DOI: 10.11992/tis.201706049
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1534.030.html

一种基于密度的 SMOTE 方法研究

王俊红, 段冰倩
（山西大学 计算机与信息技术学院, 山西 太原 030006）

摘　要: 重采样技术在解决非平衡类分类问题上得到了广泛的应用。其中, Chawla 提出的 SMOTE(Synthetic Minority Oversampling Technique) 算法在一定程度上缓解了数据的不平衡程度, 但这种方法对少数类数据不加区分地进行过抽样, 容易造成过拟合。针对此问题, 本文提出了一种新的过采样方法: DS-SMOTE 方法。DS-SMOTE 算法基于样本的密度来识别稀疏样本, 并将其作为采样过程中的种子样本; 然后在采样过程中采用 SMOTE 算法的思想, 在种子样本与其 k 近邻之间产生合成样本。实验结果显示, DS-SMOTE 算法与其他同类方法相比, 准确率以及 G 值有较大的提高, 说明 DS-SMOTE 算法在处理非平衡数据分类问题上具有一定优势。

关键词: 非平衡; 分类; 采样; 准确率; 密度
中图分类号: TP311　　**文献标志码:** A　　**文章编号:** 1673−4785(2017)06−0865−08

中文引用格式: 王俊红, 段冰倩. 一种基于密度的 SMOTE 方法研究[J]. 智能系统学报, 2017, 12(6): 865–872.
英文引用格式: WANG Junhong, DUAN Bingqian. Research on the SMOTE method based on density[J]. CAAI transactions on intelligent systems, 2017, 12(6): 865–872.

Research on the SMOTE method based on density

WANG Junhong, DUAN Bingqian
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: In recent years, over-sampling has been widely used in the field of classification of imbalanced classes. The SMOTE(Synthetic Minority Oversampling Technique) algorithm, presented by Chawla, alleviates the degree of data imbalance to a certain extent, but can lead to over-fitting. To solve this problem, this paper presents a new sampling method, DS-SMOTE, which identifies sparse samples based on their density and uses them as seed samples in the process of sampling. The SMOTE algorithm is then adopted, and a synthetic sample is generated between the seed sample and its k neighbor. The proposed algorithm showed great improvement in precision and G-mean compared with similar algorithms, and it has advantage of treating imbalanced data classification.

Keywords: imbalance; classification; sampling; precision; density

非平衡数据的分类问题广泛存在于电信诈骗检测、医疗诊断、网络入侵监控[1]、生物信息学、文本分类[2]、语言识别[3]、监测石油泄漏卫星图像[4]等领域中, 在这些实际应用中, 很多数据的结构并不是理想化、均匀、平衡地分布。在非平衡数据分类过程中, 由于正类样本数量相对稀少, 其所要表达的信息受到了限制, 从而在分类时很难正确分析出数据的分布以及内部规律, 导致少数类的分类精度降低, 所以在分类过程中非平衡数据中少数类的数据稀少是导致分类性能下降的直接原因之一[5]。如何能够在分类之前对数据进行预处理, 弥补少数类样本在分布信息方面不足的问题, 以达到将数据平衡化的目的, 从而提高分类器的性能, 是非平衡数据学习过程中的重点所在。

在目前的研究中, 用于解决非平衡数据分类问题的常用策略大致分为两种, 即数据层面的方法和算法层面的方法。算法层面的方法主要包括集成学

收稿日期: 2017−06−12.　**网络出版日期:** 2017−11−09.
基金项目: 国家自然科学基金项目(61772323, 61402272); 山西省自然科学基金项目(201701D121051).
通信作者: 王俊红. E-mail: wjhwjh@sxu.edu.cn.

习方法和代价敏感学习方法[6]; 数据层面的方法主要思想是基于重采样技术也就是对少数类样本的过采样和对多数类样本的欠采样。

作为数据层面的处理方法, 重采样方法简单、直观, 倍受研究人员青睐[7-9], 在过采样研究中, 2002 年 Chawla 提出的 SMOTE [10](synthetic minority over-sampling technique) 算法是其中的一个经典算法, 现有很多算法都是在此原型的基础上提出的。SMOTE 算法在少数类样本与其 K 个近邻之间的连线上产生随机的合成样本, 完成对少数类样本的过采样, 提高了少数类样本数量, 使分类器更好地对未知少数类样本进行预测, 有效地提高了分类精度。但是, 由于 SMOTE 算法无区分地对所有少数类样本进行过采样, 在少数类极度稀缺时很容易造成过拟合。为了解决这一问题, 文献[11]提出了基于 SMOTE 的改进方法 (adaptive SMOTE, AS-MOTE) 方法[11], 该方法根据样本集内部实际分布特性, 自适应地调整合成样本产生过程中的近邻选择策略, 避免了原始方法中样本生成的盲目性, 进而一定程度上提高了分类算法的准确率。在文献[12]提出的 Borderline-SMOTE 方法中, 改变了传统的思路, 认为边界样本具有更多的信息, 通过查找出的“危险样本”作为种子来产生新的合成样本。在此基础上, Haibo He 等[13]对以上算法进行了改进, 根据样本的“危险”程度, 也就是少数类样本在学习中的难易程度, 使用加权的方法, 构造合成样本的分布函数, 来确定这些样本合成新样本的数目。

上述文献中所涉及的采样思想大多为基于线性距离, 有一定的局限性, 易受数据集中样本分布结构的影响。本文提出了一种新的非平衡数据集处理方法——基于密度的重采样算法 DS-SMOTE, 旨在识别任意形状的簇, 并且扩大“危险样本”的范围, 对簇中的“稀疏样本”进行重采样, 达到数据集的平衡。在本文实验中, 选择了 C4.5 算法作为基准分类算法, C4.5 算法是具有代表性的决策树基准算法, 在分类数据不平衡的情况下与同类分类器相比具有良好的分类性能, 故选择 C4.5 算法作为基准分类算法。

1　基于密度的过采样方法

基于密度的方法能够用于识别任意形状的簇, 如“S”形以及椭圆形簇 [14]。为了更好地对非平衡数据集的分布进行刻画, 结合 SMOTE 的思想, 本节提出一种基于密度的过采样方法。

1.1　SMOTE 算法

SMOTE[7]是过采样方法中的经典算法。其主要思想基于 k 近邻算法, 对每个少数类样本 o, 从它的 k 个最近邻中随机选择一个样本 a, 在 $o \sim a$ 之间的连线上随机产生合成样本, 一般地, 取 k=5 [15]。

SMOTE 过采样方法通过对数据集中每一个少数类样本随机地选取 k 个最近邻, 并在该少数类样本与其选取的最近邻之间的连线上随机产生合成样本。在步骤 1) 中, 通过获取的过采样比例以及少数类样本的数量, 产生新样本数量; 在步骤 2)~5) 中, 循环遍历每一个少数类样本, 获取其 k 近邻, 并将其 k 近邻的索引保存于数组 nnarry 中, 作为合成新样本的参数获取新样本, 这里选取最近邻个数 k=5, 例如: 已知过采样比例为 300%, 那么需要从每个少数类样本的 5 个最近邻中随机选取 3 个样本, 作为合成新样本的素材; 其中 Populate 方法详细介绍了合成样本的产生过程。在步骤 3) 中, 根据少数类样本的属性数以及新样本的数量进行遍历, 步骤 5) 用于求得样本间的欧式距离; 在步骤 8) 中, 根据公式, 得到少数类样本与其近邻之间的合成样本。新生成的样本使得分类器能够从学习过程中产生更大更抽象的决策边界, 有效地使少数类的决策边界变得更加明显, 从而获得更好的分类效果。

SMOTE 算法

SMOTE(T, N, K)

输入　少数类样本数量 T; 过采样百分比 N%; 样本近邻个数 k;

输出　(N/100)*T 个少数类合成样本。

1) N=(int)(N/100);

2) for i=1:T

3) 计算第 i 个样本的 k 个最近邻, 其 k 近邻的索引保存于 nnarry 中;

4) Populate(N, i, nnarry);

5) endfor

Populate(N, i, nnarry) 方法

1) While N~=0

2) 选择 1 ~ k 之间的随机整数, 记为 nn。随机选择第 i 个样本的 k 近邻中的一个近邻。

3) for attr=1:numattrs

4) (numattrs 为样本的属性个数)

5) dif = sample[nnarry[nn][attr]] – sample[i][attr];

6) (sample[][]为原始少数类样本集合)

7) gap=rand();

8) synthetic[newinder][attr] = sample[i][attr] + gap*dif;

9) (newinder 为合成样本总量, 初始值为 0)

10) endfor

11) newindex++;

12) N=N+1;

13) endwhile

1.2 Borderline-SMOTE 算法

Borderline-SMOTE 算法在 SMOTE 算法的基础上进行了改进[12]。不同于 SMOTE 算法, Borderline-SMOTE 算法避免了在过采样过程中样本选择的随机性, 选择类边界样本, 作为种子样本合成新样本。在样本空间中, 如某一样本周围的邻居样本较多, 则这个样本较为稠密, 就决策树分类算法而言, 产生的叶子节点——规则较多, 容易产生过拟合问题, 所以不宜在这个样本与其近邻之间添加新的样本。Borderline-SMOTE 算法尝试着在训练过程中尽量地去学习边界特征, 认为边界样本在分类中比远离边界的样本更容易被错分, 一个类的边界样本携带了更多的信息, 对分类器分类性能的好坏起到了决定性的作用。因此, 在 Borderline-SMOTE 算法中, 加入了识别边界样本的过程: 若一个少数类样本的 m 近邻中, 半数以上为多数类样本, 则认为这个样本为容易被错分的危险样本; 否则为安全样本。在危险样本与其 k 近邻之间合成新样本, 完成对边界少数类样本的过采样, 以此加强少数类的决策边界, 以获得好的分类结果。

1.3 基于密度的过采样算法

设某类的样本集合为 $S=\{\boldsymbol{s}_i, i=1, 2, \cdots, n\}$, 其中 $\boldsymbol{s}_i$ 为维数为 m 的样本向量, 样本的维数代表其属性的个数。一个类的对象 o 的密度是指靠近对象 o 的对象数量。

定义 1　参数 $\varepsilon(\varepsilon>0)$ 为对象 o 的邻域区域的半径, 即 o 的邻域半径。则一个对象 o 的 ε-邻域是指以 o 为圆心、以 ε 为半径的空间, 定义为

$$\psi=\{\varepsilon,\theta|2\cdot D(\boldsymbol{s}_i,o)\cdot\cos\theta\leqslant\varepsilon\} \tag{1}$$

式中: $D(\boldsymbol{s}_i, o)$ 为对象 o 以外的样本 $\boldsymbol{s}_i$ 到对象 o 的距离, 采用欧式距离方法计算[16]; θ 为点 o 与 $\boldsymbol{s}_i$ 连线与水平轴之间的夹角。

定义 2　对象 o 的 ε-邻域的密度是指对象 o 在 ε-邻域内的对象数量。

定义 3　一个样本数量为 m 的类在邻域半径为 ε 时, 类中样本的密度阈值 Min Pts 为类中样本 i 的 ε-邻域密度 M_i 的均值。

$$\text{Min Pts}=\frac{1}{m}\sum_{i=1}^{m}M_i \tag{2}$$

为了确定对象 o 是否为稀疏点, 即对象 o 的 ε-邻域是否稀疏, 本文中使用参数 Min Pts, 作为稠密区域的密度阈值。如果一个对象 o 的 ε-邻域密度 $M\geqslant$Min Pts, 则 o 为一个稠密对象, 如果一个对象 o 的 ε-邻域密度 $M<$Min Pts, 则 o 为一个稀疏对象。一个类中的稀疏对象构成此类的种子样本集合 (seeds)。

根据 SMOTE 算法思想, DS-SMOTE 算法根据式(3)在种子样本与其近邻之间合成新样本:

$$\text{new}\Rightarrow o+D(\text{seed}_i,o)\times r \tag{3}$$

式中: o 为目标对象, $D(\text{seed}_i, o)$ 为种子样本 seed_i 与其近邻样本 o 之间的欧氏距离, r 为随机数, 且 $r\in(0, 1)$。

DS-SMOTE 算法的核心思想为: 在 SMOTE 算法的基础上, 在少数类中抽取种子样本集合后对其进行过采样。在算法中, 我们对少数类中稀疏对象进行样本采集得到一个种子样本集合, 产生种子集合的过程主要包括: 计算少数类中样本的邻域半径、计算该类的密度阈值、产生稀疏对象集合作为种子样本集。DS-SMOTE 算法产生的合成样本分布于稀疏对象及其近邻之间, 最终得到与多数类样本数量相等的少数类样本集合。

DS-SMOTE 算法

输入　训练集 T, 原始样本集合中多数类 $S_1=\{x_1, x_2, \cdots, x_{\text{maj}}\}$、少数类集合 $S_2=\{y_1, y_2, \cdots, y_{\min}\}$、邻域半径 ε、密度阈值 Min Pts。

1) 对于少数类集合 S_2, 在整个训练集 T 中对 S_2 中的每一个样本 y_i 计算其近邻。若其近邻类别都为多数类, 则认为 y_i 为噪声, 且不会出现在下一步计算中;

2) 选择少数类集合 S_3, 执行 $\varepsilon=\text{get}E(S_3)$ 得到少数类邻域半径 ε;

3) 求得邻域半径为 ε 时, 每个少数类样本的密度;

4) for j = 1, |min|

5) 令 y_i 的 ε-邻域密度=1;

6) for j = 1, |min|

7) 计算其他少数类样本与样本 y_i 的欧式距离 D;

8) if $D<=\varepsilon$

9) y_i 的 ε-邻域密度+1;

10) endfor

11) endfor

12) endfor

13) 根据式 (2), 求得密度阈值 Min Pts;

14) for j = 1, |min|

15) if y_i 的 ε-邻域密度≤密度阈值

16) y_i 为稀疏对象, 将 y_i 加入种子样本集合;

17) endfor

18) endfor

19) 产生种子集合 seed;

20) 产生随机数 r;

21) 根据式 (3) 合成新样本, 得到样本集合 new;

22) train = train∪new;

DS-SMOTE 算法中, 需要输入非平衡数据集合 T, 以及参数邻域半径 ε、密度阈值 Min Pts。首先在步骤 1) 中, 排除少数类中的噪声产生新的少数类集合, S_3; 在步骤 2) 中, 算法选取了少数类样本集合 S_3 进行操作, 根据密度的概念: 对象 o 的 ε-邻域密度是指对象 o 在 ε-邻域内的对象数量, 遍历每个少数类样本求得其 ε-邻域的密度, 并在步骤 3)~12) 中取少数类样本密度的均值作为判断少数类集合中样本是否稠密的密度阈值。

如何排除人工的方法, 为少数类设置恰当的邻域半径是基于密度的分类算法中的一个亟待解决的问题。本文选择非平衡类分类中一般性的评估标准: F-value 以及 G-mean 值进行评估, 经实验分析, 选择类内样本间平均距离作为邻域半径。表 1 与表 2 分别随机选取 5 个数据集进行实验分析, 选取了邻近样本间平均距离 ε'值的一系列大于零的点, 如 ε'*0.3、ε'*0.5、ε'*0.7 以及 ε'*1.3、ε'*1.5、ε'*1.7 进行测试。经过测试, 在图 1 与图 2 中可看出, 结果呈现出一定的规律性: ε 值的变化对不同的数据集的影响不尽相同, 在 ε 值等于 ε'或邻近取值时 F-value 以及 G-mean 值水平较高, 随着取值向 ε'的两侧远离, F-value 以及 G-mean 值或保持平稳, 或有所下降。说明选取类内样本间平均距离作为邻域半径具有一定的普适性, 并且对分类器的分类性能有一定的保证。由于其脱离了人工选择的邻域半径设置方法, 所以这种邻域半径的设置方法也提高了 DS-SMOTE 方法的可操作性。

表 1　不同邻域半径取值下的 G-mean 值

Table 1　The G-mean under different neighborhood radius

邻域半径	ε'*0.3	ε'*0.5	ε'*0.7	ε'	ε'*1.3	ε'*1.5	ε'*1.7
Germany	0.641 9	0.647 4	0.614 7	0.671 0	0.665 8	0.661 5	0.601 5
Tic	0.929 3	0.969 5	0.973 6	0.983 1	0.975 9	0.968 2	0.963 1
diabetis	0.512 5	0.669 8	0.695 5	0.804 7	0.644 6	0.623 6	0.570 9
ionosphere	0.839 0	0.881 9	0.884 8	0.900 7	0.866 0	0.857 5	0.854 9
parkinsons	0.784 5	0.836 7	0.838 5	0.839 7	0.786 8	0.721 1	0.658 3

表 2　不同邻域半径取值下的 F-value 值

Table 2　The F-value under different neighborhood radius

邻域半径	ε'*0.3	ε'*0.5	ε'*0.7	ε'	ε'*1.3	ε'*1.5	ε'*1.7
Germany	0.578 0	0.593 5	0.613 2	0.645 2	0.597 1	0.592 9	0.531 6
Tic	0.888 5	0.957 9	0.970 3	0.962 6	0.942 0	0.922 3	0.892 8
diabetis	0.545 5	0.659 7	0.685 0	0.745 6	0.627 7	0.565 0	0.585 0
ionosphere	0.786 9	0.782 0	0.880 3	0.895 6	0.850 6	0.845 1	0.819 9
parkinsons	0.820 5	0.816 0	0.818 2	0.892 6	0.824 2	0.685 7	0.612 2

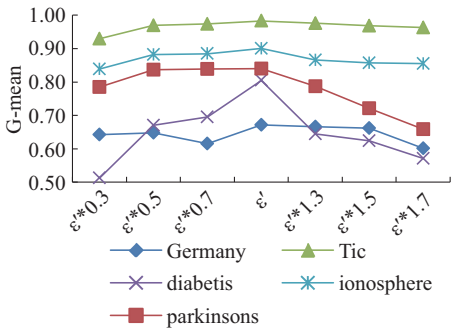


图 1　不同邻域半径取值下的 G-mean 值

Fig. 1　The G-mean under different neighborhood radius

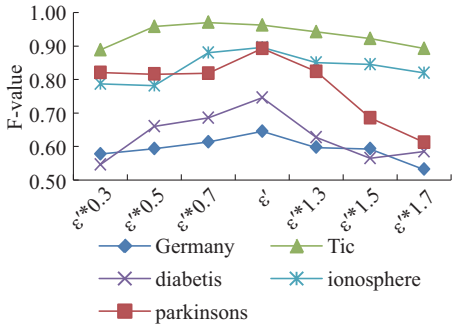


图 2　不同邻域半径取值下的 F-value 值

Fig. 2　The F-value under different neighborhood radius

在步骤 14)~18) 中, 使用循环遍历的方式判断少数类样本的稀疏性, 并将稀疏样本加入种子集合, 作为在步骤 21) 中合成新样本的素材。在步骤 21) 中引用了直观、易操作的 SMOTE 算法的思想

产生合成样本, 一定程度上避免了随机过采样方法容易造成的分类器过拟合问题, 最终在稀疏对象及其近邻之间合成新样本, 达到少数类与多数类样本在数量上的一致。

2　实验与结果分析

2.1　评价标准

一个分类器算法在二分类问题中的性能往往使用混淆矩阵来评估, 分别将两类分为正类(positive)、负类 (negative), 如表 3 所示[17]。混淆矩阵的列用来表示类的预测结果, 混淆矩阵的行用来表示类的实际类别[18]。其中, TN (true negative) 表示负类样本中被划分正确的样本数, 即真负类; TP(true positive) 表示正类样本中被划分正确的样本数, 即真正类; FN(false negative) 表示负类样本中被划分错误的样本数, 也就是负类中的样本被划分为正类的样本数, 即假负类; FP(flase positive) 表示正类样本中被划分错误的样本数, 也就是正类中的样本被划分为负类的样本数, 即假正类[19]。

表 3　二分类问题中的含混矩阵

Table 3　The confusion matrix of 2-class problem

分类	预测为正类	预测为负类
实际正类	TP	FN
实际负类	FP	TN

准确率 (Precision) 和召回率 (Recall) 是分类性能的两个最基本的指标[20]。准确率 (Precision) 也称为查准率, 召回率 (Recall) 也称为查全率, 即正类(少数类)的分类准确率。定义为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{TPR}=\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

F-value 是准确率和召回率的调和平均, 实验中令 β 值为 1, 即 F_1 度量。定义如下:(式中 β 为调整准确率 (Precision) 和召回率 (Recall) 所占比重的参数, 一般地令 β=1)。

$$\text{F-value}=\frac{(1+\beta^2)\times\text{Recall}\times\text{Precision}}{\beta^2\times\text{Recall}+\text{Precision}} \tag{6}$$

在非平衡类分类问题中, G-mean 值用来衡量分类器对于两类样本分类的平均性能[21], 是对算法性能的总体评价。

$$\text{G-mean}=\sqrt{\text{Recall}\times\text{TNR}} \tag{7}$$

本文选用 Recall(TPR)、TNR、Precision、F-value、G-mean 等值作为实验过程中算法性能指标的度量。

2.2　实验数据

为了测试文中实现的采样方法与同类方法对非平衡数据的分类效果, 文中采用了 11 个 UCI 数据集进行实验和分析, 如表 4 所示。非平衡数据中的非平衡度为正类与负类样本数量之比, 实验中所选取的数据集分别具有不同的非平衡程度, 正类的比例从 0.097 ~ 0.629 不等。这些数据集中的数据大多为数值型的两类样本数据, 其中, statimage 数据集中的样本有 7 个类别, 为了构造极其不平衡的样本集合, 人为将第 4 类样本作为少数类样本, 其余样本合为一类作为多数类样本, 从而得到一组非平衡度为 0.097 的两类数据样本; thyroid 数据集中具备 3 类样本, 通过将类别为 2 和 3 的样本合为一类, 从而获得了一组非平衡度为 0.194 的两类数据样本。表 4 同时给出了各数据集的属性个数、总样本数量、正类样本数量、负类样本数量以及正负类样本数量的比值——非平衡度。

表 4　实验所用 UCI 数据集

Table 4　The UCI datasets for experiments

数据集	属性	总样本数量	正样本数量	负样本数量	非平衡度
statimage	36	4 435	415	4 020	0.097:1
Thoracic	17	470	70	400	0.175:1
thyroid	6	215	35	180	0.194:1
parkinsons	23	195	48	147	0.327:1
ILPD	11	583	167	416	0.401:1
Germany	25	1 000	300	700	0.429:1
Echocardiogram	13	132	43	89	0.483:1
Tic	100	958	332	626	0.530:1
diabetis	9	768	268	500	0.536:1
ionosphere	35	351	126	225	0.560:1
votes	17	435	168	267	0.629:1

2.3　实验结果与分析

为了验证 DS-SMOTE 算法处理非平衡数据集的有效性, C4.5 算法是具有代表性的决策树基准算法, 在分类数据不平衡的情况下与同类分类器相比具有良好的分类性能, 实验中采用了 C4.5 算法作为分类算法, 并与 SMOTE 算法、Borderline-SMOTE 算法进行了对比。本文采用了十折交叉验证方法进行实验测试, 测试结果均为 10 次实验均值, 并针对 Recall(TPR)、TNR、Precision、F-value、G-mean 等指标进行分析。

为了对比算法的优势, 图 3 ~ 7 分别绘制了 4 种算法策略在 11 个数据集上的测试结果趋势曲线。其中, 横坐标为 4 种算法策略, 纵坐标取值在 0 ~ 1 之间, 表中加粗的数据为一系列数据中的最大值。通过以下图表可以看出, 使用 DS-SMOTE 方法进行过采样, 少数类的分类性能有所上升。

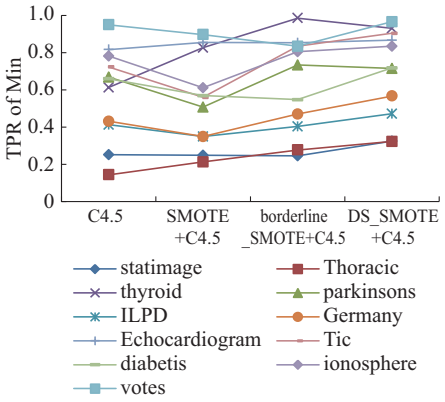


图 3　少数类准确率变化曲线图

Fig. 3　The variation curve for TPR of min

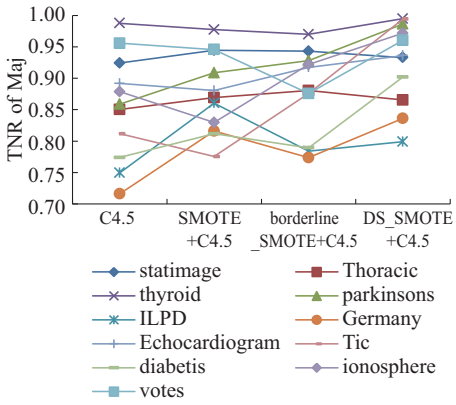


图 4　多数类准确率变化曲线图

Fig. 4　The variation curve for TNR of major

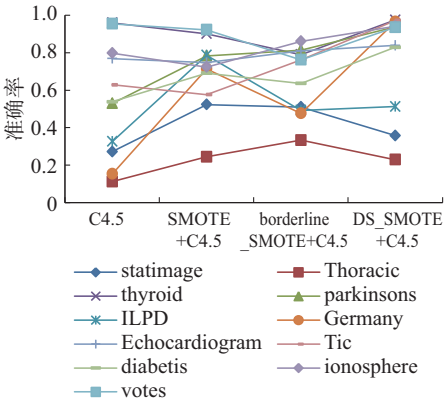


图 5　准确率变化曲线图

Fig. 5　The variation curve of precision

在表 5 中, 大部分数据集在 DS-SMOTE 算法处理后分类的 TNR 值大于使用 SMOTE 算法与 Borderline-SMOTE 算法, 对于少数类样本绝对稀少、非平衡程度较大的数据集 statimage 、thyroid 和 parkinsons 分类效果较差, 表明在处理少数类绝对稀少的非平衡类分类问题中, DS-SMOTE 算法仍有待改进; 表 6 中多数类样本的分类精度保持较高, 可见 DS-SMOTE 算法在保证多数类分类准确率的前提下对少数类的分类准确率有一定程度的改善; 在表 7 和 8 中可以观察到, DS-SMOTE 没有消除在两类的极端不平衡时对 Precision、F-value 值的影响; G-mean 值作为非平衡数据整体分类性能的评价指标, 往往能够指示一个方法在非平衡数据集的分类性能好坏, 表 9 显示出 DS-SMOTE 算法在大部分的数据集上的 G-mean 值有显著的优势, 说明本文提出的算法在这些数据集上有较好的总体分类性能。

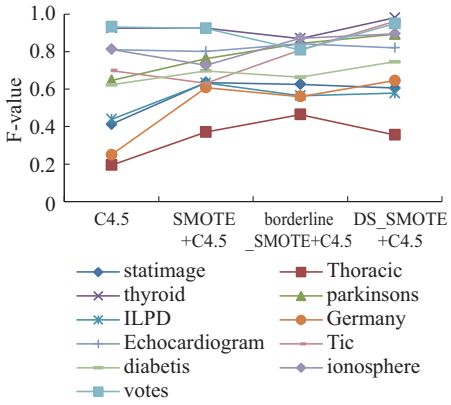


图 6　F-value 变化曲线图

Fig. 6　The variation curve of F-value

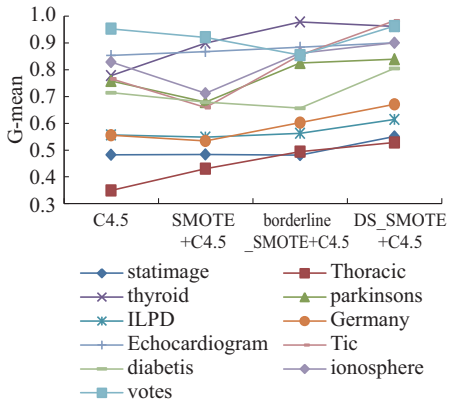


图 7　G-mean 变化曲线图

Fig. 7　The variation curve of G-mean

表 5　少数类准确率

Table 5　The TPR of min

数据集	C4.5	SMOTE+C4.5	Borderline_SMOTE+C4.5	DS-SMOTE+C4.5
statimage	0.251 4	0.248 0	0.245 3	0.324 6
Thoracic	0.143 6	0.212 9	0.277 1	0.323 1
thyroid	0.613 2	0.826 7	0.985 7	0.930 0
parkinsons	0.668 3	0.507 4	0.733 6	0.714 8
ILPD	0.413 1	0.348 8	0.403 3	0.471 5
Germany	0.431 3	0.349 7	0.469 9	0.566 9
Echocardiogram	0.816 7	0.854 8	0.852 5	0.866 9
Tic	0.723 6	0.560 0	0.834 6	0.904 3
diabetis	0.660 0	0.569 1	0.547 0	0.717 7
ionosphere	0.781 4	0.610 6	0.805 2	0.834 8
votes	0.950 1	0.897 3	0.834 6	0.965 7

表 6　多数类准确率

Table 6　The TNR of Major

数据集	C4.5	SMOTE+C4.5	Borderline_SMOTE+C4.5	DS-SMOTE+C4.5
statimage	0.924 1	0.944 5	0.943 1	0.932 9
Thoracic	0.850 3	0.869 3	0.880 6	0.865 2
thyroid	0.987 5	0.977 4	0.969 7	0.994 7
parkinsons	0.858 9	0.908 5	0.928 1	0.986 4
ILPD	0.749 6	0.860 4	0.784 0	0.799 2
Germany	0.716 1	0.816 1	0.773 7	0.836 1
Echocardiogram	0.891 9	0.880 4	0.917 4	0.936 2
Tic	0.811 6	0.775 4	0.875 9	0.994 2
diabetis	0.773 9	0.811 6	0.789 4	0.902 2
ionosphere	0.878 7	0.829 8	0.922 0	0.971 7
votes	0.955 7	0.945 6	0.875 9	0.960 4

表 7　准确率

Table 7　Precision

数据集	C4.5	SMOTE+C4.5	Borderline_SMOTE+C4.5	DS-SMOTE+C4.5
statimage	0.271 5	0.523 0	0.510 0	0.357 1
Thoracic	0.111 4	0.243 8	0.332 7	0.228 9
thyroid	0.960 0	0.901 7	0.790 0	0.975 0
parkinsons	0.529 8	0.783 3	0.814 5	0.937 9
ILPD	0.325 2	0.786 9	0.492 6	0.512 4
Germany	0.152 6	0.713 4	0.475 7	0.966 2
Echocardiogram	0.768 3	0.746 7	0.810 0	0.839 1
Tic	0.628 6	0.576 0	0.761 9	0.972 1
diabetis	0.538 5	0.690 7	0.636 2	0.829 2
ionosphere	0.797 5	0.723 8	0.861 4	0.940 9
votes	0.954 3	0.922 5	0.761 9	0.936 0

表 8　F-value

Table 8　F-value

数据集	C4.5	SMOTE+C4.5	Borderline_SMOTE+C4.5	DS-SMOTE+C4.5
statimage	0.412 2	0.635 0	0.625 0	0.605 4
Thoracic	0.194 4	0.370 6	0.464 3	0.355 4
thyroid	0.924 9	0.924 5	0.869 6	0.980 6
parkinsons	0.645 5	0.762 9	0.844 3	0.892 6
ILPD	0.438 4	0.632 2	0.564 3	0.578 3
Germany	0.249 4	0.607 5	0.559 8	0.645 2
Echocardiogram	0.810 1	0.801 2	0.842 9	0.820 5
Tic	0.698 6	0.631 8	0.808 9	0.962 6
diabetis	0.622 8	0.697 5	0.663 8	0.745 6
ionosphere	0.811 6	0.726 5	0.870 8	0.895 6
votes	0.931 3	0.923 3	0.808 9	0.949 4

表 9　G-mean

Table 9　G-mean

数据集	C4.5	SMOTE+C4.5	Borderline_SMOTE+C4.5	DS-SMOTE+C4.5
statimage	0.482 1	0.484 0	0.481 0	0.550 3
Thoracic	0.349 4	0.430 2	0.494 0	0.528 7
thyroid	0.778 1	0.898 9	0.977 7	0.961 8
parkinsons	0.757 7	0.678 9	0.825 1	0.839 7
ILPD	0.556 5	0.547 8	0.562 3	0.613 8
Germany	0.555 7	0.534 2	0.603 0	0.671 0
Echocardiogram	0.853 5	0.867 5	0.884 3	0.900 9
Tic	0.766 3	0.658 9	0.855 0	0.983 1
diabetis	0.714 7	0.679 6	0.657 1	0.804 7
ionosphere	0.828 6	0.711 8	0.861 6	0.900 7
votes	0.952 9	0.921 1	0.855 0	0.963 0

实验结果表明本文提出的算法在少数类信息不足的情况下, 分类效果有一定程度的改进, 能够在不降低多数类分类精度的同时, 保证分类器对少数类的识别, 并具有良好的适应性。

3　结束语

基于数据采样的方法是解决非平衡数据分类问题的一个重要途径, 本文在 SMOTE 算法的基础上, 结合密度的概念, 提出了基于密度的过采样方法, 以提高非平衡数据分类的准确率。实验结果表明, 本文的方法在处理非平衡数据分类问题上具有良好的效果。另外, 在本文中使用类内样本间平均距离作为邻域半径, 通过实验证明, 这种取值方法避免了人工取值的难题, 具有普适性和可操作性, 也使得分类器的分类性能得到了一定的保证。但是, 如何通过自适应方法产生类的邻域半径, 是本文进一步的研究方向。

参考文献:

[1] CHARTE F, RIVERA A J, JESUS M J D, et al. Addressing imbalance in multilabel classification: Measures and random resampling algorithms[J]. Neurocomputing, 2015, 163: 3–16.

[2] RADIVOJAC P, CHAWLA N V, DUNKER A K, et al. Classification and knowledge discovery in protein databases[J]. Journal of biomedical informatics, 2004, 37(4): 224–239.

[3] LIU Y, CHAWLA N V, HARPER M P, et al. A study in machine learning from imbalanced data for sentence bound-

ary detection in speech[J]. Computer speech and language, 2006, 20(4): 468–494.

[4] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine learning, 1998, 30(2): 195–215.

[5] QIAN H, HE G. A survey of class-imbalanced data classification[J]. Computer engineering and science, 2010, 5: 025.

[6] 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-side link and distribution density of instances[J]. Chinise journal of electronics, 2014, 42(7): 1311–1319.

[7] CHARTE F, RIVERA A J, JESUS M J D, et al. Addressing imbalance in multilabel classification: Measures and random resampling algorithms[J]. Neurocomputing, 2015, 163: 3–16.

[8] GONG C, GU L. A novel smote-based classification approach to online data imbalance problem[J]. Mathematical problems in engineering, 2016, 35: 1–14.

[9] BIAN J, PENG X G, WANG Y, et al. An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem[J]. Mathematical problems in engineering, 2016, 6: 1–9.

[10] 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.

[11] 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(B12): 22–26.
YANG Zhimin, QIAO Liyan, PENG Xiyuan. Research on datamining method for imbalanced dataset based on improved SMOTE[J]. Chinese journal of electronics, 2007, 35(B12): 22–26.

[12] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International Conference on Intelligent Computing. Springer Berlin Heidelberg, 2005, 3644(5): 878-887.

[13] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// IEEE International Joint Conference on Neural Networks. IEEE Xplore, 2008: 1322-1328.

[14] GRZYMALA-BUSSE J W, STEFANOWSKI J, WILK S. A comparison of two approaches to data mining from imbalanced data[J]. Journal of intelligent manufacturing, 2005, 16(6): 565–573.

[15] EZ J, KRAWCZYK B, NIAK M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets[J]. Pattern recognition, 2016, 57(C): 164–178.

[16] NANNI L, FANTOZZI C, LAZZARINI N. Coupling different methods for overcoming the class imbalance problem[J]. Neurocomputing, 2015, 158(C): 48–61.

[17] NAGANJANEYULU S, KUPPA M R. A novel framework for class imbalance learning using intelligent undersampling[J]. Progress in artificial intelligence, 2013, 2(1): 73–84.

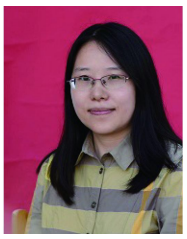
[18] ZHANG X, SONG Q, WANG G, et al. A dissimilarity-based imbalance data classification algorithm[J]. Applied intelligence, 2015, 42(3): 544–565.

[19] JIANG K, LU J, XIA K. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE[J]. Arabian journal for science and engineering, 2016, 41(8): 3255–3266.

[20] XU Y, YANG Z, ZHANG Y, et al. A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification[J]. Knowledge-based systems, 2016, 95: 75–85.

[21] ANWAR N, JONES G, GANESH S. Measurement of data complexity for classification problems with unbalanced data[J]. Statistical analysis and data mining the asa data science journal, 2014, 7(3): 194–211.

作者简介:



王俊红 女, 1979 年生, 副教授, 博士, 主要研究方向为形式概念分析、粗糙集与粒计算以及数据挖掘。

段冰倩, 女, 1991 年生, 硕士研究生, 主要研究方向为数据挖掘。