

稀疏近似最近特征空间嵌入标签传播^{*}

陶剑文¹, Fu-Lai CHUNG², 王士同^{2,3}, 姚奇富⁴

¹(浙江大学 宁波理工学院 信息科学与工程学院, 浙江 宁波 315100)

²(香港理工大学 电子计算学系, 香港)

³(江南大学 数字媒体学院, 江苏 无锡 214122)

⁴(浙江工商职业技术学院 电子与信息工程学院, 浙江 宁波 315012)

通讯作者: 陶剑文, E-mail: jianwen_tao@aliyun.com

摘要: 针对现有的基于图的半监督学习(graph-based semi-supervised learning, 简称 GSSL)方法存在模型参数敏感和数据空间判别信息不充分等问题, 受最近特征空间嵌入和数据稀疏表示思想的启发, 提出一种稀疏近似最近特征空间嵌入标签传播算法 SANFSP(sparse approximated nearest feature space embedding label propagation). SANFSP 首先利用特征空间嵌入投影点来稀疏表示原始数据; 然后, 度量原始数据和稀疏近似最近特征空间嵌入投影间的相似性; 进而提出稀疏近似最近特征空间嵌入正则化项; 最后, 基于传统 GSSL 方法的标签传播算法, 实现数据标签的平滑传播. 同时, 还将 SANFSP 算法简单拓展到 out-of-sample 学习. SANFSP 算法在人造和实际数据集(如人脸识别、可视物件识别以及手写数字分类等)上取得了有效的实验结果.

关键词: 半监督学习; 稀疏表示; 标签传播; 最近特征空间嵌入

中图法分类号: TP181

中文引用格式: 陶剑文, Chung FL, 王士同, 姚奇富. 稀疏近似最近特征空间嵌入标签传播. 软件学报, 2014, 25(6): 1239-1254. <http://www.jos.org.cn/1000-9825/4556.htm>

英文引用格式: Tao JW, Chung FL, Wang ST, Yao QF. Label propagation using sparse approximated nearest feature space embedding. Ruan Jian Xue Bao/Journal of Software, 2014, 25(6): 1239-1254 (in Chinese). <http://www.jos.org.cn/1000-9825/4556.htm>

Label Propagation Using Sparse Approximated Nearest Feature Space Embedding

TAO Jian-Wen¹, Fu-Lai CHUNG², WANG Shi-Tong^{2,3}, YAO Qi-Fu⁴

¹(School of Information Science and Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China)

²(Department of Computing, Hong Kong Polytechnic University, Hong Kong, China)

³(School of Digital Media, Jiangnan University, Wuxi 214122, China)

⁴(School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012, China)

Corresponding author: TAO Jian-Wen, E-mail: jianwen_tao@aliyun.com

Abstract: There exist several problems in existing graph-based semi-supervised learning (GSSL) methods such as model parameters sensitiveness and insufficient discriminative information in data space, etc. To address those issues, this paper proposes a sparse approximated nearest feature space embedding label propagation (SANFSP) algorithm, which is inspired by both ideas of nearest feature space embedding and that of sparse representation. SANFSP first sparsely reconstructs data from original space using its feature space embedding projection images, and then measures the similarity between original data and its sparse approximated nearest feature space embedding projection points, thus proposing a sparse approximated nearest feature space embedding regularizer. At last, SANFSP completes label propagation procedure by using classical label propagation algorithm. The study also derives an easy way to extend

* 基金项目: 教育部人文社会科学研究规划基金(13YJAZH084); 浙江省自然科学基金(LY14F020009, LY13F020011); 宁波市自然科学基金(2013A610065, 2013A610072); 香港理工大学基金(G-UA68)

收稿时间: 2013-03-02; 定稿时间: 2013-11-14

SANFSP to out-of-sample data. Promising experimental results are obtained on several toy and real-world classification tasks such as face recognition, visual object recognition and digit classification.

Key words: semi-supervised learning; sparse representation; label propagation; nearest feature space

训练学习有效的分类器通常需要一个带标签的训练数据集^[1],而在许多实际的模式分类和数据挖掘应用中,我们时常面临缺少足够数量的带标签训练数据的问题;而同时,大量无标签数据集可能会非常容易获得.为了有效利用大量无标签数据和有限的带标签数据,半监督学习(semi-supervised learning,简称 SSL)^[2,3]方法得以提出.目前,已有大量 SSL 方法被提出,详细信息见文献[4],其中取得的一个杰出成就就是基于图的 SSL (graph-based SSL,简称 GSSL)模型^[3,5].因 GSSL 所具有的优雅数学形式和独特的效能,其已发展成为机器学习领域热门研究主题之一^[3,5].GSSL 模型将整个数据集建模为图结构 $G=(V,E)$,其中, V 为顶点集, E 为边集.每个边 $e_{ij} \in E$ 被赋予一个非负权值 $w_{ij} \geq 0$,以反映数据点对 i 和 j 间的相似性.图 G 可以为有向(即 $w_{ij} \neq w_{ji}$)或无向(即 $w_{ij} = w_{ji}$),本文仅关注无向图的情况.一般来说,存在两种基于图的 SSL 任务:1) 演绎学习(或称 transduction 学习),其仅旨在预测无标签顶点的分类标签^[2,6-11];2) 归纳学习(或称 induction 学习),其试图归纳出一个在整个样本空间具有最低误差率的决策函数^[12].显然,归纳学习困难且复杂^[8].

研究表明,SSL 问题的关键是先验一致性假设^[3,6],GSSL 方法所持的基本假设为聚类假设^[3],即:如果两个数据点间连接通路仅穿过高密度区,则二者具有相同的类标签.Zhou 等人^[6]进一步研究了该假设背后的几何本质属性,即:1) 相邻点可能具有相同类标签;2) 在相同结构(如聚类或子流形)上的点也可能具有相同类标签.第 1 个假设为局部假设,第 2 个则为全局假设.聚类假设告诉我们,在学习中应充分考虑包含于数据集中的局部和全局信息.聚类假设同近年发展起来的非线性降维方法(或称流形学习方法^[13])具有直接的关联性,因为这些降维方法的中心思想是为数据集构建一个保留数据局部结构的低维全局坐标系统.LLE^[13]和 LPP^[14]是两个最具代表性的流形学习方法,其均是将数据集建模为一个如 GSSL 方法中的图模型,以便寻求一个保留数据间图结构的低维嵌入坐标系统.尽管现有的 GSSL 方法已被成功应用于许多实际领域,但是经研究发现,现有 GSSL 方法依然普遍存在如下几个关键挑战:

(1) 模型参数依赖问题

现有研究指出,GSSL 算法性能在一定程度上依赖于图边的权值^[8],其通常采用 k -最近邻(k -NN)或 Gaussian 核相似性(GKS)等方法计算取得.而在实际应用中,优化的最近邻数 k 和 Gaussian 核带宽通常是很难确定的,从而导致现有 GSSL 方法在具体的应用中难以取得最优性能.对此,Wang 等人^[8,9]提出一种线性邻居传播算法(linear neighbor propagation,简称 LNP),其首先近似构建一个由多个互重叠的线性近邻块构成的图,其中,各近邻块的边权值通过邻居线性投影重构计算得到.LNP 方法虽然能够通过修正数据点与其 k -NN 间的权值来改善传统 k -NN 方法的性能,但是其仍然依赖传统的欧式距离来预定义数据点的 k -NN,换句话说,LNP 依然未能彻底解决传统 SSL 方法中如何有效确定数据点的邻居数这个根本问题.

(2) 面对复杂的学习环境,现有 GSSL 方法所面临的鲁棒性问题

一般现有的 GSSL 方法性能通常对噪声(或桥接点)和缺失数据(如人脸数据中的遮罩)^[8]等干扰信号较敏感,会导致现有 GSSL 方法错误地将标签传播到异类,降低了现有 GSSL 方法在复杂噪声环境下的学习性能.

(3) 小样本学习问题

对于人脸图像识别等小样本学习问题,由于学习数据集中点对间的相似性判别信息有限^[15],从而在一定程度上导致现有 GSSL 方法面临判别信息不充分的问题.

现有研究显示^[16],人脸图像数据的稀疏表示具有更自然的判别力.为此,Wright 等人^[16]提出一种稀疏表示分类器(sparse representation classifier,简称 SRC).为了解决 GSSL 方法中存在的模型参数难以确定的问题,Hong 等人^[17]和 Fan 等人^[18]分别提出稀疏表示加权图正则化 SSL 方法.为了克服人脸图像分类这类小样本学习任务所面临的判别信息缺少的问题,Chen 等人^[15]从模式判别分析的角度,提出一种 NFS 嵌入人脸识别方法.

本文受 NFS 方法思想和近来在稀疏表示(sparse representation,简称 SR)理论上的研究进展^[16-20]的启发,提出一种稀疏特征空间嵌入相似性度量.其关键思想在于:利用特征空间嵌入实现数据点的稀疏重构能反映数据

点与特征空间嵌入间的本质相似性结构,从而提供一种度量数据点与特征空间嵌入间的相似性度量方法.基于此,提出一种新颖的稀疏近似最近特征空间嵌入标签传播算法(sparse approximated nearest feature space embedding label propagation,简称 SANFSP),所提方法无需依赖欧式距离来预计算数据点的 k 近邻.与传统 GSSL 方法相比,SANFSP 方法所具有的优势在于:

- (1) 在 SANFSP 方法中,标签传播图模型的构建准则为特征空间嵌入投影的核稀疏表示,而该准则一般优于传统 GSSL 方法中采用的最近邻准则,尤其是对于高维小样本数据集的学习.与传统 GSSL 方法中数据近邻和图边权值的计算分开完成不同,本文方法中邻居大小和图边权值通过 l_1 -范优化过程来进一步确定,使得不同样本具有不同的近邻数,其能增强本文方法对于复杂数据分布学习的自适应性;
- (2) 由于将原始数据分别嵌入到多个(理论上可为无限个)特征空间,扩大了原始数据空间的信息表示容量,使得 SANFSP 方法具有更丰富的模式判别能力,从而能够有效解决小样本学习问题;
- (3) 通过使用稀疏集中索引(sparsity concentration index,简称 SCI)^[16]技术来自然地消除噪声数据或恢复缺失数据,使得 SANFSP 方法在复杂的噪声环境下具有更鲁棒的学习性能;
- (4) 由于同时考虑了数据分布的局部和全局几何结构信息,使得 SANFSP 方法具有更强的泛化性能.

1 SANFSP 算法

1.1 问题描述

给定数据集 $X=\{x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+u}\} \in \mathbb{R}^d$ (d 维欧式空间)和标签集 $L=\{1, 2, \dots, c\}$,其中,前 l 个点 x_i ($1 \leq i \leq l$) 标签为 $y_i \in L$,余下的 u 个数据点 x_j ($l+1 \leq j \leq l+u=N$) 无标签,每个数据点 x_i 均采样自某个固定但未知的分布 $p(x)$,则 SANFSP 方法旨在基于由数据集 X 构成的图 G 实现标签信息的平滑传播,即满足:(1) 标签传播输出应接近图中带标签顶点的标签值;(2) 传播的标签信息应在整个图上平滑.为此,SANFSP 方法基本步骤包括:(1) 原始空间数据在特征空间的嵌入,并获得各原始数据点在特征空间的相应投影点集;(2) 利用特征空间投影点集稀疏重构各原始数据,获得稀疏重构矩阵 S ,并以此构建加权图模型 $G=(X,S)$,其中, X 为数据集, S 为图边权值;(3) 基于图模型 G 实现标签传播.

1.2 特征空间嵌入计算

考虑原始空间的对象集 $X=\{x_1, x_2, \dots, x_N\}$,给定对象 $x_i \in X$ 到特征空间的距离为

$$\|x_i - f^{(P)}(x_i)\|^2 \quad (1)$$

其中, $f^{(P)}$ 为 P ($P=1, 2, \dots, N-1$) 个对象生成的特征空间, $f^{(P)}(x_i)$ 为对象 x_i 在特征空间 $f^{(P)}$ 的嵌入(或投影).图 1 显示了 $P=2$ 和 $P=3$ 情况下的特征空间嵌入及其权值.当 $P=1$ 时,公式(1)表示点对间的距离度量,对象 x_i 在特征空间的嵌入为原始空间对象集.下面重点分析当 $P \geq 2$ 时,对象 x_i 在特征空间的嵌入 $f^{(P)}(x_i)$ 计算方法.

(1) 当 $P=2$ 时,公式(1)变成点 x_i 到特征线 $\overline{x_a x_b}$ 间的距离度量(如图 1(a)所示),表示为 $\|x_i - f_{a,b}^{(2)}(x_i)\|$,投影点可计算为

$$f_{a,b}^{(2)}(x_i) = x_a + \lambda_{a,b}(x_a - x_b) \quad (2)$$

其中,权值 $\lambda_{a,b} = (x_i - x_a)^T (x_a - x_b) / (x_a - x_b)^T (x_a - x_b)$ ($i \neq a \neq b$),则有:

$$\begin{aligned} x_i - f_{a,b}^{(2)}(x_i) &= x_i - x_a + \lambda_{a,b}(x_a - x_b) \\ &= x_i - (1 - \lambda_{a,b})x_a - \lambda_{a,b}x_b \\ &= x_i - \lambda_{a,b}x_a - \lambda_{b,a}x_b. \end{aligned}$$

其中, $\lambda_{a,b} + \lambda_{b,a} = 1$.

(2) 当 $P=3$ 时,公式(1)变成点到特征面 $E_{r,a,b}$ 间的距离度量,表示为 $\|x_i - f_{r,a,b}^{(3)}(x_i)\|$,其中,特征平面 $E_{r,a,b}$ 由对象 x_r, x_a 和 x_b 张成(如图 1(b)所示),特征平面 $E_{r,a,b}$ 的嵌入(或投影) $f_{r,a,b}^{(3)}(x_i)$ 由公式(3)计算获得:

$$\begin{aligned}
 f_{r,a,b}^{(3)}(x_i) &= \Gamma_{r,a,b}(\Gamma_{r,a,b}^T \Gamma_{r,a,b})^{-1} \Gamma_{r,a,b}^T(x_i - x_r) + x_r \\
 &= [(x_a - x_r)(x_b - x_r)][\lambda_a \lambda_b]^T + x_r \\
 &= (1 - \lambda_a - \lambda_b)x_r + \lambda_a x_a + \lambda_b x_b \\
 &= \lambda_r x_r + \lambda_a x_a + \lambda_b x_b
 \end{aligned}
 \tag{3}$$

其中, $\Gamma_{r,a,b} = [(x_a - x_r)(x_b - x_r)]$ 为一个 $r \times 2$ 矩阵, 矩阵 $(\Gamma_{r,a,b}^T \Gamma_{r,a,b})^{-1} \Gamma_{r,a,b}^T(x_i - x_r)$ 表示为 $[\lambda_a \lambda_b]^T$, 且 $\lambda_r + \lambda_a + \lambda_b = 1$. 从而, 投影点 $f_{r,a,b}^{(3)}(x_i)$ 被表示为对象 x_r, x_a 和 x_b 的线性组合.

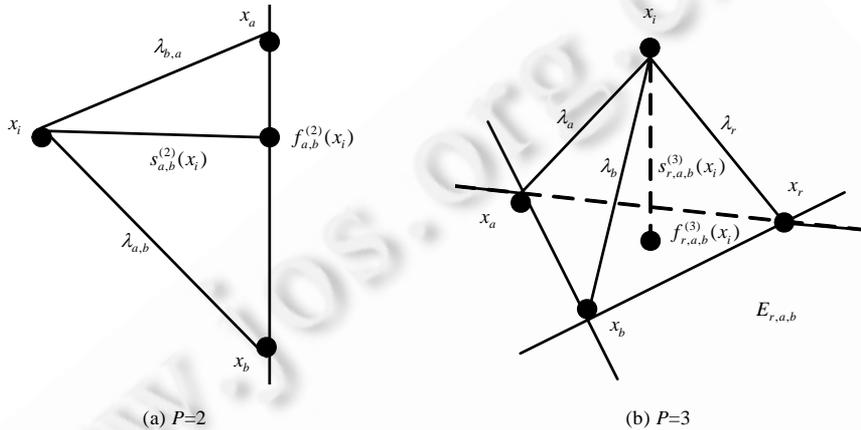


Fig.1 Feature space embedding and the weight setting for $P=2$ and $P=3$

图 1 特征空间嵌入及其权值: $P=2, P=3$

(3) 一般地, 对于 $P > 3$, 公式(1)变成点到特征空间之间的距离度量, 表示为 $x_i - f^{(P)}(x_i)$, 对象 x_i 在特征空间 $f^{(P)}$ 的嵌入 $f^{(P)}(x_i)$ 可由公式(4)计算获得:

$$f^{(P)}(x_i) = \Gamma_{1:P}(\Gamma_{1:P}^T \Gamma_{1:P})^{-1} \Gamma_{1:P}^T(x_i - x_1) + x_1 = \sum_{j=1}^P \lambda_j x_j
 \tag{4}$$

其中, $\Gamma_{1:P} = [(x_2 - x_1)(x_3 - x_1) \dots (x_P - x_1)]$, $\sum_{j=1}^P \lambda_j = 1$.

1.3 稀疏特征空间嵌入Laplacian图构造

如下两条理由能够说明为何数据的稀疏表示适于图的构建:

(1) 在典型的 k -邻居图构建中, 稀疏性具有重要地位: 一方面, 稀疏性刻画了数据分布的局部性; 另一方面, 稀疏性有效节省计算成本和存储空间. 但是, 传统的基于 k -NN 和高斯函数构建的 k -邻居图的稀疏性依赖于人工设定的邻居数和高斯核参数 σ . 文献[8]研究指出: 在只有少量标签数据的情况下, 难以可靠地选取模型参数, 即, 难以确定优化的参数 σ . 为此, 需要寻求一种更可靠、稳定的方法来构建图模型 G ;

(2) 最稀疏的表示自然地具有判别性. 因为我们的最终目标是对目标数据实现分类, 所以期望图数据包含尽可能多的判别信息, 即来自相同类的两个数据点通过边连接起来. 对于典型的 k -NN 图, 上述所期望的属性严重依赖近邻准则在原始空间实施效果的好坏^[17], 但不幸的是, 对于原始高维数据(如人脸图像数据), 最近邻准则通常不能取得好的性能^[14, 17]. 相比之下, 近来研究^[16]显示, 稀疏表示具有自然的判别力, 并能在高维数据环境下取得较好的性能; 而且, 该判别力仅与类数紧密相关, 与样本数无关. 因此, 基于 SR 构建的图模型在无需大量标签数据的情况下能包含更多的判别信息.

设 $T^{(P)}$ 为对象集 X 生成的 C_p^{N-1} 个可能的特征空间集, $T^{(P)}(x_i)$ 为对象 x_i 在特征空间集 $T^{(P)}$ 中投影点集. 本文试图避开传统的 GSSL 方法中所采用的点对间关系度量方法, 而采用特征空间投影点集 $T^{(P)}(x_i)$ 来稀疏表示原始空间对象 $x_i \in X$, 从而构建一个加权图 $G = \{T^{(P)}, S\}$. 其中, S 为稀疏表示权值, $s^{(P)}(x_i) \in S$ 代表对象 x_i 和特征空间嵌入

$f^{(P)}(x_i) \in T^{(P)}(x_i)$ 间的稀疏表示权值.即通过求解如下 l_1 范最小化问题来为各数据点寻求一个稀疏重构权值向量:

$$\min \|x_i - s^{(P)}(x_i)f^{(P)}(x_i)\|^2 + C \|s_i^{(P)}\|_1 \tag{5}$$

其中, C 为正则化参数,以控制重构稀疏性和重构补偿间的平衡; $s_i^{(P)} = [s_{i1}^{(P)}, s_{i2}^{(P)}, \dots, s_{i(j-1)}^{(P)}, 0, s_{i(j+1)}^{(P)}, \dots, s_{iC_p^{N-1}}^{(P)}]^T$ 是一个 C_p^{N-1} 维列向量,其中,第 j 个元素等于 0 表示特征空间投影点 $\tilde{x}_j \in T^{(P)}(x_i)$ 从 $T^{(P)}(x_i)$ 中移除, $s_{ij}^{(P)}$ ($i \neq j$) 表示 \tilde{x}_j 对 x_i 的重构贡献度.本文进一步约束 $\sum_j s_{ij}^{(P)} = 1$, 且 $s_{ij}^{(P)} \geq 0$. 特别地,当 $P=2$ 和 $P=3$ 时,公式(5)分别变为

$$\min \left\| x_i - \sum_{i \neq a \neq b} s_{a,b}^{(2)}(x_i) f_{a,b}^{(2)}(x_i) \right\|^2 + C \|s_i^{(2)}\|_1,$$

$$\min \left\| x_i - \sum_{i \neq r \neq a \neq b} s_{r,a,b}^{(3)}(x_i) f_{r,a,b}^{(3)}(x_i) \right\|^2 + C \|s_i^{(3)}\|_1,$$

其中, $s_i^{(2)}$ 和 $s_i^{(3)}$ 分别为由稀疏权值 $s_{a,b}^{(2)}(x_i)$ 和 $s_{r,a,b}^{(3)}(x_i)$ 构成的 C_2^{N-1} 维和 C_3^{N-1} 维列向量.

可以通过扩展特征-符号搜索法^[19,20]来求解公式(5)中稀疏编码问题.在求得所有数据点的优化重构稀疏向量 $\widehat{s}_i^{(P)}$ ($1 \leq i \leq N$)后,便可构建稀疏权值矩阵 $S = [\widehat{s}_1^{(P)}, \widehat{s}_2^{(P)}, \dots, \widehat{s}_N^{(P)}]$.值得说明的是, S 中元素 $s_{ij}^{(P)}$ 并非数据点对 x_i 和 \tilde{x}_j 间简单的相似性度量,矩阵 S 本质上有别于传统的图正则化算法(如 LPP^[14])中的权值矩阵.对于 GSSL 问题,采用稀疏矩阵 S 作为图权值矩阵具有如下有效属性:(1) 在稀疏矩阵 S 中,各权值向量 $\widehat{s}_i^{(P)}$ 均遵从重要的对称性,即旋转不变性(满足公式(5)约束)和转换不变性(满足约束 $\mathbf{1}_{C_p^{N-1}}^T \widehat{s}_i^{(P)} = 1$,其中, $\mathbf{1}_{C_p^{N-1}}$ 代表 C_p^{N-1} 维全 1 列向量),使得权值矩阵 S 在一定程度上能够反映数据的本质几何属性;(2) 即使在无类标签的情况下,权值矩阵 S 中也能自然地保留数据的判别信息.

1.4 稀疏近似最近特征空间嵌入标签传播算法

在本节,我们将利用公式(5)构建的核稀疏图和一个迭代过程来有效解决数据标签传播的问题.设 F 表示定义于样本集 X 上的分类函数集,且 $\forall f \in F$,则可赋予每个数据点 x_i 一个实值 $f_i = f(x_i)$,无标签数据 x_u 的标签由 $f_u = f(x_u)$ 的符号确定.在每次迭代中,使每个数据从其稀疏重构对象中吸收部分标签信息,且保留其初始状态的部分标签信息,这样,在第 $t+1$ 次迭代时, x_i 的标签为

$$f_i^{t+1} = \alpha \sum_{j \neq i} M_{ij} f_j^t + (1 - \alpha) y_i \tag{6}$$

其中, $0 \leq \alpha \leq 1$ 控制 x_i 从其重构对象吸收标签信息部分, $M_{ij} = (S + S^T - S^T S)_{ij}$ (或 $M_{ij} = (S)_{ij}$).令 $y = \{y_1, y_2, \dots, y_N\}^T$ 且 $y_i = L_i \in L$ ($1 \leq i \leq l$),其中, $L = \{0, 1\}$ 为标签集, $y_u = 0$ ($l+1 \leq i \leq l+u$).令 $f^t = (f_1^t, f_2^t, \dots, f_N^t)^T$ 为在第 t 次迭代的预测标签向量, $f^0 = y$.公式(6)迭代方程重写为

$$f^{t+1} = \alpha M f^t + (1 - \alpha) y \tag{7}$$

本文将采用公式(7)更新各数据对象的标签直至收敛,即数据的预测标签在经过几次迭代后不再发生变化.

下面将 SANFSP 算法拓展为多分类问题:设有 c 个分类,标签集为 $\bar{L} = \{1, 2, \dots, c\}$.令 K 为 $N \times c$ 矩阵集,其中,矩阵元素为非负的实数值.任意矩阵 $F = [F_1^T, F_2^T, \dots, F_N^T] \in K$ 对应 X 上的一个特定的分类,即,数据 x_i 分类为 $y_i = \operatorname{argmax}_{j \in \bar{L}} F_{ij}$.因此, F 也可看成一个标签函数.初始地,设 $F_0 = Y$,其中,如果 x_i 标记为 j ,则 $Y_{ij} = 1$; 否则, $Y_{ij} = 0$; 对于无标签数据点 $x_u, Y_{uj} = 0$ ($1 \leq j \leq c$).SANFSP 算法的主要步骤描述见表 1.

2 讨论

2.1 算法收敛性

定理 1^[6]. 公式(12)中计算的序列 $\{f^t\}$ 收敛于下式:

$$f^* = (1 - \alpha)(I - \alpha M)^{-1} y \tag{8}$$

根据定理 1,以 f_i 作为分类函数,使得 SANFSP 成为一站式算法,即,只需一步便能预测所有数据标签.同样,

对于多分类情况,只要将公式(8)中的 y 简单地替换为 Y ,可得如下多分类预测函数:

$$F^*=(1-\alpha)(I-\alpha S)^{-1}Y \tag{9}$$

则,每个数据对象的标签可由 $y_i = \arg \max_{j \leq c} F_{ij}^*$ 确定.

Table 1 SANFSP algorithm

表 1 SANFSP 算法

算法1.	SANFSP.
输入:	样本集 X 、参数 C, α 以及初始化标签向量 y ;
输出:	无标签数据的预测标签.
Step 1	根据公式(7)~公式(9)分别计算数据 $x_i \in X$ 的特征空间投影集 $T^{(P)}(x_i), P=2,3, \dots, N-1, i=1,2, \dots, N$
Step 2	通过求解公式(10)中 l_1 范最小化问题构建稀疏图 $G=(T^{(P)}, S)$,从而构建传播矩阵 $M=S+S^T-S^T S$.
Step 3	重复 $F^{(i+1)}=\alpha M F^{(i)}+(1-\alpha)Y$ 直到收敛.
Step 4:	令 F^* 为序列 $F^{(i)}$ 的极限,输出各数据点 x_i 的标签 $y_i = \arg \max_{j \leq c} F_{ij}^*$.

2.2 正则化框架

2.2.1 稀疏近似最近特征空间嵌入正则化

目前,许多降维方法均利用一个线性映射函数将原始特征空间的数据矩阵 X 映射到某个低维表示 $F^{[21]}$,即: $F=A^T X$,其中, A 为映射矩阵.同样,MR 方法^[12]也通过一个线性函数 $h(X)$ 来连接预测标签矩阵 F 和数据矩阵 X ,即: $F=h(X)=A^T X$.由此可知,线性技术提供了一个简单且有效的数据映射方法,其假设低维表示或预测标签可位于数据矩阵 X 所张成的空间.

考虑一个低维变换空间对象集 $Y=\{y_1, y_2, \dots, y_N\}$,其中, $y_i=a^T x_i(x_i \in X)$ 为 x_i 的某个低维表示或预测类标签向量, a 为低维变换矩阵,则给定对象 y_i 到特征空间的距离为 $\|y_i - l^{(P)}(y_i)\|_2$,其中, $l^{(P)}$ 为对象集 Y 中 P 个对象生成的特征空间, $l^{(P)}(y_i)$ 为对象 y_i 在特征空间 $l^{(P)}$ 的投影.设 $Y^{(P)}$ 为对象集 Y 生成的 C_p^{N-1} 个可能的特征空间集,则最近特征空间嵌入(NFS)人脸识别方法^[15]最小化如下目标函数:

$$L_1 = \sum_i \left\| \sum_{l^{(P)} \in Y^{(P)}} (y_i - l^{(P)}(y_i)) w^{(P)}(y_i) \right\|^2 \tag{10}$$

$$L_2 = \sum_i \sum_{l^{(P)} \in Y^{(P)}} \|y_i - l^{(P)}(y_i)\|^2 w^{(P)}(y_i) \tag{11}$$

其中, $w^{(P)}(y_i)$ 为特征空间投影点的权值.对于特定对象 y_i ,共可生成 C_p^{N-1} 个特征空间嵌入.对于 N 个对象及其相应投影对象集 $l^{(P)}(y_i)(i=1,2, \dots, N)$,权值 $w^{(P)}(y_i)$ 构成一个 $N \times C_p^{N-1}$ 的连接关系矩阵.在基于 NFS 的判别分析中, $l^{(P)}(y_i)(i=1,2, \dots, N)$ 为 k -最近邻特征空间嵌入集.当 $P=1$ 时,公式(10)变为 LLE 方法^[13],公式(11)变为 LPP 方法^[14].为了方便起见,权值 $w^{(P)}(y_i)$ 被扩展为一个 $N \times N^P$ 矩阵,其中,部分特征空间嵌入权值不存在.对于不同的 P 值,以上目标函数 L_1 和 L_2 均可变换为 Laplacian 矩阵表示形式^[15].

定理 2^[21]. 设矩阵 $X=[x_1, x_2, \dots, x_N] \in \mathbb{R}^{d \times N}$ 表示原始空间数据集矩阵, d 为数据维数.为简单起见,假定 $X \mathbf{1}_N = 0$,其中, $\mathbf{1}_N$ 表示 $N \times 1$ 维向量.分别定义总体散度矩阵 S_t 、类间散度矩阵 S_b 和类内散度矩阵 S_w 为: $S_t = X X^T, S_b = X G G^T X^T, S_w = X X^T - X G G^T X^T$,其中, $G = Y(Y^T Y)^{-1/2}$ 为一个加权类指示矩阵,且 $G^T G = I_c(c$ 为分类数).

如果 $rank(S_b) = c - 1$ 和 $rank(S_t) = rank(S_w) + rank(S_b)$,则原始空间数据的真实类指示矩阵能被其某个低维线性投影所表示,即,存在某个投影矩阵 $A \in \mathbb{R}^{d \times c}$,使得 $Y = X^T A$,其中, Y 为预测类标签矩阵.

定理 2 中的条件在实际应用中通常对于高维、小样本问题是满足的(如人脸识别应用)^[21].根据以上描述,稀疏权值矩阵 S 在一定程度上能够反映数据与特征空间嵌入点间的本质几何特性,且包含了数据的自然判别信息.根据定理 2,有如下假设:

假设 1. 与特征空间相应的标签空间具有和特征空间相似的几何特性,即原始数据的真实类标签也能被标签空间嵌入稀疏表示.

基于假设 1,本文提出一种稀疏标签空间嵌入正则项.令 $F=(f_1, f_2, \dots, f_c) \in \mathbb{R}^{N \times c}$ 为类指示矩阵,其中 $f_j = [f_j(x_1),$

$f_j(x_2), \dots, f_j(x_N)]^T (1 \leq j \leq c)$, 则可通过最小化如下目标函数来构建稀疏标签空间嵌入正则项, 其最好地保留了对象 x_i 的特征空间嵌入的稀疏权重向量:

$$F_1 = \sum_i \left\| \sum_{j: L^{(P)} \in F^{(P)}} (f_i - L^{(P)}(f_j)) s_j^{(P)}(x_i) \right\|^2 \tag{12}$$

$$F_2 = \sum_i \sum_{j: L^{(P)} \in F^{(P)}} \|f_i - L^{(P)}(f_j)\|^2 s_j^{(P)}(x_i) \tag{13}$$

其中, $s_j^{(P)}(x_i)$ 为第 j 个特征空间嵌入稀疏表示权值, $L^{(P)}$ 为 P 个标签对象生成的标签空间, $L^{(P)}(f_i)$ 为对象 f_i 在特征空间 $L^{(P)}$ 的投影, $F^{(P)}$ 为标签集生成的 C_p^{N-1} 个可能的标签空间集. 公式(12)和公式(13)称为稀疏标签空间嵌入目标函数.

定理 3. 稀疏标签空间嵌入目标函数(12)和公式(13)可分别表示为 Laplacian 矩阵形式, 即:

$$F_i = \text{tr}(FLF^T) (i=1, 2).$$

定理 3 的详细证明见附录 1.

定理 4. 设 $F = (f_1, f_2, \dots, f_c) \in \mathbb{R}^{N \times c}$ 为类指示矩阵, 定义变量 $\alpha = 1/(1+\mu)$ 和 $\beta = \mu/(1+\mu)$, 可知 $\alpha + \beta = 1$, 则 SANFSP 通过公式(9)计算的预测结果可通过如下正则化框架导出:

$$Q_{\text{SANFSP}}(F) = \min_F \text{tr}(FLF^T) + \frac{\mu}{2} \text{tr}((F - Y)^T (F - Y)) \tag{14}$$

其中, $Y \in \mathbb{R}^{N \times c}$ 为原始标签矩阵. 如果 x_i 标记为 $y_i = j$, 则 $Y_{i,j} = 1$; 否则, $Y_{i,j} = 0$.

证明: 定理 4 可仿照文献[6]推导证明, 空间所限, 在此省略.

根据定理 4, 目标数据点 $x_u \in X (l+1 \leq u \leq N)$ 的预测标签可由下式确定:

$$f_u = \arg \max_{1 \leq j \leq c} F_{u,j}, l+1 \leq u \leq N \tag{15}$$

公式(15)中, $Q_{\text{SANFSP}}(F)$ 的第 1 项称为平滑项, 其描述了相对于稀疏重构结构的数据标签的总体变化; 其中第 2 项称为拟合项, 其度量预测标签和原始标签的拟合性能.

2.2.2 局部和全局一致学习正则化

本质上来说, 稀疏特征空间嵌入正则化是一种全局学习方法. 但不幸的是, 以全局的方式选择一个分类函数可能不是一个好的策略, 因为学习空间可能并未包含一个适于全局数据集的学习函数^[10]. 然而从待学习的函数集中可以更容易地选出一些适于输入空间局部区域的更好的预测函数, 因此, 将整个输入空间分割成多个局部区域, 然后针对各个局部区域以更有效地实现局部成本函数的最小化. 相反, 采用纯粹的局部学习算法也可能存在问题, 因为在各局部区域可能不具备用以训练局部学习函数的数据^[11]. 因此, 在局部学习正则化的同时, 还需应用一个全局平滑项以根据本质数据分布来平滑预测数据标签, 使得预测标签更精确.

为了捕捉局部判别信息, 本文采用局部核岭回归函数对每个训练数据实施一次回归运算. 当 $P=1$ 时, 公式(5)变为

$$\min \|x_i - Xs_i\|^2 + C \|s_i\|_1, i=1, \dots, N \tag{16}$$

其中, 矩阵 $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{d \times N}$ 表示原始空间数据集矩阵, d 为数据维数. 公式(16)代表训练数据 x_i 的稀疏表示, 令 $X_i = \{x_j\}_{j=1}^k \in \mathbb{R}^{d \times k} \subset X$ 代表数据 x_i 的核稀疏重构对象集, 其中, k 代表数据点 x_i 的稀疏重构对象数(根据经验, 为了获得最好的有判别的 k 个重构模式, 本文设置 $s_{ij} \geq \varepsilon, \varepsilon \in (0, 1)$ 为某个用户定义的相对较小的阈值), 则局部学习正则项可以定义为

$$\min_{W_i, b_i, F_i} \sum_{i=1}^N (\|X_i^T W_i + 1_k b_i^T - F_i\|^2 + \gamma (\text{tr} W_i^T W_i)) \tag{17}$$

其中, $F_i = [f_{i_1}, \dots, f_{i_k}]^T$ 为局部类指示矩阵, W_i 为局部为投影矩阵, b_i 为局部截距向量, $1_k \in \mathbb{R}^{k \times 1}$ 代表全 1 向量. 对公式(17)分别针对变量 W_i, b_i 求导数, 并令结果等于 0, 可得如下分析解:

$$\begin{cases} b_i = \frac{1}{k} F_i^T \mathbf{1}_k \\ W_i = (X_i H_k X_i^T + \gamma I_d)^{-1} X_i H_k F_i \end{cases} \quad (18)$$

其中, $H_k = I_k - \mathbf{1}_k \mathbf{1}_k^T / k$ 为局部中心化矩阵, $I_k \in \mathbb{R}^{k \times k}$ 为单位矩阵. 将公式(18)的结果代入公式(17), 可得局部学习正则化项 $tr(FL_o F^T)$, 其中,

$$L_o = [\delta_1, \dots, \delta_N] \begin{bmatrix} L_{o_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_{o_N} \end{bmatrix} [\delta_1, \dots, \delta_N]^T \quad (19)$$

其中, $L_{o_i} |_{i=1}^N = H_k - X_i^T (X_i X_i^T + \gamma I_d)^{-1} X_i$; $\delta_i |_{i=1}^N$ 为选择矩阵, 满足 $F_i = \delta_i^T F$.

从公式(17)可看出, 局部学习 Laplacian 矩阵 L_o 能有效捕捉数据的局部判别信息.

根据局部与全局一致学习观, 在公式(14)中, 用 $G = \lambda L + (1 - \lambda)L_o$ 替换 L , 其中, $\lambda \in [0, 1]$ 为平衡参数, 从而得到一个基于局部和全局一致学习的 SANFSP 正则化形式:

$$Q_{SANFSPLG}(F) = \min_F tr(FGF^T) + \frac{\mu}{2} tr((F - Y)^T (F - Y)) \quad (20)$$

根据公式(20), 当 $\lambda = 1$ 时, $Q_{SANFSPLG}(F)$ 降级为基于全局学习正则化的形式 $Q_{SANFSP}(F)$.

2.3 Out-of-Sample推理

第1节论述了 SANFSP 算法的主要演绎学习过程, 本节介绍如何将 SANFSP 算法推广到样本外数据(out-of-sample)学习. 按照文献[22]的做法, 为了将 SANFSP 泛化到样本外数据学习, 需要做两件事情: (1) 对于新的样本外测试数据点 x_u 使用和公式(14)(或公式(20))相同的平滑准则类型; (2) 确保样本外数据 x_u 的加入不会影响训练数据集的原始 $Q_{SANFSP}(F)$ 或 $Q_{SANFSPLG}(F)$ 值. 因此, 对于新的测试数据点 x_u , 平滑准则定义为

$$Q_{SANFSP}(f(x_u)) = \sum_{j: L^{(j)} \in F^{(P)}} (f(x_u) - L^{(j)}(f_u)) s_j^{(P)}(x_u) \quad (21)$$

因为 $Q_{SANFSP}(f(x_u))$ 关于 $f(x_u)$ 为凸函数, 故在公式(22)条件下, 其能被最小化:

$$f(x_u) = \sum_{j: L^{(j)} \in F^{(P)}} s_j^{(P)}(x_u) L^{(j)}(f_u) \quad (22)$$

有趣的是, 公式(22)正好是当数据点 x_u 的标签能被标签空间嵌入对象优化重构时的公式, 即为公式(23)的优化解:

$$f(x_u) = \min_{f(x_u)} \left\| f(x_u) - \sum_{j: L^{(j)} \in F^{(P)}} L^{(j)}(f_j) s_j^{(P)}(x_u) \right\|^2 \quad (23)$$

2.4 鲁棒性

当数据集中有无效测试数据时, SANFSP 算法存在潜在问题. 因此, 在实施标签传播前须先确定无标签样本是否有效. 在实际应用场景中, 检测并排斥无效测试样本(或离群点)是模式分类方法的关键能力.

对于每个类 $i (1 \leq i \leq c)$, 令 $\Psi_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$ 为选取与第 i 类相关的稀疏重构系数的特征函数. 对于一个稀疏重构系数向量 $s \in \mathbb{R}^n$, $\Psi_i(s) \in \mathbb{R}^n$ 为一个新的向量, 其非零元素为 s 中与第 i 个类相关的元素.

定义 1(稀疏集中索引(sparsity concentration index, 简称 SCI)^[16]. 稀疏重构系数向量 $s \in \mathbb{R}^d$ 的 SCI 定义为:

$$SCI(s) = \frac{c \cdot \max_i \|\Psi_i(s)\|_1 / \|s\|_1 - 1}{c - 1} \quad (24)$$

根据定义 1, 当 $P=1$ 时, 对于由公式(10)取得的优化解 \hat{s}_i , 如果 $SCI(\hat{s}_i) = 1$, 则表示测试样本 x_i 仅由同一类的对象表示; 如果 $SCI(\hat{s}_i) = 0$, 则表示稀疏重构系数均分于所有类. 从而, 可以选择一个适当的阈值 $\tau \in (0, 1)$, 使得当 $SCI(\hat{s}_i) \geq \tau$ 时, 则认为测试样本 x_i 为有效. 因此, 在稀疏图构造之前, 我们可实施一个预处理过程, 即通过 SCI 测试方法来排除数据集中某些噪声或离群点, 从而增强 SANFSP 算法的鲁棒性能.

3 实验结论与分析

为了评价所提方法 SANFSP 在 SSL 问题上的有效性,本文将在人造数据集和实际数据集上将所提方法与几个代表性的算法进行比较.参与比较的算法包括 GFHF^[7]、LapRLS^[12]、基于稀疏重构技巧的 S-RLSC^[18]、基于线性邻居传播模型(linear neighbor propagation,简称 LNP)^[8]以及基线方法 1-近邻分类算法 1-NN^[23].对于所有算法实验,分别记录 10 次独立实验结果的平均值用于算法性能评价.

对于 GFHF 和 LapRLS 方法,Laplacian 图矩阵采用 Gaussian 函数构建,其中,Gaussian 函数参数采取 5-重交叉验证法来设置,在 GFHF 方法中的 Laplacian 图矩阵的对角元素设为零;在 LapRLS 和 S-RLSC 方法中,为了算法比较的公平起见,正则化参数 γ_A 和 γ_I 的设置与文献[18]中相同,核函数选择为 RBF 核,核带宽设置为 $\sigma=0.5$;在 LNP 方法中,最近邻参数 k 手工设置为 7,正则参数手工设置为 $\alpha=0.9$;在 SANFSP 方法中,正则参数 α 设置为与 LNP 相同.文献[15]指出,没有证据显示 $P \geq 3$ 时的 NFS 方法性能优于 $P=2$ 时的识别效果.因此,为计算简单起见,本文将参数 P 设置为 $P=2$.

3.1 Toy 数据实验

本节将采用一个二维人造数据集来显示所提方法的学习性能,以更好地理解所提方法在 SSL 问题上的标签传播过程.首先,人工生成一个包含 100 个样本的刻画两个信息类别的双月形二维数据集作为原始训练数据,其中,每个类包含 50 个样本对象,图 2(a)显示其中一个类(以“ Δ ”标识)中某个点 x (以“ \star ”标识)的特征空间嵌入投影点(以黑色“ $*$ ”标识).

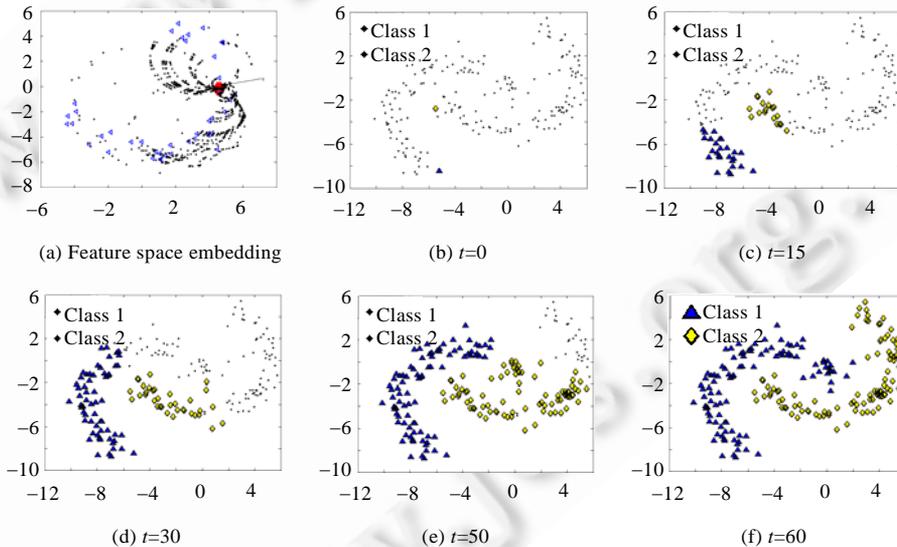


Fig.2 Label propagation results of SANFSP on the toy dataset

图 2 SANFSP 在人造数据集上标签传播结果

从图 2(a)可看出:相比于类中其他原始数据点,数据点 x 的特征空间嵌入投影点具有更丰富的表示信息,即,其在较大程度上充实了原始数据点的表示空间,从而具有更完备的判别信息,这显然能够有效增强模式识别性能.为了观测所提方法的标签传播效率,在上述人造数据集各类中分别标识 1 个标签数据点,如图 2(b)所示.图 2(c)~图 2(f)直观地显示了所提方法在双月数据集上的标签传播有效性,其中, t 为所提算法的迭代次数.

3.2 人脸识别实验

本实验将在 4 个实际人脸图像数据集上验证所提方法的鲁棒有效性.4 个人脸识别数据集包括 YALE B, CMU PIE, ORL, YALE(4 个数据集均可从网址 <http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html> 下载),

分别如图 3(a)~图 3(d)所示.扩展的 Yale B(YALE B)人脸数据库包含 38 个对象的 2 114 个正面人脸图像数据,每个数据点被裁减为 64×64 大小的灰度图像,每个对象类包括大约 60 幅图像,每幅图像由一个 1 024 维向量表示;PIE 人脸数据库包含 68 位志愿者的 41 368 张多姿态、光照和表情的面部图像,每个对象包含 170 幅人脸图像.实验中,对数据库中图像进行预处理,使其缩放到 32×32 像素大小,且每个像素为 256 灰度级,则在图像空间,每张图像由一个 1 024 维向量表示;ORL 数据库包含 40 个不同对象的 400 幅人脸图像,根据时间、灯照条件、面部表情、面部细节等不同特征,分别有 10 幅图像刻画每个对象,实验中每幅原始图像大小缩减为 40×40 像素;Yale 人脸数据库包括 15 张人脸的 165 个灰度级图像,这些图像分别显示了不同光照条件和脸部表情.实验中,对图像定位进行了预处理,使其缩放到 32×32 像素大小,且每个像素为 256 灰度级,则在图像空间,每张图像由一个 1 024 维向量表示.YALE 数据库包含 15 个不同对象的 165 幅人脸图像,根据不同的面部表情和面部配置,分别有 11 幅图像刻画每个对象,并将原始图像缩减为 32×32 像素大小.

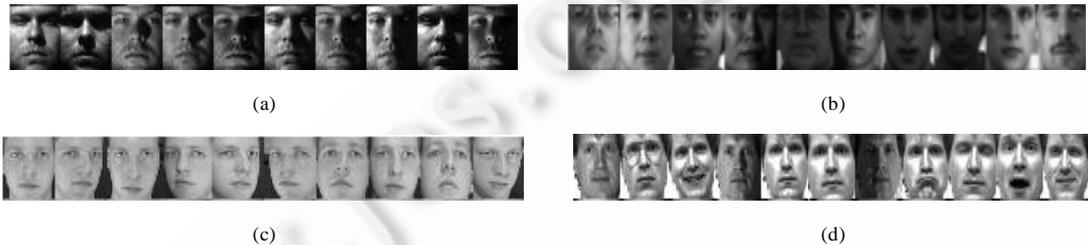


Fig.3 10 selected image samples in YALE B, PIE, ORL and YALE
图 3 从 YALE B,PIE,ORL 和 YALE 中分别选取的 10 副人脸样本

对于各人脸图像数据集 X ,随机选取 m 个数据样本作为标签化训练数据,对于 1-NN 算法,训练集仅包含 1 个标签数据.对于 LNP,GFHF,LapRLS,S-RLSC 和 SANFSP 方法,训练集包含数据集 X 中所有带标签和无标签数据.图 4(a)~图 4(d)分别显示了所提方法在 4 个人脸数据集上的识别性能比较结果,其中,水平轴表示每个对象随机标注的人脸图像数,垂直轴代表所有算法的相应的识别误差率.从图 5 显示的结果可以看出:所提方法 SANFSP 在所有 4 个人脸数据集上的识别性能均优于其他方法,其中,随着随机标注的人脸样本数增加,SANFSP,S-RLSC 和 LNP 这 3 种方法具有可比较的识别性能.

为了进一步评价所提方法在人脸识别上的鲁棒性,本文从 YALE 和 ORL 数据库中分别对每个对象随机选取 8 幅图像作为训练数据集,并在样本数据中先后加入随机噪声和遮罩信息:第 1 次在图像样本中逐渐增加高斯白噪声百分比,以逐渐减小信噪比(signal-to-noise ratio,简称 SNR),如图 5(a)所示;第 2 次不同大小的黑色方块随机覆盖在目标样本图像上,从而产生遮罩(或缺失数据)现象,如图 5(b)所示.对于大小为 32×32 像素的 YALE 图像,黑色方块大小分别为 6×6 , 10×10 , 18×18 和 22×22 像素;对于大小为 40×40 像素的 ORL 图像,黑色方块大小分别为 8×8 , 14×14 , 22×22 和 26×26 像素.图 6 和图 7 分别显示了上述两次处理数据集的实验结果,其中水平轴表示数据集中干扰信号的能量变化,垂直轴表示人脸识别精确率.由图 6 和图 7 可看出:在样本数据集包含缺损数据和噪声信息的情况下,本文方法 SANFSP 在绝大多数情况下均能取得最好的识别性能.这显示,基于稀疏近似最近特征空间嵌入相似度量的 SANFSP 算法是有效的.另外,随着噪声或遮罩信息的增加,所有算法的识别性能均表现出了不同程度的下降趋势,而本文所提方法 SANFSP 和 S-SRLC 性能下降相对缓慢,这也进一步说明了基于稀疏重构技术的图像识别方法对于噪声或缺失数据具有更强的鲁棒性.

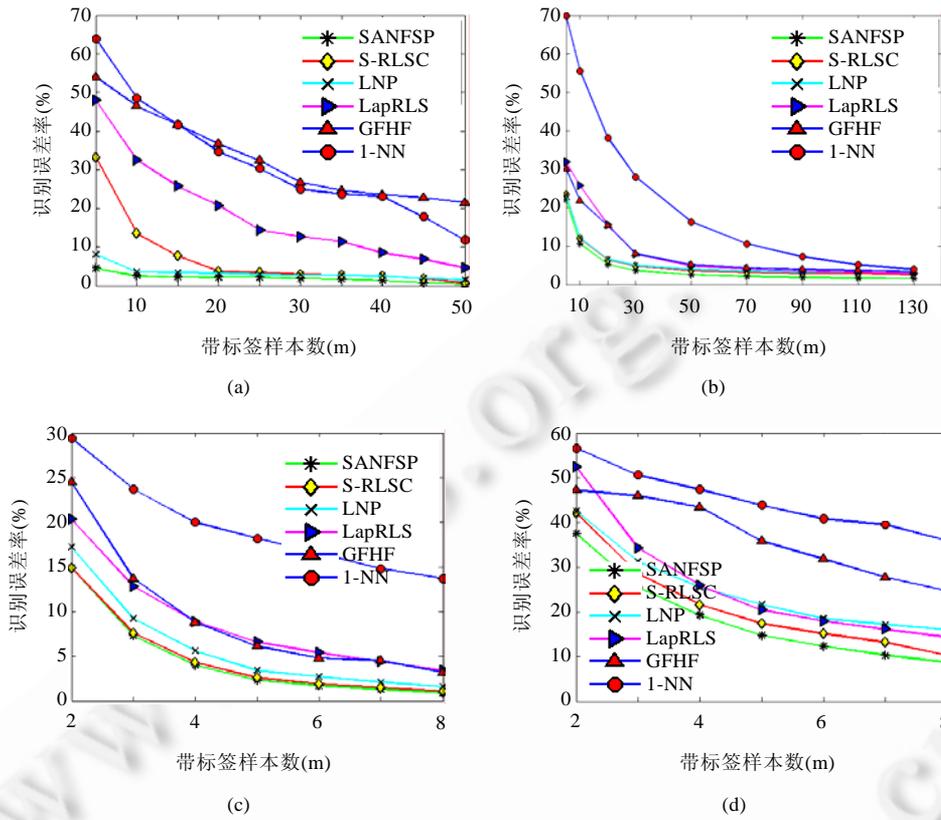


Fig.4 Face recognition results on YALE B, PIE, ORL and YALE

图 4 分别在 YALE B,PIE,ORL 和 YALE 上的人脸识别结果

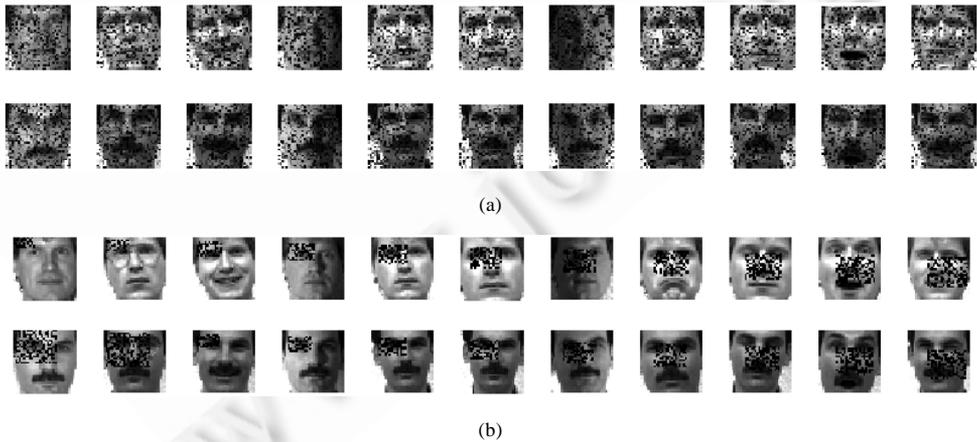


Fig.5 Face images in YALE with noise and occlusions

图 5 带有噪声和遮罩的 YALE 人脸

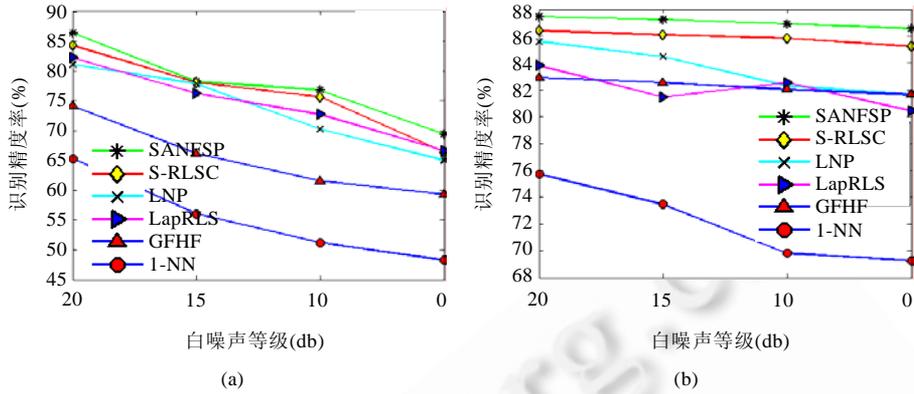


Fig.6 Recognition accuracies of with different levels of white noise on YALE and ORL
图 6 在具有不同能量等级噪声的 YALE 和 ORL 人脸数据集上的识别精度

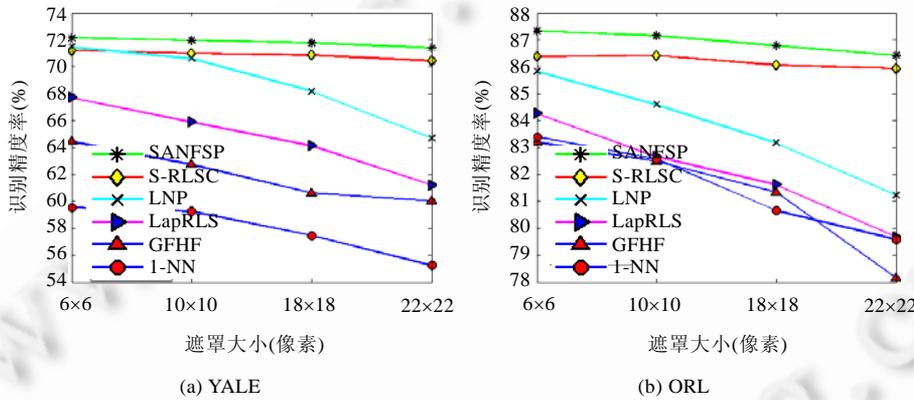


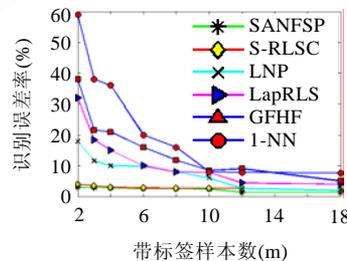
Fig.7 Recognition accuracies of all methods with different sizes of occlusion on YALE and ORL
图 7 在具有不同大小遮罩的 YALE 和 ORL 数据集上的平均识别精度

3.3 物件识别实验

本实验将所提方法 SANFSP 应用于可视物件识别任务,实验采用 Columbia 大学图像库(COIL 20)作为训练数据集^[9],该库包含 20 个类别物件的灰色图像集,对于每类对象,拥有 72 幅大小为 28×28 的图像,图 8(a)显示 20 个类别的物件.图 8(b)显示了所提方法在 COIL 20 上的识别性能比较,其中,水平轴表示随机选取的标签样本数,垂直轴代表物件识别误差率.



(a) COIL 20 数据集中 20 个物件样本



(b) 所有算法的识别结果

Fig.8 Object recognition

图 8 物件识别

从图 8(b)显示可看出,SANFSP 优于传统方法,并与 S-RLSC 具有可比较的性能.而随着标签样本的增

加,LNP,S-RLSC 和 SANFSP 三者识别性能趋于接近.

3.4 USPS数字识别实验

本实验将进一步评价所提方法在 USPS 数据集(<http://www.kernel-machines.org/data.html>)上的分类性能,该数据集包含 10 类大小为 16×16 的手写数字图像.本实验从该数据集中分别选取数字 1,2,3,4 的前 200 幅图像作为训练集,随后的 200 幅图像作为 Out-of-Sample 图像.图 9(a)显示了上述 4 个数字的 1 个类样本.实验中,记录 10 次独立实验结果的平均值作为各算法的识别精度率,图 9(b)描绘了所提方法与其他方法的性能比较.从图 9(b)显示的结果可明显看出所提方法的性能优势:随着随机标注样本的增加,LapRLS,LNP 和 S-RLSC 这 3 种方法具有可比较的稳定的识别性能.值得指出的是,随着随机标注样本的增加,所提方法 SANFSP 的识别性能非常稳定.换句话说,即使在标签样本较少的情况下,所提方法仍能取得较高的识别性能.可能的原因在于:一方面,不同类数字的图像数据处于各自的子流形结构,即同类数字的图像数据处于相同的子流形结构,而基于特征空间嵌入稀疏重构技术的 SANFSP 方法有效揭示了不同类数字的子流形结构,使得只需少量的标签样本即可预测无标签样本标签;另一方面,由于采用了特征空间嵌入技术,在一定程度上充实了数据空间的判别信息,从而只需少量标签数据便可判别测试数据.

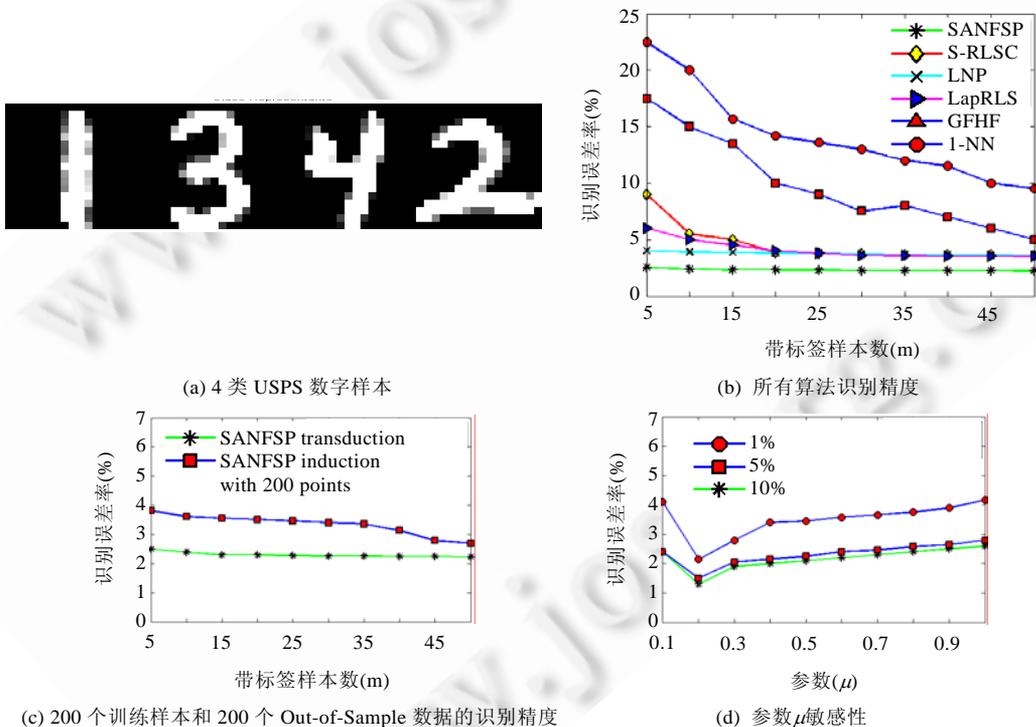


Fig.9 Digit recognition on the USPS data set

图 9 USPS 数字识别实验

本实验还评价了 SANFSP 算法的归纳学习能力:首先,在前 200 个数字图像上进行演绎(transduction)学习,其中,标签样本数随机变化;然后,利用公式(9)将习得的标签矩阵用于 Out-of-Sample 归纳学习.图 9(c)显示了 SANFSP 算法的演绎学习和 Out-of-Sample 归纳学习的结果,从该图可比较看出,SANFSP 算法依旧能够取得较高的归纳学习性能.图 9(d)显示了实验中参数 μ 变化对 SANFSP 算法性能的影响,其中,水平轴表示参数 μ 的变化区间,垂直轴表示 SANFSP 的相应识别精度.不同的曲线代表不同的标签样本数,如 5% 的曲线代表 SANFSP 在标签样本数占总训练样本数的 5% 的情况下的不同识别精度.由于参数 $\alpha = 1/(1+\mu)$,故图 9(d)也间接显示了参数 α 的敏感性.根据定理 4,正则参数 μ 起到平衡预测损失和平滑性的作用.从图 9(d)可知,参数 μ 过大或过小,SANFSP

均不能取得最优学习性能.因此,在训练 SANFSP 算法时应同时考虑预测损失和平滑性.

4 结束语

本文提出一种新颖的 GSSL 方法 SANFSP,即稀疏近似最近特征空间嵌入标签传播.基于特征空间嵌入投影点集,采用核稀疏表示技术来分别重构各原始数据,并以此构造一个稀疏图,以完成标签传播过程,从而实现最终的 SSL 任务.理论分析结果显示:基于稀疏近似最近特征空间嵌入正则化,SANFSP 所预测的数据标签具有充分的平滑性和鲁棒性.在人造和实际 SSL 任务上的实验结果验证了所提方法的有效性.对于特征空间嵌入方法,各原始数据点的特征空间嵌入投影点集的大小为 C_p^{N-1} ,则所有数据的特征空间嵌入点集的大小为 $N \times C_p^{N-1}$,为原始数据集大小的 C_p^{N-1} 倍,从而导致最近特征空间嵌入方法在大数据集上的计算复杂度较高.因此,如何有效减少最近特征空间嵌入方法的计算复杂度,是值得进一步研究的方向.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是本文的各位审稿专家表示衷心的感谢.

References:

- [1] Muller KR, Mika S, Ratsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. *IEEE Trans. on Neural Networks*, 2001, 12(2):181–201. [doi: 10.1109/72.914517]
- [2] Joachims T. Transductive inference for text classification using support vector machines. In: *Proc. of the 16th Int'l Conf. on Machine Learning (ICML)*. 1999.
- [3] Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs. In: *Proc. of the 17th Conf. on Learning Theory*. 2004. [doi: 10.1007/978-3-540-27819-1_43]
- [4] Zhu XJ. Semi-Supervised learning literature survey. *Computer Sciences Technical Report*, University of Wisconsin-Madison, 2005.
- [5] Liu W, Wang J, Chang SF. Robust and scalable graph-based semi-supervised learning. In: *Proc. of the IEEE*, 2012,100(9): 2624–2638. [doi: 10.1109/JPROC.2012.2197809]
- [6] Zhou D, Bousquet O, Lal T, Weston J, Schölkopf B. Learning with local and global consistency. In: *Proc. of the Advances in Neural Information Processing Systems*. 2004.
- [7] Zhu XJ, Ghahramani Z, Lafferty J. Semi-Supervised learning using Gaussian fields and harmonic functions. In: *Proc. of the 20th Int'l Conf. on Machine Learning*. 2003.
- [8] Wang F, Zhang C. Label propagation through linear neighborhoods. *IEEE Trans. on Knowledge and Data Engineering*, 2008,20(1): 55–67. [doi: 10.1109/TKDE.2007.190672]
- [9] Wang J, Wang F, Zhang C, Shen HC, Quan L. Linear neighborhood propagation and its applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009,31(9):1600–1615. [doi: 10.1109/TPAMI.2008.216]
- [10] Wu MR, Schölkopf B. Transductive classification via local learning regularization. In: *Proc. of the Int'l Conf. on Artificial Intelligence and Statistics*. Cambridge: MIT Press, 2007. 624–631.
- [11] Wang F, Li T, Wang G, Zhang C. Semi-Supervised classification using local and global regularization. In: *Proc. of the 23rd AAAI Conf. on Artificial Intelligence (AAAI)*. 2008.
- [12] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 2006,7:2399–2434.
- [13] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000,290:2323–2326. [doi: 10.1126/science.290.5500.2323]
- [14] He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using Laplacian faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005,27(3):328–340.
- [15] Chen YN, Han CC, Wang CT, Fan KC. Face recognition using nearest feature space embedding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2011,33(6):1073–1086. [doi: 10.1109/TPAMI.2010.197]
- [16] Wright J, Yang A, Sastry S, Ma Y. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009,31(2):210–227. [doi: 10.1109/TPAMI.2008.79]
- [17] Cheng H, Liu ZC, Yang J. Sparsity induced similarity measure for label propagation. In: *Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV)*. 2009. [doi: 10.1109/ICCV.2009.5459267]
- [18] Fan M, Gu N, Qiao H, Zhang B. Sparse regularization for semi-supervised classification. *Pattern Recognition*, 2011,44(8): 1777–1784. [doi: 10.1016/j.patcog.2011.02.013]

[19] Qiao L, Chen S, Tan X. Sparsity preserving projections with applications to face recognition. Pattern Recognition, 2010,43(1): 331–341. [doi: 10.1016/j.patcog.2009.05.005]
 [20] Zheng M, Bu J, Chen C, Wang C, Zhang LJ, Qiu G, Cai D. Graph regularized sparse coding for image representation. IEEE Trans. on Image Processing, 2011. [doi: 10.1109/TIP.2010.2090535]
 [21] Nie F, Zeng Z, Tsang IW, Xu D, Zhang C. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. IEEE Trans. on Neural Networks, 2011,22(11):1796–1808. [doi: 10.1109/TNN.2011.2162000]
 [22] Delalleu O, Bengio Y, Le Roux N. Efficient non-parametric function induction in semi-supervised learning. In: Proc. of the 10th Int'l Workshop on Artificial Intelligence and Statistics. 2005.
 [23] Zhang H, Berg AC, Maire M, Malik J. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proc. of the CVPR, Vol.2. 2006. 2126–2136. [doi: 10.1109/CVPR.2006.301]

附录 1. 定理 3 证明

证明:下面分几种情况来分别证明结论的正确性.

(1) 当 $P=1$ 时,根据文献[15]结论可知,目标函数(12)和公式(13)能表示为 Laplacian 矩阵形式.

(2) 当 $P=2$ 时,公式(12)和公式(13)变成点到各标签线间相似性度量,从标签线 $\overline{f_a f_b}$ 到 f_i 的距离计算为 $\|f_i - L_{a,b}^{(2)}(f_i)\|$,投影点可计算为 $L_{a,b}^{(2)}(f_i) = f_a + \lambda_{a,b}(f_a - f_b)$,且满足 $\lambda_{a,b} + \lambda_{b,a} = 1 (i \neq a \neq b)$,从而有:

$$\begin{aligned}
 F_1 &= \sum_i \left\| \sum_{a \neq b} (f_i - L_{a,b}^{(2)}(f_i)) s_{a,b}^{(2)}(x_i) \right\|^2 \\
 &= \sum_i \left\| \sum_{a \neq b} (f_i - \lambda_{b,a} f_a - \lambda_{a,b} f_b) s_{a,b}^{(2)}(x_i) \right\|^2 \\
 &= \sum_i \left\| f_i - \sum_j M_{i,j} f_j \right\|^2 \\
 &= \text{tr}(F(I - M)^T (I - M) F^T) \\
 &= \text{tr}(F(D - W) F^T) \\
 &= \text{tr}(FLF^T)
 \end{aligned} \tag{25}$$

其中, $M_{i,b} = \sum_a s_{a,b}^{(2)}(x_i) q_{a,b}$, $q_{a,b} = (x_i - x_a)^T (x_a - x_b) / ((x_a - x_b)^T (x_a - x_b))$ 且 $\sum_j M_{i,j} = 1$. 根据文献[15]的结果,矩阵 W 设置为:当 $i \neq j$ 时, $W_{i,j} = (M + M^T - M^T M)_{i,j}$; 否则, $W_{i,j} = 0$. 这样,目标函数(12)可表示为 Laplacian 矩阵形式.

另一方面,目标函数(20)首先分解为 C_2^{N-1} 个成分,每个成分代表各标签对象 f_i 到第 $k (1 \leq k \leq C_2^{N-1})$ 个稀疏重构标签线的距离平方和,考虑第 1 项,矩阵 $M_{i,j}(1)$ 表示 x_i 到特征线 $\overline{x_a x_b} (i, a, b = 1, 2, \dots, N \text{ 且 } i \neq a \neq b)$ 的连接关系矩阵. 两个非零项 $M_{i,b}(1)$ 和 $M_{i,a}(1)$ 在矩阵 $M_{i,j}(1)$ 的各行均存在,且满足 $\sum_j M_{i,j}(1) = 1$. 一般来说,如果特征线 $\overline{x_a x_b}$ 为点 x_i 的稀疏重构对象,即 $s_{a,b}^{(2)}(x_i) > 0$, 则 $M_{i,b}(k) = s_{a,b}^{(2)}(x_i) q_{a,b}$, $M_{i,a}(k) = s_{a,b}^{(2)}(x_i) q_{b,a} (i \neq a \neq b)$; 否则, $M_{i,b}(k) = M_{i,a}(k) = 0$. 上述 C_2^{N-1} 个成分均可分别表示为 Laplacian 矩阵形式,即

$$\begin{aligned}
 F_2 &= \sum_i \sum_{a \neq b} \|f_i - L_{a,b}^{(2)}(f_i)\|^2 s_{a,b}^{(2)}(f_i) \\
 &= \sum_i \sum_{a \neq b} \|f_i - \lambda_{b,a} f_a - \lambda_{a,b} f_b\|^2 s_{a,b}^{(2)}(f_i) \\
 &= \sum_i \left\| f_i - \sum_j M_{i,j}(1) f_j \right\|^2 + \sum_i \left\| f_i - \sum_j M_{i,j}(2) f_j \right\|^2 + \dots + \sum_i \left\| f_i - \sum_j M_{i,j}(C_2^{N-1}) f_j \right\|^2 \\
 &= \text{tr}(F(I - M(1))^T (I - M(1)) F^T) + \text{tr}(F(I - M(2))^T (I - M(2)) F^T) + \dots + \text{tr}(F(I - M(C_2^{N-1}))^T (I - M(C_2^{N-1})) F^T) \tag{26} \\
 &= \text{tr}(F(D(1) - W(1)) F^T) + \text{tr}(F(D(2) - W(2)) F^T) + \dots + \text{tr}(F(D(C_2^{N-1}) - W(C_2^{N-1})) F^T) \\
 &= \text{tr}(F(D - W) F^T) \\
 &= \text{tr}(FLF^T)
 \end{aligned}$$

其中, $W_{i,j}(k) = (M(k) + M(k)^T - M(k)^T M(k))_{i,j}$, $D = \frac{1}{C_2^{N-1}} (D(1) + D(2) + \dots + D(C_2^{N-1}))$.

从而, 目标函数(13)也能表示为 Laplacian 矩阵形式.

(3) 当 $P=3$ 时, 公式(12)、公式(13)表示对象 f_i 与标签平面的距离, 即 $f_i - L_{r,a,b}^{(3)}(f_i)$, 其中, 标签平面由对象 f_r, f_a 和 f_b 张成. 标签平面的嵌入(或投影)对象 $L_{r,a,b}^{(3)}(f_i)$ 由下式计算获得:

$$\begin{aligned} L_{r,a,b}^{(3)}(f_i) &= F_{r,a,b} (F_{r,a,b}^T F_{r,a,b})^{-1} F_{r,a,b}^T (f_i - f_r) + f_r \\ &= [(f_a - f_r) \quad (f_b - f_r)] [\lambda_a \quad \lambda_b]^T + f_r \\ &= (1 - \lambda_a - \lambda_b) f_r + \lambda_a f_a + \lambda_b f_b \\ &= \lambda_r f_r + \lambda_a f_a + \lambda_b f_b \end{aligned} \tag{27}$$

其中, $F_{r,a,b} = [(f_a - f_r) \quad (f_b - f_r)]$ 为一个 $r \times 2$ 矩阵, 矩阵 $(F_{r,a,b}^T F_{r,a,b})^{-1} F_{r,a,b}^T (f_i - f_r)$ 表示为 $[\lambda_a \quad \lambda_b]^T$, 且 $\lambda_r + \lambda_a + \lambda_b = 1$.

从而, 投影点 $L_{r,a,b}^{(3)}(f_i)$ 被表示为对象 f_r, f_a 和 f_b 的线性组合. 与 $P=2$ 情况相似, 公式(12)和公式(13)分别表示为 $\sum_i \|f_i - \sum_j M_{i,j} f_j\|^2$, 其中, 矩阵 M 中权值 $(M_{i,r}, M_{i,a}, M_{i,b})$ 赋值为

$$\begin{aligned} (M_{i,r}, M_{i,a}, M_{i,b}) &= (q_r s_{r,a,b}^{(3)}(x_i), q_a s_{r,a,b}^{(3)}(x_i), q_b s_{r,a,b}^{(3)}(x_i)), \\ [q_a \quad q_b]^T &= (X_{r,a,b}^T X_{r,a,b})^{-1} X_{r,a,b}^T (x_i - x_r), \\ X_{r,a,b} &= [(x_a - x_r) \quad (x_b - x_r)], \\ q_r + q_a + q_b &= 1. \end{aligned}$$

从而, 与 $P=2$ 情况相似, 推导可分别导出公式(12)和公式(13)的 Laplacian 矩阵表示形式.

(4) 一般地, 对于 $P>3$ 情况, 公式(12)和公式(13)表示对象 f_i 与标签空间的距离, 即 $f_i - L_{1:P}^{(P)}(f_i)$, 对象 f_i 在标签空间的投影点 $L_{1:P}^{(P)}(f_i)$ 表示为 P 个对象 f_1, f_2, \dots, f_P 的线性组合, 即:

$$L_{1:P}^{(P)}(f_i) = F_{(1:P)} (F_{(1:P)}^T F_{(1:P)})^{-1} F_{(1:P)}^T (f_i - f_1) + f_1 = \sum_{j=1}^P \lambda_j f_j \tag{28}$$

其中, $F_{(1:P)} = [(f_2 - f_1) \quad (f_3 - f_1) \quad \dots \quad (f_P - f_1)]$ 为 $c \times (P-1)$ 矩阵, 且 $\sum_{j=1}^P \lambda_j = 1$. 对象 f_i 与标签空间投影点的距离表示为

$$\sum_i \|f_i - L_{1:P}^{(P)}(f_i)\|^2 s_{1:P}^{(P)} = \sum_i \|f_i - \sum_j M_{i,j} f_j\|^2 \tag{29}$$

从而, 目标函数(12)和公式(13)亦可表示为 Laplacian 矩阵形式.

综上, 对于 $P \geq 1$, 目标函数(12)、公式(13)可表示为 Laplacian 矩阵形式. □



陶剑文(1973-), 男, 湖北武汉人, 博士, 教授, 主要研究领域为模式识别, 数据挖掘.
E-mail: jianwen_tao@aliyun.com



王士同(1964-), 男, 教授, 博士生导师, 主要研究领域为人工智能, 机器学习.
E-mail: wx_stwang@yahoo.com.cn



Fu-Lai CHUNG(1965-), 男, 博士, 副教授, 博士生导师, 主要研究领域为人工智能, 模糊神经网络.



姚奇富(1965-), 男, 教授, 博士生导师, 主要研究领域为数据挖掘.
E-mail: yqf@zjbti.net.cn