基于密度敏感最大软间隔 SVDD 不均衡数据分类算法

陶新民,李晨曦,沈 微,常 瑞,王若彤,刘艳超
(东北林业大学工程技术学院,黑龙江哈尔滨 150040)

摘 要: 为了提高传统支持向量域描述(C-SVDD)算法处理不均衡数据集的分类能力,提出一种基于密度敏感最大软间隔支持向量域描述(DSMSM-SVDD)算法. 该算法通过对多数类样本引入相对密度来体现训练样本原始空间分布对求解最优分类界面的影响,通过在目标函数中增加最大软间隔正则项,使 C-SVDD 的分类边界向少数类偏移,进而提高算法分类性能. 算法首先对每个多数类样本计算相对密度来反映样本的重要性,然后将训练样本输入到 DSMSM-SVDD 中实现数据分类. 实验部分,讨论了算法参数间的关系及其对算法分类性能的影响,给出算法参数取值建议. 最后通过与 C-SVDD 的对比实验,表明本文建议的算法在不均衡数据情况下的分类性能优于 C-SVDD 算法.

关键词: 支持向量域数据描述; 不均衡数据; 相对密度
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2725-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.20

The SVDD Classifier for Unbalanced Data Based on Density-Sensitive and Maximum Soft Margin

TAO Xin-min, LI Chen-xi, SHEN Wei, CHANG Rui, WHANG Ruo-tong, LIU Yan-chao
(*College of Engineering and Technology, University of Northeast Forestry, Harbin, Heilongjiang* 150040, *China*)

Abstract: In order to improve the conventional support vector domain description(C-SVDD) algorithm's classification performance under unbalanced datasets, a novel maximum soft margin support vector domain descriptionalgorithm based on density sensitivity(DSMSM-SVDD) is presented. The relative density informationof the majority samples is introduced to reflect the impact of original training sample's space distribution on the optimal interface, by adding the maximum soft margin regularization term in the objective function, the classification boundary of the C-SVDD algorithm is shifted tominority classes, and consequently the classification performance of the proposed algorithm is significantly improved. Firstly, the relative density of each majority sample is calculated to reflect the importance of the training samples, and then the obtained training samples with relative density are input into the proposed DSMSM-SVDD algorithm to implement the classification-task. In the experiments, the relationship of the parameters and the influence of the parameters on classification performance are investigated. Finally, the comparison results with C-SVDD algorithm demonstrate that the proposed algorithm is superior to the C-SVDD algorithm in the case of unbalanced data.

Key words: support vector domain description; unbalanced datasets; relative density

1 引言

分类问题作为一项重要的任务,在机器学习、模式识别与数据挖掘领域有着广泛的应用[1-3]. 在分类问题的模型构造方面,通常分为二分类算法和单分类算法. 以往基于二分类的算法都是在样本均衡的前提条件下进行的. 其中,基于统计学习理论结构风险最小化的支持向量机(Support Vector Machine,SVM)算法[4-5]作为一个经典的二分类算法,因其良好的非线性区分能力近年来引起了广泛关注. 然而在实际应用过程中所能得到的训练数据在数量上或质量上并不都是均衡的,因此受不均衡数据的影响,传统 SVM 分类性能严重下降[6]. 那么如何提高分类算法在不均衡数据下的分类性能成为了众多学者关注的重点. 近些年,学者们相继提出了各种处理不均衡

收稿日期:2017-09-08;修回日期:2017-12-12;责任编辑:蓝红杰
基金项目:中央高校基本科研业务费专项资金(No. 2572017EB02, No. 2572017CB07);东北林业大学双一流科研启动基金(No. 411112438);哈尔滨市科技局创新人才基金(No. 2017RAXXJ018);国家自然科学基金(No. 31570547)

数据分类的改进算法. 其中针对二分类算法的研究方向主要分为两大类:第一类是为了改进现有分类器的模型. 其中代价敏感学习方法就是通过把各类的不同错分代价应用到分类决策中去,来尽可能地降低误分类的整体代价,然而由于代价敏感值的大小在现实应用中难以确定,使得该方法在实际应用中受到限制[7]. 第二类方法是数据均衡化处理,即将不均衡数据分类问题转化为均衡的两类分类问题进行研究,最具代表性的就是利用抽样算法[8]实现数据均衡,其中又分为对多数类样本的欠抽样[9]和对少数类样本的过抽样[10]. 欠抽样算法由于只抽取了多数类的子集来训练分类器,忽略掉了多数类样本的重要结构分布信息,从而引起边界偏移. 传统过抽样算法因为复制的少数类样本没有增加任何新知识,从而导致过学习问题. 虽然后来提出的基于 SMOTE[11]过取样能在一定程度上降低非平衡度且优于传统的过抽样方法,但是由于训练样本的增加会使决策域减小,导致算法过度拟合,同时人为增加样本有可能造成噪声点增加,从而降低分类精度. 与传统两分类算法相比,单分类算法在处理不均衡数据的分类问题时,因为能有效地分辨出多数类样本与少数类样本,且具有更高的识别率与精度. 因此,近年来受到了学者们广泛关注. 其中传统支持向量域描述(C-SVDD)算法因其原理简单、训练速度快,现已被广泛应用于不均衡数据分类领域[12-13]. 然而由于 C-SVDD 在训练过程中没有考虑到训练样本在原始空间的分布差异对求解最优分类界面的影响,使得求解后的分类界面受野性噪声点的影响出现偏移,进而降低了分类器的泛化性能. 另外,C-SVDD 利用的是惩罚因子对两类样本集的错分总和进行调整,当面对线性可分问题时,两类样本集的错分样本总和为零,因此没有起到调整边界的作用.

鉴于此,本文为了充分考虑原始空间中样本的分布信息,通过引入相对密度[14]来反映训练样本的重要程度,使处于相对密度高区域的训练样本较相对密度低的训练样本更易落入超球体之内,进而消除噪声影响. 同时借鉴传统 SVM 分类间隔最大化思想,对 C-SVDD 进行改进[15],通过引入最大软间隔正则项[16]来充分利用少数类样本信息,将分类界面向少数类偏移,进而提高算法处理不均衡数据的分类性能. 本文通过将上述两种思想结合后提出一种基于密度敏感最大软间隔 SVDD 不均衡数据分类算法. 实验部分将本文提出的方法同 C-SVDD 进行比较,结果表明本文的方法在数据不均衡情况下的分类性能较 C-SVDD 有较大幅度提高.

2 理论分析

2.1 C-SVDD 算法

C-SVDD 作为直接使用目标训练数据进行分类判别的经典单分类算法,其中心思想是借助非线性映射 $\varphi(\cdot)$将训练样本映射到高维特征空间中,然后在特征空间中通过求解球心 $\boldsymbol{a}$ 和半径 R 来寻找包含大多数映射训练样本的最小超球体. C-SVDD 通过引入惩罚因子 $C=\left(\frac{c_1}{N},\frac{c_2}{\overline{N}}\right)$来控制对不同类别错分样本的惩罚力度,由于该算法不依赖少数类样本的个数,因此非常适合处理不均衡数据分类问题. 其简要数学模型如下:

$$\min_{\boldsymbol{a},R,\xi_i,\xi_j}\left\{R^2+\frac{c_1}{N}\sum_{i,y_i=1}\xi_i+\frac{c_2}{\overline{N}}\sum_{j,y_j=-1}\xi_j\right\} \tag{1}$$

约束条件为:

$$\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\|^2\leqslant R^2+\xi_i,\xi_i\geqslant 0,\forall i=1,2,\cdots,N$$
$$\|\varphi(\boldsymbol{x}_j)-\boldsymbol{a}\|^2\geqslant R^2-\xi_j,\xi_j\geqslant 0,\forall j=1,2,\cdots,\overline{N} \tag{2}$$

其中,R 为超球体半径,$\boldsymbol{a}$ 为圆心,$N,\overline{N}$ 表示多数类与少数类训练样本数目,$\frac{c_1}{N},\frac{c_2}{\overline{N}}$为用于控制多数和少数训练样本损失函数大小的惩罚因子,ξ_i,ξ_j 分别为不同类别样本实际错误的松弛变量,$\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)$表示训练样本 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 在高维特征空间的映射.

这里由拉格朗日最优化理论,可得上述问题的对偶问题:

$$\max_{\boldsymbol{\alpha}}\left\{\sum_{i=1}^{N+\overline{N}}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i,j=1}^{N+\overline{N}}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right\} \tag{3}$$

约束条件为:

$$\sum_{i=1}^{N+\overline{N}}\alpha_i y_i=1 \tag{4}$$

$$0\leqslant\alpha_i\leqslant\frac{c_1}{N},\forall i,y_i=1,i=1,2,\cdots,N \tag{5}$$

$$0\leqslant\alpha_j\leqslant\frac{c_2}{\overline{N}},\forall j,y_j=-1,j=1,2,\cdots,\overline{N} \tag{6}$$

其中,$\begin{cases}y_i=1,\boldsymbol{x}_i\in\text{多数类}\\ y_j=-1,\boldsymbol{x}_j\in\text{少数类}\end{cases}$,$k(\boldsymbol{x}_i,\boldsymbol{x}_i)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_i)$,$\alpha_i,\alpha_j$ 为拉格朗日乘子,在描述中 $0<\alpha_i<\frac{c_1}{N},0<\alpha_j<\frac{c_2}{\overline{N}}$ 所对应的 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 称为支持向量,即为位于球体边缘决定分类界面的训练样本.

2.2 DSMSM-SVDD 算法

2.2.1 密度敏感因子

C-SVDD 作为经典的单分类算法因其原理简单易实现被广泛应用于不均衡数据的分类问题. 但是由于 C-SVDD 没有充分考虑训练样本的自身分布情况,导致在反映目标数据集整体密度分布特性上具有一定的局限性. 在 C-SVDD 模型的训练过程中,尽管惩罚因子 C 对错分样本点具有一定的约束作用,但是由于 C 是一个不变的常参量,即表示对所有的训练样本同等处理,导致 C-SVDD 模型在训练过程中对异常点或者噪声点

非常敏感,甚至造成算法过分拟合. 另外,C-SVDD 在训练过程中与经典的 SVM 算法一样,最优分类界面仅仅依靠小部分被称为支持向量的数据来确定,忽略了球体内非支持向量对数据域描述的影响. 但由于具有较高密度的非支持向量周围区域比较低密度区域更重要,所以对其错分的惩罚力度应该增加,以便于更准确地识别出目标数据集的数据域描述,提高算法的分类性能. 因此,仅依靠支持向量来确定最优分类界面而不考虑密度分布可能会错过最优解.

针对 C-SVDD 的上述问题,本文通过对每个目标数据点引入相对密度,并与错分惩罚项相结合,改造 C-SVDD 优化目标的方式来体现相对密度对算法最优解的影响. 即达到使分布在相对密度较大区域的样本点尽可能地包含在描述区域边界内,而相对密度较小的噪声点因惩罚因子的作用不会对最优分类界面的求解产生影响的目的. 其中,本文采用 Parzen-window 算法来计算样本的相对密度,具体方法描述如下:

设 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$ 为给出的目标训练样本集,其中 N 代表目标训练样本的个数. 对于任意一个训练样本 $\boldsymbol{x}_i$ 其相对密度 ρ_i 的定义如下:

$$\rho_i=\exp\left\{\omega\times\frac{\mathrm{Par}(\boldsymbol{x}_i)}{\zeta}\right\},\ \forall i=1,2,\cdots,N \tag{7}$$

其中 $\mathrm{Par}(\boldsymbol{x}_i)=\dfrac{1}{N}\sum\limits_{j=1}^{N}\left(\dfrac{1}{\sqrt{(2\pi)^D}s}\right)\exp\left(-\dfrac{1}{2s}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right)$,$\zeta=\dfrac{1}{N}\sum\limits_{i=1}^{N}\mathrm{Par}(\boldsymbol{x}_i)$,$D$ 为输入数据的维度,ω 为权重,s 是 Parzen-window 的平滑参数. 若训练样本点 $\boldsymbol{x}_i$ 获取的相对密度值 ρ_i 越大,则表明 $\boldsymbol{x}_i$ 所处的区域越紧致.

2.2.2　最大软间隔项

C-SVDD 虽然通过对多数类与少数类样本增加不同惩罚因子的方式使其分类界面向少数类样本方向偏移,进而提高算法的分类性能. 但是 C-SVDD 在解决线性可分问题时,由于多数类样本与少数类样本错分总和为零,惩罚因子对算法达不到调整分类界面的效果.

因此,针对上述问题,本文在充分借助少数类样本信息的前提下,借鉴传统 SVM 最大软间隔思想,对 C-SVDD 算法的优化目标进行改进,通过增加少数类样本与多数类样本的最大软间隔正则项信息,使得算法的最优分类界面向少数类偏移,提升算法分类性能,进而提高 C-SVDD 处理不均衡数据集的能力.

2.2.3　DSMSM-SVDD 算法

针对 C-SVDD 存在的上述两方面问题,本文提出一种 DSMSM-SVDD 算法. 令 $\{(\boldsymbol{x}_i,\rho_i),i=1,2,\cdots,N+\bar{N}\}$ 为目标训练样本数据集,其中 $\boldsymbol{x}_i$ 为目标训练样本,ρ_i 为对应目标样本 $\boldsymbol{x}_i$ 的相对密度. 针对不均衡数据分类问题,当 $\boldsymbol{x}_i\in$ 多数类时,ρ_i 求法如上节所述. 考虑到少数类样本稀少,按上述方法计算后的相对密度无法体现少数类样本真实的密度分布信息,因此当 $\boldsymbol{x}_j\in$ 少数类时,令 $\rho_j=1$. 为了便于描述,这里采用 $\varphi(\boldsymbol{x}_i)$ 表示训练样本 $\boldsymbol{x}_i$ 在高维特征空间的映射. DSMSM-SVDD 算法具体的数学模型表示如下:

$$\min_{\boldsymbol{a},R,d,\xi_i,\xi_j}\left\{R^2-Md^2+\frac{c_1}{N}\sum_{i,y_i=1}^{N}\rho_i\xi_i+\frac{c_2}{\bar{N}}\sum_{j,y_j=-1}^{\bar{N}}\rho_j\xi_j\right\} \tag{8}$$

约束条件为:

$$y_i(R^2-\langle\varphi(\boldsymbol{x}_i)-\boldsymbol{a},\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\rangle)\geqslant d^2-\xi_i \tag{9}$$

$$\xi_i\geqslant 0,\ \forall i=1,2,\cdots,N+\bar{N} \tag{10}$$

其中,d 表示超球面 S 距离最近多数、少数类训练样本的距离,$M\geqslant 1$ 为调节超球面半径和间隔项的参数,$\dfrac{c_1}{N}$、$\dfrac{c_2}{\bar{N}}$为用于控制多数和少数训练样本损失函数大小的惩罚因子,ξ_i、ξ_j 为实际错误.

由上述的优化问题,得广义的拉格朗日函数:

$$\begin{aligned}L(\boldsymbol{a},R,d,\xi_i,\xi_j,\boldsymbol{\alpha},\boldsymbol{\beta})=\ &R^2-Md^2+\frac{c_1}{N}\sum_{i,y_i=1}^{N}\rho_i\xi_i+\frac{c_2}{\bar{N}}\sum_{j,y_j=-1}^{\bar{N}}\rho_j\xi_j+\\&\sum_{i=1}^{N+\bar{N}}\alpha_i[d^2-\xi_i-y_i(R^2-\langle\varphi(\boldsymbol{x}_i)\\&-\boldsymbol{a},\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\rangle)]-\sum_{i=1}^{N+\bar{N}}\beta_i\xi_i\end{aligned} \tag{11}$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_{N+\bar{N}})$,$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_{N+\bar{N}})$.

利用 KKT 条件有:

$$\frac{\partial L}{\partial R}=2R-2\sum_{i=1}^{N+\bar{N}}\alpha_i y_i R=0 \tag{12}$$

$$\frac{\partial L}{\partial d}=-2Md+2\sum_{i=1}^{N+\bar{N}}\alpha_i d=0 \tag{13}$$

$$\frac{\partial L}{\partial \boldsymbol{a}}=-2\sum_{i=1}^{N+\bar{N}}\alpha_i y_i(\varphi(\boldsymbol{x}_i)-\boldsymbol{a})=0 \tag{14}$$

$$\frac{\partial L}{\partial \xi_i}=\frac{c_1}{N}\rho_i-\alpha_i-\beta_i=0,\ \forall i,y_i=1 \tag{15}$$

$$\frac{\partial L}{\partial \xi_j}=\frac{c_2}{\bar{N}}\rho_j-\alpha_j-\beta_j=0,\ \forall j,y_j=-1 \tag{16}$$

还有松弛互补条件:

$$\alpha_i[d^2-\xi_i-y_i(R^2-\langle\varphi(\boldsymbol{x}_i)-\boldsymbol{a},\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\rangle)]=0 \tag{17}$$

$$\alpha_i\geqslant 0 \tag{18}$$

$$-\beta_i\xi_i=0 \tag{19}$$

$$\beta_i\geqslant 0 \tag{20}$$

由于 α_i,β_i 均是非负实数,由式(12)得:

$$\sum_{i=1}^{N+\bar{N}}\alpha_i y_i=1 \tag{21}$$

由式(13)得:

$$\sum_{i=1}^{N+\bar{N}}\alpha_i = M \tag{22}$$

由式(15),(18)得:

$$0\leqslant\alpha_i\leqslant\frac{c_1}{N}\rho_i,\ \forall i,y_i=1 \tag{23}$$

由式(16),(18)得:

$$0\leqslant\alpha_j\leqslant\frac{c_2}{\bar{N}}\rho_j,\ \forall j,y_j=-1 \tag{24}$$

由式(14),(21)得:

$$\boldsymbol{a} = \sum_{i=1}^{N+\bar{N}}\alpha_i y_i\varphi(x_i) \tag{25}$$

将式(15),(16),(21),(22),(25)代入 L 中,得:

$$\begin{aligned}L(\boldsymbol{a},R,d,\xi_i,\xi_j,\boldsymbol{\alpha},\boldsymbol{\beta}) =& \sum_{i=1}^{N+\bar{N}}\alpha_i y_i\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_i) - \\ & \sum_{i,j=1}^{N+\bar{N}}\alpha_i\alpha_j y_i y_j\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)\end{aligned} \tag{26}$$

做下变换 $\text{kernel}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)$,本算法引用经典高斯核函数,则 $\text{kernel}(\boldsymbol{x}_i,\boldsymbol{x}_j)=k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$.

得到原问题的对偶问题:

$$\max_{\boldsymbol{\alpha}}\left\{\sum_{i=1}^{N+\bar{N}}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x}_i) - \sum_{i,j=1}^{N+\bar{N}}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right\} \tag{27}$$

约束条件:

$$\sum_{i=1}^{N+\bar{N}}\alpha_i y_i = 1 \tag{28}$$

$$\sum_{i=1}^{N+\bar{N}}\alpha_i = M \tag{29}$$

$$\alpha_i\geqslant 0,\ \forall i=1,2,\cdots,N+\bar{N} \tag{30}$$

$$0\leqslant\alpha_i\leqslant\frac{c_1}{N}\rho_i,\ \forall i,y_i=1,i=1,2,\cdots,N \tag{31}$$

$$0\leqslant\alpha_j\leqslant\frac{c_2}{\bar{N}}\rho_j,\ \forall j,y_j=-1,j=1,2,\cdots,\bar{N} \tag{32}$$

通过求解上述问题,得到 $\boldsymbol{\alpha}$. 进而求得 $\boldsymbol{a}$ 与 R. 令 $\gamma_i^*=\alpha_i y_i,\ \forall i=1,2\cdots,N+\bar{N}$ 则:

$$\boldsymbol{a} = \sum_{i=1}^{N+\bar{N}}\gamma_i^*\varphi(\boldsymbol{x}_i) \tag{33}$$

$$\begin{aligned}R^2 =& \frac{1}{2}\left[\frac{1}{N_n}\sum_{l=1}^{N_n}\left(1-2\sum_{i=1}^{N+\bar{N}}\gamma_i^* k(\boldsymbol{x}_l,\boldsymbol{x}_i) + \sum_{i,j=1}^{N+\bar{N}}\gamma_i^*\gamma_j^* k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)\right. \\ & \left. + \frac{1}{N_a}\sum_{m=1}^{N_a}\left(1-2\sum_{i=1}^{N+\bar{N}}\gamma_i^* k(\boldsymbol{x}_m,\boldsymbol{x}_i) + \sum_{i,j=1}^{N+\bar{N}}\gamma_i^*\gamma_j^* k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)\right]\end{aligned} \tag{34}$$

其中,N_n 为属于多数类且是支持向量的样本点个数,N_a 为属于少数类且是支持向量的样本点个数. $\boldsymbol{x}_l$ 为多数类支持向量,$\boldsymbol{x}_m$ 为少数类支持向量,$k(\boldsymbol{x}_i,\boldsymbol{x}_j)$为高斯核函数.

对于待测样本 $\boldsymbol{x}_{new}$,构造决策函数对其进行分类:

$$\begin{aligned}d_{new}^2 =& \left\|\varphi(\boldsymbol{x}_{new}) - \sum_{i=1}^{N+\bar{N}}\gamma_i^*\varphi(\boldsymbol{x}_i)\right\|^2 \\ =& 1 + \sum_{i,j=1}^{N+\bar{N}}\gamma_i^*\gamma_j^* k(\boldsymbol{x}_i,\boldsymbol{x}_j) - 2\sum_{i=1}^{N+\bar{N}}\gamma_i^* k(\boldsymbol{x}_{new},\boldsymbol{x}_i)\end{aligned} \tag{35}$$

$d_{new}^2\leqslant R^2$ 表明待测样本 $\boldsymbol{x}_{new}$ 为多数类样本,$d_{new}^2>R^2$ 表明待测样本 $\boldsymbol{x}_{new}$ 为少数类样本.

根据 KKT 最优化条件得:

位于多数类间隔内的多数类样本满足:

$$\alpha_i=0\Rightarrow\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\|^2<R^2-d^2,\xi_i=0$$

位于少数类间隔内的少数类样本满足:

$$\alpha_j=0\Rightarrow\|\varphi(\boldsymbol{x}_j)-\boldsymbol{a}\|^2>R^2+d^2,\xi_j=0$$

位于多数类间隔上的多数类样本满足:

$$0<\alpha_i<\frac{c_1}{N}\rho_i\Rightarrow\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\|^2=R^2-d^2,\xi_i=0$$

位于少数类间隔上的少数类样本满足:

$$0<\alpha_j<\frac{c_2}{\bar{N}}\rho_j\Rightarrow\|\varphi(\boldsymbol{x}_j)-\boldsymbol{a}\|^2=R^2+d^2,\xi_j=0$$

位于多数类间隔外的多数类样本满足:

$$\alpha_i=\frac{c_1}{N}\rho_i\Rightarrow\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}\|^2>R^2-d^2,\xi_i>0$$

位于少数类间隔外的少数类样本满足:

$$\alpha_j=\frac{c_2}{\bar{N}}\rho_j\Rightarrow\|\varphi(\boldsymbol{x}_j)-\boldsymbol{a}\|^2<R^2+d^2,\xi_j>0 \tag{36}$$

2.2.4 M 值的取值范围

由式(36)知,带有间隔错误的多数类样本对应的 $\xi_i>0$,根据松弛互补条件 $-\beta_i\xi_i=0$,则相对应的拉格朗日乘子 $\beta_i=0$,同时由式(15)可推得 $\alpha_i=\frac{c_1}{N}\rho_i$. 同理,对于带有间隔错误的少数类样本 $\alpha_j=\frac{c_2}{\bar{N}}\rho_j$. 结合式(29),(30)可以得到:

$$M>\frac{c_1}{N}\sum_{i=1}^{P_{m_+}}\rho_i+\frac{c_2}{\bar{N}}\sum_{j=1}^{P_{m_-}}\rho_j \tag{37}$$

我们分别通过式(31)、(32)可以得到 $\alpha_i\leqslant\frac{c_1}{N}\rho_i$,$\alpha_j\leqslant\frac{c_2}{\bar{N}}\rho_j$. 并且由于只有多数类、少数类支持向量和错分的多数类、少数类样本对应的 $\alpha_i>0$,$\alpha_j>0$. 可得:

$$M<\frac{c_1}{N}\sum_{i=1}^{P_{m_+}+Q_{m_+}}\rho_i+\frac{c_2}{\bar{N}}\sum_{j=1}^{P_{m_-}+Q_{m_-}}\rho_j \tag{38}$$

综上,

$$\frac{c_1}{N}\sum_{i=1}^{P_{m_+}}\rho_i+\frac{c_2}{\overline{N}}\sum_{j=1}^{P_{m_-}}\rho_j<M<\frac{c_1}{N}\sum_{i=1}^{P_{m_+}+Q_{m_+}}\rho_i+\frac{c_2}{\overline{N}}\sum_{j=1}^{P_{m_-}+Q_{m_-}}\rho_j \tag{39}$$

其中,P_{m_+}是带有间隔错误的多数类样本的数目,Q_{m_+}是多数类支持向量的数目,P_{m_-}是带有间隔错误的少数类样本的数目,Q_{m_-}是少数类支持向量的数目.

2.2.5 DSMSM-SVDD 算法引入相对密度的必要性

由于 $P_{m_+}+Q_{m_+}<N$,$P_{m_-}+Q_{m_-}<\overline{N}$. 则由式(39)可以推得:

$$M<\frac{c_1}{N}\sum_{i=1}^{P_{m_+}+Q_{m_+}}\rho_i+\frac{c_2}{\overline{N}}\sum_{j=1}^{P_{m_-}+Q_{m_-}}\rho_j<\frac{c_1}{N}\sum_{i=1}^{N}\rho_i+\frac{c_2}{\overline{N}}\sum_{j=1}^{\overline{N}}\rho_j \tag{40}$$

假如算法没有考虑每个样本的相对密度,即 $\rho_i=1$,$\rho_j=1$. 则式(40)化为:

$$M<c_1+c_2 \tag{41}$$

由式(41)可知,在 c_1,c_2 都较小的情况下,不考虑密度的影响,M 的取值范围完全取决于惩罚因子 C 的大小. 当实验注重多数类和少数类训练样本的最大软间隔时,由于 M 受取值上界的限制,可能会导致算法的泛化能力提升不明显. 因为 $\rho_i\geqslant 1$,$\rho_j\geqslant 1$ 可以使 M 的取值更灵活,所以密度参数的提出是必要的.

3 实验与分析

3.1 不均衡数据分类性能度量指标

为了能量化地衡量分类算法对不均衡数据分类性能的影响,近些年来研究人员提出了一些适用于不均衡数据分类性能的评测指标,为了方便描述,这里首先介绍混合矩阵的定义:

表 1　混合矩阵

	预测正类	预测负类
真正正类	TP	FN
真正负类	FP	TN

其中,TP(True Positive)指使用分类算法将原本属于正类的样本正确地预测为正类的样本数;FP(False Positive)指使用分类算法将原本属于负类的样本错误地预测为正类的样本数;FN(False Negative)指使用分类算法将原本属于正类的样本错误地预测为负类的样本数;TN(True Negative)指使用分类算法将原本属于负类的样本正确地预测为负类的样本数. 因此,可得:

正类样本正确率(Sensitivity):

$$\text{Sensitivity}=\frac{\text{TP}}{(\text{TP}+\text{FN})}$$

负类样本正确率(Specificity):

$$\text{Specificity}=\frac{\text{TN}}{(\text{TN}+\text{FP})}$$

正类样本准确率(Precision):

$$\text{Precision}=\frac{\text{TP}}{(\text{TP}+\text{FP})}$$

几何平均正确率 G-Mean:

$$\text{G}=\sqrt{\text{Sensitivity}\cdot\text{Specificity}}$$

正类样本 F-Measure 指标:

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

3.2 实验数据来源

为了验证本文提出的 DSMSM-SVDD 算法在处理不均衡数据分类方面的优势,我们进行了下列实验. 其中实验数据均来源于国际机器学习标准数据库 UCI 中的 11 组不同的数据集,数据特征信息见表 2. 其中类别表示选择出来作为少数类和多数类样本的代表类别. 实验环境:Windows7 操作系统,CPU:Intel i7,3.4G 处理器,仿真软件为 Matlab2010b.

表 2　实验数据集描述

数据集	属性	少数类/多数类	类别
german	24	300/700	B/G
yeast	8	429/463	NUC: CYT
wpbc	34	46/148	R:N
pima	8	268/500	1:0
abalone	8	634/689	10:9
sonar	60	97/111	-1:1
haberman	3	81/225	2:1
ionosphere	33	126/225	Bad:Good
phome	5	317/683	1:0
breast cancer	9	239/444	4:2
wine	13	71/59	2:1

3.3 M 与 $C=\left(\frac{c_1}{N},\frac{c_2}{\overline{N}}\right)$取值范围的相互影响

为了测试参数 M 与 $C=\left(\frac{c_1}{N},\frac{c_2}{\overline{N}}\right)$取值范围之间的关系,令 DSMSM-SVDD 算法中 $\rho_i=1$,$\rho_j=1$,则式(39)化为:

$$\frac{c_1}{N}P_{m_+}+\frac{c_2}{\overline{N}}P_{m_-}<M<\frac{c_1}{N}(P_{m_+}+Q_{m_+})+\frac{c_2}{\overline{N}}(P_{m_-}+Q_{m_-}) \tag{42}$$

我们设计了如下实验:选定 $C=(0.01,1)$,通过将 M 从 1 至 1000 逐步增加,来显式观察算法的分类界面位置和支持向量分布情况,结果如图 1 所示. 其中,DSMSM-SVDD 选取核半径参数为 $\sigma=1$ 的高斯核函数.

通过图 1 对比可以发现 $M=1$ 时由于算法中最大间隔项的权重较小,算法等同于 C-SVDD. 当 M 从 1 增加至 10 时,最大间隔项的权重增大,使得分类界面向两类样本最大间隔处偏移,呈现 SVM 分类特征,降低多数类样本错分率,泛化能力有所提升. 但随着 M 值的进一步增加,其分类界面位置和支持向量分布情况并不理想. 其原因由不等式

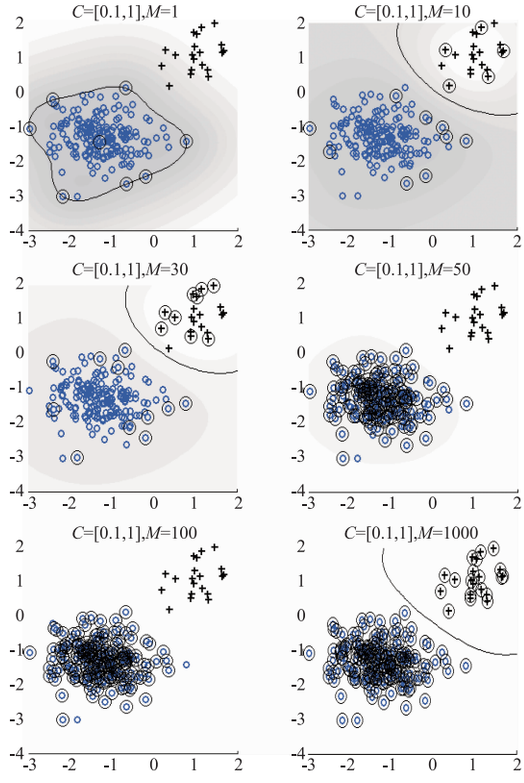

图1 DSMSM-SVDD算法的分类结果随M值变化示意图

(42)知:当 C 取值较小,随着 M 值的增大,因受不等式(42)中 M 上界的约束,导致支持向量个数增多,当 $M=50$ 时,几乎所有多数类样本成为支持向量,出现过拟合现象. 另外,当 $M=1000$ 时,最优分类界面出现 SVM 分类特征是因 M 值过大,所有的多数类样本作为支持向量也不能满足不等式(42)约束,迫使所有少数类样本也充当支持向量,虽分类界面向中间移动,但此时算法要消耗大量的时间与空间,降低算法实现速度,且算法完全过拟合. 因此可初步得出如下结论:M 值的调节范围是受 C 值的影响. 当 C 取值较小,M 的取值不能过大,否则为了满足 M 值上界的约束将导致支持向量个数增多,出现过拟合现象. 因此当 C 较小,应尽可能地选取较小合理的 M 值.

为了能够进一步量化地观察 M 与 C 值之间的相互关系,我们选取来自 UCI 的 wine 数据集作为本次实验的测试数据,其中多数类与少数类样本数目按 30:1 的比例来选取. M 值从 1 至 1000 逐渐增加,观察 C 在分别为(0.01,1),(0.1,1),(10,100)时的 G-Mean,F-Measure 性能曲线变化趋势,F-Measure,G-Mean 值越大,说明该算法的综合分类性能越好. 其他参数同上,测试结果如图 2 所示.

通过图 2 可知:在 M 值逐渐增大的过程中,较大 C 值的 G-Mean,F-Measure 性能达到稳定时的 M 值要大于较小 C 值的 G-Mean,F-Measure 性能达到稳定时的 M 值. 即当 C 值较大时,M 的取值范围扩大;同样,当 C 值

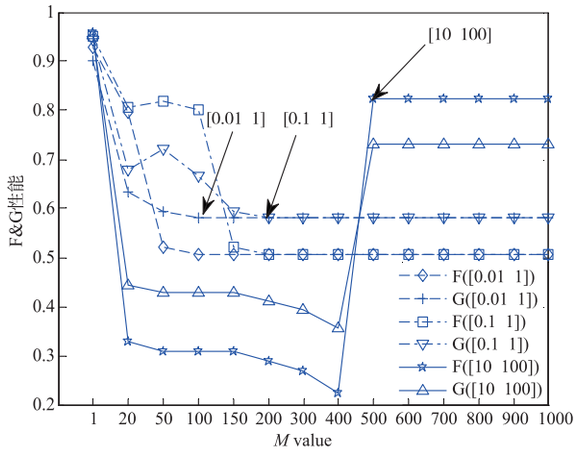

图2 wine数据集下M值对算法性能调节区间分布情况图

较小时,M 的取值范围缩小,该实验现象也进一步验证了前期实验得出的 M 值的上界受 C 值影响的初步结论. 另外,通过实验可知当 M 取值较大时,C 必须取相应较大的值,才能防止因受式(42)右半部分不等式限制导致支持向量个数增大而产生过拟合现象. 而当 M 值较小时,C 的取值范围可适当扩大,但 C 的取值不可以无限制小,仍要满足不等式(42). 该实验结论同样与前面实验得到的初步结论相吻合.

3.4 DSMSM-SVDD 与 C-SVDD 算法性能比较实验及分析

为了比较本文提出的 DSMSM-SVDD 算法与 C-SVDD 算法在处理不均衡数据分类问题上的性能,这里做了如下实验:实验数据集如表 2 所示. 我们将表 2 中的多数类数据和少数类数据按 10:1 的比例进行选取. 对于每类数据集,为了防止数据的随机影响,我们通过 10 次交叉验证法进行实验测试. DSMSM-SVDD 与 C-SVDD 的分类性能最终以 G-Mean 性能指标与支持向量个数(Support Vector,SV)的统计平均值作为评价标准. 其中,DSMSM-SVDD 与 C-SVDD 均采用核半径参数 $\sigma=10$ 的高斯核函数,$C=(0.01,1)$ 的惩罚因子,DSMSM-SVDD 算法中 $M=50$,$\omega=2$,$s=3$. 实验结果如图 3 及图 4 所示.

通过图 3 可知,本文提出的 DSMSM-SVDD 算法的 G-Mean 性能指标值均高于 C-SVDD,即本文算法在处理不均衡数据分类问题时,较 C-SVDD 具有更好的泛化能力. 这是由于 C-SVDD 只考虑包含多数类样本的超球面半径项,而没能考虑两者的间隔信息,使得到的分类界面受不均衡数据的影响产生偏移. 同时,该算法在处理训练样本整体分布特性上具有一定局限性,很可能将密度大的非支持向量区域误判,导致分类精度不高. 而本文提出的算法不仅充分考虑了两类样本的间隔信息,同时通过对多数类样本增加相对密度的方式来保证在目标函数的求解过程中考虑到数据分布差异的影

响,使求解的最优分类界面更加合理,进而提升了算法泛化性能. 另外,通过图 4 的条形图可以看到,除了 breast cancer ,wine 数据集之外,本文算法的支持向量个数远远小于 C-SVDD. 由于两种算法的复杂度均取决于支持向量的个数,因此可以说本文算法在时间和空间上的消耗都远小于 C-SVDD 算法.

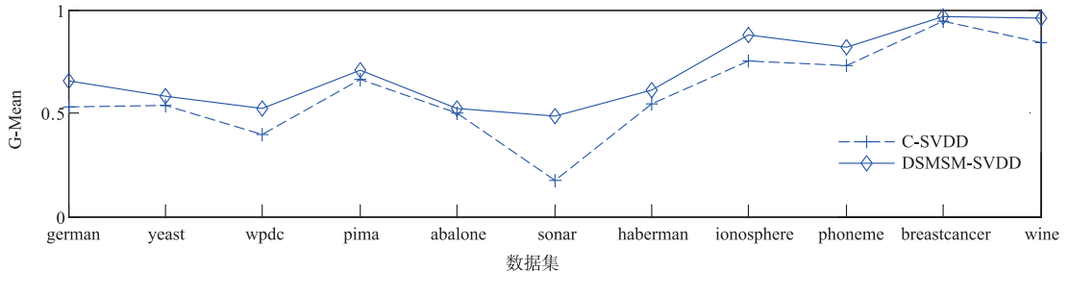

图3　DSMSM-SVDD与C-SVDD算法对数据集分类的G-Mean性能指标对比图

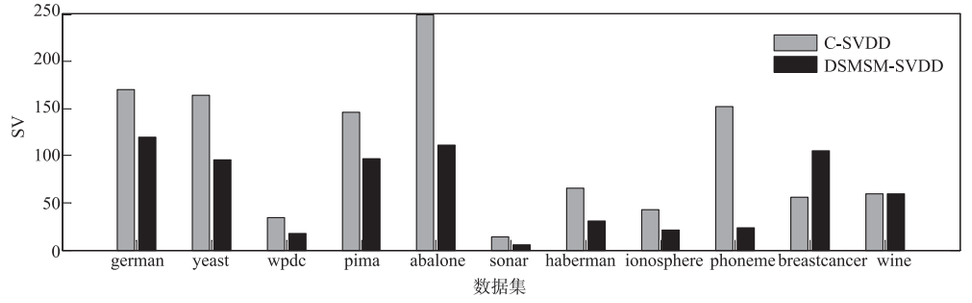

图4　DSMSM-SVDD与C-SVDD算法支持向量个数条形对比图

3.5　高斯核函数半径值对算法分类性能的影响

为了考察高斯核函数半径参数 σ 对 DSMSM-SVDD 算法分类性能的影响,我们从表 2 的数据集中选取 pima,yeast,haberman,german,ionosphere 5 类难分数据集作为本次实验数据. 每类数据按 10∶1 的比例进行多数类与少数类样本的选取,通过 10 次交叉验证方法测试. 其中,σ 的选定区间为[3,100],其他参数 $M=50$,$\omega=2$,$s=3$,$C=(0.1,1)$. 实验通过改变 σ 值来观察 DSMSM-SVDD 算法的 G-Mean 性能的变化趋势,结果如图 5 所示.

通过图 5 我们可以得出:DSMSM-SVDD 算法随着参数 σ 值的增大,各数据集的综合分类性能指标 G-Mean 值的变化趋势缓慢,彼此之间的区分不明显. 这是因为本文提出的 DSMSM-SVDD 算法中采用了最大软间隔正则项,使得优化后得到的分类界面泛化能力的提升很大程度上取决于间隔项的影响,这也在一定程度上抵消了高斯核半径对算法分类性能的影响.

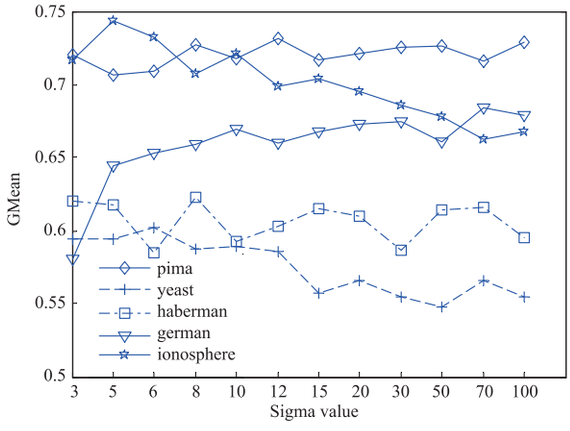

图5　DSMSM-SVDD算法G-MEAN性能随高斯核参数变化图

4　结论

针对不均衡数据的分类问题,提出一种基于密度敏感最大软间隔支持向量域描述算法. 通过实验分析得到如下结论:

(1) 为了消除噪声影响同时防止算法欠拟合,本文通过引入相对密度来反映训练样本的重要程度,使处于相对密度高的非支持向量区域的训练样本较相对密度低的训练样本更易落入超球体之内,进而消除噪声影响. 实验部分通过和 C-SVDD 进行对比的结果表明,该方法更有利于算法准确地识别出给定目标数据集的数据域描述,进而提高了算法的泛化性能.

(2) 为了消除不均衡数据集对 C-SVDD 算法性能的影响,本文在充分利用少数类样本信息的前提下,借鉴传统 SVM 最大化间隔思想,将最大软间隔加入 C-SVDD 的优化目标函数中. 实验结果表明,该方法有效提高算法处理不均衡数据的分类性能,同时使分类结果不随高斯核参数的改变而明显改变,这在一定程度上消除了高斯核参数对算法性能的影响.

(3) 实验部分将间隔项参数 M、惩罚因子 C 以及

高斯核半径参数 σ 之间的关系进行分析讨论并给出相应的结论. 值得一提的是,如何实现两个参数的联合优化对于提升本文算法处理不均衡数据分类性能而言至关重要,这也将是本课题下一阶段研究的重点.

参考文献

[1] Gu B, Sun X M, Sheng V S. Structural minimax probability machine[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(7): 1646 – 1656.

[2] Luo FF, Guo W Z, Yu Y L, et al. A multi-label classification algorithm based on kernel extreme learning machine [J]. Neurocomputing, 2017, 260: 313 – 320.

[3] H B, Ho H C, Z J, et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping[J]. Geoderma, 2016, 265: 62 – 77.

[4] Zhang X, Zhang X. Adaptive multiclass support vector machine for multimodal data analysis[J]. Pattern Recognition, 2017, 70: 177 – 184.

[5] Zuo W M, Wang F Q, Zhang D, et al. Distance metric learning via iterated support vector machines[J]. IEEE Transactions on Image Processing, 2017, 26(10): 4937 – 4950.

[6] 陶新民,李震,刘福荣,等. 基于精简集支持向量机的变压器故障检测方法[J]. 高电压技术,2016,42(10):3199 – 3206.
Tao Xinmin, Li Zhen, Liu Furong, et al. Fault detection method for power transformer based on SVM using reduced vector set[J]. High Voltage Engineering, 2016, 42(10): 3199 – 3206. (in Chinese)

[7] Zhou Y H, Zhou Z H. Large margin distribution learning with cost interval and unlabeled data[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1749 – 1763.

[8] 毕冉,李建中,高宏. 无线传感器网络中基于双阈值的分布式检测算法[J]. 电子学报,2014,42(8):1594 – 1600.
Bi Ran, Li Jianzhong, Gao Hong. Dual threshold based distributed monitoring algorithm in wireless sensor network [J]. Acta Electronica Sinica, 2014, 42(8): 1594 – 1600. (in Chinese)

[9] 陶新民,张冬雪,郝思媛,等. 基于谱聚类欠取样的不均衡数据 SVM 分类算法[J]. 控制与决策,2012,27(12):1761 – 1768.
Tao Xinmin, Zhang Dongxue, HaoSiyuan, et al. SVM classifier forunbalanced data based on spectrum cluster-based under-sampling approaches [J]. Control and Decision, 2012, 27(12): 1761 – 1768. (in Chinese)

[10] Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling techniques [J]. IEEE Transactionson Knowledgeand Data Engineering, 2016, 28(1): 238 – 251.

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

[12] 杨金鸿,邓廷权. 一种基于单簇核 PCM 的 SVDD 离群点检测方法[J]. 电子学报,2017,45(4):813 – 819.
Yang Jinhong, Deng Tingquan. A one-cluster kernel PCM based SVDD method for outlier detection[J]. Acta Electronica Sinica, 2017, 45(4): 813 – 819. (in Chinese)

[13] Lazzaretti A E, Tax D M J, Neto H V, et al. Novelty detection and multi-class classification in power distribution voltage waveforms [J]. Expert Systems with Applications, 2016, 45: 322 – 330.

[14] 程昊翔,王坚. 基于快速聚类分析的支持向量数据描述算法[J]. 控制与决策,2016,31(3):551 – 554.
Cheng Haoxiang, Wang Jian. Support vector data description based on fast clustering analysis[J]. Control and Decision, 2016, 31(3): 551 – 554. (in Chinese)

[15] Huang J, Yan X F. Related and independent variable fault detection based on KPCA and SVDD [J]. Journal of Process Control, 2016, 39: 88 – 99.

[16] 文传军,詹永照,陈长军. 最大间隔最小体积球形支持向量机[J]. 控制与决策,2010,25(1):79 – 83.
Wen Chuanjun, Zhan Yongzhao, Chen Changjun. Maximal-margin minimal-volume hypersphere support vector machine [J]. Control and Decision, 2010, 25(1): 79 – 83. (in Chinese)

作者简介

陶新民 男,1973 年生. 博士,教授,主要研究领域为智能信号处理、故障诊断、模式识别.
E-mail: taoxinmin@nefu.edu.cn

李晨曦 女,1993 年生. 硕士研究生,研究方向为故障诊断、模式识别.
E-mail: chenxili0613@163.com