

嵌入数据结构信息的单类支持向量机及其线性规划算法^{*}

冯爱民¹, 刘学军¹, 孙廷凯²

(1. 南京航空航天大学信息科学与技术学院, 江苏 南京 210016;
2. 南京理工大学计算机科学与技术学院, 江苏 南京 210094)

摘 要: 针对现有基于超平面的单类分类器未同时考虑目标数据全局与局部信息的不足, 通过在单类支持向量机 One-Class SVM (OCSVM) 算法中加入类内散度以反应目标数据的全局信息, 提出了结构化单类支持向量机 Structured OCSVM (SOCSVM), 不仅使之具有全局与局部化学习的特点, 同时也为诸多的 SVM 算法嵌入数据内在结构这类先验信息提供了统一框架。为进一步提高运算效率, 在 SOCSVM 二次规划求解基础上, 通过最小化目标数据均值到超平面的函数距离, 提出了线性规划算法, 同时也避免了 SOCSVM 必须以原点作为负类代表的不足。人工和真实数据集上的实验结果验证了嵌入目标数据结构信息的 SOCSVM 及其线性规划算法的有效性。

关键词: 单类分类器; 支持向量机; 结构信息; 二次规划; 线性规划

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0529-6579 (2009) 06-0010-08

Embedding Target Data's Structural Distribution Information into One-Class SVM and Its Linear Programming Algorithm

FENG Aimin¹, LIU Xuejun¹, SUN Tingkai²

(1. Information Science & Technology College,
Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China;
2. Computer Science & Technology College,
Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In order to distinguish the target class from outliers accurately, One-Class Classifier (OCC) should take into account the prior knowledge of the target class. However, One-Class SVM (OCSVM), the state-of-the-art OCC, neglects the data's distribution information while finding the optimal hyperplane. Structured OCSVM (SOCSVM), the novel proposed OCC, alleviates this problem by embedding the within-class scattered matrix of the target data into OCSVM. As a result, SOCSVM not only overcomes the above disadvantage of the OCSVM, but also provides a unified framework for the present SVM algorithms how to consider intrinsic structure of the data. Moreover, to improve the efficiency of SOCSVM, linear programming algorithm called SlpOCSVM is proposed to instead of the quadratic programming solving for SOCSVM. Through minimizing the functional distance of the data's mean to the hyperplane, the optimal hyperplane is attracted automatically to the place of the minimum positive half space without borrowing the origin as a representative of the outlier anymore. The experiment results on toy problem and real data sets demonstrate the advantage of SOCSVM and its linear programming algorithm.

Key words: one-class classifier; support vector machine; structured information; quadratic programming; linear programming

^{*} 收稿日期: 2008-12-25

基金项目: 国家自然科学基金资助项目 (60603029, 60703016, 60803049)

作者简介: 冯爱民 (1971 年生), 女, 副教授; E-mail: amfeng@nuaa.edu.cn

单类问题^[1]是指训练样本中只有一类目标数据, 其他非目标数据因为采集代价过高(如故障诊断, 不可能为了获取大量的非目标数据而人为破坏机器)或数据类型过多而使现有数据无代表性(如人脸检测, 任何非人脸都可作为非目标数据), 不得不放弃获取, 造成样本缺失。其中, 目标数据又称为正常类, 而非目标数据亦称为异常类或负类。针对单类问题所设计的单类分类器, 不同于两类或多类分类器, 主要用于描述数据而非判别数据, 故也称为数据描述。由于单类分类器常被用于识别正常类或拒绝异常类, 因而可用于异常检测, 网络入侵检测, 文本分类或图像抽取等。

单类分类器由于没有负类数据, 因此尽可能挖掘目标数据的先验信息变得尤为重要。然而现有的单类分类器, 尤其是基于边界的方法^[2-3], 都是从经典的两类分类器发展而来的, 而两类分类器主要关心的是数据间的可分性而非描述数据, 因此相应的单类算法性能未必最优。从两类支持向量机 ν -SVM 发展而来的单类支持向量机 One-Class SVM (OCSVM)^[3]便属于此种情形, 作为基于边界方法的典型单类分类器, 它通过最大化原点和目标数据之间的最小欧式距离来寻找最优超平面。然而由于所采用的欧氏度量并未考虑目标数据的结构分布这类全局信息, 因而很可能把正常数据所在的区域给拒绝掉, 导致所获得的超平面很可能是次优解^[4]。

同样采用超平面作为目标数据描述边界的单类最小化最大概率机 One-Class MPM (OCMPM)^[5], 亦是从两类最小化最大概率机^[6]发展而来的单类分类器。它虽然利用了目标数据均值和协方差这类样本分布信息, 然而由于它必须保证满足上述分布的数据在最坏分布情形下的数据描述, 因而忽略了现有数据自身各样本点的局部信息, 导致数据描述边界过松而包含了许多空白区域, 因而很可能把异常数据包含进来, 实验效果并不理想^[5], 且因求解过程无法采用对偶理论而丧失了解的稀疏性^[7]。

上述分析可进一步通过图1 OCSVM 和 OCMPM 所寻找的超平面示意图来说明, 简化起见, 图中各超平面均相互平行, 且为硬间隔情形, 即全体目标数据均落在正半空间, SV ($\circ$) 表示支持向量。从图中可见, 目标数据 ($\bullet$) 呈椭球分布, 由于仅有有限样本, 且抽样的不均匀性, 当前的目标数据无法保证覆盖数据分布所在的全部空间。这就意味着, 对于未见的测试样本 ($\otimes$), 按照数据分布趋势, 有可能落在图中 H_1 (虚线) 与超椭球所包围的空间中, 这里 H_1 是 OCSVM 所寻找的超平面。

显然, 由于 OCSVM 未考虑数据的分布趋势, 超平面 H_1 可能将目标数据所在的部分区域拒绝掉, 导致落在该区域的目标类被错判为异常类而使 False Negative (FN) 错误率增大。而对于 OCMPM, 由于它所寻找的超平面 H_2 (点划线) 必须保证满足当前均值和方差的最坏数据分布情形, 因而忽略了现有各目标数据点的当前位置而使数据描述区域过松。因此, 易使落入超平面 H_2 和椭球间的异常数据被错判为正常类而导致 False Positive (FP) 错误率增大。而更合理的位置, 如图中超平面 H_3 所示, 既包含了 OCSVM 所遗漏的目标数据可能出现的数据区域, 也拒绝了 OCMPM 所包含的异常数据可能出现的空白区域, 为此, 必须考虑目标数据的结构分布等先验信息。

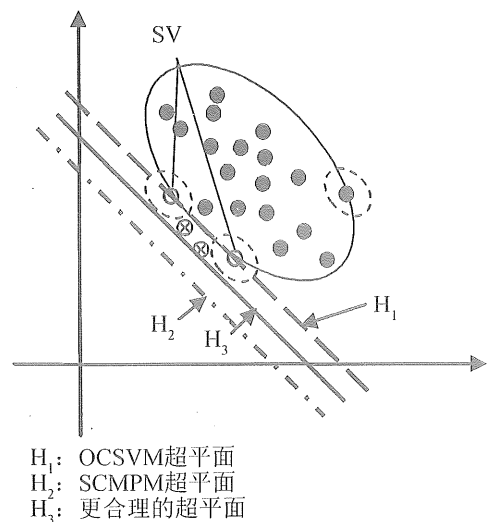


图1 输入空间中 OCSVM/OCMPM 的最优超平面及更合理的超平面示意图

Fig. 1 The optimal hyperplane of OCSVM, OCMPM and a more reasonable hyperplane in input space

受上述分析的启发, 为了得到类似于超平面 H_3 的单类分类器, 要求算法既要如 OCSVM 那样考虑各数据点的局部信息, 同时也应该如 OCMPM 算法那样考虑目标数据的全局分布, 为此, 本文提出了全局与局部化学习兼备的单类分类器, 称为结构化单类支持向量机 Structured OCSVM (SOCSVM)。SOCSVM 通过在 OCSVM 框架中嵌入目标数据的类内散度矩阵, 旨在将数据分布先验信息加入到目标函数当中, 以克服 OCSVM 未考虑全局信息的不足。事实上, 现有众多的 SVM 分类和回归算法中同样存在此不足, 因此上述思想可作为一个通用框架, 为各类 SVM 算法嵌入先验信息提供了一个新

的解决方案, 且求解过程仍保持二次规划 (Quadratic Programming: QP) 不变。

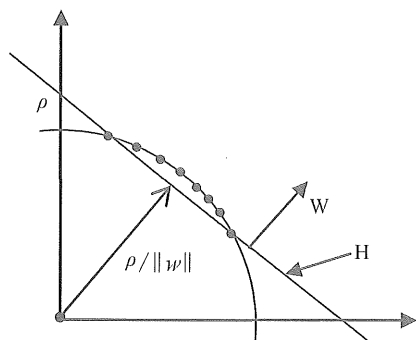


图 2 OCSVM 取高斯核时找到的最优超平面 H

Fig. 2 The optimal hyperplane of OCSVM with the RBF kernel

为进一步提高 SOCSVM 算法的求解速度, 受 OCSVM 线性规划算法 lpOCSVM 的启发^[8], 通过最小化目标数据与超平面之间的平均函数距离, 提出了结构化线性规划单类支持向量机 SlpOCSVM, 取代 QP 求解的同时, 也避免了 SOCSVM 必须采用原点作为负类样本代表的不足。显然, 这里的函数距离, 由于引入了数据结构信息, 已不再是 lpOCSVM 所采用欧式度量^[8]。

本文后续内容是按如下顺序展开的, 第二部分在简述 OCSVM 基础上详细阐述了 SOCSVM 算法, 包括其输入空间和特征空间的刻画形式; 第三部分则描述了 SlpOCSVM 的线性形式和核化算法, 并且比较了 QP 和 LP 算法的时间复杂度; 第四部分以人工和真实数据集上的实验结果来验证算法的推广性能, 结论部分对全文进行总结。

1 嵌入结构信息的单类支持向量机

1.1 单类支持向量机 OCSVM

为沿用 SVM 间隔的思想, OCSVM 巧妙地利用了原点作为异常类的代表以便设计出间隔。通过最大化原点和目标数据之间的最小欧氏距离 $\frac{\rho}{\|w\|}$ 来寻找最优超平面 $w^T x_i - \rho = 0$, 其中 w 是超平面的法向量, ρ 是超平面截距, x_i 为目标数据集 $X = \{x_1, x_2, \dots, x_n\}$ (n 为样本个数) 中位于超平面上的样本点, 称之为支持向量^[9]。为使算法具有一定的鲁棒性, 引入松弛因子 $\xi = [\xi_1, \dots, \xi_n]^T$, 目标函数如下:

$$\min_{w, \xi, \rho} \frac{1}{2} w^T w - \rho + \frac{1}{vn} \sum_{i=1}^n \xi_i \quad (1)$$

$$s. t. \quad w^T x_i \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n$$

其中, $\nu \in (0, 1]$ 是所谓的百分比估计, 和支持向量的数目有密切联系, 即 ν 是边界支持向量的上界, 是全部支持向量的下界, 称为 ν 属性^[3]。

引入向量 $\alpha = [\alpha_1, \dots, \alpha_n]^T, \mathbf{1} = [1, \dots, 1]^T$, 式的对偶形式以矩阵表示为:

$$\min_{\alpha} = \frac{1}{2} \alpha^T X^T X \alpha = \frac{1}{2} \alpha^T K \alpha \quad (2)$$

$$s. t. \quad K: = X^T X, \alpha^T \mathbf{1} = 1, 0 \leq \alpha \leq \frac{1}{vn} \mathbf{1}$$

这是一个 QP 问题, 可采用相应的软件包或者序列最小优化算法^[3]来优化。在式 (2) 中通过引入核函数, 可得到 OCSVM 算法在高维特征空间的核化形式。常用的核函数是高斯核:

$$K(x, y) = e^{-\|x-y\|^2 / \sigma^2} \quad (3)$$

事实上, 这也是 OCSVM 算法之所以能采用原点作为负类代表的原因。尽管在线性空间中, 将原点假设为负类未必可行, 因为我们无法保证目标数据与原点之间一定可分。然而对于高斯核, 由于目标数据在该核的作用下均被映射到一个圆心在原点的单位超球上 ($K(x_i, x_i) = 1$), 且各数据点间因为 $K(x, y) > 0$ 而一定位于同一象限内, 因此原点与目标数据一定可分, 尽量远离原点也就意味着优化目标数据所在的正半空间。如图 2 所示即 2 维空间中目标数据在高斯核作用下所生成的部分单位超球以及 OCSVM 算法通过最小化间隔 $\frac{\rho}{\|w\|}$ 所寻找到的最优超平面 H。

对于测试样本 z , 判别函数如下:

$$f(z) = \text{sgn}[\alpha^T X^T z - \rho] = \begin{cases} 1 & \text{目标类} \\ -1 & \text{异常类} \end{cases} \quad (4)$$

1.2 结构化单类支持向量机 Structured OCSVM

1.2.1 线性形式 从 OCSVM 的描述形式可知其在寻找超平面时仅关注原点和最近数据点的可分性而并忽略了超平面内部的数据分布, 也就是说 OCSVM 仅考虑目标数据和原点之间的类间局部信息而忽略了目标数据内部分布的全局信息。受线性判别分析 LDA 类间尽量分开且类内尽量紧凑思想的启发^[10], 拟在现有 OCSVM 算法基础上通过嵌入表示类内散度的正则化项 $w^T \Sigma w$ 以弥补全局信息的不足, 并称之为结构化单类支持向量机 (Structured OCSVM: SOCSVM), 这里, Σ 为全体目标数据的协方差矩阵。目标函数形式如下:

$$\min_{w, \xi, \rho} \frac{1}{2} w^T w + \frac{\lambda}{2} w^T \Sigma w - \rho + \frac{1}{vn} \sum_{i=1}^n \xi_i \quad (5)$$

$$s. t. \quad w^T x_i \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n$$

其中 λ 是正则化因子, 用于平衡数据的结构信息

和局部信息, 且 $\lambda \geq 0$ 。若 $\lambda = 0$ 时, 全局信息消失, SOCSVM 退化为 OCSVM 算法。

相应对偶形式:

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \alpha^T X^T (I + \lambda \Sigma)^{-1} X \alpha \\ \text{s. t. } & \alpha^T \mathbf{1} = 1, 0 \leq \alpha \leq \frac{1}{vn} \end{aligned} \quad (6)$$

从式 (6) 可看出, SOCSVM 具有如下优点:

(1) 数据描述能力更强: 和 OCSVM 的对偶形式 (2) 比较不难看出, 上式 (6) 寻找超平面的对偶空间已不再是样本内积 $X^T X$ 作用的数据输入原空间, 而是 $X^T (I + \lambda \Sigma)^{-1} X$ 所在的输出空间, 相对于原输入空间, SOCSVM 相当于把样本数据通过 $(I + \lambda \Sigma)^{-\frac{1}{2}}$ 映射到该空间中以寻找最优超平面。若将该超平面映射回输入空间, 显然是非线性的, 因而较之 OCSVM 具有更强的数据描述能力。

(2) 可 QP 求解: 尽管考虑了数据分布, 目标函数仍属 QP 问题, 可直接采用 QP 软件包或单类序列最小优化算法来优化。且该二次规划问题可进一步简化为 LP 问题, 具体参见论文第三部分。

(3) 具有稀疏解: 由于保留了 SVM 的框架, 因此根据解 α 的数值, 仍具有支持向量与非支持向量与之对应, 即对于大量的落在超平面内部正半空间的点, $\alpha = 0$, 为非支持向量; 而对于少量落在超平面及其负半空间的点, $\alpha \neq 0$, 为支持向量, 而正是其中 $0 < \alpha < \frac{1}{vn}$ 的非边界支持向量构建出相应的超平面。

对于待测样本 z , 可通过如下判别函数决定是否属于异常类:

$$f(z) = \text{sgn}[\alpha^T X^T (I + \lambda \Sigma)^{-1} z - \rho] = \begin{cases} 1 & \text{目标类} \\ -1 & \text{异常类} \end{cases} \quad (7)$$

1.2.2 核化形式 简单起见, 仍旧采用线性空间中的目标函数 (5) 及其相应对偶形式 (6) 来推导其核化形式, 显然, 问题的关键是将对偶表达式中 $(I + \lambda \Sigma)^{-1}$ 采用内积形式来表达, 由于协方差 Σ 可表示为:

$$\Sigma = \frac{1}{n} X X^T - \frac{1}{n^2} X \mathbf{1} \mathbf{1}^T X^T = \frac{1}{n} X H X^T \quad (8)$$

式中, $H = (I - \frac{1}{n} \mathbf{1} \mathbf{1}^T)$ 。借助 Woodbury 公式^[11]:

$$(A + BC)^{-1} = A^{-1} - A^{-1} B (I + CA^{-1} B)^{-1} CA^{-1}$$

以及 $HH = H$ 和 $H = H^T$ 的属性, 可得:

$$(I + \lambda \Sigma)^{-1} = I + \frac{\lambda}{n} (X H) (H X^T) =$$

$$I - \frac{\lambda}{n} X H (I + \frac{\lambda}{n} H X^T X H)^{-1} H X^T \quad (9)$$

带入式 (6) 且引入核技巧 $K = X^T X$, 可得:

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \alpha^T \hat{K} \alpha \\ \text{s. t. } & \hat{K} = (K - \frac{\lambda}{n} K H (I + \frac{\lambda}{n} H K H)^{-1} H K) \\ & \alpha^T \mathbf{1} = 1, 0 \leq \alpha \leq \frac{1}{vn} \end{aligned} \quad (10)$$

与 OCSVM 的对偶形式 (2) 相比, 会发现上式的核矩阵 $\hat{K}$ 是在原来核矩阵 K 的基础上加入了和数据分布相关的项, 或者说, 正是通过调整核矩阵使之能更为准确的反应数据的分布趋势, 从而使寻找到的数据边界能更准确的覆盖到目标数据所在的区域。事实上, 对于形如式 (10) 中的核矩阵亦可通过核学习来获得, Tsang 等即通过学习协方差矩阵的不确定性来得到更优的核^[4]。

式 (10) 仍属 QP 问题, 求解过程同 OCSVM。若 K 可导, 可进一步通过 Woodbury 公式简化得:

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \alpha^T (K^{-1} + \frac{\lambda}{n} H)^{-1} \alpha \\ \text{s. t. } & \alpha^T \mathbf{1} = 1, 0 \leq \alpha \leq \frac{1}{vn} \end{aligned} \quad (11)$$

为得到核空间的判别函数, 可将式 (9) 带入到线性空间的判别函数式 (7) 得:

$$F(Z) = \text{Sgn} \left[\left(\alpha^T \left(\tilde{K} - \frac{\lambda}{n} K H \left(I + \frac{\lambda}{n} H K H \right)^{-1} H \tilde{K} \right) \right)^T - \rho \right] \quad (12)$$

这里 $\tilde{K} = X Z$ 表示训练样本集 X 和测试样本集 $Z = \{z_1, z_2, \dots, z_m\}$ 的核矩阵。 $\rho = \rho \mathbf{1}$ 表示超平面的偏移向量, $F(Z)$ 和 $\text{Sgn}(\cdot)$ 分别表示式 (7) 中 $f(z)$ 和 $\text{sgn}(\cdot)$ 的对应向量, 即 $F(Z) = [f(z_1), f(z_2), \dots, f(z_m)]^T$, $\text{Sgn} = [\text{sgn}(z_1), \text{sgn}(z_2), \dots, \text{sgn}(z_m)]^T$, 各 $f(z)$ 的判别原则同式 (7)。

2 从二次规划到线性规划

通过对 SOCSVM 的判别函数式 (7) 分析不难发现, 若将作为训练样本的目标数据集 X 带入其中, 显然均应落入正半空间。换言之, 若将此作为目标函数, 通过最小化目标数据集与超平面之间的函数距离完全可以寻找到最优超平面, 而全体目标数据均需落在正半空间中也就成为约束之一。上述做法所带来的直接好处是将 SOCSVM 的 QP 求解进一步简化为 LP 优化, 称之为结构化线性规划单类

支持向量机 SlpOCSVM。事实上,上述思想最早出现于 Campbell 等提出的 lpOCSVM 当中 [8],显然,该算法如同 OCSVM 一样,因未考虑数据的全局信息而导致数据分布非球形分布时性能受到影响。因此,SlpOCSVM 不仅是 SOCSVM 的 LP 优化算法,同时也可看作在 lpOCSVM 中嵌入了全局结构信息。另外,由于不再采用间隔,SlpOCSVM 在提高速度的同时,也避免了 SOCSVM 必须借助原点作为负类代表来刻画算法的不足。

2.1 SlpOCSVM 及其核化形式

在保证目标数据均落在正半空间的前提下,SlpOCSVM 通过最小化全体目标数据在分类器作用下的平均输出结果,从而使最优超平面自动靠近样本数据。为避免噪声以及异常点的影响,亦引入软间隔,目标函数如下:

$$\begin{aligned} \min W(\alpha, \rho) = & \frac{1}{n} \left[(\alpha^T X^T (I + \lambda \Sigma)^{-1} X)^T - \rho + \frac{1}{v} \xi \right]^T \mathbf{1} \\ \text{s. t. } & \alpha^T \mathbf{1} = 1, \alpha \geq 0, \\ & (\alpha^T X^T (I + \lambda \Sigma)^{-1} X)^T \geq \rho - \xi, \xi \geq 0 \end{aligned} \quad (13)$$

从上式可见,SlpOCSVM 作为 SOCSVM 算法的加快速算法,是在对偶形式基础上简化出来的 LP 方法。与式 (6) 相比,所增加的约束即用于保证目标样本集均落在正半空间,向量 ρ 和 ξ 的意义同前,而 (α, ρ) 是相应的优化目标。对于式 (13) 的优化,可通过 LP 软件包或其它优化方法如内点法来实现。

对于 SlpOCSVM 的核化,仍集中在 $(I + \lambda \Sigma)^{-1}$ 的内积表达上,将其展开式 (9) 结果代入式,可得 SlpOCSVM 的核化形式:

$$\begin{aligned} \min W(\alpha, \rho) = & \frac{1}{n} \left[\left(\alpha^T \left(K - \frac{\lambda}{n} KH \left(I + \frac{\lambda}{n} HKH \right)^{-1} HK \right) \right)^T - \rho + \frac{1}{v} \xi \right]^T \mathbf{1} \\ \text{s. t. } & \alpha^T \mathbf{1} = 1, \alpha \geq 0, \\ & \left(\alpha^T \left(K - \frac{\lambda}{n} KH \left(I + \frac{\lambda}{n} HKH \right)^{-1} HK \right) \right)^T \geq \rho - \xi, \xi \geq 0 \end{aligned} \quad (14)$$

若 K 可逆,可得到如下简化形式:

$$\begin{aligned} \min_{\alpha} & \left[\left(\alpha^T \left(K^{-1} + \frac{\lambda}{n} H \right)^{-1} \right)^T - \rho + \frac{1}{v} \xi \right]^T \mathbf{1} \\ \text{s. t. } & \alpha^T \mathbf{1} = 1, \alpha \geq 0, \\ & \left(\alpha^T \left(K^{-1} + \frac{\lambda}{n} H \right)^{-1} \right)^T \geq \rho - \xi, \xi \geq 0 \end{aligned} \quad (15)$$

对于测试样本,输入空间和高维特征空间的判别函数分别对应于式 (7) 和式 (12)。

2.2 与 SOCSVM 计算复杂性比较

考虑到 SOCSVM 及其 LP 算法均有求逆运算且该结果可事先计算并存储,在此并不计入复杂度计算当中,因此两算法计算复杂度的比较,事实上就退化为 QP 和 LP 优化方法的比较。由于 LP 算法在采用形如单纯形法或内点法时,计算复杂度可渐进达到 $o(n)$ [12],而通常认为 QP 的复杂性为 $O(n^3)$,显然,采用线性规划的 lpOCSVM 算法在计算复杂性上要大大优于 QP 算法。

3 实验

3.1 正则化因子 λ 的影响

S(lp) OCSVM 所引用的参数 λ ,主要用于折中局部化和全局化信息。为观察该参数对算法的影响,图 3 比较了 SOCSVM 算法对于香蕉型数据在不同 λ 值作用下所寻找的目标数据描述边界。其中, $\lambda = 0$ 即为 OCSVM 的情形,图中以实线表示。而 $\lambda = 100, 1\,000, 10\,000$ 所寻找的数据边界则依次以点线,点划线和虚线表示。算法所采用的核函数是式 (3) 所示的高斯核(下同),从上到下三行分别表示核带宽为 $\sigma = \beta_0/2, \beta_0, 2\beta_0$ 的情形, β_0 为目标数据各样本点间的平均距离。左右两列分别表示硬间隔以及软间隔 $\nu = 0, 0.1$ 的情形。目标数据 50 个,以点 $\bullet$ 表示,支持向量以小空心圆 $\circ$ 表示。

图 3 中 σ 和 ν 所组合出的六种情形,从上向下看,随着核带宽参数 σ 的增大,分界面越来越光滑,所包围的目标数据域也越来越大。虽然因此可避免过拟合,但同时也意味着可能把更多的异常数据包含进数据域中。尤其是对于 $\lambda = 0$ 时的 OCSVM,受核带宽的影响更大。从图中可见,在 $\sigma = \beta_0/2$ 时 OCSVM 的数据域描述边界基本能反应香蕉数据分布,到 $\sigma = 2\beta_0$ 时该数据域几乎成为一超球,显然欠拟合。而对于 $\lambda \neq 0$ 情形下的 SOCSVM,随着 λ 增大,更多的数据结构信息反应在描述边界上,因而能更为准确的描述数据分布。尤其是在 $\sigma = 2\beta_0$ 的情形,在 OCSVM 已几乎失效的情况下,SOCSVM 仍能依靠 λ 参数的调整来得到较为准确的数据区域。由此可进一步看出,由于 λ 参数的引入,SOCSVM 对核带宽的选取不像 OCSVM 算法那样敏感,或者说,SOCSVM 的核带宽参数空间更宽泛。且在当前香蕉型数据分布时,核带宽取较大数值时,正则化因子 λ 也往往要取较大的数值,以便通过加入更多的全局信息来补偿因为大的核带宽所导致的欠拟合数据描述。

关于 ν 属性的影响, 比较左右两列可看出, OCSVM 随着软间隔数值的增加, 因为欧式距离的各向同性, 在各个方向都一致的向数据密集区域收缩, 因而对于数据分布方向有可能导致正常数据被拒绝而成为异常类。而对于 SOCSVM 算法, 由于

加入的协方差矩阵恰恰反应了数据分布的方向, 因此, 对于数据分布的方向, 软间隔算法并没有因为允许目标数据落在外部而有明显的收缩, 相反, 对于数据分布的正交方向, 却远比硬间隔时要紧凑的多。

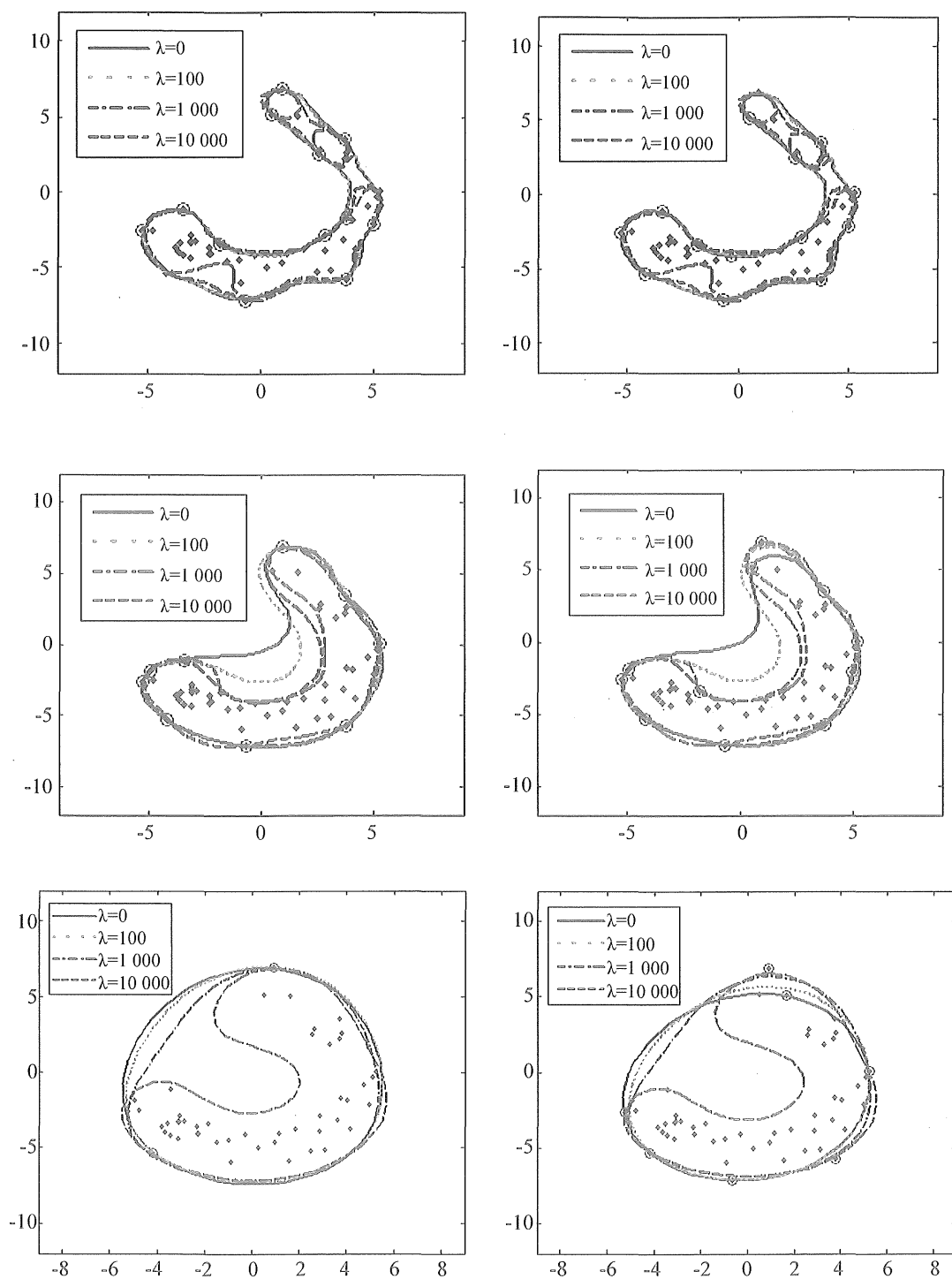


图3 SOCSVM 对于香蕉型数据在高斯核作用下不同 λ 值数据描述边界
从上到下 $\sigma = \beta_0/2, \beta_0, 2\beta_0$ 从左到右 $\nu = 0, 0.1$ (图中刻度表示当前数据的相对位置)

Fig. 3 Data description for SOCSVM with different λ using RBF kernel
from top to bottom $\sigma = \beta_0/2, \beta_0, 2\beta_0$ from left to right $\nu = 0, 0.1$

3.2 真实数据集验证

为进一步验证算法的性能, 拟在真实数据集上进行实验, 由于鲜有标准的单类数据集, 目前通用的做法大都是采用标准的两类或多类数据集并且转化为单类数据集来进行算法测试的^[4-5]。本文亦采用了 7 个 UCI 机器学习两类数据集。对于每个数据集, 考虑到实际应用场景中往往目标数据数量较大, 因此采用数目较多的一类作为目标类, 另一类作为异常类。按照通常的做法, 随机选择 80% 的目标数据作为训练样本, 剩余 20% 的目标数据和异常类分别作为正类和负类的测试数据。

试验中采用以错误率为中心的指标, 即目标类错分为异常类 FN 和异常类错分为目标类 FP 以及两者的平衡错误 Balance Loss (BL) 来评价, $BL = \frac{(FP + FN)^{[13]}}{2}$ 。其中, 表 1 比较了加入结构信息的 SOCSVM 算法与 OCSVM 及 OCMPM 的结果, 而表 2 则比较了两线性算法 (S) lpOCSVM 的性能。实验中仍采用了式 (3) 所示的高斯核, 核带宽 σ 和参数 λ 均通过网格搜索得到。区别起见, 各表中分别以黑斜体表示各指标中的最优值, 若有次优值, 则以黑体表示。

表 1 OCSVM/OCMPM/SOCSVM 三算法的 FP/ FN /BL 值

Table 1 FP/FN/BL results on UCI benchmark data sets with OCSVM, OCMPM and SOCSVM

Data Sets	OCSVM	OCMPM	SOCSVM
	FP/ FN /BL	FP/ FN /BL	FP/ FN /BL
Biomed	0.141 8/0.152 0/0.146 9	0.211 9/0.088 0/0.150 0	0.131 3/0.112 0/0.121 7
Breast cancer	0.022 8/0.060 4/0.041 6	0.062 2/0.031 9/0.047 1	0.027 4/0.047 30.037 3
Heart	0.542 4/0.278 1/0.410 3	0.769 1/0.153 1/0.461 1	0.516 5/0.253 1/0.384 8
Import	0.177 /0.335 3/0.256 4	0.233 8/0.229 4/0.231 6	0.181 7/0.229 4/0.205 6
Ionosphere	0.031 7/0.155 6/0.093 7	0.118 3/0.068 9/0.093 6	0.032 5/0.115 6/0.074 0
Sonar	0.116 5/0.659 1/0.387 8	0.656 7/0.100 0/0.378 4	0.135 1/0.540 9/0.338 0
Arrhythmia	0.461 2/0.119 1/0.290 2	0.473 2/0.006 4/0.294 1	0.379 8/0.159 6/0.269 7

分析表 1, 可看出:

(1) 比较三算法的 FN 值, 可看出 OCMPM 的值最低, SOCSVM 次之, 而 OCSVM 值最高。充分说明考虑数据结构信息, 可预测出数据的分布趋势, 因而避免目标数据落在边界外部, OCSVM 因为忽略了此信息, 导致 FN 错误率偏高。

(2) 比较三算法的 FP 值, 可看出 OCSVM 和 SOCSVM 的值相当, 而 OCMPM 则表现出较大的误差。这充分说明, 如 OCMPM 那样仅考虑数据的全局分布而忽略各样本点的信息, 使之不得不保证数据在最坏情形下的数据分布, 因而导致数据描述区域过松而可能包含更多的异常数据, 使 FP 值偏大。另外, OCSVM 及 SOCSVM 相当的 FP 值有力的说明了采用 OCSVM 框架的合理性, 使 SOCSVM 在考虑数据分布趋势的同时, 仍能够保留紧凑的数据描述边界, 因而并未降低对异常类的识别。

(3) 按照 BL 的结果, SOCSVM, OCSVM 和 OCMPM 依次递增, 显然与上述分析结果吻合。该结果进一步说明, 在单类分类器设计中, 充分考虑数据结构分布信息, 可以有效的提高分类器性能。

上述结论, 也可进一步通过表 2 两线性算法的

比较得出一致性的结果, 即考虑数据结构信息在真实数据集上一致优于占优。

表 2 (S) lpOCSVM 两算法的 FP/ FN /BL

Table 2 FP/FN/BL results on UCI benchmark data sets with (S) lpOCSVM

Data Sets	lpOCSVM	lpSOCSVM
	FP/ FN /BL	FP/ FN /BL
Biomed	0.288 1 /0.124 0/0.206 0	0.140 3/0.124 0/0.132 1
Breast cancer	0.034 4 /0.090 1/0.062 3	0.025 7/0.105 5/0.065 6
Heart	0.485 6 /0.406 3/0.445 9	0.595/0.200 0/0.397 5
Import	0.128 2/0.376 5/0.252 3	0.2141/0.229 4/0.221 7
Ionosphere	0.024 6 /0.228 9/0.126 7	0.0349/0.106 7/0.070 8
Sonar	0.029 9 /0.554 5/0.292 2	0.095 9/0.454 5/0.275 2
Arrhythmia	0.447 5 /0.142 5/0.295 0	0.392 3/0.140 4/0.266 4

4 结 论

本文提出的结构化单类支持向量机 SOCSVM, 通过在现有 OCSVM 基础上加入目标数据的协方差矩阵以反应数据的结构分布信息, 不仅克服了原算法嵌入先验信息的不足, 同时也为诸多的 SVM 算法提供了统一框架。为进一步提高运算效率, 在二

次规划求解 SOCSVM 的基础上, 提出了线性规划算法 SlpOCSVM, 从而克服了 SOCSVM 必须借助原点来刻画模型的不足。人工和真实数据集上的实验结果验证了 S (lp) OCSVM 的有效性。

参考文献:

- [1] TAX D. One-class classification: Concept-learning in the absence of counter-examples [D]. Delft University of Technology: Delft, 2001.
- [2] TAX D, DUIN R P. Support vector data description [J]. Machine Learning, 2004, 54(1): 45 - 66.
- [3] SCHOLKOPF B, PLATT J C, SHAWE-TAYLOR J. Estimating the support of a high-dimensional distribution [J]. Neural Computation, 2001, 13(7): 1443 - 1471.
- [4] 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 (IJCNN'06), 2006, Vancouver, BC: IEEE Computer Society, 1169 - 1175.
- [5] LANCKRIET G R G, GHAOUI L E, JORDAN M. Robust novelty detection with single-class MPM [C]. Advances in Neural Information Processing Systems, MIT Press, 2002: 905 - 912.
- [6] LANCKRIET G R G, GHAOUI L E, BHATTACHARYYA C, et al. A robust minimax approach to classification [J]. J Machine Learning Research, 2002, 3: 555 - 582.
- [7] JUSZCZAK P. Learning to recognise: a study on one-class classification and active learning [D]. Delft University of Technology: Delft, 2006.
- [8] CAMPBELL C, BENNETT K P. A linear programming approach to novelty detection [C]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001.
- [9] CRISTIANINI N, TAYLOR J S. An introduction to support vector machines and other Kernel-based learning methods [M]. Cambridge: Cambridge University Press, 2000: 90 - 120.
- [10] DUDA R O, HART P E, STORK D G. Pattern classification [M]. 2nd Edition, New York: John Wiley & Sons, 2001.
- [11] 张贤达, 矩阵分析与应用[M]. 北京: 清华大学出版社, 2004.
ZHANG X D. Matrix analysis and applications [M]. Beijing: Qinghua University Press, 2004.
- [12] MEGIDDO N. On the complexity of linear programming // Advances in economic theory: Fifth world congress [M]. T Bewley, ed. Cambridge: Cambridge University Press, 1987: 225 - 268.
- [13] JAMES T K, TSANG I W, ZURADA J M. A class of single-class minimax probability machines for novelty detection [J]. IEEE Transactions on Neural Networks, 2007, 18(3): 778 - 785.