

文章编号: 1001-0920(2011)11-1665-05

密度诱导型数据描述单类分类机

王万良^a, 王震宇^b, 郑建炜^a, 郑泽萍^a

(浙江工业大学 a. 计算机科学与技术学院, b. 信息与工程学院, 杭州 310023)

摘要: 为改进基于数据描述的单类分类机识别率, 将样本分布密度加入分类机的设计中, 提出采用密度诱导型数据描述单类分类机 (DISVDD). 以支撑向量域描述 (SVDD) 算法为基础, 通过一种简易的形式引入数据密度因子, 使高密度区数据对分类支撑域的作用被强化, 而低密度区数据的作用被削弱, 结果使分类超球体因靠近高密度区而提高其识别性能, 而且不增加计算复杂度. 在构造样本值与真实数据集上的实验结果表明, 所提出的算法对于不同类型的数据均具有更好的推广性.

关键词: 密度因子; 单类分类机; 超球体; 数据描述

中图分类号: TP18

文献标识码: A

Density-induced data description one-class classifier

WANG Wan-liang^a, WANG Zhen-yu^b, ZHENG Jian-wei^a, ZHENG Ze-ping^a

(a. College of Computer Science and Technology, b. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China. Correspondent: WANG Zhen-yu, E-mail: wang852211@163.com)

Abstract: The density induced support vector data description (DISVDD) is proposed to improve the recognition rate of one-class classifier by introducing the structure of the given data. Based on SVDD, DISVDD ensures that the learn ability of data domain having higher density structure is enhanced. Meanwhile, the learn ability of data domain having lower density structure is weakened. Consequently, DISVDD has better domain margin and gets higher recognition rate without the increase of computation load. Experiments with various real datasets show better promising results.

Key words: density factor; one-class classifier; hypersphere; data description

1 引言

数据描述单类分类机^[1]是指训练过程中只有一类目标数据, 其他非目标数据因为无法轻易采集 (如机械故障诊断, 往往只能获取机器正常运转时的状态值, 而异异常数值却可遇而不可求) 或者数据过多无法得到具有代表性的另类样本 (如指纹识别、说话人确认等场合, 非目标类样本数量远远多于目标数据) 不得不放弃获取, 因而必须对目标类数据进行精细描述, 使得当新样本出现时能够被算法正确认识而接受或因陌生而拒绝. 目标数据又称为正常类, 而非目标数据则称为奇异点或负类. 针对数据描述问题所设计的单类分类机, 可用于图像检索^[2]、人脸识别^[3]和手写体数字分割^[4]等.

典型的数据描述研究方法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 (GMM)^[5]首先估计某类训练数据

的概率密度函数; 然后通过计算新测试样本属于该类别数据的概率值进行判别. 如高于设定阈值, 则判为目标样本; 反之, 则拒绝. 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计算更复杂的概率密度函数, 得不偿失. 基于支撑边界的方法是引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, 而负类数据则尽可能远地隔离在区域之外, 达到识别错误率最小的目的. 这个超球体或者超平面是由不为零的支撑量张成的, 故称为支撑域.

支撑域方法是支撑向量机 (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), 而且进行正负样本分类的方法是寻找最优支撑域而非概率密度, 算法计算量较少, 适合高维、含噪、有限训练样本的数据场合, 目前已成为单类分类机研究的主要方法. 典型的支撑域方法

收稿日期: 2010-06-29; 修回日期: 2010-09-06.

基金项目: 国家自然科学基金项目(60573123); 教育部高校博士点基金项目(20060337002).

作者简介: 王万良(1957-), 男, 教授, 博士生导师, 从事计算机控制与智能化等研究; 王震宇(1985-), 男, 硕士生, 从事模式识别、语音识别等研究.

中,单类支持向量机(OCSVM)^[6]是超平面支撑域的代表模型,用原点作负类,隐含了负类数据的位置信息,对于某些实际数据而言并不合理;而超球模型的典型算法:支持向量数据描述(SVDD)^[1],由于缺乏对数据本身分布信息的考虑,在处理各向异构的数据时,往往得到次优边界。

基于已有算法的缺陷, Tsang 等人从单类最小最大概率机(OCMPM)^[7]的思想中得到启发,用马氏距离取代 OCSVM 中的欧氏距离而设计出了 MOCSVM,并将其引入核学习^[8],从而获得更优的支撑域; Dolia 等人将 SVDD 中的超球用超椭球替代^[9],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 等人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,并将算法应用于多分类场合^[10]。上述算法虽然致力于获取更加合理的支撑边界,但都没有考虑目标数据本身的分布特征,实际上,数据内部的分簇结构以及数据在不同区域的分布密度等先验信息都能更好地指导支撑域的形成。KiYoung 等人通过 k 近邻方法对每个样本点引入局部密度因子,并将其应用于模型构建^[11]。类似地,冯爱民等人也对目标数据进行分簇,并将各簇协方差矩阵引入到最优支撑域的获取^[12]。二者都通过数据本身的分布特点来寻找最优的包络边界。然而,在寻优过程中都需要借助其他算法进行先验信息获取,增加了算法的计算量,并且两种算法都引入了新的经验参数,需要通过网格搜索获取或者依经验指定,从而增加了模型构建的复杂度。

为了提高超球体模型对不同类型数据的适应能力,在 SVDD 算法的基础上引入数据类内特征权值,并融入算法最大间隔寻优公式。因为类内特性由数据间密度引出,故称为密度诱导型数据描述单类分类机(DISVDD),不仅保留了原 SVDD 算法的诸多优点,结合了数据的结构信息,带有更多的先验知识,而且没有寻求其他算法的帮助,在计算复杂度不增加的前提下获得了数据描述性能的大幅度提升。因此,与基于超球体的同类算法相比,理论上具有更好的推广性。

2 支持向量域描述

2.1 线性描述 SVDD

给定目标数据集

$$X = \{x_i\}_{i=1}^n, x_i \in R^d.$$

其中: n 为样本数目, d 为样本维度。SVDD 的目标是在空间中寻找一个超球体 (a, R) , a 为球体中心, R 为球体半径。超球在必须满足半径 R 尽可能小的同时,应尽可能多地包络目标数据,同时为适应存在稀疏偏远点的部分数据集,引入非负松弛变量 $\xi_i, i = 1, 2,$

\dots, n , 即目标问题规划如下:

$$\min \left(R^2 + \frac{1}{vn} \sum_{i=1}^n \xi_i \right); \quad (1)$$

$$\text{s.t. } \|x_i - a\|^2 \leq R^2 + \xi_i, \quad (2)$$

$$\xi_i \geq 0, i = 1, 2, \dots, n. \quad (3)$$

其中: $v \in (0, 1]$ 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,对允许超球外面存在样本的约束程度越大。

上述优化问题的解可由相应拉格朗日泛函的鞍点给出,即

$$L(R, a, \xi_i, \alpha_i, \lambda_i) = R^2 + \frac{1}{vn} \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i (R^2 + \lambda_i - \|x_i - a\|^2) - \sum_{i=1}^n \lambda_i \xi_i, \quad (4)$$

其中 $\alpha_i \geq 0$ 和 $\lambda_i \geq 0$ 为该目标类样本集的拉格朗日系数。求式(4)最小值可转化为求其 Wolf 对偶问题的最大值

$$W(\alpha) = \max \left\{ \sum_{i=1}^n \alpha_i (x_i x_i) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j (x_i x_j) \right\}; \quad (5)$$

$$\text{s.t. } \sum \alpha_i = 1, 0 \leq \alpha_i \leq \frac{1}{vn}. \quad (6)$$

记式(5)二次规划问题最优解 $\alpha^* = \{\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*\}$, 其中对应 $\alpha_i^* = 0$ 的样本位于超球内部。对应 $0 < \alpha_i^* < 1/(vn)$ 的样本落在超球面上,称作(非边界)支撑向量。余下的对应 $\alpha_i^* = 1/(vn)$ 的样本位于超球外部,称作边界支撑向量。超球球心可由非零 α_i^* 线性组合计算出 $a = \sum_{i=1}^n \alpha_i^* x_i$ 。任选一个 $\alpha_i^* \in (0, 1/(vn))$, 可求得超球半径 $R^2 = \|x_i - a\|^2$ 。对于新测试样本 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。接受条件为

$$\|z - a\|^2 = zz - 2 \sum_{i=1}^n \alpha_i^* (zx_i) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j (x_i x_j) \leq R^2. \quad (7)$$

2.2 核描述 SVDD

输入空间的样本点往往不满足球状分布,可通过核技巧将输入空间先映射到高维空间,然后在映射后的高维空间内求解,对应于原空间中的非线性包络,即将上节公式中的内积形式都变换成核函数形式

$$x_i x_j \rightarrow \phi(x_i) \phi(x_j) = K(x_i, x_j). \quad (8)$$

其中 ϕ 为非线性映射,一般无法显式地求出其具体形式。常见的核函数包括:

1) p 阶多项式核函数

$$K(x, y) = (1 + xy)^p;$$

2) RBF 核函数

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right);$$

3) Sigmoid 核函数

$$K(x, y) = \tanh[b(xy) - c].$$

引入核函数后, 原线性 SVDD 最优泛函式 (5) 改为

$$\max \sum_{i=1}^n \alpha_i K(x_i, x_i) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j). \quad (9)$$

约束条件不变, 而新测试样本决策条件变成

$$\begin{aligned} \|\phi(z) - a\|^2 = \\ K(z, z) - 2 \sum_{i=1}^n \alpha_i^* K(z, x_i) + \\ \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \leq R^2. \end{aligned} \quad (10)$$

3 密度诱导型数据描述分类机

第 2 节所述 SVDD 原型算法, 虽然出发点是为数据描述最精细的球体边界, 但并没有考虑训练样本的分布密度, 对于不呈简单团簇结构的数据集往往捉襟见肘, 即使可以采用核映射方法间接实现原输入空间的非线性包络, 但在实际特征空间中仍以球体形式区分分类内外数据. 事实上, 如果单类分类机只是简单地将全部训练数据同等对待, 反而会导致球体为包括部分稀疏偏远点而包络过大, 使得奇异点进入球体内部, 从而影响分类机的推广性能. 密度诱导型数据描述分类机的目标是由数据分布密度指引支撑边界的划分, 使之向高密度数据区域靠拢.

在原有算法框架不变的前提下, 在目标泛函中引入密度矩阵 A , 作用是凸显近邻样本点, 抑制离群样本点. A 中元素一般计算方法如下:

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\lambda^2}\right), \quad (11)$$

其中 $\lambda > 0$ 为可调整参数, 需多次验证确定其取值. 为了避免对 λ 进行网格搜索而带来的巨大计算负担, 也为了使密度矩阵 A 具有稀疏性, 引入欧氏近邻法, 将样本中的相邻点指定对应的密度因子, 而非近邻点的密度权值取为零. 如样本 x_i 先在样本空间中以欧氏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 x_j 与 x_i 的距离小于 N_i^k , 则认为 x_j 属于 x_i 的近邻, 取密度权值如下:

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{N_i^k N_j^k}\right), \quad (12)$$

其余 $A_{i,j}$ 设为零. 一般认为 k 值为 7 时能获得较优的结果^[13].

保持支撑向量数据描述算法基本框架不变, 引入

密度矩阵 A 后, 目标泛函由式 (9) 改为

$$\max \sum_{i=1}^n \alpha_i K(x_i, x_i) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j A_{i,j} K(x_i, x_j). \quad (13)$$

约束条件依旧不变, 并且模型的新测试样本决策条件也保持式 (10) 不变. 分析式 (13) 可知, 虽然引入密度因子 $A_{i,j}$ 后变换了目标泛函表达式, 但依然保持了原算法的所有优点: 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, 密度矩阵 A 本身具有对称性, 因此核矩阵 K 对称结构依然不变, 主对角元素也仍为 1, 其他矩阵元素却依数据间密度权值进行了修正. 也就是说, 在高维特征空间, 数据描述域不再分布在一个超球体上, 而是在数据局部密度的作用下构造了一个新核. 这个新核包含了数据的先验知识, 能够更好地帮助边界支撑域的形成, 所以通过式 (13) 所得到的 α^* 是经局部密度信息作用后的超球体参数. 在密度因子 $A_{i,j}$ 的作用下, 训练数据中的各个分簇都能得到公正的对待, 而零星散点则被视作奇异点不予考虑, 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. 因此, 对于新来的样本, 尽管所用的判别式 (10) 形式上与原算法相同, 但支撑域已经变更, 因此分类器的推广性能得到了提高.

从计算复杂度的角度分析 (设定输入样本数为 n), 局部密度因子 $A_{i,j}$ 与实验采用的 RBF 核函数具有基本一致的形式 (当采用式 (11) 时完全一致), 因此耗费 $O(nk - k^2)$ 计算所得的 N_i^k 即为 $A_{i,j}$ 的计算时间复杂度. 该密度因子 (如式 (13)) 作用在核矩阵上, 增加了 $O(n^2)$ 的时间复杂度, 因此所有计算复杂度为 $O(nk - k^2 + n^2)$. 原 SVDD 算法中的二次规则问题所需时间复杂度为 $O(n^3)$, 所以密度诱导型数据描述算法的计算时间复杂度仍然保持为 $O(n^3)$, 而且得到密度矩阵 A 后作用于 K , 形成一个新的核矩阵, 则算法空间复杂度也与原 SVDD 算法保持一致. 换言之, 密度因子的引入并没有增加算法的时间和空间复杂度, 但使分类机的识别率有了很大的提升.

4 实 验

为了验证所提出的密度诱导型分类机在数据呈不同分布情形下的支撑域描述能力及分类识别率, 分别采用人造数值与 UCI 公共测试数据库中的部分数据集进行直观的域描述, 以及不同算法间的识别性能对比. 算法所需核函数均采用 RBF 核, 其中核带宽 σ 由网格搜索和 5-fold 交叉验证过程获取, 所有实验结果都在 CPU 主频为 2.6 GHz, 内存 2G 的台式机中产生, 软件环境为 Windows XP 操作系统, Matlab 7.1 编译平台并内嵌 SVDD 工具箱 DD_Tools^[14].

4.1 数据描述性能对比

为了直观地了解数据的空间结构对不同域描述算法性能的影响,实验先通过 DD-Tools 构造 4 组不同分布的 2 维数据集,分别采用 SVDD 及 LPDD 描述它们的边界. 4 组数据集分别为: 1) Simple Data, 服从均值为 0, 方差为 1 的高斯分布; 2) Banana Data, 在香蕉形区域内, 方差为 1 的高斯分布与均匀分布叠加; 3) Long Data, 均值变化、方差为 1 的高斯分布并且旋转 45°; 4) Lithuanian Data, 在两瓣香肠形区域内, 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.

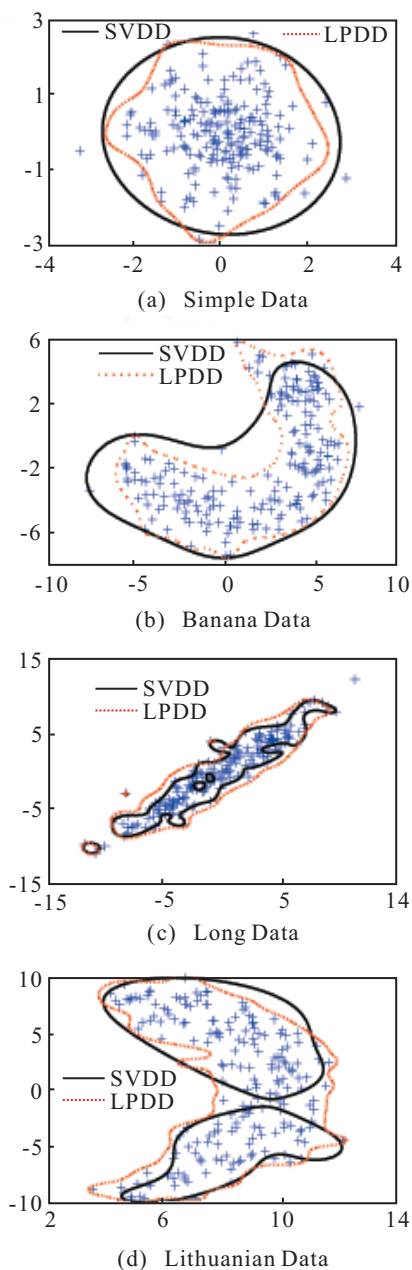


图 1 SVDD/DISVDD 数据描述边界对比显示

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, 因此密度诱导型数据描述算法

(DISVDD) 的包络边界向内收进, 更能反应数据的本质结构, 而支持向量数据域描述 (SVDD) 的包络相对偏大, 易将奇异点归入类内数据; 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) 反应了 SVDD 为了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, 而中间交叉部分数据则被隔在包络之外.

综观图 1, 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, 而密度诱导型数据描述算法在引入局部密度因子后, 则能够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 公共数据集识别实验

作为一种单类分类算法, 数据域描述分类机训练阶段仅需要目标类数据, 得到目标类的描述轮廓. 对于测试样本, 如果落入边界区域内, 则接受该样本; 否则, 拒绝. 实验采用 UCI 公共数据库中的 8 个数据集作为算法性能验证对象, 各数据集集中的第一类样本作为正样本, 余下的则为奇异点样本. 表 1 显示了各数据集参量. 其中: DimFea 为数据特征维度, TrNum 为训练样本数, TeNum 为测试样本数, P 为正测试样本数, F 为负测试样本数.

表 1 SVDD/DISVDD 性能对比采用的数据集

| | DimFea | TrNum | TeNum ($P : F$) |
|------------|--------|-------|-------------------|
| Brest | 9 | 217 | 241:458 |
| Heart | 13 | 148 | 164:139 |
| Sonar | 60 | 100 | 111:97 |
| Arrhythmia | 278 | 214 | 237:183 |
| Diabetes | 8 | 450 | 500:268 |
| Spectf | 44 | 229 | 254:95 |
| Waveform | 21 | 270 | 300:600 |
| Survival | 3 | 203 | 225:81 |

根据测试样本的真实类别和分类结果, 单类分类问题包括两种错误率, 即正确目标类被错分为异常类 (FN: false negative) 和异常类被错分为正确目标类 (FP: false positive). 显然, 这 2 个指标值越小, 相应的算法性能越好. 另外, AUC (area under ROC curve) 值也是度量单类分类机性能的一种重要指标^[15], 其值越大, 分类机性能越优. 表 2 列出了上述 8 组数据集在算法 SVDD 和 DISVDD 下得出的 FN 和 FP 错误率及 AUC 值.

由表 2 可以看出, 在对不同数据集测试过程中, 除 Breast 外, 其他数据集中由 DISVDD 算法所得到的 AUC/FN/FP 指标均优于 SVDD 算法; 而在 Breast 数据中, 虽然 SVDD 的 AUC 值略大于 DISVDD 的 AUC

表 2 SVDD 和 DISVDD 算法在 UCI 数据集上的 AUC/FN/FP

| | SVDD | | | DISVDD | | |
|------------|--------|--------|--------|--------|--------|--------|
| | AUC | FN | FP | AUC | FN | FP |
| Breast | 0.9878 | 0.1494 | 0.0218 | 0.9796 | 0.0373 | 0.0349 |
| Heart | 0.9512 | 0.1037 | 0 | 0.9683 | 0.0976 | 0 |
| Sonar | 0.8775 | 0.3153 | 0.0928 | 0.9285 | 0.0991 | 0.0412 |
| Arrhythmia | 0.8813 | 0.1603 | 0.1585 | 0.9597 | 0.0541 | 0.0546 |
| Diabetes | 0.6631 | 0.2960 | 0.4179 | 0.9248 | 0.0920 | 0.0534 |
| Spectf | 0.5140 | 0.3189 | 0.3200 | 0.9759 | 0 | 0.0316 |
| Waveform | 0.9264 | 0.2600 | 0.1183 | 0.9751 | 0.0500 | 0.0833 |
| Survival | 0.8572 | 0.2400 | 0.0741 | 0.8720 | 0.1644 | 0.0741 |

指标,且前者 FP 错误率略低于后者,但代价是 FN 指标 SVDD 远大于 DISVDD,因此这一项例外,无法否认密度诱导型数据描述算法的有效性.另外,两种算法的理论框架、训练算法均一致,只是 DISVDD 算法加入了对测试数据内部密度结构的考虑,可见,局部密度因子的引入确实可以增强 SVDD 的数据描述能力,使得算法在应对不同分布的数据集时,都能具有更精确的边界描述性能.

5 结 论

尽可能利用先验知识是提高分类器推广性能的关键.密度诱导型数据描述单类分类机(DISVDD)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,引入反映数据分布的密度因子,并构建密度矩阵,而后将此先验信息作为类内紧性嵌入经典的 SVDD 框架中,使得算法能够获得更具代表性的数据支撑域.作为 SVDD 的推广,DISVDD 不仅继承了原算法的全部优点,克服了 SVDD 并未考虑数据结构信息的不足,并且通过一种简易的方式刻画了数据的分布,算法复杂度保持不变,因此在实际数据集上表现出更好的推广性.由于考虑数据局部密度结构信息后,新算法的解稀疏性能有所下降,如数据集 Spectf 的实验中,SVDD 算法解的支持向量数为 93,而 DISVDD 算法解的支持向量数达到 156.未来工作中将重点分析此问题,希望能够加以改进,以便算法应用时具备更好的实时性.

参考文献(References)

[1] Tax D M J, Duin R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.
 [2] Hong Gu, Guangzhou Zhao, Hongbo Wang. Shape representation and recognition in high dimensional feature space[C]. Proc of the 2009 Int Conf on Computational Intelligence and Natural Computing. Washington: IEEE Computer Society, 2009: 186-189.
 [3] Changjun Chen, Yongzhao Zhan, Chuanjun Wen. Hierarchical face recognition based on SVDD and

SVM[C]. Int Conf on Environmental Science and Information Application Technology. Wuhan, 2009: 692-695.

- [4] Alves V M O, Oliveira A L I, Silva E R, et al. Handwritten digit segmentation in images of historical documents with one-class classifiers[C]. The 20th IEEE Int Conf on Tools with Artificial Intelligence. Dayton, 2008: 41-44.
 [5] Varchol P, Levicky D, Juhar J. Optimization of GMM for text independent speaker verification system[C]. The 18th Int Conf on Radioelektronika. Prague, 2008: 1-4.
 [6] Scholkopf B, Platt J C. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.
 [7] 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.
 [8] Tsang I W, James T K, Li S. Learning the kernel in Mahalanobis one-class support vector machines[C]. Proc of the Int Joint Conf on Neural Networks. Vancouver, 2006: 1169-1175.
 [9] Dolia A, Harris C, Shawe-Taylor J. Kernel ellipsoidal trimming[J]. Computational Statistics and Data Analysis, 2007, 52(1): 309-324.
 [10] Mu T, Nandi A K. Multiclass classification based on extended support vector data description[J]. IEEE Trans on System, Man and Cybernetics—Part B: Cybernetics, 2009, 39(5): 1206-1216.
 [11] Lee K Y, Kim D W, Lee K H, et al. Density-induced support vector data description[J]. IEEE Trans on Neural Networks, 2007, 18(1): 284-289.
 [12] 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.
 (Feng A M, Xue H, Liu X J, et al. Enhanced one-class SVM[J]. J of Computer Research and Development, 2008, 45(11): 1858-1864.)
 [13] Zelnik-Manor L, Perona P. Self-tuning spectral clustering[C]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2005: 1601-1608.
 [14] Tax D M J. DDtools.the data description toolbox for matlab[EB/OL]. [2010-06-15]. http://homepage.tudelft.nl/n9d04/dd_tools.html.
 [15] 陈斌, 李斌, 潘志松, 等. 流形嵌入的支持向量数据描述[J]. 模式识别与人工智能, 2009, 22(4): 548-553.
 (Chen B, Li B, Pan Z S, et al. Support vector data description with manifold embedding[J]. Pattern Recognition and Artificial Intelligence, 2009, 22(4): 548-553.)