LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

基于特征的迁移学习

Presented by Yafei Hu

2020.11.24

背景

- 域$\mathbb{D}$：特征空间$\mathfrak{X}$及其上的边缘概率分布$\mathbb{P}^X$
- 任务$\mathbb{T}$：标签空间$\mathcal{Y}$和条件概率分布$\mathbb{P}^{Y|X}$(等价于预测函数f(·))
- 同构迁移学习：$\mathfrak{X}_s \cap \mathfrak{X}_t \neq \emptyset$且$\mathcal{Y}_s = \mathcal{Y}_t$，但$\mathbb{P}^{X_s} \neq \mathbb{P}^{X_t}$或$\mathbb{P}^{Y_s|X_s} \neq \mathbb{P}^{Y_t|X_t}$

基本思路

学习一对映射函数$\{\varphi_s(\cdot), \varphi_t(\cdot)\}$，将来自源域和目标域的数据映射到共同的特征空间，然后使用映射到新的特征空间上的源域和目标域数据训练分类器。

主要内容

依据学习特征映射函数的动机和假设的不同，可将基于特征的迁移学习方法分为三类：

- 最小化域间差异
- 学习通用特征
- 特征增强

最小化域间差异

最小化域间差异

动机：学习一对映射函数$\{\varphi_s(\cdot), \varphi_t(\cdot)\}$，使得源域和目标域的数据再新的特征空间上的分布较为接近。

如何度量不同域在新的特征空间上的距离？

最小化域间差异

最大均值差异（MMD）：给定来自两个分布的域样本X_s和X_t，MMD距离的经验估计为

$$\mathrm{MMD}(\mathrm{X}_s, \mathrm{X}_\mathrm{t}) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \Phi(x_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} \Phi\left(x_i^t\right) \right\|_{\mathcal{H}}^2$$

其中，$\Phi(\mathrm{x})$将实例映射到与核$k\left(x_i, x_j\right) = \Phi(x_i)^T\ \Phi\left(x_j\right)$相关联的希尔伯特空间$\mathcal{H}$，$n_s$和$n_\mathrm{t}$分别是源域和目标域的样本数量。

最小化域间差异

将MMD展开可得：

$$\mathrm{MMD}(\mathrm{X}_s, \mathrm{X}_t) = \left\| \frac{1}{\mathrm{n}_s^2} \sum_{i=1}^{n_s} \sum_{i'=1}^{n_s} \Phi(x_i^s)^T \Phi\left(x_{i'}^s\right) - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \Phi\left(x_i^t\right)^T \Phi\left(x_{j'}^t\right) + \frac{1}{\mathrm{n}_t^2} \sum_{j=1}^{n_t} \sum_{j'=1}^{n_t} \Phi\left(x_j^t\right)^T \Phi\left(x_{j'}^t\right) \right\|_{\mathcal{H}}$$

可以使用核函数$k\left(x_i, x_j\right) = \Phi(x_i)^T \Phi\left(x_j\right)$替代上式中的各项。更进一步，使用$K_{s,s}$、$K_{t,t}$、$K_{s,t}$表示源域、目标域和交叉域的核矩阵，定义一个复合核矩阵

$$K = \begin{bmatrix} K_{s,s} & K_{s,t} \\ K_{s,t}^T & K_{t,t} \end{bmatrix}$$

最小化域间差异

矩阵K大小为$(n_s, n_t)\times(n_s, n_t)$，刚好涵盖了展开式中的所有内积，另外定义一个同样大小的矩阵L：

$$l_{ij} = \begin{cases} \frac{1}{\mathrm{n}_s^2}, & x_i, x_j \in X_s \\ \frac{1}{\mathrm{n}_t^2}, & x_i, x_j \in X_t \\ -\frac{2}{n_s n_t}, & others \end{cases}$$

显然这些刚好对于展开式中的常数项，简单的验证可以发现:

$$\mathrm{MMD}(\mathrm{X}_s, \mathrm{X}_t) = tr(KL)$$

最小化域间差异

基于MMD距离，人们提出了一种用于迁移学习的降维算法，名为最大均值差异嵌入（MMDE）。主要思想是解如下问题：

$$\min_{\varphi} MMD\left(\varphi(X_S), \varphi(X_T)\right) + \lambda\Omega(\varphi)$$

$$s.t. \qquad \varphi(X_S)\text{和}\varphi(X_T)\text{的约束条件}$$

其中φ是待学习的映射函数，第一项最小化源域数据和目标域数据之间的MMD距离， Ω是(φ)映射函数φ的正则项，约束条件用于保留原始数据的特性。

最小化域间差异

根据$\mathrm{MMD}(\mathrm{X}_s, \mathrm{X}_\mathrm{t}) = tr(KL)$，上式可以改写为：

$$\min_{\varphi} tr(KL) + \lambda\Omega(\varphi)$$

$$s.t. \quad \varphi(X_\mathrm{S})和\varphi(X_T)的约束条件$$

其中K为由核函数$k(x_i, x_j) = \psi(x_i)^T \psi(x_j)$构成的核矩阵，$\psi(\cdot) = \Phi(\varphi(\cdot))$。更进一步，可以先求K：

$$\min_{K\geq 0} tr(KL) + \lambda\Omega(\varphi)$$

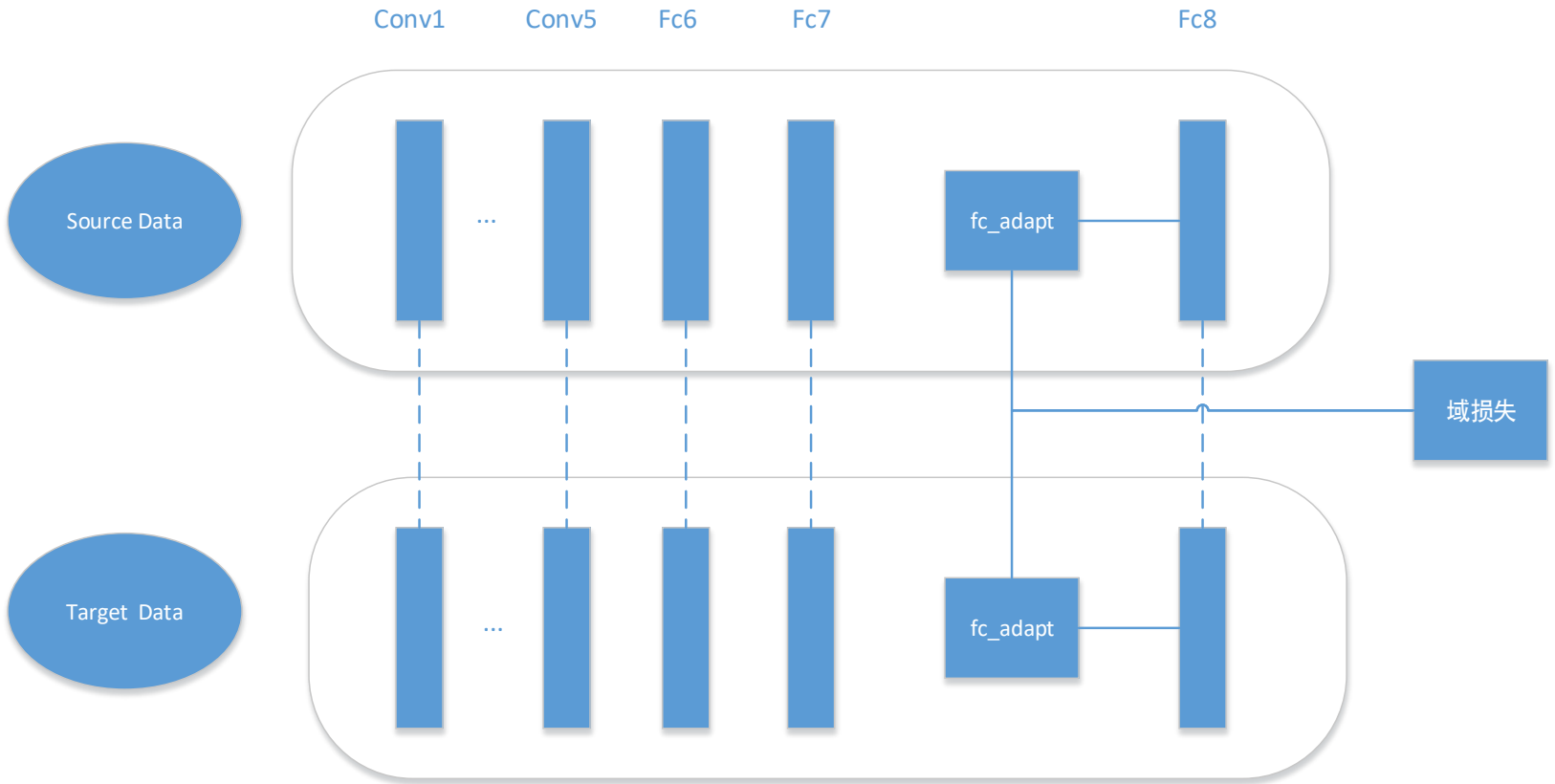
$$s.t. \quad 关于K的约束条件$$

然后在K上应用主成分分析，获得主要特征向量重构源域和目标域数据的期望映射。另外还使用数学手段提升效率，即迁移成分分析方法（TCA），详见论文：

https://www.ijcai.org/Proceedings/09/Papers/200.pdf

最小化域间差异

在深度学习的背景下，研究者提出使用深度神经网络近似前面要学习的特征映射$\varphi(\cdot)$一种做法如下图所示：



最小化域间差异

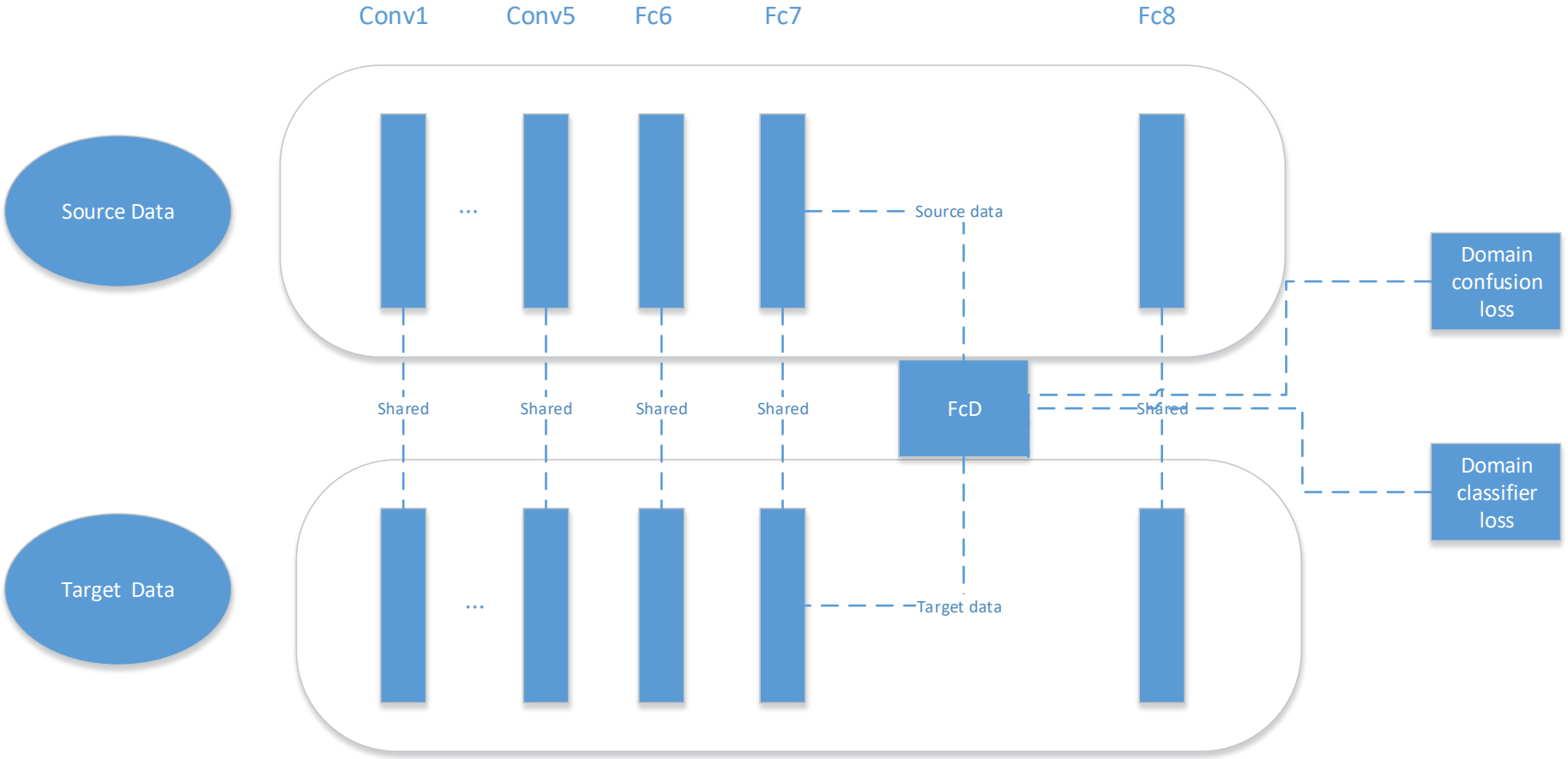
除了使用MMD距离来测量域间差异，还可以基于别的度量方法，比如Bregman散度，提出类似的优化目标：

$$\min_{\phi} F(\phi) + \lambda D_{\omega}(\phi(x_s)\|\phi(x_t))$$

其中$F(\phi)$定义了一个任务特定目标，例如最小化分类误差。 $D_{\omega}(\phi(x_s)\|\phi(x_t))$是$\phi(x_s)$和$\phi(x_t)$之间的Bregman散度。

最小化域间差异

度量域间差异有多种标准，但没有哪一种标准是通用的。基于对抗训练的思想，人们提出了如下的通用方法：



学习通用特征

学习通用特征

动机：之前的方法都是学习给定的源域和目标域中的域不变性特征，学习通用特征旨在从若干个域中学习通用的特征表示。这主要有以下两个方法：

- 学习通用编码
- 深度通用特征

学习通用特征

学习通用编码的基本思想包括两个步骤：

第一步，从大量无标签数据中学习更高级别的基向量组$\mathrm{B} = \{b_1, \cdots, b_{n_s}\}$，这些数据可以来自迁移学习环境中的多个源域，形式化为如下问题：

$$\min_{A,B} \sum_i \left\| x_i^s - \sum_j a_i^j b_j \right\|_2^2 + \beta \|a_i\|_1$$
$$s.t. \quad \|b_j\|_2 \leq 1, \qquad \forall j \in 1, \cdots, n_s$$

其中$a_i = \left(a_i^1, \cdots, a_i^{n_s}\right)^T$是$x_i^s$在基向量组B上的坐标，$\beta$是正则化参数。第一项将$x_i^s$重构为基$\{b_1, \cdots, b_{n_s}\}$的加权组合，第二项使$a_i$尽量稀疏。

学习通用特征

第二步：求解目标域在$\{b_1, \cdots, b_{n_s}\}$空间上的表示，形式化为如下问题

$$\hat{a}_i = arg\min_{a_i} \left\| x_i^t - \sum_j a_i^j b_j \right\|_2^2 + \beta \|a_i\|_1$$

最终，可以将所有数据映射到$\{b_1, \cdots, b_{n_s}\}$这个公共空间上，并进行模型的学习和训练。

学习通用特征

除了稀疏编码，人们将由深度神经网络得到的特征称为深度通用特征。这一方面的工作主要可以分为两个部分：

- 基于编码解码和重构损失的方法
- 基于聚类的方法

前者将网络分为编码器和解码器，编码器将输入映射到隐藏特征空间，解码器将隐藏特征空间还原到真实数据，使用多个域的数据以及重构损失训练网络后，可以将编码器的结果视为通用的特征表示。

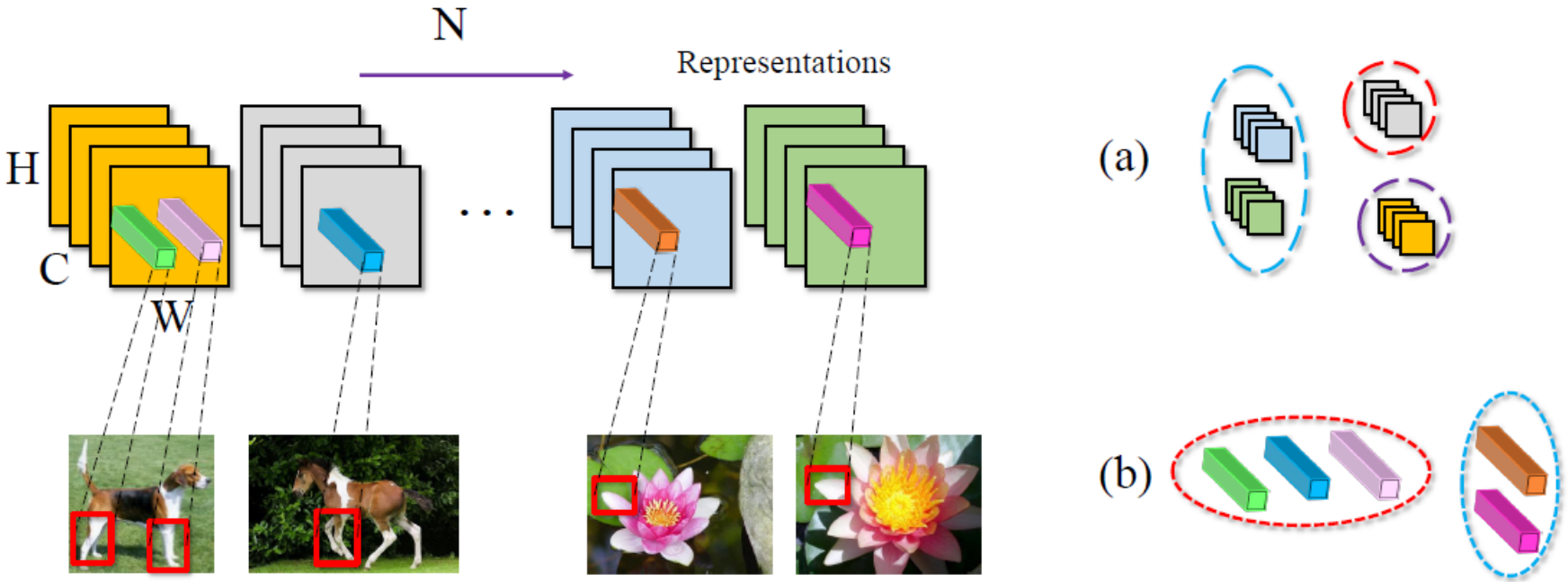
学习通用特征

与重构损失相比，基于聚类的方法在复杂性方面更轻量的无监督学习方法，而且可以增加学习到的特征的可解释性。假设神经网络输出的表示为一个四维张量 $Y \in R^{N\times C\times H\times W}$,其中N，C，H和W分别为数据批大小，隐藏单元的数量，特征表示的高度和宽度，按基于k-means的样本聚类的损失函数定义：

$$\Re_{\text{sample}}(Y,\mu)=\frac{1}{2\text{NCHW}}\sum_{n=1}^{N}\left\|T^{\{N\}\times\{C,H,W\}}(Y)_n-\mu_{z_n}\right\|^2$$

其中T为将张量Y展开成上标大小的矩阵，μ为给定的k个center组成的矩阵， z_n 为一个潜在变量，表明第n个样本所属的cluster。

学习通用特征



如上示意图，a为样本聚类，b为空间聚类。除此之外还可以使用别的聚类方法，详见论文

https://proceedings.neurips.cc/paper/2016/file/a376033f78e144f494bfc743c0be3330-Paper.pdf

特征增强

特征增强

基于特征增强的domain adaptation做法非常简单，假设原始输入空间$\mathfrak{X} \in \mathbb{R}^F$，将输入空间增强到$\tilde{\mathfrak{X}} \in \mathbb{R}^{3F}$，源域和目标域上的映射函数$\phi^s$，$\phi^t$：$\mathfrak{X} \to \tilde{\mathfrak{X}}$定义为

$$\phi^s(x) = \langle x, x, 0 \rangle, \qquad \phi^t(x) = \langle x, 0, x \rangle$$

其中0表示F维空间中的零向量，然后使用特征增强后的数据进行学习。

总结

总的来说，基于特征的迁移方法都是研究如何获得一个好的映射，使得源域和目标域的数据在映射到新的特征空间上后尽量相似，以使用源域的大量有标签数据。

Thanks