过度依赖于少数类样本,导致过拟合。因此,本文算法在大多数数据集上表现出色,具有更好的分类性能。

表9 基于决策树分类器的F_1值比较结果

Table 9 F_1-value comparision results based on decision tree classifier

算法	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8
ADASYN	0.985	0.953	0.885	0.947	0.950	0.752	0.951	0.741
Borderline1	0.986	0.964	0.913	0.928	0.973	0.717	0.964	0.736
Borderline2	0.986	0.940	0.863	0.945	0.923	0.713	0.899	0.731
ORIGINAL	0.986	0.970	0.930	**0.951**	**0.974**	0.771	0.917	0.778
SMOTE	0.986	0.965	0.904	0.948	0.960	0.756	0.972	0.750
本文算法	**0.989**	**0.983**	**0.936**	0.948	**0.974**	**0.776**	**0.980**	**0.780**

表10统计了不同的算法在各种评价指标上的获胜表现,包括AUC、Acc、Sen、Spe、Pre和F_1。其中,ADASYN算法在所有评价指标上都没有获胜,即表现最差。Borderline1算法在AUC、Acc、Spe和Pre指标上获胜,是除本文算法以外获胜次数最多的算法。Borderline2算法仅在AUC和Acc指标上获

胜且获胜次数较少。ORIGINAL算法在Sen和F1指标上获胜,但在其他指标上表现较差。SMOTE算法在AUC、Acc和Spe指标上获胜,也是在其他指标上表现较差。本文算法在所有评价指标上都获得了最多的胜利次数,表现最好。总体来说,本文算法在分类任务中表现优异,具有较高的应用价值。

表10 获胜次数比较结果

Table 10 Wining time comparision results

算法	AUC	Acc	Sen	Spe	Pre	F_1	Total
ADASYN	0	0	0	0	0	0	0
Borderline1	2	1	0	3	3	0	9
Borderline2	1	1	0	0	0	0	2
ORIGINAL	0	0	3	0	0	2	5
SMOTE	3	1	0	2	0	0	6
本文算法	3	8	5	3	5	7	31

3.6 算法时间复杂度

在新算法中,假设训练的样本数为n,则自助法的时间复杂度为$O(n^2)$,其次对最大熵优化进行分析,改进的迭代尺度法的时间复杂度为$O(n^2)$,迭代1 000次的时间复杂度为$O(1)$,最后在生成新样本时其时间复杂度为$O(sn)$,s代表生成新样本数,因此新算法总的时间复杂度为$O(n^2+sn+1)$。

用s表示生成新样本数,x表示原始数据少数类样本数,则SMOTE算法的时间复杂度为$O(sx)$,SMOTE-Borderlinel算法的时间复杂度为$O(x+kx+1)$,SMOTE-Borderline2算法的时间复杂度为$O(x+kx+1)$。ADASYN算法的时间复杂度为$O(kx+sx+1)$,k代表近邻数。

综上所述,时间复杂度最大的是本文提出的过采样算法,最小的是SMOTE-Borderline系列,新算法虽然时间复杂度较大,但是根据表10中新算法的获胜次数来看,牺牲时间复杂度换来了性能上的提升是值得的。

4 结束语

为了解决类别分类不平衡的问题,研究人员提出了许多基于抽样的预处理方法。一般这些方法的基本原理是通过具体的策略对非平衡数据进行再平衡,为了提高少数类数据的生成质量,本文提出一种基于自助法在最大熵原理下的概率增强算法。新算法的主要优点是与同样是基于统计特征的GMM-EM算法相比较而言,对原始数据分布没有限制,降低了算法对数据分布的要求;其次,新算法与其他基于几何特征的算法相比较在提高分类性能上更加出色。然而,新算法与传统方法相比,分类效果虽然得到了提高,但同时也存在不足之处。一方面,新算法在基于Bootstrap方法下,样本容量过大会增加算法的时间复杂度;另一方面,新算法没有考虑数据的几何特征。所以,将考虑用其他方法降低时间复杂度并将随机方法与几何特征相结合起来进行研究,这也是今后需要继续研究的方向。

参考文献:

[1] HE H B, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[2] SUN Y M, WONG A, KAMEL M S. Classification of imbalanced data: A review[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4): 687-719.

[3] DIZ J, MARREIROS G, FREITAS A. Applying data mining techniques to improve breast cancer diagnosis[J]. Journal of Medical Systems, 2016, 40(9): 1-7.

[4] YUEAI Z, JUNJIE C. Application of unbalanced data approach to network intrusion detection[C]//Proceedings of 2009 First International Workshop on Database Technology and Applications. [S.l.]: IEEE, 2009: 140-143.

[5] JIA J H, LIU Z, XIAO X, et al. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset[J]. Analytical Biochemistry, 2016, 497: 48-56.

[6] WEI W, LI J J, CAO L B, et al. Effective detection of sophisticated online banking fraud on extremely imbalanced data[J]. World Wide Web-Internet and Web Information Systems, 2013, 16(4): 449-475.

[7] ATTABI Y, DUMOUCHEL P. Anchor models for emotion recognition from speech[J]. IEEE Transactions on Affective Computing, 2013, 4(3): 280-290.

[8] LONGADGE R, DONGRE S. Class imbalance problem in data mining review[J]. International Journal of Computer Science & Network, 2013, 2(1):1-5.

[9] ASADI S, SHAHRABI J. ACORI: A novel ACO algorithm for rule induction[J]. Knowledge-Based Systems, 2016, 97: 175-187.

[10] ASADI S, SHAHRABI J. Complexity-based parallel rule induction for multiclass classification[J]. Information Sciences, 2017, 380: 53-73.

[11] TAHAN M H, ASADI S. EMDID: Evolutionary multi-objective discretization for imbalanced datasets[J]. Information Sciences, 2018, 432: 442-461.

[12] TAHAN M H, ASADI S. MEMOD: A novel multivariate evolutionary multi-objective discretization[J]. Soft Computing, 2018, 22(1): 301-323.

[13] 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: 321-357.

[14] COHEN G, HILARIO M, SAX H, et al. Learning from imbalanced data in surveillance of nosocomial infection[J]. Artificial Intelligence in Medicine, 2006, 37(1): 7-18.

[15] NAPIERAŁA K, STEFANOWSKI J, WILK S. Learning from imbalanced data in presence of noisy and borderline examples [C]//Proceedings of International Conference on Rough Sets and Current Trends in Computing. Berlin: Springer, 2010: 158-167.

[16] RIVERA W A. Noise reduction a priori synthetic over-sampling for class imbalanced data sets[J]. Information Sciences, 2017, 408: 146-161.

[17] DOUZAS G, BACAO F. Self-organizing map oversampling (SOMO) for imbalanced data set learning[J]. Expert Systems with Applications, 2017, 82: 40-52.

[18] ELREEDY D, ATIYA A F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance[J]. Information Sciences, 2019, 505: 32-64.

[19] SHIN K, HAN J, KANG S. MI-MOTE: Multiple imputation-based minority oversampling technique for imbalanced and incomplete data classification[J]. Information Sciences, 2021, 575: 80-89.

[20] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Proceedings of Conference on Artificial Intelligence in Medicine in Europe. Berlin: Springer, 2001: 63-66.

[21] BATISTA G E, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[22] YEN S J, LEE Y S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset[M]. Berlin: Springer, 2006: 731-740.

[23] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2009, 39(2): 539-550.

[24] PRUSA J, KHOSHGOFTAAR T M, DITTMAN D J, et al. Using random undersampling to alleviate class imbalance on tweet sentiment data[C]//Proceedings of 2015 IEEE International Conference on Information Reuse and Integration. [S.l.]: IEEE, 2015: 197-202.

[25] CATENI S, COLLA V, VANNUCCI M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems[J]. Neurocomputing, 2014, 135: 32-41.

[26] JINDALUANG W, CHOUVATUT V, KANTABUTRA S. Under-sampling by algorithm with performance guaranteed for class-imbalance problem[C]//Proceedings of 2014 International Computer Science and Engineering Conference (ICSEC). [S.l.]: IEEE,

2014: 215-221.

[27] CAO L, SHEN H. Combining re-sampling with twin support vector machine for imbalanced data classification[C]//Proceedings of 2016 17th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). [S.l.]: IEEE, 2016: 325-329.

[28] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]//Proceedings of International Conference on Intelligent Computing. Berlin: Springer, 2005: 878-887.

[29] BUNKHUMPORNPAT C, SINAPIROMSARAN K, LURSINSAP C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 475-482.

[30] LEE H, KIM J, KIM S. Gaussian-based smote algorithm for solving skewed class distributions[J]. International Journal of Fuzzy Logic and Intelligent Systems, 2017, 17(4): 229-234.

[31] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences, 2019, 501: 118-135.

[32] CHEN B Y, XIA S Y, CHEN Z, et al. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise [J]. Information Sciences, 2021, 553: 397-428.

[33] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). [S.l.]: IEEE, 2008: 1322-1328.

[34] TANG B, HE H. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning[C]//Proceedings of 2015 IEEE Congress on Evolutionary Computation (CEC). [S.l.]: IEEE, 2015: 664-671.

[35] 陈刚, 吴振家. 一种基于GMM-EM的非平衡数据的概率增强算法[J]. 控制与决策, 2020, 35(3): 763-768.
CHEN Gang, WU Zhenjia. An enhancing probability algorithm for imbalanced datasets based on GMM-EM[J]. Control and Decision, 2020, 35(3): 763-768.

[36] RIVERA W A, XANTHOPOULOS P. A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets[J]. Expert Systems with Applications, 2016, 66: 124-135.

[37] DONG Q, GONG S G, ZHU X T. Imbalanced deep learning by minority class incremental rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1367-1381.

[38] TAO X M, LI Q, GUO W J, et al. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification[J]. Information Sciences, 2019, 487: 31-56.

[39] ZHAI J H, QI J X, SHEN C. Binary imbalanced data classification based on diversity oversampling by generative models[J]. Information Sciences, 2022, 585: 313-343.

[40] EFRON B. Bootstrap methods: Another look at the jackknife[J]. The Annals of Statistics, 1979, 7(1): 1-26.

[41] JAYNES E T. Information theory and statistical mechanics[J]. Physical Review, 1957, 106(4): 620.

[42] BERGER A. The improved iterative scaling algorithm: A gentle introduction[EB/OL]. [2021-10-06].http://luthuli.cs.uiuc.edu/~daf/courses/optimization/papers/berger-iis.pdf.

[43] LI B, FRIEDMAN J, OLSHEN R, et al. Classification and regression trees (CART)[J]. Biometrics, 1984, 40(3): 358-361.

作者简介:

雷天纲(1994-),男,硕士研究生,研究方向:数据挖掘,机器学习,E-mail: Itg0302@163.com。

陈刚(1964-),**通信作者**,男,教授,研究方向:数据挖掘、机器学习、复杂系统的建模与计算,E-mail: chengang@dlmu.edu.cn。

(编辑:夏道家)