计算机科学
COMPUTER SCIENCE

一种新的代价敏感SVDD二类分类方法

吴崇明, 王晓丹, 赵振冲

引用本文

吴崇明, 王晓丹, 赵振冲.一种新的代价敏感SVDD二类分类方法[J]. 计算机科学, 2023, 50(6A): 220300202-5.

WU Chongming, WANG Xiaodan, ZHAO Zhenchong. New Cost Sensitive SVDD Binary Classification Method [J]. Computer Science, 2023, 50(6A): 220300202-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300202

一种新的代价敏感 SVDD 二类分类方法

吴崇明[1]　王晓丹[2]　赵振冲[2]
1 西京学院商学院　西安 710123
2 空军工程大学防空反导学院　西安 710051
(afeu_w@163.com)

摘　要　为提升代价敏感分类性能,通过提升较高误分代价类别的学习精度来降低总误分代价,利用支持向量域描述(Support Vector Domain Description,SVDD)实现代价敏感分类,提出一种代价敏感 SVDD 二类分类方法 CS-SVDD。该方法首先将单类 SVDD 拓展为二类分类 SVDD,对不同类别分别构建 SVDD 超球体,通过误分类代价调节 SVDD 分类器对不同类别样本的分类精度,对误分代价高的类别进行更为精确的学习,从而降低总误分代价;对于处于两个超球体之外或覆盖区域的类别属性不明确的样本,以误分代价最小为原则定义代价敏感决策规则。在人工数据集和 UCI 数据集上与同类方法进行了实验比较,实验结果表明了所提方法的有效性。

关键词:代价敏感分类;支持向量数据描述;支持向量
中图法分类号　TP391

New Cost Sensitive SVDD Binary Classification Method

WU Chongming[1],WANG Xiaodan[2] and ZHAO Zhenchong[2]
1 College of Business,Xijing University,Xi'an 710123,China
2 College of Air and Missile Defense,Air Force Engineering University,Xi'an 710051,China

Abstract　In order to improve the performance of cost-sensitive classification,this paper improves the learning accuracy of higher misclassification cost categories to reduce the total misclassification cost,uses support vector domain description(SVDD) to realize cost sensitive classification,and proposes a cost sensitive SVDD two-class classification method,CS-SVDD. This method first expands single class SVDD to two class classification SVDD,and constructs SVDD hyperspheres for different categories. by adjusting the classification accuracy of SVDD classifier for different class samples through the misclassification cost,the class with high misclassification cost can be more accurately learned,so as to reduce the total misclassification cost. For the samples with ambiguous category attributes outside the two hyperspheres or in the coverage area,cost sensitive decision rules are defined based on the principle of minimum misclassification cost. Experimental results on artificial data sets and UCI data sets show the effectiveness of the proposed method.

Keywords　Cost sensitive classification,Support vector data description,Support vector

1　引言

对于代价敏感(cost-sensitive)分类问题,不同类别的误分结果造成的损失不同。故障检测、疾病诊断、金融欺诈检测等都属于代价敏感分类问题。对代价敏感问题的研究最早始于 1974 年 Habbema 等[1]提出的常数错误代价的概念。2001 年,美国加州大学教授 Elkan[2]在国际人工智能联合会议上发表了关于代价敏感奠基性的文章,极大地促进了代价敏感分类问题的研究。代价敏感学习利用不同类别样本的误分代价调整分类边界或决策阈值,使识别的风险达到最小。其本质是识别错误率与识别风险之间的折中。传统分类方法都有其代价敏感拓展,如代价敏感支持向量机[3-5]、代价敏感贝叶斯分类器[6-7]、代价敏感决策树[8-11]、代价敏感集成[12-13]和代价敏感大间隔分布学习机[14-15]等。

支持向量数据描述(SVDD)是 Tax 等[16]提出的用于学习单类数据的方法。SVDD 对单类分类具有较好学习能力,其基本思想是:对单类样本在特征空间构建一个体积尽可能小的超球体,同时使该超球体包含尽可能多的目标样本。为了更好地描述超球体边界,Tax 等[17]对原方法进行了改进,在训练的过程中同时考虑目标类和非目标类样本,使超球体包含尽可能多的目标样本和尽可能少的非目标样本。SVDD 的优化问题由于是简单的凸规划,容易求解,几何意义明显,并且可以通过核函数将特征映射到高维空间,同时继承了 SVM 的小样本特性和强泛化能力,因此受到广泛研究。很多改进算法被提出,如 D-SVDD[18],υ-NSVDD[19], SA-SVDD[20], LSSVDD[21]和 PSVDD[22]等,被应用于了奇异值检测[23]、机器故障诊断[24]、飞机故障检测[25]等问题。

基于 SVDD 实现代价敏感分类,目前少有研究。SVDD

基金项目:国家自然科学基金项目(61876189,61273275)
This work was supported by the National Natural Science Foundation of China(61876189,61273275).
通信作者:王晓丹(afeu_w@163.com)

对单类样本具有较好的学习能力，可通过分类器学习参数调整，实现对目标类别的高精度学习。该特点可以被用来构建代价敏感 SVDD 分类器，即对不同类别分别构建 SVDD 超球体，并通过对不同类别的误分代价调节 SVDD 分类器对不同类别样本的分类精度。对于代价敏感分类问题，重点关注误分代价高的类别，对其进行更为精确的学习，从而实现总误分代价最小的目的。

基于上述思想，本文提出了一种代价敏感二类 SVDD 方法(Cost-sensitive Two Class SVDD，CS-SVDD)。首先通过重新定义 SVDD 的目标函数和约束条件，将单类 SVDD 拓展为 2 类分类 SVDD；定义了基于误分代价的代价敏感二类 SVDD(CSSVDD)优化问题的目标函数和约束条件，引入代价因子用以调节误分代价对惩罚因子 C 的影响程度，使误分代价高的类别的样本尽可能包含在超球体内，从而降低误分代价；最后，对于处于两个超球体之外或覆盖区域的类别属性不明确的样本，以误分代价最小为原则定义了代价敏感决策规则。

2 SVDD 算法

SVDD 最初是为学习单类数据提出的。设训练集 $I=\{\boldsymbol{x}_i\in R^d\mid i=1,2,\cdots,N\}$，SVDD 的目标就是要获得一个以 $\boldsymbol{a}$ 为中心的超球体，使其包含尽可能多的目标样本的同时半径 R 尽可能小，即如下优化问题[16]：

$$\begin{aligned}&\min R^2+C\sum\xi_i\\&\text{s.t. }(\boldsymbol{x}_i-\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{a})\leqslant R^2+\xi_i,i=1,2,\cdots,N\\&\quad\xi_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{1}$$

其中，ξ_i 为松弛变量，允许一部分远离训练集中心的样本处于球体外；C 为惩罚因子，控制着对错分样本的惩罚程度。对式(1)引入拉格朗日乘子 $\alpha_i\geqslant 0,\beta_i\geqslant 0$，得到拉格朗日函数：

$$L(R,\boldsymbol{a},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=R^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i(R^2+\xi_i-\|\boldsymbol{x}_i-\boldsymbol{a}\|^2)-\sum_{i=1}^{N}\beta_i\xi_i\tag{2}$$

分别令 $R,\boldsymbol{a},\boldsymbol{\xi}$ 的偏导等于零，有：

$$\begin{cases}\dfrac{\partial L}{\partial R}=0\Rightarrow\sum\limits_{i=1}^{N}\alpha_i=1\\[2mm]\dfrac{\partial L}{\partial\boldsymbol{a}}=0\Rightarrow\boldsymbol{a}=\sum\limits_{i=1}^{N}\alpha_i\boldsymbol{x}_i\\[2mm]\dfrac{\partial L}{\partial\boldsymbol{\xi}_i}=0\Rightarrow C-\alpha-\beta_i=0\end{cases}\tag{3}$$

由于 $\beta_i\geqslant 0$，因此 $C\geqslant\alpha_i\geqslant 0$。将式(3)代入式(2)，得到式(1)的对偶问题：

$$\begin{aligned}&\max_{\boldsymbol{a}}L(\alpha)=-\sum_{i,j=1}^{N}\alpha_i\alpha_j(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{N}\alpha_i(\boldsymbol{x}_i,\boldsymbol{x}_i)\\&\text{s.t. }\sum_{i=1}^{N}\alpha=1,C\geqslant\alpha_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{4}$$

其中，$(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 的内积。求解以上对偶问题，能够获得一个最优解，记为 $\boldsymbol{\alpha}^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*]$。则对于一个测试样本 $\boldsymbol{x}$，决策函数可表示为：

$$\begin{aligned}f(\boldsymbol{x})&=R^2-\|\boldsymbol{x}-\boldsymbol{a}\|^2\\&=(\boldsymbol{x}_{sv},\boldsymbol{x}_{sv})-(\boldsymbol{x},\boldsymbol{x})-2\sum_{i\in SV}\alpha_i^*((\boldsymbol{x}_i,\boldsymbol{x}_{sv})-(\boldsymbol{x}_i,\boldsymbol{x}))\end{aligned}\tag{5}$$

其中，$SV=\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in I,0<\alpha_i^*\leqslant C\}$ 为支持向量的集合，$\boldsymbol{x}_{SV}$ 为任一满足 $0<\alpha_{SV}^*<C$ 的边界样本。如果 $f(\boldsymbol{x})>0$，意味着样本 $\boldsymbol{x}$ 到球心 $\boldsymbol{a}$ 的距离小于半径 R(即在超球体内)，则 $\boldsymbol{x}$ 被分为目标类；否则 $\boldsymbol{x}$ 被识别为异常样本。

与 SVM 类似，当问题在原始空间线性不可分时，可以利用核函数 $k(x_i,x_j)$ 来替换内积，将训练数据映射到更高维的核空间，以获得对原数据更细致的描述。此时式(1)所示的优化问题可以重新表示为：

$$\begin{aligned}&\min R^2+C\sum\xi_i\\&\text{s.t. }(\varphi(\boldsymbol{x}_i)-\boldsymbol{a})^{\mathrm{T}}(\varphi(\boldsymbol{x}_i)-\boldsymbol{a})\leqslant R^2+\xi_i,i=1,2,\cdots,N\\&\quad\xi_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{6}$$

其中 $\varphi(\cdot)$ 为映射函数，且 $k(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。此时式(5)变为：

$$f(\boldsymbol{x})=k(\boldsymbol{x}_{sv},\boldsymbol{x}_{sv})-k(\boldsymbol{x},\boldsymbol{x})-2\sum_{i\in SV}\alpha_i^*(k(\boldsymbol{x}_i,\boldsymbol{x}_{sv})-k(\boldsymbol{x}_i,\boldsymbol{x}))\tag{7}$$

高斯核函数 $k(x,y)=\exp(-\|x-y\|^2/\sigma^2)$ 是 SVDD 使用最多的核函数[18,26]。

3 代价敏感二类 SVDD 算法

设二类分类问题的训练集 $I=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$ 为 d 维特征向量，$y_i\in\{+,-\}$ 为正负类的类别标签，$I_+=\{\boldsymbol{x}_i\mid y_i=+,i=1,2,\cdots,n_+\}$，$I_-=\{\boldsymbol{x}_i\mid y_i=-,i=1,2,\cdots,n_-\}$ 分别为正、负类样本集合，且 $N=n_++n_-$。

对于二类问题，可以对每一类样本定义一个超球体，使得每个超球体尽可能多地包含各自的训练样本，并同时最小化两个超球体的体积。以此定义二类 SVDD 的优化问题为：

$$\begin{aligned}&\min R_+^2+R_-^2+C_+\sum_{i=1}^{n_+}\xi_{+i}+C_-\sum_{i=1}^{n_-}\xi_{-i}\\&\text{s.t. }\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_+\|^2\leqslant R_+^2+\xi_{+i},\boldsymbol{x}_i\in I_+\\&\quad\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_-\|^2\leqslant R_-^2+\xi_{-i},\boldsymbol{x}_i\in I_-\\&\quad\xi_{+i}\geqslant 0,i=1,2,\cdots,n_+;\xi_{-i}\geqslant 0,i=1,2,\cdots,n_-\end{aligned}\tag{8}$$

其中，$R_+,R_-,\boldsymbol{a}_+,\boldsymbol{a}_-$ 为正、负类超球体的半径和球心，$C_+>0,C_->0,\xi_{+i}\geqslant 0,\xi_{-i}\geqslant 0$ 为正、负类的惩罚因子和松弛变量。

代价敏感二类分类问题的误分代价矩阵为：

$$\boldsymbol{Cost}=\begin{bmatrix}c_{++}&c_{+-}\\c_{-+}&c_{--}\end{bmatrix}\tag{9}$$

其中，$c_{ij}(i,j=+1,-1)$ 为将类别 i 中的样本误分为类别 j 的代价，且有 $c_{++}=c_{--}=0$。

SVDD 超球体的体积与 C 是正相关的关系。为降低误分代价，可以依据不同类别的误分类代价设定对应的惩罚因子的值：对于误分代价高的类别，应使其对应的 C 尽量大，从而尽量将该类样本包含在超球体内；对于误分代价小的类别，应使其对应的 C 应尽量小，以尽量避免将代价高的样本包括在内。

基于以上分析，本文引入代价因子用以调节误分代价对惩罚因子 C 的影响程度，依据不同类别的误分类代价设定对应的惩罚因子的值，具体如下：

$$\begin{cases}C_+=C*(c_{+-}/c_{-+})^{\rho}\\C_-=C*(c_{-+}/c_{+-})^{\rho}\end{cases}\tag{10}$$

其中，$\rho\geqslant 0$ 为代价因子，$C>0$ 用以控制模型的整体泛化能力。由式(10)可以看出，对于所有的 $\rho>0$，误分代价大的类的惩罚因子总是要比另一类的大。

将式(10)代入式(8)，代价敏感 SVDD(CSSVDD)的优化问题可表示为：

$$\begin{aligned}&\min\ R_+^2+R_-^2+C*(c_{+-}/c_{-+})^p\sum_{i=1}^{n_+}\xi_{+i}+\\&\quad C*(c_{-+}/c_{+-})^p\sum_{i=1}^{n_-}\xi_{-i}\\&\text{s.t.}\ \ \|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_+\|^2\leqslant R_+^2+\xi_i,\boldsymbol{x}_i\in I_+\\&\quad\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_-\|^2\leqslant R_-^2+\xi_{-i},\boldsymbol{x}_i\in I_-\\&\quad\xi_{+i}\geqslant 0,i=1,2,\cdots,n_+\\&\quad\xi_{-i}\geqslant 0,i=1,2,\cdots,n_-\end{aligned}\tag{11}$$

引入拉格朗日乘子 $\alpha_{+i}\geqslant 0,\alpha_{-i}\geqslant 0,\beta_{+i}\geqslant 0,\beta_{-i}\geqslant 0$，则式(11)对应的拉格朗日函数为：

$$\begin{aligned}&L(R_+,R_-,\boldsymbol{a}_+,\boldsymbol{a}_-,\xi_{+i},\xi_{-i})\\&=R_+^2+R_-^2+C*(c_{+-}/c_{-+})^p\sum_{i=1}^{n_+}\xi_{+i}+C*(c_{-+}/c_{+-})^p\sum_{i=1}^{n_-}\\&\quad\xi_{-i}-\sum_{i=1}^{n_+}\alpha_{+i}(R_+^2+\xi_{+i}-\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_+\|^2)-\sum_{i=1}^{n_-}\alpha_{-i}(R_-^2+\\&\quad\xi_{-i}-\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_-\|^2)-\sum_{i=1}^{n_+}\beta_{+i}\xi_{+i}-\sum_{i=1}^{n_-}\beta_{-i}\xi_{-i}\end{aligned}\tag{12}$$

分别令 L 对 $R_+,R_-,\boldsymbol{a}_+,\boldsymbol{a}_-,\xi_{+i},\xi_{-i}$ 的偏导为零，有：

$$\begin{cases}\dfrac{\partial L}{\partial R_+}=0\Rightarrow\sum\limits_{i=1}^{n_+}\alpha_{+i}=1\\[2ex]\dfrac{\partial L}{\partial\boldsymbol{a}_+}=0\Rightarrow\boldsymbol{a}_+=\sum\limits_{i=1}^{n_+}\alpha_{+i}\varphi(\boldsymbol{x}_i)\\[2ex]\dfrac{\partial L}{\partial\xi_{+i}}=0\Rightarrow C*(c_{+-}/c_{-+})^p-\alpha_{+i}-\beta_{+i}=0\end{cases}\tag{13}$$

$$\begin{cases}\dfrac{\partial L}{\partial R_-}=0\Rightarrow\sum\limits_{i=1}^{n_-}\alpha_{-i}=1\\[2ex]\dfrac{\partial L}{\partial\boldsymbol{a}_-}=0\Rightarrow\boldsymbol{a}_+=\sum\limits_{i=1}^{n_-}\alpha_{-i}\varphi(\boldsymbol{x}_i)\\[2ex]\dfrac{\partial L}{\partial\xi_{-i}}=0\Rightarrow C*(c_{-+}/c_{+-})^p-\alpha_{-i}-\beta_{-i}=0\end{cases}$$

将式(13)代入式(12)，得到式(11)的对偶问题：

$$\begin{aligned}\max_{\boldsymbol{\alpha}}L(\boldsymbol{\alpha}_+,\boldsymbol{\alpha}_-)=&-\sum_{i,j=1}^{n_+}\alpha_{+i}\alpha_{+j}k(\boldsymbol{x}_{+i},\boldsymbol{x}_{+j})-\sum_{i,j=1}^{n_-}\alpha_{-i}\alpha_{-j}\\&k(\boldsymbol{x}_{-i},\boldsymbol{x}_{-j})+\sum_{i=1}^{n_+}\alpha_{+i}k(\boldsymbol{x}_{+i},\boldsymbol{x}_{+i})+\sum_{i=1}^{n_-}\\&\alpha_{-i}k(\boldsymbol{x}_{-i},\boldsymbol{x}_{-i})\\\text{s.t.}\ \ \sum_{i=1}^{n_+}\alpha_{+i}=1,&C*(c_{+-}/c_{-+})^p\geqslant\alpha_{+i}\geqslant 0,i=1,2,\cdots,n_+\\\sum_{i=1}^{n_-}\alpha_{-i}=1,&C*(c_{-+}/c_{+-})^p\geqslant\alpha_{-i}\geqslant 0,i=1,2,\cdots,n_-\end{aligned}\tag{14}$$

利用以上对偶问题的解 $\alpha_{+i}^*,\alpha_{-j}^*,i=1,2,\cdots,n_+,j=1,2,\cdots,n_-$，得到两个超球体的球心 $\boldsymbol{a}_+,\boldsymbol{a}_-$ 和半径 R_+,R_-。

对未知类别测试样本 x 进行识别时，会出现以下情况：

$$\begin{cases}1)\ f_+(\boldsymbol{x})\geqslant 1,f_-(\boldsymbol{x})\leqslant 1\\2)\ f_+(\boldsymbol{x})\leqslant 1,f_-(\boldsymbol{x})\geqslant 1\\3)\ f_+(\boldsymbol{x})>1,f_-(\boldsymbol{x})>1\\4)\ f_+(\boldsymbol{x})<1,f_-(\boldsymbol{x})<1\end{cases}\tag{15}$$

第 1)种情况，表明测试样本 $\boldsymbol{x}$ 在正类超球体内。第 2)种情况，表明测试样本 $\boldsymbol{x}$ 在负类超球体内。对于第 1)和 2)两种情况，$\boldsymbol{x}$ 仅在一个超球体内，此类样本的可分性强，将其判别为对应的类别即可。第 3)和 4)种情况，表明 $\boldsymbol{x}$ 在两个超球体的重叠区域、在两个超球体之外，此类样本的可分性弱，要进行进一步识别。

当 $\boldsymbol{x}$ 处于两个超球体之外或重叠区域时，离某个球心的距离越小且该超球体的半径越大，属于该类别的可能性越大。因此，可根据样本与球心的距离和超球体的半径，定义样本属于不同类别的隶属度函数。令：

$$f_i(\boldsymbol{x})=\frac{R_i^2}{\|\varphi(\boldsymbol{x})-\boldsymbol{a}_i\|^2},i=+,-\tag{16}$$

则 $\boldsymbol{x}$ 属于某一类的隶属度函数可定义为：

$$P(i|\boldsymbol{x})=\frac{f_i(\boldsymbol{x})}{S_l}=\frac{R_i^2}{\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_i\|^2*S_l},i=+,-\tag{17}$$

其中，$S_l=\sum\limits_{i=+,-}f_i(\boldsymbol{x})$。结合误分类代价，情况 3)和 4)条件下将 $\boldsymbol{x}$ 识别为正负类的代价可近似表示为：

$$\begin{aligned}c_+&=P(-|\boldsymbol{x})c_{-+}=\frac{R_-^2*c_{-+}}{\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_-\|^2*S_l}\\c_-&=P(+|\boldsymbol{x})c_{+-}=\frac{R_+^2*c_{+-}}{\|\varphi(\boldsymbol{x}_i)-\boldsymbol{a}_+\|^2*S_l}\end{aligned}\tag{18}$$

综上，CS-SVDD 的决策规则可表示为：

$$\begin{cases}\text{if}:f_+(\boldsymbol{x})\geqslant 1\ \text{and}\ f_-(\boldsymbol{x})\leqslant 1\Rightarrow y_x=+\\\text{else if}:f_+(\boldsymbol{x})\leqslant 1\ \text{and}\ f_-(\boldsymbol{x})\geqslant 1\Rightarrow y_x=-\\\text{else if}:c_+<c_-\Rightarrow y_x=+\\\text{else}:y_x=-\end{cases}\tag{19}$$

4 实验与分析

分别利用 Banana-shaped 人工数据集和 10 个 UCI 数据集验证 CS-SVDD 的性能。Banana-shaped 人工数据集共 2 类、2 维特征。10 个 UCI 数据集的特征数、类别数和样本数如表 1 所列，所有数据预先进行归一化处理。对于多类数据集，将样本数少的类进行合并，直至只剩下 2 类。实验中，将第一类作为负类(Negative)，另一类作为正类(Positive)。

表 1 UCI 数据集(圆括号中的类别被合并为 1 类)

Table 1 UCI dataset (categories in parentheses are merged into one)

数据集	特征数	类别数	样本数	负/正样本数
Iris	4	3	150	(50 50)/50
Breast-w	9	2	699	241/458
Glass	10	6	214	(70 76)/(17 13 9 29)
Sonar	60	2	208	97/111
Wine	13	3	178	71/(59 48)
Vehicle	18	4	846	(199 218)/(217 212)
Thyriod	5	3	215	150/(35 30)
Page-blocks	10	5	5473	4913/(329 28 88 115)
Sat	36	6	6435	(1533 703 1358)/(626 707 1508)
shuttle	9	7	14500	11478/(13 39 2155 809 4 2)

分类器性能评价准则包括 3 种：错误率(Er)、G-mean、总误分类代价(Tc)。

设 tp 为正确分类的正类样本数，tn 为正确分类的负类样本个数；fp 为负类样本识别为正类的样本个数；fn 为正类样本识别为负类的样本个数，则错误率 Er、G-mean、总误分类代价 Tc 可分别表示为：

$$Er=\frac{fp+fn}{tp+tn+fp+fn}$$

$$G\text{-}mean=\sqrt{\frac{tp}{tp+fn}+\frac{tn}{tn+fp}}$$

$$Tc=fn\times c_{+1,-1}+fp\times c_{-1,+1}$$

除本文方法 CS-SVDD 外，选用其他 3 种作为对比方法：MRNB[6]，CSBN[7] 和 CSSVM[3]。采用高斯模型和核函数。

4.1 人工数据集结果及分析

利用 Banana-shaped 人工数据集进行实验。令 $C=0.1$，

$\rho=1$，固定 $c_{-+}=1$，表 2 列出了 $c_{+-}=0.5$，2 时，CS-SVDD 以及对比的 3 种方法在 99%置信水平下的实验结果。

表 2 不同方法在 Banana-shaped 数据集上的分类结果

Table 2 Classification results of different methods on Banana shaped dataset

c_{+-}	评价指标	MRNB	CSBN	CSSVM	CS-SVDD
0.5	*Er*	0.19±0.016	0.149±0.028	0.14±0.031	**0.047±0.027**
	G-mean	0.801±0.017	0.844±0.028	0.857±0.031	**0.952±0.028**
	Tc	12.9±1.29	8.65±2.07	8.95±2.64	**3.3±2.0**
2	*Er*	0.162±0.02	0.139±0.025	0.145±0.027	**0.049±0.027**
	G-mean	0.835±0.02	0.858±0.025	0.851±0.027	**0.95±0.027**
	Tc	21.1±3.83	17.5±3.69	17.6±4.35	**7.5±3.79**

从表 2 中可以看出，CS-SVDD 在 3 个指标中均表现出优异的性能。这是由于在 Banana-shaped 数据集上，CS-SVDD 能够有效利用误分类代价，调整超球体对不同类别的描述能力，对于处于两个超球体之间的类别属性不强的样本所在区域，能够使决策边界向误分代价小的类别偏移，进而降低整体识别风险，并能够获得比较高的识别正确率。

4.2 UCI 数据集结果及分析

为了进一步验证 CS-SVDD 的有效性，利用 UCI 标准数据集进行实验。实验中每个数据集被随机地分为样本数相等的两部分分别作为训练集和测试集，所有的算法均在相同的训练集和测试集上进行训练和测试。以上过程在每种代价比下重复 10 次，并以所得结果的均值作为最终结果。

令 $\rho=1$，对于 CSSVM 和 CS-SVDD，惩罚参数 C 固定为 1，参数 σ 在每一个数据集上利用 grid search 在[0.05，0.1，…，1，2，4，6，8，10]范围搜索最优值。固定 $c_{-+}=1$，分别令 $c_{+-}=0.5$，1，2，5，以对比不同代价比值条件下的实验结果。

表 3、表 4 分别列出了 $c_{+-}=0.5$，$c_{+-}=2$ 时，CS-SVDD 及对比方法在 UCI 数据集上置信水平为 99%的双边 t 估计结果。在每个表中，对应每个数据集分类的最优值被加黑。在每个表的最后，列出了 3 个评价指标的均值和每种方法的平均 Rank。

可以看出，在两种代价下，CS-SVDD 均能获得最低的平均总误分代价，平均错误率的平均 Rank 最低，以及最高的平均 *G-mean*。

$c_{+-}=0.5$ 时，CS-SVDD 在 3 种评价准则上的 Rank 均为最好。CS-SVDD 的 *Er*/*G-mean*/*Tc* 分别在 7/6/5 个数据集上优于其他方法；在 4 个数据集(Sonar，Thyriod，Sat，Shuttle)上，其性能均明显好于其他方法；在 7 个数据集上获得了最低错误率，在 5 个数据集上获得了最低总误分类代价。

当 $c_{+-}=2$ 时，CS-SVDD 在 2 种评价准则上的 Rank 均为最好，1 种评价准则上的 Rank 和 CSSVM 相近，同时明显优于其他方法。CS-SVDD 的 *Er*/*G-mean*/*Tc* 分别在 5/4/4 个数据集上优于其他方法。在 4 个数据集(Glass，Thyriod，Sat，Shuttle)上，CS-SVDD 的性能均明显好于其他方法。在 5 个数据集上获得了最低错误率，在 4 个数据集上获得了最低总误分类代价。

对于样本数目比较大的 3 个数据集(Page-blocks，Sat，Shuttle)，CS-SVDD 的性能均明显好于其他 3 种方法。这是因为当样本集比较充足时，训练样本的分布更能反映真实情况，由于样本数目比较大，每次训练的不确定性降低，得到的超球体边界趋于稳定，因此表现出比较好的效果。

表 3 $c_{+-}=0.5$ 时的 UCI 数据集分类结果

Table 3 Classification results of UCI datasetwhen $c_{+-}=0.5$

Dataset	Metrics	MRNB	CSBN	CSSVM	CSSVDD
Iris	*Er*/%	8.0±5.4	6.0±3.5	6.67±3.1	**5.3±3.7**
	G-mean/%	92.5±4.5	93.7±3.5	89.2±5.4	**94.8±3.5**
	Tc	1.1±0.8	0.75±0.54	**0.5±0.24**	0.7±0.56
Breast-w	*Er*/%	4.15±1.1	**2.57±1.1**	4.43±1.7	3.87±1.3
	G-mean/%	96.1±1.2	**97.7±1.1**	94.7±2.6	95.8±1.7
	Tc	2.55±0.74	**1.65±0.7**	2.2±0.76	2.1±0.78
Glass	*Er*/%	17.3±4.4	15.8±6.1	16.3±4.2	**14.1±3.5**
	G-mean/%	72.8±7.0	75.6±9.3	69.5±9.2	**79.7±6.0**
	Tc	2.25±0.78	2.1±0.98	**1.8±0.48**	1.95±0.66
Sonar	*Er*/%	31.8±8.4	22.6±5.1	22.2±7.1	**17.8±7.6**
	G-mean/%	66.9±9.1	76.2±5.6	70.3±12	**79.8±10**
	Tc	5.65±1.4	3.4±0.82	2.7±0.72	**2.3±0.89**
Wine	*Er*/%	2.81±2.8	2.25±2.1	**1.14±1.7**	5.1±4.4
	G-mean/%	96.4±4.3	97.8±2.2	**98.2±2.6**	91.5±7.5
	Tc	0.35±0.33	0.35±0.33	**0.1±0.15**	0.45±0.39
Thyriod	*Er*/%	8.31±0.05	9.26±0.03	15.7±0.05	**4.57±0.03**
	G-mean/%	86±0.07	83.5±0.07	65.4±0.18	**95.00±0.03**
	Tc	1±0.63	1.1±0.39	1.7±0.56	**0.8±0.77**
Vehicle	*Er*/%	24.3±4.3	23.9±2.1	**16.8±2.2**	17.9±2.2
	G-mean/%	75.2±4.4	75.2±2.1	**82.1±2.1**	81.5±2.6
	Tc	13.7±3.1	12.4±1.5	**7.6±1.2**	9.3±1.2
Page-blocks	*Er*/%	9.78±0.9	7.02±0.5	6.48±0.5	**5.77±0.4**
	G-mean/%	65.6±3.4	**70.2±2.3**	68.9±3.4	69.7±2.4
	Tc	38.2±4.7	24.6±2.6	20.7±2.0	**17.3±1.6**
Sat	*Er*/%	12.0±1.0	11.9±1.0	8.69±0.4	**6.68±0.3**
	G-mean/%	87.9±1.0	88.2±1.0	91.0±0.4	**93.4±0.3**
	Tc	59.8±5.2	61.1±5.7	40.5±1.1	**35.1±2.1**
Shuttle	*Er*/%	10.6±0.3	5.26±0.5	4.31±0.1	**0.21±0.1**
	G-mean/%	77.4±1.2	87.3±1.6	92.7±0.4	**99.7±0.1**
	Tc	95.7±3.8	41.0±4.3	33.4±1.3	**2.2±0.7**
Average *Er*/%/Rank		12.9/6	10.7/4.3	10.3/4.5	**8.13/2.8**
Average *G-mean*/%/Rank		81.7/5.4	84.6/3.8	82.2/5	**88.1/2.9**
Average *Tc*/Rank		22.0/5.8	14.8/4.6	11.1/3.8	**7.23/3.1**

表 4 $c_{+-}=2$ 时的 UCI 数据集分类结果

Table 4 Classification results ofUCI dataset when $c_{+-}=2$

Dataset	Metrics	MRNB	CSBN	CSSVM	CSSVDD
Iris	*Er*/%	10.6±4.5	9.33±4.6	**3.33±3.3**	4.67±3.2
	G-mean/%	90.7±4.4	91.7±5.2	**96.9±3.4**	95.8±2.1
	Tc	1.8±0.8	1.6±0.9	**0.6±0.7**	0.8±0.5
Breast-w	*Er*/%	4.01±1.2	**2.57±1.3**	4.72±2.1	3.43±1.7
	G-mean/%	96.3±1.3	**97.7±1.5**	94.8±2.8	96.2±1.8
	Tc	3.4±1.3	**2.1±1.2**	4.8±2.6	3.6±1.9
Glass	*Er*/%	17.3±7.1	16.9±6.5	16.4±6.2	**14.5±4.2**
	G-mean/%	74.8±8.9	76.3±8.9	77.4±6.7	**80.9±6.5**
	Tc	6.3±2.2	6.0±2.1	5.8±1.7	**5.0±1.6**
Sonar	*Er*/%	32.7±5.6	24.6±6.2	**14.0±6.5**	17.8±5.8
	G-mean/%	66.0±5.1	75.0±6.5	**85.6±6.2**	80.6±5.8
	Tc	8.4±1.7	7.5±1.7	**4.5±1.8**	6.4±2.0
Wine	*Er*/%	2.78±2.1	2.81±2.1	**1.56±1.2**	5.76±5
	G-mean/%	96.9±3.1	96.9±3.3	**98.5±1.3**	94.7±8.5
	Tc	0.7±0.6	0.7±0.6	**0.4±0.2**	1.4±1.7
Thyriod	*Er*/%	6.97±4.4	8.38±3.4	9.26±3.4	**5.15±4.1**
	G-mean/%	88.3±8.1	86.2±7.2	82.9±6.1	**95.0±3.5**
	Tc	2.8±1.9	3.3±1.5	4.0±1.5	**1.4±1.1**
Vehicle	*Er*/%	24.6±3.2	23.0±3.1	**13.0±2.4**	17.8±2.4
	G-mean/%	75.3±3.1	76.2±3.3	**86.5±2.3**	81.7±2.1
	Tc	32.2±4.0	33.8±4.9	**14.0±2.0**	26.4±3.8
Page-blocks	*Er*/%	10.4±0.6	7.03±0.6	6.52±0.4	**5.87±0.5**
	G-mean/%	66.3±3.2	**79.3±3.1**	70.8±2.1	69.4±3.8
	Tc	86.8±4.7	**57.7±5.6**	67.1±4.4	60.8±5.8

（续表）

Dataset	Metrics	MRNB	CSBN	CSSVM	CSSVDD
Sat	*Er*/%	12.1±0.9	11.8±0.9	8.65±0.7	**6.82±0.8**
	G-mean/%	87.9±0.9	88.4±0.9	90.1±0.7	**93.3±0.8**
	Tc	111.3±9.4	103.7±9.2	75.3±8.3	**60±8.9**
Shuttle	*Er*/%	10.8±0.3	4.04±0.3	2.51±0.1	**0.21±0.0**
	G-mean/%	78.7±0.3	90.8±0.3	94.0±0.1	**99.7±0.0**
	Tc	263±8.9	110±8.0	59.1±5.1	**4.5±1.4**
Average *Er*/%/Rank		13.2/5.8	11.1/4.6	**8.0**/3.6	8.22/**3.4**
Average *G-mean*/%/Rank		82.1/5.8	85.9/4.5	87.8/**3.7**	**88.7**/3.9
Average *Tc*/Rank		51.6/5.8	32.6/4.7	23.6/3.8	**17.0**/**3.7**

结束语 本文基于 SVDD 对单类分类具有较好学习能力，以及 SVDD 可以通过调整分类器参数的方法调节对单类别的泛化能力的特点，通过重新定义 SVDD 的目标函数和约束条件，将单类 SVDD 拓展为二类分类 SVDD；进一步，为降低代价敏感分类的误分代价，基于应该对误分代价高的类别进行更为精确的学习的思想，提出一种代价敏感 SVDD 二类方法 CS-SVDD。首先定义了基于误分代价的代价敏感二类 SVDD 优化问题的目标函数和约束条件，引入代价因子用以调节误分代价对 SVDD 惩罚因子的影响程度，使误分代价高的类别的样本尽可能包含在超球体内，从而降低误分代价；对于处于两个超球体之外或覆盖区域的类别属性不明确的样本，以误分代价最小为原则定义了代价敏感决策规则。基于人工数据集和 10 类 UCI 数据集的实验表明，提出的方法能够有效提升分类性能，降低分类代价。

参 考 文 献

[1] HABBEMA J D F, HERMANS J. Cases of doubt in allocation problems[J]. Biometrika, 1974, 61(2): 313-324.

[2] ELKAN C. The Foundations of Cost-Sensitive Learning[C]// Proceedings of Proceedings of 17th International Joint Conference on Artificial Intelligence. 2001: 973-978.

[3] DHAR S, CHERKASSKY V. Development and evaluation of cost-sensitive universum-SVM[J]. IEEE Transactions on Cybernetics, 2015, 45(4): 806-818.

[4] YUAN G L, SUN Z W, QIN X Y, et al. Object Tracking Based on Cost Sensitive Structured SVM[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3335-3341.

[5] 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.

[6] IBÁÑEZ A, BIELZA C, LARRAÑAGA P. Cost-sensitive selective naïve Bayes classifiers for predicting the increase of the h-index for scientific journals[J]. Neurocomputing, 2014, 135: 42-52.

[7] JIANG L, LI C, WANG S. Cost-sensitive Bayesian network classifiers[J]. Pattern Recognition Letters, 2014, 45: 211-216.

[8] FREITAS A, COSTA-PEREIRA A, BRAZDIL P. Cost-sensitive decision trees applied to medical data[C]// Data Warehousing and Knowledge Discovery. Springer, 2007: 303-312.

[9] LI X J, ZHAO H, ZHU W. A cost sensitive decision tree algorithm with two adaptive mechanisms[J]. Knowledge-Based Systems, 2015, 88: 24-33.

[10] XIONG B Y, WANG G Y, DENG W B. Hierarchical Cost Sensitive Decision Tree and Its Application in Mobile Phone Replacement Prediction[J]. Journal of Shandong University, 2015, 45(5): 36-42.

[11] CHEN Y L, WUB C C, TANG K. Time-constrained cost-sensitive decision tree induction[J]. Information Sciences, 2016, 354: 140-152.

[12] SUN Y M. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40: 3358-3378.

[13] XUE Y Z, WANG T. Target tracking based on cost Sensitive Adaboost[J]. Chinese Journal of Image and Graphics, 2016, 21(5): 544-555.

[14] CHENG F Y, ZHANG J, WEN C H. Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data[J]. Pattern Recognition Letters, 2016, 80: 107-112.

[15] CHENG F Y, ZHANG J. Large cost-sensitive margin distribution machine for imbalanced data classification[J]. Neurocomputing, 2017, 224: 45-57.

[16] TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20: 1191-1199.

[17] TAX D M J, DUIN R P W. Support vector domain description[J]. Machine Learning, 2004, 54: 45-66.

[18] LEE D, LEE J. Domain described support vector classifier for multi-classification problems[J]. Pattern Recognition, 2007, 40(1): 41-51.

[19] MU T T, NANDI A K. Multiclass classification based on extended support vector data description[J]. IEEE Trans on Systems, Man, and Cybernetics—Part B: Cybernetics, 2009, 39(5): 1206-1216.

[20] ZHENG S F. Smoothly approximated support vector domain description[J]. Pattern Recognition, 2016, 49: 55-64.

[21] GUO Y, XIAO H, FU Q. Least square support vector data description for HRRP-based radar target recognition[J]. Applied Intelligence, 2017, 46: 365-372.

[22] WANG C D, LAI J H. Position regularized Support Vector Domain Description[J]. Pattern Recognition, 2013, 46: 875-884.

[23] ElAZAMI M. Converting SVDD scores into probability estimates: Application to outlier detection[J/OL]. http://dx.doi.org/10.1016/j.neucom.2017.01.103.

[24] DUAN L X, XIE M Y. A new support vector data description method for machinery fault diagnosis with unbalanced datasets[J]. Expert Systems With Applications, 2016, 64: 239-246.

[25] ZHOU Y M, WU K. Fault detection of aircraft based on support vector domain description[J]. Computers and Electrical Engineering, 2017, 61: 80-94.

[26] HUANG G, CHEN H, ZHOU Z, et al. Two-class support vector data description[J]. Pattern Recognition, 2011, 44: 320-329.

WU Chongming, born in 1966, Ph.D, associate professor. His main research interests include machine learning and intelligent information processing.

WANG Xiaodan, born in 1966, Ph.D, professor. Her main research interests include machine learning and pattern recognition.