JDA joint-distribution-adaptation

cross-domain knowledge adaptation problem

在一个原则性降维过程中同时适应边缘分布和条件分布，并构造出有效且鲁棒的新特征表示法，以有效应对大量分布差异。

域间具有不同的边缘分布和条件分布，且目标域无标签.

Assumptions:

$$X_s = X_t,\ Y_s = Y_t \quad P_s(X_s) \neq P_t(X_t),\quad Q_s(y_s|x_s) \neq Q_t(y_t|x_t)$$

通过特征变换T来适应联合分布

$$\min_T \left\| E_{P(x_s,y_s)}[T(x_s), y_s] - E_{P(x_t,y_t)}[T(x_t), y_t] \right\|^2$$

$$\approx \left\| E_{P_s(x_s)}[T(x_s)] - E_{P_t(x_t)}[T(x_t)] \right\|^2 \quad \text{margin distribution}$$

$$+ \left\| E_{Q_s(y_s|x_s)}[y_s|T(x_s)] - E_{Q_t(y_t|x_t)}[y_t|T(x_t)] \right\|^2 \quad \text{condition distribution}$$

迭代伪标签优化策略 来迭代优化 变换T和分类器f

(1) PCA: 先用PCA降维，进行data reconstruction $n = n_s + n_t$ (concat)

$$\max_{A^TA=I} tr(A^T X H X^T A) \quad (2) \quad X = [x_1, x_2, \cdots, x_n] \in R^{m\times n} \ (n\text{个sample}, m\text{维})$$

$$H = I - \frac{1}{n}11^T \quad (n = n_s + n_t)$$

正交变换矩阵A

X的协方差矩阵 XHX^T 降维后的数据 $Z = [z_1, \cdots, z_n] = A^T X$ $A \in R^{m\times k}$

$Z \in R^{k\times n}$ 所以，最大化的是降维后数据的协方差矩阵 $ZHZ^T = A^T X H X^T A$

A: 正交变换矩阵(要求的矩阵)

特征分解 $XHX^TA = A\Phi$ $\Phi = diag(\phi_1, \phi_2, \cdots, \phi_k) \in R^{k\times k}$ 前k个最大特征值

(2) margin distribution Adaptation

用MMD衡量经PCA处理后的域间的边缘分布差异

$$\left\| \frac{1}{n_s}\sum_{i=1}^{n_s} z_i - \frac{1}{n_t}\sum_{j=n_s+1}^{n_s+n_t} z_j \right\|^2$$

$$\left\| \frac{1}{n_s}\sum_{i=1}^{n_s} A^T x_i - \frac{1}{n_t}\sum_{j=n_s+1}^{n_s+n_t} A^T x_j \right\|^2 = tr(A^T X M_0 X^T A) \quad (3) \ \text{最小化}$$

$$(M_0)_{ij} = \begin{cases} \frac{1}{n_s n_s}, & x_i, x_j \in D_s \\ \frac{1}{n_t n_t}, & x_i, x_j \in D_t \\ -\frac{1}{n_s n_t}, & \text{otherwise} \end{cases} \quad (4)$$

2019/10/13 20:36

$$Q_s(y_s|x_s) = \frac{Q_s(x_s|y_s)P_s(y_s)}{P_s(x_s)}$$ $\rightarrow$ (已经minimise了?)

$Q_s(y_s|x_s)$ and $Q_t(y_t|x_t)$ $\downarrow$ $Q_s(x_s|y_s)$ and $Q_t(x_t|y_t)$

(3) Conditional Distribution Adaptation

生成 target data 的伪标签

将在有标签的源域数据上训练的基本分类器,应用到无标签目标域数据

用MMD(修正)衡量类条件分布 $Q_s(x_s|y_s=c)$ 和 $Q_t(x_t|y_t=c)$ 之间的距离

$$\left\| \frac{1}{n_s^{(c)}} \sum_{x_i \in D_s^{(c)}} A^T x_i - \frac{1}{n_t^{(c)}} \sum_{x_j \in D_t^{(c)}} A^T x_j \right\|^2 = tr(A^T X M_c X^T A) \quad (5)$$

$D_s^{(c)} = \{ x_i : x_i \in D_s \wedge y(x_i) = c \}$

$D_t^{(c)} = \{ x_j : x_j \in D_t \wedge \hat{y}(x_j) = c \}$

类间 class $c \in \{1, \cdots, C\}$

MMD矩阵 M_c

$n_s^{(c)} = |D_s^{(c)}|$

$n_t^{(c)} = |D_t^{(c)}|$

$$(M_c)_{ij} = \begin{cases} \frac{1}{n_s^{(c)} n_s^{(c)}}, & x_i, x_j \in D_s^{(c)} \\ \frac{1}{n_t^{(c)} n_t^{(c)}}, & x_i, x_j \in D_t^{(c)} \\ -\frac{1}{n_s^{(c)} n_t^{(c)}}, & \begin{cases} x_i \in D_s^{(c)}, x_j \in D_t^{(c)} \\ x_j \in D_s^{(c)}, x_i \in D_t^{(c)} \end{cases} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

最小化(3)和(5)来最大化(2)

将(3).(5)式,合并到(2)式中

optimization

$$\min_{A^T X H X^T A = I} \sum_{c=0}^{C} tr(A^T X M_c X^T A) + \lambda \|A\|_F^2 \quad (7)$$

$+ tr(\lambda A^T A)$ 广义瑞利商

kernelization 核化

$x \rightarrow \phi(x)$ $\phi(X) = [\phi(x_1), \cdots, \phi(x_n)]$, $K = \phi(X)^T \phi(X) \in R^{n \times n}$

kernel-JDA

$$\min_{A^T K H K^T A = I} \sum_{c=0}^{C} tr(A^T K M_c K^T A) + \lambda \|A\|_F^2$$

$A \in R^{n \times k}$

2019/10/13 20:37

$tr(A^T X \sum_{c=0}^{C} M_c X^T A + \lambda A^T A) + tr((I - A^T X H X^T A)\Phi)$

迭代　用JDA为目标域数据生成label，然后以此作为伪标签，迭代运行JDA直到收敛

对(7)式求拉格朗日函数

$$L = tr(A^T(X\sum_{c=0}^{C} M_c X^T + \lambda I)A) + tr((I - A^T X H X^T A)\Phi) \quad (9)$$

$\Phi = diag(\phi_1, \phi_2, \cdots, \phi_k) \in R^{k\times k}$

令 $\frac{\partial L}{\partial A} = 0$，得

$$(X\sum_{c=0}^{C} M_c X^T + \lambda I)A = XHX^T A\Phi \quad (10)$$

求(10)式的前k个最小特征向量

求的是A.　$A*V = B*V*D$ 广义特征分解

$(\quad)A \quad XHX^T \quad A \quad \Phi$

Computational Complexity

(10)特征分解 ← 构造MMD矩阵

$O(Tkm^2 + TCn^2 + Tmn)$ → 训练分类器

$dL = tr(d(A^T)(X\sum_{c=0}^{C} M_c X^T + \lambda I)A) + tr(A^T(X\sum_{c=0}^{C} M_c X^T + \lambda I)dA) - tr(d(A^T)XHX^T A\Phi + A^T XHX^T dA\Phi)$

$= tr(A^T(X\sum_{c=0}^{C} M_c X^T + \lambda I))dA + tr(A^T(X\sum_{c=0}^{C} M_c X^T + \lambda I))dA - tr(\Phi A^T XHX^T)dA - tr(\Phi A^T XHX^T)dA$

$= 2tr(A^T(X\sum_{c=0}^{C} M_c X^T + \lambda I))dA - 2tr(\Phi A^T XHX^T)dA$

$= 2tr((X\sum_{c=0}^{C} M_c X^T + \lambda I)A)dA - 2tr(XHX^T A\Phi)dA$

$= 2tr((X\sum_{c=0}^{C} M_c X^T + \lambda I)A - XHX^T A\Phi)dA$

2019/10/13 20:38

充分统计量
类条件概率
瑞利商

充分统计量

对于统计量 $t=T(x)$，若数据 x 在已知 $t=T(x)$ 时的条件分布不依赖于参数 θ，则称其是关于参数 θ 的充分统计量

一个关于样本集 D 的函数 s，其中包含了能有助于估计某种参数 θ 的全部相关信息　$p(\theta|s,D)=p(\theta|s)$

扩展MMD

MMD与PCA集成

最小化输入数据的重构误差来学习

TCA是JDA的在 $C=0$ 的情况下的特例

Input: Data $X(X_s, X_t)$, Y_s, k, λ.

Output: A, Z, classifier f.

Begin

Construct MMD matrix M_0 by Eq.(4), set $\{M_c := 0\}_{c=1}^{C}$ （以1NN初始化一个伪标签来构造 $\{M_c\}$）

repeat

对eq(10)进行特征分解，取前k个最小的特征向量构成矩阵A　$Z := A^T X$

训练一个标准分类器 f 在 $\{(A^T x_i, y_i)\}_{i=1}^{n_s}$ 上训练，

在 $(A^T x_j)_{j=n_s+1}^{n_s+n_t}$ 上预测生成伪标签 $\{\hat{y}_j = f(A^T x_j)\}_{j=n_s+1}^{n_s+n_t}$ 并更新（实际预测得到的伪标签

通过(1)式更新MMD矩阵 $\{M_c\}_{c=1}^{C}$　作为下一次迭代的输入）

until Convergence

2019/10/13 20:38