

基于类内散度的粗糙 one-class 支持向量机

张 彬 朱嘉钢

(江南大学物联网工程学院 无锡 214122)

摘 要 粗糙 one-class 支持向量机(ROC-SVM)在粗糙集理论基础上通过构建粗糙上超平面和下超平面来处理过拟合问题,但是在寻找最优分类超平面的过程中,忽略了训练样本类内结构这一非常重要的先验知识。因此,提出了一种基于类内散度的粗糙 one-class 支持向量机(WSROC-SVM),该方法通过最小化训练样本类内散度来优化训练样本类内结构,一方面使训练样本在高维特征空间中与坐标原点的间隔尽可能大,另一方面使得训练样本在粗糙上超平面尽可能紧密。在合成数据集和 UCI 数据集上的实验结果表明,较原始算法,该方法有着更高的识别率和更好的泛化性能,在解决实际分类问题上更具优越性。

关键词 粗糙集,一类支持向量机,类内散度,过拟合

中图分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.024

Rough Set One-class Support Vector Machine Based on Within-class Scatter

ZHANG Bin ZHU Jia-gang

(School of IoT Engineering, Jiangnan University, Wuxi 214122, China)

Abstract Classical rough one-class support vector machine(ROC-SVM) constructs rough upper margin and rough lower margin to deal with the over-fitting problem on rough set theory. However, in the process of searching for the optimal classification hyper-plane, ROC-SVM ignores the inner-class structure of the training data which is a very important prior knowledge. Thus, a rough set one-class support vector machine based on within-class scatter(WSROC-SVM) was proposed. This algorithm optimizes the inner-class structure of the training data by minimizing the within-class scatter of the training data. It not only precipitates margin between the origin and the training data in a higher dimensional space as large as possible, but also makes the training data close around the rough upper margin as tight as possible. Experimental results carried out on one synthetic dataset and the UCI dataset indicate that the proposed method improves the accuracy as well as the generalization of the result. And it is more advantageous in solving practical classification problems.

Keywords Rough set, One-class SVM, Within-class scatter, Over-fitting

1 引言

one-class 支持向量机(one-class Support Vector Machine, OC-SVM)^[3]是由 Schölkopf 等人提出的一种新型的机器学习方法,主要用于处理一类问题或者聚类问题。首先它根据对应的核函数把目标样本点映射到高维特征空间,然后寻求一个超平面将目标样本点和原点之间以最大间隔分开,最终回归到一个决策函数来判断样本的类别。除此之外,根据文献[1,2],one-class 支持向量机还有一种形式是构造非线性映射空间中的超球体来覆盖目标样本点,称为支持向量数据描述(Support Vector Data Description)。one-class 支持向量机已经被应用在故障诊断、图像分割、异常检测和图片检索等应用领域,它主要是对目标的训练样本进行描述,然后根据训练样本的相似性来判断是否属于所描述的目标样本。

one-class 支持向量机的分类超平面仅由一小部分训练样

本点所决定,即由支持向量决定。因此它对于噪声点和异常点很敏感,影响了算法性能,甚至会导致过拟合。为了避免 one-class 支持向量机过拟合的问题,文献[5]将粗糙集理论引入到 one-class 支持向量机中,提出了新的 one-class 支持向量机,即粗糙 one-class 支持向量机(Rough One-class SVM, ROC-SVM),它可以有效地消除冗余信息,避免噪声点的干扰,并且在一定程度上能够避免过拟合问题。然而,粗糙 one-class 支持向量机在寻找最优分类的粗糙空间过程中,并未重视训练样本类内结构这一非常重要的先验知识。

本文将最小化类内散度与粗糙 one-class 支持向量机模型相融合,提出了基于最小类内散度的粗糙 one-class 支持向量机(WSROC-SVM)。该方法不仅能使训练样本在高维特征空间中与原点的间隔尽量大,而且还能使目标样本点的边界尽可能紧密。在 UCI 数据库的 6 个标准数据集上的实验结果表明,相比 ROC-SVM 和 OC-SVM,本文提出的算法取得

到稿日期:2015-10-19 返修日期:2016-01-18

张 彬(1990—),男,硕士生,CCF 学生会员,主要研究方向为人工智能与模式识别,E-mail:kayzhhb@163.com;朱嘉钢(1957—),男,博士,副教授,主要研究方向为人工智能与模式识别、软件工程。

了显著的分类效果,并且没有增大计算复杂度。

2 相关工作

2.1 粗糙 one-class 支持向量机

粗糙集 one-class 支持向量机^[5]构建了粗糙上超平面 $H(\omega, \rho_u)$ 和粗糙下超平面 $H(\omega, \rho_l)$, 并将它们与原点之间的距离定义为粗糙间隔。位于粗糙上超平面之内的样本属于正常点, 位于粗糙上下超平面之间的样本属于待判定点, 而位于粗糙下超平面之外的样本视为异常点样本。这样在寻找最优超平面的过程中更合理地考量了样本的差异性, 将置信信息引入了模型。

为了寻找最大粗糙间隔, 求解如下二次规划问题:

$$\begin{aligned} \min_{\omega, \xi, \rho} & \frac{1}{2} \omega^T \omega + \frac{1}{vL} \sum_{i=1}^L (\xi_i + \delta \xi_i^*) - (\rho_l + \rho_u) \\ \text{s. t. } & (\omega \cdot \phi(x_i)) \geq \rho_u - \xi_i - \xi_i^* \\ & 0 \leq \xi_i \leq \rho_u - \rho_l \\ & \rho_u, \rho_l \geq 0 \\ & \xi_i^* \geq 0, i=1, 2, \dots, L \end{aligned} \quad (1)$$

其中, ξ_i 和 ξ_i^* 是松弛变量, v 表示一个权衡变量, 参数 $\delta > 0$ 表示加重了对下边界样本点的惩罚, 使得这些点对下近似超平面的影响比其它处的样本点更大。为了解决这个二次规划问题, 引入以下拉格朗日函数:

$$\begin{aligned} L = & \frac{1}{2} \omega^T \omega + \frac{1}{vL} \sum_{i=1}^L (\xi_i + \delta \xi_i^*) - (\rho_l + \rho_u) - \sum_{i=1}^L \alpha_i ((\omega \cdot \phi(x_i)) - \rho_u + \xi_i + \xi_i^*) - \sum_{i=1}^L h_i (\rho_u - \rho_l - \xi_i) - \sum_{i=1}^L \gamma_i \xi_i - \\ & \mu_1 \rho_l - \mu_2 \rho_u - \sum_{i=1}^L \eta_i \xi_i^* \end{aligned} \quad (2)$$

其中, $\alpha_i \geq 0, h_i \geq 0, \gamma_i \geq 0, \eta_i \geq 0, \mu_1 \geq 0, \mu_2 \geq 0$ 为拉格朗日乘子。

通过求解各参数的偏导并使其等于 0, 可以得到这个二次规划的对偶问题如下:

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \sum_{i=1}^L \sum_{j=1}^L \alpha_i \alpha_j k(x_i, x_j) \\ \text{s. t. } & \sum_{i=1}^L \alpha_i \geq 2, 0 \leq \alpha_i \leq \frac{\delta}{vL}, i=1, 2, \dots, L \end{aligned} \quad (3)$$

当对偶问题的最优解 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_L)^T$ 求出后, 训练样本点的位置可以被 α_i 的值决定, 并且可以得到如下结论。

(1) 如果 $\alpha_i = 0$, 那么 x_i 位于粗糙上超平面 $H(\omega, \rho_u)$ 之内, 且满足 $\omega \cdot \phi(x_i) > \rho_u$, 则它是对应的目标类点, 即正常点。

(2) 如果 $0 < \alpha_i < \frac{1}{vL}$, 那么 x_i 位于粗糙上超平面 $H(\omega, \rho_u)$ 上, 且满足 $\omega \cdot \phi(x_i) = \rho_u$, 则它是对应上边界上的点, 即上近似界支持向量。

(3) 如果 $\alpha_i = \frac{1}{vL}$, 那么 x_i 位于粗糙上超平面 $H(\omega, \rho_u)$ 和粗糙下超平面 $H(\omega, \rho_l)$ 之间, 即在粗糙间隔上, 且满足 $\rho_l < \omega \cdot \phi(x_i) < \rho_u$, 则它对应一个粗糙间隔支持向量。

(4) 如果 $\frac{1}{vL} < \alpha_i < \frac{\delta}{vL}$, 那么 x_i 位于粗糙下超平面 $H(\omega, \rho_l)$ 上, 且满足 $\omega \cdot \phi(x_i) = \rho_l$, 则它是对应下边界上的点, 即下近似界支持向量。

(5) 如果 $\alpha_i = \frac{\delta}{vL}$, 那么 x_i 位于粗糙下超平面 $H(\omega, \rho_l)$ 之

外, 且满足 $\omega \cdot \phi(x_i) < \rho_l$, 则它一般对应一个异常点。

因而, 可以利用任意满足 $0 < \alpha_i < \frac{1}{vL}$ 来得到 ρ_u , 此时对应的 x_j 满足:

$$\rho_u = \omega \cdot \phi(x_j) = \sum_{i=1}^{l_1} \alpha_i K(x_i, x_j) \quad (4)$$

为了增强算法的鲁棒性以及自适应控制性能, 按式(5)计算 ρ_u 的值:

$$\rho_u' = \omega \cdot \phi(x_j) = \frac{1}{l_1} \sum_{j=1}^{l_1} \sum_{i=1}^{l_1} \alpha_i K(x_i, x_j) \quad (5)$$

其中, l_1 表示落在粗糙上超平面上点的数量, 即对应于 $0 < \alpha_i < \frac{1}{vL}$ 的点。

同理, 可以得到:

$$\rho_l' = \omega \cdot \phi(x_j) = \frac{1}{l_2} \sum_{j=1}^{l_2} \sum_{i=1}^{l_2} \alpha_i K(x_i, x_j) \quad (6)$$

其中, l_2 表示落在粗糙下超平面上点的数量, 即对应于 $\frac{1}{vL} < \alpha_i < \frac{\delta}{vL}$ 的点。

通过上面的分析, 把粗糙 one-class SVM 的决策函数定义为:

$$f(x) = \text{sgn}(\sum_{i=1}^L \alpha_i K(x_i, x) - \rho_u') \quad (7)$$

$$f'(x) = \text{sgn}(\sum_{i=1}^L \alpha_i K(x_i, x) - \rho_l') \quad (8)$$

根据上述分析, 训练样本按照它们在粗糙间隔中的相对位置获得不同的 α_i 值, 从而自适应地对决策超平面产生了影响, 并共同决定了它在高维空间中的位置。

对于一个给定的测试样本点 x_t , 可以通过以上两个决策函数得到 $f(x_t)$ 和 $f'(x_t)$ 的值, 然后按照如下结论来判别样本点的类别:

(1) 如果 $f(x_t) > 0$, 那么测试样本点 x_t 属于正样本, 即目标样本点;

(2) 如果 $f'(x_t) < 0$, 那么测试样本点 x_t 属于负样本, 即异常点;

(3) 如果 $f(x_t) < 0$ 且 $f'(x_t) > 0$, 那么测试样本点 x_t 可能属于负样本, 但因为缺少信息, 所以它的类别待定。

2.2 类内散度

类内散度^[7]是用来衡量各类样本的分布情况, 所以可以用来表示样本的类内部结构, 通常利用类内散度矩阵来表示。

已知样本集 $X = \{x_1, x_2, \dots, x_j\}$, $x_i \in R^D$, 它们分别属于两个不同的类, 样本子集 D_k 表示第 k 类, 给定分类决策平面的法向量 ω , 则类内散度表示为 $\omega^T S \omega$, $S = \sum_{i=1}^2 \sum_{x \in D_k} (x - \mu)$

$(x - \mu)^T$ 称为类内散度矩阵, $\mu = \frac{1}{n_k} \sum_{x \in D_k} x$ ($k=1, 2$) 称为样本均值。

3 基于最小类内散度粗糙 one-class 支持向量机

3.1 WSROC-SVM 算法

对一个分类器来说, 数据结构信息是一个重要的先验知识。然而, 粗糙集 one-class 支持向量机忽略了训练样本内部的结构信息, 如果让训练样本的类内散度最小化, 那么训练样本的区域会变得紧密, 从而使得样本更容易分离出原点。因

此,提出了一个基于最小类内散度的粗糙集 one-class 支持向量机(WSROC-SVM),在一个高维空间中,通过最小化类内散度,使得训练样本和原点之间的距离变得更大。为此,需要求解如下的二次规划问题:

$$\begin{aligned} \min_{\omega, \xi, \rho} & \frac{1}{2} \omega^T \omega + \frac{1}{2} \beta \omega^T S \omega + \frac{1}{\nu l} \sum_{i=1}^l (\xi_i + \delta \xi_i^*) - \rho_u - \rho_a \\ \text{s. t. } & (\omega \cdot \phi(x_i)) \geq \rho_u - \xi_i - \xi_i^*, \\ & 0 \leq \xi_i \leq \rho_u - \rho_l \\ & \rho_u, \rho_l \geq 0 \\ & \xi_i^* \geq 0, i=1, 2, \dots, l \end{aligned} \quad (9)$$

其中, ξ_i 和 ξ_i^* 是松弛变量, ν 表示一个权衡变量, 参数 $\delta > 1$, 表示加强位于粗糙下超平面样本的惩罚因子, $S = \sum_{i=1}^l (\phi(x_i) - m)(\phi(x_i) - m)^T$ 表示类内散度, $m = \frac{1}{l} \sum_{i=1}^l \phi(x_i)$ 表示训练样本的平均值, $\beta \geq 0$ 是一个权衡粗糙边缘和训练样本类内结构的参数。

根据表示理论^[3], 式(9)的解能够表示成 $\omega = \sum_{i=1}^l \alpha_i \phi(x_i)$, 换言之, 任何解都可以在高维空间中由训练样本线性表示。

第一项 $\frac{1}{2} \omega^T \omega$ 可以被改写成:

$$\frac{1}{2} \omega^T \omega = \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j k(x_i, x_j) = \frac{1}{2} \alpha^T H \alpha \quad (10)$$

其中, H 是核函数 $k(x_i, x_j)$ 组成的 $l \times l$ 的格拉姆矩阵。

第二项 $\frac{1}{2} \beta \omega^T S \omega$ 可以被改成:

$$\begin{aligned} & \frac{1}{2} \beta \omega^T S \omega \\ &= \frac{\beta}{2} \omega^T \sum_{i=1}^l (\phi(x_i) - m)(\phi(x_i) - m)^T \omega \\ &= \frac{\beta}{2} [\sum_{i=1}^l \omega^T \phi(x_i) \phi(x_i)^T \omega - l \omega^T m m^T \omega] \\ &= \frac{\beta}{2} \{ \sum_{i=1}^l [\sum_{p=1}^l \alpha_p k_{pi} \sum_{q=1}^l \alpha_q k_{qi}] - \frac{1}{l} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j \sum_{k=1}^l k_{ik} \sum_{l=1}^l k_{jl} \} \\ &= \frac{\beta}{2} (\alpha^T H H^T \alpha - \alpha^T H L H^T \alpha) \\ &= \frac{\beta}{2} \alpha^T H (I - L) H^T \alpha \end{aligned} \quad (11)$$

其中, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$, I 是一个 l 阶的单位矩阵, L 是一个所有元素都等于 $\frac{1}{l}$ 的矩阵。

因此, 根据式(10)和式(11), 可以把式(9)改写成:

$$\begin{aligned} \min_{\omega, \xi, \rho} & \frac{1}{2} \alpha^T G \alpha + \frac{1}{\nu l} \sum_{i=1}^l (\xi_i + \delta \xi_i^*) - (\rho_l + \rho_u) \\ \text{s. t. } & \sum_{j=1}^l \alpha_j k(x_j, x_i) \geq \rho_u - \xi_i - \xi_i^* \\ & 0 \leq \xi_i \leq \rho_u - \rho_l \\ & \rho_u, \rho_l \geq 0 \\ & \xi_i^* \geq 0, i=1, 2, \dots, l \end{aligned} \quad (12)$$

其中, $G = H + \beta H(I - L)H^T$ 。同样, 为了解决这个二次规划问题, 首先引入以下拉格朗日函数。

$$\begin{aligned} L = & \frac{1}{2} \alpha^T G \alpha + \frac{1}{\nu l} \sum_{i=1}^l (\xi_i + \delta \xi_i^*) - (\rho_l + \rho_u) - \sum_{i=1}^l g_i (\sum_{j=1}^l \alpha_j k(x_j, x_i) - \rho_u + \xi_i + \xi_i^*) \\ & - \sum_{i=1}^l h_i (\rho_u - \rho_l - \xi_i) - \sum_{i=1}^l \gamma_i \xi_i - \mu_1 \rho_l - \mu_2 \rho_u - \sum_{i=1}^l \eta_i \xi_i^* \end{aligned} \quad (13)$$

其中, $g_i \geq 0, h_i \geq 0, \gamma_i \geq 0, \eta_i \geq 0, \mu_1 \geq 0, \mu_2 \geq 0$ 为拉格朗日乘子。根据 KKT 条件, 参数满足以下条件:

$$\frac{\partial L}{\partial \alpha} = G \alpha - H g = 0 \quad (14)$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{\nu l} - g_i + h_i - \gamma_i = 0 \quad (15)$$

$$\frac{\partial L}{\partial \xi_i^*} = \frac{\sigma}{\nu l} - g_i + \eta_i = 0 \quad (16)$$

$$\frac{\partial L}{\partial \rho_l} = 1 + \sum_{i=1}^l h_i - \mu_1 = 0 \quad (17)$$

$$\frac{\partial L}{\partial \rho_u} = -1 + \sum_{i=1}^l g_i - \sum_{i=1}^l h_i - \mu_2 = 0 \quad (18)$$

$$g_i (\sum_{j=1}^l \alpha_j k(x_j, x_i) - \rho_u + \xi_i + \xi_i^*) = 0 \quad (19)$$

$$h_i (\rho_u - \rho_l - \xi_i) = 0 \quad (20)$$

$$\gamma_i \xi_i = 0, \mu_1 \rho_l = 0, \mu_2 \rho_u = 0, \eta_i \xi_i^* = 0 \quad (21)$$

可以得到这个二次规划问题的对偶问题如下:

$$\begin{aligned} \min_g & \frac{1}{2} g^T H^T ((H + \beta H(I - L)H^T)^{-1})^T H g \\ \text{s. t. } & \sum_{i=1}^l g_i = 2, 0 \leq g_i \leq \frac{\delta}{\nu l}, i=1, 2, \dots, l \end{aligned} \quad (22)$$

通过以上推导过程发现, 最终的优化式可以用 QP (Quadratic Programming) 优化来解决, 如同粗糙 one-class SVM 的最终式一样, 求得 g 后, 再通过式(23)得到 α :

$$\alpha = (H + \beta H(I - L)H^T)^{-1} H g \quad (23)$$

在模型中引入类内散度, 实现了对类内结构的考量, 同时没有增加计算复杂度。进一步推导可以发现, 当 $\beta = 0$ 时, 本文算法就退化为粗糙 one-class 支持向量机, 因此, 本文算法在加强类间鉴别信息处理能力的同时充分考虑了类内样本的分布情况, 在解决分类问题时更具有一般性。

3.2 WSROC-SVM 算法描述

根据以上目标函数的详细推导, 可以得到 WSROC-SVM 算法流程, 如表 1 所列。

表 1 WSROC-SVM 算法流程

算法: WSROC-SVM 算法
输入: 输入训练样本 X 和测试样本 Y
步骤 1 初始化各个参数;
步骤 2 输入训练样本, 通过高斯核函数映射得到核矩阵 H ;
步骤 3 计算 $H^T(H + \beta H(I - L)H^T)^{-1}H$;
步骤 4 使用二次规划方法求解对偶问题式(22);
步骤 5 由步骤 4 得到式(22)的解 g , 以计算优化式(9)的解 $\alpha = (H + \beta H(I - L)H^T)^{-1}H g$;
步骤 6 根据判断 g_i 的值, 分别用所对应的 α_i 根据式(5)和式(6)计算 ρ_u' 和 ρ_l' ;
步骤 7 输入测试样本, 计算式(7)和式(8);
步骤 8 根据所求式(7)和式(8)的值来判定测试样本所属类别。
输出: 识别率

4 实验与分析

为了验证本文提出的 WSROC-SVM 算法在处理分类问题时的性能, 在合成数据和 UCI 数据库的 6 个基准数据集上进行了实验。实验程序采用 PC 机上的 Matlab2013a 实现, 机器配置如下: Intel(R) Core(TM) i5-4590 CPU@3.30GHz, 8GB 内存, Windows7(64 位)操作系统。

4.1 仿真数据实验

首先, 随机生成一个二维的合成数据集来验证所提算法, 第一类由 $X_1 \sim N(3.5, 1)$ 正态分布随机生成 100 个点, 第二

类由 $X_2 \sim N(1,1)$ 正态分布随机生成 20 个点, 它们的分布如图 1 所示。在训练学习过程中, 把第一类看成正常点, 即正样本, 把第二类看成异常点, 即负样本。将正样本的 80% 作为训练样本, 将剩余的正样本和所有的负样本作为测试样本, 整个过程循环 5 次。

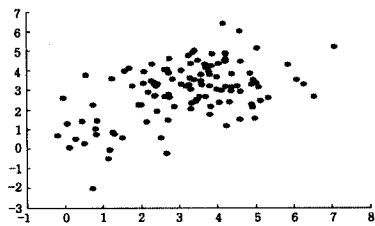


图 1 合成数据分布图

WSROC-SVM 的测试样本平均识别率与参数 β 的关系如图 2 所示。在图 2 中, 横轴表示参数 β 的值, 纵轴表示所有测试样本的平均识别率, 从中可以看出, 当确定其它参数值时, 参数 β 值的选择对识别结果的影响较大, 因而对于最小类内散度的粗糙 one-class 支持向量机, 选择合适的参数很重要。

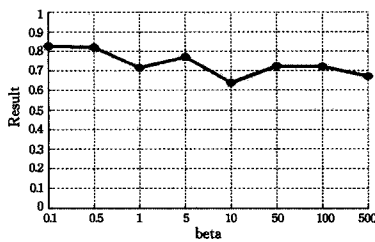


图 2 参数 β 对合成数据实验结果的影响

4.2 UCI 数据集实验

为了进一步验证本文提出的 WSROC-SVM 算法在处理分类问题时有着良好的性能, 在 UCI 数据库^[17] 的 6 个标准数据集中进行了实验。数据集基本信息如表 2 所列, 数据集包括 Surcical, Wine, Heart, Breast Cancer, Ionosphere 和 Sonar。

表 2 UCI 数据集基本信息

样本数据集	正样本数	负样本数	维数
Survival	225	81	3
Wine	71	107	13
Heart	160	137	13
Breast Cancer	148	46	33
Ionosphere	225	126	34
Sonar	97	111	60

为了比较各算法之间的性能, 对每一个样本数据集分别用 one-class 支持向量机(OC-SVM), 粗糙集 one-class 支持向量机(ROC-SVM)和基于最小类内散度的粗糙 one-class 支持向量机(WSROC-SVM)进行比较, 采用 5-fold 交叉验证的方法, 将正样本随机分成 5 组子集, 选取其中 4 组正样本作为训练样本, 剩余的 1 组正样本和所有的负样本作为测试样本, 整个过程循环 5 次, 然后取平均识别率。

在所有数据集实验中, 选用高斯核 $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / p^2)$ 作为核函数, 然后通过网格搜索法来寻找出各个算法的最优参数, 高斯核参数 p 从集合 $\{2^i | i = -4, -3, \dots, 10\}$ 中选取, 参数 v 从集合 $\{0.1, 0.2, \dots, 0.9\}$ 中选取, 参数 σ 从集合 $\{2^j | j = 1, 2, \dots, 7\}$ 中选取, 此外, 引入的新参数 β 用来控制样本边界和类内结构的权衡, 从集合 $\{0.1, 0.5, 1, 5, 10, 50,$

100, 500\} 中选取。表 3 给出了 OC-SVM, ROC-SVM 和 WSROC-SVM 在 UCI 数据集上的实验结果, 分别统计每个数据集识别过程中的所有样本的识别率(AR)以及训练和识别耗时(Time), 实验中各个算法使用 5-fold 交叉验证设置的参数如表 4 所列。

表 3 UCI 数据集识别率(%)和算法耗时(s)对比

数据集	算法	AR	Time
Survival	OC-SVM	46.51	3.87
	ROC-SVM	50.32	4.04
	WSROC-SVM	64.29	3.99
Wine	OC-SVM	80.53	0.31
	ROC-SVM	86.48	0.34
	WSROC-SVM	89.44	0.36
Heart	OC-SVM	70.13	2.07
	ROC-SVM	70.67	2.38
	WSROC-SVM	82.27	2.21
Breast Cancer	OC-SVM	38.88	1.21
	ROC-SVM	41.27	1.17
	WSROC-SVM	60.85	1.18
Ionosphere	OC-SVM	83.39	3.01
	ROC-SVM	93.33	3.15
	WSROC-SVM	90.64	3.24
Sonar	OC-SVM	70.86	0.54
	ROC-SVM	74.24	0.57
	WSROC-SVM	85.12	0.60

表 4 算法参数设置

数据集	算法	v	i	j	β
Survival	OC-SVM	2	0.3	—	—
	ROC-SVM	0.7	8	1	—
	WSROC-SVM	0.2	8	7	0.1
Wine	OC-SVM	2	0.4	—	—
	ROC-SVM	0.8	9	1	—
	WSROC-SVM	0.5	8	6	0.5
Heart	OC-SVM	2	0.6	—	—
	ROC-SVM	0.8	1	2	—
	WSROC-SVM	0.7	6	1	0.1
Breast Cancer	OC-SVM	10	0.1	—	—
	ROC-SVM	0.1	2	4	—
	WSROC-SVM	0.2	9	7	0.1
Ionosphere	OC-SVM	2	0.1	—	—
	ROC-SVM	0.1	1	2	—
	WSROC-SVM	0.3	1	5	0.1
Sonar	OC-SVM	2	0.9	—	—
	ROC-SVM	0.8	3	1	—
	WSROC-SVM	0.1	10	4	0.1

通过比较表 3 中的所有样本的识别率(AR)可以发现, WSROC-SVM 在 Surcical, Wine, Heart, Breast Cancer 和 Sonar 数据集上的 AR 结果优于 OC-SVM 和 ROC-SVM, 而 ROC-SVM 仅在 Ionosphere 数据集上的 AR 结果最优, 但是 WSROC-SVM 的 AR 结果与最优结果差距不大。说明 WSROC-SVM 具有较强的鲁棒性。

通过比较表 3 中的算法耗时(Time)可以发现, 各算法训练和识别耗时基本差不多, 尽管 WSROC-SVM 引入了一个新参数 β , 但是并没有增加 CPU 计算的时间, 即新增加的参数并没有增大 WSROC-SVM 的时间复杂度。

结束语 本文根据训练样本类内结构这样一个非常重要的先验知识, 在粗糙 one-class 支持向量机的基础上构造了一种基于类内散度的粗糙 one-class 支持向量机, 通过使训练样本的类内散度最小化的方式来优化训练样本类内结构, 从而

(下转第 172 页)

孙玲芳,李烁朋. 基于 K-means 聚类与张量分解的社会化标签推荐系统研究[J]. 江苏科技大学学报(自然科学版),2012,26(6):597-601

- [6] Wang Long, Wang Jia-lun, Cheng Zhuan-li, et al. Personalized Medicine Recommendation Based on Tensor Decomposition[J]. Computer Science, 2015, 42(5): 225-229(in Chinese)
王龙,王嘉伦,程转丽,等. 基于张量分解的药品个性化推荐[J]. 计算机科学, 2015, 42(5): 225-229
- [7] Fouzia J, Shah K, Amna M, et al. Semantics discovery in social tagging systems: A review[J]. Multimedia Tools and Applications, 2016, 75(1): 573-605
- [8] 江铭炎,袁东风. 人工鱼群算法及其应用[M]. 科学出版社,2012
- [9] Peng Yong, Tang Guo-lei, Xue Zhi-shun. Optimal operation of cascade reservoirs based on improved artificial fish swarm algorithm[J]. Systems Engineering-Theory & Practice, 2011, 36(6): 1118-1125(in Chinese)
彭勇,唐国磊,薛志春. 基于改进人工鱼群算法的梯级水库群优化调度[J]. 系统工程理论与实践, 2011, 36(6): 1118-1125
- [10] Liao Yu-lei, Liu Peng, Wang Jian, et al. Control parameter optimization for the unmanned surface vehicle with the improved artificial fish swarm algorithm[J]. Journal of Harbin Engineering University, 2014, 35(7): 800-806(in Chinese)

廖煜雷,刘鹏,王建,等. 基于改进人工鱼群算法的无人艇控制参数优化[J]. 哈尔滨工程大学学报, 2014, 35(7): 800-806

- [11] Chen Guang-zhou, Wang Jia-quan, Li Chuan-jun, et al. An improved artificial fish swarm algorithm and its applications [J]. Systems Engineering, 2009, 27(12): 105-110(in Chinese)
陈广洲,汪家权,李传军,等. 一种改进的人工鱼群算法及其应用[J]. 系统工程, 2009, 27(12): 105-110
- [12] Weng S S, Lin B S, Chen W T. Using contextual information and multidimensional approach for recommendation [J]. Expert Systems with Applications, 2007, 36(2): 1268-1279
- [13] Li Gui, Wang Shuang, Li Zheng-yu, et al. Personalized Tag Recommendation Algorithm Based on Tensor Decomposition[J]. Computer Science, 2015, 42(2): 267-273(in Chinese)
李贵,王爽,李征宇,等. 基于张量分解的个性化标签推荐算法[J]. 计算机科学, 2015, 42(2): 267-273
- [14] Zou B Y, Li C P, Tan L W, et al. Social recommendations based on user trust and tensor factorization[J]. Journal of Software, 2014, 25(12): 2852-2864(in Chinese)
邹本友,李翠平,谭力文,等. 基于用户信任和张量分解的社会网络推荐[J]. 软件学报, 2014, 25(12): 2852-2864
- [15] Milicevic A K, Nanopoulos A, Mirjana Ivanovic. Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions[J]. Artificial Intelligence Review, 2010, 33(1): 187-209

(上接第 138 页)

使训练样本在高维特征空间中与原点的时间隔尽可能大,同时使目标样本在上近似超平面尽可能紧密,充分利用训练样本的类内结构增强分类性能。

在实验中发现,不同目标函数的形式以及参数 v, p, σ, β 的取值对算法性能有较大影响。因此,如何设计合理的目标函数和选取算法中的参数将是未来的研究重点。

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] Campbell C, Bennett P. A Linear Programming Approach to Novelty Detection [M]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001
- [3] Schölkopf B, Platt J C, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471
- [4] Lin C F, Wang S D. Fuzzy Support Vector Machine [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471
- [5] Xu Yi-tian, Liu Chun-mei. A rough margin-based one class support vector machine[J]. Neural Comput & Applic, 2013(22): 1077-1084
- [6] Bishop C M. Pattern Recognition and Machine Learning [M]. Cambridge: Springer, 2007: 291-320
- [7] Jun G, Chung F, Wang S. Matrix pattern based minimum within-class scatter support vector machines[J]. Applied Soft Computing, 2011, 11(8): 5602-5610
- [8] Zhang J, Wang K, Zhu W, et al. Least Squares Fuzzy One-class Support Vector Machine for Imbalanced Data[J]. International Journal of Signal Processing, Image Processing and Pattern Re-

cognition, 2015, 8(8): 299-308

- [9] Jeong Y S, Kang I H, Jeong M K, et al. A New Feature Selection Method for One-Class Classification Problems [J]. IEEE Transactions on Systems, Man, and Cybernetics—part c: Applications and Reviews, 2012, 42(6): 1500-1509
- [10] Zhu W, Zhong P. A new one-class SVM based on hidden information[J]. Knowledge-Based Systems, 2014, 60(2): 35-43
- [11] Asharaf S, Shevade S K, Narasimha murty M. Rough support vector clustering[J]. Pattern Recognition, 2005, 38(10): 1779-1783
- [12] Xiao Y, Wang H, Xu W. Parameter Selection of Gaussian Kernel for One-Class SVM[J]. IEEE transactions on cybernetics, 2015, 45(5): 927-939
- [13] An W, Liang M. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises [J]. Neurocomputing, 2013, 110(6): 101-110
- [14] Wang L, Jia H D, Li J. Training Robust Support Vector Machine with Smooth Ramp Loss in the Primal [J]. Neurocomputing, 2008, 71: 3020-3025
- [15] Junejo I N, Bhutta A A, Foroosh H. Single-class SVM for dynamic scene modeling [J]. SIVIP, 2013, 7(1): 45-52
- [16] Mygdalis V, Iosifidis A, Tefas A, et al. Exploiting subclass information in one-class support vector machine for video summarization[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 2259-2263
- [17] Wang Lei, Yang Yi-fan, Zhou Qi-hai. Rough Set based One-class Support Vector Machine[J]. Computer Science, 2009, 36(9): 242-245(in Chinese)
王磊,杨一帆,周启海. 粗糙 one-class 支持向量机[J]. 计算机科学, 2009, 36(9): 242-245
- [18] <http://archive.ics.uci.edu/ml/datasets.html>