

机器视觉中基于界标的无人管理域自适应算法研究^①

曾远柔^{②*} 王红霞^{**}

(* 长江职业学院数据信息学院 武汉 430074)

(** 武汉理工大学计算机科学与技术学院 武汉 430070)

摘 要 针对域自适应算法需要所有源案例及目标案例的问题,提出一种基于界标的无人管理域自适应算法。首先,从源和目标(ST)中选择界标将源分布与目标分布间的映射最大化;然后,在选择好的界标上应用高斯核,以得到新的源点(KS)与新的目标点(KT);接着,在执行有关特征向量的子空间对齐之前,实行两个独立的主成分分析(PCA);最后,从被标注的源数据中研究分类器,并将其执行到目标域中。在图像域自适应的大量实验表明,所提方法选择出的界标能够降低领域之间的不一致性,用于非线性项目,能够呈现出有效子空间对齐的数据,优于其他无人管理域自适应算法。

关键词 界标,域自适应,核投影,子空间对齐,自适应算法

0 引 言

当标准机器学习提出训练集和测试集来源于相同的数据分布^[1]时,很多应用并不赞同这种说法,如机器视觉^[2]和自然语言处理等,因为分类器中的数据通常与要处理的数据不同。为了处理这种情况,可使用从源头到目标的转换,研究两种领域间的分布传输,而域自适应法^[3-5]中的两种主要类别都是可行的。

为此,很多研究者对其进行了研究。如文献[6]中显示域自适应法够接触到目标领域中的一些被标记过的例子以及数据。文献[7]集中研究统计机器翻译领域自适应问题,并提出解决此类问题的新方法,针对双语网站的识别和定位,提出一种基于全局搜索和局部分类的特定领域双语网站识别方法,实验结果表明,在相同测试集下,特定领域机器翻译系统的性能获得显著提升,验证了该方法的有效性。文献[8]研究了一种中间子空间的顺序,这种子空

间沿着测地线路径连接了源子空间和目标子空间。如文献[9]中源数据与目标数据都被输入中间线性子空间中,这种空间分布在链接两个原始空间的最短的测地线路径周围。然而,研究这些子空间的方法虽然有效,成本却很大,且会受到干扰。文献[10]围绕中文分词领域自适应的课题,针对大规模人工分词训练语料难以获得的问题,提出了基于主动学习的中文分词方法。该方法的主要思想是通过对目标领域文本的分析,选择出小规模含有语言知识较多的语料,进行人工分词标注,然后利用这些语料训练出目标领域的分词模型。从而达到通过标注少量语料,为大量语料服务的目的。Fernando 等人在文献[11]中通过优化单一线性绘图函数直接将源子集与目标子集连接起来。这种方法不仅被证明比最新的其他方法要好,而且在闭合形式下也是可计算的。

本文主要改进文献[11]的方法,但这面临着两个主要的问题。首先,该方法指出两种分布间的传输能够通过线性传输得以纠正,但这很容易被很多

① 湖北省自然科学基金(2013CFB351)资助项目。
② 男,1975年生,副教授,硕士;研究方向:图像处理,移动应用等;E-mail: 309077880@qq.com
(收稿日期:2018-01-21)

现实世界的应用所推翻。其次,此方法指出在实行适应时,需要所有的源案例及目标案例。然而,在大多数情况下,只有一个源数据的子集会与目标域分布相似,反之亦然。本文将用以下方式处理这两个问题:(1)从两种域中选择界标来减少源分布与目标分布的不一致性;(2)使用关于界标选择的高斯核函数将源数据与目标数据输入共享空间中,这使得从数据库中捕捉到非线性变得容易;(3)研究一个线性绘图函数将源子空间与目标子空间连接起来,这只需要简单地计算出源特征向量与目标特征向量之间的内积。通过实验证明本文所提出的方法优于当前的域自适应法。

1 拟定的域自适应法

本节将会详细介绍本文拟定的域自适应法。它是针对域自适应的以界标为基础的子空间对齐法,是完全无人监督的,因此在执行域自适应时,不需要任何标注。从源域中获取的被标注部分只用于随后

研究分类器。

源 S 和目标 T 被认为是分别从源分布 D_S 和目标分布 D_T 中获取而来。域自适应指出,源分布与目标分布是不同的,但它们也有一些相似之处,这使得将源域中研究所得放入目标域中成为可能。不同的是,如果有一套 L_S 的源案例,它们就能够用于研究适合目标域的分类器。

本文的方法是将两种观点结合起来,即源分布与目标分布的相似和不同之处。首先,将源案例与目标案例输入到有关选择好的界标的普通子空间中。接着,在两种域中运行子空间对齐。在 S 和 T 中选择出界标后,使用高斯核函数将其所有的点输入到界标中,用 K_S 和 K_T 重新展现了源点与目标点,并通过子空间对齐法完成了映射。

与文献[11]相比,本文的两步法在捕捉非线性时,既保持了准确性,操作又简单快捷。以下详细介绍从多尺度界标选择到子空间对齐和分类方法中的每一个步骤。整体工作流程如图1所示。

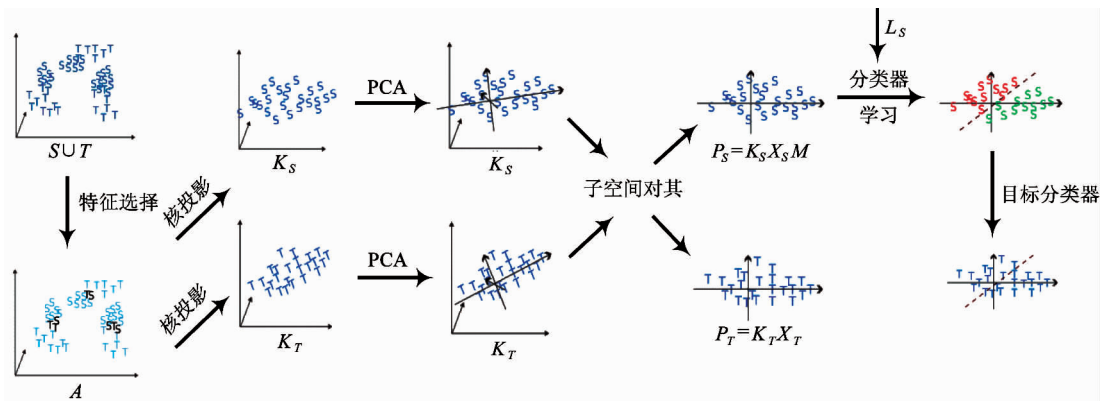


图1 本文方法的整体工作的工作流程

2 多尺度界标选择及核投影

本文方法的第一步就是选择一些点作为界标。直观来看,一套好的界标能够将源数据与目标数据输入到共享空间中,使得它们的分布更加相似。该方法从 S 和 T 中选择界标且未使用过任何其他标记。界标选择最终的输出如下: $A = \{\alpha_1, \alpha_2, \dots\}$, 其中 $A \subset S \cup T$, 为了避免昂贵的重复优化法,本文

提出了一种直接法,其能判断是否该保留某个点作为界标。

2.1 界标选择概述

实际上,本文的界标选择法把从 $S \cup T$ 的每一个 c 点都当作备选界标,并独立研究了每一个备选界标。对备选界标执行质量检测,如果检测高于阈值,就将其确定为界标。为了评估备选 C 的质量,首先要用高斯核基的标准误差 s , 计算其与 $p \in S \cup T$ 所有点的相似性,公式如下:

$$K(c, p) = \exp\left(\frac{-\|c - p\|^2}{2s^2}\right) \tag{1}$$

计算备选界标 C 的质量来作为源点与目标点中 K 值分布的重复。因此,在使用核基之后,如果源点与目标点的分布是相似的,那它就是一个好界标。

2.2 多尺度分析

式(1)中的核基半径值很重要,因为其确定了备选界标中相邻界标的大小,为给定的界标选择正确的 s 值,且能够在正确范围内捕捉到本地现象,并更好地将源分布与目标分布对齐。由于 s 的极端值会将源点分配的目标点分配完美地匹配起来: K 值会变成 0(当 s 接近 0 的时候)或 1(当 s 非常大时),故应当避免。

计算备选界标的质量实际上是做了一个多尺度分析:本文选择最佳尺度的 s 来捕捉数据的本地属性,同时避免 s 的极端值。为了达到这个目的,本文计算了所有元素对中欧几里得距离的分布,并尝试了分布的每一个百分位数。有了这个以百分位数为基础的方法,本文尝试了一串 s 值,其结果貌似都是可信的。本文计算了 s 中源分布和目标分布之间的重复,保留了备选界标中质量检测最佳的一个。

2.3 分布重复标准

对于备选界标 c 和标尺 s ,本文计算出两个 K 值集中的重复度:源点中的 KV_s 和目标点中的 KV_T 。为了降低计算成本,两种分布都被近似为普通分布并用标准误差公式进行总结: $\mu_s, \sigma_s, \mu_T, \sigma_T$ 。为了能够使用固定阈值并对其赋予意义,本文采用了一种标准重复算法,如下所示:

$$overlap(\mu_s, \sigma_s; \mu_T, \sigma_T) = \frac{N(\mu_s - \mu_T \mid 0, \sigma_{sum}^2)}{N(0 \mid 0, \sigma_{sum}^2)} \tag{2}$$

式中, $\sigma_{sum}^2 = \sigma_s^2 + \sigma_T^2$, $N(\cdot \mid 0, \sigma_{sum}^2)$ 是居中的标准分布。

为了计算两种概率密度 f 和 g 的重复,本文算出了它们的积: $\int f(x)g(x)dx$ 。通常,两种密度间的重复越多,积就越大。这种重复算法与巴塔查里亚系数相似。通过解析地结合两个普通分布的积,本文得出了如下公式:

$$\int N(x \mid \mu_s, \sigma_s^2) N(x \mid \mu_T, \sigma_T^2) dx = N(\mu_s - \mu_T \mid 0, \sigma_{sum}^2) \tag{3}$$

式中的分母与给定 σ_{sum} (由 $\mu_s = \mu_T$ 中获得)中分子的最大值一致。分母作为一种归一元素,当分布完美匹配且给出了更简单的解释时,将重复设为 1,有助于阈值 th 的选择。

2.4 选择界标的投影

$S \cup T$ 中的每一个元素 p 都通过使用有着标准误差的高斯核基被投影到了界标 $\alpha_j \in A$ 中,

$$K(p, \alpha_j) = \exp\left(\frac{-\|p - \alpha_j\|^2}{2\sigma^2}\right) \tag{4}$$

整体来看, S 和 T 中所有元素都被投影到了普通空间内。由于有了界标,这些普通空间有很多维度。仿照其他非线性法,本文在随机从 $S \cup T$ 中提取的要素对间将 σ 设置成中间距离。也通过一些交叉验证选择了 σ 的值。在投影之后,最终获得了源与目标的新代表,分别是 K_s 和 K_T 。

2.5 子空间对齐

在采用了非线性投影集 K_s 和 K_T 后,本文进行了子空间对齐法。PCA 分别用于域中,提取了拥有最大特征向量的 d 特征向量。根据文献[11]中的理论部分,能够确定 d 的最佳值,因为文献[11]中提出了一个基于标准集中不平等的关于特征向量的一致性定理,其可以在两个连续特征向量的误差中找到界限。本文利用此界限有效地调整了 PCA 中的维度数量 d 。源域和目标域中的 d 特征向量分别由 X_s 和 X_T 表示。域中的每个要素都能够分别被投影到其子空间 $K_s X_s$ 和 $K_T X_T$ 中。

子空间对齐的目的是找到线性转换 M , 能够将源特征向量最好地投影到目标特征向量中。此外,本文意在找到 M 来减小源特征向量和目标特征向量之间欧几里得距离的数量。这种减少等同于下列弗罗宾尼斯范数:

$$F(M) = \|X_s M - X_T\|_F^2 \tag{5}$$

对齐转化 M 将要素从源特征空间中映射到目标特征空间,其能够通过计算 $K_s X_s M$, 将投影好的源要素 $K_s X_s$ 作为投影好的目标要素 $K_T X_T$ 带到相同的特征空间中。此算法叫做 LSSA(以选择的界标为基础的子空间对齐),其伪代码如算法 1 所示。

算法 1 以界标选择为基础的子空间对齐和分类

输入: 源数据 S , 目标数据 T , 源标记 L_s , 阈值 th , 子空间维度 d 。

输出: L_T 是 T 中要素的预估标记

$A \leftarrow \text{choose_landmarks}(S, T, th)$

$\sigma \leftarrow \text{median_distance}(S \cup T)$

$K_S \leftarrow \text{project_using_kernel}(S, A, \sigma)$

$K_T \leftarrow \text{project_using_kernel}(T, A, \sigma)$

$X_S \leftarrow \text{PCA}(K_S, d)$

$X_T \leftarrow \text{PCA}(K_T, d)$

$M \leftarrow X'_S X_T$

$P_S \leftarrow K_S X_S M$

$P_T \leftarrow K_T X_T$

$\text{classifier} \leftarrow \text{learn_classifier}(P_S, L_S)$

$L_T \leftarrow \text{classifier}(P_T)$

3 实验与分析

本文实验有两个目的。其一,研究界标选择法的表现,即是处理与其他界标选择法相比较而言的无人管理的图像域自适应。其二,本文希望证明在与子空间对齐法联合使用时,此法在自适应领域有巨大的提升,包括文献[11,12]中描述的两个最好最新的方法。

3.1 数据集与实验设置

本文在图像域自适应的标准数据集中进行实验。所使用的办公数据集^[13]包括从摄像头中获取的图像(用 W 表示),从数字 SLR 相机中取得的图像(用 D 表示),和从百度中取得的图像(用 A 表示)。此外,还运用了一些大学实验图像^[9](用 C 表示)。每一个数据集都为 10 种类别提供了不同的图像。因此,本文能从 4 个数据集 (A, C, D, W) 中获取到 12 种域自适应子问题。其中一个数据集扮演了源 S 的角色,而另一个被看作是目标 T 。本文通过符号 $S \rightarrow T$ 表明了一个域自适应问题。目的是从被标注的源 S 中研究一个 SVM 分类器(使用 SVM 的线性核基),并将其配置到目标 T 上。本文根据文献[8,9,11,13,14]中的标准协议得到了源案例和目标案例。

界标选择法的比较:为了完成这个实验,本文将该方法(表 1 中的 MLS)域下列三种基线进行比较。

随机选择:本文随机选出了 300 个界标(每个

域中 150 个),并重复了 5 次,得到一个平均表现。

无界标选择:本文将所有源案例与目标案例作为界标。

所有的备选界标都采用相同的标准误差 σ 。 σ 被设为最标准的误差(对此条基线有利),这有助于获取两种分布之间最大重复的平均值^[15]。对于 MLS 和 $\sigma - LS$, 本文将重复率固定为 0.3 来选择界标。由于规范化,0.3 的阈值等同于 30% 的重复率。本文还将 MLS 与其他界标选择法进行比较,即用界标连接点法(CDL)。测地线流内核(GFK)^[9]中间子空间的顺序分布在连接源域与目标域的测地线路径两旁。一步子空间对齐法(SA)方法经过研究两个子空间的线性转化而提出。文献[12]中提到的转化联结匹配法(TJM),此方法是最近提出的基于特征匹配和案例权重的方法。

此外,本文在两种基线下进行实验。第一种并没有进行任何的自适应(NA);第二种在源域和目标域中实行了两个独立的 KPCA,并用 SA 算法(用 KPCA + SA 表示)研究了线性转化。

3.2 结果分析

根据表 1 所示结果,本文能做出如下结论。首先,平均来说,本文的方法(使用学生成对测试,平均精确度为 48.1%)大大优于其他方法。在 12 个域自适应任务中,MLS 在 8 个子问题中的精准度都是最佳的。其次,对于两种子问题($W \rightarrow D$ 和 $D \rightarrow W$),ALL 更好些。这意味着将所有源案例和目标案例保留在这两种对称情形中比试图寻找界标要好些。值得一提的是,这两种子问题都是最简单的问题,它们有着最高的精准度,证明了保留所有数据的好处。此外,本文方法在 12 种问题中的 10 种都比 CDL 有优势,而且 CDL 在半无人管理的域自适应情境中是专门选择界标的。最后,单一尺度法(采用固定 σ)并不是很好。这表明在 MLS 中,为每一个界标选择最好的活动半径是多么重要。图 2 给出了 MLS 为每一个域自适应子问题选择出的界标的分布,其中,黑色表示界标分布范围为 15% ~ 20%,深灰色表示 10% ~ 15%,浅灰色表示 5% ~ 10%。这表明即使没有类别信息,本文的方法仍然能够在各种类别中做出平衡选择。

表 1 关于 12 个无人管理的域自适应子问题的五种界标选择法的比较 (%)

方法	$A \rightarrow W$	$A \rightarrow D$	$A \rightarrow C$	$C \rightarrow D$	$C \rightarrow W$	$C \rightarrow A$	$W \rightarrow D$	$W \rightarrow A$	$W \rightarrow C$	$D \rightarrow W$	$D \rightarrow C$	$D \rightarrow A$	平均
RD	40.3	38.8	42.3	41.2	40.6	47.5	84.0	32.9	28.4	81.8	36.8	32.3	45.6
ALL	41.0	39.4	44.7	41.4	41.6	49.6	85.3	33.0	29.2	82.7	38.6	31.3	46.5
σ -LS	39.3	37.5	43.8	42.7	31.5	52.4	80.3	32.6	29.5	82.0	38.6	31.2	45.1
CDL	38.3	38.8	43.9	45.8	45.4	51.7	77.7	35.3	30.9	72.5	33.9	33.3	45.6
MLS	41.1	39.5	45.0	45.2	44.1	53.6	84.7	35.9	31.6	82.4	39.2	34.5	48.1

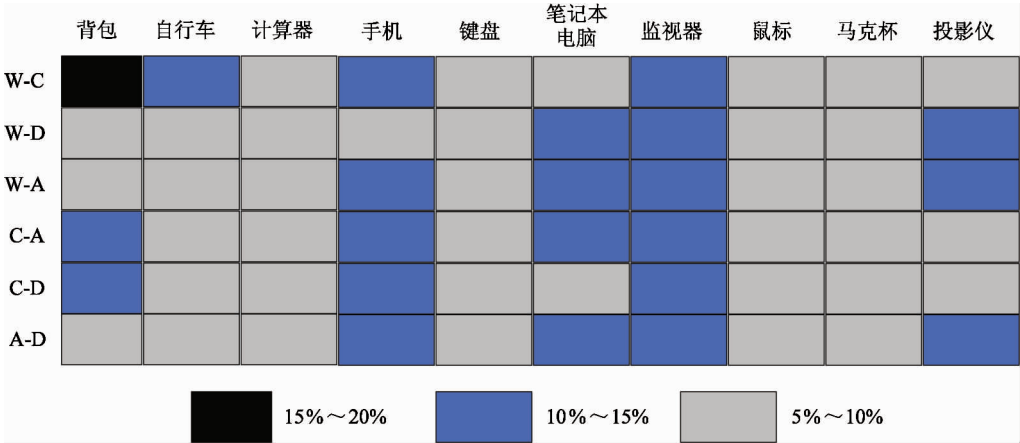


图 2 每一个域自适应子问题中十种类别的界标分布

与当前优秀的无人管理自适应法作比较。表 2 给出了最新的无人管理子空间对齐域自适应法的实验结果。值得一提的是,本文的 LSSA 法在 12 种中的 7 种子问题上都比其他方法表现好,同时 TJM 在剩下的 5 种子问题中表现更好。然而,平均来看, LSSA 大大优于 TJM (52.6%对50.5%)。此外, TJM 的时间复杂性远远大于其他方法,因为其需要解决一个较大的优化问题,而本文的方法包含的针对界标选择的贪心策略和针对子空间对齐封闭解更加有

效。TJM 和 LSSA 的精准度的差别在于:前者采用了权重机制,主要能将两种域移动得更近一些,而后者通过高斯假设,同时考虑到了界标数据分布的标准误差。从表 2 中可以看到, LSSA 远远优于 SA, LSSA 能够捕捉到非线性,这是 SA 难以比拟的。然而,考虑非线性的方式也是一个关键。如 KPCA + SA 所表示的那样,在子空间对齐之前执行两个独立的 KPCA 会导致最坏的结果。

表 2 无监督 DA 方法的比较 (%)

方法	$A \rightarrow W$	$A \rightarrow D$	$A \rightarrow C$	$C \rightarrow D$	$C \rightarrow W$	$C \rightarrow A$	$W \rightarrow D$	$W \rightarrow A$	$W \rightarrow C$	$D \rightarrow W$	$D \rightarrow C$	$D \rightarrow A$	平均
NA	31.5	40.7	45.4	38.2	30.2	50.1	80.2	32.4	32.2	67.8	28.3	30.8	42.2
KPCA + SA	10.1	5.1	7.7	7.6	10.5	10.4	7.6	10.4	11.8	7.2	8.5	7.5	8.7
GFK	38.6	35.7	40.1	44.6	39.0	54.1	81.2	36.6	28.9	80.3	39.2	33.1	45.9
SA	40.7	46.4	41.6	49.0	42.7	52.7	78.9	39.4	34.7	83.4	44.8	38.0	49.3
TJM	42.0	45.8	45.7	49.0	48.8	58.6	83.4	40.8	34.8	82.0	39.6	35.1	50.5
LSSA	42.4	47.2	44.8	54.1	48.1	58.4	87.2	39.4	34.7	87.1	45.7	38.1	52.6

4 结 论

在子空间对齐为基础的域自适应算法框架下,

推测源分布与目标分布之间的转化可以用线性函数来纠正。因此,本文优化非线性映射函数来对齐源子空间和目标子空间。本文结合了新的首次选择界

标的无人管理域自适应法(LSSA)中的两个观点,使得能够通过非线性投影来创建一个空间,在这个空间中,两种域能够变得更近,还能使子空间对齐。实验结果表明,在12个域自适应任务中,MLS在8个子问题中的精准度都是最佳的。在无监督方法的比较中,LSSA在7种子问题上比其他方法表现更佳,TJM在剩下的5种子问题中表现更佳。未来本文将依靠优化过程来共同选择界标,这样就可以避免它们之间的重复。

参考文献

- [1] 杨东日,王颖,储浩. 基于 InfiniBand 协议的云存储关键技术研究[J]. 高技术通讯, 2014, 24(1):35-40
- [2] 原魁,肖晗,何文浩. 采用FPGA的机器视觉系统发展现状与趋势[J]. 计算机工程与应用, 2010, 46(36): 1-6
- [3] 苏晨. 面向统计机器翻译的领域自适应方法研究[D]. 北京:北京交通大学, 2015
- [4] Pan S J, Yang Q. A survey on transfer learning[J]. *IEEE Transactions on Knowledge & Data Engineering*, 2010, 22(10): 1345-1359
- [5] 李明爱,崔燕,杨金福,等. 基于HHT和CSSD的多域融合自适应脑电特征提取方法[J]. 电子学报, 2013, 41(12): 2479-2486
- [6] Kumar A, Saha A, Daume H. Co-regularization based semi-supervised domain adaptation[C]. In: Proceedings of the Conference on Neural Information Processing Systems 2010, Vancouver, Canada, 2010. 478-486
- [7] 刘昊. 统计机器翻译领域自适应方法研究[D]. 苏州:苏州大学, 2016
- [8] Gopalan R, Li R, Chellappa R. Domain adaptation for object recognition: an unsupervised approach[C]. In: Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011. 999-1006
- [9] Grauman K. Geodesic flow kernel for unsupervised domain adaptation[C]. In: Proceedings of the Computer Vision and Pattern Recognition, Providence, USA, 2012. 2066-2073
- [10] 许华婷. 基于Active Learning的中文分词领域自适应方法的研究[D]. 北京:北京交通大学, 2015
- [11] Fernando B, Habrard A, Sebban M, et al. Unsupervised visual domain adaptation using subspace alignment[C]. In: Proceedings of the IEEE International Conference on Computer Vision, Columbus, USA, 2014. 2960-2967
- [12] Long M, Wang J, Ding G, et al. Transfer joint matching for unsupervised domain adaptation[C]. In: Proceedings of the Computer Vision and Pattern Recognition, Columbus, USA, 2014. 1410-1417
- [13] 张良力,吴超仲,严新平,等. 基于视觉范围的自适应阈值DR模型研究[J]. 系统仿真学报, 2008, 20(4): 826-828
- [14] Kulis B, Saenko K, Darrell T. What you saw is not what you get: domain adaptation using asymmetric kernel transforms[C]. In: Proceedings of the Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011. 1785-1792
- [15] 赵鹏,王维,刘慧婷,等. 基于迁移鲁棒稀疏编码的图像表示方法[J]. 计算机学报, 2017, 40(10): 2421-2432

Research on unmanned management domain adaptive algorithm based on landmarks in machine vision

Zeng Yuanrou*, Wang Hongxia**

(* College of Data and Information, Changjiang Polytechnic, Wuhan 430074)

(** College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070)

Abstract

To solve the problem that domain adaptive algorithm needs all source cases and target cases, an unmanaged domain adaptive algorithm based on boundaries is proposed. Firstly, the mapping between source distribution and target distribution is maximized by selecting boundaries from source and target (ST). Second, Gaussian kernel is applied to the selected boundaries to obtain new source (K_s) and new target (K_T). Then, two independent principal components analysis (PCA) are implemented before the subspace alignment of the feature vectors. Finally, classifiers are studied from the annotated source data and executed into the target domain. The results of experiments in the image domain show that the boundaries selected by the proposed method can reduce the inconsistency between domains, and can display the data of effective subspace alignment for non-linear items, which is superior to other unmanned domain adaptive algorithms.

Key words: landmark, domain adaption, kernel projection, subspace alignment, adaptive algorithm