迁移学习经典方法

zzl

2017 年 11 月 26 号

Introduction

本文介绍迁移学习的两个经典方法，TCA 与 JDA。这两篇文章各自给迁移学习提出了一个优化目标，可以成为具有开山价值的工作。这两篇文章分别是 09 年和 13 年出的，看似时间距现在比较早了，但这么多年来很少有类似价值的工作，大多数文章都是基于某个经典工作的优化目标来做些改进。

一、数学背景知识

1．矩阵的迹具有如下性质：

1. $\mathrm{tr(A)} = \mathrm{tr}(A^T)$
2. $\mathrm{tr(AB)} = \mathrm{tr(BA)}$
3. $\frac{d(tr(AB))}{d(A)} = B^T$
4. $\frac{d(tr(ABA^TC))}{d(A)} = \mathrm{CAB} + C^TAB^T$

2．从迹的角度求解 PCA 降维算法

输入数据为 X，$\mathrm{X} \in R^{m*n}$

源域数据量为n_s，目标域数据量为n_t，$n = n_s + n_t$

中心矩阵为$\mathrm{H} = \mathrm{I} - \frac{1}{n}O$，$\mathrm{H} \in R^{n*n}$，其中 O 为 n*n 值全为 1 的矩阵

定义矩阵 X 的协方差矩阵为 S，$\mathrm{XH}X^T$=n*S，n 不影响后面的计算，$\mathrm{XH}X^T$也能表现出数据之间的关系，所以这里我们用$\mathrm{XH}X^T$来表示 X 的协方差矩阵。

对于线性 PCA，转移矩阵为 A，那么我们的优化目标为：$\max_{\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}}\mathbf{A}\right)$，因为$A^TB$为线性去相关后的矩阵，所以$A^TXHX^TA$只有对角线上有值。用拉格朗日乘子法来求解这个函数，求得$XHX^TA = A\emptyset$，$\emptyset$是特征值组成的对角矩阵，如果这个方差取 k 个解，那么 k 个解就是特征值中前 k 大的数。同理，如果这里方程的目标变为求最小，那么 k 个解就是特征值中最小的 k 个数。数据降维后的结果为$\mathrm{Z} = A^TX$。

对于核 PCA，$\max_{\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}\right)$，K 为核矩阵，A 为转移矩阵，数据降维后的结果为$\mathrm{Z} = A^TK$。

关于 PCA 算法的详解，大家可以参考周志华老师的《机器学习》。本文没有给出理论证明部分，只是利用迹来求解该问题。

二、maximum mean discrepancy

最大均值平均差异，也叫 MMD 距离。最先提出的时候用于双样本的检测（two-sample test）问题，用于判断两个分布 p 和 q 是否相同。它的基本假设是：如果对于所有以分布生成的样本空间为输入的函数 f，如果两个分布生成的足够多的样本在 f 上的对应的像的均值都相等，那么那么可以认为这两个分布是同一个分布。现在一般用于度量两个分布之间的相

似性。

具体而言，基于 MMD（maximize mean discrepancy）的统计检验方法是指下面的方式：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，求不同分布的样本在 f 上的函数值的均值，通过把两个均值作差可以得到两个分布对应于 f 的 mean discrepancy。寻找一个 f 使得这个 mean discrepancy 有最大值，就得到了 MMD。最后取 MMD 作为检验统计量（test statistic），从而判断两个分布是否相同。如果这个值足够小，就认为两个分布相同，否则就认为它们不相同。同时这个值也用来判断两个分布之间的相似程度。如果用 F 表示一个在样本空间上的连续函数集，那么 MMD 可以用下面的式子表示：

$$\mathrm{MMD}\left[\mathcal{F}, p, q\right] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{y \sim q}[f(y)]\right),$$

在 domain adaptation 中，经常用到 MMD 来在特征学习的时候构造正则项来约束学到的表示，使得两个域上的特征尽可能相同。从上面的定义看，我们在判断两个分布 p 和 q 的时候，需要将观测样本首先映射到 RKHS 空间上，然后再判断。但实际上很多文章直接将观测样本用于计算，省了映射的那个步骤。

三、Transfer Feature Learning with Joint Distribution Adaptation

1.符号介绍

X_s表示源域的数据

Y_s表示源域的数据对应标签

$\mathrm{P}(X_s)$ 表示源域数据的边缘分布密度

$\mathrm{P}(Y_s|X_s)$ 表示源域数据的条件分布密度

X_t表示目标域的数据

Y_t表示目标域的数据对应标签

$\mathrm{P}(X_t)$ 表示目标域数据的边缘分布密度

$\mathrm{P}(Y_t|X_t)$ 表示目标域数据的条件分布密度

1.问题背景

在很多机器学习任务中，模型的训练和测试时所采用的样本分布是一致的，但在实际中我们模型训练时候的样本分布和实际使用时候的样本分布在很多情况下是不一致的，这就导致了领域适应性问题（Problem of Domain Adaptation）。

域适应问题用数学符号表示为$\mathrm{P}(X_s) \neq \mathrm{P}(X_t)$，$\mathrm{P}(Y_s|X_s) \neq \mathrm{P}(Y_t|X_t)$，并且目标域缺少$Y_t$。

2.算法假设

TCA 假设存在一个映射∅，使得映射后的数据分布$\mathrm{P}(\emptyset(X_s)) \approx \mathrm{P}(\emptyset(X_t))$，并且 $\mathrm{P}(Y_s|\emptyset(X_s)) \approx \mathrm{P}(Y_t|\emptyset(X_t))$ ，现在我们的工作是找到合适的映射∅。

TCA 的工作其实只是使得数据满足$\mathrm{P}(\emptyset(X_s)) \approx \mathrm{P}(\emptyset(X_t))$的条件，并默认为$\mathrm{P}(\emptyset(X_s)) \approx \mathrm{P}(\emptyset(X_t))$时$\mathrm{P}(Y_s|\emptyset(X_s)) \approx \mathrm{P}(Y_t|\emptyset(X_t))$，现在看来 TCA 的假设很强烈，但 TCA 还是取得了不错的效果。其实我们可以从另一个角度来考虑这个问题，对于简单的文本分类问题，基于词频的朴素贝叶斯在某些条件下表现的效果很不错。

3.TCA 优化目标

需要解决域适应问题的本质是缩小源域与目标域之间的差异（距离）。TCA 算法的优化目标是最小化源域与目标域之间边缘分布的差异。

这里的关键是用什么距离来度量他们之间的差异，TCA 采用的是最大均值距离（MMD，maximum mean discrepancy）。

$$dist(X'_{src}, X'_{tar}) = \left\| \frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_{src_i}) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi(x_{tar_i}) \right\|_{\mathcal{H}}$$

4.TCA 算法求解

那么我们需要求Ø使得 dist 最小，TCA 中映射Ø为核函数，当然映射也可以为神经网络，之后的很多工作就是基于网络并且优化目标和 TCA 类似。

TCA 中引入一个核矩阵 K：$K = \begin{bmatrix} K_{src,src} & K_{src,tar} \\ K_{tar,src} & K_{tar,tar} \end{bmatrix}$ 与 L：$L_{ij} = \begin{cases} \frac{1}{n_1{}^2} & x_i, x_j \in X_{src}, \\ \frac{1}{n_2{}^2} & x_i, x_j \in X_{tar}, \\ -\frac{1}{n_1 n_2} & \text{otherwise} \end{cases}$。

这里。$\text{Dist}(X'_S, X'_T = \text{tr}(\text{KL}))$

$\text{K} = (\text{K}K^{-1}{}_2)(K^{-1}{}_2\text{K})$，我们引入一个转移矩 $\text{W} \in R^{(n1+n2)*m}$，

$\widetilde{K} = (KK^{-1/2}\widetilde{W})(\widetilde{W}^{\top}K^{-1/2}K) = KWW^{\top}K$，那现在 $\text{tr}(\text{KL}) =$ $\text{tr}(\text{KW}W^T\text{KL}) = \text{tr}(\text{KW}LW^T\text{K}) = \text{tr}(W^T\text{KLKW})$，那这个问题就很核 PCA 很像了。

$$\begin{aligned} \min_W \quad & \text{tr}(W^{\top}W) + \mu\,\text{tr}(W^{\top}KLKW) \\ \text{s.t.} \quad & W^{\top}KHKW = I, \end{aligned}$$

那现在优化问题变为 ，W^TW作为正则化项，$W^TKHKW = \text{I}$为限定条件，防止优化没有下界，这里想到了 NCE 损失。然而，这个约束是使得变换后的数据的协方差为单位矩阵，原文中作者并没有给出相关解释，我认为是为了给出一个约束条件，而取的这种比较好计算的公式。

接下来用拉格朗日乘子法来求解问题,原问题化为

$\text{tr}(W^T(\text{I} + \text{uKLK})\text{W}) - \text{tr}\big((W^T\text{KHKW} - \text{I})\text{Z}\big) = 0$ 这里 Z 是拉格朗日乘子组成的对角矩阵。对上式求导得下式：$(\text{I} + \text{uKLK})\text{W} = \text{KHKWZ}$

最后得 W 为矩阵$(\text{I} + \text{uKLK})^{-1}KHK$的最大 m 个特征值对应的特征向量组成的矩阵。

变换后的数据为$X' = W^TK$。

四．Transfer Feature Learning with Joint Distribution Adaptation

这篇文章考虑了条件分布概率，给这个领域提出了一个新的优化目标，很有价值。这篇文章的主题是基于线性映射来做的，当然核映射也能推出。

1.优化目标

JDA 的优化目标是减小源域与目标域之间边缘分布的差异同时减小他们之间联合概率密度之间的差异。

减少边缘密度的差异 TCA 已经完成了这部分工作。

减少联合概率密度之间的差异需要求得 P(Y|X)，而目标域中没有标签，并且求得该值比较复杂（用贝叶斯公式）。作者用 P(X|Y)来替换这里的 P(Y|X)，这里同样没有标签，作者采用在源域上训练一个学习器并用这个学习器给目标域打上伪标签。

C 表示一共有 C 个类别，c 表示某一个具体的类别。对于每一个 c 类别中的数据，我们需要使得 $\text{P}(X_s|Y_c), \text{P}(X_t|Y_c)$ 之间的距离最小。即最小化

$$\left\| \frac{1}{n_s^{(c)}} \sum_{\mathbf{x}_i \in \mathcal{D}_s^{(c)}} \mathbf{A}^{\mathrm{T}} \mathbf{x}_i - \frac{1}{n_t^{(c)}} \sum_{\mathbf{x}_j \in \mathcal{D}_t^{(c)}} \mathbf{A}^{\mathrm{T}} \mathbf{x}_j \right\|^2 = \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}} \mathbf{X} \mathbf{M}_c \mathbf{X}^{\mathrm{T}} \mathbf{A}\right)$$

，

$$(M_c)_{ij} = \begin{cases} \frac{1}{n_s^{(c)} n_s^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s^{(c)} \\ \frac{1}{n_t^{(c)} n_t^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \frac{-1}{n_s^{(c)} n_t^{(c)}}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_s^{(c)}, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \mathbf{x}_j \in \mathcal{D}_s^{(c)}, \mathbf{x}_i \in \mathcal{D}_t^{(c)} \end{cases} \\ 0, & \text{otherwise} \end{cases}$$

其中

那么总体的优化目标为：$\min_{\mathbf{A}^{\mathrm{T}} \mathbf{X} \mathbf{H} \mathbf{X}^{\mathrm{T}} \mathbf{A} = \mathbf{I}} \sum_{c=0}^{C} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}} \mathbf{X} \mathbf{M}_c \mathbf{X}^{\mathrm{T}} \mathbf{A}\right) + \lambda \|\mathbf{A}\|_F^2$。C=0 表示边缘密度分布部分的工作。

2.算法求解

这里的求解工作与 TCA 相同。最后求解工作转换为求解

$\left(\mathbf{X} \sum_{c=0}^{C} \mathbf{M}_c \mathbf{X}^{\mathrm{T}} + \lambda \mathbf{I}\right) \mathbf{A} = \mathbf{X} \mathbf{H} \mathbf{X}^{\mathrm{T}} \mathbf{A} \mathbf{\Phi}$，其中∅为拉格朗日乘子组成的对角矩阵。

3.算法框架

基于伪标签求得的结果只能保证在伪标签条件下效果最好，而伪标签与真实标签肯定有差距。那么我们基于伪标签求得结果后，再给目标域打上一次伪标签，这里的结果更加适应目标域。然后以此循环迭代计算。这里的核心思想也就是 EM 算法的思想。

Algorithm 1: JDA: Joint Distribution Adaptation

Input: Data $\mathbf{X}$, $\mathbf{y}_s$; #subspace bases k, regularization parameter λ.
Output: Adaptation matrix $\mathbf{A}$, embedding $\mathbf{Z}$, adaptive classifier f.

begin
  Construct MMD matrix $\mathbf{M}_0$ by Eq. (4), set $\{\mathbf{M}_c := \mathbf{0}\}_{c=1}^{C}$.
  repeat
    Solve the generalized eigendecomposition problem in Equation (10) and select the k smallest eigenvectors to construct the adaptation matrix $\mathbf{A}$, and $\mathbf{Z} := \mathbf{A}^{\mathrm{T}} \mathbf{X}$.
    Train a standard classifier f on $\left\{(\mathbf{A}^{\mathrm{T}} \mathbf{x}_i, y_i)\right\}_{i=1}^{n_s}$ to update pseudo target labels $\{\widehat{y}_j := f(\mathbf{A}^{\mathrm{T}} \mathbf{x}_j)\}_{j=n_s+1}^{n_s+n_t}$.
    Construct MMD matrices $\{\mathbf{M}_c\}_{c=1}^{C}$ by Equation (6).
  until *Convergence*
  Return an adaptive classifier f trained on $\{\mathbf{A}\mathbf{x}_i, y_i\}_{i=1}^{n_s}$.

4.总结

JDA 方法比较巧妙，同时适配两个分布，然后非常精巧地规到了一个优化目标里。用弱分类器迭代，最后达到了很好的效果，值得我们去学习。和 TCA 的主要区别有两点：1）TCA 是无监督的（边缘分布适配不需要 label），JDA 需要源域有 label；2）TCA 不需要迭代，JDA 需要迭代。在此之后有工作是减少 P(Y|X)之间的差异，但也是基于这篇文章的框架所做。

参考文献

[1] S. J. Pan, I. W. Tsang, J. T. Kwok and Q. Yang, "Domain Adaptation via Transfer Component Analysis," in IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 199-210, Feb. 2011.doi: 10.1109/TNN.2010.2091281

[2] https://zhuanlan.zhihu.com/p/26764147?group_id=844611188275965952

[3] Long M, Wang J, Ding G, et al. Transfer feature learning with joint distribution adaptation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 2200-2207.

[4] https://zhuanlan.zhihu.com/p/27336930

[5] 周志华. 机器学习. 清华大学出版社. 2016.

[6] http://blog.csdn.net/a1154761720/article/details/51516273

[7] A kernel two sample test