TCA

杜云涛

2019 年 9 月 25 日

1 相关知识

0.论文讲解

https://zhuanlan.zhihu.com/p/26764147

1.整体框架

一般的迁移学习中将问题建模为以下形式的优化问题：

$$\begin{aligned} \min \quad & \text{某个衡量两个分布差异的统计指标} + \text{正则化项目} \\ s.t. \quad & \text{与具体问题相关的约束} \end{aligned}$$

在TCA算法中,该优化问题表示为:

$$\begin{aligned} \min \quad & \text{以}MMD\text{表示的边缘分布差异} + \text{降维矩阵}W\text{的正则化项} \\ s.t. \quad & \text{最大化保留映射后的样本所含信息(TCA中用方差进行衡量)} \end{aligned}$$

2.Hilbert空间

张贤达《矩阵分析与应用》第1章

表 1.3.1 几种向量空间的比较

向量空间	定义了向量的加法和向量的数乘，以向量为元素的集合 $\mathbb{R}^n$ 或 $\mathbb{C}^n$
内积向量空间	定义了内积 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ (向量的乘法) 的向量空间
赋范向量空间	定义了范数 $\|\boldsymbol{x}\|$ 的向量空间，可度量向量的长度、距离与邻域
Banach 空间	满足 $\lim\limits_{n\to\infty} \boldsymbol{v}_n \to \boldsymbol{v}, \forall \boldsymbol{v}_n, \boldsymbol{v} \in \mathbb{C}^n$ 的完备赋范向量空间
Hilbert 空间	满足 $\lim\limits_{n\to\infty} \|\boldsymbol{v}_n\| \to \|\boldsymbol{v}\|, \forall \boldsymbol{v}_n, \boldsymbol{v} \in \mathbb{C}^n$ 的完备赋范向量空间
Euclidean 空间	具有 Euclidean 范数 $\|\boldsymbol{x}\|_2$ 的赋范向量空间

Figure 1: Hilbert空间

3.核函数

$K(x, z) = \phi(x)^T \phi(z)$.

https://www.cnblogs.com/jerrylead/archive/2011/03/18/1988406.html

4.矩阵的迹的性质

$$
\begin{aligned}
& a = tr(a), \text{a为常数} \\
& tr(A^T) = tr(A) \\
& tr(ABC) = tr(BCA) = tr(CAB) \\
& ||A||_F^2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \\
& ||A||_F^2 = tr(AA^T)
\end{aligned}
$$

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

5.MMD公式推导

https://zhuanlan.zhihu.com/p/63026435

$$
\begin{aligned}
& \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_i) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(\mathbf{x}_j) \right\|^2 \\
& = \left\| \frac{1}{n_s}\left[\phi(x_1),\phi(x_2),...,\phi(x_{n_s})\right]\begin{bmatrix}1\\1\\...\\1\end{bmatrix} - \left[\phi(x_1),\phi(x_2),...,\phi(x_{n_t})\right]\begin{bmatrix}1\\1\\...\\1\end{bmatrix} \right\|^2
\end{aligned}
$$

记$\phi_(\boldsymbol{x_s}) = \left[\phi(x_1),\phi(x_2),...,\phi(x_{n_s})\right]$, $\phi(\boldsymbol{x_t})\left[\phi(x_1),\phi(x_2),...,\phi(x_{n_t})\right]$
$\phi(x_s)$的维度可以看作是$1*n_s$,也可以看作是$d*n_s$，因为每一个$\phi(x_{s_i})$的维度实际是$d*1$.
$\phi(x_t)$同理。
王晋东的公式推导中，对于$\phi(x_s)$和$\phi(x_t)$部分缺少了转置符号，因为$K(x,z) = \phi(x)^T\phi(z)$.

上式可以表示为：

$$
\begin{aligned}
&\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_i)-\frac{1}{n_t}\sum_{j=1}^{n_t}\phi(\mathbf{x}_j)\right\|^2 \\
&=\left\|\frac{1}{n_s}\left[\phi(x_1),\phi(x_2),...,\phi(x_{n_s})\right]\begin{bmatrix}1\\1\\...\\1\end{bmatrix}_{n_s*1}-\left[\phi(x_1),\phi(x_2),...,\phi(x_{n_t})\right]\begin{bmatrix}1\\1\\...\\1\end{bmatrix}_{n_t*1}\right\|^2 \\
&=\left\|\frac{1}{n_s}\phi(\boldsymbol{x_s})\mathbf{1}_{n_s*1}-\frac{1}{n_t}\phi(\boldsymbol{x_s})\mathbf{1}_{n_t*1}\right\|^2 \\
&=tr[(\frac{1}{n_s}\phi(\boldsymbol{x_s})\mathbf{1}_{n_s*1}-\frac{1}{n_t}\phi(\boldsymbol{x_s})\mathbf{1}_{n_t*1})(\frac{1}{n_s}\phi(\boldsymbol{x_s})\mathbf{1}_{n_s*1}-\frac{1}{n_t}\phi(\boldsymbol{x_s})\mathbf{1}_{n_t*1})^T] \\
&=tr[(\frac{1}{n_s}\phi(\boldsymbol{x_s})\mathbf{1}_{n_s*1}-\frac{1}{n_t}\phi(\boldsymbol{x_s})\mathbf{1}_{n_t*1})(\frac{1}{n_s}\mathbf{1}_{n_s*1}{}^T\phi(\boldsymbol{x_s})^T-\frac{1}{n_t}\mathbf{1}_{n_t*1}{}^T\phi(\boldsymbol{x_s})^T)] \\
&=tr\left(\begin{bmatrix}\phi(\mathbf{x}_s) & \phi(\mathbf{x}_t)\end{bmatrix}\begin{bmatrix}\frac{1}{n_s}\mathbf{1}_{n_s*1}\\-\frac{1}{n_t}\mathbf{1}_{n_t*1}\end{bmatrix}\begin{bmatrix}\frac{1}{n_s}\mathbf{1}_{n_s*1}{}^T & -\frac{1}{n_t}\mathbf{1}_{n_t*1}{}^T\end{bmatrix}\begin{bmatrix}\phi(\mathbf{x}_s)^T\\\phi(\mathbf{x}_t)^T\end{bmatrix}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}\phi(\mathbf{x}_s) & \phi(\mathbf{x}_t)\end{bmatrix}\begin{bmatrix}\frac{1}{n_s^2}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{-1}{n_sn_t}\mathbf{1}\mathbf{1}^{\mathrm{T}}\\\frac{-1}{n_sn_t}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{1}{n_t^2}\mathbf{1}\mathbf{1}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\phi(\mathbf{x}_s^T)^T\\\phi(\mathbf{x}_t)^T\end{bmatrix}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}\phi(\mathbf{x}_s)^T\\\phi(\mathbf{x}_t)^T\end{bmatrix}\begin{bmatrix}\phi(\mathbf{x}_s) & \phi(\mathbf{x}_t)\end{bmatrix}\begin{bmatrix}\frac{1}{n_s^2}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{-1}{n_sn_t}\mathbf{1}\mathbf{1}^{\mathrm{T}}\\\frac{-1}{n_sn_t}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{1}{n_t^2}\mathbf{1}\mathbf{1}^{\mathrm{T}}\end{bmatrix}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}<\phi(\mathbf{x}_s),\phi(\mathbf{x}_s)> & <\phi(\mathbf{x}_s),\phi(\mathbf{x}_t)>\\<\phi(\mathbf{x}_t),\phi(\mathbf{x}_s)> & <\phi(\mathbf{x}_t),\phi(\mathbf{x}_t)>\end{bmatrix}\mathbf{M}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}K_{s,s} & K_{s,t}\\K_{t,s} & K_{t,t}\end{bmatrix}\mathbf{M}\right)
\end{aligned}
$$

where

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

$$(M)_{ij}=\begin{cases}\frac{1}{n_sn_s}, & \mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}_s\\\frac{1}{n_tn_t}, & \mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}_t\\\frac{-1}{n_sn_t}, & \text{otherwise}\end{cases}\quad or\quad L=\begin{bmatrix}\frac{1}{n_s{}^2} & -\frac{1}{n_sn_t}\\-\frac{1}{n_sn_t} & \frac{1}{n_t{}^2}\end{bmatrix}$$

6.降维

对于上述存在的核矩阵K，实际的维度为$(n_s+n_t)*(n_s+n_t)$, 这个矩阵实际上可以看作是样本在新空间下的特征表示，第一个n_s+n_t 表示样本数目，此时包括源域和目标域，每个样本的维度为$1*(n_s+n_t)$ or $(n_s+n_t)*1$(一般表示为一个列向量，这时候是第一个(n_s+n_t)表示维度)。

为了进行降维，TCA 采用了一个$K(x,z)-->\phi_x^T\phi(z)-->$ 引入降维矩阵W，降维后的映射为$\phi(x)W-->(\phi(x)W)(\phi(z)W^T)=\phi(x)WW^T\phi(z)$, (这里是把转置符号放在了后面，放在前面也一样，因为根据迹的性质, 可以将这几个交换顺序，会变成一样的形式.) $-->$ K'的过程，即逆向分解核函数矩阵，再正向表示核函数矩阵。

这里要纠正一下，我之前以为直接用K'代替K会有问题，现在发现这样并没有问题，因为实际上上式将MMD距离表示$tr(KL)$,这个核函数矩阵K实际上可以是任意的核函数矩阵，这里是在降维

之后变成了另外一个核函数矩阵，但是同样可以适用，相当于此时的核函数选择是原始的核函数和矩阵W的共同复合成的，因此直接用$tr(K'L)$表示MMD是没有问题的。

7.PCA算法

用方差来衡量样本映射到新的空间下所含信息的多少。

详见西瓜书第10.3节，最大化方差

8.中心化矩阵H

H具有一个很好的性质：$H^n = H^2 = H$，样本X的方差或者协方差可以表示为$X^T HX$.

中心化矩阵可以表示https://www.cnblogs.com/yanxingang/p/10776475.html

9.矩阵求导

$a = tr(a)$, a为常数 $\frac{\partial(f(x))}{\partial x} = \frac{\partial(tr(f(x)))}{\partial x} = tr(\frac{\partial f(x)}{\partial x})$

详见张贤达《矩阵分析与应用》第3.2小节

由于标量函数 $f(\boldsymbol{X})$ 相对于 $m \times n$ 矩阵变元 $\boldsymbol{X}$ 的 Jacobian 矩阵和梯度矩阵之间存在转置关系，所以命题 3.2.1 也意味着

$$\mathrm{d}f(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\mathrm{d}\boldsymbol{X}) \iff \nabla_{\boldsymbol{X}} f(\boldsymbol{X}) = \boldsymbol{A}^{\mathrm{T}} \tag{3.2.27}$$

由于 Jacobian 矩阵 $\boldsymbol{A}$ 的唯一确定性，故梯度矩阵是唯一确定的。

考察二次型函数 $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$，其中，$\boldsymbol{A}$ 是一个正方的常数矩阵。首先将标量函数写成迹函数形式，然后利用矩阵乘积的微分易得

$$\begin{aligned}
\mathrm{d}f(\boldsymbol{x}) &= \mathrm{d}(\mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x})) = \mathrm{tr}[(\mathrm{d}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\mathrm{d}\boldsymbol{x}] \\
&= \mathrm{tr}\left([\mathrm{d}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}]^{\mathrm{T}} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\mathrm{d}\boldsymbol{x}\right) = \mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\mathrm{d}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\mathrm{d}\boldsymbol{x}) \\
&= \mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})\mathrm{d}\boldsymbol{x})
\end{aligned}$$

由命题 3.2.1 直接得二次型函数 $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$ 关于变元向量 $\boldsymbol{x}$ 的梯度向量为

$$\nabla_{\boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \left[\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})\right]^{\mathrm{T}} = (\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{A})\boldsymbol{x} \tag{3.2.28}$$

显然，若 $\boldsymbol{A}$ 为对称矩阵，则 $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$。

Figure 2: 求导规则

TCA会议版论文中，公式(11)对于矩阵W的求导过程如下:

$$\frac{\partial\left[tr(W^T(I+\mu KLK)W) - tr((W^T KHKW - I)Z)\right]}{\partial W}$$

$$tr(\frac{\partial\left[W^T(I+\mu KLK)W - W^T KHKWZ - IZ\right]}{\partial W})$$

$$\begin{aligned}
&= tr(\frac{\partial W^T}{\partial W}(I+\mu KLK)W + W^T\frac{\partial(I+\mu KLK)W}{\partial W} - (\frac{\partial W^T}{\partial W}KHKWZ + W^T\frac{\partial KHKWZ}{\partial W})) \\
&= tr((dW)^T(I+\mu KLK)W + W^T(I+\mu KLK)dW - ((dW)^T KHKWZ + W^T KHKdWZ)) \\
&= tr(W^T(I+\mu KLK)^T dW + W^T(I+\mu KLK)dW - ((KHKWZ)^T dW + ZW^T KHKdW)) \\
&= tr(W^T(I+\mu KLK)^T dW + W^T(I+\mu KLK)dW - (Z^T W^T K^T H^T K^T dW + ZW^T KHKdW)) \\
&= tr(W^T(I+\mu KLK)^T dW + W^T(I+\mu KLK)dW - (ZW^T KHKdW + ZW^T KHKdW)) \\
&= tr(2(W^T(I+\mu KLK)^T - ZW^T KHK)dW))
\end{aligned}$$

上面有一个需要注意的地方，$d(WZ) = dWZ + WdZ = dWZ$

所以,

$$\begin{aligned}
&\frac{\partial\left[tr(W^T(I+\mu KLK)W) - tr((W^TKHKW - I)Z)\right]}{\partial W} \\
&= 2(W^T(I+\mu KLK)^T - ZW^TKHK)^T \\
&= 2((I+\mu KLK) - KHKWZ)
\end{aligned}$$

10.LDA or Fisher discriminant

西瓜书第二章

11.TCA算法效果

https://github.com/jindongwang/transferlearning/blob/master/data/benchmark.md

12.Office-31 数据集

原始图片：https://people.eecs.berkeley.edu/~jhoffman/domainadapt/

做实验用的数据：http://ise.thss.tsinghua.edu.cn/~mlong/ 《44.Transfer Feature Learning with Joint Distribution Adaptation》code中的数据

2 问题梳理

1. 如何降低两个领域间的边缘分布差异？
2. TCA算法隐含的假设是什么？
3. TCA算法对源域和目标域的数据要求如何？