第 34 卷 第 9 期 自 动 化 学 报 Vol. 34, No. 9
2008 年 9 月 ACTA AUTOMATICA SINICA September, 2008

一种改善支撑向量域描述性能的核优化算法

赵 峰 [1,2] 张军英 [1] 刘 敬 [1]

摘 要 支撑向量域描述 (Support vector domain description, SVDD) 是一种重要的数据描述算法, 其性能受核参数的影响很大. 基于最优核参数应导致特征空间中映射数据的分布是一个超球形区域的思想, 提出一种核参数优化算法. 首先, 基于训练样本在特征空间所张成的子空间的一组标准正交基, 给出一种描述映射数据分布的方法, 回避了映射数据不可表示的难题; 其次, 基于最大熵原则的非高斯性测度, 构造了一个估计数据分布逼近超球形区域程度的判别准则, 用以确定最优核参数. 基于仿真数据与实测数据的实验验证了本文方法的有效性.

关键词 支撑向量域描述, 核函数, 非高斯性测度
中图分类号 TP391.4

An Optimizing Kernel Algorithm for Improving the Performance of Support Vector Domain Description

ZHAO Feng[1,2] ZHANG Jun-Ying[1] LIU Jing[1]

Abstract The support vector domain description (SVDD) is a robust data domain description method. Its performance, however, is strongly influenced by kernel parameter. In this paper, a novel parameter-optimizing algorithm is presented based on the idea that the optimal parameter can lead to the distribution of the mapped data in the feature space to be a hyper-sphere shape. Firstly, based on an orthogonal basis of the subspace spanned by the mapped data, a way is given to capture the structure of the entire mapped data, which avoids the problem that the mapped data cannot be expressed in an explicit form. Secondly, based on the maximum-entropy non-Gaussian measurement, a new criterion is presented for estimating the degree for a distribution to be closes to the hyper-sphere area and it is used to select the suitable kernel parameter. The experiments on simulated data and real-world data demonstrate the effectiveness of the proposed method.

Key words Support vector domain description (SVDD), kernel function, non-Gaussian measurement

支撑向量域描述 (Support vector domain description, SVDD)[1−2] 是在支撑向量机 (Support vector machine, SVM)[3] 基础上提出的一种数据描述算法, 有着很好的潜在应用前景. 目前, 它已经在机械故障检测、语音识别、人脸检测等方面有许多成功的应用[4−6]. SVDD 的主要思想是寻求一个最小超球, 尽可能把所有训练样本包围起来, 并以最小超球的边界作为界限对数据进行分类和描述. 类似 SVM, 由于核函数的介入, SVDD 描述数据的方式更加灵活, 但其性能受核函数的影响很大, 若核参数选择不当, SVDD 描述数据的效果将变得不理想.

由于在特征空间中, 映射数据以隐式形式存在, 其分布情况无法直接估计, 所以还没有非常有效的办法解决 SVDD 的核优化问题. 目前, 交叉验证是常用的核参数选择方法[2,7], 但该方法耗时严重, 而且是经验上的最优值, 优化标准在不同应用目的中也不尽相同.

基于上述问题, 本文从映射数据的空间结构出发, 提出一种优化核参数的新思路. 事实上, 理想的核参数应满足: 通过核函数对应的非线性映射, 训练数据被映射到特征空间, 其对应的映射数据在特征空间中所占据的区域, 应该是一个超球形区域. 换句话说, 映射数据的分布应该逼近超球形. 但映射数据以隐式形式出现, 所以如何分析映射数据在特征空间的结构以及如何描述数据分布逼近超球形区域的程度, 是新思路是否可行的关键.

针对新思路所遇到的问题, 本文给出了如下解决方案: 1) 基于线性相关性, 得到映射数据在特征空间所张成的子空间的一组标准正交基. 将所有映射数据向该组基上作投影, 则投影向量 (即投影坐标所构成的向量) 与映射数据的分布性质相同. 因此, 可以通过分析投影向量的性质, 来了解映射数据的分布. 而且投影向量完全由核函数确定, 无需明确非线性映射的具体形式. 2) 基于最大熵原则的非高斯测

收稿日期 2007-06-29 收修改稿日期 2007-12-11
Received June 29, 2007; in revised form December 11, 2007
国家自然科学基金 (60574039, 60371044), 总装预研项目 (413070501) 资助
Supported by National Natural Science Foundation of China (60574039, 60371044) and National Defense Advanced Research Foundation of China (413070501)
1. 西安电子科技大学计算机学院 西安 710071 2. 山东工商学院信息与电子工程学院 烟台 264005
1. School of Computer Science and Engineering, Xidian University, Xi′an 710071 2. School of Information and Electronics Engineering, Shandong Institute of Business and Technology, Yantai 264005
DOI: 10.3724/SP.J.1004.2008.01122

度[8], 构造了一个判别准则函数, 用以估计一个分布逼近超球形区域的程度. 按照这个策略, 最优的核参数应导致判别准则函数值最小. 实验验证了本文算法的有效性.

1 支撑向量域描述简述

设 $X=\{\boldsymbol{x}_i\}_{i=1}^n$ 为 $\mathbf{R}^d$ 空间中的一个样本集, 其中 n 表示样本数目. 通过一个非线性映射 ϕ 将样本映射到特征空间 F 中, 即 $\phi:\boldsymbol{x}_i\in X\subset\mathbf{R}^d\to\phi(\boldsymbol{x}_i)\in F$ (简述起见, 称 $\phi(\boldsymbol{x}_i)$ 为 $\boldsymbol{x}_i$ 对应的核样本). SVDD 的基本思想是: 在 F 中寻求一个体积最小的超球 $\Omega=(\boldsymbol{a},R)$, 使得核样本集 $\overline{X}=\{\phi(\boldsymbol{x}_i)\}_{i=1}^n$ 中的数据全部或尽可能多地包容在 Ω 中. 其中 $\boldsymbol{a}$ 表示球心, R 表示半径. 最小化超球体的体积是一个二次规划问题, 即

$$\min_{R,\boldsymbol{a},\varepsilon_i}(R^2+C\sum_{i=1}^n\varepsilon_i) \tag{1}$$

约束条件为

$$\|\phi(\boldsymbol{x}_i)-\boldsymbol{a}\|^2\le R^2+\varepsilon_i,\ i=1,2,\cdots,n \tag{2}$$

其中 $\varepsilon_i\ge 0$ 表示松弛变量, 以便把奇异点排除在超球体外. C 是一个指定的常数, 起到控制对错分样本惩罚程度的作用, 以实现错分样本比例与算法复杂程度之间的折中. 该优化问题的解可由下面的拉格朗日 (Lagrange) 泛函的鞍点给出

$$L(R,\boldsymbol{a},\varepsilon_i,\alpha_i,\lambda_i)=R^2+C\sum_{i=1}^n\varepsilon_i-\sum_{i=1}^n\alpha_i(R^2+\lambda_i-\|\phi(\boldsymbol{x}_i)-\boldsymbol{a}\|_2)-\sum_{i=1}^n\lambda_i\varepsilon_i \tag{3}$$

其中, $\alpha_i\ge 0$, $\lambda_i\ge 0$. 求式 (3) 的最小值可变成求其 Wolf 对偶问题的最大值

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

其中, $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$, α 的约束是 $\sum_{i=1}^n\alpha_i=1$ 和 $0\le\alpha_i\le C$. $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示核函数, 用它替代内积运算, 即 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=<\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)>$.

对于一个新样本 $\boldsymbol{x}$ 如果满足下列条件, 则接受它属于目标样本, 否则就拒绝.

$$f(\boldsymbol{x})=k(\boldsymbol{x},\boldsymbol{x})-2\sum_{i=1}^n\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x})+\sum_{i=1}^n\sum_{j=1}^n\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\le R^2 \tag{5}$$

在实际计算中, 多数 $\alpha_i=0$, 只有少数 $\alpha_i\ne 0$. 称 $\alpha_i\ne 0$ 对应的样本为支撑向量 (Support vector, SV), 只有这些支撑向量才决定了圆心 $\boldsymbol{a}$ 与半径 R 的值.

2 子空间正交基的确定

由于非线性映射 ϕ 没有明确给出, 无法直接分析数据集 $\overline{X}=\{\phi(\boldsymbol{x}_i)\}_{i=1}^n$ 在特征空间 F 中的分布情况. 因此本文采用下面的思路来了解 $\overline{X}$ 的分布情况: 令 H 表示 $\phi(\boldsymbol{x}_1)$, $\phi(\boldsymbol{x}_2)$, $\cdots,\phi(\boldsymbol{x}_n)$ 所张成的子空间, $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{br})$ 是 H 的一组标准正交基, $\boldsymbol{y}_i$ 表示 $\phi(\boldsymbol{x}_i)$ 在基 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{br})$ 下的投影向量. 从几何的角度, 数据集 $Y=\{\boldsymbol{y}_i\}_{i=1}^n$ 与 $\overline{X}$ 具有相同的空间结构, 即它们的分布相同. 因此了解 $\overline{X}$ 的分布等价于分析 $Y=\{\boldsymbol{y}_i\}_{i=1}^n$ 的分布. 下面具体讨论如何寻找 H 的一组标准正交基.

定理 1. 设 $\boldsymbol{x}_{b1}$, $\boldsymbol{x}_{b2},\cdots,\boldsymbol{x}_{bs}$ 表示 s 个训练样本, 其对应的核样本为 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs})$, 核矩阵为 $K_{ss}=(k(\boldsymbol{x}_{bi},\boldsymbol{x}_{bj}))_{1\le i,j\le s}$. 则 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs})$ 线性无关 $\Leftrightarrow\det(K_{ss})\ne 0$.

证明. 令 $A=(\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs}))$, 则 $K_{ss}=A^{\mathrm{T}}A$, 故 $\mathrm{rank}(K_{ss})=\mathrm{rank}(A)$. 又 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs})$ 线性无关 $\Leftrightarrow\mathrm{rank}(A)=s$ 即 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs})$ 线性无关 $\Leftrightarrow\mathrm{rank}(K_{ss})=s$, 即 $\det(K_{ss})\ne 0$. □

基于定理 1, 可以采用如下迭代算法寻找 H 的一组基: 假设利用 t ($t\le n$) 个样本 $\{\boldsymbol{x}_i\}_{i=1}^t$ 完成了训练, 得到子空间 $\{\phi(\boldsymbol{x}_i)\}_{i=1}^t$ 的一组基 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs})$. 对一个新样本 $\boldsymbol{x}$, 依据定理 1 判定 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs}),\phi(\boldsymbol{x})$ 是否线性无关? 如果无关, 则令 $\phi(\boldsymbol{x}_{b(s+1)})=\phi(\boldsymbol{x})$, 构成一个新的线性无关组 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{bs}),\phi(\boldsymbol{x}_{b(s+1)})$. 当遍历所有训练样本时, 所求的线性无关组 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2}),\cdots,\phi(\boldsymbol{x}_{br})$ 即为 H 的一组基. 具体算法步骤如下:

算法 1. 寻找子空间基的迭代算法

Step 1. 初始化: 在训练集 $X=\{\boldsymbol{x}_i\}_{i=1}^n$ 任选一个样本 $\boldsymbol{x}$, 满足 $k(\boldsymbol{x},\boldsymbol{x})\ne 0$. 令 $S=\{\boldsymbol{x}\}$, $D=\{\boldsymbol{x}\}$, $t=1$.

Step 2. 选择一个样本 $\boldsymbol{x}^*\in\bar{S}=X-S$. 令 $S=S\cup\{\boldsymbol{x}^*\}$, 计算核矩阵 $K_{ss}=(k(\boldsymbol{x},\boldsymbol{y}))_{\boldsymbol{x},\boldsymbol{y}\in S}$, 并验证下式是否成立

$$\det(K_{ss})=0 \tag{6}$$

Step 3. 如果式 (6) 成立, 令 $t=t+1$. 否则令 $D=D\cup\{\boldsymbol{x}^*\}$, $t=t+1$.

Step 4. 如果 $t=n$, 程序中止, 输出 D; 否则,

返回 Step 2.

程序终止时, 集合 D 中的向量所对应的核样本即为 H 的一组基. 不妨记为 $\phi(\boldsymbol{x}_{b1})$, $\phi(\boldsymbol{x}_{b2})$, $\cdots$, $\phi(\boldsymbol{x}_{br})$. 下面将其标准正交化.

定理 2. 设子空间 H 的一组基为 $\boldsymbol{x}_{b1}$, $\boldsymbol{x}_{b2}$, $\cdots$, $\boldsymbol{x}_{br}$, 则 H 的一组标准正交基为

$$(\beta_1, \beta_2, \cdots, \beta_r) = (\phi(\boldsymbol{x}_{b1}), \phi(\boldsymbol{x}_{b2}), \cdots, \phi(\boldsymbol{x}_{br}))\boldsymbol{V} \tag{7}$$

其中, $\boldsymbol{V} = (u_1/\sqrt{\lambda_1}, u_2/\sqrt{\lambda_2}, \cdots, u_r/\sqrt{\lambda_r})$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\cdots$, $\boldsymbol{u}_r$ 为核矩阵 $K_{rr} = (k(\boldsymbol{x}_{bi}, \boldsymbol{x}_{bj}))_{1\le i,j\le r}$ 的一组标准正交特征向量, $\lambda_1, \lambda_2, \cdots, \lambda_r$ 为对应的特征值.

证明. 因为 $K_{rr} = (\phi(\boldsymbol{x}_{b1}), \phi(\boldsymbol{x}_{b2}), \cdots, \phi(\boldsymbol{x}_{br}))^{\mathrm{T}} \times (\phi(\boldsymbol{x}_{b1}), \phi(\boldsymbol{x}_{b2}), \cdots, \phi(\boldsymbol{x}_{br}))$, 显然其为正定矩阵. 令 $\boldsymbol{U} = (u_1, u_2, \cdots, u_r)$, $\Lambda_1 = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$, 则 $\boldsymbol{U}^{\mathrm{T}} K_{rr} \boldsymbol{U} = \Lambda_1$. 令 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_r)$, $\Lambda_2 = \mathrm{diag}\{1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \cdots, 1/\sqrt{\lambda_r}\}$, 则 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = \boldsymbol{V}^{\mathrm{T}} K_{rr} \boldsymbol{V} = \Lambda_2 \boldsymbol{U}^{\mathrm{T}} K_{rr} \boldsymbol{U} \Lambda_2 = I$. □

由式 (7), $\phi(\boldsymbol{x}_i)$ $(1 \le i \le n)$ 在正交基 $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\cdots$, $\boldsymbol{\beta}_r$ 的投影向量为

$$\begin{aligned}\boldsymbol{y}_i =& (\beta_1, \beta_2, \cdots, \beta_r)^{\mathrm{T}} \phi(\boldsymbol{x}_i) = \\ & \boldsymbol{V}^{\mathrm{T}} (k(\boldsymbol{x}_{b1}, \boldsymbol{x}_i), k(\boldsymbol{x}_{b2}, \boldsymbol{x}_i), \cdots, k(\boldsymbol{x}_{br}, \boldsymbol{x}_i))^{\mathrm{T}}\end{aligned} \tag{8}$$

由式 (8) 可以看出, 投影向量 $\boldsymbol{y}_i$ 的表示只用到相应的核函数, 无需明确具体的非线性映射形式.

3 基于非高斯性测度的核参数优化算法

由第 2 节知, 评价数据集 $\overline{X} = \{\phi(\boldsymbol{x}_i)\}_{i=1}^n$ 的分布逼近超球形区域的程度, 可以转化为考察 $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ 逼近超球形区域的程度. 目前, 还没有一个有效的测度来定量地描述数据逼近超球形区域的程度. 因此, 本文提出一个新的想法: 如果一个多维变量 $\boldsymbol{y}$ 的各个分量独立且服从同方差的高斯分布, 显然 $\boldsymbol{y}$ 的分布是一个超球形区域. 进一步, 本文基于最大熵原则的非高斯性测度, 构造了一个用以度量多维随机变量逼近高斯分布程度的测度, 来反映该数据逼近超球形区域的程度. 需要强调的是: 尽管并不是所有超球形区域的分布都满足上述特定的高斯分布, 但总可以通过调节核参数, 使得数据的分布尽可能地满足上述条件, 达到逼近超球形区域的目的.

3.1 基于最大熵原则的多维变量非高斯性测度

文献 [8] 介绍了一种基于最大熵原则的非高斯性测度, 用于测量一维随机变量的非高斯性, 并证明了该测度比传统的基于累计量的测度[9] 精确得多, 其定义如下

$$J(z) = [\mathrm{E}\{g(z)\} - \mathrm{E}\{g(\nu)\}]^2 \tag{9}$$

其中, ν 是标准的高斯变量, 随机变量 z 假定具有零均值、单位方差. g 是非二次函数, 选取方法见文献 [8].

式 (9) 只能用于一维变量的非高斯性测度, 为此本文提出一种多维的非高斯测度. 假定 $\boldsymbol{\nu}$ 是一个 N 维的高斯随机变量, 每维服从零均值, 单位方差的高斯分布. 假定 $\boldsymbol{z}$ 是一个 N 维随机变量, 每维具有零均值, 单位方差. 定义测度如下

$$J(\boldsymbol{z}) = [\mathrm{E}\{g(\|\boldsymbol{z}\|)\} - \mathrm{E}\{g(\|\boldsymbol{\nu}\|)\}]^2 \tag{10}$$

显然, 仅当 $\boldsymbol{z}$ 也是一个 N 维的高斯分布时, 式 (10) 才有最小值零. 即 $\boldsymbol{z}$ 越逼近高斯分布, $J(\boldsymbol{z})$ 的值越小.

3.2 核参数优化测度

由前面的分析可知, 理想的核参数应导致数据集 $\overline{X} = \{\phi(\boldsymbol{x}_i)\}_{i=1}^n$ 的分布逼近超球形区域, 即其对应的投影数据 $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ 逼近超球形区域. 显然, 如果 $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ 的分布是高斯分布, 且每个分量独立、同方差, 则分布一定是超球形区域. 为此, 最优的核参数 $\hat{\theta}$ 定义如下: 最优核参数 $\hat{\theta}$, 应导致数据集 $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ 最逼近一种特定的高斯分布 (该高斯分布的各个分量独立、同方差).

一般情况下, 核参数 θ 不同, 对应的映射 ϕ 不同, 最终得到的数据集 $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ 的维数也不同. 为使不同维数的数据集逼近上述特定高斯分布的测度具有可比性, 根据式 (10), 本文定义核优化测度如下

$$J(\boldsymbol{z}) = [|\mathrm{E}\{g(\|\boldsymbol{z}\|)\}|^{\frac{1}{N}} - |\mathrm{E}\{g(\|\boldsymbol{\nu}\|)\}|^{\frac{1}{N}}]^2 \tag{11}$$

其中 $\boldsymbol{z}$ 和 $\boldsymbol{\nu}$ 的定义与式 (10) 相同, 只是非二次函数选用 $g(t) = -\exp(-t^2/2)$, 这是因为在这个非二次函数下, 如果 $\boldsymbol{z}$ 是一个 N 维变量, 则 $\mathrm{E}\{g(\|\boldsymbol{z}\|)\} = \prod_{j=1}^N \mathrm{E}\{g(z_j)\}$, 其中 z_j 表示 $\boldsymbol{z}$ 的第 j 个分量, 显然, 经过式 (11) 处理, 维数不同的变量的测度具有可比性. 其中当 $\boldsymbol{z}$ 是高斯分布时, 式 (11) 有最小值零. 即 $\boldsymbol{z}$ 越逼近高斯分布, $J(\boldsymbol{z})$ 的值越小.

最后, 给出最优化核参数算法的主要步骤:

算法 2. 最优化核参数算法

Step 1. 对于给定的核参数 θ, 求解训练样本集 $X = \{\boldsymbol{x}_i\}_{i=1}^n$ 所对应的数据集 $Y_\theta = \{\boldsymbol{y}_i\}_{i=1}^n$. 具体求解见第 2 节.

Step 2. 将数据 $\boldsymbol{y}_i$ $(i = 1, 2, \cdots, n)$ 按照如下方式处理: $\boldsymbol{z}_i = (\boldsymbol{y}_i - \overline{\boldsymbol{y}})/\sigma$. 其中 $\overline{\boldsymbol{y}}$ 是 $Y_\theta = \{\boldsymbol{y}_i\}_{i=1}^n$ 的均值, σ 则是各个分量方差的平均. 因此, 如果 $Y_\theta = \{\boldsymbol{y}_i\}_{i=1}^n$ 是高斯分布, 且各分量独立、具有相同的方差, 则 $\boldsymbol{z}_i = (\boldsymbol{y}_i - \overline{\boldsymbol{y}})/\sigma$ 服从均值为零、单位方差的高斯分布.

Step 3. 用式 (11) 度量 $\boldsymbol{z}_i$ $(i = 1, 2, \cdots, n)$ 的非高斯性程度 J_θ.

Step 4. 按照下式确定最优的核参数

$$\hat{\theta} = \min_{\theta} J_\theta \tag{12}$$

4 实验

4.1 数据的空间分布对 SVDD 性能的影响

为了直观了解数据的空间结构对 SVDD 性能的影响, 本文构造了 3 组 2 维数据集, 采用线性 SVDD 描述它们的边界. 3 组数据集的产生方式为: 1) DATA1: x 服从 $[-1, 1]$ 上的均匀分布, $y = x^2 + \xi$, 其中 ξ 是服从均值为 0, 方差为 0.2 的高斯分布; 2) DATA2: x 服从 0 均值, 方差为 0.1 的高斯分布, y 服从 0 均值, 方差为 0.5 的高斯分布; 3) DATA3: x 与 y 均服从 0 均值, 方差为 0.1 的高斯分布. 图 1 显示了这些数据集的散点分布图以及线性 SVDD 的描述边界 (图中 "·" 表示数据点, 实曲线表示描述边界, J 表示数据集的非高斯性测度 (见式 (11))).

由图 1 可以看出: 1) 数据集的空间结构直接决定了 SVDD 描述边界的效果. 当数据集的分布不是球形区域时 (如 DATA1 是一个 "弧形区域", DATA2 是一个 "椭圆形区域"), 采用 SVDD 表述数据边界的效果很差. 原因在于 SVDD 的出发点是: 用一个半径最小的球体尽可能地包容目标数据. 因此如果数据的分布不是球形区域, 而用一个球面去描述其边界, 效果势必很差. 反之, 当数据集是一个球形区域时 (如 DATA3), SVDD 描述边界的效果比较理想. 2) 本文所提出的核参数优化测度, 可以很好地反映数据逼近球形区域的程度. 如 DATA3, 该数据的两个分量相互独立, 服从相同方差的高斯分布, 其非高斯性测度 J 最小. 3) 该实验也间接地说明了优化核参数的取向: 对于输入空间中不是超球形区域的数据, 通过特定的非线性变换 (该变换取决于核函数的形式及核参数的取值), 其映射数据的分布应该逼近超球形. 只有在这种情形下, SVDD 描述边界的效果才会理想.

4.2 核参数对 SVDD 性能的影响

为了能从输入空间直观地分析核参数对 SVDD 性能的影响, 我们以数据集 DATA1 (见第 4.1 节) 为描述对象, 研究核参数变化时, SVDD 对数据的描述效果. 实验中核函数形式采用高斯形式 $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$. 图 2 (见下页) 显示了核参数 σ 取不同值时, SVDD 所生成的描述边界. 其中 SV 表示相应的支撑向量数目, 其他符号含义同图 1. 由图 2 可以看出: 1) 不同核参数下的数据描述边界变化很大. 一般情况下, 就高斯核而言, σ 过大, 支撑向量数目愈少, 区域边界变得越松弛, 许多非目标区域将被接受. σ 过小, 支撑向量数目愈多, 区域边界过于紧凑, 存在 "过适应" 问题. 说明核参数的变化对 SVDD 的性能影响很大. 2) 本文所提出的核参数优化测度, 可以较好地指导核参数的选择. 如图 2 所示, 当 $\sigma = 0.8$ 时, SVDD 所描述的边界, 最能反映目标数据的特点, 此时其所对应的非高斯性测度也最小.

4.3 基于实测数据的核参数优化

从模式分类的角度考虑, SVDD 是一种单值分类算法. 即 SVDD 训练阶段仅需要目标类数据, 得到目标类的描述边界. 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. 下面采用本文所提的核参数优化算法, 来改善 SVDD 的识别性能.

本文采用 MIT 的 CBCL 人脸库[10] 以及 UCI 数据库[11] 中的 2 个数据集: IRIS 和 YEAST 进行实验. 其中 CBCL 数据集的每个样本均为 19×19

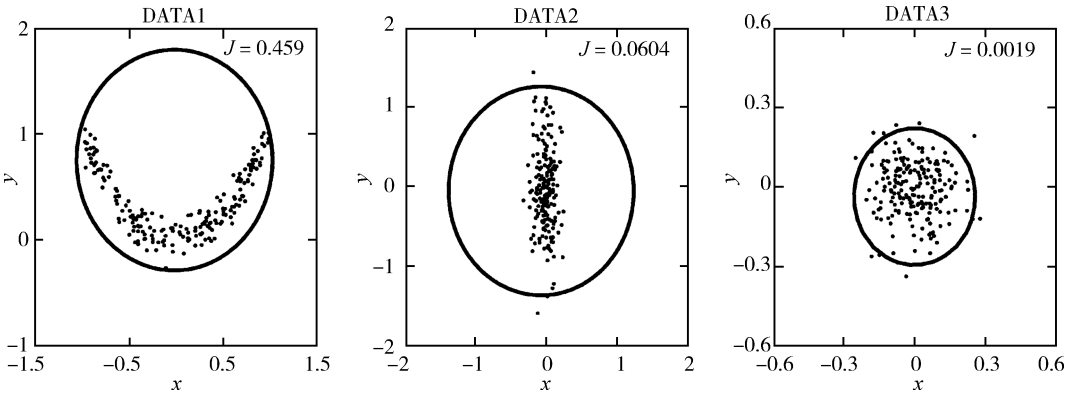


图 1　线性 SVDD 的描述边界

Fig. 1　The description of data boundary using linear support vector domaina description

的标准图像, 实验中选取 400 个人脸样本用于训练, 测试中使用人脸样本 100 个, 非人脸样本 100 个. IRIS 数据源自 3 种类型的鸢尾属植物, 每类 50 个样本, 共 150 个样本. 数据特征为 4 维, 分别表示植物萼片长度、宽度以及花瓣长度、宽度. YESAT 是 3 类真实的真核生物酵母菌基因数据, 共 277 个样本. 每个样本含有 17 个基因的表达水平, 各类分别有 67、135、75 个样本. 显然, IRIS 与 YEAST 不是单分类问题. 本文把每个数据集的第 1 类看作目标类, 其他类看作非目标类, 转化为单分类问题. 在 IRIS 中, 选取第 1 类中的 30 个样本用于训练, 测试则使用第 1 类剩余的 20 个样本和第 2、3 类中各 10 个样本, 共 40 个测试样本. 在 YEAST 中, 选取第 1 类中 40 个样本用于训练, 选取第 1 类中的 20 个样本和第 2、3 类中各 10 个样本, 共 40 个测试样本. 实验中核函数采用高斯核. 图 3 描绘了核优化测度 J 随参数 σ 变化的曲线. 表 1 显示了不同核参数下的识别结果.

由表 1 可以看出: 在不同的核参数下, 实验的识别效果有很大差异. 总的来说, 核参数优化测度 J 最

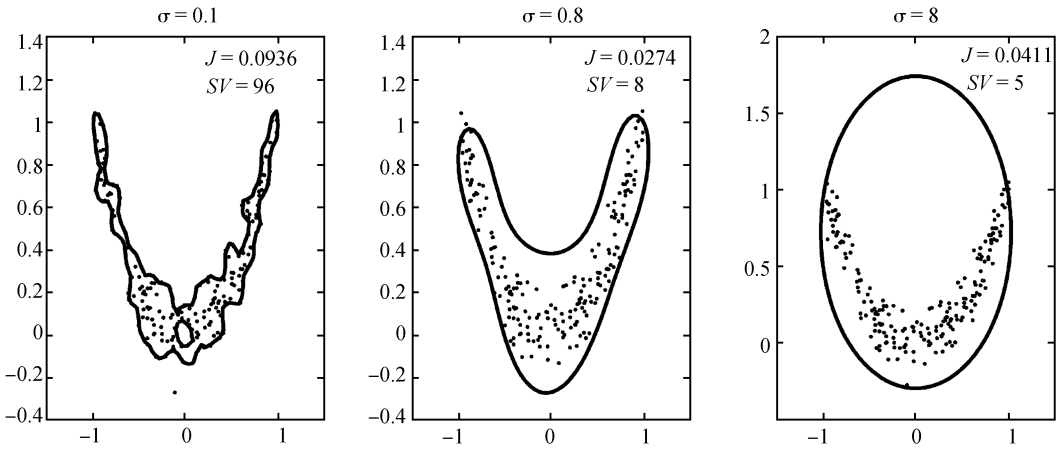


图 2 核参数变化时 SVDD 的描述边界

Fig. 2 The description boundary of SVDD under different kernel parameters

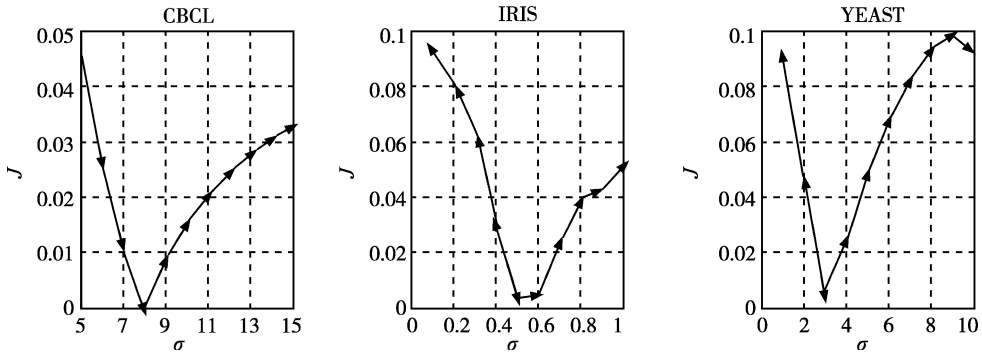


图 3 核优化测度 J 随参数 σ 变化的曲线

Fig. 3 The curve of the kernel optimizing measure J varying with parameter σ

表 1 不同核参数下训练数据与测试数据的识别率 (%)

Table 1 The recognition rates of training and testing data under different kernel parameters

CBCL				IRIS				YEAST			
σ	J	训练	测试	σ	J	训练	测试	σ	J	训练	测试
5	0.0463	95	83	0.1	0.945	88	93	1	0.919	87	70
7	0.0113	99	90	0.5	0.0128	1	93	3	0.0056	93	87
8	0	99	93	0.6	0.014	95	95	5	0.0487	95	83
11	0.0205	97	90	0.8	0.0458	93	95	7	0.0823	92	81
15	0.0328	1	87	1	0.566	88	95	10	0.0935	92	74

小时, 数据的识别效果 (从训练数据与测试数据的识别率综合考查) 最优或者接近最优, 证明了本文所提的算法可以用来指导核参数的选择. 同时由图 3 可以看出: σ 的变化, 直接影响了映射数据的空间结构变化, 进而导致其逼近超球形区域的程度有很大差异, 其对应的非高斯性测度 J 也不相同. J 值较小的, 其对应的识别效果最优, 说明本文所构造的核参数优化测度, 能够比较准确描述数据的分布逼近特定超球形区域的程度 (所谓特定超球形区域, 即数据的各个分量独立, 服从同方差的高斯分布); 也说明当映射数据的分布逼近超球形区域时, SVDD 描述边界的效果较好.

5 结论

针对 SVDD 的核参数调节问题, 本文首次提出从映射数据在特征空间的几何结构考虑核参数的选择问题. 基于这个思路, 本文做了如下两个工作: 1) 基于映射数据在特征空间所张成的子空间的一组标准正交基, 将映射数据的分布研究转化为投影坐标向量的分布研究, 并给出详细的求解过程, 解决了映射数据不可直接表达的条件下, 如何了解其分布信息的问题. 2) 基于最大熵原则下的非高斯性测度, 构造了一个评价数据分布逼近超球形区域程度的判决准则, 并用该准则指导核参数的优化问题.

本文算法的设计思路完全可以推广到其他核方法的参数调节问题. 特别是基于特征空间中一组标准正交基, 将映射数据转化为显式形式来研究, 为了解特征空间的数据结构提供了一种新思路. 但需要指出的是, 本文所给出的核参数优化测度, 本质上是描述数据分布逼近一种特殊超球形区域的程度 (数据分布服从高斯分布, 且各分量独立、同方差). 如何全面描述球形区域的分布特点, 有待进一步研究和探讨.

References

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

2 Lee K Y, Kim D W, Lee D, Lee K H. Improving support vector data description using local density degree. *Pattern Recognition*, 2005, **38**(10): 1768−1771

3 Vapnik V N. *The Nature of Statistical Learning Theory*. New York: John Wiley and Sons, 2000

4 Xin D, Wu Z H, Zhang W F. Support vector domain description for speaker recognition. In: Proceedings of the 2001 IEEE Signal Processing Society Workshop on Neutral Networks for Signal Processing XI. North Falmouth, USA: IEEE, 2001. 481−488

5 Seo J, Ko H. Face detection using support vector domain description in color images. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, USA: IEEE, 2004. 729−732

6 Ypma A, Tax D M J, Duin R P W. Robust machine fault detection with independent component analysis and support vector data description. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop on Neutral Networks for Signal Processing IX. Madison, USA: IEEE, 1999. 67−76

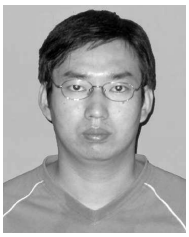
7 Ban T, Abe S. Implementing multi-class classifiers by one-class classification methods. In: Proceedings of IEEE International Joint Conference on Neural Networks. Vancouver B C, Canada: IEEE, 2006. 327−332

8 Hyvarinen A. New approximations of differential entropy for independent component analysis and projection pursuit. In: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10. Cambridge, USA: MIT Press, 1998. 273−279

9 Amari S, Cichocki A, Yang H H. A new learning algorithm for blind source separation. In: Proceedings of Advances in Neural Information Processing System. Cambridge, USA: MIT Press, 1996. 757−763

10 MIT Center for Biological and Computation Learning. CBCL face database [Online], available: http://www.ai.mit.edu/projects/cbcl, January 6, 2007

11 Asuncion A, Newman D J. Machine learning repository [Online], available: http://mlearn.ics.uci.edu/MLRepository.html, January 6, 2007



赵 峰 西安电子科技大学计算机学院博士研究生. 主要研究方向为模式识别. 本文通信作者.
E-mail: zhaofeng1016@126.com
(**ZHAO Feng** Ph. D. candidate at the School of Computer Science and Engineering, Xidian University. His main research interest is pattern recognition. Corresponding author of this paper.)

张军英 西安电子科技大学计算机学院教授. 主要研究方向为模式识别.
E-mail: jyzhang@xidian.edu.cn
(**ZHANG Jun-Ying** Professor, Ph.D. at the School of Computer Science and Technology, Xidian University. Her main research interest is pattern recognition.)

刘 敬 西安电子科技大学计算机学院博士研究生. 主要研究方向为模式识别.
E-mail: liujing@126.com
(**LIU Jing** Ph. D. candidate at the School of Computer Science and Engineering, Xidian University. Her main research interest is pattern recognition.)