文章编号: 1001-0920 (2010) 06-0000-00

不均衡数据下基于阴性免疫的过抽样新算法

陶新民[1], 徐 晶[2], 童智靖[1], 刘 玉[1]

(1. 哈尔滨工程大学 信息与通信工程学院，哈尔滨 150001；2. 黑龙江省科技学院 数力系，哈尔滨 150027)

摘 要: 为提高不均衡数据集下算法分类性能, 提出一种基于阴性免疫的过抽样算法. 该算法利用阴性免疫实现少数类样本空间覆盖, 以生成的检测器中心为人工生成的少数类样本. 由于该算法利用的是多数类样本信息生成少数类样本, 避免了人工少数类过抽样技术(SMOTE)生成的人工样本缺乏空间代表性的不足. 通过实验将此算法与 SMOTE 算法及其改进算法进行比较, 结果表明, 该算法不仅有效提高了少数类样本的分类性能, 而且总体分类性能也有了显著提高.

关键词: 不均衡数据；阴性免疫；过抽样算法；人工少数类过抽样技术

中图分类号: TP18 **献标识码:** A

Over-sampling algorithm based on negative immune in imbalanced data sets learning

TAO Xin-min[1], *XU Jing*[2] , *TONG Zhi-jing*[1] , *LIU Yu*[1]

(1. College of Information and Communication Engineering，Harbin Engineering University，Harbin 150001，China；
2. Department of Mathematics and Mechanics，Heilongjiang Institute of Science and Technology，Harbin 150027，China.
Correspondent: TAO Xin-min，E-mail: taoxinmin@hrbeu.edu.cn)

Abstract: To improve the classification performance of minority class, a over-sampling based on negative immune principle is proposed. In this approach, the negative immune algorithm is induced to generate a set of resprentive detectors to implement the overlapping of minority class space based on learning majority samples. The centers of resprentive detectors are regarded as the synthetic minority samples in order to resolve the imbalance problem. The majority samples are used to generate the synthetic minority samples, which can address the problem of synthetic minority over-sampling technique (SMOTE)'s lacking the ability of overlapping whole minority space using the existing minority samples. Comparing the performance of the proposed approach with SMOTE and other improved algorithms, the experimental results show that the proposed method can not only effectively improve the classification performance of minority samples, but also significantly enhance the whole classification performance.

Key words: Imbalanced data sets；Negative immune principle；Over-sampling technique；SMOTE

1 引 言

当机器学习从理论转入实践, 不平衡数据分类问题便成为一个新的研究领域并广泛应用于商业、工业以及科学研究中[1, 2]. 不均衡数据集是指, 数据集中某些类的样本比其他类多很多, 样本多的类为多数类, 样本少的类为少数类. 目前不均衡类数据挖掘领域的研究已变得越来越重要[3], 这是因为人们认识到现实中的数据集基本上都是不均衡的, 并且这种不均衡性已经严重影响到分类算法的性能. 例如, 欺骗信用卡检测、医疗诊断等[4, 5]应用中, 往往少数类的识别率显得更为重要. 在欺骗信用卡检测中, 合法信用卡用户 (多数类) 比欺骗信用卡用户 (少数类) 多得多, 当合法信用卡被误分类为欺骗信用卡时, 银行需要额外的人力与物力来验证; 但是, 如果欺骗信用卡被误分类为合法信用卡, 则所带来的经济损失要比前一种情况大很多. 而传统的分类方法都倾向于对多数类有较高的识别率, 而对于少数类的识别率却很低. 因此, 不均衡数据集的分类问题需要寻求新的方

收稿日期: 2009-06-08；**修回日期:** 2009-08-20.

基金项目: 中国博士后科学基金项目(20090450119)；中国博士点新教师基金项目(20092304120017)；黑龙江省博士后基金项目(LBH-Z08227).

作者简介: 陶新民(1973−), 男, 安徽蚌埠人, 副教授, 从事智能信号处理、智能计算等研究；徐晶(1974−), 女, 黑龙江鸡西人, 副教授, 从事模式识别的研究.

法.

目前不均衡数据集分类问题的解决方法分为两种, 即算法层的方法和数据层的方法. 算法层的方法对分类算法进行操作, 即修改已有的分类算法或提出新的算法. 对于已有的算法, 通过调节不同类样本之间的成本函数、改变概率密度、调整分类边界等措施, 使其更有利于少数类的分类[6]. 数据层的方法又称为重抽样, 它直接对训练集进行操作, 处理后的样本用来训练分类器.

重抽样又分为欠抽样和过抽样. 最简单的欠抽样方法是, 随机地去掉某些多数类样本以减小多数类的规模, 缺点是会丢失多数类的一些重要信息; 改进的欠抽样方法是有选择地去掉对分类作用不大, 即远离分类边界或引起数据重叠的多数类样本, 以得到更理想的分类效果[7]. 欠抽样方法虽然可以降低训练集的非平衡度, 但由于其只抽取多数类的一个子集来训练分类, 很可能会忽略掉多数类中的一些有用信息. 过抽样方法通过复制少数类样本或利用已有少数类信息生成人工样本, 以增加少数类的规模, 缺点是没有增加任何新的知识, 使分类器学到的决策域变小, 从而导致过学习[8-10].

针对以上问题, 本文借鉴阴性免疫原理[11]的思想, 从过抽样角度出发, 提出一种利用阴性免疫算法生成人工少数类样本, 从而实现训练样本数据均衡的过抽样新算法. 算法通过对多数类样本学习, 生成覆盖少数类样本空间的人工少数类样本, 由于算法利用了多数类样本的先验知识, 而在不均衡数据应用中, 多数类样本数据又容易得到, 避免了通过学习少数类样本生成人工样本缺乏空间代表性的不足.

2 基于阴性免疫的过抽样算法

2.1 不均衡数据下传统过抽样算法

不均衡数据集的分类问题是机器学习领域中新的研究热点, 解决它对于完善机器学习体系、提出新的机器学习思想具有很高的理论和应用价值. 在解决不均衡数据集分类问题数据层面的方法中, 由于欠采样只取多数类的一个子集来训练分类, 很可能会忽略掉多数类中的一些有用信息, 导致分类算法性能减低. 因此人们主要集中在对过抽样算法的研究上[12-14]. 其中, 最简单的过抽样方法是通过复制少数类样本来增加少数类的规模, 其缺点是没有增加任何新的知识, 使分类器学到的决策域变小, 从而导致了过学习问题. SMOTE[8]方法扩充了传统的过抽样方法, 它不再是简单重复少数类样本, 而是通过向少数类中插入成对的邻近样本来合成新的、非重复的少数类实例. SMOTE 方法能在一定程度上降低非平衡度, 且优于传统的过抽样方法, 但由于其引入了额外的训练数据, 分类器的建立时间将会延长. 为解决该问题, 文献 [9] 对其进行了改进, 通过对少数类样本进行选择性重复, 只选择那些处于少数类与多数类边界之间的少数类样本, 根据最近邻选择方法的不同, 分别提出了 Borderl-SMOTE 算法和 Border 2-SMOTE 两种过抽样算法. [15] 提出了利用求最近邻样本均值点进而生成人工样本的 D-SMOTE 算法. 最近, [16] 利用周围空间结构信息的邻居计算公式, 提出了 N-SMOTE 过抽样算法.

上述算法都不同程度地提高了不均衡数据下分类器的性能. 然而, 这些算法都只利用已有的少数类样本信息进行学习, 如果该少数类样本集缺乏样本空间的代表性且噪音较多, 则会导致算法过度拟合. 极端情况下, 会使分类规则只涵盖一个被重复多次的样本规则, 甚至还是错误的规则. 因此, 如何解决这一问题一直是众多学者关注的重点.

2.2 基于阴性免疫的过抽样算法

通过以上分析可知, 过抽样算法由于是通过对少数类样本的学习生成人工少数类样本来实现训练数据间的均衡, 无可避免地受到已有少数类样本好坏的限制.而在非均衡数据应用中, 少数类样本是极难收集的, 这就使得人们对已有的少数类样本无法进行自由选择.

文献 [11] 提出的阴性免疫算法是通过对已类样本学习生成覆盖非己空间的检测器集合, 利用检测器实现类别的判断. 受此算法的启发, 本文提出一种新的基于阴性免疫原理的过抽样新算法 (NI-OVERSAMPLING). 该算法利用阴性免疫算法对多数类样本进行学习, 通过相互之间移动克隆变异, 生成覆盖少数类样本空间的检测器集合, 最终利用生成的具有代表性的检测器的中心点集作为人工少数类样本, 实现不均衡数据下的多数类与少数类训练数据样本间的均衡. 具体算法描述如下.

2.2.1 检测器表示及亲和度定义

算法中检测器定义为

$$d=(c,r_d),\ \ c=(c_1,c_1,\ldots,c_m),r_d\in \boldsymbol{R}. \tag{1}$$

其中: c 为检测器中心点, r_d 为检测器的半径.

检测器的亲和度反映了检测器 (抗体) 和多数类样本 (抗原) 之间的结合力, 度量公式为

$$D(x,y)=\left(\sum|x_i-y_i|^{\lambda}\right)^{1/\lambda}. \tag{2}$$

其中: $x=(x_1,x_2,\ldots,x_m),y=(y_1,y_2,\ldots,y_m)$. 本文取 $\lambda=2$ 时的欧氏距离作为亲和度的度量标准. 距离越小, 则亲和度越高.

2.2.2 检测器移动操作

设检测器 $d=(c,r_d)$ 与训练集合中多数类样本最近点的距离是 D, 即检测器的亲和度. 如果检测器的 $(D-r_s)$ 是负值, 则表示该检测器位于多数类样本空间内部, 于是该检测器将被移走或删除. 其中: r_s 为距离的阈值, $r_d=D-r_s$ 为检测器的半径. 算法依据下列公式移动检测器:

$$c^{\text{new}}=c+\alpha\frac{\text{dir}}{\|\text{dir}\|},\ \text{dir}=c-c^{\text{nearest}}. \tag{3}$$

$d=(c,r_d)$ 为当前检测器; c^{nearest} 为多数类样本空间中与当前检测器最近的抗原点, 或是与当前检测器 $d=(c,r_d)$ 距离最近的检测器 $d^{\text{nearest}}=(c^{\text{nearest}},r_d^{\text{nearest}})$ 的中心点. 这是由于当一个检测器与其他检测器大部分覆盖时, 为了保留具有代表性的少数类空间的样本, 需要对这个检测器进行移动, α 控制收敛变量和移动方向. 这种做法允许寻找具有代表性的检测器覆盖少数类样本空间.

2.2.3 检测器间覆盖度的测量标准

为了保证检测器的多样性, 防止检测器重复, 定义检测器的覆盖度的度量标准为

$$W(d)=\sum_{d\neq d'}w(d,d^{'}), \tag{4}$$

$$w(d,d^{'})=(\exp(\delta)-1)^m, \tag{5}$$

$$\delta=\Big(\frac{2r_d-r_d^{'}}{2r_d}\Big). \tag{6}$$

其中: $W(d)$ 为当前检测器 $d=(c,r_d)$ 的覆盖度, $w(d,d^{'})$ 为当前检测器 $d=(c,r_d)$ 和检测器 $d^{'}=(c^{'},r_d^{'})$ 之间的覆盖度测量标准, m 为空间的维度. 依据这种判定标准, 具有大的半径和小的覆盖度的检测器将会具有较小的值. 如果当前检测器的覆盖度超过覆盖度阈值 ξ 时, 则按式 (3) 执行检测器的移动操作.

2.2.4 检测器克隆增殖操作

在算法的每一次循环过程中, 选择少数的好的检测器构成记忆体集合 A_m, 并进行克隆操作. 克隆操作定义为

$$P_c(x_i)=({p_c}^1(x_i),{p_c}^2(x_i),\ldots,{p_c}^{\text{NC}_i}(x_i))^{\text{T}}. \tag{7}$$

其中: $P_c(x_i)$ 称为克隆算子; NC_i 为克隆的个数. $\text{NC}_i=kN/i$, $k\in(0,1)$ 为克隆常数, N 是产生检测器的规模, i 是按照检测器亲和度大小的排序号. 亲和度越低的检测器, 其克隆数目越多. 为了使程序快速收敛, 算法中需要增加克隆成熟度判断算子, 即

$$D>\text{ThresClone}M, \tag{8}$$

其中 ThresClone$M$ 为克隆成熟度判断阈值. 当亲和度 D 达到一定程度时, 可认为该检测器达到克隆成熟, 将不在下一次循环中参与克隆与变异.

2.2.5 检测器变异操作

为了保证检测器的多样性, 需要在检测器的克隆群体中随机地进行变异操作, 即

$$P_m(x_i)=({p_c}^{'1}(x_i),{p_c}^{'2}(x_i),\ldots,{p_c}^{'\text{NC}_i}(x_i))^{\text{T}}, \tag{9}$$

其中 $P_m(x_i)$ 称为变异算子. 这里引用一个自适应变异算子

$$c^{\text{new}}=c^{\text{old}}+\theta\,p. \tag{10}$$

其中: p 为随机变异方向; θ 为变异常数, 用来调整变异强度, 它与搜索空间的大小和种群规模相关, 这里取

$$\theta=\frac{1}{\sqrt{N}}\frac{1}{D}. \tag{11}$$

显然, 检测器的变异率与其亲和度成正比, 即亲和度越低 (距离越大), 变异率越小. 检测器在每次迭代过程中, 根据亲和度自适应地调整变异步长, 使其能在亲和度高的检测器周围集中搜索以提高收敛速度, 同时保持种群的多样性.

2.2.6 检测器克隆选择及消亡操作

为了能将变异后的检测器代替原有的克隆母体, 需要增加一个克隆选择算子. 若在 $x_i\in A_m$ 的变异群体中存在检测器 ${p_c}^{'}$, $D({p_c}^{'})<D(x_i)$, 则有

$$P_s(x_i)={p_c}^{'},\ x_i\in A_m, \tag{12}$$

其中: $P_s(x_i)$ 称为选择算子. 选择算子实际是将亲和度更低的检测器选入记忆集, 代替亲和度较高的母体检测器.

此外, 为了能够最大限度地覆盖少数类样本空间, 每一次循环时需要对一些差的检测器利用随机产生的检测器代替. 因此引入消亡算子

$$P_d=\text{rand}()\cdot(U-L)+L, \tag{13}$$

其中 U,L 为变量的上限和下限. 消亡算子将检测器重新初始化为定义域中的值, 在检测器的记忆集之外的集合 A_γ 中取 $1/3\times\text{Num}(A_r)$ 个亲和度最高的检测器, 运用消亡算子将其抛弃, 并利用重新初始化的检测器进行代替, 可保持种群的多样性.

2.3 基于阴性免疫的过抽样算法流程

基于阴性免疫的过抽样算法流程如下:

设 r_s 为距离的阈值;

ThresClone$M$ 为克隆成熟度阈值;

α 为收敛变量;

t 为年龄阈值, 一旦检测器达到这个年龄, 则被认为是成熟的;

ξ 为覆盖度阈值;

随机产生的检测器的群体 E_0;

成熟的检测器的群体 $M_0=\phi$;

记忆检测器的群体 $A_m=0$.

当循环次数 $i <$ num_iter 不满足停止条件时, 对于每一个检测器 $d \in E_i$, 循环以下步骤:

如果距离 $\text{dist}(d, \text{NearestSelf}) < r_s$, 则:

如果 $\text{age}_d > t$, 则检测器 d 是老的,

用一个新的随机检测器代替 d,

$$\text{age}_d = 0,$$

$$r_d = \text{Dist}(d, \text{NearestSelf});$$

否则, $r_d = \text{Dist}(d, \text{NearestSelf})$,

增加 age_d++,

计算 $\text{dir} = c - c^{\text{nearest}}$,

调整中心点

$$c^{\text{new}} = c + \alpha \frac{\text{dir}}{\|\text{dir}\|};$$

否则, 计算 $W(d)$,

如果 $W(d) > \xi$, 则检测器 d 重叠, 需要移动,

计算 $\text{dir} = c - c^{\text{nearest}}$,

调整中心点

$$c^{\text{new}} = c + \alpha \frac{\text{dir}}{\|\text{dir}\|};$$

否则, 增加 d 到集合 M_i,

$$M_i \leftarrow M_i \bigcup \{d\},$$

循环结束.

依照距离的降序排序 M_i 中的检测器成员, 选择前 $n < 1/3 \times \text{Num}(M_i)$ 的检测器作为记忆集合 A_m, 另一部分为 A_r. 再选择前 $1/2 \times \text{Num}(A_m)$ 检测器并且克隆成熟度小于 ThresClone$M$ 的检测器进行克隆, 并根据式 (9) 对克隆群体进行变异. 利用式 (11) 重新选择新的变异体到记忆集合 A_m, 在检测器的记忆集之外的 A_r 中取得 $1/3 \times \text{Num}(A_r)$ 亲和度最高 (距离小) 的检测器, 运用消亡算子予以抛弃, 然后再将其重新初始化.

$$M_i = A_m + A_r, \quad E_i = M_i,$$

循环结束.

最终输出成熟的检测器的群体 E_i 的中心点集作为覆盖少数类样本空间的代表性的人工样本.

2.4 对标称值属性的处理

在上述阴性克隆选择过抽样算法中, 计算亲和度时采用了欧氏距离, 而在现实应用中, 有很多属性是标称值, 因此对这些属性需要进行特殊处理. 本文算法对标称值属性亲和度的计算采用 VDM 距离, 对属性 i 的两个标称值 u 和 v 的距离计算如下:

$$\text{VDM}_i(u, v) = \sum_{c=1}^{C} (N_{iu}^c / N_{iu} - N_{iv}^c / N_{iv})^2. \quad (14)$$

其中: C 为类别数, N_{iu}^c 为 c 类样本中属性 i 标称值是 u 的个数, N_{iu} 为整个样本属性 i 标称值是 u 的个数, 其他类似.

在进行检测器移动时, 由于标称值数据的特殊性, 需利用如下公式实现新中心点属性 i 的标称值的确定:

$$c_i^{\text{new}} = v, \min(\text{VDM}(c_i, v) - \alpha \times \text{dir}_i). \quad (15)$$

其中: c_i^{new} 为新中心点属性 i 的标称值, c_i 为旧检测器中心点属性 i 的标称值, 新中心点属性 i 的标称值取决于与旧检测器中心点属性 i 的标称值距离相差 $\alpha \times \text{dir}_i$ 最小的标量值.

2.5 非均衡数据下的评价准则

传统的性能评估方法为了提高整个数据的分类准确性而倾向于将数据归到多数类中, 从而忽略了少数类, 因此已不再适用于非均衡数据分布的情况. 为此, 本文采用非均衡数据下的评价准则. 为便于理解, 下面给出混合矩阵定义. 在此假设使用二分类法, 将训练集分为少数类和多数类, 并将少数类称为正类, 多数类称为反类.

表 1 二分类问题的混合矩阵

	预测为正样本数	预测为反样本数
实际正类样本数	TP	FN
实际反类样本数	FP	TN

$$\text{G-Mean} = \sqrt{\frac{\text{TP}}{\text{TP+FN}} \cdot \frac{\text{TN}}{\text{TN+FP}}}, \quad (16)$$

$$\text{F-Measure} = \frac{(1+\beta^2) \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \text{Recall} + \text{Precision}}, \quad (17)$$

$$\text{Precision} = \text{TP}/(\text{TP+FP}), \quad (18)$$

$$\text{Recall} = \text{TP}/(\text{TP+FN}). \quad (19)$$

几何均值 G-Mean 是不均衡数据集学习中常用的评价准则, 它是少数类的精确度与多数类的精确度乘积的平方根. 当二者的值都大时, 几何均值才会大. 因此, 几何均值能合理地评价不均衡数据集的总体分类性能.

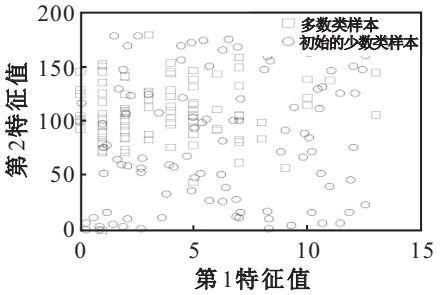
少数类的 F-Measure 也是不均衡数据集学习中有效的评价准则, 它是查准率 Precision 和查全率 Recall 的组合, 其中 β 通常取值为1. 只有当少数类的查全率和查准率的值都大时, 少数类的 F-Measure 才会大. 因此, 它能正确地反映少数类的分类性能.

3 实验结果及分析

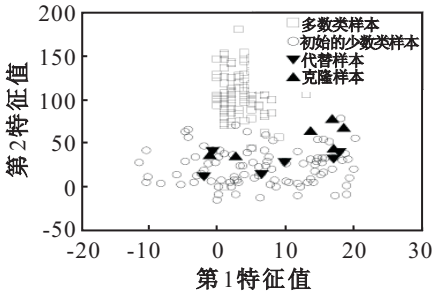
3.1 NI-OVERSAMPLING 算法生成人工样本的实验

为了直观地描述基于阴性免疫原理的过抽样算法生成人工样本的实验过程, 选自 UCI 数据库中的 PIMA 数据集作为测试数据. 其中: 数据集中多数类数目为 100 个, 少数类样本为 10 个, 属

性为8个. 实验参数设置如下: 需产生人工少数类样本的个数为90个, 根据实验经验设置距离的阈值r_s为数据集平均距离的0.012倍, 收敛变量$\alpha = 0.3\mathrm{e}^{-i/200}$, 克隆成熟度阈值为$13r_s$, 检测器成熟度阈值$t = 6$, 覆盖度阈值$\xi$为0.1, 停止准则为200次循环[11]. 实验结果如图1所示. 其中：图1(a)显示了初始化时第1特征值和第2特征值的数据分布情况；图1(b)显示了循环为100时的数据分布情况.



(a) 初始化时的数据分布



(b) 循环为100时的数据分布

图 1　PIMA 数据集 NI 过抽样算法收敛性能对比

实验结果表明, 本文提出的基于NI过抽样算法具有较好的收敛能力. 同时, 从图1中可以直观地看出, NI过抽样算法是通过对多数类样本学习来生成人工少数类样本的.

3.2　不同过抽样算法的性能对比实验

为了验证本文的基于NI的过抽样算法的性能, 利用来自UCI的不同数据集进行性能对比实验, 其中数据集的描述如表2所述.

表 2　数据集描述

数据集	样本数	属性数	类标号	少数类比例/%
Haberman	306	3	Die:Survive	26.47
Pima	768	8	1:0	34.77
Glass	214	9	1:0	15.68
Satimage	6435	36	4:remainder	9.73

过抽样算法选择SMOTE算法, Border1-SMOTE和Border2-SMOTE过抽样算法[9], D-SMOTE算法[15], N-SMOTE(K-NCN)算法[16], 随机抽样算法, 不进行抽样的算法以及本文基于NI的过抽样算法. 分类器选择C4.5和支持向量机(SVM). SVM算法参数设置为: 核函数为高斯函数, 其中核宽数为1. SMOTE算法, Border1-SMOTE, Border2-SMOTE, D-SMOTE, N-SMOTE算法的最近邻算法$K = 5$. 本文基于NI的过抽样算法的参数设置同上, 生成的人工样本数目根据不同的数据集分别计算, 采用10次交叉验证. 实验结果如表3, 表4所示.

表 3　数据集少数类 F-Measure 性能比较

方　法	数　据　集			
	Haberman	Pima	Glass	Satimage
SVM	0.061 7	0.566 4	0.612 1	0.241 2
SMOTE+SVM	0.099 2	0.715 3	0.638 4	0.645 7
Border1SMOTE+SVM	0.125 6	0.723 5	0.643 2	0.651 2
Border2SMOTE+ C4.5	0.132 1	0.709 0	0.665 6	0.637 7
DSMOTE+SVM	0.116 7	0.713 4	0.649 7	0.641 9
NSMOTE+SVM	0.129 7	0.714 0	0.668 1	0.652 2
NI+SVM	**0.420 3**	**0.729 0**	**0.754 0**	**0.718 3**
C4.5	0.453 4	0.613 4	0.621 5	0.417 2
SMOTE+ C4.5	0.475 1	0.671 2	0.666 8	0.652 1
Border1SMOTE+ C4.5	0.476 7	0.692 7	0.667 3	0.673 3
Border2SMOTE+ C4.5	0.432 2	0.705 6	0.697 3	0.647 9
DSMOTE+ C4.5	0.463 1	0.697 4	0.681 2	0.665 2
NSMOTE+ C4.5	0.459 8	0.704 9	0.698 1	0.677 1
NI +C4.5	**0.521 1**	**0.719 0**	**0.813 2**	**0.714 8**

表 4　数据集少数类 G-Mean 性能比较

方　法	数　据　集			
	Haberman	Pima	Glass	Satimage
SVM	0.373 6	0.642 3	0.531 3	0.547 8
SMOTE+SVM	0.378 5	0.809 4	0.596 7	0.613 7
Border1SMOTE+SVM	0.386 1	0.804 1	0.612 5	0.601 9
Border2SMOTE+ C4.5	0.364 8	0.812 3	0.607 7	0.634 7
DSMOTE+SVM	0.371 1	0.822 3	0.614 2	0.652 9
NSMOTE+SVM	0.383 3	0.805 3	0.632 1	0.663 0
NI+SVM	**0.421 4**	**0.831 1**	**0.763 9**	**0.767 8**
C4.5	0.416 2	0.663 3	0.699 4	0.724 3
SMOTE+ C4.5	0.475 1	0.688 9	0.748 9	0.733 7
Border1SMOTE+ C4.5	0.476 7	0.693 1	0.756 7	0.778 1
Border2SMOTE+ C4.5	0.483 1	0.703 7	0.697 3	0.748 5
DSMOTE+ C4.5	0.493 9	0.712 1	0.682 2	0.711 7
NSMOTE+ C4.5	0.517 7	0.700 9	0.728 9	0.762 2
NI +C4.5	**0.567 0**	**0.729 3**	**0.821 0**	**0.793 3**

从实验结果可以看出, 对于这4种数据集而言, 本文提出的基于阴性免疫过抽样算法的F-measure性能及G-mean都优于其他基于SMOTE及其改进的算法, 这种性能的提高是由两种算法的本质不同所决定的. 本文算法是根据多数类样本先验知识进行学习, 而SMOTE算法则是利用已知的少数类样本分布来生成人工少数类样本. 少数类样本在具体的不均衡数据应用中很难得到, 而多数类样本则很容易收集到, 在这一点上, 本文算法较其他算法具有明显优势.

3.3 不同不均衡数据比例的性能对比实验

为了显示基于NI的过抽样算法与经典SMOTE算法的本质区别, 本文选择UCI数据库中的abalone数据集和NCI数据集中的mcd作为测试数据. 其中: abalone数据集中样本总数为4 280, 属性数为8, 少数类样本个数为105; mcd数据集样本总数为29 508, 属性数为6, 少数类样本数为292, 其中两个数类样本的个数按5:1, 10:1, 15:1, 20:1, 25:1, 30:1, 35:1, 40:1 比例分别进行选取. 测试分类器选择C4.5. 为了消除随机影响, 实验进行10次交叉验证. SMOTE过抽样算法中最近邻算法 $K = 5$. NI过抽样算法中的参数设置同上述实验. 实验评测准则选用F-measure和G-mean, 实验结果如图2, 图3所示.

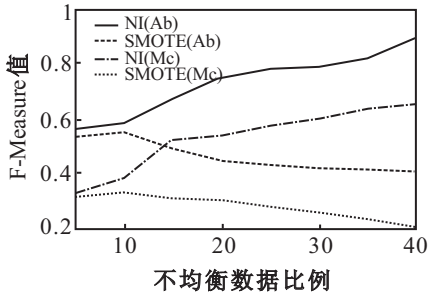


图 2 不同均衡数据比例F-Measure性能对比

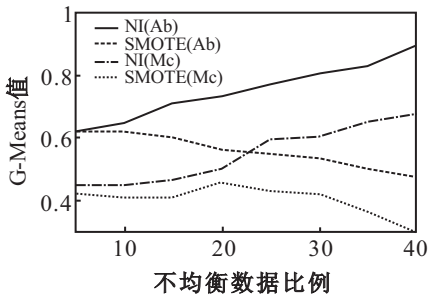


图 3 不同均衡数据比例G-Mean性能对比

从实验结果可以看出, 本文算法无论对于abalone数据集还是mcd数据集而言, 其F-measure性能及G-mean都优于SMOTE算法, 同时随着不平衡数据比例的增大, 基于本文算法的分类器的性能也随之提高. 这是因为基于NI的过抽样算法是通过多数类样本先验知识的学习实现少数类样本空间的覆盖, 所以随着多数类样本的增加, 对多数类空间的先验知识随之增强, 因此通过它们生成的少数类空间检测器的代表性和覆盖能力也有所提高, 进而基于这些样本训练后的分类器性能也就随之增大. 而SMOTE过抽样算法是基于少数类样本进行学习的. 随着不均衡比例的增大, 少数类样本的先验知识没有增加, 而需要生成的人工少数类样本反而加大. 少数类样本的自身局限性限制了生成的人工少数类样本的代表性和覆盖性, 因此随着不均衡比例的增加, 基于SMOTE算法的分类器性能反而降低.

4 结 论

本文提出一种新的不均衡数据下基于阴性免疫原理的过抽样算法. 通过分析和实验可得如下结论:

1) 借鉴阴性免疫算法的思想, 新算法利用多数类样本先验知识生成人工少数类样本, 实现了训练样本间的数据均衡. 这与传统的基于SMOTE算法的少数类样本信息进行学习有本质上的不同.

2) 利用不同的数据, 对SMOTE算法及其改进算法与本文算法进行性能对比. 实验结果表明, 本文算法在少数类的F-Measure性能和全局G-Mean性能都有显著的提高.

3) 通过实验发现, 本文算法的分类器性能随着不均衡数据比例的增大(即多数类样本数目增加)而增强, 这是由本文算法的本质所决定的. 在不均衡数据应用领域中多数类样本很容易收集, 因此该算法具有很强的实际应用价值.

参考文献(References)

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

[2] Cohen G, Hilario M, Hugonet Sax H S, et al. Learning from imbalanced data in surveillance of nosocomial infection[J]. Artificial Intelligence in Medicine, 2006, 37(5): 7-18.

[3] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction[J]. Expert Systems with Applications, 2009, 36(3): 4626-4636.

[4] Zhou Z H, Liu X Y. The influence of class imbalance on cost-sensitive learning: An empirical study[C]. Proc of the 6th IEEE Int Conf on Data Mining. Hong Kong: IEEE Press, 2006: 970-974.

[5] Liu X Y, Wu J X, Zhou Z H. Exploratory under-sampling for class-imbalance learning[C]. Proc of the 6th IEEE Int Conf on Data Mining. Hong Kong: IEEE Press, 2006: 965-969.

[6] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[7] Liu X Y, Wu J, Zhou Z H. Exploratory under-sampling for class-imbalance learning[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 539-550.

[8] Chawla N V, Bowyer K W, Hall L O. SMOTE: Synthetic minority over-sampling technique[J]. J of Artificial Intelligence Research, 2002, 16(5): 321-357.

[9] Han H, Wang W Y. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learnings[C]. Int Conf on Intelligent Computing. Hefei: IEEE Press, 2005, 3644: 878-887.

[10] 韩慧, 王文渊, 毛炳寰. 不均衡数据集中基于Adaboost的过抽样算法[J]. 计算机工程, 2007, 33(10): 207-209.
(Han H, Wang W Y, Mao B H. Over-sampling algorithm based on adaboost in unbalanced data set[J]. Computer Engineering, 2007, 33(10): 207-209.)

[11] 陶新民, 杜宝祥.基于高阶统计特征实值阴性克隆选择算法的轴承故障检测[J]. 机械工程学报, 2008, 44(7): 230-236.
(Tao X M, Du B X. Bearing fault detection using real-valued negative clone selection algorithm based on higher order statistics[J]. Chinese J of Mechanical Engineering, 2008, 44(7): 230-236.)

[12] Yen S, Lee Y, Lin C. Investigating the effect of sampling methods for imbalanced data distributions[C]. IEEE Int Conf of Systems, Man and Cybernetics, Taipei: IEEE Press, 2006: 4163-4168.

[13] Hulse J V, Khoshgoftaar T M. Experimental perspectives on learning from imbalanced data[C]. Proc of the 24th Int Conf on Machine Learning. Corvalis: IEEE Press, 2007: 935-942.

[14] Seiffert C, Khoshgoftaar T M. Hybrid sampling for imbalanced data[C]. IEEE Int Conf on Information Reuse and Integration. Las Vegas: IEEE Press, 2008: 202-207.

[15] Jorge de la C, Olac F. A distance-based over-sampling method for learning from imbalanced data sets[C]. Proc of the 20th Int Florida Artificial Intelligence Research Society Conf. Florida: IEEE Press, 2007: 634-635.

[16] García V, Sánchez J S. On the use of surrounding neighbors for synthetic over-sampling of the minority class[C]. Proc of the 8th Conf on Simulation, Modeling and Optimization. Cantabria: IEEE Press, 2008: 389-394.