doi:10.11887/j.cn.201506023 http://journal.nudt.edu.cn

面向监督学习的稀疏平滑岭回归方法*

任维雅,李国辉
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:岭回归是监督学习中的一个重要方法,被广泛用于多目标分类和识别。岭回归中一个重要的步骤是定义一个特殊的多变量标签矩阵,以实现对多类别样本的编码。通过将岭回归看作是一种基于图的监督学习方法,拓展了标签矩阵的构造方法。在岭回归的基础之上,进一步考虑投影中维度的平滑性和投影矩阵的稀疏性,提出稀疏平滑岭回归方法。对比一系列经典的监督线性分类算法,发现稀疏平滑岭回归在多个数据集上有着更好的表现。另外,实验表明新的标签矩阵构造方法不会降低原始岭回归方法的表现,同时还可以进一步提升稀疏平滑岭回归方法的性能。

关键词:岭回归;多分类;全局维度平滑性;监督学习
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2015)06-121-08

Sparse smooth ridge regression method for supervised learning

REN Weiya, LI Guohui
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)

Abstract: Ridge regression is an important method in supervised learning. It is wide used in multi-class classification and recognition. An important step in ridge regression is to define a special multivariate label matrix, which is used to encode multi-class samples. By regarding the ridge regression as a supervised learning method based on graph, methods for constructing multivariate label matrix were extended. On the basis of ridge regression, a new method named sparse smooth ridge regression was proposed by considering the global dimension smoothness and the sparseness of the projection matrix. Experiments on several public datasets show that the proposed method performs better than a series of state-of-the-art supervised linear algorithms. Furthermore, experiments show that the proposed label matrix construction methods do not reduce the performance of the original ridge regression. Besides, it can further improve the performance of the proposed sparse smooth ridge regression.

Key words: ridge regression; multi-class classification; global dimension smoothness; supervised learning

监督学习是机器学习和模式识别中一个重要的学习内容,被应用于包括人脸识别、文本识别及图像分类等诸多领域。在大数据应用需求的背景下,监督学习面临两个重要问题:一是提高分类器的分类准确率问题;二是能够给出对新样本的显式映射,即解决"out-of-sample"问题。

为解决以上两个问题,近年来涌现出一系列基于线性投影的机器学习方法。这些方法包括:基于流形学习的方法,如局部保持投影(Locality Preserving Projections, LPP)[1]和邻域保持嵌入(Neighborhood Preserving Embedded, NPE)[2]等;度量学习(metric learning)方法,如KISS(Keep It Simple and Straightforward)方法[3]、最大边界近邻学习(Large Margin Nearest Neighbor learning, LMNN)[4]和信息论度量学习(Information Theoretic Metric Learning, ITML)[5]等;其他一些著名的机器学习方法,如线性判别分析(Linear Discriminant Analysis, LDA)[6-7]、局部敏感判别分析(Locality Sensitive Discriminant Analysis, LSDA)[8]和间隔判别分析(Marginal Fisher Analysis, MFA)[9]等。

岭回归(Ridge Regression, RR)方法[10-13]是一种利用正则化的最小二乘法方法,最早只设计了单变量标签[11-13]。文献[10]推广了原始岭回归方法,将单变量标签扩展成多变量标签,以解决多分类问题。岭回归方法[10]是一种监督学习方法,由于其出色的学习性能,目前正受到越来越为广泛的关注。它主要包括以下步骤:①生成训练样本点的多变量标签矩阵;②学习线性分类器,即投影矩阵;③对新样本进行分类识别。

文献[10]指出岭回归的多变量标签矩阵方

* **收稿日期**:2014-12-26
基金项目:国家自然科学基金资助项目(611701586);数学工程与先进计算国家重点实验室开放资助项目(Grant 2013A08)
作者简介:任维雅(1988—),男,河南周口人,博士研究生,E-mail:weiyren.phd@gmail.com;
李国辉(通信作者),男,教授,博士,博士生导师,E-mail:gli2010a@163.com

法是特定的。然而,通过将岭回归学习方法纳入基于图(graph-based)的监督学习方法,发现多变量标签矩阵的构造方法是可以灵活设定的,学习投影矩阵的稀疏性往往是一个优良投影矩阵必备的潜在特征。因此在岭回归学习方法中引入投影矩阵的稀疏性约束就得到稀疏平滑岭回归方法。

1 岭回归

岭回归方法[10]使用正则单形顶点(regular simplex vertices)[14]作为训练样本的多变量标签,将高维特征空间映射到低维特征空间,并使样本投影到这些正则单形顶点的周围。记训练样本为 $\boldsymbol{X}=[x_1,\cdots,x_n]\in\boldsymbol{R}^{m\times n}$,对应标签为 $\boldsymbol{L}=[l_1,\cdots,l_n]$,其中 $l_i\in\{1,2,\cdots,k\}$,代表训练样本共有 k 个类别。

记 $\boldsymbol{T}_i\in\boldsymbol{R}^{k-1}(i=1,2,\cdots,k)$ 为一个正则 k 单形的顶点,$\boldsymbol{T}=[\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_k]\in\boldsymbol{R}^{(k-1)\times k}$。$\boldsymbol{T}$ 构造方法如下:

1)$\boldsymbol{T}_1=[1,0,\cdots,0]^{\mathrm{T}}$ 且 $T_{1,i}=-1/(k-1)$,$i=2,\cdots,k$。

2)当 $1\leqslant g\leqslant k-2$,有

$$T_{g+1,g+1}=\sqrt{1-\sum_{i=1}^{g}T_{i,g}^2};$$

$$T_{g+1,j}=-\frac{T_{g+1,g+1}}{k-g-1},\ j=g+2,\cdots,k;$$

$$T_{i,g+1}=0,\ g+2\leqslant i\leqslant k-1。$$

这 k 个顶点分布在以原点为圆心的超球面上,是 $k-1$ 维空间中最平衡和对称的分隔点,任意两点之间的距离相等。

进一步构造多变量标签矩阵 $\boldsymbol{Y}$:当 x_i 属于第 j 类时,$\boldsymbol{Y}^{(i)}=\boldsymbol{T}_j^{\mathrm{T}}(i=1,\cdots,n;j=1,\cdots,k)$。其中,$\boldsymbol{Y}^{(i)}$ 为 $\boldsymbol{Y}(\boldsymbol{Y}\in\boldsymbol{R}^{n\times(k-1)})$ 的第 i 行,$\boldsymbol{T}_j$ 为 $\boldsymbol{T}$ 的第 j 列。

岭回归方法最小化如式(1)所示的目标函数:

$$\begin{aligned}J(\boldsymbol{P})&=\sum_{i=1}^{n}\|\boldsymbol{P}^{\mathrm{T}}x_i-\boldsymbol{Y}^{(i)}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2\\&=\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}^{\mathrm{T}}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2\end{aligned}\tag{1}$$

其中,$\boldsymbol{P}\in\boldsymbol{R}^{m\times d}$ 是待学习的线性投影矩阵,$\lambda_1>0$ 是平衡式中两项的正则化参数,$\|\boldsymbol{P}\|_F$ 表示 $\boldsymbol{P}$ 的 Frobenius 范数,$\|\boldsymbol{P}\|_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{d}P_{ij}^2\right)^{\frac{1}{2}}$($P_{ij}$ 为 $\boldsymbol{P}$ 的第 i 行第 j 列的元素)。

直接求导可得:

$$\boldsymbol{P}=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\lambda_1\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{Y}\tag{2}$$

式中,$\boldsymbol{I}$ 为单位矩阵。

在分类阶段,对于新样本 z,通过比较 $\boldsymbol{P}_z^{\mathrm{T}}$ 和训练样本中 $\boldsymbol{P}_{x_i}^{\mathrm{T}}(i=1,2,\cdots,n)$ 的距离来确定新样本的类别,即使用最近邻分类器判别法。

2 多变量标签矩阵

岭回归方法实质上是一种基于图的监督线性学习方法,在此基础之上,可以拓展多变量标签矩阵的构造方法。首先考察两个经典的基于图的线性投影算法:局部保持投影 LPP[1] 和邻域保持嵌入 NPE[2]。LPP 和 NPE 优化如式(3)所示的目标函数:

$$\begin{aligned}\boldsymbol{P}&=\operatorname{argmin}\frac{1}{2}\sum_{i,j=1}^{n}(\boldsymbol{P}^{\mathrm{T}}x_i-\boldsymbol{P}^{\mathrm{T}}x_j)^2\tilde{w}_{ij}\\&\text{s.t.}\quad\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}=\boldsymbol{I}\end{aligned}\tag{3}$$

其中,$\tilde{\boldsymbol{W}}=[\tilde{w}_{ij}]$ 对应一种特定的图构建方法[1-2]。$\tilde{w}_{ij}$值较大,意味着 x_i 与 x_j 有着较高的相似度,反之亦然。在监督学习中,可以认为标签相同的样本比标签不同的样本有着更高的相似度,因此优化式(3)的结果意味着相同标签的样本在投影后要比不同标签的样本投影后更加相似。但是在 LPP 和 NPE 中,不同标签的样本在投影后没有直接约束,这将导致其间隔可以任意大,因此需要通过正交约束 $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}=1$ 来确保解的良态性。

如果认为具有相同标签的样本是相互相似的,则岭回归学习方法符合基于图的学习方法对于相似样本的约束。实际上,观察式(1),可以发现岭回归的标签矩阵约束了相同标签的样本在投影后的距离,使之趋于接近。另外,不同于 LPP 和 NPP 方法,岭回归方法通过标签矩阵的约束避免了解的病态性问题。在 LPP 和 NPP 方法中,如果没有正交约束,不同标签的样本在投影后的距离将趋于无穷大;而在岭回归方法中,标签矩阵的约束使得不同标签的样本在投影后的距离将趋于一个固定间隔。

因此,岭回归的多变量标签矩阵只要满足以上对同标签样本的约束和不同标签样本的约束,即可纳入为基于图的监督学习方法。在基于图的监督学习方法的框架下,岭回归的多变量标签矩阵可以通过如下方法构造:

记多变量标签矩阵 $\boldsymbol{Y}\in\boldsymbol{R}^{n\times d}$($d$ 是样本投影后的新维度大小)。岭回归方法[10]使用正则单形顶点,且 $d=k-1$。这种构造方法较为严格,实际上,只需在 d 维空间中构造 k 个相互正交、长度为 1 的顶点就可以满足基于图的学习方法的要求。d 的大小是可以定义的,这意味着投影后样本的维度也是可以预先定义的。

标签矩阵的具体构造步骤为:

1)在 d 维空间中构造 k 个相互正交、长度为 1 的顶点,记为 $\boldsymbol{T}=[\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_k]\in\boldsymbol{R}^{d\times k}$。

2)每一类训练样本对应一个顶点,构造标签矩阵 $\boldsymbol{Y}$,当 x_i 属于第 j 类时,$\boldsymbol{Y}^{(i)}=\boldsymbol{T}_j^{\mathrm{T}}(i=1,\cdots,n;j=1,\cdots,k)$。其中,$\boldsymbol{Y}^{(i)}$ 为 $\boldsymbol{Y}$ 的第 i 行,$\boldsymbol{T}_j$ 为 $\boldsymbol{T}$ 的第 j 列。

根据上述步骤,提出两种构造 $\boldsymbol{T}$ 的方法:

1)构造方法 1:当 $i=j$ 时,$T_{ij}=1$,否则 $T_{ij}=0$。要求 $d\geqslant k$,通常可取 $d=k$。

2)构造方法 2:在 d 维空间中生成 k 个随机顶点,使用施密特正交化方法生成 k 个新顶点,以构造 $\boldsymbol{T}$。

构造方法 1 最直观简单,构造方法 2 可以控制维度。在第五节中将给出不同构造方法对岭回归多分类识别率的影响。

3　稀疏平滑岭回归

将所有样本点在维度上的坐标记为一个维度点 $\boldsymbol{d}^{(i)}$($\boldsymbol{X}$ 的第 i 行),可以使用多种权重[15]度量方法度量其相似性。使用核权重对它们的相似性进行衡量,即如果点 $\boldsymbol{d}^{(i)}$ 是点 $\boldsymbol{d}^{(j)}$ $(i\neq j)$ s 个最近点之一或点 $\boldsymbol{d}^{(j)}$ 是点 $\boldsymbol{d}^{(i)}$ 的 s 个最近点之一,则:

$$W_{ij}=\mathrm{e}^{\frac{-(\boldsymbol{d}^{(i)}-\boldsymbol{d}^{(j)})^2}{2\sigma^2}} \tag{4}$$

否则 $W_{ij}=0$。通常 $\sigma=\sum_{i,j=1}^{m}\sqrt{D_E^2(\boldsymbol{d}^{(i)},\boldsymbol{d}^{(j)})/m^2}$,$D_E^2(\boldsymbol{d}^{(i)},\boldsymbol{d}^{(j)})$ 是 $\boldsymbol{d}^{(i)}$ 和 $\boldsymbol{d}^{(j)}$ 的欧式距离。

线性投影矩阵 $\boldsymbol{P}\in\boldsymbol{R}^{m\times d}$ 将样本 x 从 m 维投影到 k 维($m\gg k$),在这一过程中,维度得到了一定程度的变形和压缩,一个合理的全局假设是:如果空间中所有的样本点在维度 i 和维度 j 上具有相似的坐标,记为 $\boldsymbol{d}^{(i)}\rightarrow\boldsymbol{d}^{(j)}$,则在投影时这两个维度应当受到相似的操作,即对应的权重系数相似,记为 $\boldsymbol{P}^{(i)}\rightarrow\boldsymbol{P}^{(j)}$($\boldsymbol{P}^{(i)}$ 是投影矩阵 $\boldsymbol{P}$ 的第 i 行)。

将这个假设称为全局维度平滑性假设,其数学的表示为最小化如式(5)所示的正则化项:

$$\begin{aligned}R&=\frac{1}{2}\sum_{i,j=1}^{m}(\boldsymbol{P}^{(i)}-\boldsymbol{P}^{(j)})^2w_{ij}\\&=\sum_{i=1}^{m}(\boldsymbol{P}^{(i)})^{\mathrm{T}}\boldsymbol{P}^{(i)}D_{ii}-\sum_{i,j=1}^{m}(\boldsymbol{P}^{(i)})^{\mathrm{T}}\boldsymbol{P}^{(j)}W_{ij}\\&=\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{P})-\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{P})\\&=\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P})\end{aligned} \tag{5}$$

其中,trace(·)为矩阵的迹,$\boldsymbol{L}$ 为图 W 的拉普拉斯矩阵,即 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$,$\boldsymbol{D}$ 是一个对角矩阵,且 $D_{ii}=\sum_{j=1}^{n}W_{ij}$。

通过最小化 $\boldsymbol{R}$,希望初始相似的维度点投影后依旧相似,即 $\boldsymbol{d}^{(i)}\rightarrow\boldsymbol{d}^{(j)}$,则 $\boldsymbol{P}^{(i)}\rightarrow\boldsymbol{P}^{(j)}$。

考虑正则化项 R,岭回归最小化目标变为:

$$\boldsymbol{P}=\mathrm{argmin}\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}^{\mathrm{T}}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2+\lambda_2\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P}) \tag{6}$$

式中,$\lambda_1\lambda_2>0$ 是平衡各正则化项的参数。

比起大多线性学习方法,经典的 KISS 度量学习方法和 MFA 方法学习得到的投影矩阵往往具有较好的稀疏性,较好的稀疏度有利于提高投影的鲁棒性,提高模型的泛化能力。因此,进一步对岭回归投影矩阵增加稀疏度要求,式(6)变为最小化如式(7)所示的目标函数:

$$\begin{aligned}\boldsymbol{P}=\mathrm{argmin}&\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}^{\mathrm{T}}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2+\\&\lambda_2\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P})+\lambda_3\|\boldsymbol{P}\|_1\end{aligned} \tag{7}$$

式中,$\|\boldsymbol{P}\|_1$ 表示 $\boldsymbol{P}$ 的 l_1 范数,即 $\|\boldsymbol{P}\|_1=\sum_{i=1}^{m}\sum_{j=1}^{d}|P_{ij}|$。$\lambda_1,\lambda_2,\lambda_3$(都大于0)是平衡各正则化项的参数。

将解决式(7)所示问题(问题(7))的方法称为稀疏平滑岭回归(Sparse smooth Ridge Regression,SRR)方法。

4　算法实现

通过变量分别优化的方法解决问题(7),即通过固定其他参数求解某一个参数。采用 Inexact ALM[16](augmented Lagrange multiplier)方法,通过一个附属变量拆分目标函数的变量,式(7)可以重写为:

$$\begin{aligned}\boldsymbol{P}=\mathrm{argmin}&\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}^{\mathrm{T}}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2+\\&\lambda_2\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P})+\lambda_3\|\boldsymbol{H}\|_1\\&\text{s.t. }\boldsymbol{P}=\boldsymbol{H}\end{aligned} \tag{8}$$

式(8)的拉格朗日函数为:

$$\begin{aligned}\mathcal{L}=&\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{Y}^{\mathrm{T}}\|_F^2+\lambda_1\|\boldsymbol{P}\|_F^2+\lambda_2\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P})+\\&\lambda_3\|\boldsymbol{H}\|_1+<\boldsymbol{Q},\boldsymbol{P}-\boldsymbol{H}>+\frac{\mu}{2}(\boldsymbol{P}-\boldsymbol{H})^2\end{aligned} \tag{9}$$

式中,Q 是拉格朗日乘子,$\mu\geqslant0$ 是惩罚参数。

固定其他变量,求 $\boldsymbol{P}$:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{P}}=\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}-\boldsymbol{Y})+\lambda_1\boldsymbol{P}+\lambda_2\boldsymbol{L}\boldsymbol{P}+\boldsymbol{Q}+\mu(\boldsymbol{P}-\boldsymbol{H})=0 \tag{10}$$

于是,

$$\boldsymbol{P}=\left(\frac{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\lambda_1\boldsymbol{I}+\lambda_2\boldsymbol{L}}{\mu}+\boldsymbol{I}\right)^{-1}\left(\frac{\boldsymbol{X}\boldsymbol{Y}-\boldsymbol{Q}}{\mu}+\boldsymbol{H}\right) \tag{11}$$

固定其他变量,求 $\boldsymbol{H}$:

$$\boldsymbol{H} = \operatorname{argmin}\lambda_3 \|\boldsymbol{H}\|_1 + \frac{\mu}{2}\left\|\boldsymbol{H} - \left(\boldsymbol{P} + \frac{\boldsymbol{Q}}{\mu}\right)\right\|_F^2 = \boldsymbol{\Theta}_{\frac{\lambda_3}{\mu}}\left(\boldsymbol{P} + \frac{Q}{\mu}\right) \tag{12}$$

其中,$\boldsymbol{\Theta}_\beta(x) = \operatorname{sign}(x)\max(|x| - \beta, 0)$ 是软阈值操作子[17],且有:

$$\operatorname{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & \text{其他} \end{cases} \tag{13}$$

通过 Inexact ALM[16] 方法解决问题(7)的完整算法见算法 1。

算法 1 解决问题(7)的完整算法

Alg. 1 The complete algorithm for solving problem (7)

输入: 数据 $\boldsymbol{X}$,参数 $\lambda_1 > 0, \lambda_2 > 0$ 和 $\lambda_3 > 0$。

初始化: $\boldsymbol{P} = \boldsymbol{H} = \boldsymbol{Y}_1 = \boldsymbol{0}, \mu = 10^{-3}, \mu_{\max} = 10^{10}, \rho = 1.05$ 和 $\varepsilon = 10^{-6}$。

如果不收敛则**循环**

1)固定其他变量更新 $\boldsymbol{P}$

$$\boldsymbol{P}_{k+1} = \left(\frac{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda_1 \boldsymbol{I} + \lambda_2 \boldsymbol{L}}{\mu} + \boldsymbol{I}\right)^{-1}\left(\frac{\boldsymbol{X}\boldsymbol{Y} - Q_k}{\mu} + \boldsymbol{H}_k\right)$$

2)固定其他变量更新 $\boldsymbol{H}$

$$\boldsymbol{H}_{k+1} = \operatorname{argmin}_{\boldsymbol{H}}\lambda_3 \|\boldsymbol{H}\|_1 + \frac{\mu}{2}\left\|\boldsymbol{H} - \left(\boldsymbol{P}_k + \frac{\boldsymbol{Y}_1}{\mu}\right)\right\|_F^2 = \boldsymbol{\Theta}_{\frac{\lambda_3}{\mu}}\left(\boldsymbol{P}_k + \frac{\boldsymbol{Y}_1}{\mu}\right)$$

3)更新 μ

$$\mu = \min(\mu_{\max}, \rho\mu)$$

4)检查收敛条件

$$\|Q_k - Q_{k+1}\|_\infty < \varepsilon$$
$$\|\boldsymbol{P}_k - \boldsymbol{P}_{k+1}\|_\infty < \varepsilon$$
$$\|\boldsymbol{P} - \boldsymbol{H}\|_\infty < \varepsilon$$

结束循环

输出: $(\boldsymbol{P}, \boldsymbol{H})$.

5 实验

本节面向监督学习进行多分类实验,通过对测试样本的识别准确率来衡量不同算法的水平。实验用的线性投影方法共 8 种,包括:LPP、NPE、KISS、LSDA、MFA、LDA、RR、SRR。同时,实验分析了不同标签矩阵对岭回归方法的影响。数据集包括图像数据集、人脸数据集、手写体数据集和文本数据集,表 1 给出了 4 个数据集的统计指标,图 1 展示了一些数据集的原始图像示例。

表 1 4 个数据集的统计指标

Tab. 1 Statistics of the four data sets

数据集	大小 n	维度 m	类别数 k
COIL20	1440	1024	20
Yale	165	1024	15
TDT2	750	36771	15
USPS	9298	256	10

5.1 数据集

1)COIL20 数据集。COIL20 数据集[18] 包括 20 个类别图像,每类图像包含 72 张不同视角的图像。每张图像降采样后的大小是 32 × 32 像素,被表示为一个 1024 维的向量。

2)Yale 数据集。Yale 数据集[19] 包含 15 个人物,共 165 张灰度照片。每个人物有 11 张表情和外形不同的照片,每张图片降采样后的大小是 32 × 32 像素,由一个 1024 维的向量表示。

3)TDT2 数据集。TDT2 数据集[20] 是一个文本数据集,包括 9394 个文本文件。每个文本文件被一个 36771 维的向量表示。样本点最多的前 15 类数据的各自前 50 个样本点作为实验数据集使用。

4)USPS 数据集。USPS 数据集[21] 是一个手写体数据集,包括 9298 张图片,来自 10 个类别。每张图片大小为 16 × 16 像素,由一个 256 维的向量表示。

通常可采用主成分分析(Principal Component Analysis,PCA)将数据先降维至一个合适的维数以提高运算效率。另外,数据的预处理方法是对数据进行平方和归一化操作。

5.2 实验流程

5.2.1 监督分类学习实验

选择一个数据集,确定在每类样本中要挑选的训练样本个数 NL,实验流程如下:

1)在每类样本中随机选择 NL 个样本组成训练集,余下样本作为测试集;

2)用不同方法学习线性投影矩阵;

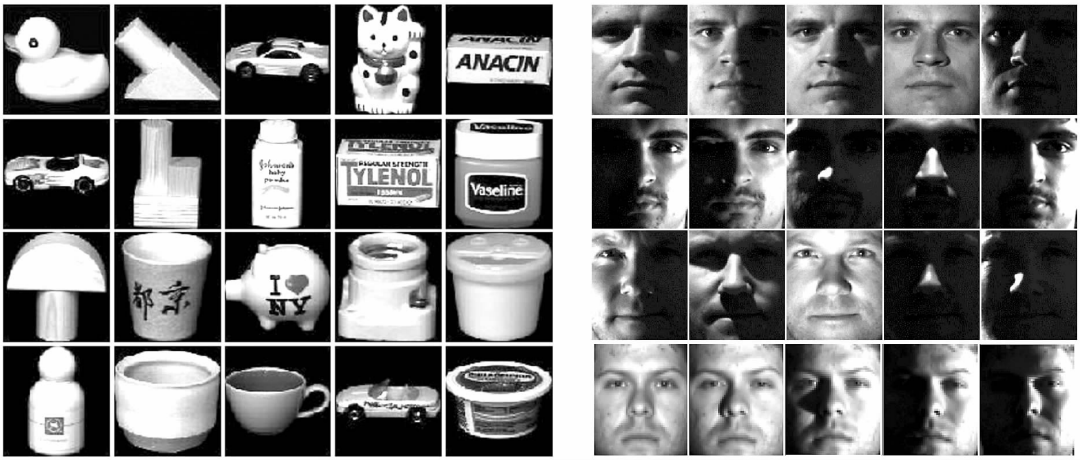
3)对测试集样本进行投影;

4)通过最近邻方法(1 - NN)确定测试样本的预测标签,计算每类方法在测试样本上的识别准确率;

5)重复以上流程 50 次。

5.2.2 标签矩阵实验

构造 5 个不同的标签矩阵,对比这些标签矩阵对 RR 和 SRR 方法的影响。这些标签矩阵包括:



(a) COIL20　　(b) Yale

(c) USPS

图 1　COIL20,YaleB 和 USPS 数据库上的图片示例

Fig. 1　Sample images in COIL20, Yale and USPS database

1) $\boldsymbol{Y}_1$:原始岭回归构造法[10]。

2) $\boldsymbol{Y}_2$:使用第 2 节的构造法 1,取 $d=k$(d 是构造顶点 $\boldsymbol{T}$ 的维度,k 是样本类别数目)。$\boldsymbol{Y}_2$ 是一个 0-1 矩阵,每行只有一个 1,其余为 0。

3) $\boldsymbol{Y}_3$:通过 $\boldsymbol{T}$ 构造法 2 构建标签矩阵,令 $d=2k$。

4) $\boldsymbol{Y}_4$:使用 $\boldsymbol{T}$ 构造法 2,令 $d=3k$。

5) $\boldsymbol{Y}_5$:使用 $\boldsymbol{T}$ 构造法 2,令 $d=m$。其中,m 是样本数据 $\boldsymbol{X}$ 的原始维度。

5.3　实验结果

多分类实验结果如表 2 ~ 5 所示。SRR 方法在实验数据集上表现良好,特别在 TDT2 文本数据库和 COIL20 图像数据库上表现优异。观察 USPS 数据库和 Yale 数据库,如表 2、表 3 所示,当训练集数目逐渐增加时,部分经典方法识别效果反而下降,这可能是因为训练出现了过拟合现象。与此同时,SRR 方法依然表现良好,体现出较好的泛化能力。

在标签矩阵实验中(见表 6),标签矩阵并没有降低 RR 方法的识别率,这说明将岭回归方法看作是一种基于图的学习方法并由此设计标签矩阵是合理的。这意味着标签矩阵的作用是尽量使投影后的样本同类聚集,异类等距分隔。另外,设计的标签矩阵在 SRR 方法上比原始标签矩阵有一定的提升,这验证了拓展标签矩阵设计的价值。

表 2　不同方法在 USPS 数据集上的识别率

Tab. 2　Recognition rate of different methods on USPS dataset

NL	LPP/%	NPE/%	KISS/%	LSDA/%	MFA/%	LDA/%	RR/%	SRR/%
4	63.89	63.95	67.88	62.19	68.43	63.41	69.09	**74.66**
8	62.55	62.33	66.59	59.36	66.40	61.07	75.77	**81.48**
12	61.59	60.10	64.92	57.13	64.94	58.24	79.58	**83.94**
16	55.91	52.91	61.34	50.54	63.50	52.01	81.73	**85.58**
20	48.04	44.54	60.06	43.48	63.76	45.09	83.10	**86.40**

表3 不同方法在Yale数据集上的识别率

Tab. 3 Recognition rate of different methods on Yale dataset

NL	LPP/%	NPE/%	KISS/%	LSDA/%	MFA/%	LDA/%	RR/%	SRR/%
2	60.74	61.39	**62.22**	56.04	56.96	60.75	60.66	58.87
4	73.61	71.44	**77.35**	71.37	66.85	74.22	74.13	73.79
6	77.89	76.40	**82.48**	76.18	70.13	76.98	81.38	81.94
8	72.35	71.68	81.73	71.60	68.08	71.51	82.53	**83.77**
10	37.86	36.26	59.86	37.33	61.86	36.26	82.26	**86.40**

表4 不同方法在COIL20数据集上的识别率

Tab. 4 Recognition rate of different methods on COIL20 dataset

NL	LPP/%	NPE/%	KISS/%	LSDA/%	MFA/%	LDA/%	RR/%	SRR/%
4	74.25	73.87	80.60	74.39	80.35	73.93	75.58	**83.40**
8	75.55	75.08	85.47	75.09	88.42	74.83	84.08	**91.96**
12	78.57	78.33	78.15	77.55	90.95	77.60	87.72	**94.27**
16	89.58	92.68	89.27	89.10	93.24	89.01	90.47	**96.25**
20	93.69	96.61	93.76	93.26	94.93	93.22	92.73	**97.38**

表5 不同方法在TDT2数据集上的识别率

Tab. 5 Recognition rate of different methods on TDT2 dataset

NL	LPP/%	NPE/%	KISS/%	LSDA/%	MFA/%	LDA/%	RR/%	SRR/%
4	85.32	83.82	83.90	84.24	78.79	85.18	85.74	**87.37**
8	84.53	84.26	84.73	84.23	82.55	84.38	86.16	**89.79**
12	70.16	71.42	77.09	69.47	81.55	69.72	83.48	**91.04**
16	77.35	77.78	73.36	76.93	81.14	76.99	84.32	**91.43**
20	85.84	85.81	84.89	85.44	84.68	85.47	87.17	**92.22**

表6 使用不同标签矩阵的岭回归方法在各数据集上的识别率($NL=5$)

Tab. 6 Recognition rate of ridge regression methods with different multivariate label matrices on all datasets ($NL=5$)

%

	RR					SRR				
	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$	$\boldsymbol{Y}_3$	$\boldsymbol{Y}_4$	$\boldsymbol{Y}_5$	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$	$\boldsymbol{Y}_3$	$\boldsymbol{Y}_4$	$\boldsymbol{Y}_5$
USPS	71.78	71.78	71.78	71.78	71.78	77.22	77.37	77.84	78.16	**78.80**
Yale	77.05	77.05	77.05	77.05	77.05	76.72	**77.16**	76.27	76.50	75.83
COIL20	76.67	76.68	76.68	76.68	76.68	85.53	85.69	86.85	**87.09**	85.97
TDT2	87.02	87.05	87.05	87.05	87.05	88.69	88.93	89.31	89.24	**89.40**

5.4 算法分析

参数选择是一项重要的工作,文中所使用的对比方法采用其文献所提议的最佳参数。对于SRR方法,可通过有限网格法[22]选择参数。实验采取的参数为:对于USPS,TDT2和Yale数据库,$\lambda_1=0.01$,$\lambda_2=0.01$,$\lambda_3=0.01$;对于COIL20数据库,$\lambda_1=0.001$,$\lambda_2=0.01$,$\lambda_3=0.1$。使用核权重(式(4))来度量维度间的相似度,所有实验取$s=5$。简单起见,文中使用$\boldsymbol{Y}_2$作为SRR的标签矩阵。

分析表示投影矩阵$\boldsymbol{P}$的稀疏度,投影矩阵的稀疏度可定义如式(14):

$$sparsity(\boldsymbol{P}) = \frac{\sum_{i=1}^{m} sparsity(\boldsymbol{P}^{(i)})}{m} \tag{14}$$

式中,行向量$\boldsymbol{P}^{(i)}$的稀疏度$sparsity(\boldsymbol{P}^{(i)})$可由向量稀疏度[23]计算得到:

$$sparsity(\boldsymbol{P}^{(i)}) = \frac{\sqrt{d} - \sum |P_{ij}| \Big/ \sqrt{\sum P_{ij}^2}}{\sqrt{d}-1} \times 100\% \tag{15}$$

式中,P_{ij}是$\boldsymbol{P}^{(i)}$的第j个元素。

当一个向量所有值相同时,其稀疏度则为 0%,当一个向量只有一个元素不为 0 时,其稀疏度达到最大,取值为 100%。

表 7 为不同算法得到的投影矩阵的平均稀疏度。由表可看出,SRR 方法得到的投影矩阵比 RR 和其他大多对比方法得到的投影矩阵具有更高的稀疏度。KISS 度量学习方法往往可以得到具有最大稀疏度的投影矩阵。对比表 2 ~ 5 和表 7,发现投影矩阵稀疏性的提高往往带来识别率上的提升。KISS 方法要求相似样本尽量聚集,其对异类样本间的距离没有约束,这可能是其投影矩阵稀疏性高但其识别率不如 SRR 方法的原因。

表 7　不同算法得到的投影矩阵的平均稀疏度($NL=5$)

Tab. 7　The mean sparseness of the projection matrix obtained by different methods on datasets($NL=5$)

%

	LPP	NPE	KISS	LSDA	MFA	LDA	RR	SRR
USPS	23.73	23.04	**88.16**	23.00	23.91	25.70	26.89	55.30
Yale	24.52	24.33	**79.72**	24.92	28.09	25.14	27.52	50.42
COIL20	23.79	24.15	**75.63**	24.29	27.51	24.17	27.10	74.55
TDT2	25.34	25.61	**83.93**	24.32	25.87	25.22	31.52	59.53

5.5　稀疏约束拓展

投影矩阵的稀疏性对算法性能有着一定的影响,除了约束外,还可以考察如式(16)所示的正则化项:

$$\|\boldsymbol{P}\|_{q,1} = \sum_{i=1}^{m} \|\boldsymbol{P}^{(i)}\|_q \tag{16}$$

式中:$\boldsymbol{P}^{(i)}$ 是 $\boldsymbol{P}$ 的第 i 行向量,$\|\boldsymbol{P}^{(i)}\|_q = \left(\sum_{j=1}^{d} |P_{ij}|^q\right)^{1/q}$;当 $q=2$ 时,$\|\boldsymbol{P}\|_{2,1} = \sum_{i}^{m}\left(\sum_{j}^{d} P_{ij}^2\right)^{1/2}$ 为矩阵的组稀疏(group sparse)约束[24];当 $0<q<1$ 时,$\|\boldsymbol{P}^{(i)}\|_q$ 是 $\boldsymbol{R}^d$ 空间中的 L_q 拟范数(quasi-norm)[25-26];特别地,当 $q=\frac{1}{2}$ 时,$\|\boldsymbol{P}\|_{1/2,1} = \sum_{i}^{m}\left(\sum_{j}^{d} |P_{ij}|^{1/2}\right)^2$ 是矩阵的 $l_{1/2}$ 稀疏约束[27]。

分别考虑 $\|\boldsymbol{P}\|_{2,1}$ 和 $\|\boldsymbol{P}\|_{1/2,1}$ 约束代替 $\|\boldsymbol{P}\|_1$ 约束,式(7) 变为式(17)、式(18)。

$$\boldsymbol{P} = \operatorname{argmin}\|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{Y}^{\mathrm{T}}\|_F^2 + \lambda_1\|\boldsymbol{P}\|_F^2 + \lambda_2\operatorname{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P}) + \lambda_3\|\boldsymbol{P}\|_{2,1} \tag{17}$$

$$\boldsymbol{P} = \operatorname{argmin}\|\boldsymbol{P}^{T}\boldsymbol{X} - \boldsymbol{Y}^{\mathrm{T}}\|_F^2 + \lambda_1\|\boldsymbol{P}\|_F^2 + \lambda_2\operatorname{trace}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{P}) + \lambda_3\|\boldsymbol{P}\|_{1/2,1} \tag{18}$$

求解式(17) 和式(18) 可参考求解式(7) 的算法,相应地,只需将式(12) 分别替换为式(19)、式(20)。

$$\boldsymbol{H} = \operatorname{argmin}\lambda_3\|\boldsymbol{H}\|_{2,1} + \frac{\mu}{2}\left\|\boldsymbol{H} - \left(\boldsymbol{P} + \frac{\boldsymbol{Q}}{\mu}\right)\right\|_F^2 = \boldsymbol{\Gamma}_{\frac{\lambda_3}{\mu}}\left(\boldsymbol{P} + \frac{\boldsymbol{Q}}{\mu}\right) \tag{19}$$

$$\boldsymbol{H} = \operatorname{argmin}\lambda_3\|\boldsymbol{H}\|_{1/2,1} + \frac{\mu}{2}\left\|\boldsymbol{H} - \left(\boldsymbol{P} + \frac{\boldsymbol{Q}}{\mu}\right)\right\|_F^2 = \boldsymbol{\Omega}_{\frac{\lambda_3}{\mu}}\left(\boldsymbol{P} + \frac{\boldsymbol{Q}}{\mu}\right) \tag{20}$$

其中,$\boldsymbol{\Gamma}$ 是 $l_{2,1}$ 范数(行稀疏) 操作子(参照文献[28] 的列稀疏操作子),$\boldsymbol{\Omega}$ 是 $l_{1/2,1}$, 范数操作子[27]。

将由 $\|\boldsymbol{P}\|_1$、$\|\boldsymbol{P}\|_{2,1}$ 和 $\|\boldsymbol{P}\|_{1/2,1}$ 作为稀疏约束项的稀疏平滑岭回归算法分别记为 SRR_1,SRR_2 和 SRR_3。对这三种算法进行对比实验,结果如表 8 所示。除参数 $\lambda_i(i=1,2,3)$ 变化外,实验的其他参数设定同上文描述。

表 8 中列出了 SRR 系列算法在不同数据库上所达到的识别率和对应的参数值 $\lambda_i(i=1,2,3)$,其中,参数选择是通过有限网格法[22] 进行的,网格值为{0.0001, 0.001, 0.01, 0.1, 1, 10}。就识别率而言,SRR_1,SRR_2 和 SRR_3 表现相近,总体来说,SRR_2 表现最好,SRR_1 次之,SRR_3最差。

6　结论

扩展了岭回归方法中多变量标签矩阵的构造方法,使同类样本在投影后相互聚集,使类别不相同的样本在投影后实现固定间隔分割。通过投影过程中对维度操作的分析,得出全局维度平滑性,同时引入投影矩阵的稀疏性,拓展了 RR 方法,形成 SRR 方法。实验分析表明:SRR 方法在多个数据集上具有良好的表现,其投影矩阵具有良好的稀疏性,另外,新的标签矩阵构造方法可以进一步提高 SRR 方法的性能。

表8 不同稀疏约束的SRR方法在4个数据集上的识别率

Tab. 8 Recognition rate of different sparse constrained SRR methods on 4 datasets

数据集	SRR_1(λ_1,λ_2,λ_3)	SRR_2(λ_1,λ_2,λ_3)	SRR_3(λ_1,λ_2,λ_3)	*NL*	SRR_1/%	SRR_2/%	SRR_3/%
USPS	(0.01,0.1,0.01)	(0.1,0.1,0.01)	(0.1,0.1,0.001)	5	79.20	**79.36**	79.34
				10	84.45	**84.57**	84.55
Yale	(0.01,0.01,0.01)	(0.001,0.001, 0.1)	(0.01,0.001,0.001)	5	78.33	**78.86**	78.84
				10	**86.40**	86.00	83.60
COIL20	(0.1, 0.1,0.1)	(0.1, 0.1, 0.1)	(0.1,1,0.001)	5	87.47	**87.76**	87.24
				10	**94.72**	94.50	94.47
TDT2	(0.01, 0.1, 0.01)	(0.001, 0.01, 0.1)	(0.01, 0.01, 0.001)	5	89.15	**89.45**	88.82
				10	**92.12**	92.11	90.46

参考文献(References)

[1] He X F, Niyogi P. Locality preserving projections [J]. Advances in Neural Information Processing Systems, 2004, 16:153 – 160.

[2] He X F, Cai D, Yan S C, et al. Neighborhood preserving embedding [C]//Proceedings of IEEE International Conference on Computer Vision, 2005:1208 – 1213.

[3] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints [C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012:2288 – 2295.

[4] Weinberger K Q, Saul L K. Fast solvers and efficient implementations for distance metric learning [C]// Proceedings of the 25th International Conference on Machine Learning, 2008:1160 – 1167.

[5] Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning [C]//Proceedings of the 24th International Conference on Machine Learning, 2007:209 – 216.

[6] Lu J W, Plataniotis K N, Venetsanopoulos A N. Face recognition using LDA-based algorithms [J]. IEEE Transactions on Neural Networks, 2003, 14(1):195 – 200.

[7] Welling M. Fisher linear discriminant analysis [J]. Department of Computer Science, 2008, 16(94):237 – 280.

[8] Cai D, He X F, Zhou K, et al. Locality sensitive discriminant analysis [C]//Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007: 708 – 713.

[9] Xu D, Yan S C, Tao D C, et al. Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval[J]. IEEE Transactions on Image Processing, 2007, 16(11): 2811 – 2821.

[10] An S, Liu W Q, Venkatesh S. Face recognition using kernel ridge regression [C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2007:1 – 7.

[11] Saunders C, Gammerman A, Vovk V. Ridge regression learning algorithm in dual variables[C]// Proceedings of the 15th International Conference on Machine Learning (ICML98), 1998: 515 – 521.

[12] Hoerl A E, Kennard R W. Ridge regression: applications to nonorthogonal problems[J]. Technometrics, 1970, 12(1): 69 – 82.

[13] Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems [J]. Technometrics, 1970, 12(1):55 – 67.

[14] Parks H R, Wills D C. An elementary calculation of the dihedral angle of the regular n-simplex[J]. The American Mathematical Monthly (Mathematical Association of America), 2002, 109 (8): 756 – 758.

[15] Ren W Y, Li G H, Tu D, et al. Nonnegative matrix factorization with regularizations [J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2014, 4(1): 153 – 164.

[16] Lin Z, Chen M, Wu L, et al. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices[R]. Technical Report, UILU-ENG – 09 – 2215, 2009.

[17] Candès E J, Li X D, Ma Y, et al. Robust principal component analysis[J]. Journal of the ACM, 2011, 58(3): 1 – 37.

[18] Nene S A, Nayar S K, Murase H. Columbia object image library (COIL – 20)[R]. Technical Report CUCS – 005 – 96, 1996.

[19] Belongie S, Kriegman D, Ramamoorthi R. UCSD computer vision[EB/OL]. [2014 – 07 – 02]. http://vision. ucsd. edu/content/yale-face-database.

[20] Cieri C, Graff D, Liberman M, et al. The TDT – 2 text and speech corpus [C]//Proceedings of the DARPA Broadcast News Workshop, 1999: 57 – 60.

[21] Hull J J. A database for handwritten text recognition research[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(5): 550 – 554.

[22] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]//Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005: 57 – 64.

[23] Hoyer P O. Non-negative matrix factorization with sparseness constraints [J]. Journal of Machine Learning Research, 2004, 5: 1457 – 1469.

[24] Vogt J, Roth V. A complete analysis of the l_1, p Group-Lasso[C]//Proceedings of the 29th International Conference on Machine Learning, 2012.

[25] Chartrand R. Exact reconstruction of sparse signals via nonconvex minimization[J]. IEEE Signal Processing Letters, 2007, 14(10):707 – 710.

[26] Chartrand R, Staneva V. Restricted isometry properties and nonconvex compressive sensing[J]. Inverse Problems, 2008, 24(3):1 – 14.

[27] Xu Z B, Chang X Y, Xu F M, et al. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(7): 1013 – 1027.

[28] Liu G C, Lin Z C, Yu Y. Robust subspace segmentation by low-rank representation [C]// Proceedings of the 27th International Conference on Machine Learning, 2010:663 – 670.