耦合负类样本裁剪与非对称错分惩罚的非均衡 SVM 算法

高雷阜,赵世杰,于冬梅,徒 君
(辽宁工程技术大学优化与决策研究所,辽宁阜新 123000)

摘 要: 针对标准支持向量机(SVM)识别非均衡数据往往会出现最优超平面倾向性和正类样本大量错分的现象,探讨 SVM 识别非均衡数据失效的原因及对策;考虑到 SVM 最优超平面仅由少量支持向量完全决定的特性,提出一种基于负类边界样本裁剪策略的 SVM 数学模型. 鉴于该模型需经多次负类数据的"训练-裁剪"过程才能较好地识别正类样本且较为费时,以等效的一次性裁掉更多样本的裁截面技术作为替代,提出一种耦合负类样本裁剪与非对称错分惩罚的非均衡 SVM 算法,并利用改进正余弦优化算法优化裁剪偏移量以提高算法的非均衡数据处理能力. 数值实验结果验证了裁剪偏移量的优化必要性、改进正余弦优化算法的较强优化性能和改进 SVM 算法对非均衡数据的较好识别性能.

关键词: 非均衡数据; 支持向量机; 边界样本; 裁截超平面; 非对称错分惩罚; 正余弦优化算法
中图分类号: TP181, TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)12-2978-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.12.021

Unbalanced Support Vector Machine Coupling Negative-Samples Cutting with Asymmetric Misclassification Cost

GAO Lei-fu, ZHAO Shi-jie, YU Dong-mei, TU Jun
(*Institute of Optimization and Decision, Liaoning Technical University, Fuxin, Liaoning* 123000, *China*)

Abstract: Optimal hyperplane tendency and a large number of positive sample misclassifications often appear when the standard support vector machine (SVM) is employed to classify unbalanced data. So several causes and corresponding countermeasures for the perspective of SVM misclassifying unbalanced data are discussed. Considering the characteristics of SVM that optimal hyperplane is only decided by a small amount of support vectors, a novel SVM mathematical model based on negative boundary sample cutting strategy is constructed. However, this model has better recognition performance on positive samples only when the "training-cutting" step of negative samples is carried out many times, which is a time-consuming process. To replace it with the equivalent cutting hyperplane technique which can cut more negative samples at one time, an unbalanced SVM algorithm coupling negative-samples cutting with asymmetric misclassification cost is proposed. To further enhance the classification ability of this algorithm on unbalanced data, an improved sine cosine algorithm (ISCA) is presented to optimize the biased constant of the cutting hyperplane. Experimental results verify the optimized necessity of the biased constant of the cutting hyperplane, the advanced optimization performance of ISCA algorithm and the outstanding recognition performance of the proposed algorithm on unbalanced datasets, respectively.

Key words: unbalanced data; support vector machine; boundary sample; cutting hyperplane; asymmetric misclassification cost; sine cosine algorithm

1 引言

支持向量机(Support Vector Machine, SVM)[1]是 Vapnik 等根据统计学习理论提出的一种机器学习方法,以最优超平面技术实现推广泛化性能的控制、软间隔改善机器的容错性能、核思想保证线性、非线性问题

收稿日期:2016-07-20;修回日期:2016-10-14;责任编辑:梅志强
基金项目:教育部高等学校博士学科点专项科研基金联合资助项目(No. 20132121110009);国家自然科学基金青年基金项目(No. 51704140);辽宁省教育厅基金项目(No. L2015208, No. LJYL043)

的适用性,在处理小样本、非线性和高维数据时有效克服传统学习算法的局部极值、维数灾难等问题而在垃圾邮件识别[2]、文本分类[3]、大规模分类[4]等领域得到广泛应用. 标准 SVM (Standard SVM,S^2VM)在处理类间数据量(/分布)相当的数据时具有较好的识别性能;但对医疗诊断、故障识别等领域存在的类间数据量不相当的非均衡数据(定义少数类为正类,多数类为负类)时,则会出现明显的分类倾向性[5]—最优超平面过于趋近或贯穿于正类数据集而造成正类样本的大量错分. 鉴于正类样本所蕴含的重要潜在价值,若错分将会造成较大的潜在损失或风险. 因此,如何提高 SVM 对正类样本的识别性能已成为非均衡问题研究的一个重要方向.

目前,处理非均衡数据的策略主要包括平衡类间均衡度的数据层面方法与改进 S^2VM 的算法层面方法等. 数据层面方法主要通过重采样方式平衡数据集的类间均衡度,如增加正类样本的过采样(Over sampling)[6,7]、减少负类样本的欠采样(Under sampling)[8]和混合采样[9]等. 算法层面的方法则是通过 S^2VM 的改进以适应非均衡数据识别:Zhang 等[10]以保角变换构造缩放核提出适定性的改进 SVM 算法;Maldonado 等[11]通过无穷范数松弛 Vapnik-Chervonenkis(VC)维的界提出 SOCP-SVM 算法;Datta 等[12]将决策边界平移与不同正则项惩罚相组合提出 NBSVM 算法;Jian Chuanxia 等[13]基于支持向量与非支持向量的差异性贡献提出 DCS 算法.

鉴于 SVM 利用最优超平面(Optimal Hyperplane,OH)实现数据的有效分类,且 OH 仅由稀疏的、少量支持向量所完全决定[1],因此,为改善 S^2VM 的非均衡数据识别性能,需要调整支持向量样本位置以实现 OH 的平移. 当前数据层面算法中的过/欠采样技术,由于采样过程存在较强随机性,采样前后原支持向量位置可能未发生实质性改变,继而导致 OH 移动不明显且不能显著改善正类样本识别性能;S^2VM 识别非均衡数据失效的直观原因是 OH 位置不当——过于趋近或嵌入正类. 鉴于此,提出一种耦合负类边界样本裁剪与代价敏感化方法的改进 SVM 算法:通过负类边界样本的裁剪实现支持向量改变以向远离正类方向移动 OH,通过代价敏感的非对称性惩罚机制降低正类样本的错分概率.

2　支持向量机(SVM)理论

支持向量机[1,14]通过构造最优超平面 OH 将待分类输入样本划分为两类,即位于 OH 两侧的样本各为一类. 设训练样本集为 $(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$,$\boldsymbol{x}\in R^N$,$\boldsymbol{y}\in\{-1,1\}$. 线性问题 SVM 对应的凸二次规划为:

$$\begin{aligned}&\min_{\boldsymbol{\omega}}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{N}\xi_i\\&\text{s.t. } y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\\&\xi_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{1}$$

引入 Lagrange 函数可将式(1)转化为:

$$\begin{aligned}L(\boldsymbol{\omega},b,\boldsymbol{\lambda})&=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{N}\xi_i\\&-\sum_{i=1}^{N}\lambda_i[y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+b)-1+\xi_i]-\sum_{i=1}^{N}\alpha_i\xi_i\end{aligned}\tag{2}$$

其中,$\lambda_i\geqslant 0,i=1,2,\cdots,N$. 式(1)的对偶形式为:

$$\begin{aligned}&\max Q(\boldsymbol{\lambda})=\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{N}y_iy_j\lambda_i\lambda_j(x_i\cdot x_j)\\&\text{s.t }\sum_{i=1}^{N}y_i\lambda_i=0,0\leqslant\lambda_i\leqslant C\end{aligned}\tag{3}$$

根据 Karush-Kuhn-Tucker(KKT)条件可得线性 SVM 的分类决策函数为:

$$f(x)=\operatorname{sgn}\Big(\sum_{SV}y_i\lambda_i(x_i\cdot x_j)+b\Big)\tag{4}$$

对非线性问题,利用非线性映射 $\boldsymbol{\Phi}$ 将低维空间数据映射到高维 Hilbert 空间中以实现数据的线性可分性和相似性度量. 根据 Mercer 核与内积运算的关系有 $K(x_i,x)=(\boldsymbol{\Phi}(x_i),\boldsymbol{\Phi}(x))$,可有效避免高维特征空间中数据内积运算的"维数灾难". 目前常用的核函数有线性核 $K_L=\boldsymbol{x}'\cdot z$、多项式核 $K_P=(g\cdot\boldsymbol{x}'\cdot z+b)^D$ 和 RBF 核 $K_R=\exp^{-g\|x-z\|^2}$. 非线性问题的 Lagrange 对偶形式为:

$$\begin{aligned}&\max Q(\boldsymbol{\lambda})=\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{N}y_iy_j\lambda_i\lambda_j(\boldsymbol{\Phi}_{x_i}\times\boldsymbol{\Phi}_{x_j})\\&\text{s.t }\sum_{i=1}^{N}y_i\lambda_i=0,0\leqslant\lambda_i\leqslant C\end{aligned}\tag{5}$$

则非线性 SVM 的分类决策函数为:

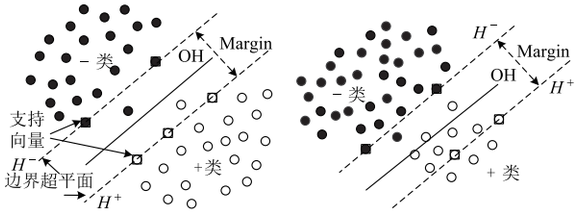
$$f(x)=\operatorname{sgn}\Big(\sum_{SV}y_i\lambda_iK(x_i,x_j)+b\Big)\tag{6}$$

3　改进的非均衡 SVM 算法

3.1　S^2VM 处理非均衡数据的失效原因

S^2VM 同其他常见机器学习方法一样(如贝叶斯分类器),一般以分类准确率最大化为评价准则,且分类数据需满足如下假设:正负类样本服从(近似)相同的概率分布、且具有(近似)相等的数据量和分散程度等特点,此时 S^2VM 的最优超平面 OH 示意见图 1(*a*). 但在处理非均衡数据时,由于数据难以满足 S^2VM 适应的假设前提而造成 OH 过于趋近数据量小、分散程度大的正类并引起正类样本的严重错分,如图 1(*b*)所示.

由图 1 分析可知:S^2VM 错分正类样本的直观原因是 OH 位置不当—过于趋近或嵌入正类样本,对应的理论原因则是由 OH 数学表达式 $\boldsymbol{\omega}^*\cdot\boldsymbol{x}+b^*=0$ 中平移偏置量



(*a*) S²VM对均衡数据的最优超平面位置　(*b*) S²VM对非均衡数据的最优超平面位置

图1　S²VM处理均衡/非均衡数据的OH位置示意

$b^* = b_0 - \dfrac{b^+ + b^-}{2} = y_j - \sum_{i=1}^{N} y_i \lambda_i^* K(x_i, x_j)$所引起的:

(1)由于数据的非均衡特性,S²VM 分类决策函数中偏置 b^* 因边界超平面的偏移量 b^+ 和 b^- 优化不当而造成 OH 偏移向正类. 因此,为弱化数据集的非均衡性,常采取的方法是过采样、欠采样及融合采样等.

(2)由 KKT 条件的互补条件知: $\sum_{i\in I^+} \lambda_i = \sum_{i\in I^-} \lambda_i$.

由于正类数据量相对较少,可能造成正类支持向量数目也相对较少而无法较好支撑 OH 并偏移向正类;同时由于正负类支持向量的 Lagrange 乘子之和相等,表明正类支持向量应具有更大的乘子值而表现出更强的偏移敏感性. 因此,需要赋予乘子值较大的正类样本更大的错分惩罚以降低正类样本的错分可能性.

(3)核函数 $K(\cdot)$ 的选取与设计也将影响 S²VM 的 OH 位置,不同核函数将原始非均衡数据映射到不同的高维 Hilbert 空间中并表现出不同的线性可分性和数据相似性度量. 因此,可通过选择/构造适定的核函数 $K(\cdot)$ 或核矩阵以改善 S²VM 的非均衡数据处理能力.

3.2　适应于非均衡数据的 S²VM 改进策略

S²VM 识别非均衡数据时造成大量正类样本错分的外延是 OH 过于趋近或贯穿于正类数据,因此,改进 S²VM 的直观策略是调整正负类边界超平面的偏移量 b^+ 和 b^- 以增大 OH 与正类边界的距离并缩小与负类间距. 该策略理论上等价于改变式(1)的约束条件,对应的 SVM 数学模型为:

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C^+ \sum_{i\in I^+} \xi_i + C^- \sum_{i\in I^-} \xi_i \\ &\text{s.t. } y_i(\boldsymbol{\omega}^{\mathrm{T}} x_i + b) \geqslant b^+ - \xi_i, \forall x_i \in I^+ \\ &\quad y_i(\boldsymbol{\omega}^{\mathrm{T}} x_i + b) \geqslant b^- - \xi_i, \forall x_i \in I^- \\ &\quad \xi_i \geqslant 0, i = 1,2,\cdots,N \end{aligned} \tag{7}$$

其中,$b^+ \in [1,2)$,$b^- \in (0,1]$且满足 $b^+ + b^- = 2$;C^+ 和 C^- 分别为正和负类样本的错分惩罚且满足 $C^+/N^- = C^-/N^+ = C_0$(N^+ 和 N^- 分别为正、负类样本数目,$C_0 \in (0, +\infty)$为常数). 当数据均衡时有 $N^+ \approx N^-$,$b^+ = b^- = 1$,此时,式(7)将退化为 S²VM.

按式(7)改进 S²VM(记作 ISVM)等效于在 S²VM 正负类边界超平面间隔内移动 OH,而非线性情形则是在相应特征空间内进行 OH 平移。由于 S²VM 的 OH 只由稀疏的、少量支持向量决定[1],远离异类的样本对其几乎无任何支撑作用,因此,为改善 ISVM 的非均衡数据识别性能,需要将负类样本的当前支持向量从训练集中删除以向外推移 H^- 并增大正负类的边界间距,即完成一次“训练-删除”过程. 由于负类样本数据量大且分散程度小,样本删除前后相邻两次 ISVM 训练所得的负类样本支持向量差异性可能不显著,需经多次“删除-再训练”过程才能较好地识别正类样本。但在实际算法设计中终止准则通常也难以预先设置:若以训练次数为终止却无法预估算法执行次数;若以识别准确率为终止也无法预知 S²VM 对特定数据集的分类阈值.

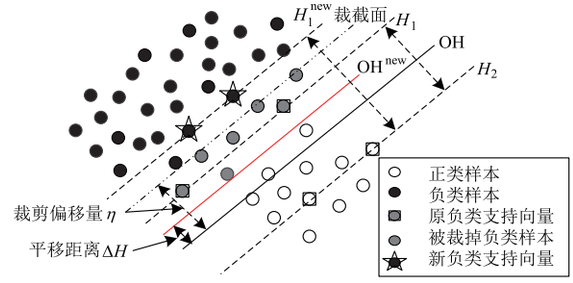


图2　基于裁截面技术的ISVM裁剪示意图

鉴于支持向量的反复“训练-删除-再训练”过程实质上是对近正类的负类样本进行适度删减,等效于基于特定裁截面的一次性负类样本裁剪过程,即一次“训练-裁剪”过程。该裁剪技术可一次性裁掉更多的负类样本,有效减少 ISVM 的“训练-删除-再训练”过程以降低算法的执行复杂度. 图 2 为基于裁截面技术的 ISVM 裁剪示意(假定裁剪中 OH 法方向保持不变).

3.3　基于耦合策略的非均衡 SVM 算法

为提高 S²VM 的非均衡数据识别性能,利用裁截面技术对近正类的负类样本进行裁剪并融入错分惩罚机制而提出耦合负类样本裁剪与非对称错分惩罚的 SVM 算法(SVM Coupling negative-samples Cutting with asymmetric misclassification Cost, C³SVM). C³SVM 算法利用裁截面技术裁剪负类样本,不仅可弱化正负类样本间的非均衡度、减少训练数据量、提高算法执行效率,而且可有效扩大正负类间的边界间隔、实现 OH 向负类数据集移动以改善正类样本识别率;非对称错分惩罚则通过增大正类的错分惩罚以降低正类样本的错分概率和敏感程度、改善正类的识别效率. 虽然负类样本的裁剪可能会造成一定的负类信息丢失,但相较于具有重要价值的正类识别率提高而言这种信息丢失是可以容忍的,如医疗诊断中,健康人因某些信息丢失被初判为病人,仍可通过进一步检查再辨为非病人,但病人均已被

准确识别而避免误判所造成的严重后果.

C³SVM 识别非均衡数据的执行伪码见算法 1.

算法 1　C³SVM

1:设定初始惩罚参数 C_0,得到正负类初始错分惩罚 C^+ 和 C^-;
2:训练带非对称错分惩罚的 SVM 模型,得到当前最优超平面 OH 和法向量 $\boldsymbol{w}$;
3:计算负类样本到超平面 OH 的距离向量并按升序排列;
4:选择适当的裁剪偏移量 η 得到实现负类样本裁剪的裁截面;
5:根据偏移量 η 对负类样本进行裁剪(距离小于 η 的负类样本被裁剪掉),得到裁剪后的新的负类样本集;
6:更新正负类样本的错分惩罚 C^+ 和 C^-;
7:再次训练 SVM 模型,得到预测 model 并输出相关数据.

C³SVM 算法中裁剪偏移量 η 的选择影响负类样本的裁剪效果,继而影响 ISVM 对非均衡数据的识别性能:当 η 选取过小时,因负类样本的数据量大且分散程度小,可能导致裁减效果不佳而使 OH 向负类方向移动较小且未能显著改善 SVM 的正类识别率;当 η 选取过大时将会导致大量负类样本被裁掉,虽然能够有效识别正类样本,但因负类样本过度裁剪而使 OH 过于倾向于负类并引起负类样本的大量错分和样本属性的逆转现象,即由原数据量多的负类逆变为裁剪后数据量少的假"正"类. 因此,为提高 C³SVM 算法对非均衡数据的正类识别率和整体分类性能,需在裁剪偏移量 η 的有效裁剪区间内进行优化选择.

4　ISCA 算法优化 C³SVM 的裁截偏移量 η

4.1　SCA 算法及其改进算法

正余弦优化算法(Sine Cosine Algorithm,SCA)[15]是 Mirjalili 利用正-余弦函数提出的一种基于种群的元启发式优化算法. 它通过多搜索体的并行迭代寻优以获得优化问题的近似最优解,每个搜索体类似于遗传算法的染色体或粒子群优化算法[16]的粒子. SCA 算法的迭代优化过程由探索和开采 2 个阶段并耦合 r_1、r_2、r_3 和 r_4 等 4 个关键参数来实现,其执行伪码见算法 2.

算法 2　正余弦优化算法(SCA)

1:初始化搜索体种群(解)($\boldsymbol{X}$):
2:执行
3:　　计算每个搜索体的目标函数值:
4:　　更新当前最优解 $P=\boldsymbol{X}*$
5:　　更新模型参数 r_1,r_2,r_3,r_4
6:　　更新搜索体种群($\boldsymbol{X}$)
7:判断 当前迭代次数 t 是否小于预设最大迭代次数
8:输出全局最优解

在探索阶段,SCA 算法以较高概率将随机解与原解集相组合以实现最优解潜在区域的探索;在开采阶段,随机解是递变的且弱于探索阶段. 为便于刻画这 2 个阶段,构造如下的个体位置更新方程:

$$X_i^{t+1}=\begin{cases}X_i^t+r_1\times\sin(r_2)\times|r_3P_i^t-X_i^t|, & r_4<0.5\\ X_i^t+r_1\times\cos(r_2)\times|r_3P_i^t-X_i^t|, & \text{otherwise}\end{cases}\tag{8}$$

其中,X_i^t 和 X_i^{t+1} 分别为第 t 次和第 $t+1$ 次迭代时第 i 维分量,P_i^t 为第 t 次迭代时最优解的第 i 维分量.

显然,式(8)包含 SCA 算法 4 个关键参数,且每个参数均有其特定作用:参数 r_1 控制子代搜索体的移动区间界值,参数 r_2 表示移向或远离当前最优解的移动距离,参数 r_3 实现对当前最优解的随机赋权以增强($r_3>1$)或减弱($r_3<1$)目标点对移动距离的影响,参数 r_4 通过阈值的比较判断实现正-余弦函数的选择性继承. 特别地,参数 r_1 折衷算法的探索和开采性能,控制迭代时正余弦函数的波动振幅限值,其表达式为

$$r_1=a-t\frac{a}{T}\tag{9}$$

其中,t 为当前迭代次数,T 为最大迭代次数,a 为常数. 图 3 点线间线表示 SCA 算法中参数 r_1 的限制作用.

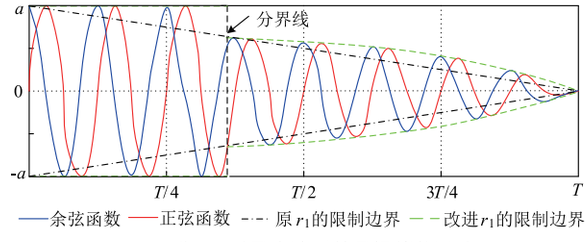


图3　改进r_1对正余弦函数区间值的影响

由 SCA 算法的基本知识可知:参数 r_1 直接影响算法的探索和开采性能,不同的 r_1 递变形式将引起 SCA 算法的优化性能差异. 在基本 SCA 算法中,参数 r_1 是线性递减的,但迭代寻优前期由于当前最优解往往离全局最优解相对较远,较快的 r_1 递减形式并不利于搜索空间的充分探索,甚至会增加陷入局部极值的可能性. 因此,为保证算法较好的全局寻优性能,参数 r_1 的递变过程应遵循的规则为:在迭代前期,参数 r_1 应维持在一个相对较大值以保证 SCA 算法对搜索空间的充分探索;随迭代次数的增加直至后期,r_1 的递减趋势应越趋增大以保证算法后期较好的开采性能和收敛效率.

为改善 SCA 算法的整体优化性能,遵循上述递变规则提出一种非分段的线性参数 r_1 递变函数为:

$$r_1=\begin{cases}a, & t<\dfrac{1-\zeta}{a}T\\ -(a+\zeta)^{\mp}+a+\zeta, & \text{otherwise}\end{cases}\tag{10}$$

其中,$\zeta\in[0,1]$为可调节常数. 不同 ζ 值将导致不同的 r_1

递变过程,本文设 $\zeta=0.6$. 图 3 展示 ISCA 算法中 r_1 对正余弦函数的界值限制作用. 基于式(10)的 r_1 函数提出一种改进 SCA 算法(Improved SCA algorithm,ISCA),并将其用于 C^{3}SVM 算法中偏移量 η 的优化选择中.

4.2 基于 ISCA 算法的 C^{3}SVM 优化模型

C^{3}SVM 算法待优化参数有影响负类样本裁剪效果与算法优化性能的偏移量 η、控制推广泛化性能的核参数 g、折衷机器结构风险与泛化性能的罚参数 C 等. 不同参数组合将引起模型分类性能差异,且参数优化选择同算法本身的设计同等重要. C^{3}SVM 算法的参数优化选择可看作是一个组合优化问题:

$$\begin{aligned}&\max/\min\ F(\boldsymbol{X})\\&\text{s.t. } x^i\in[x^i_{\min},x^i_{\max}]\\&i=1,2,\cdots,l\end{aligned}\tag{11}$$

其中,$\boldsymbol{X}$ 表示模型中待优化的参数向量,l 表示待优化的参数数目. 本文主要优化裁剪偏移量 η.

利用 ISCA 优化 C^{3}SVM 参数,并将二者相融合构造 ISCA-C^{3}SVM 模型(IC^{3}SVM)以提高 SVM 的非均衡数据识别性能,其算法执行流程图见图 4,主要步骤为:

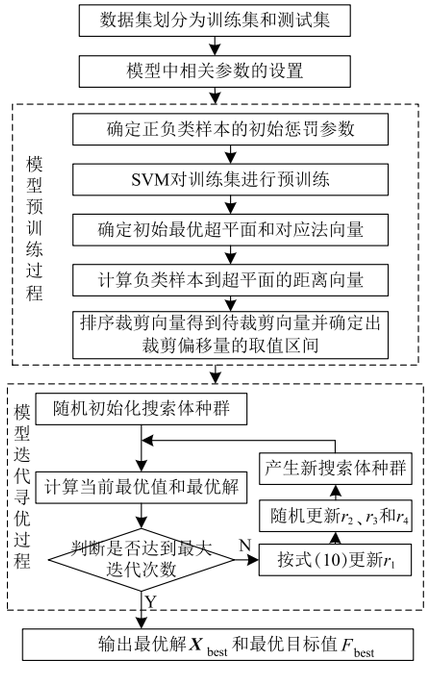


图4 IC^{3}SVM模型流程图

Step 1:数据集的划分和模型参数设置. 将数据集划分为训练集和测试集以分别用于模型的训练和测试;模型中需要设置的参数主要有 C^{3}SVM 预训练的初始参数、ISCA 算法的搜索体数目、迭代次数等.

Step 2:C^{3}SVM 算法的预训练和负类待裁剪样本集的确定. 根据 C^{3}SVM 算法的初始参数,进行模型预训练以确定初始最优超平面 OH、计算负类样本到 OH 的距离并排序以得到待裁剪样本集和 η 有效裁剪区间.

Step 3:ISCA 算法优化 C^{3}SVM 的裁剪偏移量 η. 以 C^{3}SVM 的识别性能作为 ISCA 的优化目标,利用多个搜索体进行并行迭代寻优过程.

Step 4:输出最优参数 $\boldsymbol{X}_{best}$ 和最优目标值 F_{best}.

5 数值实验

为验证本文算法的有效性和可行性,共进行 3 组实验:第 1 组实验以人造数据验证不同裁剪偏移量对 C^{3}SVM 算法的性能影响;第 2 组实验以基准函数验证 ISCA 算法的优越寻优性能;第 3 组实验以标准数据库的非均衡数据验证 IC^{3}SVM 算法的较好识别性能.

5.1 裁剪偏移量 η 对 C^{3}SVM 算法的性能影响分析

(1)人造实验数据

本节 2D 实验数据采用正态分布通过人工方式随机生成,以验证裁剪偏移量 η 对 C^{3}SVM 的性能影响和优化必要性:正类样本数据 $\text{data}^+\sim N(2,1)$,数据量为 100;负类样本数据 $\text{data}^-\sim N(0,1)$,数据量为 400.

(2) 评价指标

均衡数据的分类评价指标——分类准确率不再适用于非均衡数据:因为正类样本在整个数据集中占少数,即使被全部错分也将表现出较高的分类精度. 因此,需选择适于非均衡数据的分类器评价指标,为便于后续描述首先定义混淆矩阵见表 1.

表 1 混淆矩阵

	预测为正类	预测为负类
实际正类	TP(True Positives)	FN(False Negatives)
实际负类	FP(False Positives)	TN(True Negatives)

正类查准率 Precision = TP/(TP + FP);正类召回率 Sensitivity = TP/(TP + FN);负类召回率(也称真负率) Specificity = TN/(FP + TN);进而可得:

$$G_{\text{mean}}=\sqrt{\text{Sensitivity}\times\text{Specificity}}$$

$$F_{\text{measure}}=\frac{2\times\text{Sensitivity}\times\text{Precision}}{\text{Sensitivity}+\text{Precision}}$$

其中,G_{mean} 指标是 Sensitivity 和 Specificity 的几何均值,衡量分类器的整体分类性能;F_{measure} 指标是 Precision 和 Sensitivity 的调和均值,衡量分类器的正类识别性能.

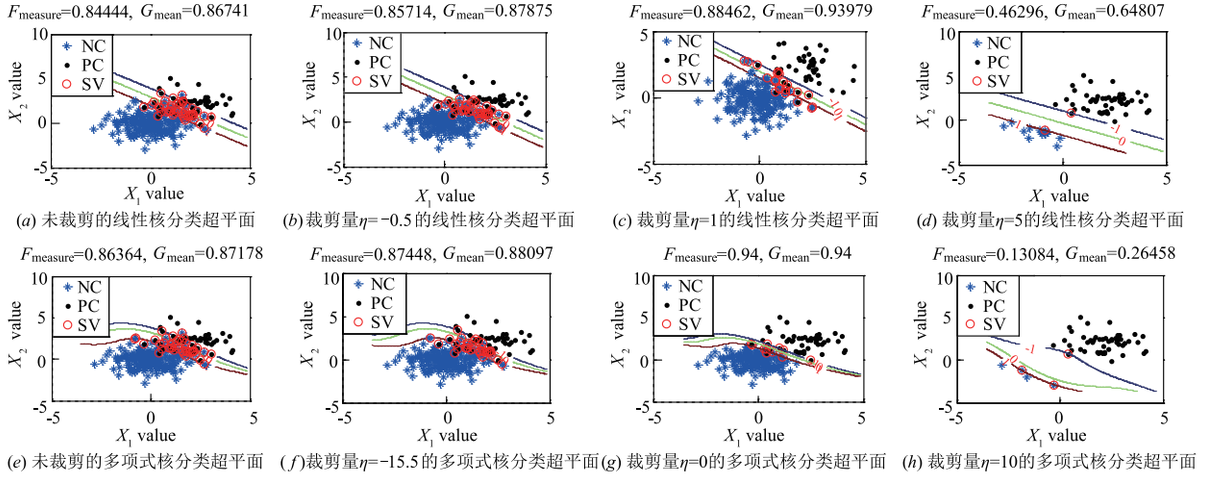
(3) 实验结果和分析

以线性核 K_{L}、多项式核 K_{P} 和 RBF 核 K_{R} 分别进行实验,对人造数据集按类别比例均匀随机选取正负类样本的 1/2 作为训练集,剩余样本为测试集. SVM 预训练得到基于 3 种核函数 C^{3}SVM 算法的裁剪偏移量 η 的有效裁剪区间分别为 $[-1.7,6.7]$、$[-31,17]$ 和 $[-1.8,3]$(负值表示该样本跨过 OH 而混入正类一侧). 为可视化不同裁剪量 η 对基于不同核 C^{3}SVM 的

OH 偏移影响,绘制图 5 (NC、PC 和 SV 分别为 Negative-class、Positive-Class 和 Support Vectors 样本).

由图 5 分析可知,基于不同核函数的 C^{3}SVM 算法的最优超平面 OH 对训练集的裁剪效果各有差异,OH 位置与形状、SV 数目与分布形态、η 有效裁剪区间等均与核函数的选取有关;基于不同裁剪偏移量 η 的 OH 对测试集的识别效果有显著性差异:当裁剪量 η 较小时,裁掉的负类样本数目较少而使训练集非均衡度未发生本质改变且 OH 的分类效果改善不明显,即 F_{measure} 和 G_{mean} 指标均无显著性提高,见图 5(b)、(f)、(j);随裁剪量 η 的不断增大,被裁掉的负类样本不断增多而使训练集非均衡度越趋减弱,分类器对正类识别性能增强,F_{measure} 和 G_{mean} 指标均明显提高,见图 5(c)、(g)、(k);当裁剪量 η 超过一定阈值并继续增大时则会导致负类样本的过度裁剪而发生正负类样本的"属性逆转"现象,此时分类器 OH 将过于偏向于假"正类"的负类样本,表现为 F_{measure} 和 G_{mean} 指标的再次减小,见图 5(d)、(h)、(l).

为进一步分析不同裁剪量 η 对分类器评价指标的递变影响,基于 3 种核函数的 C^{3}SVM 分别按 0.1、0.5 和 0.05 为裁剪间隔 $\Delta\eta$ 进行实验,记录并绘制 F_{measure}、G_{mean} 等指标的实验结果对比见图 6.

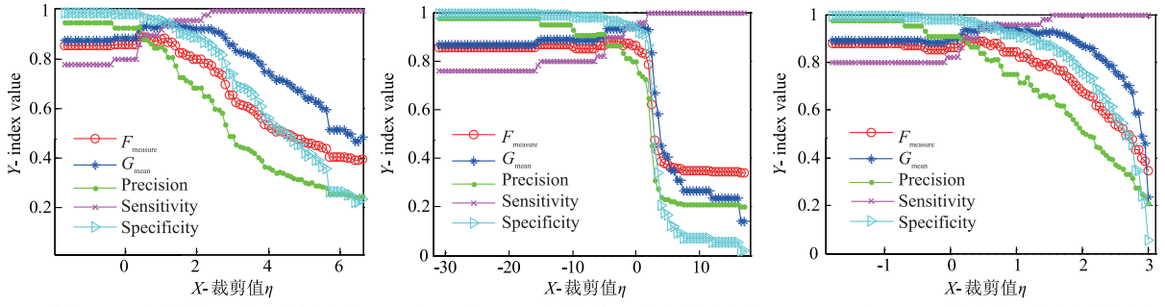


(*a*) 未裁剪的线性核分类超平面　(*b*)裁剪量η=−0.5的线性核分类超平面　(*c*) 裁剪量η=1的线性核分类超平面　(*d*) 裁剪量η=5的线性核分类超平面

(*e*) 未裁剪的多项式核分类超平面　(*f*)裁剪量η=−15.5的多项式核分类超平面　(*g*) 裁剪量η=0的多项式核分类超平面　(*h*) 裁剪量η=10的多项式核分类超平面

(*i*) 未裁剪的线性核分类超平面　(*j*) 裁剪量η=−1的RBF核分类超平面　(*k*) 裁剪量η=0.5的RBF核分类超平面　(*l*) 裁剪量η=2.5的RBF核分类超平面

图5　不同裁剪量η对3种核C^{3}SVM的OH偏移影响(C=1;多项式核中g=1,D=3;RBF核中g=0.05)



(*a*) 线性核C^{3}SVM随着裁剪量η的分类性能变化　(*b*) 多项式核C^{3}SVM随着裁剪量η的分类性能变化　(*b*) RBF核C^{3}SVM随着裁剪量η的分类性能变化

图6　裁剪量η对不同核C^{3}SVM的分类性能递变影响

由图6分析可知,(1)C^3SVM的Sensitivity值随裁剪量η的递增而增大,当超过一定阈值时可实现正类样本的完全识别,此时C^3SVM可看作SVDD算法的一种替代.(2)其它指标随裁剪量η的递增表现出近似相同的波动趋势:在递变初期指标值呈平稳状态,表明裁剪量η较小时C^3SVM算法的非均衡识别性能改善不明显;中期各指标值均有所提高,表明适度裁剪有利于提高C^3SVM的非均衡识别性能;后期各指标值则再次明显减小,表明过度裁剪不利于C^3SVM的非均衡分类.(3)裁剪过程中Precision指标变化最为剧烈,是因为裁剪后期数据集发生样本属性的逆转现象且假"正类"的负类样本错分速率大于正类样本的识别效率.(4)基于3种核函数的C^3SVM算法中多项式核的变化曲线最为陡峭,对裁剪量η的变化最为敏感;RBF核的变化最为平缓,具有较好的优化适应性.

总体来说,裁剪偏移量η对不同核函数C^3SVM算法的非均衡数据识别性能有差异性影响,数据集中负类样本的适度裁剪有利于改善C^3SVM的分类识别性能,因此,裁剪偏移量η的优化选择对提高C^3SVM的非均衡数据识别性能有其必要性.

5.2 ISCA算法的优化性能分析

为验证ISCA算法的较好优化性能,选用6组基准测试函数,以粒子群算法(Particle Swarm Optimization, PSO)[16]、人工蜂群算法(Artificial Bee Colony, ABC)[17]、多元宇宙优化算法(Multi-verse optimizer, MVO)[18]和传统SCA算法为对比算法进行实验.基准测试函数的信息见表2(各函数目标值均为0),其中前3组为单峰函数以检验算法的开采性能与收敛效率、后3组为多峰函数以检验算法的探索性能与局部极值规避性.

表2 基准测试函数信息

Function	$\text{Range}^{\text{Dim}}$
$f_1(\boldsymbol{x}) = \sum_{i=1}^{n} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$	$[-10,10]^{10}$
$f_2(\boldsymbol{x}) = \sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2$	$[-100,100]^{10}$
$f_3(\boldsymbol{x}) = \sum_{i=1}^{n} \lvert x_i \rvert^{i+1}$	$[-1,1]^{10}$
$f_4(\boldsymbol{x}) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos(\frac{x_i}{\sqrt{i}}) + 1$	$[-600,600]^{10}$
$f_5(\boldsymbol{x}) = \sum_{i=1}^{n} \lvert x_i \sin(x_i) + 0.1x_i \rvert$	$[-10,10]^{10}$
$f_6(\boldsymbol{x}) = \sum_{i=1}^{n-1} (0.5 + \frac{\sin^2(\sqrt{100x_i^2 + x_{i+1}^2}) - 0.5}{1 + 0.001(x_i^2 - 2x_i x_{i+1} + x_{i+1}^2)^2})^2$	$[-100,100]^{10}$

各智能算法均设定搜索体数目和最大迭代次数分别为20和2000,各实验组均独立运行30次并以实验统计结果的均值(ave)、标准差(std)、最大值(max)和最小值(min)为评价指标,具体统计结果见表3.

由表3分析可知:6组单/多峰测试函数的优化结果中,ISCA算法几乎均取得最优的ave指标表明ISCA具有较好的平均优化性能、最小std指标说明ISCA具有较好的算法稳定性/鲁棒性、最优max指标证明算法在最坏极端情况下仍具有相对较好的优化性能、最小min指标验证算法极强的迭代寻优性能.改进算法对单峰函数均寻得最优的4项指标,表明ISCA算法具有较好的开采性能和算法收敛效率;对多峰函数,ISCA也获得相对较优的指标值,验证了其对搜索空间的较好探索能力与局部极值的良好规避性.上述结果验证了ISCA算法优越的寻优性能.下节将利用ISCA算法优化C^3SVM的裁剪偏移量η(IC^3SVM)以改善其非均衡数据处理能力.

表3 不同算法的实验结果对比

f		PSO	ABC	MVO	SCA	ISCA
f_1	ave	3.3	3.8E-16	1.2E-2	6.3E-34	**8.4E-37**
	std	1.1E+1	9.5E-17	4.4E-3	2.7E-33	**2.7E-36**
	max	5.0E+1	5.2E-16	2.5E-2	1.5E-32	**1.2E-35**
	min	2.3E-2	1.5E-16	4.9E-3	2.7E-41	**2.8E-45**
f_2	ave	6.1E+4	9.5E-17	1.4E+4	1.3E-46	**1.3E-51**
	std	4.2E+4	3.6E-17	1.1E+4	4.9E-46	**7.0E-51**
	max	1.5E+5	2.0E-16	3.5E+4	2.5E-45	**3.8E-50**
	min	5.2E+3	4.7E-17	1.6E+3	6.5E-63	**8.0E-65**
f_3	ave	2.6E-7	1.7E-17	7.0E-9	1.1E-86	**4E-108**
	std	4.3E-7	6.0E-18	8.3E-9	6.1E-86	**2E-107**
	max	1.4E-6	3.1E-17	4.4E-8	3.3E-85	**8E-107**
	min	**0.0**	5.8E-18	3.9E-10	9E-127	1E-137
f_4	ave	4.0	7.0E-3	2.3E-1	2.3E-2	**3.1E-3**
	std	3.4	7.3E-3	6.7E-2	7.5E-2	**6.7E-3**
	max	1.6E+1	**2.7E-2**	3.5E-1	3.9E-1	9.6E-2
	min	8.4E-2	1.1E-16	1.1E-1	**0.0**	**0.0**
f_5	ave	2.6E-1	1.0E-9	9.5E-2	2.3E-16	**7.2E-23**
	std	2.4E-1	5.3E-9	8.3E-2	1.3E-15	**3.4E-22**
	max	1.2	2.9E-8	3.4E-1	7.0E-15	**1.9E-21**
	min	4.2E-2	4.45E-16	8.3E-3	3.1E-39	**1.3E-49**
f_6	ave	1.0	**2.1E-1**	9.6E-1	1.1	2.4E-1
	std	3.1E-1	**8.9E-2**	1.8E-1	3.9E-1	2.9E-1
	max	1.5	**3.6E-1**	1.3	1.6	1.4
	min	2.0E-1	1.5E-2	5.3E-1	2.0E-1	**7.0E-3**

5.3 IC^3SVM对非均衡数据的优化性能分析

为验证IC^3SVM算法对非均衡数据的较好识别性

能,以 UCI 数据库和 KEEL 数据库[19]的 5 组非均衡数据集为实验数据,其物理属性信息见表 4;以 S^2VM[20]、Over sampling SVM (OSVM)[7]、Under sampling SVM (USVM)[8]和 Smote SVM (SSVM)[6]为对比算法进行实验.

鉴于 RBF 核在无先验知识前提下对线性、非线性问题的较好适用性以及 5.1 节中表现出的较好非线性处理性能,本节以 RBF 核为 SVM 核函数,各算法中惩罚参数 C 和核参数 g 均设定为 10 和 0.5,SSVM 算法中 k-NN 近邻数目 $k=5$,IC^3SVM 算法中种群规模和最大迭代次数分别为 5 和 100 且以 G_{mean} 指标最大化为优化目标. 随机选取各数据集的 1/2 为训练集剩余样本为测试集. 评价指标选用 G_{mean}(G)、Sensitivity(S)指标和 $F_{measure}$(F)指标,具体实验结果见表 5.

表 4　非均衡数据集的物理属性信息

数据集	样例规模	正例数	非均衡度	属性数
3TE	958	332	1.886	9
Cli	540	46	10.739	18
L7D	443	37	10.973	7
E4	336	20	15.80	7
Y5	1484	44	32.727	8

注: ClimateModelSimulationCrashes (Cli)、TicTacToeEndgame (3TE)、Ecoli4(E4)、Led7digit_02456789VS1(L7D)、Yeast5(Y5).

表 5　5 种算法的实验结果对比

Data		S^2VM	OSVM	USVM	SSVM	IC^3SVM
3TE	G	0.862	0.866	0.847	0.846	**0.874**
	S	0.753	0.759	**0.807**	0.747	0.783
	F	0.848	0.851	0.800	0.819	**0.855**
Cli	G	0.686	0.686	0.804	0.716	**0.864**
	S	0.478	0.478	0.783	0.522	**0.870**
	F	0.579	0.579	0.429	**0.615**	0.513
L7D	G	0.850	0.844	0.860	0.850	**0.891**
	S	0.737	0.737	0.790	0.737	**0.947**
	F	**0.757**	0.700	0.638	**0.757**	0.514
E4	G	0.894	0.943	0.894	0.889	**0.984**
	S	0.800	0.900	0.800	0.800	**1.0**
	F	0.857	0.857	**0.889**	0.800	0.800
Y5	G	0.213	0.965	0.967	0.968	**0.975**
	S	0.046	**1.0**	**1.0**	**1.0**	**1.0**
	F	0.087	0.473	**0.571**	0.489	0.557

由表 5 分析可知,IC^3SVM 算法与其他算法相比均取得最优 G_{mean},表明改进算法具有较好的非均衡数据整体识别性能和良好的负类样本过拟合规避性. IC^3SVM 算法几乎均获得最优 Sensitivity,说明其具有较好的正类召回率,在一定程度上可有效提高非均衡数据的正类识别性能,有利于降低实际应用中因正类样本错分而造成的潜在损失与危害. 表 5 中改进算法的 $F_{measure}$ 指标改善效果有限,是因为 ISCA 算法以 G_{mean} 最大化为优化目标. 为提高 IC^3SVM 识别非均衡数据的 $F_{measure}$ 性能,可将 $F_{measure}$ 最大化为优化目标以保证 IC^3SVM 算法对正类样本的完全识别(见 5.1 节分析). 而在实际应用中,ISCA 算法的优化目标可根据具体分类需求以 G_{mean}、$F_{measure}$ 等指标或综合多指标的特定加权指标进行针对性设置. 表 6 为以 G_{mean} 与 $F_{measure}$ 1/2 加权为优化目标时的实验结果.

表 6　1/2 加权条件下 $F_{measure}$ 对比结果

Data	3TE	Cli	L7D	E4	Y5
IC^3SVM	**0.855**	0.513	0.514	0.800	0.557
$IC^3SVM^{1/2}$	**0.855** *	**0.604**	**0.789** *	**0.857**	**0.750** *

注: $IC^3SVM^{1/2}$ 中带 * 角标的表示其也优于表 5 中其他对比算法

由表 6 分析可知,在优化目标中融入 $F_{measure}$ 指标的 $IC^3SVM^{1/2}$ 算法能够较好地调和 IC^3SVM 对非均衡数据的正类查准率与召回率,有效提高分类器的正类识别率,表明在实际非均衡应用中可根据具体需求自适应设定优化目标以改善分类器的特定指标性能,进一步验证改进算法在实际应用中具有较好的特定需求适定性.

综上所述,IC^3SVM 算法具有较好的非均衡数据整体分类效果与正类识别率及在实际应用中较好的特定需求适定性,验证了改进算法对非均衡数据的较好识别性能.

6　结论

针对 S^2VM 处理非均衡数据出现的正类样本大量错分现象,分析知其直观原因是最优超平面过于趋近或嵌入正类样本,理论原因则是最优超平面的偏置量 b 优化不当所引起的,继而得出 S^2VM 对非均衡数据分类失效的 3 种原因及相应对策. 根据 SVM 最优分类超平面仅由少量支持向量决定的特性,提出一种基于负类支持向量样本裁剪策略的改进 SVM 数学表达. 考虑到负类样本数据量大需经多次反复训练-裁剪过程才能实现正类数据的较高识别率且较为费时,鉴于数据样本的反复“训练-删除”过程等效于基于特定裁截面的一次性裁剪,提出一种基于特定裁截面的负类样本裁剪和非对称错分惩罚机制相耦合的 C^3SVM 算法,并利用 ISCA 算法优化其裁截偏移量 η 以提高其非均衡数据处理性能,实验结果验证 IC^3SVM 算法对非均衡数据的较好识别性能.

本文的改进思想主要是从 SVM 对非均衡数据分类失效的第 1 种原因即平移偏置量的角度并融合错分惩罚机制提出的,接下来的工作重点将是从第 2 或 3 种原因展开相关研究:正负类 Lagrange 乘子的代价敏感程度及其与错分惩罚机制间的关系,以及融入数据特性的针对特定非均衡数据的核函数构造与设计等.

参考文献

[1] Vapnik V. The Nature of Statistical Learning Theory [M]. New York:Wiley,1998.

[2] 王友卫,刘元宁,凤丽洲,等. 基于用户兴趣集的在线垃圾邮件快速识别新方法[J]. 电子学报,2015,43(10):1963 - 1970.
WANG Y W,LEU Y N,et al. A novel quick online spam identification method based on user interest set[J]. Acta Electronica Sinica,2015,43(10):1963 - 1970. (in chinese)

[3] Shafiabady N,Lee L H,Rajkumar R,et al. Using unsupervised clustering approach to train the support vector machine for text classification[J]. Neurocomputing,2016,211:4 - 10.

[4] Shao Y H,Chen W J,Wang Z,et al. Weighted linear loss twin support vector machine for large-scale classification [J]. Knowledge-Based Systems,2015,73:276 - 288.

[5] Japkowic Z N,Stephen S . The class imbalance problem:A systematic study[J]. Intelligent Data Analysis,2002,6(5):429 - 449.

[6] 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.

[7] Drummond C,Holte R C. C4. 5,class imbalance,and cost sensitivity:why under-sampling beats over-sampling [A]. Workshop on Learning from Imbalanced Datasets II[C]. Washington DC:Citeseer,2003:1 - 8.

[8] Kubat M,Matwin S. Addressing the curse of imbalanced training sets:one-sided selection[A]. Proc 14th ICML [C]. Nashville:Morgan Kaufmann Publishers,1997. 179 - 186.

[9] Sain H,Purnami S W. Combine sampling support vector machine for imbalanced data classification[J]. Procedia Computer Science,2015,72(1):59 - 66.

[10] Zhang Y,Fu P,Liu W,et al. Imbalanced data classification based on scaling kernel-based support vector machine [J]. Neural Computing and Applications,2014,25(3 - 4):927 - 935.

[11] Maldonado S,López J. Imbalanced data classification using second-order cone programming support vector machines[J]. Pattern Recognition,2014,47(5):2070 - 2079.

[12] Datta S,Das S. Near-bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs[J]. Neural Networks,2015,70:39 - 52.

[13] Jian C,Gao J,Ao Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble[J]. Neurocomputing,2016,193:115 - 122.

[14] Wang X,Huang F,Cheng Y. Super-parameter selection for gaussian-kernel SVM based on outlier-resisting[J]. Measurement,2014,58:147 - 153.

[15] Mirjalili S. SCA:a sine cosine algorithm for solving optimization problems[J]. Knowledge-Based Systems,2016,96:120 - 133.

[16] Eberhart R C,Kennedy J. A new optimizer using particle swarm theory[A]. Proc. Sixth International Symposium on MICRO Machine and Human Science[C]. Nagoya,Japan:IEEE Press,2002. 39 - 43.

[17] Karaboga D. An idea based on honey bee swarm for numerical optimization[R]. Kayseri:Erciyes University,Engineering Faculty,Computer Engineering Department,2005.

[18] Mirjalili S,Mirjalili S M,Hatamlou A. Multi-verse optimizer:a nature-inspired algorithm for global optimization [J]. Neural Computing and Applications,2016,27(2):495 - 513.

[19] Fernández A,García S,del Jesus M J,et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets[J]. Fuzzy Sets and Systems,2008,159(18):2378 - 2398.

[20] Chang C C,Lin C J. LIBSVM:a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST),2011,2(3):27.

作者简介

高雷阜 男,1963 年出生,辽宁阜新人,辽宁工程技术大学教授,博士生导师,主要研究领域为最优化理论与应用、非线性动力系统等.
E-mail:gaoleifu@163.com

赵世杰 男,1987 年出生,山东五莲人,辽宁工程技术大学博士研究生,主要研究领域为人工智能与数据挖掘、优化与管理决策等.
E-mail:zhao2008shijie@126.com

于冬梅 女,1986 出生,辽宁鞍山人,博士,辽宁工程技术大学讲师,主要研究领域为最优化理论与应用.

徒 君 男,1982 出生,安徽全椒人,博士,辽宁工程技术大学讲师,主要研究领域为智能算法、供应链管理等.