

稀疏正则化最小类散度半监督分类机

刘建华, 吴冬燕

(浙江工商职业技术学院 电子与信息工程学院, 浙江 宁波 315012)

摘要: 基于稀疏表示理论提出一种稀疏正则化最小类散度半监督分类机(SRMCV), 且对于模式分类问题, SRMCV通过引入稀疏Laplacian正则化项和类内散度信息以实现预测空间函数在全局稀疏表示结构下平滑变化, 同时通过类内数据散度结构进一步优化决策函数的判别方向, 此方法能解决现有SSL方法对模型参数敏感和在噪声学习环境下缺乏鲁棒性等问题, 其有效性已在实际数据集上通过实验验证.

关键词: 半监督学习; 稀疏表示; 支持向量机; 最小类散度; 核技巧

中图分类号: TP181

文献标志码: A

文章编号: 1001-5132 (2014) 02-0044-05

基于流形学习理论, Belkin 等^[1]提出一种流形正则化(MR)框架作为一种统一架构范式, MR能用于从无监督学习到半监督学习(SSL), 再到有监督学习等几乎所有机器学习范畴问题^[2]. MR框架建立于某个再生核 Hilbert 空间(RKHS)^[3], 并由此提出一种新的用于正则化框架优化的 Representer 理论^[1]. 相较于仅关注函数空间中函数复杂度的传统正则化理论, MR框架的最大贡献在于其基于流形假设理论(即高维输入数据位于某个本质低维流形空间), 在传统正则化框架中额外增加了一个流形正则化惩罚项, 并用该惩罚项来度量基于流形数据习得的预测函数的平滑性, 该流形正则化项有效地增强了流形学习机的性能. Belkin 等^[1]基于MR框架提出了两种经典算法, 包括Laplacian正则化最小平方分类(LapRLSC)和Laplacian支持向量机(LapSVM), 这些算法均能有效用于有监督或半监督学习问题.

一直以来, 在分类机中控制数据分布结构是一个重要的研究主题. 线性判别分析(LDA)具备了保持数据的全局判别信息及全局数据结构的能力, 但是它也存在小样本问题(SSS), 而且投影之后的结果也无法直接用于分类, 正因为此, Zafeiriou 等^[4]通过最小化类内方差以便构建正则化框架. 针对

MR框架未能充分考虑全局数据的分布信息, 近来皋军等^[5]结合数据的全局散度信息和局部流形信息, 提出一种全局和局部保持的半监督支持向量机(GLSSVM), 其本质上是MCSVM和MR框架的融合体.

尽管现有的大多数SSL方法均已成功应用于机器学习或模式识别的许多实际领域, 但经研究发现, 上述SSL方法在具体应用中未能较好地解决模型参数选择、流形假设、鲁棒性等问题.

笔者基于稀疏表示理论^[6], 从充分考虑类内数据分布结构信息和模型鲁棒性的角度, 提出一种有判别的稀疏正则化最小类散度半监督学习机(SRMCV). 与现有方法相比, SRMCV的主要创新之处体现在:

(1) 源于稀疏表示的本质鲁棒性和自然判别力^[6], 所提方法针对噪声学习环境具有更强的鲁棒性和学习判别力.

(2) 与现有基于稀疏表示的半监督学习方法不同, 所提方法除了具有 Out-of-Sample 推理能力外, 还充分考虑了数据类内分布结构信息, 从而优化学习机的判别方向, 这将在一定程度上能提升所提方法的分类性能.

(3) 通过稀疏表示来构建数据关系图模型, 同

时实现了近邻集构建和图边权值的计算, 避免了传统方法在模型参数选取上的敏感性.

1 SRMCV 方法

1.1 问题描述

给定一个二类训练数据集 $X = \{x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+u}\} \in \mathbb{R}^d$ (d 为数据维数) 和标签集 $Y = \{1, 2\}$, 其中, 前 l 个点构成数据集 X_l , 数据点 $x_i \in X_l$ ($1 \leq i \leq l$) 标签为 $y_i \in Y$, 余下的 u 个数据点构成数据集 X_u , 数据点 $x_j \in X_u$ ($l+1 \leq j \leq l+u = N$) 无标签. 半监督学习方法的主要目标是利用少量有标签数据和大量无标签数据来学习一个在所有数据集上具有强泛化能力的决策函数 $f(x) = w^T x$, 其中, w 为决策平面的法向量, T 代表矩阵或向量转置. 为描述方便, 先作如下定义:

定义 1 设带标签数据集 X_l 分属 2 个类集 C_1 和 C_2 , 其中, 类 C_k ($k=1, 2$) 数据集大小为 N_k , $\sum_{k=1}^2 N_k = l$, 则数据集 X_l 的类内散度为 $\Omega_w = w^T S_w w$, 其中, 类内散度矩阵 $S_w = \sum_{k=1}^2 \sum_{x \in C_k} (x - u_k)(x - u_k)^T$, 均值 $u_k = (\sum_{x \in C_k} x) / N_k$ ($k=1, 2$).

定义 2 任意数据点 $x_i \in X$ ($i=1, \dots, N$) 的最优稀疏表示系数向量^[6]由如下优化目标确定:

$$\min_S \|x_i - Xs_i\|^2 + \lambda \|s_i\|_1, \quad (1)$$

其中, S 为数据集 X 的稀疏表示矩阵, 列向量 $s_i = [s_{i1}, \dots, s_{ii-1}, 0, s_{ii+1}, s_{iN}]^T$ 为数据点 $x_i \in X$ 的稀疏表示系数向量, $\lambda > 0$ 为平衡参数, $\|\cdot\|_1$ 表示 l_1 -范数.

(1) 式中 l_1 -范优化问题可通过现有算法求解^[6]. 根据定义 2, 笔者构建 1 个稀疏图模型以表示数据点间稀疏表示关系, 或称数据稀疏图 $G(X, S)$, 图中每个顶点都对应 1 个原始数据点, 图边权值矩阵 S 表示数据点间的稀疏表示关系.

定义 3 稀疏正则化 Laplacian 矩阵为 $L_s = D - S$, 其中, D 为对角矩阵,

$$D = \sum_{j=1}^N S_{ij}.$$

SRMCV 方法旨在基于稀疏图 G 寻求一个优化的分类函数 $f(x)$, 且满足: (1) 优化函数的输出应接近图中带标签顶点的标签值; (2) 优化函数的输出应在整个图上平滑; (3) 数据集类内散度最小化. 从而, SRMCV 方法的最小化目标函数为:

$$f^* = \arg \min_{f \in H_k} \left\{ \frac{1}{l} \sum_{i=1}^l \text{loss}(x_i, y_i, f) + \right.$$

$$\left. \gamma_w \Omega_w + \gamma_s \|f\|_s^2 \right\}, \quad (2)$$

其中, $\text{loss}(\cdot)$ 为损失函数^[2] (如 Hing 损失和平方损失); H_k 为由某个 Mercer 核 k (如线性核、Gaussian 核等) 诱导的再生核 Hilbert 空间(RKHS)^[2]; $\|f\|_s^2$ 则为稀疏正则项, 定义为 $\|f\|_s^2 = f^T L_s f$; 模型参数 $\gamma_w, \gamma_s \geq 0$. 根据 SVM 的良好性能表现, 笔者拟采用 Hing 损失函数, 且综合定义, SRMCV 目标函数可描述为:

$$f^* = \arg \min_{f \in H_k} \left\{ \frac{1}{l} \sum_{i=1}^l \zeta_i + \frac{\gamma_w}{2} w^T S_w w + \frac{\gamma_s}{2} f^T L_s f \right\}, \quad (3)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \zeta_i, i=1, \dots, l.$$

以下分线性和非线性两种情况探讨 SRMCV 目标函数的优化算法.

1.2 线性 SRMCV

定理 1 线性 SRMCV 方法的原始目标优化问题(3)的对偶问题为:

$$\min_{\beta} \frac{1}{2} \beta^T M \beta - 1_k^T \beta,$$

$$\text{s.t. } \sum_{i=1}^l y_i \beta_i = 0, 0 \leq \beta_i \leq \frac{1}{l}, \quad (4)$$

其中, $\beta = [\beta_1, \dots, \beta_l]^T$ 为拉格朗日系数, $M = (M_{ij})_{l \times l}$, $M_{ij} = y_i y_j x_i^T (\gamma_w S_w + \gamma_s X L_s X^T)^{-1} x_j$, Δ^{-1} 表示矩阵 Δ 的逆阵, 1_k 为 k 维全 1 向量.

可按照文献[5]同样方法来证明定理 1, 且根据定理 1 的证明, 可知分类平面法向量 w 为:

$$w = \sum_{i=1}^l \beta_i y_i (\gamma_w S_w + \gamma_s X L_s X^T)^{-1} x_i.$$

优化问题(4)为标准的凸二次规划问题, 可采用标准 SVM 优化算法来求解, 以获得优化变量 β , 从而获得优化法向量 w^* . 此外, 按照 SVM 方法, 优化偏置量 b^* 计算公式为:

$$b^* = \frac{1}{l} \sum_{i=1}^l (y_i - \sum_{j=1}^l y_j \beta_j x_j^T (\gamma_w S_w + \gamma_s X L_s X^T)^{-1} x_i).$$

因此, SRMCV 决策函数为:

$$f(x) = w^T x + b = \sum_{i=1}^l \beta_i y_i (\gamma_w S_w + \gamma_s X L_s X^T)^{-1} x_i + \frac{1}{l} \sum_{i=1}^l (y_i - \sum_{j=1}^l y_j \beta_j x_j^T (\gamma_w S_w + \gamma_s X L_s X^T)^{-1} x_i), \quad (5)$$

其中, $\langle \cdot, \cdot \rangle$ 表示向量内积.

1.3 非线性 SRMCV

针对非线性学习环境采用核技巧^[3], 笔者提出非线性 SRMCV 方法. 为方便计算, 在定义 1 中令 $\tilde{x} = x - u_k$, $x \in C_k$ ($k=1, 2$), 重新定义训练数据集 X_l 的类内散度为 $\Omega_w = w^T \tilde{S}_w w$, 其中, 类内散度矩阵 $\tilde{S}_w = \sum_{k=1}^2 \sum_{x \in C_k} \tilde{x} \tilde{x}^T$. 根据 Representer 定理^[1], 决策平面法向量 w 可表示为 $w = \sum_{i=1}^N \alpha_i \phi(x_i)$, 其中, $\alpha = [\alpha_1, \dots, \alpha_N]^T$ 为权值向量, ϕ 为 Mercer 核 K 映射, 则上述各正则化项可分别表示为:

$$\Omega_w = w^T \tilde{S}_w w = \alpha^T K_l K_l^T \alpha,$$

$$f^T L_s f = \alpha^T K_N L_s K_N \alpha,$$

其中, K_l 为 $N \times l$ 矩阵, K_N 为 $N \times N$ 矩阵. 因此非线性 SRMCV 原始优化目标为:

$$\arg \min_{\alpha, b, \xi} \left\{ \frac{1}{l} \sum_{i=1}^l \xi_i + \frac{\gamma_w}{2} \alpha^T K_l K_l^T \alpha + \gamma_s \alpha^T K_N L_s K_N \alpha / 2 \right\}, \quad (6)$$

$$\text{s.t. } y_i \left(\sum_{j=1}^N \alpha_j K(x_i, x_j) + b \right) \geq 1 - \xi_i, i=1, \dots, l.$$

定理 2 非线性 SRMCV 方法的原始优化问题(6)的对偶问题为:

$$\begin{aligned} \min_{\beta} & \frac{1}{2} \beta^T \tilde{M} \beta - 1_k^T \beta, \\ \text{s.t. } & \sum_{i=1}^l y_i \beta_i = 0, 0 \leq \beta_i \leq \frac{1}{l}, \end{aligned} \quad (7)$$

其中, $\tilde{M} = \tilde{Y} K_l^T (\gamma_w K_l K_l^T + \gamma_s K_N L_s K_N)^{-1} K_l \tilde{Y}$, $\tilde{Y} = \text{diag}(y_1, \dots, y_l)$ 表示对角矩阵.

同样, 可按照文献[5]方法来证明定理 2, 且根据定理 2 的证明, 可知:

$$\alpha = (\gamma_w K_l K_l^T + \gamma_s K_N L_s K_N)^{-1} K_l \tilde{Y} \beta.$$

从而通过目标问题(7)的优化解 β^* 来求解目标问题(6)中优化变量 α^* , 进而获得优化法向量 w^* . 按照线性 SRMCV 方法求得非线性方法偏置 $\tilde{b}$ 为:

$$\tilde{b}^* = \frac{1}{l} \sum_{i=1}^l (y_i - \sum_{j=1}^N \alpha_j^* K(x_i, x_j)).$$

因此对于测试数据 $x \in X_u$, 非线性 SRMCV 的决策函数为:

$$\tilde{f}(x) = \sum_{i=1}^N \alpha_i K(x_i, x_j) + \tilde{b}^*.$$

2 实验分析

为评价本文方法的有效性, 将所提方法分别在 UCI 数据集^[7]和人脸数据集上进行测试, 并与相关方法 SVM、MCVSVM^[4]、GLSSVM^[5]和 S-RLSC^[8]进行比较. 所有的实验样本先归一化为 $[-1, 1]$. 而非线性映射使用高斯核函数 $K(x, y) = \exp(-\|x - y\|^2 / \sigma^2)$ ^[9], 其中 σ 是训练样本的平均范数平方根, γ 值在集合 $\{\sigma^2/8, \sigma^2/4, \sigma^2/2, \sigma^2, \sigma^2*2, \sigma^2*4\}$ 之中选取; 方法 SVM、MCVSVM、GLSSVM 还有 S-RLSC 参数的确定将分别使用与各个方法的参考文献中同样的策略, 即优化的实验参数是根据最好交叉验证参数集的平均值来确定. 同时, 基于数据集总体样本五重交叉验证精度来评价全部分类方法的泛化性能, 取其平均值作为最后实验结果. 实验中, LibSVM 软件实现 SVM 算法, 并在 Matlab 2010b 环境下实现其他算法.

2.1 UCI 数据集实验

测试模式分类方法的性能通常使用 UCI 数据

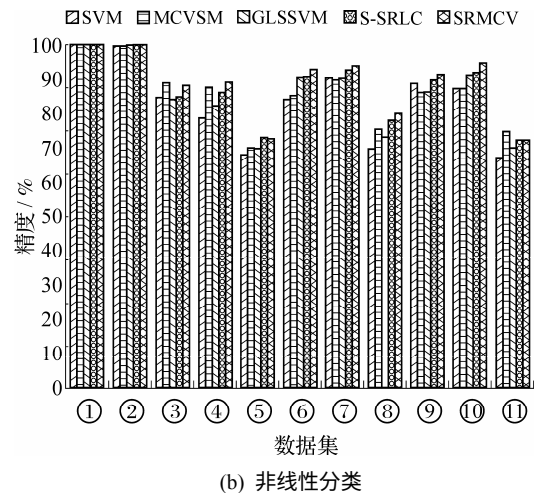
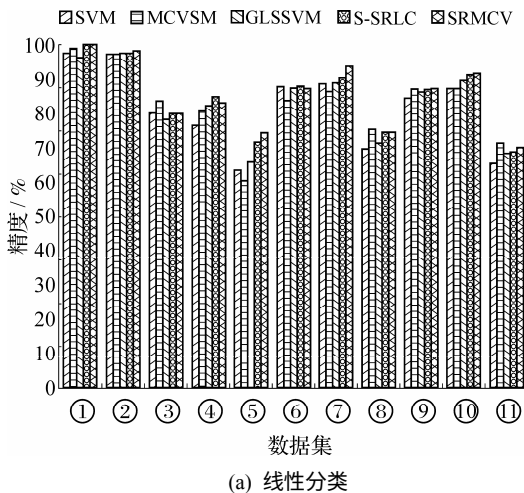


图 1 UCI 数据集分类精度比较

集. 笔者是采用文献[9]中的UCI数据集, 通过选取其中 11 个典型的数据集来进行实验, 并分别测试线性及非线性情况下 5 种方法(SVM、MCVSVM、GLSSVM、S-SRLC、SRMCV)的分类性能. 图 1 显示在线性和非线性情况下所提方法与相关方法的模式分类精度比较(图中带圈数字代表的数据集分别如下: ①Iris; ②Breast; ③Heart; ④Vehicle; ⑤Glass; ⑥Ionosphere; ⑦Waveform; ⑧Sonar; ⑨Hepatitis; ⑩Biomed; ⑪Liver). 从图 1 可看出: (1)采用核方法能明显增强 4 种方法的分类性能; (2)在线性或非线性情况下, 所提方法 SRMCV 均取得了优于其他方法的性能, 几种方法在绝大多数数据集上的分类效果在一定程度均优于传统的 SVM 方法, 这也验证了如果忽视数据空间的结构信息尤其是稀疏表示信息, 而仅仅是考虑数据整体类之间的最大间隔, 并不能取得较优的分类性能; (3)与 SVM、MCVSVM、GLSSVM 等方法相比, 所提核方法具有明显的模式分类优势, 而与 S-SRLC 方法相比, 文中方法具有可比较的性能, 这进一步说明, 在充分考虑数据空间的稀疏几何结构信息和类内分布信息的情况下, 分类方法能取得更优的模式分类性能.

2.2 人脸识别实验

实验将在 2 个人脸识别数据库(ORL 数据库和 Yale 数据库^[10])上进一步评价所提方法的优化性能, 其中 ORL 人脸数据集使用 40 个对象的人脸作为数据, 由不同表情的 10 张图片构成每类对象. YALE 人脸数据集包含 15 个类别 165 幅灰度级的图像, 同类是由不同脸部表情和光照条件的 11 张人脸数

据构成. 在进行实验前, 对图像集先进行预处理, 缩放到 32×32 的像素大小, 且每个像素 256 灰度级, 1 个 1024 维向量表示每张图像, 图 2(a)和(b)分别显示 2 个数据库中 1 个对象的人脸图像样本.

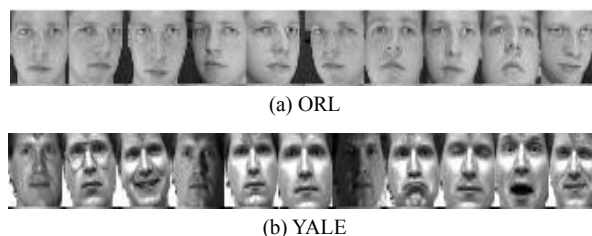


图 2 从 2 个人脸图像数据库中分别选取的人脸样本

在实验中, 经该核函数高维映射后, 本文方法、GLSSVM 和 MCVSVM 中类内散度矩阵均为奇异型, 因此, 本实验先利用 PCA 或 KPCA^[11]将原始空间数据投影到某个子空间, 以使得上述奇异型矩阵变成非奇异型. 为了对各方法进行有效比较, 实验前将数据降维到同样维度. 由于训练样本之间线性无关, 以 N 为训练样本数, 将样本投射到 $N-2$ 维空间.

图 3 显示所提方法与几种分类方法在 ORL 和 YALE 数据集上的识别精度在不同维度下的比较. 从图 3(a)和(b)可看出, 较其他方法, 文中方法在所有维度下均具有优于或可比较的识别性能. 随着维度的增大, 所有方法的识别率均上升. 另外, 即使在低维段, 文中方法也具备一定的性能优势.

3 结语

针对半监督学习环境, 基于稀疏表示技术, 从充分考虑最小类内散度和全局数据稀疏表示结构

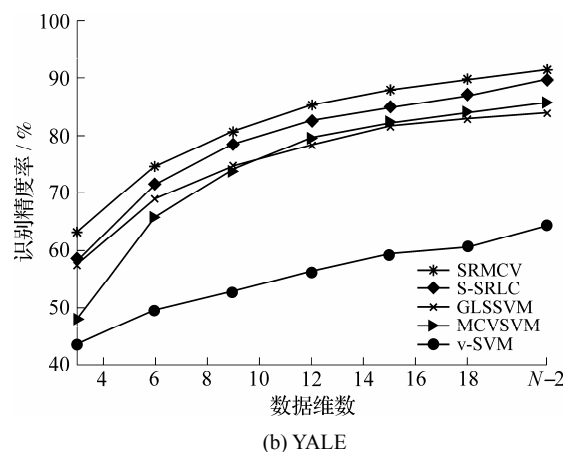
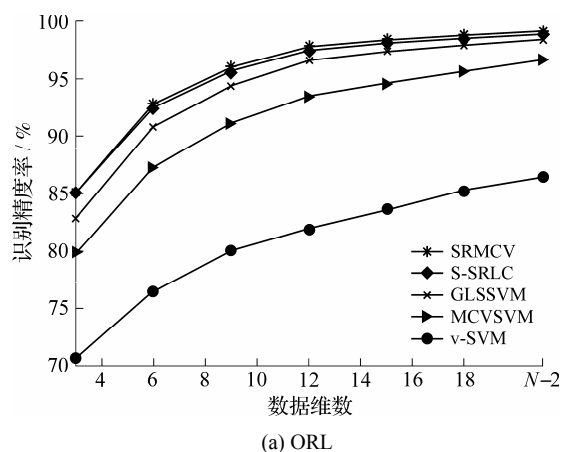


图 3 ORL 和 YALE 人脸数据集上的识别性能比较

信息的角度, 提出一种稀疏正则化最小类半监督分类机, 其能有效克服现有基于图模型半监督学习方法存在的问题. 在实际 UCI 和人脸数据集上实验证实了所提方法的鲁棒有效性. 试验中, 如何能更加有效地协调选取参数 γ_w 和 γ_s 的取值优化是有待进一步研究的问题.

参考文献:

- [1] Belkin M, Niyogi P, Sindhvani V. Manifold regularization: A geometric framework for learning from examples [J]. Journal of Machine Learning Research, 2006(7): 2399-2434.
- [2] Zhu Xiaojin. Semi-supervised learning literature survey [R]. University of Wisconsin-Madison: Computer Sciences Technical Report 1530, 2005:1-60.
- [3] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge: Cambridge University Press, 2000:235-248.
- [4] Zafeiriou S, Tefas A, Pitas I. Minimum class variance support vector machines[J]. IEEE Transactions on Image processing, 2007, 16(10):2551-2564.
- [5] 皋军, 王士同, 邓赵红. 基于全局和局部保持的半监督支持向量机[J]. 电子学报, 2010(7):1626-1633.
- [6] Wright J, Ma Y, Mairal J, et al. Sparse representation for computer vision and pattern recognition[J]. Proceedings of IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation, 2010, 98(6):1031-1044.
- [7] Shivaswamy P, Jebara T. Maximum relative margin and data-dependent regularization[J]. Journal of Machine Learning Research, 2010, 11:747-788.
- [8] Fan Mingyu, Gu Nannan, Qiao Hong, et al. Sparse regularization for semi-supervised classification[J]. Pattern Recognition, 2011, 44(8):1777-1784.
- [9] 陶剑文, 王士同. 局部保留最大信息差 v -支持向量机 [J]. 自动化学报, 2012, 38(1):97-108.
- [10] He Xiaofei, Yan Shuicheng, Hu Yuxiao, et al. Face recognition using Laplacianfaces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (3):328-340.
- [11] Scholkopf B, Mika S, Burges C J C, et al. Input space vs feature space in Kernel-based methods[J]. IEEE Transactions on Neural Network, 1999, 10(5):1000-1017.

Sparse Regularization for Minimum Class Variance Semi-supervised Classifier

LIU Jian-hua, WU Dong-yan

(School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012, China)

Abstract: This paper presents a sparse regularization minimum class scatter semi-supervised classifier (SRMCV) based on the theory of sparse representation. For the problems of pattern classification, SRMCV aims at achieving smooth changes of the predicted spatial function in the structure of the global sparse representation by introducing sparse Laplacian regularization term and the information of inner class scatter. In addition, through the scatter structure of inner class data, it is intended to further optimize the discriminant direction of the decision-making function. This method can solve some problems existing in the SSL method such as sensitivity to the model parameters, the lack of robustness in the noisy learning environment, etc. On the real data set, the experimental results manifest the effectiveness of the proposed method.

Key words: semi-supervised learning; sparse representation; support vector machine; minimum class scatter; kernel trick

(责任编辑 章践立)