半监督边缘判别嵌入与局部保持的维度约简①

兰远东, 高 蕾, 曾少宁, 曾树洪

(惠州学院 计算机科学系, 惠州 516007)

摘 要: 为了对高维数据进行降维处理, 提出了半监督学习的边缘判别嵌入与局部保持的维度约简算法. 通过最小化样本与其所属类别的中心点之间的距离, 使得样本在投影子空间中能够保持其领域的拓扑结构; 再通过最大化不同类别边缘间的距离, 使得类别间的分离度在投影子空间中得到增强. 实验结果表明: 半监督边缘判别嵌入与局部保持的维度约简算法能够获得初始特征空间的较好的投影子空间.

关键词: 降维; 半监督学习; 局部保持; 分类; 机器学习

Semi Supervised Marginal Discriminant Embedding and Local Preserving for Dimensionality Reduction

LAN Yuan-Dong, GAO Lei, ZENG Shao-Ning, ZENG Shu-Hong

(Department of Computer Science, Huizhou University, Huizhou 516007, China)

Abstract: In order to reduce the dimension of high-dimensional data, raised edge semi-supervised marginal discriminant embedding and local preserving algorithm for dimensionality reduction is proposed. By minimizing the distance between sample and the center of its category, the local topology of samples is maintained in the projection subspace. And by maximizing the distance between the edges of different categories, the inter scatter of classes is increased in the projection subspace. Experimental results show that the dimensionality reduction algorithm of semi supervised marginal discriminant embedding and local preserving can get a better projection subspace of the initial feature space.

Key words: dimensionality reduction; semi supervised learning; local preserving; classification; machine learning

在对高维复杂数据直接进行分类时, 分类器的时间复杂度往往难以接受[1,2]. 而且有些分类算法在低维数据中表现良好, 但在高维数据中却表现较差. 因此在对高维数据进行分类前, 对数据的维度进行合适的约简就显得非常重要[3,4].

线性判别分析[5](Linear Discriminant Analysis, LDA)是对高维数据进行维度约简的一种经典算法, LDA 寻找原始数据的一个线性投影空间, 在该空间中能够最大化类间间隔(Interclass Scatter)和最小化类内间距(Intra-class Scatter), 在增加类间分离度的同时也增强了类内聚合度. LDA 主要考虑类别间的判别信息, 而忽略类内样本的结构信息. 由于 LDA 假定每一个类别的样本都服从高斯分布, 因此 LDA 不能够有效的处理非高斯分布的数据. 基于这个事实, Yan 等提出了一个新的半监督维度约简算法: Marginal Fisher Analysi(MFA)[6]. MFA 使用本征图来描述类内样本之间的紧致关系, 每个样本与其同类近邻样本相连. 但是 MFA 不能够保证样本的近邻在降维后仍然保持原始的结构. Roweis 和 Saul[7]引入了局部线性嵌入(Locally Linear Embedding, LLE), LLE 是一种无监督的学习方法, 是为了得到高维数据的一个低维嵌入空间, 在低维空间中, 样本的领域得以保持. 与局部降维的聚类算法不同, LLE 将初始数据映射到一个低维度的坐标体系, 不会导致局部最小值. 通过利用局部对称和线性重构, LLE 能够获得初始高维数据的非线性流行的全局结构.

本文基于 LGE 和 MFA 给出了一种新的子空间学习方法: 半监督边缘判别嵌入与局部保持的维度约简

① 基金项目:惠州市科技计划(2011B020006002,2013w10,2012B020004005,2013W15,A511.0220);惠州学院校立自然科学基金(2012YB14)
收稿时间:2014-02-15;收到修改稿时间:2014-03-17

(Semi Supervised Marginal Discriminant Embedding and Local Preserving for Dimensionality Reduction, SS_MDELP_DR)算法. 基于 LGE 和 MFA 构建子空间的意义在于: 1)通过最小化样本与其所属类别的中心点之间的距离, 使得样本在投影子空间中能够保持其领域的拓扑结构; 2)通过最大化不同类别边缘间的距离, 使得类别间的分离度在投影子空间中得到增强; 3)与 LDA 和 MFA 相比, 本文的算法不但考虑了类间边缘信息, 同时通过领域保持还保留了原始数据的邻域结构.

1 MFA与邻域保持

具有 n 个样本的训练集: $X=\{x_1,x_2,...,x_n\}$, 每个样本都属于 C 个类别中的某个类. 类内局部结构之间的关系可定义为:

$$d^w=\sum_{i=1}^{C}\sum_{j=N(i)}\left\|a^TX_j-a^Tm_i\right\|^2 \tag{1}$$

其中, d^w 是一种距离度量方法; $N(i)$是类别为 i 的样本结合; a 是投影方向; m_i是类别 i 的均值向量, 公式(1)可以改写为:

$$\begin{aligned}d^w&=\sum_{i=1}^{C}\sum_{j\in N(i)}a^T(x_j-m_i)(x_j-m_i)^Ta\\&=a^T\left(\sum_{i=1}^{C}\sum_{j\in N(i)}(x_jx_j^T-2x_jm_i^T+m_im_i^T)\right)a\\&=a^T\left(\sum_{i=1}^{C}\sum_{j\in N(i)}\left(x_jx_j^T-\sum_{i=1}^{C}n_im_im_i^T\right)\right)a\\&=a^T\left(XX^T-XWX^T\right)a\\&=a^TX(I-X)X^Ta\end{aligned} \tag{2}$$

其中, $W=\begin{bmatrix}W^1&0&\Lambda&0\\0&W^2&\Lambda&0\\M&M&O&M\\0&0&\Lambda&W^C\end{bmatrix}$, W 中的 $W^i=\left[1/n_i\right]_{n_i\times n_i}$ 是一个 $n_i\times n_i$ 的矩阵, 每一个元素是 $1/n_i$. 通过最小化每一个类别中的样本与其所属类别的中心点的距离, 样本在投影后就会更加靠近其类别中心. 使用类与类之间的边缘距离来刻画类间分离度, 即:

$$\begin{aligned}d^b&=\sum_{i\in N_{k_2}^-(j)}\sum_{j\in N_{k_2}^-(i)}\left\|a^Tx_i-a^Tx_j\right\|^2\\&=2a^TX(D^b-W^b)X^Ta\end{aligned} \tag{3}$$

为了确保类间和类内距离度量的平衡性, 使用正则化的 Laplacian 矩阵来作为类间距离度量矩阵, 定义如下:

$$\begin{aligned}d^b&=\frac{1}{2}\sum_{i\in N_{k_2}^-(j)}\sum_{j\in N_{k_2}^-(i)}\left\|\frac{a^Tx_i}{\sqrt{D_{ii}^b}}-\frac{a^Tx_j}{\sqrt{D_{jj}^b}}\right\|^2W_{ij}^b\\&=a^TX\left(I-(D^b)^{-\frac{1}{2}}W^b(D^b)^{-\frac{1}{2}}\right)X^T\end{aligned} \tag{4}$$

其中, $I-(D^b)^{-\frac{1}{2}}W^b(D^b)^{-\frac{1}{2}}$ 是正则化的 Laplacian 矩阵, 因此目标函数可以做如下定义:

$$a_{opt}=\arg\max_a\frac{a^TXL^bX^Ta}{a^TX(I-W)X^Ta} \tag{5}$$

在投影空间中, 需要保持样本空间中的邻域关系, 也就是要保持样本原特征空间中的拓扑结构和领域关系. 在本文中, 选取 Roweis 和 Saul[7]给出的 LLE 来保证样本原特征空间中的拓扑结构和领域关系的不变性. 在文献[7]中, Roweis 和 Saul 通过最小化公式(6)来得到权值矩阵.

$$\begin{aligned}&\min\sum_i\left\|x_i-\sum_{j\in N_k(i)}W_{ij}x_j\right\|^2\\&s.t.\ \sum_{j\in N_k(x_i)}W_{ij}=1\quad\forall i\end{aligned} \tag{6}$$

LLE 是无监督学习方法, 不需要考虑类别的先验信息. 但是, 在监督学习中, 需要充分利用样本的标记信息, 也就是只能在样本所属类别中寻找它的邻居. 因此, 可以通过最小化公式(7)来得到权值矩阵.

$$\begin{aligned}&\min\sum_{i\in C_p}\left\|x_j-\sum_{j\in N_{k_1}(i)}W_{ij}^{'}x_j\right\|^2\\&s.t.\ \sum_{j\in N_{k_1}(x_i)}W_{in}^{'p}=1\quad\forall i\in C_p\end{aligned} \tag{7}$$

其中, $N_{k_1}(i)$ 是属于 x_i 的邻居的 k_1 个样本的集合; C_p 是第 p 个类别. 与 LLE 的主要区别是, 本文只在样本所处类别中重构样本. 当每个类别的样本数量较少时, k 可以选取为该类别的样本数量减 1, 即: k 可以约等于 n_p-1, n_p 是类别 p 的样本数量. 因此, 可以通过定义流行正则项 J(a)来保持邻域结构.

$$\begin{aligned}J(a)&=\sum_{p=1}^{C}\sum_{i\in C_p}\left\|a^Tx_i-\sum_{j\in N_{k_1}(x_i)}W_{ij}^{'p}a^Tx_j\right\|^2\\&=\sum_{p=1}^{C}a^TX^p(I^p-W^{'p})^T(I^p-W^{'p})(X^p)^Ta\\&=a^TX(I-W^{'})^T(I-W^{'})X^Ta\end{aligned} \tag{8}$$

其中 $X=(X^1,X^2,...,X^C)$,

$W^{'}=\begin{bmatrix} W^{'1} & 0 & \Lambda & 0 \\ 0 & W^{'2} & \Lambda & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \Lambda & W^{'C} \end{bmatrix}$，$X^p$ 是类别为 p 的样本集合，定义：

$$M_{ij}=\begin{cases}(W'+W'^{T}-W'W'^{T}) & if\ i\neq j \\ 0 & else\end{cases} \tag{9}$$

可以很容易验证：

$$\sum_j M_{ij}=\sum_i M_{ij}=1 \tag{10}$$

设 $L'=(I-W'^{T})(I-W')=I-M$，某些学者指出 L' 约等于迭代正则化 Laplacian 矩阵[8]，即：$L'f=1/2L^2f$，f表示从k维流行到1维流行的映射. 综上所述，可以得到目标函数：

$$a_{opt}=\arg\max_a \frac{a^TXL^bX^Ta}{a^TX(I-W)X^Ta+\beta a^TXL'X^Ta} \tag{11}$$

通过这个目标函数，使用局部流行算法来保证了初始样本空间中的邻域结构，不但缩小了样本到其类内中心的距离，还增加了不同类别边缘之间的距离.

2　算法描述

本文的维度约简算法基于领域保持和边缘判别分析，缩小了样本与类内样本间的距离，使得样本的类内聚合度在投影空间中得到增强；增加了不同类别边缘之间的距离，能够获得更好的类间分离度. 具体算法步骤描述如下：

算法名称: SS_MDELP_DR

输入: 样本集 $X=\{x_1,x_2,...,x_n\}$，$x_i\in R^n$

输出: 样本集 $X=\{x_1,x_2,...,x_n\}$，$x_i\in R^d$

01: For i = 1 to C　//C为样本类别数量

　寻找样本x_i的k_2个近邻：如果x_j属于x_i的k_2近邻或者x_i属于x_j的k_2近邻，$w_{ij}^b=1$，否则 $w_{ij}^b=0$；

02:　End For

03: $d^w=a^TX(I-W)X^Ta$ //计算样本与其类中心之间的距离

04: $J(a)=a^TX(I-W^{'})^T(I-W^{'})X^Ta$　//W'的流行表达

05:　Maximize Eq.11

06: $XL^bX^Ta=\lambda X[(I-W)+\beta L']X^Ta$　//得到d维特征空间

3　实验结果

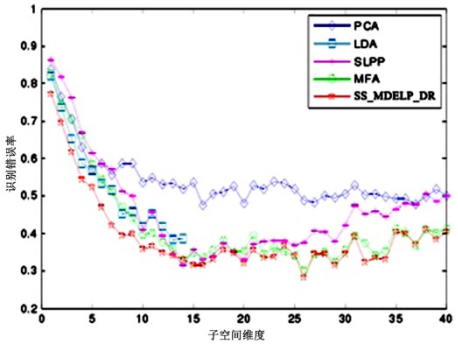
应用本文的维度约简算法来做面部图像识别，在实验中使用比较流行的基准数据集 Yale[9,10]，图 1 显示了 Yale 数据集中的部分数据. 在 Yale 数据集中有 15 个人的面部图像，每一个人有 11 张图像. 在实验中，将本文提出的维度约简算法与 PCA[11]、LDA[12]、SLPP[13]、MFA[6]等算法进行对比. 实验过程中，将所有的面部图像缩减到 32×32 像素.

PCA 是一种降维的统计方法，它借助于一个正交变换，将其分量相关的原随机向量转化成其分量不相关的新随机向量，这在代数上表现为将原随机向量的协方差阵变换成对角形阵，在几何上表现为将原坐标系变换成新的正交坐标系，使之指向样本点散布最开的 p 个正交方向，然后对多维变量系统进行降维处理. LDA 的基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性. 使用这种方法能够使投影后模式样本的类间散布矩阵最大，并且同时类内散布矩阵最小. SLPP 通过使用一个调和函数来保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离，即模式在该空间中有最佳的可分离性.

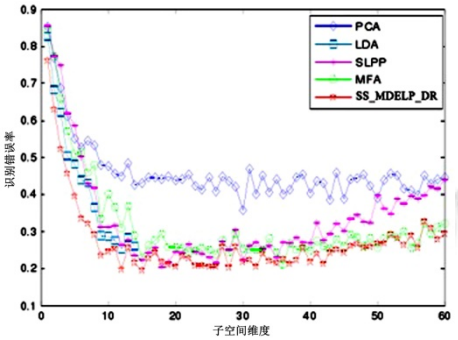
图 1　Yale 数据集中的部分人脸图像

为了确保所有算法同处于一个公平的测试环境，将所有样本分为 3 部分，分别作为训练集、验证集和测试集. 训练集用于学习得到算法的投影子空间；验证集用于获得最佳投影子空间的维度；测试集用于测试最佳投影子空间的识别率. 在验证和测试阶段，使用最近邻标准来分类.

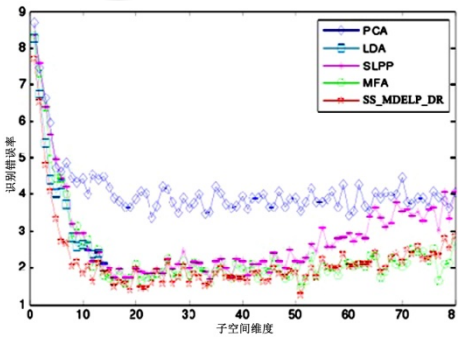
对于 Yale 数据集，由于其包含了每个人的 11 个面部图像数据，实验中随机选择(3 或 5 或 7)个样本作为训练集，将余下的一半作为验证集一半作为测试集. 独立运行每个算法 10 次，得到最终的平均识别错误率. 图 2 显示了当训练集样本为(3,5 或 7)时，在各个相应维度的子空间中，算法在验证集上的识别错误率. 表 1 显示了各个算法在其最佳维度子空间中对测试集的识别率，可以看出本文提出的算法的识别错误率要低于其他几种算法的识别错误率.



(a)训练集为数据集的 3/11



(b) 训练集为数据集的 5/11



(c) 训练集为数据集的 7/11

图 2 不同维度子空间中在验证集上的识别错误率

表 1 测试集上的最佳识别错误率(均值±标准差)

训练样本数量	3	5	7
PCA	50.16±4.71	40.44±3.79	34.33±6.86
LDA	39.50±4.78	27.56±6.39	23.67±7.24
SLPP	33.83±5.45	25.34±7.33	21.67±8.38
MFA	34.17±4.10	22.44±4.38	17.33±7.24
SS_MDELP_DR	33.67±3.99	20.00±5.24	16.33±7.76

4 总结

在分类问题中, 经常会遇到高维复杂数据, 对这样的数据直接进行分类处理, 将消耗大量的时间和能源. 因此, 分类前对数据进行降维处理, 显得非常重要. 本文提出了一种新的半监督线性维度约简算法, 该算法基于邻域保持和边缘判别分析. 该算法在邻域保持效果上, 性能优于 LDA 和 MFA. 通过添加流行正则化项, 使得本文的维度约简算法能够很好的刻画类内样本的分布. 另外, 通过最大化不同类别边缘之间的距离, 获得了更好的类间分离度.

参考文献

1 李正欣,张凤鸣,张晓丰,等.多元时间序列特征降维方法研究.小型微型计算机系统,2013,34(2):338–344.

2 李惠君,张利辉,刘雪飞,等.复杂仿真增量数据可视化中的降维研究.系统仿真学报,2013,25(10):2278–2284.

3 赵东红,王来生,张峰.遗传算法的粗糙集理论在文本降维上的应用.计算机工程与应用,2013,48(36):125–128.

4 张春红,胡清源,程时端.基于降维算法的分布式语义资源搜索.北京邮电大学学报,2013,36(2):74–78.

5 Xanthopoulos P, Pardalos PM, Trafalis TB. Linear Discriminant Analysis. Robust Data Mining. Springer New York, 2013: 27–33.

6 Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40–51.

7 Roweis S, Sau L. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(22): 2323–2326.

8 Lin H, Hong Y, Shu J. Some relations between the eigenvalues of adjacency, laplacian and signless laplacian matrix of a graph. Graphs and Combinatorics, 2013: 1–9.

9 刘剑,龚志恒,吴成东,等.一种基于改进高斯过程隐变量模型的多角度人脸识别算法.电子与信息学报,2013,35(9):2033–2039.

10 郭丽,郑忠龙,贾炯,等.一种有监督的线性降维人脸识别算法.计算机工程,2013,39(11):169–173.

11 Campbell MC, Markham J, Flores H, et al. Principal component analysis of PiB distribution in Parkinson and Alzheimer diseases. Neurology, 2013,81(6): 520–527.

12 Asafu-Adjei JK, Sampson AR, Sweet RA, et al. Adjusting for matching and covariates in linear discriminant analysis. Biostatistics, 2013, 14(4): 779–791.

13 Zhang WQ, Yang HZ. A method of multiple soft-sensors based on SLPP. Journal of East China University of Science and Technology, 2012, 38(6): 724–728.