保局性数据域描述单类分类器

郑建炜 蒋一波 王万良
(浙江工业大学计算机学院 杭州 310023)

摘 要 由于缺少对数据结构信息的考虑,现有的域描述型单类分类器得到的支撑面往往是次优解。因此,以支持向量数据描述(SVDD)算法为基础,通过一种简易的形式引入数据亲和因子以保持样本局部特性,提出保局性数据域描述分类器(LPDD),使成簇的数据作用被强化,而呈零星分布的数据影响力被削弱,引导分类支撑面自动靠近数据高密区而提高算法性能。此外,为适应大样本应用场合,采用序列最小优化算法进行模型参数调整。实验证明,所提算法无论在训练速率还是在分类性能上都优于SVDD。

关键词 亲和因子,支持向量域描述,序列最小优化,单类分类器
中图法分类号 TP393.08 **文献标识码** A

Locality Preserving Data Domain Description One-class-classifier

ZHENG Jian-wei JIANG Yi-bo WANG Wan-liang
(School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract In a support vector data description (SVDD), the compact description of target data was given in a hyper spherical model which was determined by a small portion of data called support vectors. Despite the usefulness of the conventional SVDD, however, it may not identify the optimal solution of target description due to neglecting the structure of the given data. In order to mitigate this problem, a novel one-class-classifier named locality preserving data domain description (LPDD) was proposed which takes the data density into account by using of affine factor. Besides, the sequential minimal optimization was adopted to adjust model parameters for applying in the large sample occasions. Experiments with various real data sets show promising results.

Keywords Affine factor, Support vector domain description, Sequential minimal optimization, One-class-classifier

1 引言

在故障诊断、入侵检测、疾病判定、身份认证等众多模式识别应用领域中,分类器的设计过程往往只能得到正类(目标类)样本,而负类(异常类)数据由于较为稀少或采样成本过高,不得不放弃获取,由此产生了所谓的单类分类器[1]。

典型的单类分类器可分为基于密度估计和基于支撑边界两种类型。其中,基于密度估计的方法,如高斯混合模型[2],首先估计某类训练数据的概率密度函数,然后通过计算测试样本属于该类数据的概率值来进行判别。如高于设定阈值,则将其判为目标样本,反之拒绝。它的缺点是需要大量训练样本,而且为了解决一个判别正负的二值问题而去计算更复杂的概率密度函数,得不偿失。基于支撑域边界的方法是指用一个最小超球或最贴近正样本的超平面将目标数据包络为一个封闭的超球体或者正半空间,而负类数据尽可能远地隔离在区域之外,从而达到识别错误率最小的目的。这个超球体或超平面是由不为零的支撑量张成的,故称为支撑域。

支撑域方法是支持向量机[2](Support Vector Machines, SVM)在无监督学习领域的扩展,继承了SVM的全部优点:大间隔、稀疏性、核映射、全局最优,而且算法运行时实时度较高,适合高维、含噪、有限训练样本的应用场合。典型的支撑域方法中,单类支持向量机[3](One-Class SVM, OCSVM)是超平面支撑域的代表模型,采用原点作异常点,隐含了负类数据的位置信息,对于有些实际数据来说并不合理;超球模型的典型算法——支持向量数据描述[1](Support Vector Data Description, SVDD)由于对数据本身分布信息缺乏考虑,因此在处理各向异构的数据时,往往只能得到次优边界。

基于已有算法的缺陷,Tsang等从单类最小最大概率机[4](OCMPM)思想中得到启发,用马氏距离取代OCSVM中欧氏距离,设计出MOCSVM,并将其引入到核学习中,获得了更优的支撑域[5];Dolia等[6]以超椭球替代SVDD中的超球,使得包络边界更加灵活,能适应更多种分布结构的数据类型;Tingting等[7]为获得泛化能力更好的描述边界,在模型训练过程中引入负类样本,并且将算法应用到多分类场合。上述算法虽然致力于获取更加合理的支撑边界,却都没有考虑目标数据本身的分布特征,而实际上数据内部的分簇结构以及数据在不同区域的松散度等先验信息都能更好地指导支撑域的形成。KiYoung等[8]先对输入样本进行K近邻分簇,再

到稿日期:2010-12-17 返修日期:2011-04-16 本文受国家自然科学基金(61070043)资助。
郑建炜(1982—),男,博士,主要研究方向为模式识别、人工智能,E-mail:zjw@zjut.edu.cn;蒋一波(1982—),男,博士,主要研究方向为网络控制多智能体;王万良(1957—),男,教授,博士生导师,主要研究方向为人工智能网络控制。

据此对每个样本点引入局部密度因子并将之应用到模型构建中；类似地，冯爱民等[9]也对目标数据进行分簇，并将各簇协方差矩阵引入最优支撑域获取中。两者都通过数据本身分布特点去寻找最优的包络边界，然而在寻优过程中都需要借助其它算法进行先验信息的获取，这增加了算法计算量。此外，两种算法都引入了新的经验参数，需要通过网格式搜索获取或者依经验指定，这增加了模型构建复杂度。

SVDD 的目标泛函是一个凸二次规划问题，保证了模型具有全局最优解，然而其常规的求解过程却需要 $O(n^3)$ 的时间复杂度以及 $O(n^2)$ 的空间复杂度。因此，SVDD 在应用于大样本场合时，会遇到运算量过大和内存溢出问题。S. Calvin 等[10]采用核心集的方式解决这个问题，但算法的初值选择过程比较复杂；陆从德等[11]引入乘性规则进行模型训练，虽然速度得到了大幅度提升，但结果却不再是最优解；赵英刚等[12]提出的约减型快速训练算法中需要用到核矩阵逆操作，虽然在小样本时效率极高，但随着输入数据的增多，其性能会急剧下降。

本文旨在提高超球体模型的分类性能，在 SVDD 算法基础上，将数据分布松散度以亲和因子的形式融入原算法中的最大间隔寻优公式，提出保局性数据域描述单类分类器(Locality Preserving Data Domain Description，LPDD)。LPDD 保持原算法框架不变，亲和因子的计算也无需借助其它算法。为适应大样本任务，把在 SVM 和 OCSVM 中得到成功应用的序列最小优化算法(Sequential Minimal Optimization，SMO)引入 LPDD 的模型构建中。SMO 算法初值选择简单合理，运算量远小于二次规划求解方法，且仍能保证全局最优解。

2 支持向量数据描述

给定目标数据集 $X=\{\bar{x}_i\}_{i=1}^n$，$\bar{x}_i\in R^p$，其中 n 代表样本数目，p 为样本维度。通过一个非线性映射 ϕ 将样本映射到特征空间 F 中，即 $\phi:\bar{x}_i\in X\subset R^p\rightarrow\phi(\bar{x}_i)\in F$。SVDD 的目标是在空间 F 中寻找一个超球体(a,R)，其中 a 为球体中心，R 为球体半径。超球必须满足半径 R 尽可能小的同时，尽可能多地包含目标数据。同时为适应稀疏偏远点的部分样本，引入非负松弛变量 ξ_i，$i=1,2,\cdots,n$，即目标问题规划如下：

$$\min(R^2+\frac{1}{vn}\sum_{i=1}^{n}\xi_i) \tag{1}$$

$$s.t.\ \|\bar{x}_i-a\|^2\leqslant R^2+\xi_i,i=1,\cdots,n \tag{2}$$

式中，$v\in(0,1]$是全部支撑向量的下界，用来控制超球半径与它所能包围的样本数目之间的折衷。v 越小，允许超球外面存在样本的约束程度就越大。该优化问题的解可由相应拉格朗日泛函的鞍点给出：

$$L(R,a,\xi_i,\alpha_i,\lambda_i)=R^2+\frac{1}{vn}\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(R^2+\lambda_i-\|\bar{x}_i-a\|^2)-\sum_{i=1}^{n}\lambda_i\xi_i \tag{3}$$

式中，$\alpha_i\geqslant 0$，$\lambda_i\geqslant 0$ 为拉格朗日系数。求解式(3)的最小值，可转化为求其对偶问题的最大值：

$$W(\alpha)=\max\{\sum_{i=1}^{n}\alpha_i k(\bar{x}_i,\bar{x}_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\bar{x}_i,\bar{x}_j)\} \tag{4}$$

$$s.t.\sum_{i=1}^{n}\alpha_i=1,0\leqslant\alpha_i\leqslant\frac{1}{vn} \tag{5}$$

式中，$k(\bar{x}_i,\bar{x}_j)$代表核函数，用以替代内积运算，即 $k(\bar{x}_i,\bar{x}_j)=\langle\phi(\bar{x}_i),\phi(\bar{x}_j)\rangle$，本文简写为 k_{ij} 并且只采用高斯核，即：

$$k_{ij}=\exp(-\frac{\|\bar{x}_i-\bar{x}_j\|^2}{\sigma^2}) \tag{6}$$

记式(4)二次规划问题的最优解为 $\alpha^*=\{\alpha_1^*,\alpha_2^*,\cdots,\alpha_n^*\}$，其中对应 $\alpha_i^*=0$ 的样本位于超球内部，称为内点；对应 $0<\alpha_i^*<\frac{1}{vn}$的样本落在超球面上，称作支持向量或边界点；余下对应 $\alpha_i^*=\frac{1}{vn}$的样本位于超球外部，称为野点或奇异点。超球球心 a 可由非零 α_i^* 线性组合计算出：

$$a=\sum_{i=1}^{n}\alpha_i^*\bar{x}_i \tag{7}$$

任选一个 $\alpha_i^*\in(0,\frac{1}{vn})$可求得超球半径 $R^2=\|\bar{x}_i-a\|^2$。对于测试样本 z，如果其距超球球心的距离小于半径 R，则接受该样本，否则拒绝。接受条件表达为：

$$\|\phi(z)-a\|^2=k(z,z)-2\sum_{i=1}^{n}\alpha_i^*k(z,\bar{x}_i)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i^*\alpha_j^*k_{ij}\leqslant R^2 \tag{8}$$

3 保局性数据域描述分类器

虽然 SVDD 的目标是为数据描述最精细的球体边界，但它却并没有考虑样本的分布情况，对于不呈简单团簇结构的数据集，往往无法找到最佳的支撑边界。事实上，如果单类分类器只是简单地将全部训练数据同等对待，则会导致球体为包括部分稀疏偏远点而包络过大，使得负类样本进入球体内部，影响分类器的推广性能。因此，如果通过数据分布状况引导支撑边界的形成，使之向高密度数据区域靠拢，必然会提升分类器性能。

3.1 亲和因子

局部特性可以通过密度因子的形式表达，先对训练样本进行分簇聚类，再对各簇计算相应的密度因子并作用于模型构建[13]。实际上，数据局部分布状况的反映并不需要如此复杂的操作，任意两两样本之间的距离即可在一定程度上体现其分布结构。局部保持投影算法[14](Locality Preserving Projection，LPP)即采用了这个基本思想。LPP 将样本间的亲和力引入最佳投影方向的搜寻中，使投影后的数据能最大程度地保持其原来的局部分布特性。其中，样本间的亲和力又称为亲和因子，一般计算方法如下：

$$A_{ij}=\exp(-\frac{\|\bar{x}_i-\bar{x}_j\|^2}{\sigma^2}) \tag{9}$$

式中，$\sigma>0$ 是可调整参数，需要多次验证确定取值。为了避免对 σ 进行网格式定值搜索带来的巨大计算负担，也为了使元素为 A_{ij} 的亲和矩阵 A 具有稀疏性，仅对样本中的近邻点指定相应的亲和因子，而非近邻点的亲和力值为零。如样本 $\bar{x}_i$，先在样本空间中以欧式距离搜寻其最近邻的 k 个样本，记第 k 近邻距离为 N_i^k，假如另一样本 $\bar{x}_j$ 与 $\bar{x}_i$ 的距离小于 N_i^k，则认为 $\bar{x}_j$ 属于 $\bar{x}_i$ 的近邻，取亲和因子如下：

$$A_{ij}=\exp(-\frac{\|\bar{x}_i-\bar{x}_j\|^2}{N_i^k\times N_j^k}) \tag{10}$$

而其余 A_{ij} 设为零。一般取 k 值为 7 能获得较优的结果[15]。

3.2 保局性数据域描述

得到亲和因子后，即可将之引入 SVDD 的目标泛函中，

起到凸显近邻样本点，抑制离群样本点的作用。式(4)调整为

$$W(\alpha)=\sum_{i=1}^{n}\alpha_i k_{ii}-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j A_{ij}k_{ij} \tag{11}$$

而约束条件和判决公式都保持不变。

分析上式可知，引入亲和因子后仅仅变换了目标泛函表达式的核函数数值。再从高斯核函数表达式(6)可以看出，当亲和因子取式(9)时，两者形式完全一致。因此，在这种选择下，目标函数可直接表达为：

$$W(\alpha)=\sum_{i=1}^{n}\alpha_i k_{ii}-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_{ij}^2 \tag{12}$$

将采用式(10)作为亲和因子的保局性数据域描述算法称为 LPDD1，采用式(9)则称为 LPDD2。无论采用哪种亲和因子的计算方法，由于目标泛函表达式基本框架保持不变，LPDD 必定保持着 SVDD 的所有优点：二次规划问题的全局最优性、解的稀疏性等。从核函数矩阵的角度看，由于亲和矩阵 A 本身具有对称性，因此高斯核矩阵的对称结构依然不变，主对角元素仍为 1，其它矩阵元素却依数据间亲和力值进行了修整。也就是说，在高维特征空间中，数据描述域仍然分布在一个超球体上，并且在数据局部特性的引导下构造了一个新核。这个新核包含了数据的先验知识，因此能够更好地帮助边界支撑域的形成，所以通过式(11)所得到的 α^* 是经局部结构信息作用后的超球体参数。在 A_{ij} 的作用下，训练数据中的各个分簇都能得到正确的对待，而零星散点则被视作野值，不予考虑，使得支撑域自动靠近了高密度区，更好地覆盖了正类数据所在的区域。对于新测样本，尽管所用的判别式和原算法相同，但由于支撑域变得更为精细，因此分类器的推广性能提高了。

3.3 SMO 训练算法

SMO 的基本流程是先对偶化原目标函数，如式(11)所示，再从条件式(5)出发确定其最优化条件，然后对违反该条件的拉格朗日乘子进行两两优化，直到收敛。

既然目标函数已经对偶化，那么就可以直接搜寻其最优条件。先写出式(11)的拉格朗日表达式：

$$L=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j A_{ij}k_{ij}-\sum_{i=1}^{n}\alpha_i-\beta\sum_{i=1}^{n}\alpha_i+\beta \tag{13}$$

将 L 对 α_i 微分，得到：

$$\frac{\partial L}{\partial\alpha_i}=2\sum_{j=1}^{n}\alpha_j A_{ij}k_{ij}-1-\beta=0 \tag{14}$$

设：

$$H_i=2\sum_{j=1}^{n}\alpha_j A_{ij}k_{ij} \tag{15}$$

则式(14)可简述为 $H_i=\beta+1$。定义：

$$i\_\text{up}=\arg\max_i H_i,\ i=1,2,\cdots,n \tag{16}$$

$$i\_\text{low}=\arg\min_i H_i,\ i=1,2,\cdots,n \tag{17}$$

最后可得 LPDD 的优化目标为：

$$H_{i\_\text{up}}=H_{i\_\text{low}}=\beta+1 \tag{18}$$

实际 SMO 更新过程中，先搜索满足 $H_{i\_\text{up}}\neq H_{i\_\text{low}}$ 的 α_{i_up} 和 α_{i_low}，并对它们进行调整。由于必须满足条件 $\sum_{i=1}^{n}\alpha_i=1$，因此调整公式可以描述为：

$$\alpha_{i\_\text{up}}'=\alpha_{i\_\text{up}}+t,\alpha_{i\_\text{low}}'=\alpha_{i\_\text{low}}-t$$
$$\alpha_k'=\alpha_k,\ \forall k\neq i\_\text{up},i\_\text{low} \tag{19}$$

又由于必须满足条件 $0\leqslant\alpha_i\leqslant\frac{1}{vn}$，因此上式调整过程中 t 应满足：

$$t\in[\max(-\alpha_{i\_\text{up}},-\frac{1}{vn}+\alpha_{i\_\text{low}}),\min(\frac{1}{vn}-\alpha_{i\_\text{up}},\alpha_{i\_\text{low}})] \tag{20}$$

为计算出 t 的具体取值，先改写式(11)为关于 t 的目标函数，假定 $i\_\text{up}=1$，$i\_\text{low}=2$，则：

$$W(t)=(\alpha_1+t)^2+2(\alpha_1+t)(\alpha_2-t)A_{12}k_{12}+(\alpha_2-t)^2+2(\alpha_1+t)\sum_{j=3}^{n}\alpha_j A_{1j}k_{1j}+2(\alpha_2-t)\sum_{j=3}^{n}\alpha_j A_{2j}k_{2j}-(\alpha_1+t)-(\alpha_2-t)+q_1-q_2 \tag{21}$$

式中，$q_1=\sum_{i=3}^{n}\sum_{j=3}^{n}\alpha_i\alpha_j A_{ij}k_{ij}$，$q_2=\sum_{i=3}^{n}\alpha_i$，这两者都不依赖于 t，可视为常数。

然后计算式(21)的一阶和二阶微分：

$$\frac{\partial W}{\partial t}=2(\alpha_1+t)+2(\alpha_2-\alpha_1-2t)A_{12}k_{12}-2(\alpha_2-t)+2\sum_{j=3}^{n}\alpha_j A_{1j}k_{1j}-2\sum_{j=3}^{n}\alpha_j A_{2j}k_{2j} \tag{22}$$

$$\frac{\partial^2 W}{\partial t^2}=4-4A_{12}k_{12}=\theta \tag{23}$$

当 $\bar{x}_1\neq\bar{x}_2$ 时，必定有 $\theta>0$，则 $W(t)$ 存在唯一极小值。令 $\frac{\partial W}{\partial t}=0$，并联合 H_i 的表达式，则由式(22)可得：

$$t=\frac{H_2-H_1}{\theta} \tag{24}$$

综上所述，保局性数据域描述单类分类器的训练过程如下：

Step 1　依条件式(5)初始化 α，本文选取 $\alpha_i=\frac{1}{n}$，并计算出相应的 H_i 值；

Step 2　设定 $numChang=0$，搜索所有的边界点，如果其中存在不同索引对 (i,i')，使得 $H_i\neq H_{i'}$，则依条件式(16)、式(17)选出相应的 i_up 以及 i_low；

Step 3　依式(19)调整系数 α，其中 t 依式(24)计算，如果违反了约束式(20)，则设定 t 为相应的边界值。如果调整后的 $\alpha_i=0$ 或者 $\alpha_i=\frac{1}{vn}$，则标定相应 $\bar{x}_i$ 为非边界点；

Step 4　更新所有的 H_i 值，继续 Step 2 操作，直到任意两个边界点满足 $H_i=H_{i'}$；

Step 5　依次选择各非边界点，记为 $\bar{x}_i$，选择使 $|H_i-H_j|$ 最大的点作为 $\bar{x}_{i'}$，依式(19)调整相应系数；如果调整后 $\alpha_i\in(1,\frac{1}{vn})$，则标定 $\bar{x}_i$ 为边界点，$numChang=numChang+1$；

Step 6　回到 Step 2 继续，当操作至 Step5 时所有点都满足 $H_i=H_{i'}$ 且 $numChang=0$，则迭代结束。

具体实现过程中有两个注意点：

1) 条件 $H_i=H_{i'}$ 不能严格满足，因此必须给出相对宽松的条件，本文以 $H_{\max}-H_{\min}<=1e-6$ 替代。

2) H_i 函数是迭代优化的对象，且每次 α 更新过程都需要用到，因此必须进行及时的更新保存，并应用于下一次迭代操作，具体更新表达式为：

$$H_i^{\text{new}}=H_i^{\text{old}}+2t-2tA_{ii'}k_{ii'} \tag{25}$$

$$H_{i'}^{\text{new}}=H_{i'}^{\text{old}}+2tA_{i'i}k_{i'i}-2t \tag{26}$$

$$H_k^{\text{new}}=H_k^{\text{old}}+2tA_{ki}k_{ki}-2tA_{ki'}k_{ki'} \tag{27}$$

式中，$i=i\_\text{up}$，$i'=i\_\text{low}$，$k\neq i,i'$。

3.4 计算复杂度分析

SMO 训练过程都是元素操作，因此无需存储核矩阵。亲和因子只与核元素相关，不占用额外空间量。而本文为方便后续计算，仍然保存包含亲和因子的核矩阵，其空间复杂度为 $O(n^2)$。算法迭代时需要保存 H_i 值，其空间复杂度为 $O(n)$。因此，总体空间复杂度仍然是 $O(n^2)$。

在时间复杂度方面，对于 LPDD1 来说，初值 H_i 需要计算 n 次亲和因子和核函数。其中亲和因子需耗费复杂度为 $O(nk-k^2)$的时间，核函数则为 $O(p)$，总计时间复杂度为 $O(n^2k-nk^2+np)$。之后 SMO 需要遍历所有的边界点和非边界点（总数为 n），用以搜索两个系数 α_{i_up} 和 α_{i_low} 并进行调整，每次调整都需要更新 n 个 H_i 值，因此时间复杂度为 $O(n^2)$。由于本文选取 $k=7$，若优化过程共进行 l 次迭代才使算法收敛，则总的时间复杂度为 $O((7+l)n^2-(49-p)n)$。对于 LPDD2 来说，其算法流程与 LPDD1 完全一致，且亲和因子的计算完全融于核函数中，因此总时间复杂度为 $O(n^2l+np)$。

4 实验分析

为验证所提保局性数据域描述分类器在输入样本呈不同分布情形下的支撑域形成能力和识别性能，实验分别采用人造数值与公共数据集进行直观的支撑域描述显示以及不同算法间的识别性能对比。高斯核带宽 σ 经网格式搜索、5-fold 交叉验证过程获取，目标函数中的 $v=0.1$。所有实验结果都在 CPU 主频为 2.6GHz、内存 4G 的台式机中产生，软件环境为 Windows XP 操作系统，Matlab7.1 编译平台并内嵌 SVDD 工具箱 DD_Tools[16]。

4.1 支撑域描述性能对比

为了直观了解数据的空间结构对不同域描述算法性能的影响，本文先通过 DD_Tools 构造 4 组不同分布的 2 维数据集，分别采用 SVDD 及 LPDD1 描述它们的边界。4 组数据集分别为：1）SimpleData：服从均值为 0、方差为 1 的高斯分布；2）BananaData：香蕉形区域内，方差为 1 的高斯分布与均匀分布叠加；3）LongData：均值变化、方差为 1 的高斯分布并且旋转 45°；4）LithuanianData：两瓣香肠形区域内，方差为 1 的高斯分布与均匀分布叠加。图 1 为上述 4 组数据点分布图以及相应两种算法的描述包络线。

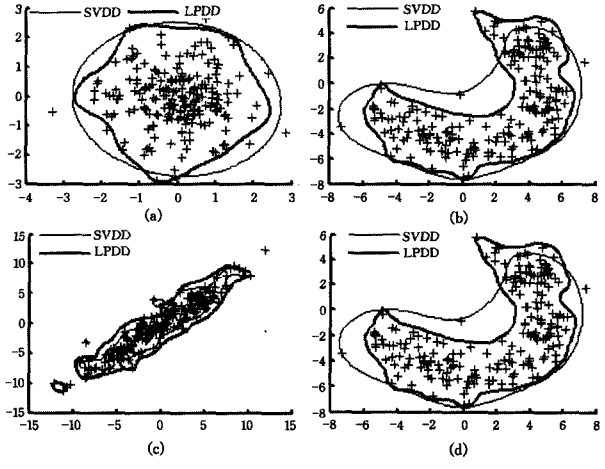

图 1 SVDD/LPDD 数据描述边界对比显示

图 1(a)为常规团簇结构的数据，周围零星分布点数据密度明显偏低，因此保局性数据域描述算法（LPDD1）的包络边界向内收进，更能反映数据的本质结构。而支持向量数据域描述（SVDD）的包络相对偏大，容易将野点归入类内数据；图 1(b)中 SVDD 为了几个散点，致使边界线在左侧过于庞大，而上侧真实类内数据却被划于包络之外；图 1(c)反映了 SVDD 为尽量细致地描述完整斜长形数据而导致过学习。图 1(d)上下两部分香肠形数据中，SVDD 对上部边界线描述过大，下部却又偏紧，中间交叉部分数据还被隔在包络之外。综合全图，经典支持向量域描述算法虽然能大致概括不同分布结构的数据，但结果往往是次优的。而保局性数据域描述算法在引入亲和因子后，能更好地获取不同数据的分布特征，对 4 组结构各异的数据集都给出了较支持向量数据描述算法更优的边界包络。

4.2 公共数据集识别

以上仅仅是 SVDD 与 LPDD 在二维简单数据上的支撑域描述能力对比，接下来以公共数据集为对象，对两者实际识别能力进行比较测试。实验采用由 David M J Tax 整理所得的 UCI 公共数据库中的 8 个数据集[17]作为算法性能验证对象，各数据集都已经由 David 进行预处理并正确标定正负类样本。单类问题的识别结果有两类错误，即目标类被错分为异常类 FN 和异常类被错分为目标类 FP，显然这两个值越小说明算法性能越好。这里比较了 SVDD 和 LPDD1 以及 LPDD2 算法，表 1 列出了以传统二次规划问题求解为训练方法的识别结果及训练用时（Tm：单位为秒）。表 2 则是以 SMO 算法进行训练的相应数据值，其中黑体为相同数据下 3 种算法的最佳结果。数据集括号内的 Dim：Tr：TeT：TeO 依次表示该数据集的维度、训练样本、目标类和异常类测试样本。

联合表 1 和表 2 进行分析，可看出：

1）在对不同数据集测试过程中，除 Breast 外，其它数据集中由 LPDD 算法得到的 FN/FP 指标都优于 SVDD 算法。而在 Breast 数据中，虽然 SVDD 的 FP 值略低于 LPDD 算法，但代价是在 FN 指标中 SVDD 远远高于 LPDD，因此仅从这一项例外不能否认所提保局性数据域描述算法的高效性。在两种算法理论框架、训练算法都一致的情况下，仅对 LPDD 算法加入了对训练样本间亲和力的考虑，可见亲和因子确实能够反映数据分布结构，并提升算法支撑域描述性能。

2）在训练速度方面，SMO 远远快于二次规划求解算法，其中 Diabetes 数据集上，SMO 在 3 种模型中的速度提升分别达 216、37 和 53 倍有余。说明两种方法确实不在同一级别。引入 SMO 后，LPDD 能够胜任大样本应用场合。

3）进一步对比 LPDD1 和 LPDD2。在基于二次规划问题的求解算法中（表 1），两者识别率各有胜负，其中 FN 指标 LPDD1 占优，而 FP 指标则是 LPDD2 略胜。训练速度方面，则是 LPDD1 明显更优，其中 Breast 数据集中 LPDD1 比 LPDD2 提升了 2.76 倍。究其原因，是 LPDD1 采用了稀疏的亲和矩阵，加速了后续的矩阵运算。在 SMO 训练算法中（表 2），两者 FP 错误率基本持平，而 FN 指标则是 LPDD2 较为优越。并且在训练速度上 LPDD2 也全面快于 LPDD1，其主要原因还是亲和矩阵的稀疏性。过多的零值影响了 LPDD1 的收敛速度，也使最优条件变得模糊。

综上所述，由于实际应用中输入数据内部分布结构往往比较复杂，引入亲和因子后的 LPDD 能更多地保留局部特性，因此识别性能也更强。在具体的 A_{ij} 选取上，如果是小样本应用情况，则采用二次规划训练法的 LPDD1；一旦碰到大数据量场合，那么 LPDD2 应该是首选策略，且训练算法必须采取 SMO。

表 1　SVDD 和 LPDD 算法在 UCI 数据集上的 FN/FP(采用凸二次规划求解)

Data Set	SVDD			LPDD1			LPDD2		
Dim:Tr:Tet:Teo	FN	FP	Tm	FN	FP	Time	FN	FP	Tm
Breast(9:217:241:458)	0.1494	**0.0218**	5.42	**0.0373**	0.0349	**1.80**	0.0498	0.0284	6.77
Heart(13:148:164:139)	0.1037	0	1.28	**0.0976**	0	**0.89**	**0.0976**	0	0.92
Sonar(60:100:111:97)	0.3153	0.0928	0.81	**0.0991**	**0.0412**	**0.57**	**0.0991**	**0.0412**	0.64
Arrhythmia(278:214:237:183)	0.1603	0.1585	3.94	0.0541	**0.0546**	**1.55**	**0.0506**	0.0601	1.56
Diabetes(8:450:500:268)	0.2960	0.4179	49.83	**0.0920**	0.0534	**10.56**	0.1040	**0.0522**	10.73
Spectf(44:229:254:95)	0.3189	0.3200	**10.2**	**0**	0.0316	11.37	0.0591	**0.0211**	11.83
Waveform(21:270:300:600)	0.2600	0.1183	9.63	**0.05**	**0.0833**	2.94	**0.05**	**0.0833**	**2.86**
Survival(3:203:225:81)	0.2400	**0.0741**	8.47	**0.1644**	**0.0741**	**8.45**	0.1733	**0.0741**	9.28

表 2　SVDD 和 LPDD 算法在 UCI 数据集上的 FN/FP(采用 SMO 求解)

Data Set	SVDD			LPDD1			LPDD2		
Dim:Tr:Tet:Teo	FN	FP	Tm	FN	FP	Time	FN	FP	Tm
Breast(9:217:241:458)	0.2158	**0.0328**	0.16	**0.0373**	0.0349	0.16	**0.0373**	0.0502	**0.14**
Heart(13:148:164:139)	0.25	0.0288	0.16	0.1585	**0**	0.13	**0.1524**	**0**	**0.11**
Sonar(60:100:111:97)	0.2072	0.0414	0.13	**0.0991**	0.0412	0.14	0.1081	**0.0411**	**0.11**
Arrhythmia(278:214:237:183)	0.1688	0.0328	0.16	0.0591	**0.0318**	0.19	**0.0422**	0.1257	**0.17**
Diabetes(8:450:500:268)	0.1920	0.1978	0.23	0.1	**0.05**	0.28	**0.0920**	0.0933	**0.2**
Spectf(44:229:254:95)	0.1063	0.1158	0.14	**0.0748**	0.0316	0.16	**0.0748**	**0.0211**	**0.14**
Waveform(21:270:300:600)	0.2600	0.09	0.17	**0.05**	**0.0833**	0.2	**0.05**	**0.0833**	**0.14**
Survival(3:203:225:81)	0.2089	0.0988	0.14	**0.1689**	**0.0741**	0.14	**0.1689**	**0.0741**	**0.13**

结束语　尽可能利用先验知识，是提高分类器推广性能的关键所在。本文提出的保局性数据域描述单类分类器(LPDD)，便是在现有 SVDD 框架下嵌入先验信息的单类算法。为了能够更有效地处理不同分布类型的数据，算法在不对数据进行分簇聚类的情况下，引入反映数据局部状态的亲和因子，将此先验信息作为类内紧性嵌入到经典的 SVDD 框架中，使得算法能够获得更具代表性的数据支撑域。并且，为了适合大数据量应用，本文将 SMO 引入 LPDD 的训练算法中，使模型构建速度大幅度提升，也使 LPDD 在实际应用中更为灵活。

参 考 文 献

[1] Tax D M J, Duin R P W. Support Vector Data Description[J]. Machine Learning, 2004, 54(1): 45-66

[2] You C H, Lee K A, Li H Z. GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(6): 1300-1312

[3] Schölkopf B, Platt J C. Estimating the Support of a High-dimensional Distribution[J]. Neural Computation, 2001, 13(7): 1443-1471

[4] Lanckriet G R G, Ghaoui L E, Jordan M. Robust novelty detection with single-class MPM[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 905-912

[5] Tsang I W, James T K, Li S. Learning the kernel in Mahalanobis one-class support vector machines[C]// Proceedings of the International Joint Conference on Neural Networks. Vancouver, Canada, 2006: 1169-1175

[6] Dolia A, Harris C, Shawe-Taylor J. Kernel ellipsoidal trimming [J]. Computational Statistics and Data Analysis, 2007, 52(1): 309-324

[7] Mu Tintin, Nandi A K. Multiclass classification based on extended support vector data description[J]. IEEE Transaction on System, Man and Cybernetics-Part B: Cybernetics, 2009, 39(5): 1206-1216

[8] Lee K Y, Kim D W, Lee K H, et al. Density-induced support vector data description[J]. IEEE Transaction on Neural Networks, 2007, 18(1): 284-289

[9] 冯爱民，薛晖，刘学军，等. 增强型单类支持向量机[J]. 计算机研究与发展，2008，45(11)：1858-1864

[10] Chu C S, Tsang I W, Kwok J T. Scaling Up Support Vector Data Description by Using Core-sets[C]// Proceedings of IEEE International Joint Conference on Neural Networks. Budapest, Hungary, 2004: 425-430

[11] 陆从德，张太镒，胡金燕. 基于乘性规则的支持向量域分类器[J]. 计算机学报，2004，27(5)：690-694

[12] 赵英刚，陈奇，何钦铭. 基于支持向量域数据描述的快速学习算法[J]. 仪器仪表学报，2006，27(6)：798-800

[13] 冯爱民，陈斌. 基于局部密度的单类分类器 LP 改进算法[J]. 南京航空航天大学学报，2006，38(6)：727-731

[14] 王文俊，张军英. 基于核的类别非局保留投影[J]. 模式识别与人工智能，2009，22(5)：769-773

[15] Zelnik-Manor L, Perona P. Self-tuning spectral clustering[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 1601-1608

[16] Tax D M J. DDtools: the Data Description Toolbox for Matlab (version 1.7.3) [CP/OL]. http://ict.ewi.tudelft.nl/~davidt/dd_tools.html, 2010-2-20

[17] Tax D M J. One-class classifier results [DB/OL]. http://ict.ewi.tudelft.nl/~davidt/occ/ index.html, 2010-2-20