最大局部加权均值差异嵌入

皋 军[1,2],黄丽莉[3]
(1.苏州大学江苏省计算机信息处理技术重点实验室,江苏苏州 215006;
2.盐城工学院信息工程学院,江苏盐城 224001;3.安徽理工大学电气与信息工程学院,安徽淮南 232001)

摘 要: 最大均值差异嵌入(Maximum Mean Discrepancy Embedding,MMDE)作为一种基于最大均值差异(Maximum Mean Discrepancy,MMD)度量的特征提取方法被成功地运用.然而通过分析得知,该方法在处理原始输入空间上的特征提取问题时一定程度上缺乏适应性.因此本文在 MMD 准则的基础上,并结合已经被广泛研究和探讨的局部学习方法,提出一个新的评价度量:最大局部加权均值差异(Maximum Local Weighted Mean Discrepancy, MLMD),该度量反映源域和目标域分布差异时能充分考虑两个区域内在的局部结构,同时还能通过局部分布差异去反映全局分布差异.本文还在此度量的基础上提出一种能实现迁移学习任务并具有一定局部学习能力的特征提取方法:最大局部加权均值差异嵌入(Maximum Local Weighted Mean Discrepancy Embedding,MWME).该方法不但能完成传统意义上的特征提取,同时还能完成在两个分布存在差异但相关的两个区域上实现领域适应学习,从而表明该特征提取方法具有较好的鲁棒性和适应性.实验证明 MLMD 准则和 MWME 方法具有上述优势.

关键词: 最大均值差异嵌入; 最大局部均值差异; 最大局部加权均值差异嵌入; 特征提取; 迁移学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2013) 08-1462-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.08.002

Maximum Local Weighted Mean Discrepancy Embedding

GAO Jun[1,2], HUANG Li-li[3]
(1. *Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu* 215006, *China*
2. *School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu* 224001, *China*
3. *School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, Anhui* 232001, *China*)

Abstract: MMDE, regarded as a MMD-based feature extraction method, has been successfully used. However, when the feature extraction problems of the original input space have been solved, the MMDE lacks the suitability to some extent. Therefore, we propose Maximum Local Weighted Mean Discrepancy(MLMD) by integrating the theory and technique of local learning methods. The measurement considers fully the internal local structure between domains; at the same time, the global distribution discrepancy can be reflected by the local distribution discrepancy. We also, based on the above measurement, propose Maximum Local Weighted Mean Discrepancy Embedding(MWME), which not only fulfills transfer learning task but also has certain local learning capability. The MWME can complete traditional feature extraction as well as domain adaptation learning in two domains whose distributions are different but relative, thus indicating its better robustness and adaptation. Tests show the above-proposed advantages of the MLMD criterion and the MWME method.

Key words: maximum mean discrepancy embedding; maximum local weighted mean discrepancy; maximum local weighted mean discrepancy embedding; feature extraction; transfer learning

1 引言

在处理具有明显高维特征的数据时,我们往往首先需要使用特征提取技术对待识别数据进行预处理,以提高相应识别方法的精度和效率.所谓特征提取就是将高维特征空间通过变换转化为相应的低维表示空间[1],而该低维空间要尽可能地保持原有空间的判别信息.传统的特征提取技术,比如主成分分析(Principal Component Analysis, PCA)[2],线性判别分析(Linear Discrimination Analysis, LDA)[3]和局部保持投影(Locality Preserving Pro

收稿日期:2012-10-11;修回日期:2013-01-03;责任编辑:马兰英
基金项目:国家自然科学基金(No.61272210);江苏省自然科学基金(No.BK2011417);苏州大学江苏省计算机信息处理技术重点实验室开放课题(No.KJS1126);江苏省新型环保重点实验室开放课题(No.AE201068);江苏省高校优秀中青年教师和校长境外研修计划

jections,LPP)[4]等,一般都基于同一个设想:即训练样本和测试样本是独立同分布(Identically and Independently Distributed,I.I.D)的.但随着迁移学习(Transfer Learning,TL)方法[5]的发展,一些传统的特征提取技术可以被推广用于处理一些非独立同分布(non-I.I.D)的识别数据.特别是 Pan 等人结合传统 PCA 和最大均值差异(Maximum Mean Discrepancy,MMD)[6],提出一种具有迁移学习能力的特征提取方法:最大均值差异嵌入(Maximum Mean Discrepancy Embedding,MMDE)[7],从而使得该方法一定程度上实现了通过在源域上所获得的知识去有效构造适合目标域的特征提取技术[8,9].

MMD 作为一种特殊的迁移学习方法:领域适应学习方法(Domain Adaptation Learning,DAL)的有效度量,它一定程度上可以表明两个分布不同但相关区域之间的分布差异.然而通过分析得知,MMD 度量中由于使用了源域和目标域的总体均值之差来表示两个区域的分布差异,而根据相应的统计学理论,总体均值一般反应的是样本空间总体的分布信息和全局结构信息.因此从这一层面上讲,MMDE 方法可以认为是一种全局方法,以致一定程度上忽视了样本空间内在的局部结构和局部信息.

因此,本文通过使用具有一定局部学习能力均值[10]概念并结合 MMD 度量中提出一种新度量:最大局部加权均值差异(Maximum Local Weighted Mean Discrepancy,MLMD),并依据该度量提出一种具有一定局部学习能力的领域适应特征提取方法:最大局部加权均值差异嵌入(Maximum Local Weighted Mean Discrepancy Embedding,MWME).与现有的方法比较,本文方法具有如下优势:

(1)提出了一种新度量:MLMD.该度量由于引入了具有局部学习能力的局部加权均值,使得该度量在反映源域和目标域分布差异时能充分考虑两个区域内在的局部分布差异,同时还能通过局部分布差异去反映全局分布差异,而且通过理论分析该度量还可以作为 MMD 度量的泛化形式.

(2)基于 MLMD 度量提出了一种具有局部学习能力和特征提取功能的领域适应学习方法:MWME.在本文中我们将首先给出该方法对应的线性形式:LMWME,并通过对线性形式的核化构造非线性形式:Ker-MWME.这样做可以在一定程度上避免使用类似于 MMDE 方法中的半定优化程序(Semi-Definite Program,SDP),从而降低了算法的时间和空间复杂度.

(3)通过在相应数据集上的扩展实验来表明本文的 MLMD 度量和 MWME 方法具有的有效性.

2　相关工作

为了便于描述本文方法,在本节我们简单回顾一下 MMD 度量和 MMDE 方法.

2.1　最大均值差异:MMD

MMD 作为一种无参度量被现有的领域适应方法广泛运用的主要原因在于该度量计算简单、含义直观.

定义 1　(MMD)[6].假设分别存在一个满足分布为 P 源域 $D_s=\{x_{si}\}|_{i=1}^{n_s}$ 和一个满足分布为 Ψ 目标域 $D_t=\{z_{tj}\}|_{j=1}^{n_t}$,则在 RKHS 中源域 D_s 与目标域 D_t 的 MMD 可以用表示为:

$$\mathrm{dist}^2(D_s,D_t)=\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\varphi(x_{si})-\frac{1}{n_t}\sum_{j=1}^{n_t}\varphi(z_{tj})\right\|_H^2 \tag{1}$$

其中:$\varphi(\cdot)$是一个从原始输入空间到高维 Hilbert 空间 H 上的非线性映射.

从式(1)看出,MMD 度量就是使用源域数据集和与目标域数据集的总体均值之差来表示源域与目标域之间的分布差异.

2.2　最大均值差异嵌入:MMDE

基于 MMD,Pan 等人提出了 MMDE.

定义 2　(MMDE)[7].假设 $D_s'=\psi(D_s)$ 和 $D_t'=\psi(D_t)$分别表示源域 D_s 与目标域 D_t 对应的低维嵌入子空间,则 MMDE 方法的目标函数为:

$$\arg\min_{\varphi,\psi}\mathrm{dist}^2(D_s',D_t')$$

$$=\arg\min_{\varphi,\psi}\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\varphi(\psi(x_{si}))-\frac{1}{n_t}\sum_{j=1}^{n_t}\varphi(\psi(z_{tj}))\right\|_H^2 \tag{2}$$

如果 RKHS 中选取合适的核函数 k,则式(2)可以转化为下式的最小化:

$$\mathrm{dist}^2(D_s',D_t')=tr(\boldsymbol{KL}) \tag{3}$$

其中:$\boldsymbol{K}=\begin{pmatrix}\boldsymbol{K}_{ss} & \boldsymbol{K}_{st}\\ \boldsymbol{K}_{ts} & \boldsymbol{K}_{tt}\end{pmatrix}\in\boldsymbol{R}^{(n_s+n_t)\times(n_s+n_t)}$,$\boldsymbol{K}_{ss}$、$\boldsymbol{K}_{tt}$和 $\boldsymbol{K}_{st}$,

$$l_{ij}=\begin{cases}1/n_s^2, & x_{si},x_{sj}\in D_s\\ 1/n_t^2, & z_{ti},z_{tj}\in D_s\\ -1/n_sn_t, & \text{otherwise}\end{cases},\ \boldsymbol{L}=(l_{ij})_{(n_s+n_t)\times(n_s+n_t)},$$

是分别定义在源域、目标域和跨域(across-domain)的核矩阵,$tr(\boldsymbol{A})$表示矩阵 $\boldsymbol{A}$ 的迹运算.

由于式(3)并不能被直接用来求解得到满足式(2)的特征投影映射,为此 MMDE 方法使用了 SDP 优化程序,从而得到源域和目标域对应的低维嵌入子空间.

正是由于 MMD 是一种全局度量,一定程度上使得 MMDE 方法存在局部学习能力不足的问题,同时 MMDE 方法中由于需要使用 SDP 优化,这也一定程度地降低了该方法的执行效率,为此,本文提出具有一定局部学习能力的 MLMD 度量,并在此基础上提出 MWME 特征提取方法.

3　最大局部加权均值差异嵌入:MWME

3.1　最大局部均值差异:MLMD

根据以上流形学习理论,我们知道对于任意一个

非高斯或流形分布的数据可以被分成若干个分块,而每一个分块可以被认为是呈现局部高斯分布的.然而在实际使用该理论时还需要解决两个问题:(1)如何对非高斯分布的区域进行有效划分,产生相应的局部分块;(2)如何有效构造每一个样本所对应的局部权值.幸运的是,文献[11]中的方法尽管是用来处理传统意义上的识别问题,但也为我们提供一个思路.

定义 3 假设 $D_1=\{x_{1q}\}\big|_{i=1}^{N}$是一任意分布的区域,对于$\forall x_{1q}\in D_1$,则称 x_{1q}对应的 k 个近邻样本组成的局部区域 $D_{1q}=\{x_{1q}^{(c)}\}\big|_{c=1}^{k}\in D_1$ 为区域 D_1 上的局部分块.

同时如果存在一个投影映射 ψ 且令 $D_{1q}^{'}=\psi(D_{1q})$,那么 D_{1q},$D_{1q}^{'}$的局部加权均值可以分别被写为:

$$\sum_{c=1}^{k}\frac{\beta_{1q}^{(c)}x_{1q}^{(c)}}{\sum_{p=1}^{k}\beta_{1q}^{(p)}},\ \sum_{c=1}^{k}\frac{\beta_{1q}^{(c)}\psi(x_{1q}^{(c)})}{\sum_{p=1}^{k}\beta_{1q}^{(p)}},$$

其中 $\beta_{1q}^{(c)}=\exp\left(-\frac{\|x_{1q}-x_{1q}^{(c)}\|^2}{h}\right)$分别表示样本 $x_{1q}^{(c)}$在局部分块 D_{1q}上的权值,h 是热核函数 $\exp\left(-\frac{d^2}{h}\right)$上的热核参数.

定义 4 假设 $D_1=\{D_{1q}\}|_{q=1}^{N}$和 $D_2=\{D_{2d}\}|_{d=1}^{M}$是两个区域,其中 $D_{1q}=\{v_{1q}^{(c_1)}\}|_{c_1=1}^{k_1}$和 $D_{2d}=\{u_{2d}^{(c_2)}\}\big|_{c_2=1}^{k_2}$分别为上述两区域任意的两个局部分块,对于$\forall D_{1q}$如果存在一个 $D_{2d'}\in D_2$ 使得下式成立,那么就称 $D_{2d'}$是 D_{1q}在区域 D_2 上的局部最近邻分块(Nearest Local Patch,NLP).

$$\begin{aligned}\mathrm{dist}_{\mathrm{NLP}}(D_{1q},D_{2d'})&=\left\|\sum_{c_1=1}^{k_1}\frac{\beta_{1q}^{(c_1)}v_{1q}^{(c_1)}}{\sum_{p_1=1}^{k_1}\beta_{1q}^{(p_1)}}-\sum_{c_2=1}^{k_2}\frac{\beta_{2d'}^{(c_2)}u_{2d'}^{(c_2)}}{\sum_{p_2=1}^{k_2}\beta_{2d'}^{(p_2)}}\right\|\\&=\min_{d=1,\cdots,M}\left\|\sum_{c_1=1}^{k_1}\frac{\beta_{1q}^{(c_1)}v_{1q}^{(c_1)}}{\sum_{p_1=1}^{k_1}\beta_{1q}^{(p_1)}}-\sum_{c_2=1}^{k_2}\frac{\beta_{2d}^{(c_2)}u_{2d}^{(c_2)}}{\sum_{p_2=1}^{k_2}\beta_{2d}^{(p_2)}}\right\|\end{aligned}\tag{4}$$

根据上述定义,我们给出本文的最大局部均值差异:MLMD.

定义 5 假设 $D_1=\{D_{1q}\}|_{q=1}^{N}$和 $D_2=\{D_{2d}\}|_{d=1}^{M}$是具有一定分布差异的两个区域,其中 $D_{1q}=\{v_{1q}{}^{(c_1)}\}|_{c_1=1}^{k_1}$和 $D_{2d}=\{u_{2d}{}^{(c_2)}\}|_{c_2=1}^{k_2}$分别为上述两区域任意的两个局部分块,如果存在一个投影映射 ψ 使得 $D_1^{'}=\{\psi(D_{1q})\}|_{q=1}^{N}$和 $D_2^{'}=\{\psi(D_{2d})\}|_{d=1}^{M}$,则 MLMD 可以表示为:

$$\begin{aligned}&\mathrm{dist}_{\mathrm{MLMD}}{}^2(D_1^{'},D_2^{'})\\&=\sum_{q=1}^{N}\sum_{d=1}^{M}\gamma_{qd}\left\|\sum_{c_1=1}^{k_1}\frac{\beta_{1q}{}^{(c_1)}\varphi(\psi(v_{1q}{}^{(c_1)}))}{\sum_{p_1=1}^{k_1}\beta_{1q}{}^{(p_1)}}\right.\\&\left.-\sum_{c_2=1}^{k_2}\frac{\beta_{2d}{}^{(c_2)}\varphi(\psi(u_{2d}{}^{(c_2)}))}{\sum_{p_2=1}^{k_2}\beta_{2d}{}^{(p_2)}}\right\|_H^2\end{aligned}\tag{5}$$

其中 $\gamma_{ij}=\begin{cases}1, & D_{1q}\text{是}D_{2d}\text{在}D_1\text{中的 NLP 或者}\\ & D_{2d}\text{是}D_{1q}\text{在}D_2\text{中的 NLP}\\0, & \text{否则}\end{cases}$

为两个区域上局部分块之间的关联系数.

如果对 MLMD 通过适当地简化(比如令 $N=1$,$M=1$,所有权值赋值为 1),则可以退化为标准的 MMD,因此从这一层面上讲,MLMD 可以看成是 MMD 度量的泛化形式.然而,我们知道在如果直接使用类似于 MMDE 方法去求解等式(5),那么必须使用 SDP 优化程序($O(n_s+n_t)^{6.5}$),这就使得该方法在处理大容量数据(比如文本数据)时不具有适应性,因此,本文使用一些非线性特征提取方法[4,12,13]构造技巧,首先通过对等式(5)进行线性化方法:LMWME,然后在 LMWME 基础上使用 Represent theory 提出非线性核化的最大局部加权均值嵌入方法:Ker-MWME,从而一定程度上避免了使用 SDP 和迭代优化方法带来的不足.

3.2 线性最大局部加权均值嵌入:LMWME

同非线性领域适应方法相比,线性方法在解决一些非线性问题时一定程度上缺乏更好模式识别效果,但作为一种简单、有效的技术在领域适应学习方法中时常被讨论[14].

定义 6 假设 $D_s=\{D_{si}\}\big|_{i=1}^{n_s}$,$D_t=\{D_{tj}\}\big|_{j=1}^{n_t}$分别表示源域和目标域,其中 D_{si},D_{tj}分别是样本 $\boldsymbol{x}_{si}=(x_{si_1},\cdots,x_{si_n})^{\mathrm{T}}\in D_s$,$\boldsymbol{z}_{tj}=(z_{tj_1},\cdots,z_{tj_n})^{\mathrm{T}}\in D_t$ 所对应的局部分块,同时令 D_s,D_t 的嵌入子空间分别为 $D_s^{'}$,$D_t^{'}$,则根据式(5),LMWME 的目标函数可以写为:

$$\begin{aligned}&\arg\min_{\boldsymbol{\omega}\,\boldsymbol{\omega}=1}\mathrm{dist}^2_{\text{Linear-MLMD}}(D_s^{'},D_t^{'})\\&=\arg\min_{\boldsymbol{\omega}\,\boldsymbol{\omega}=1}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\gamma_{ij}\left\|\sum_{c_1=1}^{k_1}\frac{\beta_{si}{}^{(c_1)}(\boldsymbol{\omega}^{\mathrm{T}}x_{si}{}^{(c_1)})}{\sum_{p_1=1}^{k_1}\beta_{si}{}^{(p_1)}}\right.\\&\left.-\sum_{c_2=1}^{k_2}\frac{\beta_{tj}{}^{(c_2)}(\boldsymbol{\omega}^{\mathrm{T}}z_{tj}{}^{(c_2)})}{\sum_{p_2=1}^{k_2}\beta_{tj}{}^{(p_2)}}\right\|^2\end{aligned}\tag{6}$$

其中 $\gamma_{ij}=\begin{cases}1, & D_{si}\text{是}D_{tj}\text{在}D_s\text{中的 NLP 或者}\\ & D_{tj}\text{是}D_{si}\text{在}D_t\text{中的 NLP}\\0, & \text{否则}\end{cases}$

为局部分块关联系数,$\boldsymbol{\omega}\in\mathcal{R}^{n\times 1}$是投影变换矢量.

定理 1 等式(6)所对应的 LMWME 方法的目标函数可以简化为如下形式:

$$\arg\min_{\boldsymbol{\omega}\,\boldsymbol{\omega}=1}\mathrm{dist}^2_{\text{Linear-MLMD}}(D_s^{'},D_t^{'})=\arg\min_{\boldsymbol{\omega}\,\boldsymbol{\omega}=1}\mathrm{tr}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\omega})\tag{7}$$

其中:$\boldsymbol{X}=D_s\cup D_t$ 为样本集,$\boldsymbol{L}=\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\boldsymbol{R}_{ij}\boldsymbol{L}_{ij}$为全局分布差异权矩阵.

证明　如果令 $\boldsymbol{X}=D_s\cup D_t$ 为样本集,对于任意 D_{si},D_{tj},我们将定义在上述局部分块上的权值扩充到整个数据集 $\boldsymbol{X}$,则有:

$$\boldsymbol{\beta}_{si}=(\beta_{si}^{(1)}/\sum_{p_1=1}^{n_s}\beta_{si}^{(p_1)},\cdots,\beta_{si}^{(n_s)}/\sum_{p_1=1}^{n_s}\beta_{si}^{(p_1)},0,\cdots,0)^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{\beta}_{tj}=(0,\cdots,0,\beta_{tj}^{(1)}/\sum_{p_2}^{n_t}\beta_{tj}^{(p_2)},\cdots,\beta_{tj}^{(n_t)}/\sum_{p_2}^{n_t}\beta_{tj}^{(p_2)})^{\mathrm{T}} \tag{9}$$

其中 $\beta_{tj}^{(c)}=\begin{cases}\exp(\dfrac{\|x_{si}^{(c)}-x_{si}\|^2}{h_1}), & x_{si}^{(c)}\in D_{si}\\ 0, & \text{否则}\end{cases}$,

$\beta_{tj}^{(c)}=\begin{cases}\exp(\dfrac{\|z_{tj}^{(c)}-z_{tj}\|^2}{h_2}), & z_{tj}^{(c)}\in D_{tj}\\ 0, & \text{否则}\end{cases}$,

由此,式(6)中的 $\sum_{c_1=1}^{k_1}\dfrac{\beta_{si}^{(c_1)}(\boldsymbol{\omega}^{\mathrm{T}}x_{si}^{(c_1)})}{\sum_{p_1=1}^{k_1}\beta_{si}^{(p_1)}}$, $\sum_{c_2=1}^{k_2}\dfrac{\beta_{tj}^{(c_2)}(\boldsymbol{\omega}^{\mathrm{T}}z_{tj}^{(c_2)})}{\sum_{p_2=1}^{k_2}\beta_{tj}^{(p_2)}}$

可分别被改写成 $\boldsymbol{\beta}_{si}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{\omega}$,$\boldsymbol{\beta}_{tj}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{\omega}$,则式(6)可以转化为:

$$\begin{aligned}&\arg\min_{\omega'\omega=1}\mathrm{dist}^2_{\text{Linear-MLMD}}(D_s',D_t')\\&=\arg\min_{\omega'\omega=1}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\gamma_{ij}\|\boldsymbol{\beta}_{si}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{\omega}-\boldsymbol{\beta}_{tj}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{\omega}\|^2\\&==tr(\boldsymbol{\omega}^{\mathrm{T}}X\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\gamma_{ij}(\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{si}^{\mathrm{T}}+\boldsymbol{\beta}_{tj}\boldsymbol{\beta}_{tj}^{\mathrm{T}}-2\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{tj}^{\mathrm{T}})X^{\mathrm{T}}\omega)\end{aligned} \tag{10}$$

如果令 $\boldsymbol{L}_{ij}=\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{si}^{\mathrm{T}}+\boldsymbol{\beta}_{tj}\boldsymbol{\beta}_{tj}^{\mathrm{T}}-2\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{tj}^{\mathrm{T}}$为局部分块权值矩阵,$\boldsymbol{R}_{ij}=\mathrm{diag}(\underbrace{r_{ij},\cdots,r_{ij}}_{n_s+n_t})$ 为局部分块关联系数矩阵.如果令全局的权值矩阵 $\boldsymbol{L}=\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}R_{ij}L_{ij}$,则定理得证.

根据定理 1,判断局部近邻分块的式(4)可以简化为:

$$\mathrm{dist}_{NLP}(D_{1q},D_{2d'})=\min_{d=1,\cdots,M}tr(\boldsymbol{X}\boldsymbol{L}_{qd}\boldsymbol{X}^{\mathrm{T}})^{1/2} \tag{11}$$

其中 $\boldsymbol{L}_{qd}=\boldsymbol{\beta}_{1q}\boldsymbol{\beta}_{1q}^{\mathrm{T}}+\boldsymbol{\beta}_{2d}\boldsymbol{\beta}_{2d}^{\mathrm{T}}-2\boldsymbol{\beta}_{1q}\boldsymbol{\beta}_{2d}^{\mathrm{T}}$.

类似于其它方法,LMWME 方法的关键就是要求解矩阵 $\boldsymbol{XLX}^{\mathrm{T}}$ 的 l 个最小非 0 特征值所对应的单位特征向量.然而当原始样本空间是高维空间时,求解上述矩阵的特征值有 $O(n^3)$时间复杂度,为此将 QR 分解方法引入到 LMWME 方法中以降低该方法时间复杂度.

根据 QR 分解的基本原理,可以得到 $\boldsymbol{X}=\boldsymbol{QR}$,其中 $\boldsymbol{Q}\in\mathcal{R}^{n\times r}$由一组正交列向量组成,$\boldsymbol{R}\in\mathcal{R}^{r\times(n_s+n_t)}$是上三角矩阵,$r=\mathrm{rank}(\boldsymbol{X})$.则求解式(7)转化为求解下式:

$$\begin{aligned}&\arg\min_{a'a=1}\mathrm{dist}^2_{\text{Linear-MLMD}}(D_s',D_t')\\&=\arg\min_{a'a=1}tr(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{L}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{a})\end{aligned} \tag{12}$$

并且(7)中的 $\boldsymbol{\omega}$ 与(12)中的 $\boldsymbol{a}$ 关系为:$\boldsymbol{\omega}=\boldsymbol{Qa}$.

上述 QR 分解方法表明求解矩阵 $\boldsymbol{XLX}^{\mathrm{T}}$ l 个最小非 0 特征值所对应的正交单位特征向量转换成求解矩阵 $\boldsymbol{RLR}^{\mathrm{T}}$ l 个最小非 0 特征值所对应的正交单位特征向量,这样做的好处就是使得原方法的时间复杂度降低为 $O(r^3)$,特别相对于高维小样本数据更有优势.

由此,根据上述的定义和性质,我们得到线性最大局部加权均值差异嵌入算法.

算法:LMWME

Input:源域 D_s 和目标域 D_t,且令 $\boldsymbol{X}=D_s\cup D_t$,热核参数 h_1 和 h_2,样本 k-NN 参数 k_1 和 k_2;

Output:样本集 $\boldsymbol{X}$ 的线性低维嵌入集 $\boldsymbol{Y}$;

Step1:根据定义 3 构造 D_s 和 D_t 的局部分块;

Step2:对于 $\forall D_{si}\in D_s$ 和 $\forall D_{tj}\in D_t$:

　Step2.1:分别使用式(8)(9)计算相对于数据集 X 权值 $\boldsymbol{\beta}_{si}$ 和 $\boldsymbol{\beta}_{tj}$;

　Step2.2:构造局部区域 D_{si} 和 D_{tj} 的 $L_{ij}=\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{si}^{\mathrm{T}}+\boldsymbol{\beta}_{tj}\boldsymbol{\beta}_{tj}^{\mathrm{T}}-2\boldsymbol{\beta}_{si}\boldsymbol{\beta}_{tj}^{\mathrm{T}}$

　Step2.3:根据式(11)分别计算 D_{si} 在 D_t 中和 D_{tj} 在 D_s 中的局部最近邻局部分块,并计算矩阵 $R_{ij}=\mathrm{diag}(\underbrace{r_{ij},\cdots,r_{ij}}_{n_s+n_t})$;

Step3:计算矩阵 $\boldsymbol{L}=\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\boldsymbol{R}_{ij}\boldsymbol{L}_{ij}$;

Step4:根据 QR 分解的基本原理,计算得到 $\boldsymbol{X}=\boldsymbol{QR}$;

Step5:求解式(12)并计算得到由 l 个最小非 0 特征值所对应的正交单位特征向量组成的变换矩阵 $\boldsymbol{A}=(a_1,\cdots,a_l)$;

Step6:使用公式 $\boldsymbol{W}=\boldsymbol{QA}$ 得到变换矩阵 $\boldsymbol{w}$;

Step7:根据 $\boldsymbol{Y}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ 计算得到 X 的低维嵌入集 $\boldsymbol{Y}$.

3.3　非线性最大局部加权均值嵌入方法:Ker-MWME

猜想存在一个非线性特征映射 φ 将原始输入空间投影到 RKHS,那么原始输入样本集 $\boldsymbol{X}$ 可以通过非线性映射 φ 被表示为 $\varphi(\boldsymbol{X})$,则在式(7)的基础上,Ker-MWME 的目标函数可以写为:

$$\arg\min_{\omega^{\varphi}\omega^{\varphi}=1}tr(\boldsymbol{\omega}^{\varphi\mathrm{T}}\varphi(\boldsymbol{X})\boldsymbol{L}\varphi(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\omega}^{\varphi}) \tag{13}$$

根据 Represent theory,$\boldsymbol{\omega}^{\varphi}=\varphi(\boldsymbol{X})\boldsymbol{\alpha}$,其中 $\boldsymbol{\alpha}=(\alpha_{s1},\cdots,\alpha_{sn_s},\alpha_{t1},\cdots,\alpha_{tn_t})^{\mathrm{T}}\in\mathcal{R}^{n_s+n_t}$是由 n_s+n_t 个系数组成的列向量,同时存在一个核函数 $k(v_i,v_j)=(\boldsymbol{\varphi}(v_i),\boldsymbol{\varphi}(v_j))=\boldsymbol{\varphi}(v_i)^{\mathrm{T}}\boldsymbol{\varphi}(v_j)$,则等式(13)可以改写为:

$$\arg\min_{\alpha^{\mathrm{T}}K\alpha=1} tr(\alpha^{\mathrm{T}}KLK\alpha) \tag{14}$$

其中 $\boldsymbol{K}=\begin{pmatrix}\boldsymbol{K}_{ss} & \boldsymbol{K}_{st}\\ \boldsymbol{K}_{ts} & \boldsymbol{K}_{tt}\end{pmatrix}\in\boldsymbol{R}^{(n_s+n_t)\times(n_s+n_t)}$，$\boldsymbol{K}_{ss}$、$\boldsymbol{K}_{tt}$ 和 $\boldsymbol{K}_{st}$ 是分别定义在源域、目标域和跨域(across-domain)的核矩阵.

假设列向量 $\boldsymbol{\alpha}^1,\cdots,\boldsymbol{\alpha}^l$ 是满足等式(14)的解，同时令 $\boldsymbol{\Lambda}=\{\boldsymbol{\alpha}^1,\cdots,\boldsymbol{\alpha}^l\}$，则输入样本集 $\boldsymbol{X}$ 对应的非线性嵌入子空间可以表示为：$\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{K}$.

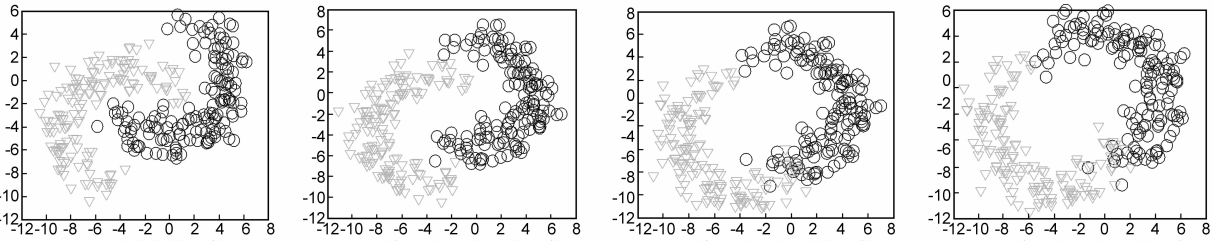
4 实验

为了说明本文方法的有效性，我们使用 two-moons 数据集来进行测试，同时结合与其它相似的特征提取方法进行比较来说明本文方法的优越性.

通过测试 two-moons 数据集来说明本文方法具有较强的局部学习能力；测试两个真实的高维文本数据集来说明本文方法在处理实际问题时所具有的特征提取的能力；通过测试人脸数据集来说明本文方法处理多类数据的特征提取能力，同时使用图像表明提取结果对分类精度的影响. 本测试过程在 Intel Core2，2.0GHz 主频，2G RAM，Vista 系统，Matlab2007 平台上实现.

4.1 测试准备

人造 two-moon 数据集具有典型的局部流形结构，因此经常被用来测试相应方法的局部保持功能. 在本阶段为了测试本文的线性 LMWME 和非线性 Ker-MWME 方法在实现迁移学习的同时还尽可能的保持样本的局部结构，我们使用一个包含有 300 样本的 two-moon 数据集作为源域，该数据集可以分为正、负 2 类，每一类包含有 150 个样本，见图 1(a). 将源域数据分别按逆时针旋转 10 次，可以得到 10 个分布不同但相关的目标域. 图 1(b)、图 1(c)和图 1(d)分布表示旋转 30°、60° 和 80° 得到的目标域. 可以从图 2 看出，当旋转的角度越大，所产生的目标域与源域的分布差异越大，从而说明相对应的领域适应问题就越复杂. 为此，我们在该测试过程中使用四种经典的特征提取方法：LDA、PCA、LPP 和 LMMDE(MMDE 的线性核函数形式)与本文的线性方法 LMWME 进行测试比较. 同时也使用 KLDA、KPCA、KLPP 和 MMDE 与本文非线性方法 Ker-MWME 进行测试比较.



(a) 源域中的数据集 (b) 旋转30°的目标域数据集 (c) 旋转60°的目标域数据集 (d) 旋转80°的目标域数据集

图1 基于two-moon型数据集构造的源域和目标域

由于该实验过程需要测试两种方法，因此，本测试实验可以设计为：

(1)在测试比较 LMWME 方法的过程中，我们令本文 LMWME 方法中的热核参数 $h_1=h_2=[2^{(-10)},2^{(-8)},2^{(-6)},\cdots,2^0,2^2,\cdots,2^6,2^8,2^{10}]$，近邻样本数参数 $k_1=k_2=[2,4,6,8]$. LDA、PCA 和 LMMDE 由于是无参方法，不需要设定参数，而 LPP 方法中的热核参数 h 和近邻样本数参数 k 设定值范围等同于本文 LMWME 方法中的热核参数和近邻样本数参数.

(2)在测试比较本文的非线性 Ker-MWME 方法的过程中，所有的测试比较方法都使用高斯核函数，其中高斯核函数的带宽 σ 统一设定为训练样本平均范数的平方根[15]. 同时令 Ker-MWME 方法中的热核参数 h_1，h_2、近邻样本参数 k_1，k_2、KLPP 方法中的热核参数 h 和近邻样本数参数 k 设定值范围等同于上述测试线性方法过程中参数的设定范围.

(3)该测试过程使用 5 次 10-折交叉验证，使用平均精度加方差的形式来度量测试方法的性能. 所有测试方法的都使用 1-NN 分类器得到测试精度. 实验结果见表 1、表 2、图 2.

4.2 实验结果与分析

根据上述实验结果，我们可以得到如下结论：

(1)从表 1、表 2 和图 2 可以看出，随着旋转的角度增大，5 种测试方法对应的特征提取效果随之降低，这一趋势是符合实际的. 因为随着旋转角度的变化，目标域的复杂差异程度也发生了变化，旋转角度增大，目标域与源域的分布差异就越大，从而导致了算法的适应性变差. 同时我们还可以看出，LDA、PCA 和 LPP 方法在处理分布差异较大的目标域时表现出适应性不强的现象，因此从这一层面上讲，上述三种方法更适合源域和目标域具有同分布的特征提取问题.

(2)根据上述结果，我们也可以得到另外一个结论，即在处理具有明显局部流形结构的数据时，当目标域和源域分布差异不大的情况下，LMMDE/MMDE 与 LPP/KLPP 相比一定程度上没有明显的优势，比如 LPP 与 LMMDE 相比，当源域数据集旋转 20° 以内时，LPP 的特征提取效果还略好于 LMMDE，而 KLPP 则表现为不差于 MMDE. 这样现象说明在目标域和源域分布差异不大

的情况下,由于 LPP 方法更注重于目标域区域中样本内在的局部流形结构,从而使得该方法具有较好的特征提取效果.因此,可以说明一个问题,那就是 MMDE 方法使用总体均值去反映源域和目标域的分布差异时,仅仅反映了源域样本和目标域样本之间的全局结构差异,而一定程度上没有考虑两个域之间的局部结构差异.这一点充分说明了提出本文方法的合理性.

表 1　5 种线性方法对 10 种不同分布的 two-moon 数据集的测试比较(mean ± std)

Target Domain (rotation angle)	10°	15°	20°	25°	30°	40°	50°	60°	70°	80°
Algorithm	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
LDA	0.8625 ± 0.03576	0.8429 ± 0.02853	0.8385 ± 0.04012	0.7917 ± 0.05172	0.7728 ± 0.04518	0.7356 ± 0.02368	0.6871 ± 0.01964	0.6319 ± 0.03027	0.4881 ± 0.04291	0.3538 ± 0.02651
PCA	0.8617 ± 0.01032	0.8310 ± 0.03271	0.8263 ± 0.0117	0.7901 ± 0.0132	0.7529 ± 0.0093	0.7120 ± 0.01002	0.6733 ± 0.01329	0.6238 ± 0.0082	0.4679 ± 0.01897	0.3201 ± 0.0172
LPP	0.9031 ± 0.05824	0.8772 ± 0.04819	0.8533 ± 0.04624	0.8393 ± 0.06201	0.8059 ± 0.01327	0.7858 ± 0.03921	0.7429 ± 0.02804	0.6987 ± 0.04181	0.5067 ± 0.03104	0.4787 ± 0.05824
LMMDE	0.9023 ± 0.02067	0.8691 ± 0.03645	0.8521 ± 0.03892	0.8319 ± 0.06201	0.8217 ± 0.03182	0.8103 ± 0.00197	0.7881 ± 0.03923	0.7134 ± 0.0291	0.6537 ± 0.02018	0.508 ± 0.02183
LMWME	0.9681 ± 0.01749	0.9542 ± 0.02985	0.9315 ± 0.01258	0.9057 ± 0.02017	0.8816 ± 0.02636	0.8672 ± 0.03921	0.8397 ± 0.03182	0.7865 ± 0.02519	0.6819 ± 0.03104	0.5782 ± 0.03246

表 2　5 种非线性方法对 10 种不同分布的 two-moon 数据集的测试比较(mean ± std)

Target Domain (rotation angle)	10°	15°	20°	25°	30°	40°	50°	60°	70°	80°
Algorithm	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
KLDA	1 ± 0	0.9867 ± 0.0028	0.9639 ± 0.02018	0.9368 ± 0.03819	0.9173 ± 0.03025	0.8605 ± 0.01406	0.7622 ± 0.0412	0.7109 ± 0.02945	0.6397 ± 0.04902	0.6138 ± 0.03209
KPCA	1 ± 0	0.9751 ± 0.0043	0.9557 ± 0.01982	0.9140 ± 0.02986	0.9082 ± 0.0201	0.8230 ± 0.0056	0.7369 ± 0.0129	0.7018 ± 0.0191	0.6385 ± 0.0063	0.6001 ± 0.0058
KLPP	1 ± 0	1 ± 0	0.9974 ± 0.0013	0.9767 ± 0.02118	0.9524 ± 0.01002	0.8784 ± 0.03928	0.8237 ± 0.02918	0.7643 ± 0.02874	0.7442 ± 0.04413	0.6987 ± 0.04738
MMDE	1 ± 0	1 ± 0	0.9928 ± 0.0052	0.9857 ± 0.0089	0.9802 ± 0.02411	0.9789 ± 0.0109	0.9746 ± 0.0039	0.8559 ± 0.03957	0.7537 ± 0.02320	0.7408 ± 0.01805
Ker-MWME	1 ± 0	1 ± 0	1 ± 0	1 ± 0	0.9910 ± 0.0039	0.9847 ± 0.0181	0.9820 ± 0.0091	0.8701 ± 0.02989	0.799 ± 0.02835	0.7682 ± 0.03976

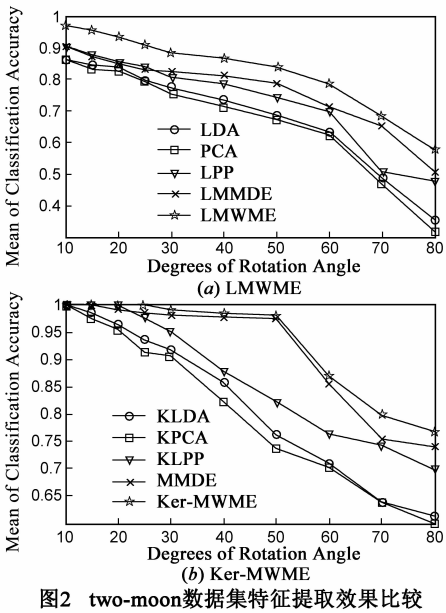


图2　two-moon数据集特征提取效果比较

5　总结

本文通过分析 MMDE 方法在处理线性问题时存在的不足,通过使用传统局部学习方法的理论和技术并结合 MMD 度量,提出一种具有局部学习能力的新度量:MLMD,并在该度量上提出一种能实现特征提取的领域适应方法:MWME.MLMD 度量在反映源域和目标域分布差异时能充分考虑两个区域内在的局部结构,同时还能通过局部分布差异去反映全局分布差异,从而使得 MWME 方法在一定成上提高了 MMDE 方法泛化能力和较强的局部学习能力.然而,本文所提度量和方法也存在着一些缺陷,比如如何进一步提高本文所提方法的执行效率将是我们以后急需解决的问题.

参考文献

[1] 边肇祺,张学工.模式识别[M].北京:清华大学出版社,2001.

Bian Z Q, Zhang X G. Pattern Recognition[M]. Beijing: TsingHua University Press, 2001. (In Chinese)
[2] Jolliffe I T. Principal Component Analysis[M]. New York : Springer-Verlag, 1986.
[3] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(2): 179 – 188.
[4] He X F, Niyogi P. Locality preserving projections[C/OL]. http://peples.cs.uchic ago.edu/xiaofei/LPP_NIPS03.pdf, 2003.
[5] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345 – 1359.
[6] Borgwardt K M, Gretton A, Rasch M J, Kriegel H P, Schöolkopf B, Smola A J. Integrating structured biological data by kernel maximum mean discrepancy [J]. Bioinformatics, 2006, 22(14): 49 – 57.
[7] Pan J L, Kwok J T, Yang Q. Transfer learning via dimensionality reduction [C/OL]. http://www.aaai.org/Papers/AAAI/2008/AAAI08 – 108.pdf.
[8] 王雪松,潘杰,程玉虎.基于知识迁移的 Ant-Q 算法[J].电子学报,2011,39(10):2359 – 2365.
Wang X S, Pan J, Cheng Y H. Ant-Q algorithm based knowledge transfer[J]. Acta Electronica Sinica, 2011, 39(10): 2359 – 2365. (in Chinese)
[9] 于重重,田蕊,谭励,涂序彦.非平衡样本分类的集成迁移学习算法[J].电子学报,2012,40(7):1358 – 1363.
Yu C C, Tian H, Tan L, Tu X Y. Integrated transfer larning algorithmic for unbalance samples classification[J]. Acta Electronica Sinica, 2012, 40(7): 1358 – 1363. (in Chinese)
[10] Christopher G. Atkeson, Andrew W. Moore, Stefan Schaal. Locally weighted learning[J]. Artifical Intelligence Review, 1997, 11(1 – 5): 11 – 73.
[11] Zhao D L, Lin Z C, Xiao R, Tang X O. Linear Laplacian discrimination for feature extraction[C/OL]. http://resear ch.microsoft.com/en-us/um/people/zhoulin/publications/2007-cvpr-lld.pdf.
[12] Li J, Li X L, Tao D C. KPCA for semantic object extraction in images[J]. Pattern Recognition, 2008, 41(10): 3244 – 3250.
[13] Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach[J]. Neural Computation, 2000, 12(10): 2385 – 2404.
[14] Wang Y Y, Chen S C, Zhou Z H. New semi-supervised classification method based on modified cluster assumption[J]. IEEE Transcation Neural Network and Learning System, 2012, 23(5): 689 – 702.
[15] Tao J W, Chung F L, Wang S T. On minimum distribution discrepancy support vector machine for domain adaptation[J]. Pattern Recognition, 2012, 45(11): 3962 – 3984.

作者简介

皋 军 男,1971 年生于江苏阜宁,博士,副教授,硕士生导师.主要研究方向:数据挖掘、人工智能、模式识别和模糊系统.
E-mail: gj0104211@163.com

黄丽莉 女,1986 年生于江苏苏州,安徽理工大学硕士研究生.主要研究方向:人工智能、模式识别.
E-mail: llhuang135@163.com