SVDD 的快速实时决策方法

胡文军 [1,2] 王士同 [1]

摘 要 为了提高一类支持向量数据描述 (Support vector data description, SVDD) 对未知样本的决策速度, 本文从样本的核特征空间出发, 利用核超球球心在原始样本特征空间中的原像, 提出一种 SVDD 的快速决策方法 (Fast decision approach of SVDD, FDA-SVDD), 使得 SVDD 的决策复杂度从 O(n) 降低到 O(1). 同时, 对球心原像所在空间进行了分析, 并在此基础上给出了两种原像逼近方法. 多种真实数据集实验表明, FDA-SVDD 方法在保证测试精度的同时, 能快速实现对未知样本的决策.

关键词 异常检测, 支持向量数据描述, 快速决策, 核超球, 球心原像

DOI 10.3724/SP.J.1004.2011.01085

Fast Real-time Decision Approach of Support Vector Data Description

HU Wen-Jun[1,2] WANG Shi-Tong[1]

Abstract For improving the decision speed of one-class SVDD, a fast decision approach is proposed in this paper, called FDA-SVDD, by utilizing the preimage in original feature space corresponding to the center of sphere in kernel feature space, by which the decision complexity of SVDD is reduced from O(n) to O(1). Meanwhile, two approximate algorithms for finding preimage are presented based on analyzing its position in the original space. Experimental results on extensive datasets show that the proposed method can not only guarantee testing accuracy but also fast classify unknown samples.

Key words Novelty detection, support vector data description (SVDD), fast decision, kernelized sphere, the preimage of sphere center

异常检测是对正常样本 (或目标样本) 学习并获得其特征属性, 并依此特征属性实现检测, 是属于一类分类问题, 广泛应用于现实生活中, 如假币识别、用户识别、医疗诊断[1] 和机器故障诊断等[1−3]. 近几十年, 异常检测得到了广泛地研究, 并提出了许多相应的学习算法, 如基于密度估计方法[4]、基于神经网络方法[5] 和基于空间深度方法[6] 等.

本文主要针对一类核化的支持向量数据描述 (Support vector data description, SVDD)[2]. SVDD 在一类分类问题中表现出很好的性能[2], 特别是在故障数据较难获得的应用场合[6]. 同时, 一些学者为了改善其测试精度, 提出了一些改进版本, 如小球体大间隔 (Small sphere and large margin approach, SSLM) 方法[1] 等; 而一些学者将一类硬分割的 SVDD 等价于最小包含球 (Minimum enclosing ball, MEB) 问题[8−9], 并通过核心集向量机 (Core vector machine, CVM) 算法[8−11] 实现对大样本的训练. 然而, 我们注意到对于 SVDD 决策效率方面的研究不多, 而实际上 SVDD 决策 1 个未知样本的复杂度为 O(n) (n 是训练样本数), 当未知样本为 N 时, 其复杂度为 O(Nn), 这对于一些高实时性或大量未知样本的数据监测场合是非常不适应的. 为此, 本文利用核超球球心在原始样本空间中的原像, 提出了一种快速决策的 SVDD 方法 (Fast decision approach of SVDD, FDA-SVDD).

1 一类 SVDD

为了简洁, 在未加特殊说明外, 下文中 SVDD 指的是一类 SVDD.

1.1 SVDD 算法

首先, 定义一个训练样本集 $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, 其中 $\boldsymbol{x}_i \in \mathbf{R}^d$ $(1 \le i \le n)$ 为列向量, SVDD 的思想是在样本特征空间找到一个超球, 记为 $B(\boldsymbol{c}, r)$, 其中 $\boldsymbol{c}$ 和 r 分别为超球的球心和半径, 并使得 $B(\boldsymbol{c}, r)$ 尽可能紧地包络目标样本, 也就是超球体积尽可能小. 因此, 该优化问题描述为

$$\begin{aligned} &\min_{r,\boldsymbol{c},\xi_i} r^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad \|\boldsymbol{x}_i - \boldsymbol{c}\|^2 \le r^2 + \xi_i,\ \ \xi_i \ge 0,\ \ 1 \le i \le n \end{aligned} \tag{1}$$

其中, ξ_i 是引入的松弛项, 当 $\boldsymbol{x}_i$ 在 $B(\boldsymbol{c}, r)$ 超球内或

收稿日期 2010-11-12 录用日期 2011-03-02
Manuscript received November 12, 2010; accepted March 2, 2011
国家自然科学基金 (60903100, 60975027) 资助
Supported by National Natural Science Foundation of China (60903100, 60975027)
1. 江南大学信息工程学院 无锡 214122 2. 湖州师范学院信息与工程学院 湖州 313000
1. School of Information Engineering, Jiangnan University, Wuxi 214122 2. School of Information and Engineering, Huzhou Teachers College, Huzhou 313000

面上时, $\xi_i = 0$, 否则 $\xi_i > 0$; $C > 0$ 是控制参数, 调节错分训练样本数 (球外样本数) 和 r 的大小, 根据实际应用调节此参数.

通常引入核函数即所谓的核技巧改善算法的适应性, 实质是找到一个合适的映射 ϕ 将样本特征空间 $\mathbf{O} \subseteq \mathbf{R}^d$ 映射到一个尽可能高维的特征空间 $\boldsymbol{\Phi}$ 中, 即 ϕ: $\boldsymbol{x} \in \mathbf{O} \mapsto \phi(\boldsymbol{x}) \in \boldsymbol{\Phi}$, 当给定正定核函数 k: $\mathbf{R}^d \times \mathbf{R}^d \in \mathbf{R}$, 用此核诱导下的内积表示 $\boldsymbol{\Phi}$ 空间中的内积形式, 即 $\phi^{\mathrm{T}}(\boldsymbol{x}_i)\,\phi(\boldsymbol{x}_j)$, 则在 $\boldsymbol{\Phi}$ 空间中的 SVDD 模型为

$$\begin{aligned} &\min_{r,\boldsymbol{c}_\Phi,\xi_i} r^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad \|\phi(\boldsymbol{x}_i) - \boldsymbol{c}_\Phi\|^2 \le r^2 + \xi_i,\ \ \xi_i \ge 0,\ \ 1 \le i \le n \end{aligned} \tag{2}$$

上述模型所对应超球记为 $B(\boldsymbol{c}_\Phi, r)$. 构造上述模型的拉格朗日方程:

$$\begin{aligned} L(r,\boldsymbol{c}_\Phi,\xi_i,\boldsymbol{\alpha},\boldsymbol{\beta}) = r^2 + C\sum_{i=1}^{n}\xi_i - \\ \sum_{i=1}^{n}\alpha_i(r^2 + \xi_i - \|\phi(\boldsymbol{x}_i) - \boldsymbol{c}_\Phi\|^2) - \sum_{i=1}^{n}\beta_i\xi_i \end{aligned} \tag{3}$$

其中, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^{\mathrm{T}} \ge \mathbf{0}$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)^{\mathrm{T}} \ge \mathbf{0}$ 是拉格朗日乘子向量. 式 (3) 分别对优化问题的原始变量 r, $\boldsymbol{c}_\Phi$ 和 ξ_i 求偏导数并置为 0, 得:

$$\frac{\partial L}{\partial r} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i = 1 \tag{4}$$

$$\frac{\partial L}{\partial \boldsymbol{c}_\Phi} = 0 \Rightarrow \boldsymbol{c}_\Phi = \sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}_i) \tag{5}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i + \beta_i = C \tag{6}$$

由式 (6) 可知, $\alpha_i = C - \beta_i$, 当取 $0 \le \alpha_i \le C$ 时, $\beta_i \ge 0$ 必定成立, 故可省略此约束. 此时将式 (4) ∼ (6) 代入式 (3), 得到原始问题的对偶形式:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i = 1,\ \ 0 \le \alpha_i \le C,\ \ 1 \le i \le n \end{aligned} \tag{7}$$

1.2 决策

根据 KKT (Karush-Kuhn-Tucker) 定理, 当拉格朗日乘子 > 0 时, 式 (2) 中的不等式约束变成等式约束, 即: 当 $\beta_i > 0$ (此时 $\alpha_i < C$) 时, $\xi_i = 0$; 当 $\alpha_i > 0$ 时, $\|\phi(\boldsymbol{x}_i) - \boldsymbol{c}_\Phi\|^2 = r^2 + \xi_i$. 则有如下结论: 1) 当 $\alpha_i = 0$ 时, $\phi(\boldsymbol{x}_i)$ 在 $B(\boldsymbol{c}_\Phi, r)$ 内; 2) 当 $0 < \alpha_i < C$ 时, $\phi(\boldsymbol{x}_i)$ 在 $B(\boldsymbol{c}_\Phi, r)$ 球面上; 3) 当 $\alpha_i = C$ 时, $\phi(\boldsymbol{x}_i)$ 在 $B(\boldsymbol{c}_\Phi, r)$ 球外. 根据结论 2) 和式 (5) 可得:

$$\begin{aligned} r = \Bigg(k(\boldsymbol{x}_k,\boldsymbol{x}_k) - 2\sum_{i=1}^{n}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_k) + \\ \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\Bigg)^{0.5} \end{aligned} \tag{8}$$

其中, $\boldsymbol{x}_k$ 对应的拉格朗日乘子 $0 < \alpha_k < C$. 给定未知样本 $\boldsymbol{x} \in \mathbf{R}^d$, 通过下面的函数进行决策:

$$f(\boldsymbol{x}) = r^2 - \|\phi(\boldsymbol{x}) - \boldsymbol{c}_\Phi\|^2 \tag{9}$$

当给定某一核函数后, 如高斯核 $k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|^2/h)$ (h 为高斯核的带宽参数, $\|\cdot\|$ 是欧氏 2 范数), 式 (9) 可简写为

$$f(\boldsymbol{x}) = 2\sum_{i=1}^{n}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}) - v \tag{10}$$

其中

$$v = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) + 1 - r^2$$

是可计算的某一常量. 当 $f(\boldsymbol{x}) \ge 0$ 时, 该样本为目标样本, 否则为异常样本.

1.3 决策复杂度

根据第 1.2 节的式 (10) 可知, 给定核后, 决策 1 个未知样本的计算复杂度为 O(n), 但拉格朗日乘子中有部分 $\alpha_i = 0$, 其可以不参与式 (10) 计算, 因此其计算复杂度应低于 O(n). 为了分析其复杂度, 将 $\alpha_i > 0$ 所对应样本构成一个集合, 因为 $\alpha_i > 0$ 对应的样本称为支持向量[1−2], 故将此集合称为支持向量集 (Support vectors set, SVs). 此时, 式 (10) 改写为

$$f(\boldsymbol{x}) = 2\sum_{i \in SV_s}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}) - v \tag{11}$$

可见, SVDD 决策 1 个未知样本的复杂度为 O($|SV_s|$), 当给定 N 个未知样本时, 总体决策复杂度为 O($N|SV_s|$). 因此, SV_s 或 N 很大时计算量很大, 这对于一些实时数据监测场合的应用非常不适用. 而一般地, SV_s 又不能太小, 否则目标 (或异常) 样本被错分的可能性会更大 (见文献 [2] 中的图 4). 为此, 本文提出一种 SVDD 的快速决策方法 (Fast decision approach of SVDD, FDA-SVDD).

2 FDA-SVDD

为了引出本文方法的思路, 先给出 1 个定义和 3 个定理.

定义 1. 整个 d 维样本空间定义为 $\mathbf{O}$. 假设存在一个映射 ϕ: $\mathbf{O}\mapsto\mathbf{\Phi}$, 且 $\mathbf{\Phi}$ 可被超球 $B(\boldsymbol{c}_\Phi, r)$ 分割成球内, 球面和球外三个子空间, 记为 $\mathbf{\Phi}_{\rm in}$, $\mathbf{\Phi}_{\rm on}$, $\mathbf{\Phi}_{\rm out}$; 同时, 根据 $\phi(\mathbf{O})$ 属于 $\mathbf{\Phi}$ 中哪个子空间也将 $\mathbf{O}$ 分为 3 个子空间, 记为 $\mathbf{O}_{\rm in}$, $\mathbf{O}_{\rm on}$, $\mathbf{O}_{\rm out}$; 给定训练样本集 $\mathcal{X}=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbf{O}\}$, 同理将 $\mathcal{X}$ 分成 3 个子集, 记为: $\mathcal{X}_{\rm in}$, $\mathcal{X}_{\rm on}$, $\mathcal{X}_{\rm out}$, 可知 $\mathcal{X}_{\rm in}\subset\mathbf{O}_{\rm in}$, $\mathcal{X}_{\rm on}\subset\mathbf{O}_{\rm on}$, $\mathcal{X}_{\rm out}\subset\mathbf{O}_{\rm out}$.

定理 1. 给定训练样本集 $\mathcal{X}=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbf{O}, 1\le i\le n\}$, 则 $\mathbf{\Phi}$ 空间中必定存在超球 $B(\boldsymbol{c}_\Phi, r)$ 且将 $\mathbf{\Phi}$ 空间分割成 $\mathbf{\Phi}_{\rm in}$, $\mathbf{\Phi}_{\rm on}$,$\mathbf{\Phi}_{\rm out}$ 三个子空间.

证明. 根据第 1.1 节和第 1.2 节的 SVDD 算法可知, $B(\boldsymbol{c}_\Phi, r)$ 必定存在, 且 $\boldsymbol{c}_\Phi$ 和 r 可根据式 (5) 和式 (8) 计算获得, 只是 $\boldsymbol{c}_\Phi$ 形式不确定 (因 ϕ 形式不确定). □

定理 2. 给定训练样本集 $\mathcal{X}=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbf{O}, 1\le i\le n\}$, 则 $\mathbf{\Phi}$ 空间中超球 $B(\boldsymbol{c}_\Phi, r)$ 的球心 $\boldsymbol{c}_\Phi\in\mathbf{\Phi}_{\rm in}$.

证明. 根据式 (5) 可知,

$$\boldsymbol{c}_\Phi=\sum_{\alpha_i=0}\alpha_i\phi(\boldsymbol{x}_i)+\sum_{0<\alpha_i<C}\alpha_i\phi(\boldsymbol{x}_i)+\sum_{\alpha_i=C}\alpha_i\phi(\boldsymbol{x}_i)=\mathbf{0}+\phi_{\rm on}(\cdot)+\phi_{\rm out}(\cdot)\tag{12}$$

根据第 1.2 节结论 2) 和 3) 以及本节中定义, 可知: $\phi_{\rm on}(\cdot)\in\mathbf{\Phi}_{\rm on}$, $\phi_{\rm out}(\cdot)\in\mathbf{\Phi}_{\rm out}$. 1) 若 $\boldsymbol{c}_\Phi\in\mathbf{\Phi}_{\rm on}$, 那么 $\boldsymbol{c}_\Phi-\phi_{\rm on}(\cdot)\in\mathbf{\Phi}_{\rm on}$, $\phi_{\rm out}(\cdot)\in\mathbf{\Phi}_{\rm on}$, 这与 $\phi_{\rm out}(\cdot)\in\mathbf{\Phi}_{\rm out}$ 矛盾, 故假设不成立. 2) 同理, $\boldsymbol{c}_\Phi\in\mathbf{\Phi}_{\rm out}$ 亦不成立. 故定理 2 成立. □

定理 3. 若 $\boldsymbol{c}_\Phi$ 在 $\mathbf{O}$ 中的原像 $\boldsymbol{c}$ 存在, 即 $\boldsymbol{c}=\phi^{-1}(\boldsymbol{c}_\Phi)\in\mathbf{O}$, 则 $\boldsymbol{c}\in\mathbf{O}_{\rm in}$.

证明. 若 $\boldsymbol{c}\notin\mathbf{O}_{\rm in}$, 则 $\boldsymbol{c}\in\mathbf{O}_{\rm on}\bigcup\mathbf{O}_{\rm out}$. 根据定义, $\phi(\boldsymbol{c})\in\mathbf{\Phi}_{\rm on}\bigcup\mathbf{\Phi}_{\rm out}$, 这与定理 2 矛盾. □

2.1 本文思想及其决策复杂度

若 $\boldsymbol{c}_\Phi$ 原像 $\boldsymbol{c}$ 存在, 那么 $\boldsymbol{c}_\Phi=\phi(\boldsymbol{c})$. 根据定理 3, 若在 $\mathbf{O}_{\rm in}$ 空间中找到原像 $\boldsymbol{c}$, 则决策函数为

$$f(\boldsymbol{x})=r^2-\|\phi(\boldsymbol{x})-\phi(\boldsymbol{c})\|^2=2k(\boldsymbol{x},\boldsymbol{c})-v'\tag{13}$$

其中, $v'=k(\boldsymbol{x},\boldsymbol{x})+k(\boldsymbol{c},\boldsymbol{c})-r^2$ 是某一常量.

原像 $\boldsymbol{c}$ 和未知样本 $\boldsymbol{x}$ 均在 $\mathbf{O}$ 空间中, 计算 $k(\boldsymbol{x}, \boldsymbol{c})$ 的复杂度为 O(1), 对于决策过程而言, v' 是一常量. 因此, 式 (13) 的计算复杂度为 O(1). 而第 1.3 节中指出了 SVDD 决策一个未知样本的复杂度为 O($|SV_s|$) (一般地, 考虑到机器噪声及机器误差, SVDD 决策时常采用所有训练样本, 此时 SVDD 决策一个未知样本的复杂度为 O(n)), 显然 O(1) 远远小于 O($|SV_s|$) 和 O(n). 所以, 若能在 $\mathbf{O}$ 空间中找到 $\boldsymbol{c}_\Phi$ 的原像 $\boldsymbol{c}$, 这将提高 SVDD 的决策速度. 下面将重点讨论原像的获取.

2.2 原像获取

一个众所周知的假设: 空间中的点可通过其邻域内点的线性组合近似表示, 如局部线性嵌入 (Locally linear embedding, LLE) 等[12−13]. 因此, 给定 $\boldsymbol{c}$ 的某一邻域 $\boldsymbol{\delta}$, 则 $\boldsymbol{c}\approx\sum_i w_i\boldsymbol{\delta}_i$, 其中, $\boldsymbol{\delta}_i\in\boldsymbol{\delta}$, $\boldsymbol{w}=(w_1,\cdots,w_{|\boldsymbol{\delta}|})$ 为权向量, 且 $w_i\ge0$ 和 $\sum_i w_i=1$. 根据定理 3 知 $\boldsymbol{c}\in\mathbf{O}_{\rm in}$, 又因为 $\mathcal{X}_{\rm in}\subset\mathbf{O}_{\rm in}$, 所以本文将 $\boldsymbol{c}$ 的邻域选为 $\mathcal{X}_{\rm in}$, 则

$$\hat{\boldsymbol{c}}=\sum_{\boldsymbol{x}_i\in\mathcal{X}_{\rm in}}w_i\boldsymbol{x}_i\tag{14}$$

现在的问题是如何选定权向量 $\boldsymbol{w}=(w_1,\cdots,w_{|\mathcal{X}_{\rm in}|})$ 使得损失函数 $\|\hat{\boldsymbol{c}}-\boldsymbol{c}\|\to\min$. 根据中值定理, 可知

$$\begin{aligned}&\phi(\hat{\boldsymbol{c}})\approx\phi(\boldsymbol{c})+\phi'(\boldsymbol{\xi})(\hat{\boldsymbol{c}}-\boldsymbol{c})\Leftrightarrow\\&\phi(\hat{\boldsymbol{c}})-\phi(\boldsymbol{c})\approx\phi'(\boldsymbol{\xi})(\hat{\boldsymbol{c}}-\boldsymbol{c})\Rightarrow\\&\|\,\phi(\hat{\boldsymbol{c}})-\phi(\boldsymbol{c})\,\|\ge\|\,(\hat{\boldsymbol{c}}-\boldsymbol{c})\,\|\min(\phi'(\boldsymbol{\xi}))\end{aligned}\tag{15}$$

从上式可知, 为了使 $\|\,\hat{\boldsymbol{c}}-\boldsymbol{c}\,\|$ 达到最小下界, 我们可以通过使 $\|\,\phi(\hat{\boldsymbol{c}})-\phi(\boldsymbol{c})\,\|$ 达到下界来近似求解. 所以, 只需构造 $\boldsymbol{w}$ 的累积平方误差 (Integrated squared error, ISE), 即 $ISE(\boldsymbol{w})=\|\phi(\hat{\boldsymbol{c}})-\phi(\boldsymbol{c})\|^2$, 并使其尽可能小, 则有

$$\begin{aligned}\hat{\boldsymbol{w}}=&\min_{\boldsymbol{w}}ISE(\boldsymbol{w})=\\&\min_{\boldsymbol{w}}\Bigg(\sum_{\boldsymbol{x}_i\in\mathcal{X}_{\rm in}}\sum_{\boldsymbol{x}_j\in\mathcal{X}_{\rm in}}w_iw_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)-\\&2\sum_{\boldsymbol{x}_i\in\mathcal{X}_{\rm in}}w_i\sum_{\boldsymbol{x}_j\in SV_s}\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)+\\&\sum_{\boldsymbol{x}_i\in SV_s}\sum_{\boldsymbol{x}_j\in SV_s}\alpha_i\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)\Bigg)\end{aligned}\tag{16}$$

上式右侧第 3 项在求解 SVDD 时已知, 故可省略. 因此, $\boldsymbol{w}$ 可以通过下列优化模型求解获得, 即

$$\begin{aligned}&\max_{\boldsymbol{w}}\sum_{\boldsymbol{x}_i\in\mathcal{X}_{\rm in}}w_i\sum_{\boldsymbol{x}_j\in SV_s}\alpha_j2k(\boldsymbol{x}_i,\boldsymbol{x}_j)-\\&\sum_{\boldsymbol{x}_i\in\mathcal{X}_{\rm in}}\sum_{\boldsymbol{x}_j\in\mathcal{X}_{\rm in}}w_iw_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\quad\boldsymbol{w}^{\rm T}\mathbf{1}=1,\ w_i\ge0,\ 1\le i\le|\mathcal{X}_{\rm in}|\end{aligned}\tag{17}$$

2.3 权向量讨论

易知, 式 (17) 是一个 QP 问题, 其时间复杂度不小于 $\mathrm{O}(|\mathcal{X}_{\mathrm{in}}|^2)$, 空间复杂度为 $\mathrm{O}(|\mathcal{X}_{\mathrm{in}}|^2)$[14], 虽然 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 在 SVDD 中已获得, 但若 $|\mathcal{X}_{\mathrm{in}}|$ 很大时, 计算时间也非常大. 为此, 通过下面的定理给出一种直接求解方法. 定义下式:

$$w_k = \frac{\sum\limits_{\boldsymbol{x}_j \in SV_s} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_k)}{\sum\limits_{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{in}}} k(\boldsymbol{x}_i, \boldsymbol{x}_k)} \tag{18}$$

定理 4. 式 (18) 是式 (16) 的一个优化解.

证明. 置 $ISE(\boldsymbol{w})$ 对 w_k 的偏导为 0, 可得:

$$\begin{aligned}&\frac{\partial ISE(\boldsymbol{w})}{\partial w_k} = 2\phi^{\mathrm{T}}(\boldsymbol{x}_k) \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{in}}} w_i \phi(\boldsymbol{x}_i) - \\ &2\phi^{\mathrm{T}}(\boldsymbol{x}_k) \sum_{\boldsymbol{x}_j \in SV_s} \alpha_j \phi(\boldsymbol{x}_j) = \\ &2\phi^{\mathrm{T}}(\boldsymbol{x}_k)\left(\sum_{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{in}}} w_i \phi(\boldsymbol{x}_i) - \sum_{\boldsymbol{x}_j \in SV_s} \alpha_j \phi(\boldsymbol{x}_j)\right) = 0\end{aligned} \tag{19}$$

将式 (18) 代入式 (19) 可知, 定理 4 成立. □

可知, 定理 4 给出了权向量的另外一种求解方法. 在 SVDD 求解过程中核矩阵已经计算过, 故上述两种方法都不需要再次计算 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

2.4 算法实现

根据第 2.1 节 ∼ 第 2.3 节可总结出 FDA-SVDD 的实现算法, 包括 2 个过程, 如下 FDA-SVDD 算法:

(训练过程)

步骤 1. 初始化核宽度参数 h 和调节参数 C;

步骤 2. 求解式 (7) 的二次规划问题;

步骤 3. 根据式 (8) 计算超球半径 r;

步骤 4. 根据式 (17) 或式 (18) 求解权向量 $\boldsymbol{w}$;

步骤 5. 根据式 (14) 计算核空间中超球球心的原像 $\boldsymbol{c}$;

(决策过程)

步骤 6. 对于未知样本, 根据式 (13) 决策.

3 实验结果与分析

3.1 SVDD测试效率实验

本节实验主要分析 SVDD 算法的测试效率与 $|SV_s|$ 大小间的关系. 数据集为人造半月形数据, 如图 1 所示, 其中, "×" 是目标样本 (200 个), "△" 是异常样本 (50 个), "o" 是 SVDD 算法获得的 SV_s, 等高线是基于样本到球心的距离 $\|\phi(\boldsymbol{x}) - \boldsymbol{c}_\Phi\|$ 绘出. 训练样本为全部目标样本, 测试样本为全部目标和异常样本, 选择高斯核为核函数, 图 1 对应参数 $h = s^2/8$ (s 是所有训练样本 2 范数的平均值), $C = 0.01$. 通过调整参数 h 和 C 改变 $|SV_s|$ 大小, 从而获得不同 $|SV_s|$ 下的测试时间 Testing (单位: ms), 结果绘制成图 2 所示的曲线. 从图 2 可以看出, $|SV_s|$ 越大 SVDD 测试效率越低, 并且其与测试时间基本成线性关系, 验证了 SVDD 的测试复杂度为 $\mathrm{O}(|SV_s|)$.

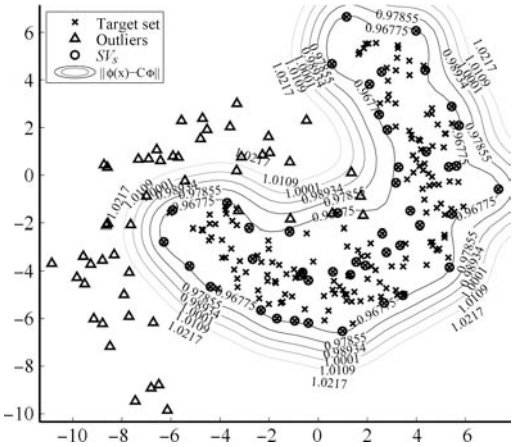


图 1 半月形数据集

Fig. 1 Half-moon dataset

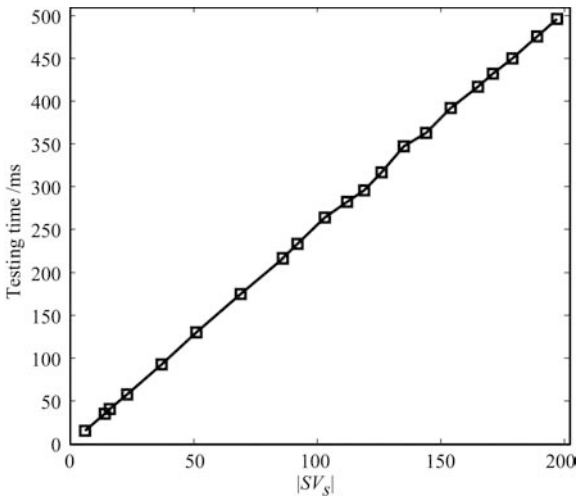


图 2 SVDD 测试效率与 $|SV_s|$ 关系曲线图

Fig. 2 The relation curve between testing time and $|SV_s|$ of SVDD

3.2 测试 UCI 数据集

本节利用表 1 中的 12 种 UCI[15] 数据集测试 SVDD 和 FDA-SVDD 的整体性能, 分别从测试精度 g (%), 训练时间 Training (单位: s) 和测试时间 Testing (单位: s) 三方面进行比较, 最终结果采用均值和标准差给出. 并根据权向量的

不同求解方法将 FDA-SVDD 分为 FDA-SVDD-I 和 FDA-SVDD-II, 前者对应 QP 求解即式 (17) 而后者对应直接求解方法即式 (18). 为了方便, 表 1 中每个数据集给定一个缩写字符, 对应方式为: W. (Wine), I. (Iris), B.C. (B.cancer), C.B. (Connectionist bench), A. (Arrhythmia), H.N.T. (Hill valley without noise training), B.S. (Balance scale), W.F. (Waveform), S.B. (Spambase), L.S. (Landsat satellite), P.R.H. (Pen based recognition of handwritten digits), B.S.C. (Blood transfusion service center). 表 1 中 B.C. 具体指 Diagnostic wisconsin breast cancer database 数据集; L.S. 是由类号为 1 和 7 样本构成; P.R.H. 由手写数字 0 和 1 样本构成. 同时 W.F., S.B., L.S. 和 P.R.H. 等 4 种数据集相对于其他 8 种数据集样本量较大.

训练样本和测试样本: 每次实验从一个数据集中选择 1 类作为目标类, 其他类作为异常类, 并从目标类中随机抽取 50 % 构成训练样本, 剩余 50 % 和异常类所有样本一起构成测试样本. 当选好参数后, 10 次随机运行, 统计性能指标并以均值和标准差输出. 关于参数选择: 控制参数 C 在 $\{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$中选择, 并选择高斯核进行实验, 带宽参数 h 在 $\{s^2/128, s^2/64, s^2/32, s^2/16, s^2/8, s^2/4, s^2/2, s^2, 2s^2, 4s^2, 8s^2\}$中选取, 其中 s 是所有训练样本 2 范数的平均值. 为公平起见, 每次实验中被比较算法的训练和测试样本以及参数完全相同. 实验环境: Pentium Core 2 2.6 GHz CPU, 2 G RAM, Windows XP, Matlab 2009a.

表 1 UCI 实验数据集 (12 种)

Table 1 Twelve UCI datasets

数据集	维数	目标类	样本数	数据集	维数	目标类	样本数
W.	13	1	59	B.S.	4	1	288
		2	71			2	49
		3	48			3	288
I.	4	1	50	W.F.	21	1	1 657
		2	50			2	1 647
		3	50			3	1 696
B.C.	30	1	357	S.B.	57	1	1 813
		2	212			2	2 788
C.B.	60	1	111	L.S.	36	1	1 533
		2	97			2	1 508
A.	278	1	237	P.R.H.	16	1	779
		2	183			2	780
H.N.T.	100	1	305	B.S.C.	4	1	570
		2	301			2	178

3.2.1 UCI 上的精度

表 2 给出了实验结果, 表中第 2 列是训练的目标类, 加粗字体标明 FDA-SVDD-I(II) 精度高于 SVDD. 从表 2 可知, 在 12 种数据集 28 个目标类中, FDA-SVDD-I 和 II 精度优于或略优于 SVDD 的

表 2 SVDD 和 FDA-SVDD 在 UCI 上的精度比较

Table 2 Comparison of accuracy on UCI datasets

数据集	目标类	SVDD	FDA-SVDD-I	FDA-SVDD-II	数据集	目标类	SVDD	FDA-SVDD-I	FDA-SVDD-II
W.	1	94.90 ± 0.91	$\mathbf{95.08 \pm 0.55}$	90.20 ± 3.78	B.S.	1	92.64 ± 0.58	84.57 ± 0.25	84.07 ± 0.30
	2	85.43 ± 2.28	$\mathbf{87.45 \pm 1.17}$	$\mathbf{86.81 \pm 0.00}$		2	97.36 ± 0.29	97.20 ± 0.22	$\mathbf{98.00 \pm 0.04}$
	3	89.36 ± 1.64	$\mathbf{91.24 \pm 0.69}$	$\mathbf{89.39 \pm 1.75}$		3	92.92 ± 0.43	84.54 ± 0.21	84.19 ± 0.43
I.	1	97.36 ± 1.35	$\mathbf{97.40 \pm 0.95}$	89.44 ± 0.00	W.F.	1	86.29 ± 1.01	$\mathbf{93.01 \pm 0.14}$	$\mathbf{91.65 \pm 0.08}$
	2	95.19 ± 1.59	$\mathbf{95.66 \pm 0.79}$	89.53 ± 0.19		2	91.30 ± 0.47	$\mathbf{93.45 \pm 0.16}$	90.71 ± 0.13
	3	96.91 ± 1.21	96.08 ± 0.91	87.86 ± 1.24		3	91.70 ± 0.23	$\mathbf{93.48 \pm 0.04}$	90.34 ± 0.16
B.C.	1	93.13 ± 0.65	92.55 ± 0.57	$\mathbf{93.30 \pm 0.74}$	S.B.	1	52.95 ± 0.45	$\mathbf{86.97 \pm 0.08}$	$\mathbf{86.56 \pm 0.32}$
	2	87.81 ± 5.32	$\mathbf{94.39 \pm 3.99}$	39.08 ± 2.19		2	73.28 ± 0.12	$\mathbf{74.58 \pm 0.19}$	$\mathbf{73.52 \pm 0.11}$
C.B.	1	79.63 ± 0.88	$\mathbf{80.36 \pm 1.74}$	$\mathbf{80.46 \pm 2.67}$	L.S.	1	91.79 ± 0.56	91.60 ± 0.45	91.59 ± 0.43
	2	84.56 ± 0.94	83.82 ± 0.40	83.55 ± 0.00		2	92.64 ± 0.77	$\mathbf{92.83 \pm 0.75}$	$\mathbf{94.19 \pm 0.41}$
A.	1	83.42 ± 0.85	79.80 ± 0.37	77.97 ± 0.00	P.R.H.	1	96.66 ± 0.87	95.28 ± 0.74	94.70 ± 0.36
	2	79.05 ± 1.11	$\mathbf{84.64 \pm 0.17}$	$\mathbf{85.00 \pm 0.00}$		2	96.78 ± 0.87	95.44 ± 0.67	94.99 ± 0.75
H.N.T.	1	67.29 ± 0.58	$\mathbf{79.86 \pm 1.30}$	$\mathbf{67.97 \pm 0.73}$	B.S.C.	1	75.74 ± 3.77	75.05 ± 3.52	$\mathbf{78.76 \pm 1.20}$
	2	68.16 ± 1.01	$\mathbf{81.39 \pm 0.22}$	68.00 ± 0.46		2	66.95 ± 3.88	$\mathbf{90.55 \pm 0.79}$	$\mathbf{87.45 \pm 1.24}$

目标类分别有 17 个和 13 个且较稳定. 而在其他目标类中 (除 B.S. 数据集外), 虽然本文方法比 SVDD 精度低, 但差别不大且较稳定. 为评价算法在整个数据集上的效果, 将每个数据集的各个目标类精度进行平均, 并命名为总平均精度 (Total average accuracy, TAA), 并用直方图绘出, 图 3 给出了结果, 横轴 1, 2, $\cdots$ 分别对应于表 2 的数据集. 从图 3 可以看出, FDA-SVDD-I 方法除第 7 和第 11 数据集外均比 SVDD 好; 对于 FDA-SVDD-II 方法, 除在第 2 和第 3 数据集精度较差外, 其他高于 SVDD 或与 SVDD 差别不大; 而 FDA-SVDD-I 整体较好于 FDA-SVDD-II. 因此, 说明了本文 FDA-SVDD 方法与 SVDD 相比在精度上具有较好的优势. 但本文方法是找出超球球心在原始空间中的原像, 这是一种逼近且一定会存在误差, 那为什么本文方法较 SVDD 有更好的优势? 可能原因是 SVDD 获得的最小 $B(\boldsymbol{c}_\Phi, r)$ 并不能实现最优分类, 而本文中的这种逼近误差或许可以改善 SVDD 的分类缺陷, 又或许其他原因, 因本文重点研究 SVDD 决策的快速性, 关于此问题不进行深入讨论, 这将作为我们近期的研究工作.

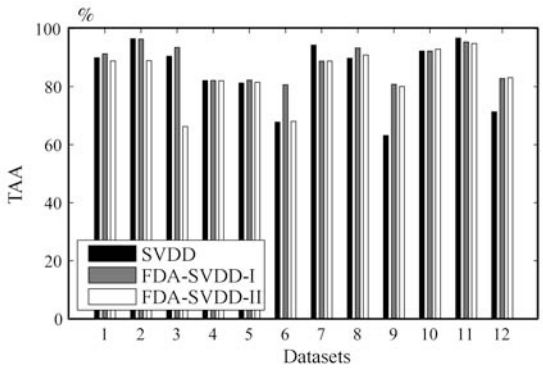


图 3 SVDD 和 FDA-SVDD 在 UCI 上的总平均精度比较

Fig. 3 Comparison of total average accuracies of SVDD and FDA-SVDD on UCI datasets

3.2.2 UCI 上的训练时间

本节实验比较 SVDD 和 FDA-SVDD 的训练效率, 同样对所有数据集的每个目标类进行实验, 表 3 给出了实验结果, 表中加粗字体标明各数据集总平均训练时间 (AV). 从表 3 可以看出: 1) 在所有数据集上, SVDD 的训练效率均高于本文方法 (如 Wine 数据集, SVDD: 0.0854 ± 0.1524 s, FDA-SVDD-I: 0.0940 ± 0.1592 s, FDA-SVDD-II: 0.0857 ± 0.1531 s), 这是因为本文方法在实现过程中除了需要运行 SVDD 算法外, 还要运行球心求解方法; 2) FDA-SVDD-II 与 SVDD 的训练时间相差不大, 这是因为 FDA-SVDD-II 采用式 (17) 直接求解权向量, 而不像 FDA-SVDD-I 需要求解 QP 问题, 在 W.F. 等 4 个较大数据集中 FDA-SVDD-I 的训练时间明显加大. 因此, 当样本较小时, FDA-SVDD-I 或 II 均可采用, 大样本时建议采用 FDA-SVDD-II 或其他方法.

3.2.3 UCI 上的测试时间

本节实验比较 SVDD 和 FDA-SVDD 的测试效率, 表 4 给出了实验结果, 表中加粗字体标明各个数据集的总平均测试时间 (AV). 从表 4 可以看出: 1) SVDD 的测试效率远低于 FDA-SVDD (如: 对于 W. 数据集, FDA-SVDD 快于 SVDD 算法近 50 倍; 对于 P.R.H. 数据集, 近 650 倍), 这是因为对 1 个待测样本而言, SVDD 的计算复杂度为 O($|SV_s|$), 而 FDA-SVDD 为 O(1); 2) FDA-SVDD-I 和 II 基本相同, 因为二者计算复杂度均为 O(1). 因此, 实验表明了本文 FDA-SVDD 方法在决策速度上远远快于 SVDD, 特别适用于在线数据监测和对大量样本数据测试的场合.

3.3 测试 PIE 数据集

本节实验利用 Pose-Illumination-Expression (PIE) 人脸图像比较 SVDD 和 FDA-SVDD 算法的性能. 可从网站 http://people.cs.uchicago.edu/~xiaofei 下载此数据集, 亦可参考文献 [16]. 实验数据集的构成: 从 PIE 数据库中选择标号为 1, 2, 35 和 38 的人脸图像构成 4 个目标类, 其中 2 人为男性, 2 人为女性, 38 号人脸图像为 164 张, 其余为 170 张, 如图 4 所示, 每张图像用一个 1 024 (32×32 个像素点) 维向量表示. 训练样本和测试样本: 每次实验从中选择 1 类作为目标类, 其他类作为异常类, 并从目标类中随机抽取 70 % 构成训练样本, 剩余 30 % 和异常类所有样本一起构成测试样本. 参数同第 3.2 节, 表 5 给出了实验结果.

图 4 PIE 数据集

Fig. 4 PIE datasets

从表 5 可以看出: SVDD 在 PIE 数据集上精度最好, 而 FDA-SVDD-I 略低于 SVDD, FDA-SVDD-II 相比更低; 在训练时间上, SVDD 最快, FDA-SVDD-II 次快, 而 FDA-SVDD-I 最慢, 其原因同 UCI 中的分析一样; 而在测试时间上, 本文方法明显快于 SVDD 算法. 因此, 本文方法在 PIE 数据集上亦能快速实现对未知样本的决策.

表 3 SVDD 和 FDA-SVDD 在 UCI 上的训练时间比较 (s)

Table 3 Comparison of training time of SVDD and FDA-SVDD on UCI datasets (s)

数据集	目标类	SVDD	FDA-SVDD-I	FDA-SVDD-II
W.	1	0.0899 ± 0.1547	0.1039 ± 0.1618	0.0902 ± 0.1554
	2	0.0952 ± 0.1524	0.1018 ± 0.1592	0.0955 ± 0.1531
	3	0.0712 ± 0.1501	0.0762 ± 0.1566	0.0714 ± 0.1508
	AV	$\mathbf{0.0854 \pm 0.1524}$	$\mathbf{0.0940 \pm 0.1592}$	$\mathbf{0.0857 \pm 0.1531}$
I.	1	0.0734 ± 0.1501	0.0843 ± 0.1570	0.0737 ± 0.1508
	2	0.0724 ± 0.1494	0.0825 ± 0.1560	0.0726 ± 0.1500
	3	0.0729 ± 0.1499	0.0880 ± 0.1553	0.0732 ± 0.1505
	AV	$\mathbf{0.0729 \pm 0.1498}$	$\mathbf{0.0849 \pm 0.1561}$	$\mathbf{0.0732 \pm 0.1504}$
B.C.	1	1.5164 ± 0.1761	2.1461 ± 0.2229	1.5180 ± 0.1768
	2	0.5095 ± 0.1634	0.6470 ± 0.1741	0.5103 ± 0.1640
	AV	$\mathbf{1.0130 \pm 0.1698}$	$\mathbf{1.3966 \pm 0.1985}$	$\mathbf{1.0142 \pm 0.1704}$
C.B.	1	0.2223 ± 0.2357	0.2266 ± 0.2422	0.2225 ± 0.2364
	2	0.1323 ± 0.1500	0.1553 ± 0.1609	0.1325 ± 0.1507
	AV	$\mathbf{0.1773 \pm 0.1929}$	$\mathbf{0.1910 \pm 0.2016}$	$\mathbf{0.1775 \pm 0.1936}$
A.	1	0.8284 ± 0.1598	0.8768 ± 0.1680	0.8293 ± 0.1605
	2	0.5252 ± 0.1458	0.5473 ± 0.2089	0.5255 ± 0.1465
	AV	$\mathbf{0.6768 \pm 0.1528}$	$\mathbf{0.7121 \pm 0.1885}$	$\mathbf{0.6774 \pm 0.1535}$
H.N.T.	1	1.3755 ± 0.1612	1.5436 ± 0.1696	1.3765 ± 0.1618
	2	1.3110 ± 0.1788	1.4735 ± 0.1838	1.3121 ± 0.1794
	AV	$\mathbf{1.3433 \pm 0.1700}$	$\mathbf{1.5086 \pm 0.1767}$	$\mathbf{1.3443 \pm 0.1706}$
B.S.	1	0.8022 ± 0.1623	0.8700 ± 0.1691	0.8031 ± 0.1629
	2	0.0695 ± 0.1466	0.0911 ± 0.2095	0.0698 ± 0.1472
	3	0.8034 ± 0.1615	0.8711 ± 0.1766	0.8043 ± 0.1622
	AV	$\mathbf{0.5584 \pm 0.1568}$	$\mathbf{0.6107 \pm 0.1851}$	$\mathbf{0.5591 \pm 0.1574}$
W.F.	1	204.0310 ± 21.7677	251.9950 ± 22.4666	204.0487 ± 21.7679
	2	209.1785 ± 7.8858	258.7930 ± 5.7377	209.1957 ± 7.8862
	3	243.1602 ± 36.8966	297.3589 ± 41.0658	243.1784 ± 36.8969
	AV	$\mathbf{218.7899 \pm 22.1834}$	$\mathbf{269.3823 \pm 23.0900}$	$\mathbf{218.8076 \pm 22.1837}$
S.B.	1	$1\,685.5320 \pm 420.8364$	$2\,997.1749 \pm 655.7414$	$1\,685.5575 \pm 420.8361$
	2	$6\,887.7102 \pm 1\,673.4445$	$12\,767.4279 \pm 3\,972.7983$	$6\,887.7727 \pm 1\,673.4364$
	AV	$\mathbf{4\,286.6211 \pm 1\,047.1405}$	$\mathbf{7\,882.3014 \pm 2\,314.2699}$	$\mathbf{4\,286.6651 \pm 1\,047.1363}$
L.S.	1	155.8472 ± 10.2147	197.8094 ± 10.8485	155.8623 ± 10.2150
	2	145.7836 ± 12.1826	183.8975 ± 15.2936	145.7981 ± 12.1832
	AV	$\mathbf{150.8154 \pm 11.1987}$	$\mathbf{190.8535 \pm 13.0711}$	$\mathbf{150.8302 \pm 11.1991}$
P.R.H.	1	19.5727 ± 1.2866	20.5551 ± 1.2526	19.5759 ± 1.2864
	2	19.3997 ± 1.2187	20.3623 ± 1.2885	19.4029 ± 1.2189
	AV	$\mathbf{19.4862 \pm 1.2527}$	$\mathbf{20.4587 \pm 1.2706}$	$\mathbf{19.4894 \pm 1.2527}$
B.S.C.	1	8.9961 ± 1.1779	9.1793 ± 1.1896	8.9974 ± 1.1782
	2	0.4432 ± 0.1907	0.5316 ± 0.2070	0.4437 ± 0.1913
	AV	$\mathbf{4.7197 \pm 0.6843}$	$\mathbf{4.8555 \pm 0.6983}$	$\mathbf{4.7206 \pm 0.6848}$

表 4 SVDD 和 FDA-SVDD 在 UCI 上的测试时间比较 (s)

Table 4 Comparison of testing time of SVDD and FDA-SVDD on UCI datasets (s)

数据集	类	SVDD	FDA-SVDD-I	FDA-SVDD-II	数据集	类	SVDD	FDA-SVDD-I	FDA-SVDD-II
W.	1	0.0502 ± 0.0004	0.0010 ± 0.0000	0.0010 ± 0.0000	B.S.	1	0.7615 ± 0.0035	0.0032 ± 0.0000	0.0032 ± 0.0001
	2	0.0578 ± 0.0006	0.0010 ± 0.0000	0.0010 ± 0.0000		2	0.1647 ± 0.0009	0.0040 ± 0.0000	0.0040 ± 0.0000
	3	0.0421 ± 0.0004	0.0011 ± 0.0000	0.0011 ± 0.0000		3	0.7582 ± 0.0040	0.0032 ± 0.0000	0.0032 ± 0.0000
	AV	**0.0500 ± 0.0005**	**0.0010 ± 0.0000**	**0.0010 ± 0.0000**		**AV**	**0.5615 ± 0.0028**	**0.0035 ± 0.0000**	**0.0035 ± 0.0000**
I.	1	0.0345 ± 0.0007	0.0009 ± 0.0001	0.0008 ± 0.0000	W.F.	1	40.2536 ± 0.0697	0.0291 ± 0.0002	0.0290 ± 0.0002
	2	0.0345 ± 0.0002	0.0008 ± 0.0000	0.0008 ± 0.0000		2	39.9566 ± 0.0914	0.0292 ± 0.0000	0.0288 ± 0.0002
	3	0.0345 ± 0.0002	0.0008 ± 0.0000	0.0008 ± 0.0000		3	40.7273 ± 0.0540	0.0289 ± 0.0000	0.0286 ± 0.0001
	AV	**0.0345 ± 0.0004**	**0.0008 ± 0.0000**	**0.0008 ± 0.0000**		**AV**	**40.3125 ± 0.0717**	**0.0291 ± 0.0001**	**0.0288 ± 0.0002**
B.C.	1	0.8418 ± 0.0076	0.0028 ± 0.0000	0.0028 ± 0.0000	S.B.	1	44.5599 ± 0.0798	0.0297 ± 0.0001	0.0296 ± 0.0002
	2	0.5835 ± 0.0024	0.0034 ± 0.0000	0.0033 ± 0.0000		2	59.6994 ± 0.0990	0.0259 ± 0.0001	0.0254 ± 0.0000
	AV	**0.7127 ± 0.0050**	**0.0031 ± 0.0000**	**0.0031 ± 0.0000**		**AV**	**52.1297 ± 0.0894**	**0.0278 ± 0.0001**	**0.0275 ± 0.0001**
C.B.	1	0.1116 ± 0.0007	0.0012 ± 0.0001	0.0012 ± 0.0000	L.S.	1	21.6684 ± 0.1214	0.0171 ± 0.0001	0.0168 ± 0.0002
	2	0.1027 ± 0.0005	0.0013 ± 0.0000	0.0012 ± 0.0000		2	21.3916 ± 0.1031	0.0171 ± 0.0001	0.0167 ± 0.0004
	AV	**0.1072 ± 0.0006**	**0.0013 ± 0.0001**	**0.0012 ± 0.0000**		**AV**	**21.5300 ± 0.1123**	**0.0171 ± 0.0001**	**0.0168 ± 0.0003**
A.	1	0.8234 ± 0.0038	0.0044 ± 0.0000	0.0040 ± 0.0001	P.R.H.	1	5.2252 ± 0.0324	0.0081 ± 0.0001	0.0080 ± 0.0000
	2	0.7148 ± 0.0047	0.0047 ± 0.0001	0.0045 ± 0.0002		2	5.2209 ± 0.0172	0.0081 ± 0.0001	0.0079 ± 0.0001
	AV	**0.7691 ± 0.0043**	**0.0046 ± 0.0001**	**0.0043 ± 0.0002**		**AV**	**5.2231 ± 0.0248**	**0.0081 ± 0.0001**	**0.0080 ± 0.0001**
H.N.T.	1	1.0456 ± 0.0045	0.0043 ± 0.0002	0.0040 ± 0.0000	B.S.C.	1	1.4578 ± 0.0123	0.0031 ± 0.0001	0.0031 ± 0.0001
	2	1.0381 ± 0.0047	0.0043 ± 0.0000	0.0041 ± 0.0000		2	0.6409 ± 0.0014	0.0043 ± 0.0000	0.0043 ± 0.0001
	AV	**1.0419 ± 0.0046**	**0.0043 ± 0.0001**	**10.0041 ± 0.0000**		**AV**	**1.0494 ± 0.0069**	**0.0037 ± 0.0001**	**0.0037 ± 0.0001**

表 5 SVDD 和 FDA-SVDD 在 PIE 数据集上的比较

Table 5 Comparison of SVDD and FDA-SVDD on PIE datasets

算法	目标类	精度 g (%)	训练时间 (s)	测试时间 (s)
SVDD	1	94.58 ± 1.08	0.6575 ± 0.2578	1.4836 ± 0.0719
	2	96.14 ± 0.83	0.6099 ± 0.1628	1.6342 ± 0.0514
	35	92.52 ± 0.64	0.6347 ± 0.1992	2.0717 ± 0.2545
	38	94.99 ± 0.80	0.5798 ± 0.1714	1.5785 ± 0.0618
	AV	**94.56 ± 0.84**	**0.6205 ± 0.1978**	**1.6920 ± 0.1099**
FDA-SVD-D-I	1	91.14 ± 0.20	0.8480 ± 0.2649	0.0216 ± 0.0004
	2	90.92 ± 0.13	3.7497 ± 4.8841	0.0214 ± 0.0001
	35	91.30 ± 0.19	0.8773 ± 0.2160	0.0234 ± 0.0035
	38	91.11 ± 0.09	0.7391 ± 0.1876	0.0217 ± 0.0002
	AV	**91.12 ± 0.15**	**1.5535 ± 1.3882**	**0.0220 ± 0.0010**
FDA-SVD-D-II	1	86.47 ± 1.34	0.6606 ± 0.2586	0.0215 ± 0.0002
	2	90.74 ± 0.74	0.6131 ± 0.1636	0.0213 ± 0.0002
	35	90.77 ± 0.11	0.6481 ± 0.2168	0.0232 ± 0.0037
	38	89.66 ± 0.51	0.5828 ± 0.1722	0.0217 ± 0.0002
	AV	**89.41 ± 0.68**	**0.6262 ± 0.2028**	**0.0219 ± 0.0011**

3.4 测试 USPS 数据集

本节实验利用 USPS 手写数字图像数据集比较 SVDD 和 FDA-SVDD 算法的性能, USPS 包含 7 291 训练和 2 007 测试样本, 并能从 http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/ datasets 下载得到, 每幅图像用 256 (16×16 个像素点) 维向量表示, 如图 5 所示, 关于 USPS 的更多信息请参考文献 [17]. 本节实验选择 USPS 的训练样本为实验数据集, 数字 0 ∼ 9 的手写图像分别有 1 194, 1 005, 731, 658, 652, 556, 664, 645, 542 和 644 张, 并分别选择数字 1, 3, 5, 7 和 9 为目标类, 每次实验从目标类中随机抽取 50 % 构成训练样本, 剩余 50 % 和其他数字样本一起构成测试样本. 参数选择同第 3.2 节, 表 6 给出了实验结果.

从表 6 可以看出: 在 5 个目标类的平均精度上, 本文方法略优于 SVDD; 训练速度上, FDA-SVDD-II 略慢于 SVDD, 而 FDA-SVDD-I 较慢于 SVDD; 分类测试速度上, 本文方法快于 SVDD 近 600 倍. 此实验中, 测试样本是相对较大 (如当选择数字 1 为目标样本时, 训练样本约 500 个, 测试样本约 6 800

个), 表 6 中的测试时间 (第 5 列) 说明了本文方法在处理大量未知样本时具有很好的性能.

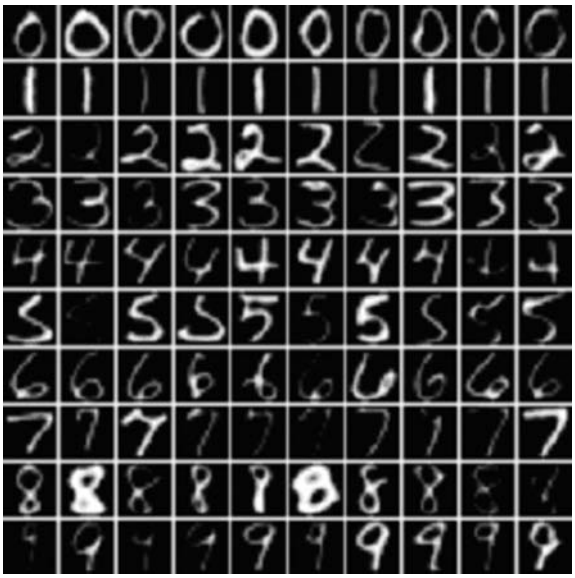


图 5　USPS 数据集

Fig. 5　USPS datasets

表 6　SVDD 和 FDA-SVDD 在 USPS 数据集上的比较

Table 6　Comparison of SVDD and FDA-SVDD on USPS datasets

算法	目标类	精度 g (%)	训练时间 (s)	测试时间 (s)
SVDD	1	99.58 ± 0.19	41.2156 ± 1.5851	89.5948 ± 0.6116
	3	98.51 ± 0.40	11.8668 ± 0.1886	57.1451 ± 0.3601
	5	92.65 ± 2.65	7.3298 ± 0.0232	48.0155 ± 0.3566
	7	97.69 ± 0.65	11.3417 ± 0.4467	55.9716 ± 0.2743
	9	96.31 ± 1.43	12.3271 ± 1.1809	57.8683 ± 0.8466
	AV	$\mathbf{96.95 \pm 1.06}$	$\mathbf{16.8162 \pm 0.6849}$	$\mathbf{61.7191 \pm 0.4898}$
FDA-SVD-D-I	1	99.50 ± 0.22	52.9450 ± 1.1713	0.1088 ± 0.0011
	3	98.84 ± 0.12	13.2492 ± 0.1002	0.1111 ± 0.0021
	5	97.75 ± 0.41	8.0131 ± 0.0314	0.1122 ± 0.0007
	7	98.74 ± 0.20	12.6684 ± 0.4120	0.1066 ± 0.0007
	9	96.94 ± 0.63	13.9694 ± 1.0953	0.1124 ± 0.0005
	AV	$\mathbf{98.35 \pm 0.32}$	$\mathbf{20.1690 \pm 0.5620}$	$\mathbf{0.1102 \pm 0.0010}$
FDA-SVD-D-II	1	99.40 ± 0.22	41.2254 ± 1.5867	0.1083 ± 0.0002
	3	98.64 ± 0.10	11.8714 ± 0.1872	0.1103 ± 0.0015
	5	97.95 ± 0.42	7.3334 ± 0.0220	0.1122 ± 0.0011
	7	98.08 ± 0.34	11.3460 ± 0.4461	0.1067 ± 0.0008
	9	97.31 ± 0.59	12.3317 ± 1.1806	0.1133 ± 0.0010
	AV	$\mathbf{98.28 \pm 0.33}$	$\mathbf{16.8216 \pm 0.6845}$	$\mathbf{0.1102 \pm 0.0009}$

4　结论

由于受到支持向量个数的影响, SVDD 很难快速实现对未知样本的决策. 为此, 本文在分析映射核空间中超球球心位置的基础上, 提出利用超球球心原像的 FDA-SVDD 方法, 使得 SVDD 的决策复杂度从 O(n) 降低到 O(1), 从而实现快速决策. UCI、PIE 和 USPS 数据集的实验结果表明: 本文提出的 FDA-SVDD 算法在精度上高于或略低于 SVDD 算法; 在训练速度方面, 对于较大样本而言 FDA-SVDD-I 不如 SVDD 快, 但 FDA-SVDD-II 基本能保持与 SVDD 相同的水平, 因此, 可以根据训练样本的大小选择合适的 FDA-SVDD 方法; 而在测试速度方面, 本文提出的 FDA-SVDD 算法明显快于 SVDD, 因此可以应用于实时性很高的数据监测或需要实现对大量样本测试的场合. 总体而言, FDA-SVDD 改善了 SVDD 的决策速度, 但 FDA-SVDD 是采用逼近算法获得 SVDD 球心原像, 总体精度为何会优于 SVDD? 同时, 如何提高 FDA-SVDD-I 训练效率? 关于这两个问题本文未进行深入地研究和讨论, 这将作为我们近期的研究重点.

References

1 Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(11): 2088−2092

2 Tax D M J, Duin R P W. Support vector data description. *Machine Learning*, 2004, **54**(1): 45−66

3 Zhao Feng, Zhang Jun-Ying, Liu Jing. An optimizing kernel algorithm for improving the performance of support vector domain description. *Acta Automatica Sinica*, 2008, **34**(9): 1122−1127
(赵峰, 张军英, 刘敬. 一种改善支撑向量域描述性能的核优化算法. 自动化学报, 2008, **34**(9): 1122−1127)

4 Roberts S, Tarassenko L. A probabilistic resource allocation network for novelty detection. *Neural Computation*, 1994, **6**(2): 270−284

5 Towell G G. Local expert autoassociators for anomaly detection. In: Proceedings of the 17th International Conference on Machine Learning. Stanford, USA: Morgan Kaufmann Publishers, 2000. 1023−1030

6 Chen Y X, Dang X, Peng H X, Bart H L. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(2): 288−309

7 Xie Lei, Liu Xue-Qin, Zhang Jian-Ming, Wang Shu-Qing. Non-Gaussian process monitoring based on NGPP-SVDD. *Acta Automatica Sinica*, 2009, **35**(1): 107−112
(谢磊, 刘雪芹, 张建明, 王树青. 基于 NGPP-SVDD 的非高斯过程监控及其应用研究. 自动化学报, 2009, **35**(1): 107−112)

8 Tsang I W, Kwok J T, Cheung P. Core vector machines: fast SVM training on very large data sets. *Journal of Machine Learning Research*, 2005, **6**: 363−392

9 Tsang I W, Kwok J T, Zurada J M. Generalized core vector machines. *IEEE Transactions on Neural Networks*, 2006, **17**(5): 1126−1140

10 Deng Z H, Chung F L, Wang S T. FRSDE: fast reduced set density estimator using minimal enclosing ball approximation. *Pattern Recognition*, 2008, **41**(4): 1363−1372

11 Chung F L, Deng Z H, Wang S T. From minimum enclosing ball to fast fuzzy inference system training on large datasets. *IEEE Transactions on Fuzzy Systems*, 2009, **17**(1): 173−184

12 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, **290**(5500): 2323−2326

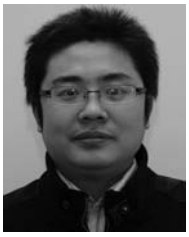
13 Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

14 Collobert R, Bengio S, Bengio Y. A parallel mixture of SVMs for very large scale problems. *Neural Computation*, 2002, **14**(5): 1105−1114

15 Frank A, Asuncion A. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml, May 3, 2010

16 He X F, Cai D, Niyogi P. Laplacian score for feature selection. In: Proceedings of the 18th Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2006. 507−514

17 Hull J J. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, **16**(5): 550−554



胡文军 江南大学信息工程学院博士研究生. 主要研究方向为模式识别和人工智能. 本文通信作者.
E-mail: hoowenjun@yahoo.com.cn
(**HU Wen-Jun** Ph. D. candidate at the School of Information Engineering, Jiangnan University. His research interest covers pattern recognition and artificial intelligence. Corresponding author of this paper.)

王士同 江南大学信息工程学院教授. 主要研究方向为人工智能、模式识别、模糊系统、医学图像处理和生物信息学.
E-mail: wxwangst@yahoo.com.cn
(**WANG Shi-Tong** Professor at the School of Information Engineering, Jiangnan University. His research interest covers artificial intelligence, pattern recognition, fuzzy system, medical image processing and bioinformation.)