文章编号: 1001-0920 (2011) 10-1535-07

基于 ODR 和 BSMOTE 结合的不均衡数据 SVM 分类算法

陶新民, 童智靖, 刘 玉, 付丹丹

(哈尔滨工程大学 信息与通信工程学院，哈尔滨 150001)

摘 要: 针对传统的支持向量机 (SVM) 算法在数据不均衡的情况下分类效果不理想的缺陷, 为了提高 SVM 算法在不均衡数据集下的分类性能, 提出一种新型的逐级优化递减欠采样算法. 该算法去除样本中大量重叠的冗余和噪声样本, 使得在减少数据的同时保留更多的有用信息, 并且与边界人工少数类过采样算法相结合实现训练样本数据集的均衡. 实验表明, 该算法不但能有效提高 SVM 算法在不均衡数据中少数类的分类性能, 而且总体分类性能也有所提高.

关键词: 不均衡数据；支持向量机算法；边界人工少数类过采样算法；逐级优化递减

中图分类号: TP18 **文献标识码:** A

SVM classifier for unbalanced data based on combination of ODR and BSMOTE

TAO Xin-min, *TONG Zhi-jing*, *LIU Yu*, *FU Dan-dan*

(College of Information and Communication Engineering，Harbin Engineering University，Harbin 150001，China. Correspondent：TAO Xin-min，E-mail：taoxinmin@hrbeu.edu.cn)

Abstract: The classification result of classical support vector machine(SVM) algorithm in the case of unbalanced data set is not satisfactory. In order to improve the SVM algorithm's classification performance under unbalanced data set, a novel under-sampling algorithm based on optimization of decreasing reduction(ODR) is presented. The algorithm is applied to under-sample the majority class instances for removal of a large number of overlapping samples of redundant and noise samples, which consequently makes reservations for the majority class instances with more useful information, and the ODR under-sampling algorithm is combined with border synthetic minority over-sample technique(BSMOTE) to achieve a balanced training sample data set. The experimental results show that the proposed method can not only improve classification performance of SVM in the minority class data, but also increase the overall classification performance.

Key words: unbalanced data；support vector machine；BSMOTE；optimization of decreasing reduction

1 引 言

支持向量机 (SVM) 是以统计学习理论为基础的一种新型机器学习方法[1]. 它克服了神经网络和传统分类器过学习、局部极值点和维数灾难等诸多缺点, 具备较强的泛化能力, 目前已成为机器学习领域的一个新的研究热点.

由于 SVM 方法属于有监督分类算法, 需要数目相同的不同类别样本进行训练才能获得较好的泛化能力. 但是, 现实生活中的很多数据样本均是不均衡的, 例如商业欺诈[2]、疾病诊断[3]、文本分类[4]等数据集. 针对不均衡数据集进行分类时, 各个类别的样本数目存在较大的差异, 导致不同类别的样本对于训练算法提供的信息不对称, 使得利用 SVM 算法处理不均衡数据时[5], 训练后得到的分类面会向少数类样本偏移, 从而使 SVM 过度拟合多数类样本点而低估了少数类样本点, 导致算法错分率增大. 因此, 如何实现 SVM 算法在不均衡数据下的正确分类便成为众多学者关注的重点.

目前, 提高不均衡数据下 SVM 算法性能的研究主要集中在算法层面和数据层面. 算法层面的方法是指对分类算法本身进行操作, 即修改已有的分类算法或者提出新的算法. 对于已有的算法, 通过调节不同

收稿日期: 2010-05-27；**修回日期:** 2010-07-26.

基金项目: 国家自然科学基金项目(61074076)；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, 副教授, 从事智能信号处理、智能计算等研究；童智靖(1986−), 男, 硕士生, 从事模式识别的研究.

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]. 数据层面研究较多的是如何将高级采样方法与SVM算法相结合. 文献[7]提出了基于过采样的代价敏感SVM算法和基于欠采样的代价敏感SVM, 文献[8]提出了基于SMOTE(synthetic minority over-sampling technique)的代价敏感SVM算法. 但是, 基于欠采样的代价敏感SVM算法由于其只随机选取了多数类的一个子集来训练分类, 在数据严重不均衡时很可能会忽略多数类中的一些有用信息, 因此分类效果很不理想[9-12]; 而基于SMOTE的代价敏感SVM算法和基于过采样的代价敏感SVM算法以及KSMOTE-SVM算法[13]虽然能解决数据的不均衡问题, 但过采样算法利用的是已有少数类信息来增加规模, 没有增加任何新的知识, 同样在训练样本严重不均时, 由于少数类样本缺乏空间代表性使得分类器学到的决策域变小, 从而导致过学习. 因此, 如何在保证数据均衡的同时, 使得保存的样本信息对决策界面的生成更加有效是提高不均衡数据下SVM分类算法性能的关键.

本文鉴于以上分析, 提出一种新的基于逐级优化递减的欠采样算法, 在实现多数类样本欠采样的同时去除多数类样本中存在的噪声和冗余信息. 进一步将该方法与只对边界样本进行过采样的BSMOTE算法相结合, 在实现训练样本均衡的同时, 又能去除训练样本中噪声样本和重复信息, 保留有用信息, 提高有效数据的利用率, 最终实现提高不均衡数据下SVM算法分类性能的目的. 将建议的ODR-BSMOTE-SVM算法与其他算法进行比较, 结果表明, 建议的算法在数据不均衡情况下分类性能较其他算法有较大幅度提高.

2 理论分析

2.1 支持向量机

SVM建立在统计学习理论中结构风险最小化原理基础上, 根据有限的样本信息, 在模型复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别样本的能力)之间寻求最佳折衷, 以期获得最好的推广能力. 它通过核函数将原始特征空间中的非线性分类界面映射到更高维的特征空间中, 以便分类界面在高维特征空间中变得线性可分, 使分类效果更好.

以两类训练样本集为例, 设给定的训练样本集为 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, $y_i\in\{+1,-1\}$, $i=1,2,\cdots,n$ 为样本类别, 核函数为 K. 构造代价函数

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i;\qquad(1)$$

$$\text{s.t. } y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\ \xi_i\geqslant 0,$$

$$i=1,2,\cdots,n.\qquad(2)$$

其中: ξ_i 为松弛变量, 表示训练样本的错分程度; C 为惩罚常数, 控制对错分样本的惩罚程度; w 和 b 分别为判决函数 $f(x)=(w\cdot x)+b$ 的权向量和阈值. 拉格朗日函数为

$$L(w,b,\alpha)=\|w\|^2/2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(w^{\mathrm{T}}x_i+b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i,\qquad(3)$$

其中 α_i 和 β_i 是拉格朗日算子. 根据KKT条件

$$\alpha_i[y_i(w^{\mathrm{T}}x_i+b)-1+\xi_i]=0,\ i=1,2,\cdots,n;$$

$$(C-\alpha_i)\xi_i=0,\ i=1,2,\cdots,n.\qquad(4)$$

对偶优化后取最优解 α^* 时应满足以下条件:

$$\alpha_i^*[y_i(w^{*\mathrm{T}}x_i+b^*)-1+\xi_{i_i}^*]=0,\ i=1,2,\cdots,n;$$

$$(C-\alpha_i^*)\xi_i^*=0,\ i=1,2,\cdots,n.\qquad(5)$$

可得 $\alpha_i^*>0$ 的样本是支持向量. 判别函数为

$$f(x)=\operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i^*y_iK(x,x_i)+b^*\Big).\qquad(6)$$

2.2 不均衡数据对SVM算法分类性能的影响

为了测试数据不均衡对SVM分类器的影响, 选用高斯函数生成的数据集作为测试样本集, 其中一类样本中心为 $(0.3,0.5)$, 另一类样本中心为 $(-0.3,-0.5)$, 方差定为0.5. SVM算法的参数设置如下: 选择高斯核函数, 核宽度为1, 惩罚常数选择为 $C=1\,000$, 两类样本数目比例分别为1:1和10:1, SVM算法的分类情况如图1所示.

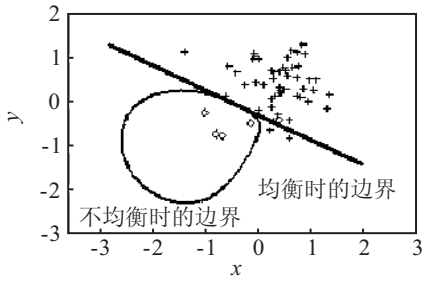


图1 数据样本比例为1:1和10:1的分类边界

由图1可见, 当数据集中两类样本数目相同时, SVM分类器的分类边界是较为理想的; 当两类样本数目相差较大时, SVM的分类边界会靠近少数类样本, 从而容易造成少数类样本错分. 此外, 靠近分类边界的多数类样本中的噪声样本对SVM分类器的训练和分类影响也很大.

上述测试表明, 数据集中不同类别样本数目的不均衡导致SVM分类器的分类边界偏向少数类, 使得错分率大大增加. 而现实中少数类样本的错分代价往

往远高于多数类样本, 所以为了提高 SVM 分类器的分类性能, 必须解决算法分类边界偏向于少数类样本的问题.

3 基于 ODR 和 BSMOTE 相结合的 SVM 分类器的设计

3.1 BSMOTE 过采样算法

过采样算法通过增加少数类样本数来降低数据集中不同样本类别之间的不均衡程度. 由于 SVM 算法的分类边界只决定于支持向量, 而支持向量存在于分类边界附近, 在复制样本时, 只有复制离边界近的样本, 才能对分类器分类性能有所改进. 因此, 为了提升有效样本的利用率和 SVM 分类器的性能, BSMOTE 过采样算法在原有 SMOTE 算法的基础上, 通过只复制分类边界少数类样本来实现训练样本均衡. 具体算法描述如下 (文中设训练样本集为 T, 少数类样本为 $F=\{f_1,f_2,\cdots,f_n\}$):

Step 1: 计算少数类样本集 F 中每一个样本在训练样本集 T 中的 k 个最近邻, 根据这 k 个最近邻对 F 中的样本进行归类. 若 k 个最近邻均是多数类样本, 则将该样本定为噪声样本, 并将其放入 N' 集合中; 反之, 若 k 个最近邻均是少数类样本, 则该样本是远离分类边界样本, 将其放入 B 集合中; 若 k 个最近邻既有多数类样本又有少数类样本, 则认为是边界样本, 存入 S 集合中.

Step 2: 设 $B=\{f_1',f_2',\cdots,f_b'\}$, 计算 B 集合中每个样本 $f_i'(i=1,2,\cdots,b)$ 在少数类样本 F 中的 k' 个最近邻 f_{ij}, 并随机选出 $s(1<s<b)$ 个最近邻, 计算它们各自与该样本间全部属性的差值 $d_{ij}=f_i'-f_{ij}(j=1,2,\cdots,s)$, 然后乘以一个随机数 $r_{ij},r_{ij}\in(0,1)$(如果 f_{ij} 是 N 集合或 S 集合中的样本, 则 $r_{ij}\in(0,0.5)$), 生成的人工少数类样本为

$$h_{ij}=f_i'+r_{ij}d_{ij},\ j=1,2,\cdots,s. \tag{7}$$

Step 3: 重复 Step 2 的过程, 直至生成人工少数类样本的个数满足要求后, 算法结束.

3.2 逐级优化递减欠采样算法 (ODR)

在多数类样本中存在着噪声样本和大量的重复信息, 这些冗余信息会严重影响 SVM 分类器的界面生成, 因此, 如何剔除这些冗余样本保留有效信息十分重要. 传统的欠采样算法只随机选取多数类的一个子集, 没能考虑采样后子集的信息是否有效. 为了实现对多数类样本集合有目的地筛选, 在 KNN 算法的基础上, 利用多数类样本对于邻域样本的影响对欠采样算法进行改进, 提出一种新的逐级优化递减欠采样算法. 该算法通过 KNN 算法评价多数类样本对邻域内样本分类影响的好坏程度, 依次删除对分类效果有负面影响或影响不大的样本, 通过删除多数类中的冗余样本来达到使样本均衡的目的. 具体算法描述如下: 设训练样本集为 T, 其中多数类样本集为 N, p 为 N 集中的样本. 定义样本 p 的关联集是指 N 集中其他样本 k 个最近邻中含有 p 的样本集, 即关联集 $A_p=\{n_{p1},n_{p2},\cdots,n_{pn}\}$, 其中样本 p 是 $n_{pi}(n_{pi}\in A_p)$ 的最近邻, 对立样本是指样本类型与 p 不同的样本. 逐级优化递减算法的流程如下:

Step 1: 针对训练样本集 T 中的每一个样本, 找到它在 T 中的 k 个最近邻, 组成该样本的最近邻链表, 根据 T 中所有样本的最近邻链表建立多数类样本集 N 中样本的关联集链表.

Step 2: 对于多数类样本集 N 中的每个样本 p, 将其关联集中的样本利用 KNN 算法进行分类, 能够正确分类的个数记为 with_p; 然后从这些样本的最近邻中去除 p, 将第 $k+1$ 个最近邻加入, 计算此时利用 KNN 算法能够被正确分类的个数, 记为 without_p.

Step 3: 比较 with_p 和 without_p 的值, 若 $\text{with}_p\leqslant\text{without}_p$, 则认为删除样本 p 对训练样本集 T 分类的影响很小, 反之认为样本 p 对分类器影响较大.

Step 4: 计算 N 中所有样本在训练样本集 T 中最近的对立样本, 并求出两者之间的欧氏距离 d_p'.

Step 5: 根据 $\text{with}_p-\text{without}_p$ 的值从大到小排列 (只在 $\text{with}_p-\text{without}_p\geqslant 0$ 的情况下), 若 $\text{with}_p-\text{without}_p$ 的值相同, 则按 d_p' 从小到大的顺序优化排列, 然后依次递减删除多数类样本, 直到多数类样本数目递减到指定的数目时, 算法结束.

算法中, 当 $\text{with}_p-\text{without}_p$ 的值为负时, 认为样本 p 是噪声样本; 当值为 0 时, 认为 p 是安全样本 (离分类边界较远的样本); 当值为正数时, 认为 p 是边界样本. 该算法能够去除多数类样本中噪声样本和安全样本, 保留边界样本, 减少欠采样时去除的有用信息, 有利于提高 SVM 分类器的分类性能.

3.3 ODR-BSMOTE-SVM 算法

虽然过采样和欠采样算法均能达到使样本均衡的目的, 但保存的样本信息对于决策界面的生成不一定有效, 因此单纯地将它们与 SVM 相结合并不能从根本上改善 SVM 算法对于少数类样本的分类性能. 因此, 本文将上述两种采样方法相结合以实现数据均衡, 给出一种基于逐级优化递减 (ODR) 并与 BSMOTE 相结合的 SVM 算法 (ODR-BSMOTE-SVM). 该算法既能去除多数类样本的噪声和重复信息, 提高数据的利用率, 又能在只增加少数类中有效位置样本信息, 保留多数类样本有用信息的情况下, 实现样本均衡. 算法的基本思想是: 首先, 设参数 α 为需要

删除的多数类样本个数与多数类和少数类样本数之间差值的比值, 根据α的初值和训练样本中多数类与少数类样本数目间的差值确定所需要删除和增加的样本数目; 然后, 分别利用逐级优化递减欠采样算法和BSMOTE算法, 按照预定值减少多数类样本, 增加少数类样本, 再将处理后的训练集利用SVM算法进行分类; 最后, 调整α值, 使分类性能达到最佳, 从而使分类器对不均衡数据具有更好的泛化能力. 算法的流程如图2所示, 最后的输出结果是在不同α值下的最优分类.

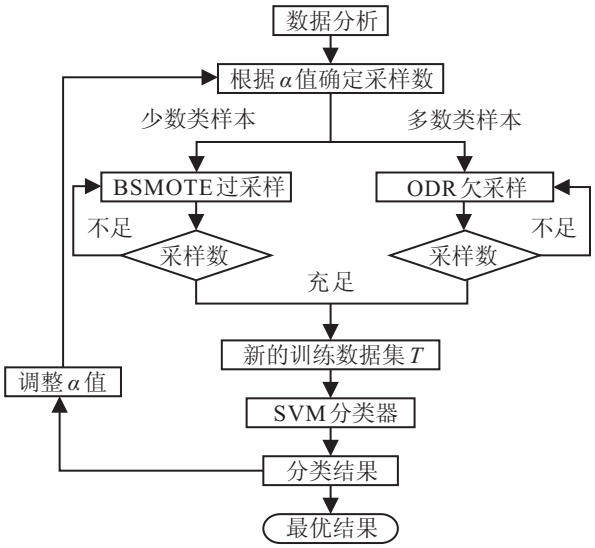


图 2 ODR-BSMOTE-SVM 算法流程

4 实验分析

4.1 实验数据

本文选用5组不同的数据集对算法进行实验, 并与SVM算法, 随机欠采样与BSMOTE结合的SVM算法(RU-BSMOTE-SVM), 核SMOTE的SVM算法(KSMOTE-SVM)和BSMOTE的SVM算法(BSOMTE-SVM)进行比较. 5组数据来自国际机器学习标准数据库UCI中的Balance-scale, Contraceptive, Haberman, Hepatitis和Pima. 其中: Balance-scale中的L和R类样本作为一类, 即多数类样本, B类样本作为少数类样本, Contraceptive中选用类别1和类别2的数据. 数据特征信息见表1.

表 1 实验数据集描述

数据集	属性	少数类数目	多数类数目	类别数
Balance-scale	4	49	576	3
Contraceptive	9	333	629	2
Haberman	3	81	225	2
Hepatitis	19	32	123	2
Pima	8	268	500	2

4.2 评估指标

在不均衡数据集分类器中, 传统的性能评估指标已不适合使用, 因为传统的性能评估是从分类器整体考虑的. 但在不均衡数据分类中, 容易将少数类样本错分, 而少数类样本数目所占比例不大, 总体的分类性能指标变化不大. 因此, 针对传统性能指标存在的缺陷, 很多学者在研究不均衡数据集分类时常使用如下指标: 定义在不均衡数据集中少数类为P, 多数类为N, FP为将多数类样本错分成少数类的数目, FN为将少数类样本错分成多数类的数目, TP和TN分别为少数类和多数类样本被正确分类的个数. 由此得到少数类样本正确率为

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}). \tag{8}$$

多数类样本正确率为

$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN}). \tag{9}$$

少数类查准率为

$$\text{Precision} = \text{TP}/(\text{FP}+\text{TP}). \tag{10}$$

几何平均正确率G_{mean}为

$$G = \sqrt{\text{Sensitivity}\cdot\text{Specificity}}. \tag{11}$$

少数类的F_{measure}为

$$F = \frac{2\text{Sensitivity}\cdot\text{Precision}}{\text{Sensitivity}+\text{Precision}}. \tag{12}$$

性能指标G综合考虑了少数类和多数类两类样本的分类性能, 如果分类器分类偏向于其中一类, 则会影响另一类的分类正确率, 从而G值会很小. 性能指标F考虑了少数类样本的查全率与查准率相结合的程度, 其中任何一个值均能影响F值的大小, 所以它能综合地体现出分类器对于少数类的分类效果.

4.3 不同算法的性能比较

为了验证本文提出的ODR-BSMOTE-SVM算法的性能, 对实验数据分别进行SVM, KSMOTE-SVM, BSMOTE-SVM, RU-BSMOTE-SVM, ODR-BSMOTE-SVM分类比较. 分类器SVM参数设置为: 核函数为高斯函数, 核宽度数为1, 惩罚因子$C = 1\,000$, KSMOTE, BSMOTE和ODR算法中最近邻算法参数k为5, 初始α值为0.2, 利用10次交叉验证法. 实验结果如表2和表3所示.

由表2和表3可见, 对于文中的5种数据, ODR-BSMOTE-SVM算法的F_{measure}性能和G_{mean}性能均优于其他的SVM改进算法. 与BSMOTE-SVM算法比较, 表明了ODR算法去除了多数类样本中的冗余和噪声, 提高了SVM算法的分类性能; 与RU-BSMOT-SVM算法的结果对比显示, 处理多数类样本时, ODR算法比随机欠采样算法更有利于提高SVM的分类性能, 这是因为随机欠采样算法删除的样本是随机选择的, 可能会丢失有用的信息, 而噪声信号和冗余信息并没有被完全删除, 不利于后续SVM分类. 从表2和

表 2　不同数据集的少数类 $F_{Measure}$ 性能比较

方法	Balancescale	Contraceptive	Haberman	Hepatitis	Pima
SVM	0.614 5	0.567 7	0.290 9	0.616 9	0.619 5
KSMOTE- SVM	0.651 3	0.569 5	0.361 6	0.621 8	0.653 4
BSMOTE- SVM	0.647 5	0.577 1	0.301 3	0.636 7	0.637 2
RU-BSMOTE-SVM	0.688 4	0.602 3	0.484 6	0.627 1	0.667 7
ODR-BSMOTE-SVM	0.760 9	0.603 5	0.550 5	0.642 1	0.675 3

表 3　不同数据集的 G_{mean} 性能比较

方法	Balancescale	Contraceptive	Haberman	Hepatitis	Pima
SVM	0.737 4	0.657 4	0.426 0	0.711 6	0.692 8
KSMOTE-SVM	0.765 0	0.658 6	0.489 1	0.723 6	0.727 3
BSMOTE-SVM	0.763 3	0.664 9	0.437 3	0.730 9	0.700 7
RU-BSMOTE-SVM	0.800 1	0.684 7	0.612 3	0.730 9	0.738 8
ODR-BSMOTE-SVM	0.936 8	0.686 9	0.676 8	0.742 6	0.744 9

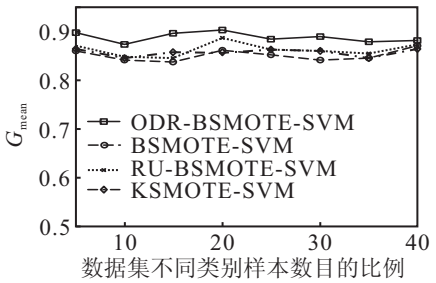
表3可以得出, ODR-BSMOTE-SVM算法的性能比KSMOTE算法有所提高, 这是由于KSMOTE算法并没有对多数类样本中的噪声和冗余信息进行处理, 而多数类中存在的噪声样本对后续SVM算法性能的影响是至关重要的.

综上所述, 本文算法利用逐级优化递减欠采样算法去除了多数类样本中的噪声和冗余信息后, 比未作处理的SVM算法的分类性能有了很大的提高.

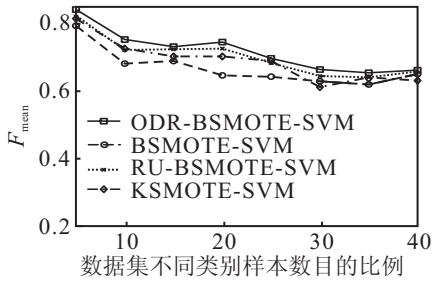
4.4　不同比例下不均衡数据分类的性能比较

为了验证ODR-BSMOTE-SVM算法在不同比例下不均衡数据中的性能, 选择UCI数据库中的page blocks classification数据集作为测试数据. 选取其中50个少数类样本数, 样本总数为5 473, 样本属性为9. 在保持少数类样本数目不变的情况下, 将两类测试数据的个数按5:1, 10:1, 15:1, 20:1, 25:1, 30:1, 35:1, 40:1的比例进行选取; 然后利用10次交叉验证法进行测试, 将测试结果与BSMOTE-SVM, RU-BSMOTE-SVM, KSMOTE-SVM算法的结果进行比较, 其中ODR-BSMOTE, BSMOTE, KSMOTE算法的k值为5, 初始α值为0.2. 比较算法在不同比例下的F_{measure}性能和G_{mean}性能. 测试结果如图3所示. 从图3中可以看出, 对于不同比例的page blocks classification数据集而言, ODR-BSMOTE-SVM算法的G_{mean}性能和F_{measure}性能均比BSMOTE算法优越, 这是因为ODR-BSMOTE-SVM算法综合了BSMOTE算法和ODR算法的优势形成的.

由图3可见, 当不均衡数据比例较小时, ODR-BSMOTE-SVM与BSMOTE-SVM算法相比优越性能显著, 体现在ODR算法去除多数类样本对分类性能影响较大. 而随着样本比例的增加, 由于α值固定, 需要利用ODR算法进行删除的多数类样本数随之增多, 从而导致一些有用的多数类信息被去除, 对后续的SVM算法不利. 所以在α值固定的情况下, ODR算法随着样本类别比例的增大对于SVM算法分类性能的提升影响较小. 而由于RU-BSMOTE-SVM算法存在较大的随机性, 在不同样本比例下, ODR-BSMOTE-SVM算法的分类性能总是明显优于RU-BSMOTE-SVM算法.



(a) G性能

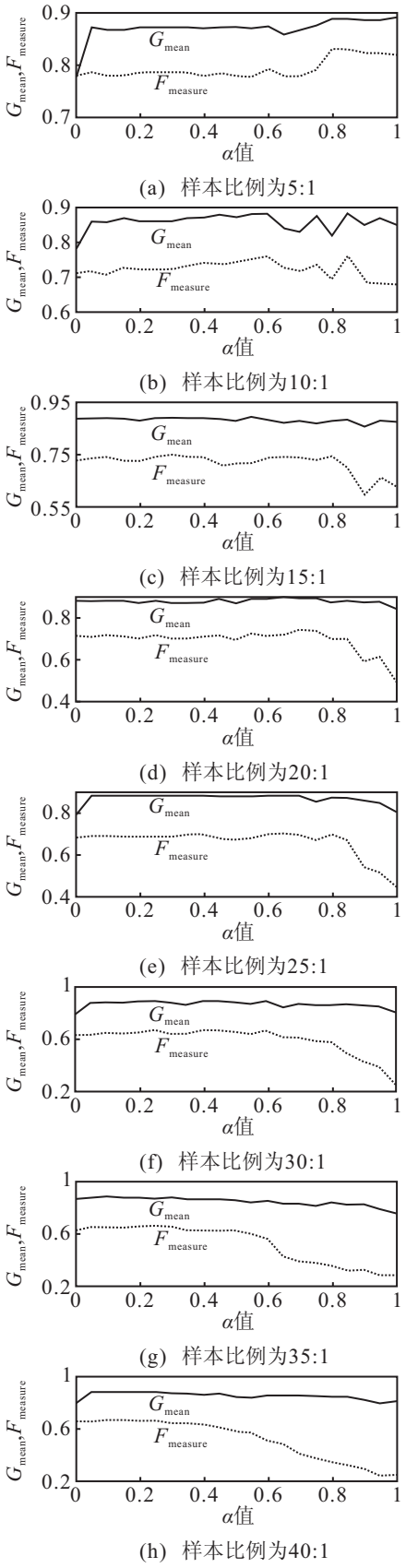


(b) F性能

图 3　不同比例训练数据各种方法 G 和 F 性能比较

4.5　参数 α 对算法性能的影响

为了测试参数α在不同样本比例的情况下对算法的影响, 选用page blocks classification数据集作为测试数据, 用于测试的少数类样本数为50, 将两类测试数据的个数按5:1, 10:1, 15:1, 20:1, 25:1, 30:1, 35:1, 40:1的比例进行选取, 利用10次交叉验证法测试在(0,1)区间中参数α的值以步长0.05均匀增加时算法分类性能的变化. 测试结果显示了在不同比例的情况下参数α对于算法分类的F_{measure}性能和G_{mean}性能的影响, 如图4所示.



(a) 样本比例为5:1

(b) 样本比例为10:1

(c) 样本比例为15:1

(d) 样本比例为20:1

(e) 样本比例为25:1

(f) 样本比例为30:1

(g) 样本比例为35:1

(h) 样本比例为40:1

图 4 不同样本比例下的 α 对分类性能的影响

由图4可见, 当α在0附近时, 随着α的增大, 算法分类性能有所提高, 因为当α为0时, 整个算法就是BSMOTE算法, 没有去除多数类样本中噪声的影响, 所以算法性能没有显著提高. 随着α值的增大(即采用ODR算法进行欠采样), SVM算法分类性能逐渐提高, 这也进一步验证了本文将ODR和BSMOTE方法相结合思想的正确性和有效性. 当α在中间区间时, 算法随α值的变化分类性能变化不大; 在α值接近1的一段区间内, 随着α值的增大, 整体分类性能有所降低, 少数类分类性能急剧降低, 尤其对于大比例不均衡数据而言. 这是由于在该区间内, 根据ODR算法按参数逐级递减优化的机制, 若α值增大利用ODR算法进行大量多数类样本欠采样, 则会导致部分有用信息损失, 降低了原有多数类样本的空间代表性, 进而影响SVM分类算法的性能. 实验结果可以表明, 当选取合适的α值时, 可以得到分类效果最优的ODR-BSMOTE-SVM算法, 同时也进一步说明了单独采用欠采样或过采样方法来提高不均衡数据下SVM算法的分类性能, 其实现效果均不理想。

5 结 论

在数据层面上改善不均衡数据下SVM算法分类性能的过程中, 利用过采样算法在均衡数据的同时会增加大量重复信息, 进而增加分类器的计算量和训练时间. 极端情况下, 若该少数类样本集缺乏空间代表性, 则会导致算法过学习. 另外, 欠采样在极端情况下会导致部分有用信息丢失, 影响SVM算法的分类性能. 因此, 本文将两种采样方法相结合, 避免只使用一种采样方法缺陷的发生. 利用提出的逐级优化递减欠采样方法可以去除多数类样本中存在的噪声样本和冗余信息, 同时利用BSMOTE算法对少数类的边界样本进行过采样, 对多数类和少数类样本同时进行采样, 可以提高有效数据的利用率, 实现数据均衡. 实验结果表明, 该方法不仅能提升不均衡数据下SVM算法对于少数类的分类性能, 而且使用新数据集训练的SVM分类器整体分类效果更加理想. 需要说明的是, 本文没有对ODR算法中KNN参数对于算法性能的影响以及ODR与核SMOTE算法的结合进行研究, 这是下一步的研究方向.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory[M] . New York: Springer, 2000: 138-167.

[2] Ezawa K J, Singh M, Norton S W. Learning goal oriented Bayesian networks for telecommunications management[C]. Proc of the 13th Int Conf on Machine Learning. San Fransisco: Morgan Kaufmann, 1996: 139-147.

[3] Chawla N V, Bowyer K W, Hall L O. Synthetic minority over-sampling technique[J]. J of Artificial Intelligence Research, 2002, 16(3): 321-357.

[4] Zheng Z H, Wu X Y, Srihar I R. Feature selection

for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004, 6(1): 80-89.

[5] Van H J, Khoshgoftaar T M. Experimental perspectives on learning from imbalanceed data[C]. Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 143-146.

[6] 林智勇, 郝志峰, 杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究, 2008, 25(2): 332-336.
(Lin Z Y, Hao Z F, Yang X W. Current state of research on imbalanced data sets classification learning[J]. Application Research of Computers, 2008, 25(2): 332-336.)

[7] 李刚. 代价敏感的支持向量机监督学习研究[D]. 南京: 南京师范大学教育科学学院, 2007.
(Li G. The research of cost sensitive SVM supervised learning[D]. Nan Jing: Institute of Education Science, Nanjing Normal University, 2007.)

[8] Gong M G, Du H F, Jiao L C. Optimal approximation of linear systems by artificial immune response[J]. Science in China, 2006, 49(1): 63-79.

[9] Jo T, Japkow I N. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 203-231.

[10] Li P P, Chan P K, Fang W. Hybrid kernel machine ensemble for imbalanced data sets[C]. Proc of the 18th Int Conf on Pattern Recognition. Washington: IEEE Computer Society, 2006: 1108-1111.

[11] 张琦, 吴斌, 王柏. 非平衡数据训练方法概述[J]. 计算机科学, 2005, 32(10): 181-183.
(Zhang Q, Wu B, Wang B. A survey on imbalanced data learning method[J]. Computer Science, 2005,32(10): 181-183.)

[12] Hongshik A, Hojin M, Melissaj F. Classification by ensembles from random partitions of high-dimensional data[J]. Computational Statistics and Data Analysis, 2007, 51(12): 6166-6179.

[13] 曾志强, 吴群, 廖备水. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报, 2009, 39(11): 2489-2495.
(Zeng Z Q, Wu Q, Liao B S. A classfication method for imbalance data set based on kernel smote[J]. Acta Electronica Sinica, 2009, 39(11): 2489-2495.)

[14] Xiao M S, Guo L F. SVM-based data editing for enhanced one-class classification of remotely sensed imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2008, 2(5): 189-193.

[15] Yi L, Lee Y. Support vector machines for classification in nonstandard situations[J]. Machine Learning, 2002, 46(2): 191-202.

[16] Rehan A, Stephen K, Nathalie J. Applying support vector machines to imbalanced datasets[J]. Machine Learning: ECML, 2004, 3201(12): 39-50.

[17] 姚程宽. 不平衡样本集中 SVM 的应用综述[J]. 计算机应用与软件, 2008, 29(9): 1-2.
(Yao C K. A survey of SVM in unbalanced data sets[J]. Computer Applications and Softwar, 2008, 29(9): 1-2.)

(上接第1534页)

[4] Oort van E R, Sonneveldt l, Chu Q P, et al. Modular adaptive input-to-state stable backstepping of a nonlinear missile model[C]. AIAA Guidance, Navigation, Control Conference and Exhibit. Hilton Head: AIAA Inc, 2007.

[5] Morse A S. Global stability of parameter-adaptive control systems[J]. IEEE Trans on Automatic Control, 1980, 25(3): 433-439.

[6] 张梅. 一类随机系统自适应控制的模块化设计[D]. 合肥: 中国科学技术大学自动化系, 2007: 1-62.
(Zhang M. Modular design of adaptive controller for stochastic nonlinear systems[D]. Hefei: Department of Automation, University of Science and Technology of China, 2007: 1-62.)

[7] Goodwin G, Mayne D. A parameter estimation perspective of continuous time model reference adaptive control[J]. Automatica, 1987, 23(1): 55-70.

[8] Sousa E. Insights on a sign-preserving numerical method for the advection-diffusion equation[J]. Int J for Numerical Methods in Fluids, 2008, 61(8): 864-887.

[9] Marc L S, Anthony B P. Nonlinear adaptive flight control with genetic algorithm design optimization[J]. Int J of Robust and Nonlinear Control, 1999, 9(14): 1097-1115.

[10] Sun Y, Zhang W G, Zhang M. Modified particle swarm optimization based on immune clone principle and for analog-matching of equivalent system[C]. IEEE Int Conf on Automation and Logistics. Shenyang, 2009: 1136-1138.

[11] Snell S A, Enns D F. Nonlinear inversion flight control for a supermaneuverable aircraft[J]. J of Guidance, Control and Dynamics, 1992, 15(4): 976-984.