基于三支决策的不平衡数据过采样方法

胡 峰,王 蕾,周 耀
(计算智能重庆市重点实验室(重庆邮电大学),重庆 400065)

摘 要: 采样是解决不平衡数据分类问题的一个有效途径. 文中结合三支决策理论,根据样本分布将样本划分成三个区域:正域、边界域和负域;在此基础上,分别对边界域和负域中的小类样本进行不同的过采样处理,提出了一种基于三支决策的不平衡数据过采样算法(TWD-IDOS 算法). 实验结果表明,在 C4.5、KNN 和 CART 等分类器上,文中提出的算法能有效解决不平衡数据的二分类问题,在 Recall、F-value、AUC 等指标上优于文献中的过采样算法.

关键词: 三支决策; 邻域粗糙集; 边界采样; 不平衡数据; SMOTE
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2018)01-0135-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.01.019

An Oversampling Method for Imbalance Data Based on Three-Way Decision Model

HU Feng, WANG Lei, ZHOU Yao
(*Chongqing Key Laboratory of Computational Intelligence* (*Chongqing University of Posts and Telecommunications*), *Chongqing* 400065, *China*)

Abstract: Sampling is an effective way to solve the problem of unbalanced data classification. According to the distribution of samples, we employ the three-way decision model to divide the universe into three parts: positive region, boundary region and negative region. After that, we oversample the minority class samples in boundary region and negative region respectively. Then, a novel oversampling algorithm for imbalance data based on three-way decision model, namely TWD-IDOS, is developed. The experimental results show that the proposed method can effectively solve the two-class classification problems of imbalanced data and has a better performance in such measures (Recall、F-value、AUC) on C45, KNN and CART classifiers than other oversampling methods.

Key words: three-way decision; neighborhood rough set; boundary sampling; imbalanced data; SMOTE

1 引言

不平衡数据集的分类问题是机器学习和模式识别领域中的一个热点问题,迄今为止,针对此问题的解决方法主要分为两大类:一类是数据预处理方法,目的是降低类别之间的不平衡性,在此层面上主要的方法是重采样,增加小类样本的数目(过采样)或减少大类样本的数目(欠采样);另一类则是在分类算法上着手,提出新的有效的分类算法或改进现有的分类算法以适应对不平衡数据分类的目的,主要包括单类学习、集成学习、代价敏感学习等方法. 但是它们没有改变类别之间的不平衡性,限制了算法的广泛应用. 所以在实际应用场景中,采用更多的是数据采样的方法.

过采样的本质就是通过各种手段增加小类样本数目,最简单的方法就是随机复制小类样本,但这种采样容易导致过拟合[1]. 因此,许多学者提出了一些更高效的过采样方法. 最具代表性的则是 SMOTE[2] 算法,该算法是 Chawla 等人提出的一种简单有效的智能过采样方法,能够有效避免分类器的过拟合现象[3]. 但是,SMOTE 算法对每个小类样本的采样存在一定的盲目性,导致有些合成的小类样本影响大类样本的泛化空间,从而降低其分类效果. 针对 SMOTE 在过采样过程中存在的问题,许多学者提出了不同的改进方法. 比如:Borderline-SMOTE 方法[4],只对边界点采样,在一定程度上,避免了合成冗余样本;ASMOTE 方法[5],考虑了大类样本的分布信息,避免了新合成的小类样本落在大

收稿日期:2016-05-10;修回日期:2016-10-31;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61309014, No. 61379114, No. 61472056);教育部人文社科规划(No. 15XJA630003);重庆市基础与前沿研究计划(No. cstc2013jcyjA40063, No. cstc2014jcyjA40049);重庆市教委科学技术研究(No. KJ1500416)

类样本的近邻区域;SMOTE-RSB * 方法[6],结合了粗糙集理论是一种混合采样方法,通过筛选,把影响大类样本泛化空间的新合成的小类样本剔除,从而保证了大类样本的识别率;KSMOTE[7]对 SMOTE 算法进行扩展,通过在特征空间中合成新样本,以解决不同空间处理训练样本所带来的不一致问题,提高所合成样本的质量;OSLDD SMOTE[8]通过单边选择链遴选出处于分类边界的小类样本,根据这些样本的动态分布密度生成新样本,有效提高了小类样本的分类准确率. 总之,对不平衡数据的处理,有效的过采样方法既要增加小类样本的分布信息,又要尽可能的避免小类样本对大类样本决策空间的影响.

三支决策理论是由加拿大学者 Yao 首次提出,主要思想就是将整体划分为三个部分,分别称为 L 域、M 域和 R 域. 分别对这三个域采用不同的处理方法,为复杂问题的求解提供了一种有效的策略与方法. 近年来,众多学者都在思考怎样将三支决策思想转换为一个理论系统、信息处理模式、和计算方法. 关于三支决策理论的应用研究获得了一定的进展,如:王磊等[9],提出了基于主题特征与三支决策理论相融合的多标记情感分类方法;Li 等[10]提出了基于三支决策的代价敏感人脸识别方法;Liu 等[11]提出了基于 logistics 回归的多分类三支决策方法;Yu 等[12]针对聚类学习中类与类之间的重叠问题,提出了基于三支决策的重叠聚类方法;Liu 等[13]将三支决策理论应用在语义分析上;Liu 等[14]将三支决策理论应用在不完备信息系统;Chen 等[15]将三支决策应用在邻域系统用来做约减;Liu[16]和 Zhou[17]结合决策粗糙集理论,给出了一种多分类问题的解决途径. 若将三支决策理论应用到不平衡数据处理方面,利用它的三个域,将训练集划分三个部分,对每一部分采用不同的采样方法,有可能是一条不平衡数据采样的有效途径.

本文结合三支决策理论,提出了一种不平衡数据的过采样算法(TWD-IDOS 算法). 首先,利用三支决策理论将样本总体划分成正域样本、边界域样本和负域样本. 其次,保留正域样本,对正域中的样本不做采样处理. 再次,对边界域和负域中的小类样本分别进行过采样处理:①对边界域中小类样本进行 SMOTE 过采样,然后对新合成的样本筛选;②对负域小类样本进行一种有别于边界域采样的过采样处理,最后得到采样后的新样本集,通过采样,能有效解决不平衡数据的二分类问题.

2 相关概念

2.1 邻域模型

1988 年,Lin T Y[18]提出了邻域模型,该模型通过空间点的邻域来粒化论域空间,将邻域理解为基本信息粒子,用来描述空间中的其他概念.

定义 1[19] 给定任意 $x_i \in U, B \subseteq C, x_i$ 在属性子集 B 上的邻域 $\delta_B(x_i)$ 定义为:

$$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta_B(x_i, x_j) \leqslant \delta\} \tag{1}$$

这里 δ 为度量函数. 定义 x_1, x_2 为两个 N 维空间样本 $A = \{a_1, a_2, \cdots, a_N\}$, $f(x, a_i)$ 表示 x 在属性 a_i 上值,则两个样本的 Minkowsky 距离可定义为:

$$\Delta_p(x_1, x_2) = \left(\sum_{i=1}^{N} | f(x_1, a_i) - f(x_2, a_i) |^p\right)^{1/p} \tag{2}$$

当 $p = 2$ 时,即欧拉距离.

欧拉距离只适用计算连续型属性,无法计算分类型属性. 对分类型属性的计算,Stanfill 和 Waltz[20]提出 Value Dierence Metric. 假设样本 x_1, x_2 在分类型属性的两个值 V_1, V_2,他们之间的距离定义为:

$$f(x_1, V_1) - f(x_2, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^{\mu} \tag{3}$$

C_1 是所有样本中该属性值为 V_1 的个数,C_{1i} 则为其中类别为 i 个数,C_2 是所有样本中该属性值为 V_2 的个数,C_{2i} 则为其中类别为 i 个数. μ 为常数,通常是 1.

2.2 邻域三支决策模型

Yao 在粗糙集和决策粗糙集理论的基础上提出了三支决策理论,该理论为粗糙集的三个域提供了合理的语义解释. 该理论是一种信息不确定或不完整的条件下进行决策的方法. Yao 在文献[21]中给出了有关三支决策问题的形式化定义.

定义 2 给定实数空间上的非空有限样本集合 U, $\forall x \in U, x$ 的邻域可表示为 $\delta(x) = \{y \mid y \in U, \Delta(x, y) \leqslant \delta\}$. 令 N_0 和 N_1 分别表示 $\delta(x)$ 内的大类样本类别和小类样本类别,则 x 的邻域 $\delta(x)$ 内大类样本个数和小类样本个数可分别定义为:

$$N_0(\delta(x)) = |\{y \mid y \in \delta(x), y \in N_0\}| \tag{4}$$

$$N_1(\delta(x)) = |\{y \mid y \in \delta(x), y \in N_1\}| \tag{5}$$

根据文献[22],为了实现三支决策:首先,需要引入实体的评价函数 $f(x)$,也称为决策函数,它的值称为决策状态值,其大小反映实体的好坏程度;其次,引入一对阈值 α 和 β 来定义正域、边界域和负域中的事件对象;再次,根据决策状态值和阈值将论域中事件对象划分到正域、边界域和负域中,构造出相应的三支决策规则. 本文结合邻域模型与三支决策模型,给出了邻域三支决策模型的相关定义.

定义 3 给定实数空间上的非空有限样本集合 $U = \{x_1, x_2, \cdots, x_n\}$, $\forall x \in U$,给定目标函数 $f(x)$,则邻域三支决策如下:

$$\begin{aligned} &(\mathrm{P})\ \text{如果 } f(x) \geqslant \alpha, \text{则 } x \in \mathrm{POS}(X) \\ &(\mathrm{B})\ \text{如果 } \beta < f(x) < \alpha, \text{则 } x \in \mathrm{BND}(X) \\ &(\mathrm{N})\ \text{如果 } f(x) \leqslant \beta, \text{则 } x \in \mathrm{NEG}(X) \end{aligned} \tag{6}$$

式(6)中,$\alpha=k,\beta=\frac{-k}{k+1}$. 其中,$k$ 表示在样本 x 的邻域内进行采样的样本个数,本文参考 SMOTE 算法[2]和实验经验结果,取 $k=5$,即,$\alpha=k=5,\beta=\frac{-k}{k+1}=-\frac{5}{6}$.

决策(P)表示当$f(x)$不小于 α 时,将 x 划分到 X 正域;决策(B)表示当$f(x)$大于 β 且小于 α 时,将 x 划分到边界域;决策(N)表示当$f(x)$不大于 β 时,将 x 划分到负域.

式(6)中,$f(x)$的计算公式如式(7)所示:

$$f(x)=\begin{cases}\dfrac{N_1(\delta(x))-N_0(\delta(x))}{N_0(\delta(x))+1}, & x\in X_{\min}\\ \dfrac{N_0(\delta(x))-N_1(\delta(x))}{N_1(\delta(x))+1}, & x\in X_{\max}\end{cases}\tag{7}$$

式(7)中,$f(x)$表示:在样本 x 的邻域内,不同类别样本个数差与异类样本个数的一种比例关系. 为了避免出现分母为 0 的情况,这里采用了对 $N_0(\delta(x))$ 和 $N_1(\delta(x))$加 1 的处理策略. 当 $x\in X_{\min}$时,$f(x)$越大,说明 x 邻域内的小类样本多,x 属于正域的可能性越大;当 $x\in X_{\max}$时,$f(x)$越小,说明 x 邻域内的大类样本越多,x 属于负域的可能性越大.

为了进一步解释邻域三支决策模型,我们给出一个例子(如图 1 所示).

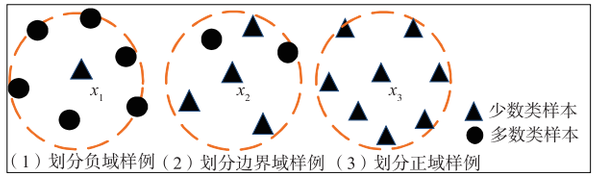

图1　划分数据集原理图

在图 1 中,$\alpha=k=5,\beta=\frac{-k}{k+1}=-\frac{5}{6}$,分 3 种示例分别解释负域、边界域和正域.

示例(1):x_1 的邻域范围内小类个数为 0,大类个数为 6. 即 $N_0(\delta(x_1))=6,N_1(\delta(x_1))=0$;评价函数 $f(x_1)=\frac{0-6}{6+1}=-\frac{6}{7}<-\frac{5}{6}$,即 $f(x_1)<\beta$,则 $x_1\in \mathrm{NEG}(X)$,即把 x_1 划分到负域.

示例(2):x_2 的邻域范围内小类个数为 3,大类个数为 2. 即 $N_0(\delta(x_2))=2,N_1(\delta(x_2))=3$,评价函数 $f(x_2)=\frac{3-2}{2+1}=\frac{1}{3}$,即 $\beta<f(x_2)<\alpha$,则 $x_2\in \mathrm{BND}(X)$,即把 x_2 划分到边界域.

示例(3):x_3 的邻域范围内小类个数为 7,大类个数为 0. 即 $N_0(\delta(x_3))=0,N_1(\delta(x_3))=7$,评价函数 $f(x_3)=\frac{7-0}{0+1}=7$,即 $f(x_3)>\alpha$,则 $x_3\in \mathrm{POS}(X)$,即把 x_3 划分到正域.

3　基于三支决策的不平衡数据的过采样算法

在过采样过程中,容易产生冗余数据,影响大类的泛化空间,且导致采样倍率不足. 为解决这些问题,本文提出了一种基于三支决策的不平衡数据的过采样算法. 该算法结合邻域三支决策模型,从样本总体分布来定义正域样本、边界域样本和负域样本,并对边界域样本和负域样本进行相应的过采样操作,从而实现对整个训练样本集的采样.

3.1　算法思路

3.1.1　确定样本的邻域半径

对训练集进行划分的一个重要前提是确定样本的邻域半径. 如果邻域半径太大,则会导致所有小类样本都被划分为边界域样本,容易合成冗余数据. 若邻域半径太小,则会导致小类样本的邻域范围内全部是大类样本,划分过程中就会把这些小类样本划分到负域中,达不到边界采样的效果.

在基于三支决策的不平衡数据的过采样算法中计算邻域半径 δ 的方法[19]:

$$\delta=\min(\Delta(x_i,s))+w\times\mathrm{range}(\Delta(x_i,s)),0\leqslant w\leqslant 1\tag{8}$$

其中,$\min(\Delta(x_i,s))$表示在训练集中样本 s 和距离其最近样本点之间的距离,$\mathrm{range}(\Delta(x_i,s))$表示在训练集中样本 s 与其他样本之间距离的取值范围.

3.1.2　对不同区域的样本过采样

在采样过程中,我们认为划分到正域中的样本点都是安全的,即在分类过程中不容易分错,为避免合成冗余数据,对其不做采样处理,其采样处理主要针对边界域和负域中的样本.

第 1 步　边界域中的小类样本分布在决策边界,容易被误分,需要进行过采样. 首先,我们采用 SMOTE 算法对这类样本进行过采样;其次,对新合成的小类样本进行判断,若其对大类样本的泛化空间无影响,则保留该样本,否则删除该新合成的小类样本.

第 2 步　对负域中的样本进行采样. 划分到负域中的样本不一定全部是噪声数据,部分数据可能对最终的分类起作用,需要考虑负域中小类样本的重要性,可通过放大负域中小类样本的邻域半径,筛选非噪声点,进行一种新的过采样处理操作. 采样原理如图 2 ~ 3 所示.

(1)对边界域进行过采样

图 2 中,首先对边界域内(椭圆区域内的样本为边界域样本,椭圆区域外的样本为正域样本)小类样本进行采样. 假设对 x_1 进行采样,首先,找到离 x_1 距离最近的 $k(k=5)$个同类样本,即 $\{x_2,x_3,x_4,x_5,x_6\}$,利用这 5

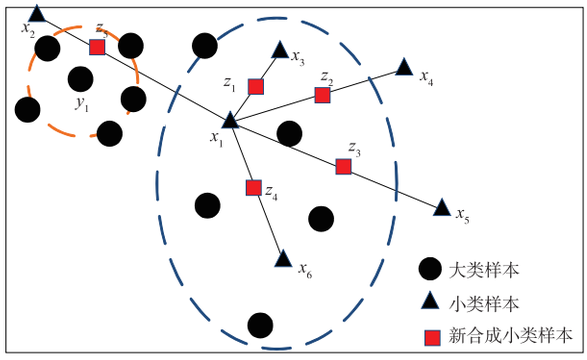

图2 边界域过采样原理图

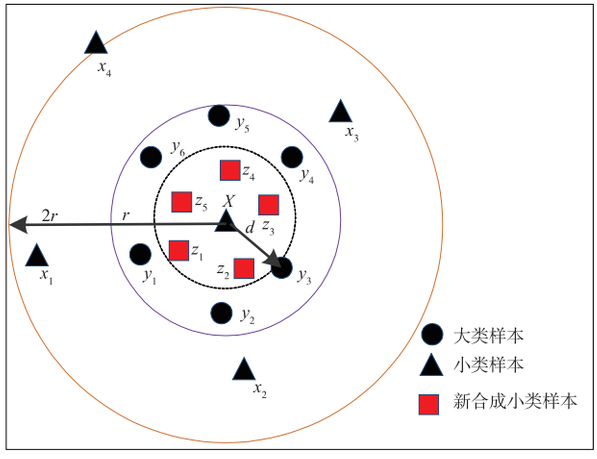

图3 负域采样过原理图

个样本合成新的样本$\{z_5,z_1,z_2,z_3,z_4\}$;其次,对新的合成样本进行检测,容易发现,$\{z_1,z_2,z_3,z_4\}$周围的大类样本都属于边界域,它们不影响正域内大类样本的分类,可以保留,而z_5会影响正域内大类样本y_1的分类,故需要删除新合成样本z_5. 通过以上处理,可有效减少新合成的小类样本对正域内的大类样本泛化空间的影响.

(2)对负域进行过采样

图3中,紫色圆形区域表示小类样本x的邻域(邻域半径为r),显然,x邻域内的样本集合$\{y_1,y_2,y_3,y_4,y_5,y_6\}$全是大类,即$N_0(\delta(x))=6,N_1(\delta(x))=0$. 根据式(7)和(8),可知$f(x)=-6/7,\beta=-5/6$,即$f(x)<\beta$,故$x$被当作噪声点划分到负域中. 分两种情况处理.

情况(1) 将x的邻域半径放大一倍为$2r$(x的新邻域见最外层圆形区域),可知新邻域内包含小类样本点$\{x_1,x_2,x_3,x_4\}$,在这种情况下,本文认为样本点x是非噪声点,需要在x的邻域内进行采样. 采样的方法如下:找到距离x最近的样本y_3,在以$d=\Delta(x,y_3)$为邻域半径的邻域内合成$k(k=5)$个新的小类样本点$\{z_1,z_2,z_3,z_4,z_5\}$.

情况(2) 将x的邻域半径放大一倍为$2r$,x在新邻域内仍没有同类样本,则将x从训练集中删除.

通过以上处理过程,一方面可保障新合成样本的数量,另一方面,可有效删除训练样本集中的噪声数据.

3.2 算法描述

基于三支决策的不平衡数据过采样算法具体步骤如下所示.

(1)根据邻域三支决策模型对训练集划分,输入邻域半径权重w,将训练集划分成正域样本集,边界域样本集和负域样本集,如算法1所示.

算法1 训练集划分算法

输入:训练样本集 TrainSet,邻域半径权重w,阈值k
输出:正域样本集(PosSet),边界域样本集(BndSet),负域样本集(NegSet))
Step1:(初始化)
BndSet=∅,PosSet=∅,NegSet=∅;
设置阈值k($k=5$),邻域半径权重w($w=0.01$ to 0.05);
Step2:(计算大类样本集和小类样本集)
按照决策属性D将 TrainSet 划分为N个等价类:$X_1,X_2\cdots X_N$;
计算小类样本集和大类样本集;
Step3:(计算每个样本的邻域)
FOR EACH x_i IN TrainSet DO
利用式(2)和(3),计算样本x_i与样本x_j之间的距离;
利用式(8)计算样本x_i的邻域半径δ_i;
计算x_i的邻域$\delta(x_i)$,其中,
$\delta(x_i)=\{x|x\in U,\Delta(x,x_i)\leqslant\delta_i\}$;
END FOR
Step4:(返回)
根据定义4,将训练样本集划分成正域样本集 PosSet,边界域样本集 BndSet,负域样本集 NegSet 并输出.

(2)根据算法1划分的训练集,对边界域样本集和负域样本集分别过采样处理,最后得到采样后的新的训练集,如算法2所示.

算法2 基于邻域三支决策的不平衡数据的过采样算法

输入:正域样本集 PosSet,边界域样本集 BndSet,负域样本集 NegSet,阈值k
输出:新训练样本集 TrainSetNew
Step1:(对边界域样本集中的小类过采样)
根据文献[2],采用 SMOTE 算法对该区域内的小类样本点进行过采样,得到合成的新的小类样本集 NewSet;
Step2:对新合成的每一个样本x_{new}判断新合成样本对正域内大类样本的影响,如果有影响将其从 NewSet 中删除.
Step3:(对负域样本集中的小类过采样)
3.1 对负域中的每一个小类样本放大的邻域半径,使$\delta_i'=2\times\delta_i$,在以$\delta_i'$为半径的邻域范围内搜索小类样本点;
3.2 IF $\delta_i'(x_i)\cap X_{\min}\neq\varnothing$ THEN
根据式(2)计算x_i与其邻域范围内所有样本的距离;令d

表示最小距离,在以 d 为邻域半径的邻域范围内生成 k ($k=5$)个小类样本点,其中,

$x_{new}=x_i+\text{rand}(0,1)\times(x_j-x_i)$;

ELSE

NegSet = NegSet-$\{x_i\}$;

Step4:(合并新训练集)

NewTrainSet = PosSet ∪ BndSet ∪ NegSet ∪ NewSet;

Step5:(返回)

输出新训练样本集 NewTrainSet.

算法的复杂度分析:令训练样本数目为 n,样本的属性数目为 m. 在算法 1 中,Step1 的时间复杂度为 $O(1)$;Step2 的时间复杂度为 $O(n)$;Step3 的时间复杂度为$O(m\times n^2)$;Step4 的时间复杂度为 $O(n)$;故算法 1 的时间复杂度为 $O(m\times n^2)$. 算法 1 的空间复杂度为 $O(m\times n)$. 在算法 2 中,Step1 的时间复杂度为 $O(k\times m\times n^2)$;Step2 的时间复杂度为 $O(m\times n^2)$;Step3 的时间复杂度为 $O(n)$;Step4 的时间复杂度为 $O(n)$;故算法 2 的时间复杂度为 $O(k\times m\times n^2)$. 算法 2 的空间复杂度为 $O(m\times n)$.

综上,本文采样算法的时间复杂度为 $O(k\times m\times n^2)$,空间复杂度为 $O(m\times n)$.

4　实验评价

4.1　数据集

为了验证算法的有效性,本文从 UCI 机器学习数据库中选取 15 个数据集进行实验,其中 12 个为两类别数据集,3 个为多类别数据集. 对于多类别数据集,合并某些数量较多的类别定义为大类,同时合并某些数量较少的类别定义为小类,则可将原来的多类别数据集重新定义为两类别数据集. 表 1 中类别是指原始数据集中数据类别的标签,即有几种类型的数据,比如“1:0”是指原始数据集中有两种类别数据,一种是类别为 1 的数据,另一种是类别为 0 的数据;“其他:1”是指原始数据集中有多种类别数据,类别数目最少的是类别 1(小类),“其他”代表剩余的其他类别(大类),所有数据集详细信息如表 1 所述. 从表 1 看每个数据集的类分布是不平衡的,按照不平衡度从小到大进行排序,不平衡度从 1.25 到 17.54.

4.2　实验方法

前文所述,邻域半径大小决定采样效果的好坏,根据式(9)可知,计算邻域半径必须确定权值 w. 本文通过实验对比确定权值 w 进而确定邻域半径的大小. 我们对部分数据集进行试验,w 的取值区间设定为[0,0.2],步长设为 0.01,通过 10 折交叉验证法计算出 w 取不同值时对应的 F-value[23],通过对比最佳的 w 值在 0.01 到 0.05 之间,故本文的 w 的取值为 0.01 ~ 0.05.

为了考察文中算法 TWD-IDOS 的性能,本文与 SMOTE[2]、Borderline-SMOTE[4]、ASMOTE[5] 和 SMOTE-RSB * [6] 等过采样算法进行了对比. 首先,确定分类器 C4.5,将本文算法与 5 个过采样算法进行了对比实验;其次,增加了两个分类器 KNN 和 CART,以验证本文算法在不同分类器下的实验效果;再次,将本文算法与欠采样算法 NCL[24],ENN[25] 在 C4.5,KNN 和 CART 等分类器下进行了对比实验;最后,将本文算法与结合集成学习的采样算法 EasyEnsemble[26] 进行了对比试验,然后,对算法 2 采样后的训练集进行了集成学习,并与 EasyEnsemble 算法进行了对比.

表 1　实验数据集(C:Continuous,N:Nominal)

数据集	属性	类别	大小	类分布
Austra	14C	1:0	690	307/383
Hert-s	6C 7N	Present:absent	270	120/150
Bupa	6C	1:2	345	145/200
Auto-mpg	7C 2N	其他:1	398	149/249
Colic	7C 15N	No: yes	368	136/232
Ionosphere	34C	b: g	351	126/225
Machine	7C	其他:2	209	74/135
Pima	8C	1:0	768	268/500
VC	7C	Normal:abnormal	370	100/210
German	24C	2:1	1000	300/700
Haberman	3C	2:1	306	81/225
Transfusion	4C	1:0	748	178/570
credit card	24C	1:0	30000	6636:23364
Yeast	8C	MIT:其他	1484	244/1240
Wilt	6C	w: n	4889	261:4578

本文采用 Weka[27] 平台下的 SMOTE[3] 算法对边界域小类样本进行过采样,使用该平台的 J48 (C4.5),KNN 和 SimpleCart (CART)算法作为分类器.

所选分类器参数均为 Weka 平台下算法的默认值,为使对比客观,将所有算法的上采样倍率 N 设为 1 倍,最近邻 k 值都设为 5. 所有实验结果均为采用 10 - 折交叉验证后的结果;对于 EasyEnsemble 算法的代码采用南京大学机器学习与数据挖掘研究所共享的 Matlab 代码[28],实验步骤与参数设置均与文献[26]中相同,其中,基分类器为 CART. 为增加可比性,本文集成学习基分类器同样选取 CART,集成算法同样采用 AdaBoost 算法.

4.3　实验结果

表 2 给出不同采样算法采样后的类别分布,表 3、表 4 给出了不同采样算法在 C4.5 算法上分类后的 F-value 和 AUC 的实验对比结果. 其中未采样表示在原始

数据集上分类得到的指标. 表中 Average 指评价指标在15个数据集上的均值;Rank 指不同采样算法在同一数据集上根据实验结果,按照1、2、3、4、5、6的次序进行Rank排序评分,最终得到的6种算法在15个数据集上Rank平均值.

从表3、表4可知,TWD-IDOS算法在大多数数据集上都取得了比较好的效果,F-value的平均值从0.6644提高到0.7012,AUC的平均值从0.7824提高到0.8083;在评价指标F-value和AUC上Rank值都小于其他算法.

分析以上实验数据及实验过程可发现:SMOTE算法对所有小类样本不加区别的采样,导致有些新合成小类样本落在了大类样本附近,影响大类样本的识别率. 虽然能够提高小类样本的F值,但在AUC上表现不理想;ASMOTE算法相比于SMOTE算法在选择 k 近邻时考虑了大类样本信息,并在采样过程加入了大类样本信息,使得有些新合成样本极其接近于大类样本,同样无法避免对大类样本识别率的影响;对于Borderline-SMOTE和SMOTE-RSB $*$ 算法,虽然减低了采样后新合成样本对大类样本识别率的影响,但是对进行采样的小类样本筛选条件较苛刻,会导致采样样本数不足,对小类样本识别率提高不是特别明显,因此,这两种算法在AUC上表现突出,在Recall和F值上提高不明显. 本文算法(TWD-IDOS)在采样过程中,删除了影响大类识别率的新合成样本,此外,还通过扩大小类样本的邻域半径进行采样,既有效减少了新合成样本对大类样本识别率影响,又保证了采样的数量,故TWD-IDOS在Recall、F值、AUC等指标上表现出了比较好的结果.

表2 不同采样算法采样后的类别分布

数据集	原始数据类分布	SMOTE采样后类分布	ASMOTE采样后类分布	BorderlineSmote采样后类分布	SMOTE-RSB采样后类分布	TWD-IDOS采样后类分布
Austra	307:383	614:383	583:383	381:383	310:383	423:383
Hert-s	120:150	240:150	228:150	152:150	122:150	195:150
Bupa	145:200	290:200	275:200	245:200	175:200	256:200
Auto-mpg	149:249	298:249	283:249	205:249	159:249	221:249
Colic	136:232	272:232	258:232	159:232	141:232	211:232
Ionosphere	126:225	252:225	239:225	174:225	158:225	209:225
Machine	74:135	148:135	140:135	94:135	75:135	120:135
Pima	268:500	500:500	509:500	418:500	298:500	510:500
VC	100:210	200:210	190:210	140:210	129:210	191:210
German	300:700	600:700	570:700	500:700	402:700	537:700
Haberman	81:225	162:225	153:225	139:225	102:225	157:225
Transfusion	178:570	356:570	338:570	270:570	278:570	341:570
Credit card	6636:23364	12608:23364	11562:23364	10218:23364	8647:23364	12764:23364
Yeast	244:1240	488:1240	463:1240	340:1240	274:1240	465:1240
Wilt	261:4578	495:4578	435:4578	387:4578	361:4578	499:4578

此外,为了直观展示不同采样算法在不同分类器(C4.5,KNN,CART)上的效果,下面用柱状图展示了不同方法在各个指标上的平均值,如图4、图5所示. 其中,纵坐标表示15个数据集在不同采样算法下对应评价指标的均值.

然后,分别在15个UCI数据集上与欠采样方法NCL[24]和ENN[25]做了简单的对比,实验结果如图6所示,其中纵坐标表示15个数据集在不同采样算法下对应评价指标的均值.

最后,本文算法分别在15个UCI数据集上与采用集成机制的EasyEnsemble采样算法进行比较,所有采样算法均以CART为分类器,以F值和AUC值作为评价指标;为了增加可比性,将本文采样算法TWD-IDOS与集成学习算法AdaBoost进行结合(先用本文算法2进行采样,然后采用AdaBoost算法集成),比较结果如表5、表6所示.

从表5可以看出,本文算法在15个数据集上的F均值为0.6946,略高于EasyEnsemble算法以及其他采样算法,其原因在于通过对边界域小类样本有监督处理,扩大了小类样本的泛化空间,从而保证小类的召回率;同时,又删除对大类分类有影响的小类,即变向提高精度;对负域样本进行采样,删除了噪声数据,同样也有利于提高精度. 因此,本文算法TWD-IDOS在F值上表现较为突出. 从表6可以看出,本文算法在15个数

据集上的 AUC 均值为 0.8234,低于 EasyEnsemble 算　　法,高于其他采样算法.

表 3　F-value 值

Dataset	未采样	SMOTE	ASMOTE	Borderline-SMOTE	SMOTE-RSB	TWD-IDOS
Austra	0.8132	**0.8459**	0.8352	0.8355	0.8093	0.8449
Heart-s	0.7364	0.7712	0.7265	0.7510	0.7364	**0.7980**
Bupa	0.5878	0.6045	0.5732	0.5590	0.5878	**0.6226**
Auto-mpg	0.8271	0.8491	0.8481	0.8381	0.8367	**0.8903**
Colic	0.7520	0.7445	0.7458	0.7473	0.7460	**0.7983**
Ionosphere	**0.8740**	0.8504	0.8730	0.8526	**0.8740**	0.8710
Machine	0.8690	0.8571	0.8816	0.8980	0.8690	**0.9109**
Pima	0.6142	0.6489	0.6397	0.6301	0.6142	**0.6643**
VC	0.6919	0.7163	0.6330	0.6961	0.6919	**0.7643**
German	0.5280	0.4544	0.4444	0.5159	0.5127	**0.5322**
Haberman	0.3582	0.4828	0.4972	**0.5116**	0.3944	0.4531
Transfusion	0.4813	0.4749	0.4745	0.4852	0.4551	**0.4885**
Credit card	0.4335	**0.4795**	0.4569	0.4722	0.4335	0.4457
Yeast	0.5577	0.5732	**0.5813**	0.5274	0.5577	0.5805
Wilt	0.8419	0.8286	0.8284	0.8402	0.8419	**0.8529**
Average	0.6644	0.6788	0.6693	0.6773	0.6640	**0.7012**
Rank	4.3333	3.3333	4.0667	3.4000	4.1333	**1.7333**

表 4　AUC 值

Dataset	未采样	SMOTE	ASMOTE	Borderline-SMOTE	SMOTE-RSB	TWD-IDOS
Austra	0.8483	0.8739	0.8433	0.8635	0.8458	**0.8980**
Heart-s	0.7443	0.7947	0.7223	0.7458	0.7448	**0.8354**
Bupa	0.6650	0.6468	0.6132	0.6355	0.6652	**0.6818**
Auto-mpg	0.9282	0.8973	0.8874	0.9015	0.9303	**0.9389**
Colic	0.7873	0.7740	0.8170	0.7971	0.7855	**0.8419**
Ionosphere	0.8923	0.8879	**0.9130**	0.8778	0.8930	0.9045
Machine	0.9359	0.9199	0.9268	**0.9538**	0.9359	0.9346
Pima	0.7514	0.7417	0.7321	0.7158	0.7513	**0.7674**
VC	0.8380	0.8107	0.7688	0.8044	0.8380	**0.8532**
German	**0.6884**	0.6333	0.6115	0.6853	0.6620	0.6805
Haberman	0.6087	0.6255	**0.6570**	0.6536	0.6174	0.6464
Transfusion	0.7001	0.6813	0.6836	**0.7075**	0.7007	0.7062
Credit card	0.6549	0.7007	0.6882	0.6702	0.6549	**0.7058**
Yeast	0.7500	**0.7801**	0.7698	0.7649	0.7500	0.7684
Wilt	0.9432	0.9338	0.9468	0.9414	0.9432	**0.9611**
Average	0.7824	0.7801	0.7721	0.7812	0.7812	**0.8083**
Rank	3.8667	4.1333	4.1333	3.6000	3.5333	**1.7333**

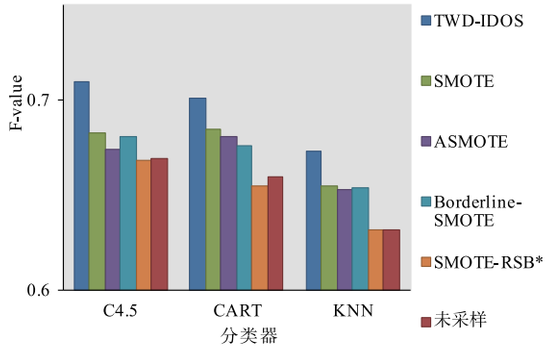

图4 本文算法与其他过采样算法在不同分类器下F-value值的对比

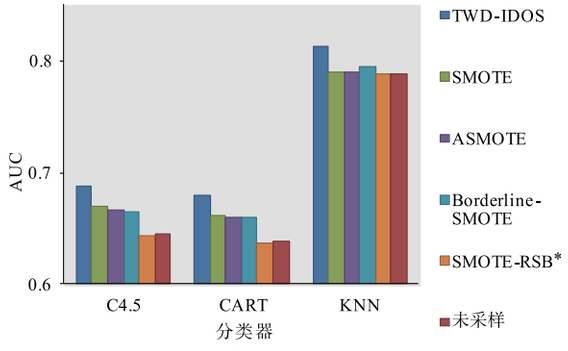

图5 本文算法与其他过采样算法在不同分类器下AUC值的对比

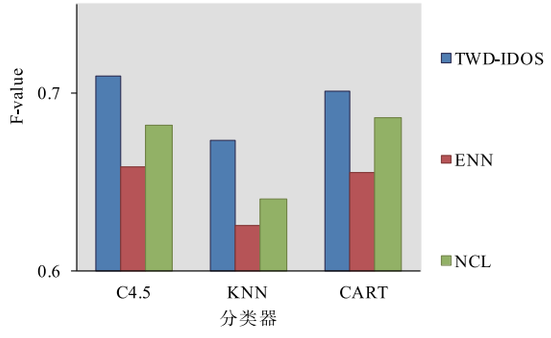

图6 本文算法与欠采样算法在不同分类器下F-value值的对比

当本文采样算法与 AdaBoost 结合后,F 均值从 0.6946 提高到 0.7004,提高了 11 个数据集的 F 值;AUC 均值从 0.8234 提高到 0.8564,近似于 EasyEnsemble 算法的 AUC 均值,提升较为明显,其中,提高了 13 个数据集的 AUC 值.

本文算法侧重点在于过采样,更多关注 F 值的提升;但是,算法过程略显复杂,计算复杂性较高. EasyEnsemble 算法有效结合了欠采样和集成学习,更侧重于集成学习,同时,文献[26]指出:EasyEnsemble 算法继承了集成学习方法的弱点,可理解性不强.

表 5 算法对比 F 值

DataSet	TWD-IDOS	Easy	TWD-IDOS (Adaboost)
Austra	0.8479	0.8571	0.8687
Hert-s	0.7920	0.7640	0.8000
Bupa	0.6560	0.6456	0.6581
Auto-mpg	0.8867	0.8864	0.8925
Colic	0.7899	0.7769	0.7862
Ionosphere	0.8506	0.8829	0.8893
Machine	0.8875	0.9072	0.9411
Pima	0.6673	0.6585	0.6258
VC	0.7248	0.7406	0.7757
German	0.5235	0.5852	0.5556
Haberman	0.4388	0.4454	0.4133
Transfusion	0.4817	0.4647	0.3843
Credit card	0.4602	0.4993	0.4820
Yeast	0.5687	0.5312	0.5804
Wilt	0.8429	0.7393	0.8529
Average	0.6946	0.6923	0.7004

表 6 算法对比 AUC 值

DataSet	TWD-IDOS	Easy	TWD-IDOS (Adaboost)
Austra	0.8817	0.9349	0.9257
Hert-s	0.8369	0.8704	0.8869
Bupa	0.7016	0.7596	0.7710
Auto-mpg	0.9560	0.9693	0.9759
Colic	0.8490	0.8254	0.8951
Ionosphere	0.9049	0.9711	0.9680
Machine	0.9584	0.9914	0.9892
Pima	0.8065	0.8064	0.7863
VC	0.8489	0.9122	0.9080
German	0.7141	0.7776	0.7716
Haberman	0.6428	0.6650	0.7140
Transfusion	0.7102	0.6956	0.6720
Credit card	0.7848	0.7452	0.7856
Yeast	0.7866	0.8491	0.8177
Wilt	0.9681	0.9918	0.9785
Average	0.8234	0.8510	0.8564

5 结束语

本文提出了一种基于三支决策的不平衡数据过采样方法,用于解决不平衡数据的二分类问题. 首先,根据邻域三支决策模型从样本总体分布来定义边界域样

本、正域样本和负域样本;其次,结合邻域三支决策模型,对划分的三个区域的样本分别进行不同的过采样处理.一方面,有效减少了新合成的小类样本与大类样本的重叠交叉,降低了噪声点对分类的影响;另一方面,解决了样本极度不平衡的条件下,小类样本被当作噪声点舍去的问题.15 个 UCI 数据集上的实验结果表明,相对于本文提到的其他采样方法,本文提出的方法对不平衡数据分类的各个指标都有明显的提高.但是,本文算法计算复杂性较大,当数据量偏大的时候导致算法运行时间较长,研究更加高效的算法将是今后的研究重点.

参考文献

[1] DRUMMOND C, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling[A]. Proceedings of ICML Workshop on Learning from Im-Balanced Datasets II[C]. New York: ACM, 2003. 1 – 8.

[2] CHAWLA N V, BOWYER K W, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, (16): 321 – 357.

[3] YEN S J, et al. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718 – 5727.

[4] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[A]. Proceedings of International Conference on Intelligent Computing (ICIC)[C]. Germany: Springer, 2005. 878 – 887.

[5] 杨智明,乔立岩,彭喜元. 基于改进 SMOTE 的不平衡数据挖掘方法研究[J]. 电子学报,2007,35(12):22 – 26.
YANG Zhiming, QIAO Liyan, PENG Xiyuan. Research on data ming method for imbalanced dataset based on improved SMOTE[J]. Acta Electronica Sinica, 2007, 35(12): 22 – 26. (in Chinese)

[6] RAMENTOL E, CABALLERO YAILé, BELLO R, et al. SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory[J]. Knowledge and Information Systems, 2012, 33(2): 245 – 265.

[7] 曾志强,吴群,廖备水,等. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报, 2009, 37(11): 2489 – 2495.
ZENG Zhiqiang, WU Qun, LIAO Beishui, et al. A classfication method for imbalance data set based on Kernel SMOTE[J]. Acta Electronica Sinica, 2009, 37(11): 2489 – 2495. (in Chinese)

[8] 翟云,王树鹏,马楠,等. 基于单边选择链和样本分布密度融合机制的非平衡数据挖掘方法[J]. 电子学报, 2014, 42(7): 1311 – 1319.
ZHAI Yun, WANG Shupeng, MA Nan, et al. A data mining method for imbalanced datasets based on one-sided link and distribution density of instance[J]. Acta Electronica Sinica, 2014, 42(7): 1311 – 1319. (in Chinese)

[9] 王磊,黄河笑,吴兵,等. 基于主题与三支决策的文本情感分析[J]. 计算机科学,2015,42(6):93 – 96.
WANG Lei, HUANG Hexiao, WU Bing, et al. Emotion analysis of text based on topics and three-way decisions[J]. Computer Science, 2015, 42(6): 93 – 96. (in Chinese)

[10] LI H X, et al. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. Knowledge-Based Systems, 2016, 91(1): 241 – 251.

[11] LIU D, LI T R, et al. Incorporating logistic regression to decision-theoretic rough sets for classifications[J]. International Journal of Approximate Reasoning, 2014, 55(1): 197 – 210.

[12] YU H, ZHANG C, WANG G Y. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-Based Systems, 2016, 91(1): 189 – 203.

[13] LIU S L, LIU X W. A novel three-way decision basedon linguistic evaluation[A]. Proceedings of 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)[C]. Istanbul: IEEE, 2015. 1 – 7.

[14] LIU D, LIANG D C, et al. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91(1): 32 – 45.

[15] CHEN Y M, ZENG Z Q, et al. Three-way decision reduction in neighborhood systems[J]. Applied Soft Computing, 2016, 38(1): 942 – 954.

[16] LIU D, LI T R, et al. A multiple-category classification approach with decision-theoretic rough sets[J]. Fundamenta Informaticae, 2012, 115(2 – 3): 173 – 188.

[17] ZHOU B. Multi-class decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55(1): 211 – 224.

[18] LIN T Y. Neighborhood systems and approximation in relational databases and knowledge bases[A]. Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems[C]. Charlotte NC: Oak Ridge National Laboratory, 1989. 75 – 86.

[19] HU Q H, et al. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866 – 876.

[20] STANFILL C, WALTZ D. Toward memory-based reasoning[J]. Communications of the ACM, 1986, 29(12): 1213 – 1228.

[21] YAO Y Y. An outline of a theory of three-way decisions[A]. Proceedings of Eighth International Conference of

RSCTC[C]. Germany:Springer,2012. 1 – 17.
[22] 刘盾,李天瑞,苗夺谦,等. 三支决策与粒计算[M]. 北京:科学出版社,2013. 12 – 30.
LIU Dun, LI Tianrui, MIAO Duoqian, et al. Three-Way Decision and Granular Computing[M]. Beijing: Science Press,2013. 12 – 30. (in Chinese)
[23] HU F, LI H. A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE[J/OL]. Mathematical Problems in Engineering, 2013, Article ID 694809, doi:10. 1155/2013/694809.
[24] LAURIKKALA J. Improving identification of difficult small-classes by balancing class distribution[A]. Proceedings of Eighth Conference on Artificial Intelligence in Medicine in Europe (AIME)[C]. Germany:Springer,2001. 63 – 66.
[25] TOMEK I. An experiment with the edited nearest-neighbor rule[J]. IEEE Transactions on Systems, Man, and Cybernetics,1976,6(6):448 – 452.
[26] LIU X Y, WU J, ZHOU Z H. Exploratory under-sampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009,39(2):539 – 550.
[27] Wikipedia Weka (machine learning)[CP/OL]. http://en. wikipedia. org/wiki/Weka,2010.
[28] Learning and Mining from Data (LAMDA)[CP/OL]. http://lamda. nju. edu. cn/CH. Data. ashx, 2016 – 10 – 31.

作者简介

胡 峰 男,1978 年 7 月出生,湖北天门人,教授、硕士生导师. 2000 年、2003 年和 2011 年分别在重庆大学、武汉大学和西南交通大学获得理学学士、工学硕士和工学博士学位,现为重庆邮电大学教师. 主要研究方向为数据挖掘、Rough 集和粒计算等.
E-mail:hufeng@ cqupt. edu. cn

王 蕾 男,1989 年出生于山东德州,重庆邮电大学在读硕士研究生. 主要研究方向为数据挖掘、三支决策、Rough 集.