

# 最优间隔分布脊回归

陈加略 姜远

(计算机软件新技术国家重点实验室(南京大学) 南京 210023)

(软件新技术与产业化协同创新中心 南京 210023)

(chenjl@lamda.nju.edu.cn)

## Optimal Margin Distribution Ridge Regression

Chen Jialüe and Jiang Yuan

(National Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023)

(Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023)

**Abstract** Ridge regression (RR) has been one of the most classical machine learning algorithms in many real applications such as face detection, cell prediction, etc. The ridge regression has many advantages such as convex optimization objection, closed-form solution, strong interpretability, easy to kernelization and so on. But the optimization objection of ridge regression doesn't consider the structural relationship between instances. Supervised manifold regularized (MR) method has been one of the most representative and successful ridge regression regularized methods, which considers the instance structural relationship inter each class by minimizing each class's variance. But considering the structural relationship interclasses alone is not a very comprehensive idea. Based on the recent principle of optimal margin distribution machine (ODM) learning with a novel view, we find the optimization object of ODM can include the local structural relationship and the global structural relationship by optimizing the margin variance interclasses and the margin variance intraclasses. In this thesis, we propose a ridge regression algorithm called optimal margin distribution machine ridge regression (ODMRR) which fully considers the structural character of the instance. Besides, this algorithm can still contain all the advantages of ridge regression and manifold regularized ridge regression. Finally, the experiments validate the effectiveness of our algorithm.

**Key words** ridge regression (RR); manifold regularization; optimal margin distribution machine (ODM); margin variance; global structure

**摘要** 脊回归(ridge regression, RR)是经典的机器学习算法之一,广泛应用于人脸识别、基因工程等诸多领域. 它具有优化目标凸、存在闭合解、可解释性强以及易于核化等优点,但是脊回归的优化目标并没有考虑样本之间的结构关系. 监督流形正则化学习是最具代表性的、最成功的脊回归正则化方法之一,其通过最小化每类类内方差来考虑样本之间的类内结构关系,可是单纯地只考虑类内结构仍然不够全面. 以一种全新的视角重新审视最近提出的“最优间隔分布学习”原理,发现了最优间隔分布的目标可以同时优化类内间隔方差和类间间隔方差,从而同时优化了局部的类内结构和全局的类间结构. 基于此提出了一种充分考虑数据结构化特征的脊回归算法——最优间隔分布脊回归(optimal margin distribution

收稿日期:2017-05-23;修回日期:2017-06-25

基金项目:国家自然科学基金项目(61673201)

This work was supported by the National Natural Science Foundation of China (61673201).

通信作者:姜远(jiangy@lamda.nju.edu.cn)

machine ridge regression, ODMRR)算法,该算法具有RR以及MRRR(manifold regularization ridge regression)的各种优势.最后通过实验验证了该方法具有优越的性能.

**关键词** 脊回归;流形正则化;最优间隔分布;间隔方差;全局结构

**中图分类号** TP181

脊回归(ridge regression, RR)<sup>[1]</sup>的优化目标致力于最小化模型输出以及样本标记之间的二次误差,在人脸识别<sup>[2]</sup>、基因工程<sup>[3]</sup>、肝脏致癌预测<sup>[4]</sup>等诸多领域<sup>[5]</sup>有着广泛的应用.脊回归具有优化目标凸、存在闭式解、可解释性强以及易于核化等优点.但其优化目标并没有体现出数据的结构(structural)特征,这就意味着原始的脊回归算法没有考虑分类样本之间存在的潜在相互关系.在分类模型的优化目标中加入正则化约束来假设数据之间的结构关系通常可以有效地提升性能.流形正则化(manifold regularization, MR)<sup>[6]</sup>是一种常见的做法,在诸多领域得到了广泛应用,如多标记学习<sup>[7]</sup>、集成学习<sup>[8]</sup>、特征提取<sup>[9]</sup>、谱图理论<sup>[10]</sup>、网页图像标记<sup>[11]</sup>等. MR是一种基于半监督<sup>[12]</sup>流形假设的适用于有标记和无标记样本学习的正则化框架,它的优化目标是 minimized 每类样本的类内方差,即要求分类器输出类内紧凑.将MR加入到有监督RR后的优化目标也是凸的,并且存在闭合解,因此具备脊回归算法所具有的各种优势.

相比于最大化最小间隔<sup>[13]</sup>,最优间隔分布学习理论<sup>[14-15]</sup>的核心思想在于通过优化“间隔分布”来有效提升模型的泛化性能.由此提出的ODM<sup>[16]</sup>、LDM<sup>[17]</sup>和cisLDM<sup>[18]</sup>算法都取得了很好的性能<sup>①</sup>.在最优间隔分布学习理论中,优化间隔分布是通过最大化间隔均值并同时最小化间隔方差<sup>[13]</sup>来实现的.本文选择了一种崭新的视角,从优化数据结构特征的角度重新分析了最优间隔理论中优化间隔分布的目标,并发现最小化间隔方差这个目标可以拆解成最小化类内间隔方差和最大化类间间隔方差2项.针对分类模型进行分析,可以发现,MR中最小化类内方差的目标和最优间隔分布学习理论中最小化间隔方差的目标是相同的,都要求类内紧凑,即两者都是在优化数据的局部类内结构.进一步对最优间隔学习理论中最大化间隔均值和最小化类间间隔方差这2项的目标进行分析,可以发现两者共同作用的目的都是为了使分类器输出类间松散.而类间

松散可以看成是优化一种全局的结构关系.所以,MR的优化目标只优化了类内紧凑这一种局部分类特性,而最优间隔分布学习理论不仅考虑了类内紧凑,还进一步优化了类间松散这一种全局分类特性.换句话说,相比于MR,最优间隔分布学习理论考虑了更加全面的数据结构信息.

在本文中,我们借鉴了最优间隔分布学习理论思想,并提出了一种新的正则化方法即最优间隔分布正则化.流形正则化方法可看作为最优间隔正则化方法的特例.我们将其应用于脊回归中得到了最优间隔分布脊回归(optimal margin distribution machine ridge regression, ODMRR)算法,ODMRR算法的优化目标也是凸的,并且具有解析解,在多个真实数据集上取得很好效果.

## 1 研究背景

假设 $\mathcal{X} \in \mathbb{R}^d$ 是样本空间, $\mathcal{Y} = \{+1, -1\}$ 是标记集合, $\mathcal{D}$ 是 $\mathcal{X} \times \mathcal{Y}$ 上的潜在分布,训练集是从 $\mathcal{D}$ 上随机独立同分布采样的 $n$ 个样本,即:

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}.$$

机器学习的目标就是通过训练集 $S$ 学习一个分类器,使其能够预测未来从未见过的样本标记.

假设 $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b$ 是一个线性判别模型.其中 $\mathbf{w}$ 是线性分类面, $b$ 是截距, $\phi(\mathbf{x})$ 是 $\mathbf{x}$ 在核 $k$ 上的一个特征投影,并满足 $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ .样本 $(\mathbf{x}_i, y_i)$ 的间隔可以定义为

$$\gamma_i = y_i (\mathbf{w}^\top \phi(\mathbf{x}_i) + b), \quad \forall i = 1, 2, \dots, n.$$

为了方便表示,可以将分类面 $\mathbf{w}$ 以及截据 $b$ 拼接成一个增广的向量: $\mathbf{w}' = [\mathbf{w}; b]$ ,将所有样本的特征尾部拼上一个常数1形成增广的特征 $\phi(\mathbf{x}_i)' = [\phi(\mathbf{x}_i); 1]$ .令 $\mathbf{X}' = [\phi(\mathbf{x}_1)', \phi(\mathbf{x}_2)', \dots, \phi(\mathbf{x}_n)']$ , $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$ .

对于线性判别分类器,其解 $\mathbf{w}^*$ 通常存在表示形式:

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) = \mathbf{X} \boldsymbol{\alpha}, \quad (1)$$

① 此类方法以往被称为“大间隔分布学习”方法<sup>[17-18]</sup>,但由于“间隔分布”本身并没有“大小”之别,因此“最优间隔分布”这个称谓比“最大间隔分布”更合适<sup>[16]</sup>.

其中,  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$  是系数向量. 可以将系数向量  $\alpha$  和截距  $b$  拼成一个增广的向量:  $\alpha' = [\alpha; b]$ .

为了方便将算法推广到核空间上, 我们用  $K$  表示  $n \times n$  的核矩阵, 其第  $(i, j)$  个位置的元素为  $k(x_i, x_j)$ , 可以发现  $K = X^T X$ . 在  $K$  的最后一行拼上  $K' = [K; \mathbf{1}_{1 \times n}]$ , 其中  $\mathbf{1}_{m \times n}$  表示维度为  $m \times n$  的全 1 矩阵.

### 1.1 脊回归和流形正则化

脊回归 (ridge regression, RR) 的优化目标是找到一个分类面  $w$  和截距  $b$ , 使得样本输出  $f(x_i)$  和样本标记  $y_i$  之间的二次误差之和尽量小. 其优化目标可以形式化为

$$\min_w \sum_{i=1}^n (y_i - w'^T \phi(x_i))'^2 + \lambda w'^T A w', \quad (2)$$

其中,  $\lambda$  是正则化因子,  $A = \begin{pmatrix} \mathbf{I}_{d \times d} & 0 \\ 0 & 0 \end{pmatrix}_{(n+1) \times (n+1)}$ . 第 1 项的目标即是二次误差项, 第 2 项则是为了减小模型的结构风险.

式(2)中的目标可以通过式(1)的表示定理转化成等价的形式:

$$\min_{\alpha'} y^T y - 2y^T K' \alpha' + \alpha'^T K' K'^T \alpha' + \lambda \alpha'^T B \alpha'. \quad (3)$$

式(3)的优化目标是凸的, 并且具有闭合的解析解, 其解的形式为

$$\alpha' = (\lambda B + K' K'^T)^{-1} K' y,$$

其中,  $B = \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix}_{(n+1) \times (n+1)}$ .

流形假设数据采样于高维欧式空间的一个低维流形, 处于局部领域的示例具有相似的性质, 即它们的标记也应该相似. 对于分类问题, 流形正则化的目标是减小 2 类样本的类内方差. 所以, 监督型流形正则化的脊回归 (manifold regularization ridge regression, MRRR) 的优化目标表示为

$$\begin{aligned} \min_w \sum_{i=1}^n (y_i - w'^T \phi(x_i))'^2 + \lambda_1 w'^T A w' + \\ \lambda_2 \left( \sum_{i \in C_1} \sum_{j \in C_1} (w'^T \phi(x_i))' - w'^T \phi(x_j) \right)' ^2 + \\ \sum_{i \in C_2} \sum_{j \in C_2} (w'^T \phi(x_i))' - w'^T \phi(x_j) \right)' ^2, \quad (4) \end{aligned}$$

其中,  $\lambda_1$  和  $\lambda_2$  是正则化因子,  $C_1$  和  $C_2$  分别代表第 1 和第 2 类样本. 式(4)中最后 2 项就是流形正则化项, 它们分别优化了第 1, 2 类样本的类内方差. 可以发现, 如果式(4)中的  $\lambda_2 = 0$ , MRRR 将会退化为 RR 算法, 即 RR 是 MRRR 的特例.

式(4)中的目标也可以通过式(1)中的表示定理转化成等价形式:

$$\min_{\alpha'} y^T y - 2y^T K' \alpha' + \alpha' K' (I_n + 2\lambda_2 M) K'^T \alpha' + \lambda_1 \alpha'^T B \alpha', \quad (5)$$

其中,  $I_n$  表示维度为  $n$  的单位阵, 而:

$$M = \begin{pmatrix} n_1 I_{n_1} - \mathbf{1}_{n_1 \times n_1} & 0_{n_1, n_2} \\ 0_{n_2, n_1} & n_2 I_{n_2} - \mathbf{1}_{n_2 \times n_2} \end{pmatrix}_{n \times n},$$

$n_1$  和  $n_2$  分别表示第 1, 2 类样本的数目.

式(5)的目标也是凸的, 并且也具有闭合的解析解, 其解的形式为

$$\alpha' = (\lambda_1 B + K' (I_n + 2\lambda_2 M) K'^T)^{-1} K' y.$$

### 1.2 最优间隔分布学习理论

相比于最大化最小间隔, 最优间隔分布学习理论揭示了优化间隔分布对于提升模型的泛化性能将更加显著. 最优间隔分布学习理论指出优化间隔分布具体表现为最大化间隔均值和最小化间隔方差. 其中, 间隔均值定义为

$$\frac{1}{n} \sum_{i=1}^n y_i w'^T \phi(x_i)' = \frac{1}{n} y^T X' w'.$$

间隔方差定义为

$$\begin{aligned} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (y_i w'^T \phi(x_i)' - y_j w'^T \phi(x_j)')^2 = \\ \frac{2}{n^2} (n w'^T X' X'^T w' - w'^T X' y y^T X'^T w'). \end{aligned}$$

根据间隔方差的定义可以发现, 当  $y_i = y_j$  时, 间隔方差的形式和流形正则化项的形式相同, 即对于分类模型, 类内方差即是类内的间隔方差. 两者的目的都是为了使分类器输出类内紧凑.

当  $y_i \neq y_j$  时, 间隔方差的形式为

$$\frac{2}{n^2} \sum_{i \in C_1} \sum_{j \in C_2} (w'^T \phi(x_i))' + w'^T \phi(x_j)')^2,$$

经过分析可以发现, 这一项实际上代表了类间间隔方差. 最小化间隔方差和最大化间隔均值两者共同作用的目标是为了使分类器输出类间分散.

所以最优间隔分布学习理论既考虑了使分类器输出类内紧凑这一局部分类特性, 还考虑了使分类器输出类间分散这一全局分类特性.

## 2 最优间隔分布脊回归

最优间隔分布 (ODM) 的主要思想就是利用最优间隔分布学习理论中同时考虑分类器输出类内紧凑、类间松散这 2 个分类特性的优势. 将最大化间隔均值和最小化间隔方差作为正则化项加入到分类模型的优化目标中.

脊回归的二次误差项表示为等价的矩阵形式:

$$\sum_{i=1}^n (y_i - \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i))'^2 = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}'^T \mathbf{w}' + \mathbf{w}'^T \mathbf{X}' \mathbf{X}'^T \mathbf{w}',$$

可以发现,最小化 $-2\mathbf{y}^T \mathbf{X}'^T \mathbf{w}'$ 的目标就是最大化间隔均值,即脊回归的优化目标中本身就包含了最大化间隔均值。所以,对于脊回归问题,最优间隔分布只需要考虑最小化间隔方差一项正则化项就可以了。于是,我们将 ODM 运用于脊回归问题,并得到了 ODMRR 的优化目标:

$$\min_{\mathbf{w}'} \sum_{i=1}^n (y_i - \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i))'^2 + \lambda_1 \mathbf{w}'^T \mathbf{A} \mathbf{w}' + \lambda_2 \sum_{i=1}^n \sum_{j=1}^n s_{ij} (y_i \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i)' - y_j \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_j))'^2 = \min_{\mathbf{w}'} \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}'^T \mathbf{w}' + \mathbf{w}'^T \mathbf{X}' (\mathbf{I}_n + 2\lambda_2 (\mathbf{D} \circ \text{diag}(\mathbf{y} \mathbf{y}^T) - \mathbf{S} \circ \mathbf{y} \mathbf{y}^T)) \mathbf{X}'^T \mathbf{w}' + \lambda_1 \mathbf{w}'^T \mathbf{A} \mathbf{w}', \quad (6)$$

其中,diag 函数的作用是用原矩阵的对角元素组成同维的新矩阵,“ $\circ$ ”表示 Hadamard 乘积, $s_{ij}$  表示样本  $i$  和样本  $j$  之间的相似性度量。 $s_{ij}$  可以采用等价距离,即  $s_{ij} = 1$ ,那么此时第 3 项就是间隔方差项。也可以采用高斯距离,其定义为  $s_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$ 。 $\mathbf{S}$  是  $n \times n$  维的相似性矩阵,其第  $(i, j)$  个元素即为  $s_{ij}$ 。 $\mathbf{D}$  是  $n \times n$  维的对角矩阵,其第  $i$  个对角元素为  $\sum_{j=1}^n s_{ij}$ 。

**定理 1.** 优化目标(6)中的最优解  $\mathbf{w}^*$  存在式(1)的表示形式。

证明。 $\mathbf{w}^*$  可以分解成有  $\boldsymbol{\phi}(\mathbf{x}_i)$  伸张以及和  $\boldsymbol{\phi}(\mathbf{x}_i)$  正交 2 个部分,形式为

$$\mathbf{w} = \sum_{i=1}^n \boldsymbol{\alpha}_i \boldsymbol{\phi}(\mathbf{x}_i) + \mathbf{v} = \mathbf{X} \boldsymbol{\alpha} + \mathbf{v},$$

其中,  $\mathbf{X}^T \mathbf{v} = 0$ 。可以发现:

$$\mathbf{X}'^T \mathbf{w}' = \mathbf{X}'^T \mathbf{w} + \mathbf{b} = \mathbf{X}'^T (\mathbf{X} \boldsymbol{\alpha} + \mathbf{v}) + \mathbf{b} = \mathbf{X}'^T \mathbf{X} \boldsymbol{\alpha} + \mathbf{b},$$

所以,在优化目标式(6)中,只有  $\mathbf{w}'^T \mathbf{A} \mathbf{w}'$  与  $\mathbf{v}$  相关,其他项都与  $\mathbf{v}$  独立,可以证明:

$$\mathbf{w}'^T \mathbf{A} \mathbf{w}' = \mathbf{w}'^T \mathbf{w} = (\mathbf{X} \boldsymbol{\alpha} + \mathbf{v})^T (\mathbf{X} \boldsymbol{\alpha} + \mathbf{v}) = \boldsymbol{\alpha}^T \mathbf{X}'^T \mathbf{X} \boldsymbol{\alpha} + \mathbf{v}^T \mathbf{v} \geq \boldsymbol{\alpha}^T \mathbf{X}'^T \mathbf{X} \boldsymbol{\alpha}$$

且不等式当且仅当  $\mathbf{v} = 0$  时才成立。

所以,设置  $\mathbf{v} = 0$  不会影响到优化目标式(6)的最优值,即  $\mathbf{w}^*$  可以表示成式(1)的表示形式。证毕。

于是,通过定理 1,可以将优化目标式(6)转换为等价的形式:

$$\min_{\mathbf{w}'} \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{K}'^T \boldsymbol{\alpha}' + \boldsymbol{\alpha}'^T \mathbf{K}' (\mathbf{I}_n + 2\lambda_2 (\mathbf{D} \circ \text{diag}(\mathbf{y} \mathbf{y}^T) - \mathbf{S} \circ \mathbf{y} \mathbf{y}^T)) \mathbf{K}'^T \boldsymbol{\alpha}' + \lambda_1 \boldsymbol{\alpha}'^T \mathbf{B} \boldsymbol{\alpha}'. \quad (7)$$

式(7)的优化目标是凸的,也具有闭合和解析解,其解的形式为

$$\boldsymbol{\alpha}' = (\lambda_1 \mathbf{B} + \mathbf{K}' (\mathbf{I}_n + 2\lambda_2 (\mathbf{D} \circ \text{diag}(\mathbf{y} \mathbf{y}^T) - \mathbf{S} \circ \mathbf{y} \mathbf{y}^T)) \mathbf{K}'^T)^{-1} \mathbf{K}'^T \mathbf{y}.$$

为了能够控制类内间隔方差和和类间间隔方差的权重,也可以把优化目标等价表示为

$$\min_{\mathbf{w}'} \sum_{i=1}^n (y_i - \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i))'^2 + \lambda_1 \mathbf{w}'^T \mathbf{A} \mathbf{w}' + \lambda_2 \sum_{i \in C_1} \sum_{j \in C_1} s_{ij} (\mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i)' - \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_j))'^2 + \sum_{i \in C_2} \sum_{j \in C_2} s_{ij} (\mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i)' - \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_j))'^2 + 2\lambda_3 \sum_{i \in C_1} \sum_{j \in C_2} s_{ij} (\mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_i)' + \mathbf{w}'^T \boldsymbol{\phi}(\mathbf{x}_j))'^2, \quad (8)$$

其中, $\lambda_1, \lambda_2, \lambda_3$  全是正则化因子,式(8)中第 2 项是第 1 类的类内间隔分布,第 3 项是第 2 类的类内间隔分布,最后一项是 2 类样本的类间间隔分布,我们将式(8)的优化目标记为 ODMRR-3h。可以发现,如果  $\lambda_3 = 0$ ,式(8)的优化目标就会退化为式(4)流形正则化方法的优化目标,如果  $\lambda_2 = 0$  和  $\lambda_3 = 0$ ,式(8)的优化目标将会退化为式(2)的脊回归的优化目标。所以脊回归算法以及流形正则化的脊回归算法都可以看作最优间隔正则脊回归算法的特例。

根据定理 1 的证明,式(8)也存在等价的表示形式:

$$\min_{\boldsymbol{\alpha}'} \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{K}'^T \boldsymbol{\alpha}' + \boldsymbol{\alpha}'^T (\lambda_1 \mathbf{B} + \mathbf{K}' (\mathbf{I}_n + \mathbf{P} + \mathbf{Q}) \mathbf{K}'^T) \boldsymbol{\alpha}', \quad (9)$$

其中,

$$\mathbf{P} = \begin{bmatrix} 2\lambda_2 (\mathbf{D}_{11} \circ \text{diag}(\mathbf{y}_1 \mathbf{y}_1^T) - \mathbf{S}_{11} \circ \mathbf{y}_1 \mathbf{y}_1^T) & & \\ & 0 & \\ & & \ddots & \\ & & & 0 & \\ & & & & 2\lambda_2 (\mathbf{D}_{22} \circ \text{diag}(\mathbf{y}_2 \mathbf{y}_2^T) - \mathbf{S}_{22} \circ \mathbf{y}_2 \mathbf{y}_2^T) \end{bmatrix}_{n \times n_1} + \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & 2\lambda_2 (\mathbf{D}_{22} \circ \text{diag}(\mathbf{y}_2 \mathbf{y}_2^T) - \mathbf{S}_{22} \circ \mathbf{y}_2 \mathbf{y}_2^T) \end{bmatrix}_{n \times n_2},$$

$$\mathbf{Q} = \begin{bmatrix} -2\lambda_3 (\mathbf{S}_{12} \circ \mathbf{y}_1 \mathbf{y}_2^T) & & \\ & & \\ & & \\ & & \\ & & 2\lambda_3 \mathbf{D}_{12} \circ \text{diag}(\mathbf{y}_2 \mathbf{y}_2^T) \end{bmatrix}_{n \times n_1} + \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & 2\lambda_3 \mathbf{D}_{21} \circ \text{diag}(\mathbf{y}_1 \mathbf{y}_1^T) - 2\lambda_3 (\mathbf{S}_{12} \circ \mathbf{y}_1 \mathbf{y}_2^T) \end{bmatrix}_{n \times n_2},$$

$\mathbf{y}_i$  为第  $i$  类样本的标记向量, $\mathbf{S}_{ij}$  为第  $i$  类样本和第  $j$  类样本之间的相似性矩阵,矩阵  $\mathbf{D}$  是一个对角矩阵,其第  $j$  个对角元上的元素是  $\mathbf{S}$  矩阵中第  $j$  列中所有行下标  $i$  类别为  $p$  的元素之和,所以其定义式可以

$$\text{写为 } \mathbf{D}_{pq} = \begin{bmatrix} \sum_{i \in C_p} s_{i1} & & \cdots & \\ & \ddots & & \\ \vdots & & \sum_{i \in C_p} s_{ij} & \vdots \\ & & & \ddots & \\ & & & & \sum_{i \in C_p} s_{in_q} \end{bmatrix}_{n_p \times n_p}.$$

式(9)的优化目标为凸,其解也为解析解,其形式为

$$\alpha' = (\lambda_1 \mathbf{B} + \mathbf{K}'(\mathbf{I}_n + \mathbf{P} + \mathbf{Q})\mathbf{K}'^T)^{-1} \mathbf{K}' \mathbf{y}.$$

## 4 实验测试

本节将先介绍实验的设置,然后给出实验的结果,并对结果进行简单地分析.

### 4.1 实验设置

实验共选择了 14 个数据集,其中 digit1 和 g241n 是半监督数据集. austra, isolet, vehicle 和 wdbc 都是 UCI 数据集. Letter 是多标记数据集,从中选出

8 对相对比较难分类的字符组成了 8 组 2 分类数据集,分别记为 DvsP, EvsF, IvsJ, IvsL, MvsN, UvsY, VvsY, DvsO. 表 1 中列出了所有数据集的样本和特征维度. 并且所有数据集都通过随机采样以 1:1 的比例平分成训练数据和测试数据.

实验 1 共选择了 6 组对比算法: RR, MRRR, ODMRR, ODMRR\_3h, ODMRR\_S1, ODMRR\_S1\_3h. 其中,在 ODMRR 算法中,带有\_S1 后缀代表相似性矩阵  $\mathbf{S}$  采用全 1 矩阵,没有\_S1 后缀的相似性矩阵  $\mathbf{S}$  则采用高斯距离矩阵. 带有\_3h 的后缀表示优化目标采用式(9)的表示形式,否则采用式(6)的表示形式.

Table 1 Dataset and the Dimension of Its Instances and Features

表 1 数据集及其样本和特征维度

Datasets	# Instance × # Feature	Datasets	# Instance × # Feature
letterDvsP	1 610 × 16	austra	690 × 12
letterEvsF	1 543 × 16	digit1	1 500 × 241
letterIvsJ	1 502 × 16	g241n	1 500 × 241
letterIvsL	1 608 × 16	isolet	600 × 617
letterMvsN	1 575 × 16	vehicle	435 × 18
letterUvsV	1 577 × 16	wdbc	569 × 30
letterVvsY	1 550 × 16	letterDvsO	1 558 × 16

### 4.2 实验结果

表 2 中的实验结果是在每个数据集上进行 10 次实验所得平均精度和方差. 其中,深色背景都是我们算法的实验结果. 我们使用加粗字体来标明每

个数据集中测试精度最高的一组算法的结果.

在每轮实验中,所有参数全部都是采用固定其余所有参数,然后通过 10 次交叉验证的方法选出的最优值. 在交叉验证的过程中,选择 80% 的训练数

Table 2 Experiment Result

表 2 实验结果

Datasets	Total Costs (mean ± std. )					
	RR	MRRR	ODMRR	ODMRR_3h	ODMRR_S1	ODMRR_S1_3h
austra	0. 864 1 ± 0. 01	<b>0. 864 9 ± 0. 01</b>	0. 863 2 ± 0. 01	0. 863 5 ± 0. 02	0. 862 3 ± 0. 01	0. 864 3 ± 0. 01
digit1	0. 947 9 ± 0. 008	0. 947 7 ± 0. 008	<b>0. 949 3 ± 0. 008</b>	0. 948 1 ± 0. 008	0. 947 7 ± 0. 008	0. 946 1 ± 0. 009
g241n	0. 856 5 ± 0. 01	0. 856 9 ± 0. 01	<b>0. 858 8 ± 0. 01</b>	0. 857 9 ± 0. 01	0. 850 8 ± 0. 01	0. 850 8 ± 0. 02
isolet	<b>0. 996 7 ± 0. 004</b>	0. 996 3 ± 0. 004	0. 993 3 ± 0. 009	0. 996 3 ± 0. 005	0. 995 ± 0. 005	0. 993 7 ± 0. 006
letterDvsO	<b>0. 964 8 ± 0. 004</b>	0. 964 6 ± 0. 004	0. 964 2 ± 0. 004	<b>0. 964 8 ± 0. 005</b>	<b>0. 964 8 ± 0. 004</b>	0. 964 4 ± 0. 005
letterDvsP	<b>0. 977 2 ± 0. 005</b>	0. 977 1 ± 0. 005	0. 977 ± 0. 005	0. 976 6 ± 0. 006	0. 977 1 ± 0. 005	0. 977 1 ± 0. 005
letterEvsF	0. 958 1 ± 0. 007	0. 958 2 ± 0. 007	0. 958 4 ± 0. 007	<b>0. 958 8 ± 0. 007</b>	0. 958 ± 0. 007	0. 958 4 ± 0. 005
letterIvsJ	0. 922 ± 0. 007	0. 921 2 ± 0. 007	0. 920 8 ± 0. 009	0. 921 1 ± 0. 007	0. 921 9 ± 0. 007	<b>0. 922 7 ± 0. 006</b>
letterIvsL	0. 977 7 ± 0. 007	0. 977 5 ± 0. 007	0. 977 5 ± 0. 007	0. 977 4 ± 0. 007	<b>0. 977 8 ± 0. 007</b>	0. 977 3 ± 0. 007
letterMvsN	0. 965 8 ± 0. 006	<b>0. 966 2 ± 0. 006</b>	0. 965 7 ± 0. 007	<b>0. 966 2 ± 0. 006</b>	0. 966 1 ± 0. 006	0. 966 1 ± 0. 006
letterUvsV	0. 987 1 ± 0. 003	0. 987 4 ± 0. 003	0. 987 2 ± 0. 003	<b>0. 987 7 ± 0. 003</b>	0. 987 2 ± 0. 003	0. 987 2 ± 0. 002
letterVvsY	0. 943 7 ± 0. 009	0. 943 7 ± 0. 009	0. 943 9 ± 0. 009	<b>0. 944 4 ± 0. 007</b>	0. 943 2 ± 0. 009	0. 943 5 ± 0. 009
vehicle	0. 958 1 ± 0. 02	<b>0. 959 4 ± 0. 02</b>	0. 959 ± 0. 02	0. 958 5 ± 0. 02	0. 9599 ± 0. 02	0. 959 9 ± 0. 02
wdbc	0. 954 2 ± 0. 01	0. 956 3 ± 0. 009	0. 968 ± 0. 009	<b>0. 968 7 ± 0. 01</b>	0. 9563 ± 0. 01	0. 958 5 ± 0. 009

据用于训练模型,用剩下的 20% 的数据来验证模型参数. 10 次交叉验证又包含先粗调后细调的过程,粗调的参数取值为 $[10^{-5}, 10^{-4}, \dots, 10^5]$ ,细调的参数取值为 $[2^{-4}, 2^{-3}, \dots, 2^4]$ . 而 ODMRR 和 ODMRR\_3h 算法中使用到的高斯距离矩阵中的  $\sigma$  参数则采用  $\sigma=1$  的默认值设置.

由于 3 个脊回归的算法都需要计算一次矩阵的求逆计算,所以这 3 个算法的速度都不是特别快. 在这 3 个算法中,RR 包含一个参数,MRRR 包含 2 个参数,而 ODMRR 算法包含 3 个参数. 为了减少调参的时间,我们采用固定其余参数,只调整一个参数的方法. 从表 2 的实验结果中可以发现,即便在如此简单的调参方法下,ODMRR 算法在 14 个数据集中有 10 个达到最优结果,其中 2 个和最优结果打平. 并且可以很明显地发现对于给类内间隔分布和类间间隔分布分配不同参数的 ODMRR\_3h 方法相比使用相同参数的 ODMRR 方法取得了更多的最优结果. 所以这也一定程度上说明,类内结构关系和类间结构关系在对于分类器的贡献在部分数据集上可能是不同的.

## 5 结束语

脊回归算法及其流形正则化算法是机器学习领域的经典方法. 这 2 种算法都没有充分考虑数据全局的结构化关系. 我们以一种全新的视角分析了最优间隔分布学习理论对于优化间隔分布的优化目标,提出了一种充分考虑类内结构关系和类间结构关系的正则化脊回归算法——最优间隔分布脊回归算法(ODMRR). 由于脊回归算法和流形正则化的脊回归算法都是 ODMRR 算法的特例,所以 ODMRR 算法保有了脊回归和流形正则化的脊回归算法中所共有的优化目标为凸函数、有解析解、易于核化等优势. 考虑到类内结构关系和类间结构关系对于模型的贡献不同,我们也提出了 ODMRR 算法的一种简单变形 ODMRR\_3 算法方便用户调节两者之间的权重. 最后,我们在真实数据集上进行了实验且在大部分数据集上取得了最优的结果.

## 参 考 文 献

- [1] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems [J]. *Technometrics*, 1970, 12(1): 55-67
- [2] Xue Hui, Zhu Yulian, Chen Songcan. Local ridge regression for face recognition [J]. *Neurocomputing*, 2009, 72(4): 1342-1346
- [3] Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data [J]. *BMC Bioinformatics*, 2011, 12(1): 372-378
- [4] Liu Shujie, Kawamoto T, Morita O, et al. Discriminating between adaptive and carcinogenic liver hypertrophy in rat studies using logistic ridge regression analysis of toxicogenomic data: The mode of action and predictive models [J]. *Toxicology and Applied Pharmacology*, 2017, 318(9): 79-87
- [5] Marquardt D W, Snee R D. Ridge regression in practice [J]. *The American Statistician*, 1975, 29(1): 3-20
- [6] Belkin M, Niyogi P, Sindhvani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. *Journal of Machine Learning Research*, 2006, 7(11): 2399-2434
- [7] Luo Yong, Tao Dacheng, Xu Chang, et al. Multiview vector-valued manifold regularization for multilabel image classification [J]. *IEEE Trans on Neural Networks and Learning Systems*, 2013, 24(5): 709-722
- [8] Geng Bo, Tao Dacheng, Xu Chao, et al. Ensemble manifold regularization [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2012, 34(6): 1227-1233
- [9] Xu Zenglin, King I, Lyu M R T, et al. Discriminative semi-supervised feature selection via manifold regularization [J]. *IEEE Trans on Neural Networks*, 2010, 21(7): 1033-1047
- [10] Luo Siwei, Zhao Lianwei. Manifold learning algorithms based on spectral graph theory [J]. *Journal of Computer Research and Development*, 2006, 43(7): 1173-1179 (in Chinese)  
(罗四维, 赵连伟. 基于谱图理论的流形学习算法[J]. *计算机研究与发展*, 2006, 43(7): 1173-1179)
- [11] Liu W, Liu H, Tao D, et al. Manifold regularized kernel logistic regression for Web image annotation [J]. *Neurocomputing*, 2016, 172(C): 3-8
- [12] Zhu Xiaojin. Semi-supervised learning literature survey [J]. *Computer Science*, University of Wisconsin-Madison, 2006, 2(3): 4-64
- [13] Vapnik V. *The Nature of Statistical Learning Theory* [M]. Berlin: Springer, 1995
- [14] Zhou Zhihua. Large margin distribution learning [C] //Proc of the Workshop on Artificial Neural Networks in Pattern Recognition. Berlin: Springer, 2014: 1-11
- [15] Gao Wei, Zhou Zhihua. On the doubt about margin explanation of boosting [J]. *Artificial Intelligence*, 2013, 203(5): 1-18
- [16] Zhang Teng, Zhou Zhihua. Multi-class optimal distribution machine [C] //Proc of the 34th Int Conf on Machine Learning. 2017 [2017-04-01]. <http://cs.nju.edu.cn/zhoush/zhoush.files/publication/icml17mcODM.pdf>
- [17] Zhang Teng, Zhou Zhihua. Large margin distribution machine [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 313-322

- [18] Zhou Yuhang, Zhou Zihua. Large margin distribution learning with cost interval and unlabeled data [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(7): 1749-1763



**Chen Jialie**, born in 1991. PhD. His main research interests include machine learning and data mining.



**Jiang Yuan**, born in 1976. Received her PhD degree in computer science from Nanjing University in 2004. Currently professor and PhD supervisor at the Department of Computer Science & Technology, Nanjing University. Her main research interests include machine learning and data mining.

## 2018年《计算机研究与发展》专题(正刊)征文通知

### ——面向硬件的数据管理

数据特性和硬件特征是高效数据管理的两个关键因素。一方面,随着大数据时代的到来,数据管理系统需要支持越来越海量、高速、多样的数据,各种新的数据管理技术纷纷涌现,不仅包括对传统的关系型数据库系统的扩展,而且包括多种新型的大数据系统,例如 NoSQL 系统, NewSQL 系统等。数据的种类不仅有传统的关系型数据,而且有图类型、JSON 树状数据等多种非传统数据类型。另一方面,计算机系统的硬件正在经历着深远的变化,更大容量的主存、NVM 技术等正改变着存储系统的面貌。GPU、FPGA、专用加速硬件正逐渐由专用领域走向通用计算,对数据的运算和管理发挥着越来越重要的影响。因此,如何把这两者相结合,利用新型的存储和计算硬件高效地支持数据管理的需求成为一个重要的学术课题,许多基本问题还有待探索,包括系统架构、数据存储、索引、查询处理、能源效率等。

《计算机研究与发展》将于 2018 年 2 月出版面向新型硬件的数据管理专题,侧重面向新硬件的数据库技术、面向新硬件的大数据系统、软硬件协同设计(Co-Design)的数据处理等方面,这将涉及数据管理、大数据系统、计算机体系结构等多个交叉学术领域,探讨面向新型硬件的数据管理基础理论研究及其应用,讨论该领域内最新的突破性进展,交流新的学术思想和新方法,展望未来的发展趋势。

欢迎相关领域的专家学者和科研人员踊跃投稿,现将专题论文征集的有关事项通知如下。

#### 征文范围(但不限于)

- 新型存储硬件与计算硬件的数据管理
- 适应新型存储硬件特性的数据管理
- 多种硬件共用的数据管理
- 新的数据存储和索引方法
- 基于新型计算硬件的数据运算与查询处理
- 多种存储硬件(NVM, HBM, DRAM 等)的一体化管理
- 软硬件协同设计
- 软硬件协同的查询处理
- 软硬件协同的大数据复杂分析
- 软硬件协同的数据管理理论与架构
- 软硬件特征分析评估
- 硬件配置检测
- 性能瓶颈分析
- 效率评估
- 高效的数据处理
- 面向异构硬件的数据处理

#### 征文要求

- 1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表,不存在一稿多投问题。作者在投稿时,需向编辑部提交投稿声明。
- 2) 论文一律用 Word 格式排版,论文格式体例参考近期出版的《计算机研究与发展》的要求(<http://crad.ict.ac.cn/>)。
- 3) 论文通过期刊网站(<http://crad.ict.ac.cn/>)投稿,投稿时提供作者的联系方式,备注中务必注明“面向硬件的数据管理 2018 专题”(否则按自由来稿处理)。

#### 重要日期

征文截止日期:2017 年 9 月 15 日

录用通知日期:2017 年 10 月 10 日

作者修改稿提交日期:2017 年 11 月 10 日

出版日期:2018 年 2 月

#### 特邀编委

孟小峰 教授 中国人民大学 xfmeng@ruc.edu.cn

陈世敏 研究员 中国科学院计算所 chensm@ict.ac.cn

罗琼 副教授 香港科技大学 luo@cse.ust.hk

#### 联系方式

编辑部:crad@ict.ac.cn 010-62620696,010-62600350

通信地址:北京 2704 信箱《计算机研究与发展》编辑部

邮政编码:100190