特征迁移中投影子空间的分析与构造

程朝阳, 明 杨†, 洪奕光

(1. 中国科学院 数学与系统科学研究院 系统控制重点实验室, 北京 100190; 2. 中国科学院大学, 北京 100190)

摘要: 近年来, 针对实际应用场景中可匹配的训练数据不足的问题, 科研人员发展出了迁移学习的概念, 希望通过提取源域数据的特征信息进行迁移, 从而提升目标域的学习效果. 本文根据迁移学习所处理的不同数据类型, 构造了两种典型的模型: 单类别投影基构造模型与监督多类别投影模型. 由于子空间投影可以在一定程度上反映原始样本空间的特征性质. 因此, 本文应用线性判别分析的技巧以及最大均值差异的思想, 分别构造了上述模型的求解算法并对相应的非线性核方法进行了推广.

关键词: 迁移学习; 子空间投影; 线性判别分析; 最大均值差异

引用格式: 程朝阳, 明杨, 洪奕光. 特征迁移中投影子空间的分析与构造. 控制理论与应用, 2019, 36(11): 1834 – 1843

DOI: 10.7641/CTA.2019.90478

Subspace projection techniques in feature transfer learning

CHENG Zhao-yang, MING Yang†, HONG Yi-guang

(1. Key Lab of Systems and Control, Academy of Mathematics and Systems Science,
Chinese Academy of Science, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In recent years, researchers have developed the concept of transfer learning for lack of effective training data in practice, hoping to improve the performance in target domain by learning the feature information in source domain. According to different types of data in transfer learning, we build two typical models in this paper, which is referred to as one-class projective base construction and supervised multi-class projection. Since the essential properties of the raw data space can be characterized by taking projection in a proper subspace, we study the above models using techniques from linear discriminant analysis (LDA) and maximum mean discrepancy (MMD). Also, we extend those methods to the corresponding nonlinear kernel cases.

Key words: transfer learning; subspace projection; linear discriminant analysis; maximum mean discrepancy

Citation: CHENG Zhaoyang, MING Yang, HONG Yiguang. Subspace projection techniques in feature transfer learning. *Control Theory & Applications*, 2019, 36(11): 1834 – 1843

1 引言

目前, 机器学习及其相关技术已广泛的应用于图像处理[1]、目标检测[2]、自然语言处理[3]等各类科研领域, 并取得了丰富成果. 其中监督学习在很多实际场景的解决方案中发挥了关键作用. 然而在某些特定问题中, 由于隐私保护、收集成本、正负样本均匀性等因素的制约, 往往难以获取足够数量的有效数据, 所以采用通常的监督学习模型难以训练达到较好的效果. 针对这种情况, 科研人员在现有算法的基础上, 发展出了迁移学习[4–5]的概念.

在迁移学习中, 数据集一般被分为源域和目标域. 通过对源域内相关的数据信息进行迁移, 从而提升目标域的学习效果. 根据迁移的具体内容, 迁移学习一般可以分为“基于实例的迁移学习[6]”、“基于特征的迁移学习[7]”、“基于参数(模型)的迁移学习[8]”和“基于相关知识的迁移学习[9]”. 在本文中, 笔者主要研究基于子空间投影的特征迁移学习. 事实上, 在传统的机器学习方法中, 子空间投影是一种常用的降维技术. 当处理高维数据时, 如果能构造合适的投影, 就可以在不损失有效信息的同时降低计算量. 例如,

收稿日期: 2019−06−25; 录用日期: 2019−11−14.

†通信作者. E-mail: mingyang15@mails.ucas.ac.cn.

本文责任编委: 柯良军.

国家自然科学基金项目(61164015, 61305132), 江西省自然科学基金项目(20151BAB207043)资助.

Supported by the National Natural Science Foundation of China (61164015, 61305132) and the National Natural Science Foundation of Jiangxi Priovince (20151BAB207043).

常用的主成分分析(principal components analysis, PCA)[10]即可用来构造无监督型数据的子空间投影. 在此基础上对于监督型的降维问题, 科研人员提出了线性判别分析(linear discriminant analysis, LDA)方法. 同时为了处理线性不可分问题, 上述方法均产生了核化版本KPCA[11], KLDA[12].

因为子空间投影可以一定程度上表征原样本空间的性质[13–14], 所以本文基于以上子空间投影方法和最大均值差异思想(maximum mean discrepancy, MMD), 研究子空间投影在迁移学习中的应用. 在相关领域的研究中, Pan等人[15]基于迁移学习和主成分分析提出了迁移成分分析(transfer component analysis, TCA)方法, 分析了迁移学习的降维问题. Lu等人[16]根据线性判别分析和神经网络的思想提出了LDADA方法, 实现有选择迭代的迁移学习算法. 而Long等人[17]运用核方法和最大均值差异思想总结出了迁移学习的框架模型ARTL.

本文的主要成果在于, 首先根据迁移学习源域的数据类型和目标域的标签特性, 构造了两种不同的模型: 单类别投影基构造模型与有监督多类别投影模型. 其次, 在线性判别分析和迁移学习等方法的基础上, 应用最大均值差异的思想对原目标函数进行转化, 构造了以上模型的求解方法. 进一步针对源域或目标域线性不可分样本, 采用核技巧构造出了上述模型相应的核化版本. 最后通过模拟实验, 对两种模型的求解方法进行验证, 分析对比了相应的实验效果.

本文其余部分结构如下: 第2部分介绍了模型构建以及解决方案中涉及的基础知识, 包括迁移学习的基本框架、子空间投影方法及核方法等; 第3部分描述了我们所考虑的特征迁移的问题背景和模型中涉及的数据集的基本类型; 第4部分根据上述模型中优化目标的结构特点, 提出了相应的解决方案和算法, 进一步引入核方法解决源域或目标域数据线性不可分的情况; 第5部分通过实验仿真, 模拟实现以上算法, 并分析实验效果; 第6部分回顾总结了本文的工作与结果, 分析了实验中遇到的困难及其原因, 并指出了未来工作方向.

2　相关研究

本节介绍了一些背景知识和相关研究, 包含迁移学习的符号与定义、最大均值差异、子空间投影的线性判别分析与主成分分析等.

2.1　迁移学习

一般机器学习的两个要素是域(domain)和任务(task). 而域又由两个要素组成: 特征样本空间$\mathcal{X}$和边缘概率分布$P(X)$, 其中$X = \{x_1, \cdots, x_n\} \in \mathcal{X}$. 因此, 一般域可以表示为$\mathcal{D} = \{\mathcal{X}, P(X)\}$. 另外, 给定一个域, 对应的任务也由两个要素组成: 标签空间$\mathcal{Y}$和目标预测函数$f(\cdot)$. 因此, 任务也可以表示为$\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$.

在处理机器学习的问题时, 可以在训练样本中找出一组$\{x_i, y_i\}$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, 通过这组训练样本, 可以估计目标预测函数$f(\cdot)$, 目标预测函数也可以表示为$P(y|x)$.

迁移学习中涉及两个域, 一个是源域$\mathcal{D}_{\rm S}$, 另外一个是目标域$\mathcal{D}_{\rm T}$. 例如, 如果源域有标签时, 可以定义源域

$$\mathcal{D}_{\rm S} = \{(x_{{\rm S}_1}, y_{{\rm S}_1}), \cdots, (x_{{\rm S}_{n_{\rm S}}}, y_{{\rm S}_{n_{\rm S}}})\},$$

其中: $x_i \in \mathcal{X}_{\rm S}$为源域中的样本; $y_i \in \mathcal{Y}_{\rm S}$为该样本对应的标签. 当目标域有标签时, 可以定义目标域

$$\mathcal{D}_{\rm T} = \{(x_{{\rm T}_1}, y_{{\rm T}_1}), \cdots, (x_{{\rm T}_{n_{\rm T}}}, y_{{\rm T}_{n_{\rm T}}})\},$$

其中: $x_i \in \mathcal{X}_{\rm T}$为目标域中的样本; $y_i \in \mathcal{Y}_{\rm T}$为该样本对应的标签. 一般情况下, 源域样本数量远大于目标域样本数量, 即$0 \leqslant n_{\rm T} \leqslant n_{\rm S}$.

文献[4]中迁移学习的定义如下:

定义 1(迁移学习)　给定一个源域$\mathcal{D}_{\rm S}$和源域上的任务$\mathcal{T}_{\rm S}$, 以及一个目标域$\mathcal{D}_{\rm T}$以及目标域上的任务$\mathcal{T}_{\rm T}$. 迁移学习是为了提升在目标域上估计目标预测函数$f_{\rm T}(\cdot)$的精度, 而使用源域$\mathcal{D}_{\rm S}$和源域上的任务$\mathcal{T}_{\rm S}$的信息的学习方法, 其中$\mathcal{D}_{\rm S} \neq \mathcal{D}_{\rm T}$ 或$\mathcal{T}_{\rm S} \neq \mathcal{T}_{\rm T}$.

上述定义强调了迁移学习的源域和目标域的样本和任务至少有一个不相同. 其中样本不相同可能是样本域不同, 即$\mathcal{X}_{\rm S} \neq \mathcal{X}_{\rm T}$, 另外也可能是样本域上的边缘分布不同, 即$P_{\rm S}(X) \neq P_{\rm T}(X)$. 同样的, 任务不同可能是任务域不同, 即$\mathcal{Y}_{\rm S} \neq \mathcal{Y}_{\rm T}$, 或者任务域基于样本的预测函数或者条件分布不同, 即$f_{\rm S}(\cdot) \neq f_{\rm T}(\cdot)$或$P(Y_{\rm S}|X_{\rm S}) \neq P(Y_{\rm T}|X_{\rm T})$. 当然迁移学习的源域和目标域必须存在一定联系, 才可以将源域中的信息迁移到目标域之中.

对于迁移学习的源域和目标域上两个不同的分布, 文献[18–19]通过最大均值差异来衡量两个域中数据分布的距离. 基于两个分布的样本, 通过寻找在样本空间上的连续函数f, 求两分布的样本在f上函数值的均值, 将两个均值作差, 即可得到两个分布对应f的均值差异. 以最大均值差异作为检验统计量, 从而判断两个分布是否相同, 最大均值差异是目前比较常用的比较源域和目标域分布差异的方法.

$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (E_{x \sim p}[f(x)] - E_{y \sim q}[f(y)]). \tag{1}$$

2.2　子空间投影

在机器学习数据的预处理中, 子空间投影是一种有效的降维技术, 常用于降低处理高维数据时的计算量并且进一步估计原样本空间的性质. 主成分分析[10]和线性判别分析是两个常用的线性降维方法.

主成分分析是一种无监督线性降维方法, 其主要任务是找出数据的最主要特征成分, 以找到原数据的一个低维表征. 数据集为

$$X = \{x_1, \cdots, x_n\} \in \mathbb{R}^{m\times n},\ x_i \in \mathbb{R}^m,\ \forall i = 1, \cdots, n.$$

笔者希望将这n个数据从m维降到m'维, 并且这n个m'维的数据尽可能代表原始数据集, 即找到一个映射(在这里为基变换矩阵P), 作用在X上得到$Y = PX$. 然后投影之后的数据Y可以一定程度上表征原样本性质, 并在此基础上进行分析学习.

除了主成分分析之外, 线性判别分析也是一种常见的降维方法, 只是线性判别分析是监督线性降维方法, 处理有标签的数据. 线性判别分析针对一组有类别标签的样本, 希望找到样本数据的一个投影空间, 使得样本在这个空间中的投影对不同类间数据的差异较为敏感, 而对同一类别内数据之间的差异不敏感, 即希望不同类别样本在这个投影空间上投影的距离尽可能远, 而同一类别的样本在这个投影空间中投影的距离尽可能近, 以便于对新的样本进行分类. 由此得到目标函数:

$$\max_W\ J(W) = \max_W\ \frac{\mathrm{tr}(W^{\mathrm{T}} S_{\mathrm{b}} W)}{\mathrm{tr}(W^{\mathrm{T}} S_{\mathrm{w}} W)}. \tag{2}$$

该优化目标的解W取$S_{\mathrm{w}}^{-1}S_{\mathrm{b}}$的前$d$个最大特征值对应的特征向量, 其中$d$为低维投影空间的维度.

然而线性判别分析和主成分分析处理样本均为线性可分的, 对于线性不可分的样本, 通常采用核方法进行分析. 核方法是希望找到一个高维空间, 使得样本投影到这个高维空间后得到的数据线性可分, 在该高维空间中进行分析, 进而解决低维空间中数据线性不可分的问题.

核方法首先假设存在一个高维希尔伯特空间$\mathcal{H}$和一个从样本域$\mathcal{X}$到$\mathcal{H}$的映射ϕ, 使得样本域经过映射ϕ后在空间$\mathcal{H}$线性可分. 即ϕ将x_i映射为$\phi(x_i)$, 在该高维希尔伯特空间$\mathcal{H}$中x_i, x_j映射后的内积为$\phi(x_i)^{\mathrm{T}}\phi(x_j)$. 定义核函数

$$k(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j). \tag{3}$$

则通过核函数, 可以避免计算高维空间的内积, 降低了维度复杂度和投影选取带来的对计算方面的影响.

同时, 科研人员将核方法应用到主成分分析和线性判别分析中也得到了KPCA和KLDA方法, 以处理样本数据线性不可分的降维问题.

3 子空间特征迁移模型

本文假设源域样本和目标域样本在同一个域中, 只是分布不同, 即$\mathcal{X}_{\mathrm{S}} = \mathcal{X}_{\mathrm{T}}$, $P_{\mathrm{S}}(X) \neq P_{\mathrm{T}}(X)$, 且$\mathcal{X}_{\mathrm{S}} = \mathcal{X}_{\mathrm{T}} = \mathbb{R}^n$. 同时, 假设存在一个投影空间$\mathcal{Z}$和一个到该投影空间的映射$\varphi$,

$$\begin{aligned}\varphi:\ & \mathcal{X} \to \mathcal{Z},\\ & X \to W^{\mathrm{T}}X,\end{aligned} \tag{4}$$

使得源域和目标域在该投影空间上分布相同, 即$P_{\mathrm{S}}(\varphi(X)) = P_{\mathrm{T}}(\varphi(X))$. 假设在该投影空间上, 源域和目标域基于样本的预测函数是相同的, 即$P_{\mathrm{S}}(Y|\varphi(X)) = P_{\mathrm{T}}(Y|\varphi(X))$.

本文进行迁移学习的目的是尽可能找出该投影空间, 并对目标域的预测更精确. 本文针对目标域和源域不同数据类型, 提出了两种模型, 如下所示:

1) 模型1.

设源域数据集

$$D_{\mathrm{S}} = \{(x_{\mathrm{S}_1}, y_{\mathrm{S}_1}), \cdots, (x_{\mathrm{S}_{n_{\mathrm{S}}}}, y_{\mathrm{S}_{n_{\mathrm{S}}}})\},$$

其中: $x_{\mathrm{S}_i} \in \mathcal{X}_{\mathrm{S}}$, $\forall i{=}1, \cdots, n_{\mathrm{S}}$, y_i为x_i的标签, 可以是离散的也可以是连续的. 目标域数据集

$$D_{\mathrm{T}} = \{x_{\mathrm{T}_1}, \cdots, x_{\mathrm{T}_{n_{\mathrm{T}}}}\}.$$

目标域数据没有标签, $x_{\mathrm{T}_i} \in \mathcal{X}_{\mathrm{T}}$, $\forall i = 1, \cdots, n_{\mathrm{T}}$.

2) 模型2.

考虑源域$\mathcal{D}_{\mathrm{S}}$内有标签的数据

$$\mathcal{D}_{\mathrm{S}} = \{(x_{\mathrm{S}_1}, y_{\mathrm{S}_1}), \cdots, (x_{\mathrm{S}_{n_{\mathrm{S}}}}, y_{\mathrm{S}_{n_{\mathrm{S}}}})\},$$

其中: $x_{\mathrm{S}_i} \in \mathcal{X}_{\mathrm{S}}$, $\forall i{=}1, \cdots, n_{\mathrm{S}}$, y_i为x_i的标签, 代表的是x_i的类别. 目标域$\mathcal{D}_{\mathrm{T}}$内的数据部分有标签, 部分无标签.

$$\begin{aligned}\mathcal{D}_{\mathrm{T}} = \{&(x_{\mathrm{T}_1}, y_{\mathrm{T}_1}), \cdots, (x_{\mathrm{T}_{n_{\mathrm{T}_1}}}, y_{\mathrm{T}_{n_{\mathrm{T}_1}}}),\\ &x_{\mathrm{T}_{n_{\mathrm{T}_1}+1}}, \cdots, x_{\mathrm{T}_{n_{\mathrm{T}_1}+n_{\mathrm{T}_2}}}\},\end{aligned}$$

$x_{\mathrm{T}_i} \in \mathcal{X}_{\mathrm{T}}$, $\forall i = 1, \cdots n_{\mathrm{T}_1} + n_{\mathrm{T}_2}$. 而目标域中的标签也代表类别. 方便起见, 令

$$\begin{aligned}\mathcal{D}_{\mathrm{T}_1} &= \{(x_{\mathrm{T}_1}, y_{\mathrm{T}_1}), \cdots, (x_{\mathrm{T}_{n_{\mathrm{T}_1}}}, y_{\mathrm{T}_{n_{\mathrm{T}_1}}})\},\\ \mathcal{D}_{\mathrm{T}_2} &= \{x_{\mathrm{T}_{n_{\mathrm{T}_1}+1}}, \cdots, x_{\mathrm{T}_{n_{\mathrm{T}_1}+n_{\mathrm{T}_2}}}\}.\end{aligned}$$

假设源域与目标域类别数目相同, 类别个数均为K.

4 子空间投影基的构造

该部分利用线性判别分析, 最大均值差异等方法的思想, 对上一部分提出的两种模型进行分析, 得到最终的优化目标和实现算法.

4.1 模型 1: 单类别投影基的构造

源域数据集

$$D_{\mathrm{S}} = \{(x_{\mathrm{S}_1}, y_{\mathrm{S}_1}), \cdots, (x_{\mathrm{S}_{n_{\mathrm{S}}}}, y_{\mathrm{S}_{n_{\mathrm{S}}}})\},$$

其中: $x_{\mathrm{S}_i} \in \mathbb{R}^m$, $\forall i{=}1, \cdots, n_{\mathrm{S}}$, y_i为x_i的标签. 目标域数据集

$$D_{\mathrm{T}} = \{x_{\mathrm{T}_1}, \cdots, x_{\mathrm{T}_{n_{\mathrm{T}}}}\}.$$

目标域数据没有标签, $x_{\mathrm{T}_i} \in \mathbb{R}^m$, $\forall i = 1, \cdots, n_{\mathrm{T}}$.

在进行分析之前, 先给出一些定义. 定义源域样本矩阵$X_{\mathrm{S}} = [x_{\mathrm{S}_1}\ \cdots\ x_{\mathrm{S}_{n_{\mathrm{S}}}}]$和目标域样本矩阵$X_{\mathrm{T}} =$

$[x_{\mathrm{T}_1}\ \cdots\ x_{\mathrm{T}_{n_\mathrm{T}}}]$. 定义

$$\mu_\mathrm{S}=\frac{1}{n_\mathrm{S}}\sum_{i=1}^{n_\mathrm{S}}x_{\mathrm{S}_i},\ \mu_\mathrm{T}=\frac{1}{n_\mathrm{T}}\sum_{i=1}^{n_\mathrm{T}}x_{\mathrm{T}_i},$$

分别为源域和目标域的中心; 定义

$$\begin{aligned}\Sigma_{W_\mathrm{S}}&=\sum_{i=1}^{n_\mathrm{S}}(x_{\mathrm{S}_i}-\mu_\mathrm{S})(x_{\mathrm{S}_i}-\mu_\mathrm{S})^\mathrm{T},\\ \Sigma_{W_\mathrm{T}}&=\sum_{i=1}^{n_\mathrm{T}}(x_{\mathrm{T}_i}-\mu_\mathrm{T})(x_{\mathrm{T}_i}-\mu_\mathrm{T})^\mathrm{T},\end{aligned}\tag{5}$$

分别为目标域和源域的全局散度矩阵.

由于假设存在一个投影空间$\mathcal{Z}$和一个到该投影空间的映射φ,

$$\begin{aligned}\varphi:\ &\mathcal{X}\to\mathcal{Z},\\ &X\to W^\mathrm{T}X,\end{aligned}\tag{6}$$

使得源域和目标域在该投影空间上分布相同, 即$P_\mathrm{S}(\varphi(X))=P_\mathrm{T}(\varphi(X))$. 假设在该投影空间上, 源域和目标域基于样本的预测函数是相同的, 即$P_\mathrm{S}(Y|\varphi(X))=P_\mathrm{T}(Y|\varphi(X))$.

设$W=[w_1\ w_2\ \cdots\ w_d]$为由$\mathcal{X}$到空间$\mathcal{Z}$的投影矩阵, 其中: $w_i\in\mathbb{R}^m$, $\forall i=1,2,\cdots,d$, 为该投影空间各个基的向量的方向. 接下来则是求出该投影矩阵, 使得源域和目标域再该投影空间上分布尽可能相同, 并且尽可能表征源域和目标域相同的性质.

现在以$d=1$为例, 本文的目标是: 首先, 源域和目标域二者的中心在这个投影上的距离尽可能近, 即$(w_1^\mathrm{T}\mu_\mathrm{S}-w_1^\mathrm{T}\mu_\mathrm{T})^2$尽可能小. 其次, 笔者希望源域和和目标域各自在这个投影上内部之间各自的距离尽可能大, 以各个域内各个样本到中心点的距离来衡量, 即令

$$\sum_{i=1}^{n_\mathrm{S}}(w_1^\mathrm{T}x_{\mathrm{S}_i}-w_1^\mathrm{T}\mu_\mathrm{S})^2,\ \sum_{i=1}^{n_\mathrm{T}}(w_1^\mathrm{T}x_{\mathrm{T}_i}-w_1^\mathrm{T}\mu_\mathrm{T})^2$$

都尽可能大.

首先, 分析第1个目标:

$$\begin{aligned}&(w_1^\mathrm{T}\mu_\mathrm{S}-w_1^\mathrm{T}\mu_\mathrm{T})^2=\\ &(w_1^\mathrm{T}\mu_\mathrm{S}-w_1^\mathrm{T}\mu_\mathrm{T})^\mathrm{T}(w_1^\mathrm{T}\mu_\mathrm{S}-w_1^\mathrm{T}\mu_\mathrm{T})=\\ &w_1^\mathrm{T}(\mu_\mathrm{S}-\mu_\mathrm{T})(\mu_\mathrm{S}-\mu_\mathrm{T})^\mathrm{T}w_1.\end{aligned}\tag{7}$$

令$\Sigma_\mathrm{b}=(\mu_\mathrm{S}-\mu_\mathrm{T})(\mu_\mathrm{S}-\mu_\mathrm{T})^\mathrm{T}$, 可得

$$w_1^\mathrm{T}\Sigma_\mathrm{b}w_1=(w_1^\mathrm{T}\mu_\mathrm{S}-w_1^\mathrm{T}\mu_\mathrm{T})^2,$$

于是本文的目标是

$$\min_{w_1}\ w_1^\mathrm{T}\Sigma_\mathrm{b}w_1.\tag{8}$$

另外, 分析第2个目标:

$$\begin{aligned}&\sum_{i=1}^{n_\mathrm{S}}(w_1^\mathrm{T}x_{\mathrm{S}_i}-w_1^\mathrm{T}\mu_\mathrm{S})^2=\\ &\sum_{i=1}^{n_\mathrm{S}}(w_1^\mathrm{T}x_{\mathrm{S}_i}-w_1^\mathrm{T}\mu_\mathrm{S})(w_1^\mathrm{T}x_{\mathrm{S}_i}-w_1^\mathrm{T}\mu_\mathrm{S})^\mathrm{T}=\\ &w_1^\mathrm{T}\sum_{i=1}^{n_\mathrm{S}}(x_{\mathrm{S}_i}-\mu_\mathrm{S})(x_{\mathrm{S}_i}-\mu_\mathrm{S})^\mathrm{T}w_1=\\ &w_1^\mathrm{T}\Sigma_{W_\mathrm{S}}w_1.\end{aligned}\tag{9}$$

$$\begin{aligned}&\sum_{i=1}^{n_\mathrm{T}}(w_1^\mathrm{T}x_{\mathrm{T}_i}-w_1^\mathrm{T}\mu_\mathrm{T})^2=\\ &\sum_{i=1}^{n_\mathrm{T}}(w_1^\mathrm{T}x_{\mathrm{T}_i}-w_1^\mathrm{T}\mu_\mathrm{T})(w_1^\mathrm{T}x_{\mathrm{T}_i}-w_1^\mathrm{T}\mu_\mathrm{T})^\mathrm{T}=\\ &w_1^\mathrm{T}\sum_{i=1}^{n_\mathrm{T}}(x_{\mathrm{T}_i}-\mu_\mathrm{T})(x_{\mathrm{T}_i}-\mu_\mathrm{T})^\mathrm{T}w_1=\\ &w_1^\mathrm{T}\Sigma_{W_\mathrm{T}}w_1.\end{aligned}\tag{10}$$

由式(9)和式(10)得, 本文的目标是希望

$$w_1^\mathrm{T}(\Sigma_{W_\mathrm{S}}+\alpha\Sigma_{W_\mathrm{T}})w_1\tag{11}$$

尽可能大.

综上, 设

$$J(w_1)=\frac{w_1^\mathrm{T}\Sigma_\mathrm{b}w_1}{w_1^\mathrm{T}(\Sigma_{W_\mathrm{S}}+\alpha\Sigma_{W_\mathrm{T}})w_1}.\tag{12}$$

由于笔者希望式(8)尽可能小, 希望式(9)尽可能大, 所以这里是希望$J(w_1)$尽可能小, 于是可以采取和LDA相反的解决方式得到w_1的解为$(\Sigma_{W_\mathrm{S}}+\alpha\Sigma_{W_\mathrm{T}})^{-1}\Sigma_\mathrm{b}$的最小的特征值对应的特征向量.

当$d>1$时, 最终的优化目标类似, 为

$$\min_W\ \frac{\mathrm{tr}(W^\mathrm{T}\Sigma_\mathrm{b}W)}{\mathrm{tr}(W^\mathrm{T}(\Sigma_{W_\mathrm{S}}+\alpha\Sigma_{W_\mathrm{T}})W)}.\tag{13}$$

定理 1　定义

$$J(W)=\frac{\mathrm{tr}(W^\mathrm{T}MW)}{\mathrm{tr}(W^\mathrm{T}NW)},\tag{14}$$

则优化目标

$$\min_W\ J(W)\tag{15}$$

的解W取$N^{-1}M$的前d个最小特征值对应的特征向量.

证　因为W的每个向量的大小不影响最终结果, 只在意W各个分量的方向, 且各个分量正交, 则优化目标(15)等价于

$$\begin{aligned}&\min_W\ \mathrm{tr}(W^\mathrm{T}MW),\\ &\text{s.t. }W^\mathrm{T}NW=I_d.\end{aligned}\tag{16}$$

由拉格朗日乘子法设

$$F(W)=\mathrm{tr}(W^\mathrm{T}MW)+\mathrm{tr}((I_d-W^\mathrm{T}NW)\Theta).\tag{17}$$

对$F(W)$中的W求导, 令导数等于0, 得到

$$MW=NW\Theta.\tag{18}$$

两边同乘以W, 由于$W^\mathrm{T}NW=I_d$, 则

$$W^\mathrm{T}MW=\Theta,\tag{19}$$

则Θ除了对角线外全是0.

由于笔者希望最小化$\mathrm{tr}(W^\mathrm{T}MW)$, 则$\mathrm{tr}(\Theta)$越小越好. 式(18)两边同时乘以$N^{-1}$得

$$N^{-1}MW=W\Theta,\tag{20}$$

则Θ对角线取值为$N^{-1}M$对应的前d个最小的特征

值, 而W取值为$N^{-1}M$的前d个最小特征值对应的特征向量. 证毕.

由定理1知, 优化目标(13)的解W的取值为$(\Sigma_{W_\mathrm{S}}+\alpha\Sigma_{W_\mathrm{T}})^{-1}\Sigma_\mathrm{b}$的前$d$个最小的特征值对应的特征向量.

接下来, 为了方便代码实现和引入核函数, 将目标函数进行变形.

令$1_\mathrm{S}\in\{0,1\}^{n_\mathrm{S}+n_\mathrm{T}}$, $1_\mathrm{T}\in\{0,1\}^{n_\mathrm{S}+n_\mathrm{T}}$分别为源域和目标域的指示向量, 即$1_\mathrm{S}$的第$j$个分量为1当且仅当$x_j\in\mathcal{X}_\mathrm{S}$, 否则为0; 1_T的第j个分量为1当且仅当$x_j\in\mathcal{X}_\mathrm{T}$, 否则为0. 定义$I_\mathrm{S}\in\mathbb{R}^{(n_\mathrm{S}+n_\mathrm{T})\times(n_\mathrm{S}+n_\mathrm{T})}$表示左上角为$n_\mathrm{S}\times n_\mathrm{S}$的单位向量, 其余地方均为0的矩阵; $I_\mathrm{T}\in\mathbb{R}^{(n_\mathrm{S}+n_\mathrm{T})\times(n_\mathrm{S}+n_\mathrm{T})}$表示右下角为$n_\mathrm{T}\times n_\mathrm{T}$的单位向量, 其余地方均为0的矩阵. 则

$$\mu_\mathrm{S}=\frac{1}{n_\mathrm{S}}X1_\mathrm{S},\ \mu_\mathrm{T}=\frac{1}{n_\mathrm{T}}X1_\mathrm{T}. \tag{21}$$

令

$$D_\mathrm{b}=(\frac{1}{n_\mathrm{S}}1_\mathrm{S}-\frac{1}{n_\mathrm{T}}1_\mathrm{T})(\frac{1}{n_\mathrm{S}}1_\mathrm{S}-\frac{1}{n_\mathrm{T}}1_\mathrm{T})^\mathrm{T}, \tag{22}$$

则Σ_b可表示为

$$\begin{aligned}&\Sigma_\mathrm{b}=(\mu_\mathrm{S}-\mu_\mathrm{T})(\mu_\mathrm{S}-\mu_\mathrm{T})^\mathrm{T}=\\&(\frac{1}{n_\mathrm{S}}X1_\mathrm{S}-\frac{1}{n_\mathrm{T}}X1_\mathrm{T})(\frac{1}{n_\mathrm{S}}X1_\mathrm{S}-\frac{1}{n_\mathrm{T}}X1_\mathrm{T})^\mathrm{T}=\\&X(\frac{1}{n_\mathrm{S}}1_\mathrm{S}-\frac{1}{n_\mathrm{T}}1_\mathrm{T})(\frac{1}{n_\mathrm{S}}1_\mathrm{S}-\frac{1}{n_\mathrm{T}}1_\mathrm{T})^\mathrm{T}X^\mathrm{T}=\\&XD_\mathrm{b}X^\mathrm{T}.\end{aligned} \tag{23}$$

定义

$$D_{W_\mathrm{S}}=I_\mathrm{S}-\frac{1}{n_\mathrm{S}}1_\mathrm{S}1_\mathrm{S}^\mathrm{T}, \tag{24}$$

于是Σ_{W_S}可以表示为

$$\begin{aligned}&\Sigma_{W_\mathrm{S}}=\sum_{x\in\mathcal{X}_\mathrm{S}}(x-\mu_\mathrm{S})(x-\mu_\mathrm{S})^\mathrm{T}=\\&\sum_{x\in\mathcal{X}_\mathrm{S}}xx^\mathrm{T}-\sum_{x\in\mathcal{X}_\mathrm{S}}x\mu_\mathrm{S}^\mathrm{T}-\sum_{x\in\mathcal{X}_\mathrm{S}}\mu_\mathrm{S}x^\mathrm{T}+\sum_{x\in\mathcal{X}_\mathrm{S}}\mu_\mathrm{S}\mu_\mathrm{S}^\mathrm{T}=\\&XI_\mathrm{S}X^\mathrm{T}-n_\mathrm{S}\mu_\mathrm{S}\mu_\mathrm{S}^\mathrm{T}=X(I_\mathrm{S}-\frac{1}{n_\mathrm{S}}1_\mathrm{S}1_\mathrm{S}^\mathrm{T})X^\mathrm{T}=\\&XD_{W_\mathrm{S}}X^\mathrm{T}.\end{aligned} \tag{25}$$

定义

$$D_{W_\mathrm{T}}=I_\mathrm{T}-\frac{1}{n_\mathrm{T}}1_\mathrm{T}1_\mathrm{T}^\mathrm{T}, \tag{26}$$

则Σ_{W_T}可以类似地表示为

$$\begin{aligned}&\Sigma_{W_\mathrm{T}}=\sum_{x\in\mathcal{X}_\mathrm{T}}(x-\mu_\mathrm{T})(x-\mu_\mathrm{T})^\mathrm{T}=\\&\sum_{x\in\mathcal{X}_\mathrm{T}}xx^\mathrm{T}-\sum_{x\in\mathcal{X}_\mathrm{T}}x\mu_\mathrm{T}^\mathrm{T}-\sum_{x\in\mathcal{X}_\mathrm{T}}\mu_\mathrm{T}x^\mathrm{T}+\sum_{x\in\mathcal{X}_\mathrm{T}}\mu_\mathrm{T}\mu_\mathrm{T}^\mathrm{T}=\\&XI_\mathrm{T}X^\mathrm{T}-n_\mathrm{T}\mu_\mathrm{T}\mu_\mathrm{T}^\mathrm{T}=X(I_\mathrm{T}-\frac{1}{n_\mathrm{T}}1_\mathrm{T}1_\mathrm{T}^\mathrm{T})X^\mathrm{T}=\\&XD_{W_\mathrm{T}}X^\mathrm{T}.\end{aligned} \tag{27}$$

通过以上转化, 优化目标(13)可以变形为

$$\min_W\frac{\mathrm{tr}(W^\mathrm{T}XD_\mathrm{b}X^\mathrm{T}W)}{\mathrm{tr}(W^\mathrm{T}X(D_{W_\mathrm{S}}+\alpha D_{W_\mathrm{T}})X^\mathrm{T}W)}, \tag{28}$$

则根据定理1, 最终优化目标的结果W取值为

$$(X(D_{W_\mathrm{S}}+\alpha D_{W_\mathrm{T}})X^\mathrm{T})^{-1}XD_\mathrm{b}X^\mathrm{T}$$

的前d个最小特征值对应的特征向量组成的矩阵. 则模型算法实现如算法1所示.

算法 1 单类别投影基构造算法.

Input: 源域X_S, Y_S, 目标域样本X_T, 参数α, 特征子空间维度d

Output: 目标域标签Y_T

1) 根据式(22)(24)(26)计算D_b, D_{W_S}, D_{W_T};

2) 计算$X=[X_\mathrm{S}\ \ X_\mathrm{T}]$;

3) 求出$(X(D_{W_\mathrm{S}}+\alpha D_{W_\mathrm{T}})X^\mathrm{T})^{-1}XD_\mathrm{b}X^\mathrm{T}$的前$d$个最小的特征值对应的特征向量, 组成矩阵$W$;

4) 在投影空间$W^\mathrm{T}X$上运用监督学习方法判断目标域标签Y_T.

4.2 模型2: 有监督多类别投影

源域$\mathcal{D}_\mathrm{S}$内的数据有标签

$$\mathcal{D}_\mathrm{S}=\{(x_{\mathrm{S}_1},y_{\mathrm{S}_1}),\cdots,(x_{\mathrm{S}_{n_\mathrm{S}}},y_{\mathrm{S}_{n_\mathrm{S}}})\},$$

其中: $x_{\mathrm{S}_i}\in\mathcal{X}_\mathrm{S}$, $\forall i=1,\cdots,n_\mathrm{S}$, y_i为x_i的标签, 代表的是x_i的类别. 目标域$\mathcal{D}_\mathrm{T}$内的数据部分有标签, 部分无标签:

$$\begin{aligned}\mathcal{D}_\mathrm{T}=\{&(x_{\mathrm{T}_1},y_{\mathrm{T}_1}),\cdots,(x_{\mathrm{T}_{n_{\mathrm{T}_1}}},y_{\mathrm{T}_{n_{\mathrm{T}_1}}}),\\&x_{\mathrm{T}_{n_{\mathrm{T}_1}+1}},\cdots,x_{\mathrm{T}_{n_{\mathrm{T}_1}+n_{\mathrm{T}_2}}}\},\end{aligned}$$

其中: $x_{\mathrm{T}_i}\in\mathcal{X}_\mathrm{T}$, $\forall i=1,\cdots,n_{\mathrm{T}_1}+n_{\mathrm{T}_2}$. 而目标域中的标签也代表类别. 令

$$\begin{aligned}&\mathcal{D}_{\mathrm{T}_1}=\{(x_{\mathrm{T}_1},y_{\mathrm{T}_1}),\cdots,(x_{\mathrm{T}_{n_{\mathrm{T}_1}}},y_{\mathrm{T}_{n_{\mathrm{T}_1}}})\},\\&\mathcal{D}_{\mathrm{T}_2}=\{x_{\mathrm{T}_{n_{\mathrm{T}_1}+1}},\cdots,x_{\mathrm{T}_{n_{\mathrm{T}_1}+n_{\mathrm{T}_2}}}\}.\end{aligned}$$

设源域与目标域类别数目相同, 均为K个类别. 本问题同样假设源域样本和目标域样本在同一个域中, 只是分布不同, 即$\mathcal{X}_\mathrm{S}=\mathcal{X}_\mathrm{T}$, $P_\mathrm{S}(X)\neq P_\mathrm{T}(X)$, 这里假设$\mathcal{X}_\mathrm{S}=\mathcal{X}_\mathrm{T}=\mathbb{R}^m$.

在进行分析之前, 如上一小节进行一些定义, 定义$X_\mathrm{S}=[x_{\mathrm{S}_1}\ \cdots\ x_{\mathrm{S}_{n_\mathrm{S}}}]$, $X_\mathrm{T}=[x_{\mathrm{T}_1}\ \cdots\ x_{\mathrm{T}_{n_{\mathrm{T}_1}}}]$, $X=[X_\mathrm{S}\ \ X_\mathrm{T}]=[x_{\mathrm{S}_1}\ \cdots\ x_{\mathrm{S}_{n_\mathrm{S}}}\ \ x_{\mathrm{T}_1}\ \cdots\ x_{\mathrm{T}_{n_{\mathrm{T}_1}}}]$. 定义$\mathcal{X}_{\mathrm{S}_k}$为源域第$k$类样本的集合, $\mathcal{X}_{\mathrm{T}_k}$为目标域有标签的第$k$类样本的集合.

$\mu_\mathrm{S}=\frac{1}{n_\mathrm{S}}\sum\limits_{i=1}^{n_\mathrm{S}}x_{\mathrm{S}_i}$, $\mu_\mathrm{T}=\frac{1}{n_\mathrm{T}}\sum\limits_{i=1}^{n_\mathrm{T}}x_{\mathrm{T}_i}$ 分别为源域和目标域的中心. 定义$n_{\mathrm{S}_k}=|\mathcal{X}_{\mathrm{S}_k}|$, $n_{\mathrm{T}_k}=|\mathcal{X}_{\mathrm{T}_k}|$为第$k$类源域和目标域的样本个数. 定义$\mu_{\mathrm{S}_k}$与$\mu_{\mathrm{T}_k}$分别为源域和目标域第$k$类样本的的中心. 即$\mu_{\mathrm{S}_k}=\dfrac{\sum\limits_{x\in\mathcal{X}_{\mathrm{S}_k}}x}{n_{\mathrm{S}_k}}$, $\mu_{\mathrm{T}_k}=\dfrac{\sum\limits_{x\in\mathcal{X}_{\mathrm{T}_k}}x}{n_{\mathrm{T}_k}}$.

由于上一章假设存在一个投影空间$\mathcal{Z}$和一个到该

投影空间的映射φ,

$$\begin{aligned}\varphi:\ & \mathcal{X}\to\mathcal{Z},\\ & X\to W^{\mathrm{T}}X,\end{aligned}\tag{29}$$

使得源域和目标域在该投影空间上的分布相同, 即$P_{\mathrm{S}}(\varphi(X))=P_{\mathrm{T}}(\varphi(X))$. 假设在该投影空间上, 源域和目标域基于样本的预测函数是相同的, 即$P_{\mathrm{S}}(Y|\varphi(X))=P_{\mathrm{T}}(Y|\varphi(X))$.

笔者希望找到这样的低维投影空间, 使得源域和目标域在该投影空间上的分布尽可能相似, 并且希望该投影空间上分类尽可能精确. 假设该投影空间的基为$\{w_1,\cdots,w_d\}$, $w_i\in\mathbb{R}^m$, 令$W=[w_1\ \cdots\ w_d]$.

为了实现这个目标, 笔者要求: 首先, 采取类似于MMD的定义, 希望源域和目标域在投影空间上的中心尽可能近, 以达到源域和目标域在该投影空间上的分布尽可能相似. 笔者要求源域和目标域的中心在该投影空间上的投影尽可能小, 即

$$\min_W\ \|W^{\mathrm{T}}\mu_{\mathrm{S}}-W^{\mathrm{T}}\mu_{\mathrm{T}}\|_2^2.\tag{30}$$

定义$\Sigma_{\mathrm{MMD}}=(\mu_{\mathrm{S}}-\mu_{\mathrm{T}})(\mu_{\mathrm{S}}-\mu_{\mathrm{T}})^{\mathrm{T}}$, 可得

$$\begin{aligned}&\|W^{\mathrm{T}}\mu_{\mathrm{S}}-W^{\mathrm{T}}\mu_{\mathrm{T}}\|_2^2=\\ &\mathrm{tr}(W^{\mathrm{T}}\mu_{\mathrm{S}}-W^{\mathrm{T}}\mu_{\mathrm{T}})(W^{\mathrm{T}}\mu_{\mathrm{S}}-W^{\mathrm{T}}\mu_{\mathrm{T}})^{\mathrm{T}}=\\ &\mathrm{tr}(W^{\mathrm{T}}(\mu_{\mathrm{S}}-\mu_{\mathrm{T}})(\mu_{\mathrm{S}}-\mu_{Y})^{\mathrm{T}}W)=\\ &\mathrm{tr}(W^{\mathrm{T}}\Sigma_{\mathrm{MMD}}W),\end{aligned}\tag{31}$$

即笔者希望

$$\min_W\ \mathrm{tr}(W^{\mathrm{T}}\Sigma_{\mathrm{MMD}}W).\tag{32}$$

同时, 为了使最终的模型的分类尽可能精确, 采取线性判别分析里面的思路, 让源域和目标域相同类别的投影尽可能集中, 不同类别的投影尽可能分散.

首先, 对于源域和目标域中不同类别的投影尽可能分散这一目标, 以源域和目标域内不同类别对用的中心点之间的距离衡量, 即希望

$$\begin{aligned}&\max_W\ \sum_{i\neq j}\|W^{\mathrm{T}}\mu_{\mathrm{S}_i}-W^{\mathrm{T}}\mu_{\mathrm{S}_j}\|_2^2,\\ &\max_W\ \sum_{i\neq j}\|W^{\mathrm{T}}\mu_{\mathrm{T}_i}-W^{\mathrm{T}}\mu_{\mathrm{T}_j}\|_2^2.\end{aligned}\tag{33}$$

令$\Sigma_{d_{\mathrm{S}}}=\sum\limits_{i\neq j}(\mu_{\mathrm{S}_i}-\mu_{\mathrm{S}_j})(\mu_{\mathrm{S}_i}-\mu_{\mathrm{S}_j})^{\mathrm{T}}$, 可以得到

$$\begin{aligned}&\sum_{i\neq j}\|W^{\mathrm{T}}\mu_{\mathrm{S}_i}-W^{\mathrm{T}}\mu_{\mathrm{S}_j}\|_2^2=\\ &\mathrm{tr}((W^{\mathrm{T}}\mu_{\mathrm{S}_i}-W^{\mathrm{T}}\mu_{\mathrm{S}_j})(W^{\mathrm{T}}\mu_{\mathrm{S}_i}-W^{\mathrm{T}}\mu_{\mathrm{S}_j})^{\mathrm{T}})=\\ &\mathrm{tr}(W^{\mathrm{T}}(\mu_{\mathrm{S}_i}-\mu_{\mathrm{S}_j})(\mu_{\mathrm{S}_i}-\mu_{\mathrm{S}_j})^{\mathrm{T}}W)=\\ &\mathrm{tr}(W^{\mathrm{T}}\Sigma_{d_{\mathrm{S}}}W).\end{aligned}\tag{34}$$

同样令$\Sigma_{d_{\mathrm{T}}}=\sum\limits_{i\neq j}(\mu_{\mathrm{T}_i}-\mu_{\mathrm{T}_j})(\mu_{\mathrm{T}_i}-\mu_{\mathrm{T}_j})^{\mathrm{T}}$, 则

$$\begin{aligned}&\sum_{i\neq j}\|W^{\mathrm{T}}\mu_{\mathrm{T}_i}-W^{\mathrm{T}}\mu_{\mathrm{T}_j}\|_2^2=\\ &\mathrm{tr}((W^{\mathrm{T}}\mu_{\mathrm{T}_i}-W^{\mathrm{T}}\mu_{\mathrm{T}_j})(W^{\mathrm{T}}\mu_{\mathrm{T}_i}-W^{\mathrm{T}}\mu_{\mathrm{T}_j})^{\mathrm{T}})=\\ &\mathrm{tr}(W^{\mathrm{T}}(\mu_{\mathrm{T}_i}-\mu_{\mathrm{T}_j})(\mu_{\mathrm{T}_i}-\mu_{\mathrm{T}_j})^{\mathrm{T}}W)=\\ &\mathrm{tr}(W^{\mathrm{T}}\Sigma_{d_{\mathrm{T}}}W).\end{aligned}\tag{35}$$

经过如上转化, 式(33)可以转化为

$$\begin{aligned}&\max_W\ \mathrm{tr}(W^{\mathrm{T}}\Sigma_{d_{\mathrm{S}}}W),\\ &\max_W\ \mathrm{tr}(W^{\mathrm{T}}\Sigma_{d_{\mathrm{T}}}W).\end{aligned}\tag{36}$$

同时, 笔者希望源域和目标域相同类别在投影空间上的映射尽可能近, 以各类样本点距该类中心点的距离为衡量标准, 即希望

$$\begin{aligned}&\min_W\ \sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{S}_i}}\|W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{S}_i}\|_2^2,\\ &\min_W\ \sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{T}_i}}\|W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{T}_i}\|_2^2.\end{aligned}\tag{37}$$

定义源域的类内散度矩阵$\Sigma_{W_{\mathrm{S}}}=\sum\limits_{i=1}^{K}\Sigma_{W_{\mathrm{S}_i}}$, 其中$\Sigma_{W_{\mathrm{S}_i}}=\sum\limits_{x\in\mathcal{X}_{\mathrm{S}_i}}(x-\mu_{\mathrm{S}_i})(x-\mu_{\mathrm{S}_i})^{\mathrm{T}}$, 则

$$\begin{aligned}&\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{S}_i}}\|W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{S}_i}\|_2^2=\\ &\mathrm{tr}(\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{S}_i}}(W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{S}_i})(W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{S}_i})^{\mathrm{T}})=\\ &\mathrm{tr}(\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{S}_i}}W^{\mathrm{T}}(x-\mu_{\mathrm{S}_i})(x-\mu_{\mathrm{S}_i})^{\mathrm{T}}W)=\\ &\mathrm{tr}(W^{\mathrm{T}}\Sigma_{W_{\mathrm{S}}}W).\end{aligned}\tag{38}$$

定义目标域的类内散度矩阵$\Sigma_{W_{\mathrm{T}}}=\sum\limits_{i=1}^{K}\Sigma_{W_{\mathrm{T}_i}}$, 其中$\Sigma_{W_{\mathrm{T}_i}}=\sum\limits_{x\in\mathcal{X}_{\mathrm{T}_i}}(x-\mu_{\mathrm{T}_i})(x-\mu_{\mathrm{T}_i})^{\mathrm{T}}$, 则

$$\begin{aligned}&\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{T}_i}}\|W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{T}_i}\|_2^2=\\ &\mathrm{tr}(\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{T}_i}}(W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{T}_i})(W^{\mathrm{T}}x-W^{\mathrm{T}}\mu_{\mathrm{T}_i})^{\mathrm{T}})=\\ &\mathrm{tr}(\sum_{i=1}^{K}\sum_{x\in\mathcal{X}_{\mathrm{T}_i}}W^{\mathrm{T}}(x-\mu_{\mathrm{T}_i})(x-\mu_{\mathrm{T}_i})^{\mathrm{T}}W)=\\ &\mathrm{tr}(W^{\mathrm{T}}\Sigma_{W_{\mathrm{T}}}W).\end{aligned}\tag{39}$$

综上, 式(37)可以转化为

$$\begin{aligned}&\min_W\ \mathrm{tr}(W^{\mathrm{T}}\Sigma_{W_{\mathrm{S}}}W),\\ &\min_W\ \mathrm{tr}(W^{\mathrm{T}}\Sigma_{W_{\mathrm{T}}}W).\end{aligned}\tag{40}$$

因此最终的优化目标可以写为

$$\max_W\ \frac{\mathrm{tr}(W^{\mathrm{T}}(\Sigma_{d_{\mathrm{S}}}+\alpha\Sigma_{d_{\mathrm{T}}})W)}{\mathrm{tr}(W^{\mathrm{T}}(\Sigma_{\mathrm{MMD}}+\beta\Sigma_{W_{\mathrm{S}}}+\gamma\Sigma_{W_{\mathrm{T}}})W)}.\tag{41}$$

定理 2　定义

$$J(W)=\frac{\mathrm{tr}(W^{\mathrm{T}}MW)}{\mathrm{tr}(W^{\mathrm{T}}NW)}, \tag{42}$$

则优化目标

$$\max_{W}\ J(W) \tag{43}$$

的解W取$N^{-1}M$的前d个最大特征值对应的特征向量.

证 因为W的每个向量的大小不影响最终结果, 只在意W各个分量的方向, 且各个分量正交, 则优化目标(43)等价于

$$\begin{aligned}&\min_{W}\ -\mathrm{tr}(W^{\mathrm{T}}MW),\\&\text{s.t. } W^{\mathrm{T}}NW=I_d.\end{aligned} \tag{44}$$

由拉格朗日乘子法设

$$F(W)=-\mathrm{tr}(W^{\mathrm{T}}MW)+\mathrm{tr}((I_d-W^{\mathrm{T}}NW)\Theta), \tag{45}$$

对$F(W)$中的W求导, 令导数等于0, 得到

$$-MW=NW\Theta. \tag{46}$$

两边同乘以W, 由于$W^{\mathrm{T}}NW=I_d$, 则

$$-W^{\mathrm{T}}MW=\Theta, \tag{47}$$

则Θ除了对角线外全是0.

由于笔者希望最小化$-\mathrm{tr}(W^{\mathrm{T}}MW)$, 则$\mathrm{tr}(\Theta)$越小越好. 式(46)两边同时乘以$N^{-1}$得

$$-N^{-1}MW=W\Theta, \tag{48}$$

则Θ对角线取值为$-N^{-1}M$对应的前d个最小的特征值, 即$N^{-1}M$对应的前d个最大的特征值, 而W取值为$N^{-1}M$的前d个最大特征值对应的特征向量.

证毕.

则根据定理2, 该优化问题(41)的解W取

$$(\Sigma_{\mathrm{MMD}}+\beta\Sigma_{W_{\mathrm{S}}}+\gamma\Sigma_{W_{\mathrm{T}}})^{-1}(\Sigma_{d_{\mathrm{S}}}+\alpha\Sigma_{d_{\mathrm{T}}})$$

的前d个最大特征值对应的特征向量.

同样, 为了方便代码实现和核方法转化, 同上一个模型一样, 对最终优化目标进行变形.

令$1_{\mathrm{S}}\in\{0,1\}^{n_{\mathrm{S}}+n_{\mathrm{T}}}$, $1_{\mathrm{T}}\in\{0,1\}^{n_{\mathrm{S}}+n_{\mathrm{T}}}$分别为源域和目标域的指示向量, 即$1_{\mathrm{S}}$的第$j$个分量为1当且仅当$x_j\in\mathcal{X}_{\mathrm{S}}$, 否则为0; 1_{T}的第j个分量为1当且仅当$x_j\in\mathcal{X}_{\mathrm{T}}$, 否则为0. 同理定义$1_{\mathrm{S}_i}\in\{0,1\}^{n_{\mathrm{S}}+n_{\mathrm{T}}}$, $1_{\mathrm{T}_i}\in\{0,1\}^{n_{\mathrm{S}}+n_{\mathrm{T}}}$分别为源域和目标域第$i$类样本的指示向量, 即$1_{\mathrm{S}_i}$的第$j$个分量为1当且仅当$x_j\in\mathcal{X}_{\mathrm{S}_i}$, 否则为0; 1_{T_i}的第j个分量为1当且仅当$x_j\in\mathcal{X}_{\mathrm{T}_i}$, 否则为0.

定义$I_{\mathrm{S}}\in\mathbb{R}^{(n_{\mathrm{S}}+n_{\mathrm{T}})\times(n_{\mathrm{S}}+n_{\mathrm{T}})}$表示左上角为$n_{\mathrm{S}}\times n_{\mathrm{S}}$的单位向量, 其余地方均为0的矩阵; $I_{\mathrm{T}}\in\mathbb{R}^{(n_{\mathrm{S}}+n_{\mathrm{T}})\times(n_{\mathrm{S}}+n_{\mathrm{T}})}$表示右下角为$n_{\mathrm{T}}\times n_{\mathrm{T}}$的单位向量, 其余地方均为0的矩阵.

则

$$\begin{cases}\mu_{\mathrm{S}}=\dfrac{1}{n_{\mathrm{S}}}X1_{\mathrm{S}},\\ \mu_{\mathrm{T}}=\dfrac{1}{n_{\mathrm{T}}}X1_{\mathrm{T}},\\ \mu_{\mathrm{S}_i}=\dfrac{1}{n_{\mathrm{S}_i}}X1_{\mathrm{S}_i},\\ \mu_{\mathrm{T}_i}=\dfrac{1}{n_{\mathrm{T}_i}}X1_{\mathrm{T}_i}.\end{cases} \tag{49}$$

同上一节定义,

$$D_{\mathrm{MMD}}=(\frac{1}{n_{\mathrm{S}}}1_{\mathrm{S}}-\frac{1}{n_{\mathrm{T}}}1_{\mathrm{T}})(\frac{1}{n_{\mathrm{S}}}1_{\mathrm{S}}-\frac{1}{n_{\mathrm{T}}}1_{\mathrm{T}})^{\mathrm{T}}, \tag{50}$$

$$D_{d_{\mathrm{S}}}=\sum_i n_{\mathrm{S}_i}(\frac{1}{n_{\mathrm{S}_i}}1_{\mathrm{S}_i}-\frac{1}{n_{\mathrm{S}}}1_{\mathrm{S}})(\frac{1}{n_{\mathrm{S}_i}}1_{\mathrm{S}_i}-\frac{1}{n_{\mathrm{S}}}1_{\mathrm{S}})^{\mathrm{T}}, \tag{51}$$

$$D_{d_{\mathrm{T}}}=\sum_i n_{\mathrm{T}_i}(\frac{1}{n_{\mathrm{T}_i}}1_{\mathrm{T}_i}-\frac{1}{n_{\mathrm{T}}}1_{\mathrm{T}})(\frac{1}{n_{\mathrm{T}_i}}1_{\mathrm{T}_i}-\frac{1}{n_{\mathrm{T}}}1_{\mathrm{T}})^{\mathrm{T}}, \tag{52}$$

$$D_{W_{\mathrm{S}}}=(I_{\mathrm{S}}-\sum_i\frac{1}{n_{\mathrm{S}_i}}1_{\mathrm{S}_i}1_{\mathrm{S}_i}^{\mathrm{T}}), \tag{53}$$

$$D_{W_{\mathrm{T}}}=(I_{\mathrm{T}}-\sum_i\frac{1}{n_{\mathrm{T}_i}}1_{\mathrm{T}_i}1_{\mathrm{T}_i}^{\mathrm{T}}). \tag{54}$$

则类似于模型1中的计算方法, 最终优化模型可表示为

$$\max_{W}\ \frac{\mathrm{tr}(W^{\mathrm{T}}X(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})X^{\mathrm{T}}W)}{\mathrm{tr}(W^{\mathrm{T}}X(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})X^{\mathrm{T}}W)}. \tag{55}$$

则根据定理2, 最终优化目标的解W取值为

$$(X(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})X^{\mathrm{T}})^{-1}\cdot X(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})X^{\mathrm{T}}$$

的前d个最大特征值对应的特征向量组成的矩阵. 最终算法实现如算法2所示.

算法 2 有监督多类别投影算法.

Input: 源域X_{S}, Y_{S}; 目标域有标签的样本及标签X_{T_1}, X_{T_2}, Y_{T_1}; 参数α, β, γ; 特征子空间维度d.

Output: 目标域无标签样本X_{T_2}的标签Y_{T_2}.

1) 根据式(50)–(54)计算D_{MMD}, $D_{d_{\mathrm{S}}}$, $D_{d_{\mathrm{T}}}$, $D_{W_{\mathrm{S}}}$, $D_{W_{\mathrm{T}}}$;

2) 计算$X=[X_{\mathrm{S}}\ \ X_{\mathrm{T}_1}]$;

3) 求出$(X(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})X^{\mathrm{T}})^{-1}X(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})X^{\mathrm{T}}$的前$d$个最大的特征值对应的特征向量, 组成矩阵$W$;

4) 在投影空间$W^{\mathrm{T}}X$上目标域无标签样本X_{T_2}每个数据的标签Y_{T_2}取距离其最近的有标签样本的类别.

4.3 核方法的引入

基于模型2有监督多类别投影引入核方法.

当$d=1$时, 假设可以通过某种映射$\phi:\ \mathcal{X}\to\mathcal{H}$,

将样本映射到一个高维希尔伯特空间中并线性可分, 让在该空间$\mathcal{H}$中执行线性判别分析$h(x)=w^{\mathrm{T}}\phi(x)$. 则优化问题转化为

$$\max_{w}\ \frac{w^{\mathrm{T}}X^{\phi}(D^{\phi}_{d_{\mathrm{S}}}+\lambda D^{\phi}_{d_{\mathrm{T}}})X^{\phi\mathrm{T}}w}{w^{\mathrm{T}}X^{\phi}(D^{\phi}_{\mathrm{MMD}}+\beta D^{\phi}_{W_{\mathrm{S}}}+\gamma D^{\phi}_{W_{\mathrm{T}}})X^{\phi\mathrm{T}}w},\tag{56}$$

$X^{\phi}=[\phi(x_1)\ \cdots\ \phi(x_{n_{\mathrm{S}}})\ \ \phi(x_{n_{\mathrm{S}}+1})\ \cdots\ \phi(x_{n_{\mathrm{T}_1}})]$为$X$中每个样本到$\mathcal{H}$空间中的映射.

命题 1(表示定理)[20–21]　设$\mathcal{H}$为核函数K对应映射后的空间(RKHS), $\|h\|_{\mathcal{H}}$表示$\mathcal{H}$空间中h的范数, 则对于任意单调递增的函数Ω和任意非负损失函数L, 优化问题

$$\min_{h\in\mathcal{H}}\ L(h(x_1),\cdots,h(x_N))+\Omega(\|h\|_{\mathcal{H}})$$

的解总可以表述为核函数K的线性组合$h^{\star}(x)=\sum\limits_{i=1}^{N}\alpha_i K(x,x_i)$.

令$J(w)$作为损失函数, 令$\Omega=0$, 由核函数表示定理得, $h(x)$可写为

$$h(x)=\sum_{i=1}^{n_{\mathrm{S}}+n_{\mathrm{T}}}\alpha_i k(x_i,x),\tag{57}$$

则

$$w=\sum_{i=1}^{n_{\mathrm{S}}+n_{\mathrm{T}}}\alpha_i\phi(x_i)=X^{\phi}\alpha,\tag{58}$$

其中

$$\alpha=\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_{n_{\mathrm{S}}+n_{\mathrm{T}}}\end{bmatrix}.\tag{59}$$

定义核函数$k(x_i,x_j)=\phi^{\mathrm{T}}(x_i)\phi(x_j)$, 则核矩阵$K=X^{\phi\mathrm{T}}X^{\phi}$. 则优化目标(56)可以表示为

$$\begin{aligned}&\frac{w^{\mathrm{T}}X^{\phi}(D^{\phi}_{d_{\mathrm{S}}}+\lambda D^{\phi}_{d_{\mathrm{T}}})X^{\phi\mathrm{T}}w}{w^{\mathrm{T}}X^{\phi}(D^{\phi}_{\mathrm{MMD}}+\beta D^{\phi}_{W_{\mathrm{S}}}+\gamma D^{\phi}_{W_{\mathrm{T}}})X^{\phi\mathrm{T}}w}=\\ &\frac{\alpha^{\mathrm{T}}X^{\phi\mathrm{T}}X^{\phi}(D^{\phi}_{d_{\mathrm{S}}}+\lambda D^{\phi}_{d_{\mathrm{T}}})X^{\phi\mathrm{T}}X^{\phi}\alpha}{\alpha^{\mathrm{T}}X^{\phi\mathrm{T}}X^{\phi}(D^{\phi}_{\mathrm{MMD}}+\beta D^{\phi}_{W_{\mathrm{S}}}+\gamma D^{\phi}_{W_{\mathrm{T}}})X^{\phi\mathrm{T}}X^{\phi}\alpha}=\\ &\frac{\alpha^{\mathrm{T}}K(D^{\phi}_{d_{\mathrm{S}}}+\lambda D^{\phi}_{d_{\mathrm{T}}})K\alpha}{\alpha^{\mathrm{T}}K(D^{\phi}_{\mathrm{MMD}}+\beta D^{\phi}_{W_{\mathrm{S}}}+\gamma D^{\phi}_{W_{\mathrm{T}}})K\alpha}.\end{aligned}\tag{60}$$

则优化函数可以写为

$$\max_{W}\ \frac{\alpha^{\mathrm{T}}K(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})K\alpha}{\alpha^{\mathrm{T}}K(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})K\alpha},\tag{61}$$

则优化目标的解α取值为$(K(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})K)^{-1}K(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})K$的最大的特征值对应的特征向量组成的矩阵.

对于$d>1$的情况, 优化函数可以表示为

$$\max_{W}\ \frac{\mathrm{tr}(\alpha^{\mathrm{T}}K(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})K\alpha)}{\mathrm{tr}(\alpha^{\mathrm{T}}K(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})K\alpha)},\tag{62}$$

则优化目标的解α取值为$(K(D_{\mathrm{MMD}}+\beta D_{W_{\mathrm{S}}}+\gamma D_{W_{\mathrm{T}}})K)^{-1}K(D_{d_{\mathrm{S}}}+\alpha D_{d_{\mathrm{T}}})K$的前$d$个最大的特征值对应的特征向量组成的矩阵.

对于另外第1个模型, 笔者采取类似的做法, 可以求出其核方法如下:

单类别投影基核方法的构造优化目标可以转化为

$$\min_{\alpha}\ \frac{\mathrm{tr}(\alpha^{\mathrm{T}}KD_{\mathrm{b}}K\alpha)}{\mathrm{tr}(\alpha^{\mathrm{T}}K(D_{W_{\mathrm{S}}}+\alpha D_{W_{\mathrm{T}}})K\alpha)},\tag{63}$$

则α取值为$(K(D_{W_{\mathrm{S}}}+\alpha D_{W_{\mathrm{T}}})K)^{-1}KD_{\mathrm{b}}K$的前$d$个最小的特征值对应的特征向量组成的矩阵.

5　模拟实验

该部分笔者对以上两种模型进行实验仿真, 在随机生成的数据集上观察投影基选取的效果.

5.1　单类别投影基的构造

随机生成一组数据如图1所示, 十字为源域样本, 星状为目标域样本, 根据该算法求得的投影基向量W如直线所示. 其中参数选取$\alpha=1$, $d=1$. 源域样本服从$\mu=[0,0]$, $\sigma=\begin{bmatrix}10&0\\0&10\end{bmatrix}$的高斯分布, 目标域样本服从$\mu=[5,5]$, $\sigma=\begin{bmatrix}10&0\\0&10\end{bmatrix}$的高斯分布. 由数据分布可知, 本次仿真源域和目标域均为单类别的高斯模型, 并且具有相同的方差矩阵, 只是源域和目标域的均值矩阵不同.

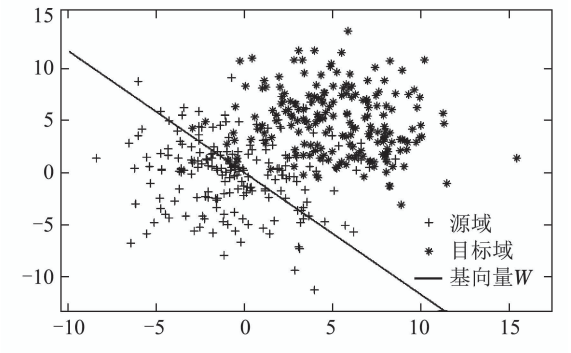

图 1　单类别投影基的构造

Fig. 1　One class projection

从图1中可以看出源域和目标域在该基W上的投影的分布相似程度很高, 并且该投影很好的降低了源域和目标域均值的差异, 并且保留了源域和目标域的方差信息. 因此源域和目标域在该投影基上可以进行一个表现良好的特征迁移学习.

在仿真实验中, 对分布中心有较大差异的源域和目标域进行模拟, 找出其投影基向量W, 从实验结果中可以发现, 源域和目标域在投影基上的投影的分布相似度很高, 并且极大程度上降低了源域和目标域由于中心不同产生的差异, 将源域和目标域有关联或者

相似的信息尽可能地投影到投影基上, 提取出相似特征.

同时也注意到, 本模型的优化问题和线性判别分析的优化问题非常类似, 除了参数选取外, 在优化目标求最大最小处有所差异. 所以本模型可以将源域和目标域视为两类, 优化目标是找出一个组投影基, 使得两类在投影基上类内散度尽可能大, 类间散度尽可能小, 即与线性判别分析有着相反的优化目标. 所以本模型的优化目标的解也可以看作线性判别分析优化目标的解取最小的d个特征值对应的特征向量. 所以在二维情况下, d 取1 时, 本模型的解正好与线性判别分析的解正交.

然而在处理高维问题时, 会出现模型分类精确度降低, 投影基向量W的选取不准确的现象. 这是因为在求解过程中$S_{\rm b}$的秩等于1, 即最终结果只有一个不为0的特征值, 因此当选取前多个特征值对应的特征向量时, 可能会出现随机选取的情况, 并且在选取的子空间投影中进行特征学习的效果会比较差.

5.2 有监督多类别投影

随机生成一组数据如图2所示, 十字为源域数据, 并且样本有不同的两类; 星状为目标域样本, 同样包含不同的两类. 根据LDA算法, 得出的投影W为虚线所示, 根据本文中的算法得出的基向量W为黑色实线所示. 参数选取为

$$\alpha = 10,\ \beta = 0.2,\ \gamma = 1,\ d = 1.$$

源域样本第1类服从$\mu = [2, 3],\ \sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$的高斯分布, 源域样本第2类服从$\mu = [20, 3],\ \sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. 而目标域的两类样本是在源域两类样本上均加上服从$\mu = [10, -5],\ \sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$的高斯分布的偏度, 最终本文模型结果准确率为0.99.

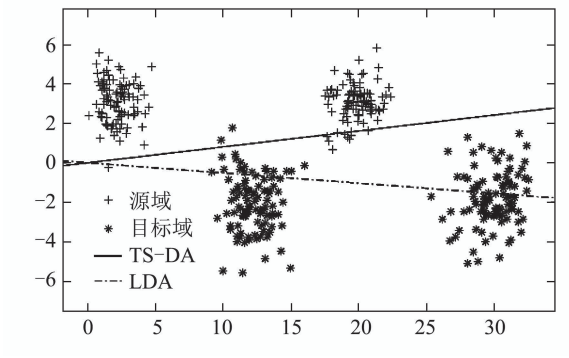

图 2 监督多类别投影

Fig. 2 Supervised multi-class projection

从生成的数据来看源域和目标域的两个类别之间的差别相似, 只是源域和目标域每类之间都有一个服从高斯分布的偏度. 从分类结果图2来看, 基于线性判别分析得到的基W虽然分别对源域和目标域的分类有良好的效果, 但是在处理源域和目标域综合的信息上效果很差. 而本文模型中, 得到的基向量W则将源域和目标域的信息综合处理, 准确度为0.99, 准确率得到明显提升. 同时也可以从图像中看出, 该方法投影基向量相对于线性判别分析的基向量, 发生了偏转, 产生了对源域和目标域共同适用的投影空间. 另外该方法选取的投影基对于源域和目标域内部的不同类投影后的差异较好地体现出来, 达到了笔者希望的效果. 同时, 也在该模型上运用核方法, 选取高斯核, 并且在相同数据下进行模拟, 结果依旧为0.99. 当然处理线性不可分问题时, 核方法效果更为明显.

本模型根据多类别分类, 目标域部分样本类别已知的情况下, 提出了有监督多类别投影模型. 并且基于最大均值差异和线性判别分析, 写出了有监督多类别投影的优化目标, 并给出了解决方法和求解算法. 然而在实验中, 本模型虽然得到很好的效果, 但是优化目标中各个优化量的重要程度不同, 单纯运用参数和分数综合实现优化目标有可能使最终效果不理想.

6 结论与展望

本文根据源域和目标域不同数据类型, 提出两种模型: 单类别投影基的构造与有监督多类别投影, 并且根据迁移学习的基本思想, 并结合最大均值差异和线性判别分析的思想, 构造出了相应的算法, 同时针对线性不可分数据, 给出了相应的核方法, 最后通过模拟实验, 观察到算法取得了较好的效果, 得到了一个对目标域和源域共同适用的基, 为这两种数据情形下的迁移学习的方式提供了参考. 同时也注意到文本的方法仍有可以改进的地方. 首先, 根据最大均值差异可以一定程度上估计两个域的分布差异, 可以改进基于最大均值差异的优化目标, 衡量子空间投影后两个分布的差异, 最终达到找出使两分布相同的投影空间的目的. 其次, 可以对优化目标的实现方式加以改进. 针对多组优化量, 可以采取加减形式优化目标, 以减少由于秩较低造成的较多特征值为0的情况, 另外也可以根据优化量的重要程度, 采取分段式优化方式, 逐次优化各个优化分量, 以达到最终的优化目标. 最后, 笔者希望未来对于更复杂的模型进行实验.

参考文献:

[1] MENG G, WANG Y, DUAN J, et al. Efficient image dehazing with boundary constraint and contextual regularization. *The IEEE International Conference on Computer Vision (ICCV)*. Sydney: IEEE, 2013: 617 – 624.

[2] WANG R J, LI X, AO S, et al. Pelee: A real-time object detection system on mobile devices. *Computing Research Repository*, vol. abs/1804.06882, 2018. [Online].

[3] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing. *Computing Research Repository*, vol. abs/1708.02709, 2017. [Online].

[4] PAN S J, YANG Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345 – 1359.

[5] WEISS K, KHOSHGOFTAAR T M, WANG D. A survey of transfer learning. *Journal of Big Data*, 2016, 3(9): 1 – 40.

[6] CHATTOPADHYAY R, SUN Q, FAN W, et al. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data*, 2012, 6(4): 18-1 – 18-26. [Online].

[7] GONG B, SHI Y, SHA F, et al. Geodesic flow kernel for unsupervised domain adaptation. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE, 2012: 2066 – 2073.

[8] DUAN L, XU D, CHANG S. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE, 2012: 1338 – 1345.

[9] LI F, PAN S J, JIN O, et al. Cross-domain co-extraction of sentiment and topic lexicons. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers–Volume 1, ser. ACL '12*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012: 410 – 419. [Online].

[10] WOLD S, ESBENSEN K, GELADI P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(1/2/3): 37 – 52.

[11] SCHöLKOPF B, SMOLA A, MüLLER K R. Kernel principal component analysis. *International Conference on Artificial Neural Networks*. Berlin, Heidelberg: Springer, 1997: 583 – 588.

[12] MIKA S, RATSCH G, WESTON J, et al. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*. Madison, WI, USA: IEEE, 1999: 41 – 48.

[13] CANDèS E J, LI X, MA Y, et al. Robust principal component analysis? *Journal of ACM*, 2011, 58(3): 11-1 – 11-37. [Online].

[14] LU C, FENG J, CHEN Y, et al. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.567.

[15] PAN S J, TSANG I W, KWOK J T, et al. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011, 22(2): 199 – 210.

[16] LU H, SHEN C, CAO Z, et al. An embarrassingly simple approach to visual domain adaptation. *IEEE Transactions on Image Processing*, 2018, 27(7): 3403 – 3417.

[17] LONG M, WANG J, DING G, et al. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(5): 1076 – 1089.

[18] GRETTON A, BORGWARDT K, RASCH M, et al. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2007: 513 – 520.

[19] GRETTON A, BORGWARDT K, RASCH M J, et al. A kernel method for the two-sample problem. *Computing Research Repository*, arXiv preprint arXiv: 0805.2368, 2008.

[20] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, 7(11): 2399 – 2434.

[21] SCHöLKOPF B, HERBRICH R, SMOLA A J. A generalized representer theorem. *International Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, 2001: 416 – 426.

作者简介:

程朝阳 博士研究生, 目前研究方向为迁移学习, E-mail: zhaoyang9735@163.com;

明 杨 博士研究生, 目前研究方向为机器学习, E-mail: mingyang15@mails.ucas.ac.cn;

洪奕光 研究员, 第4届"关肇直奖"(1997年)获奖论文作者, IEEE Fellow和国家杰出青年基金获得者, 研究方向为多自主体系统、分布式优化、非线性系统、社会网络、机器人等, E-mail: yghong@iss.ac.cn.