

Discovering Low-Rank Shared Concept Space for Adapting Text Mining Models

[Pattern Analysis and Machine Intelligence, IEEE Transactions on](#)

Author(s): [Chen, B.](#)

The Chinese University of Hong Kong, Hong Kong

[Lam, W.](#); [Tsang, I.](#); [Wong, T.](#)

Volume: PP , [Issue: 99](#)

Page(s): 1



还是考虑域迁移的问题。给一个源域都是标记了的样本 $D_S = \{(x_i, y_i)\}_{i=1}^{n_1}$ ，还有一个目标域都是没标记的 $D_T = \{x_i\}_{i=n_1+1}^{n_1+n_2}$ 。现在就是想用源域样本学习一个分类器来分类目标的样本。

咋设计分类器呢？

首先我们要本样本映射到一个新的空间 $x \rightarrow \tilde{x}$ ：

$$\tilde{x} = \begin{cases} x, \\ \mathbf{K}(\cdot, x) \end{cases}$$

- 可以是线性的 d 维原始空间: x

- 或者是跟所有样本的核函数值的 N 维空间：

$$x \mapsto \mathbf{K}_\Psi(\cdot, x)|_{x_1, \dots, x_N} = (\mathbf{K}_\Psi(x_1, x), \dots, \mathbf{K}_\Psi(x_N, x))$$

假设有 m 个类别，我们为每个类设计一个分类器，这分类器由两部分组成，

1. 首先一部分是直接乘上一个分类器系数 $u_l^\top \tilde{x}$ ，
2. 然后还有一部分是先把 $\tilde{x}$ (d 维) 通过一个降维矩阵 $\Theta (r \times d)$ 进行降维到 $\Theta \tilde{x}$ (r 维)，然后再乘上一个分类系数 $v_l^\top \Theta \tilde{x}$ 。

这样每个类就有 m 个分类器 $\{f_\theta^l(x)\}_{l=1}^m$ ：

$$w_l = u_l + \Theta^\top v_l$$

$$f_\theta^l(x) = u_l^\top \tilde{x} + v_l^\top \Theta \tilde{x} = w_l^\top \tilde{x},$$

现在就是如何学习分类器参数 u_l ， v_l 和 Θ 的问题了。

为了学习参数，要考虑以下三个问题：

- 首先给出源域的样本的标号 $Y_{il} = \begin{cases} 1 & \text{if the } i\text{-th sample belongs to the } l\text{-th class} \\ -1 & \text{if it is labeled as others} \end{cases}$

我们要让分类的结果 $w_l^\top \tilde{x}_i$ 跟 Y_{il} 之间的误差 $L(w_l^\top \tilde{x}_i, Y_{il})$ 尽量的小：

$$\text{Min} \sum_{l=1}^m \sum_{i=1}^{n_1} L(w_l^\top \tilde{x}_i, Y_{il})$$

分类误差有各种口味的测度

- Square loss.

$$L(f(x), y) = (y - f(x))^2 = (1 - yf(x))^2$$

- Hinge loss.

$$L(f(x), y) = \max(0, 1 - yf(x)).$$

- Logistic loss.

$$L(f(x), y) = \ln(1 + e^{-yf(x)})$$

- 其次为了控制分类器的复杂度，还要对每个分类器的第一个系数进行规整：

$$\text{Min} \|u_l\|^2$$

- 最后，要减少降维后的源域和目标域中的样本的分布的差异，要求降维后的俩域的数据 $\phi(x) = \Theta \tilde{x}$ 的均值尽量接近：

$$\text{Min} \left\| \frac{1}{n_2} \sum_{x' \in D_T} \phi(x') - \frac{1}{n_1} \sum_{x \in D_S} \phi(x) \right\|^2$$

$$= \text{tr}(KD) \quad D_{ij} = \begin{cases} \frac{1}{n_1^2} & \text{when } x_i, x_j \in D_S \\ \frac{1}{n_2^2} & \text{when } x_i, x_j \in D_T \\ \frac{-1}{n_1 n_2} & \text{otherwise.} \end{cases}$$

$$K_{ij} = \phi(x_i)^\top \phi(x_j)$$

$$K = \tilde{X}^\top \Theta^\top \Theta \tilde{X}$$

$$= \text{tr}(\tilde{X}^\top \Theta^\top \Theta \tilde{X} D)$$

$$= \text{tr}(\Theta \tilde{X} D \tilde{X}^\top \Theta^\top)$$

$$\bar{K} = \begin{cases} XDX^\top, & (\text{linear}) \\ KDK, & (\text{kernel}) \end{cases}$$

$$= \text{tr}(\Theta \bar{K} \Theta^\top)$$

综合上面的三项，得到目标函数，注意这里通过 $w_l = u_l + \Theta^T v_l$ 把 u_l 偷换成了 w_l

$$\begin{aligned} \min_{\Theta, w, v} & \sum_{l=1}^m \left(\sum_{i=1}^{n_1} L(w_l^T \tilde{x}_i, Y_{il}) + \alpha \|w_l - \Theta^T v_l\|^2 \right) + \beta \text{tr}(\Theta \bar{K} \Theta^T) \\ \text{s.t.} & \Theta \Theta^T = I_{r \times r}, \end{aligned} \quad (21)$$

➤ fixed (Θ, v) ，然后 Solving optimal w ，问题变成了

$$\arg \min_{w_l} \sum_{i=1}^{n_1} L(w_l^T \tilde{x}_i, Y_{il}) + \alpha \|w_l - \Theta^T v_l\|^2$$

这个问题可以通过 conjugate gradient (CG), stochastic gradient descending (SGD) 等算法优化，如果是 $L(f(x), y) = (y - f(x))^2$ ，那直接就有闭合解。

➤ fixed w ，然后 Solving optimal (Θ^*, v^*) ，问题变成了：

$$\begin{aligned} \min_{\Theta, v} & \sum_{l=1}^m \alpha \|w_l^* - \Theta^T v_l\|^2 + \beta \text{tr}(\Theta \bar{K} \Theta^T) \\ \text{s.t.} & \Theta \Theta^T = I_{r \times r}, \end{aligned}$$

对任意的 Θ 求 v_l^* ，直接求导置零，加上 $\Theta \Theta^T = I_{r \times r}$ 的约束，那么 $v_l^* = \Theta w_l^*$ ，这样就把 v_l^* 变成了 Θ 的函数，消去了一个变量 v_l^* ，只剩下了 Θ 。

把 $v_l^* = \Theta w_l^*$ 带进去 $\|w_l^* - \Theta^T v_l\|^2$

$$\|w_l^* - \Theta^T v_l\|^2$$

$$= \|w - \Theta^T \Theta w\|^2$$

$$= \|(I - \Theta^T \Theta) w\|^2$$

$$= \text{Tr}[w' (I - \Theta^T \Theta)' (I - \Theta^T \Theta) w]$$

$$= \text{Tr}[w' (I - 2 \Theta^T \Theta + \Theta^T \Theta \Theta^T \Theta) w]$$

$$= \text{Tr}[w' (I - 2 \Theta^T \Theta + \Theta^T \Theta) w]$$

$$= \text{Tr}[w' (I - \Theta^T \Theta) w]$$

$$= \text{Tr}[w' w] - \text{Tr}[w' \Theta^T \Theta w]$$

$$= \text{Tr}[w' w] - \text{Tr}[\Theta w w' \Theta^T]$$

$$= \text{Tr}[w' w] - \text{tr}(\Theta w w^* \Theta^T)$$
 去掉跟 Θ 无关的第一项

$$\min_{\Theta} - \sum_{l=1}^m \alpha \text{tr}(\Theta w_l^* w_l^{*\top} \Theta^\top) + \beta \text{tr}(\Theta \bar{K} \Theta^\top), \text{ s.t. } \Theta \Theta^\top = I$$



$$\min_{\Theta} \text{tr}(\Theta(\beta \bar{K} - \alpha w^* w^{*\top}) \Theta^\top), \text{ s.t. } \Theta \Theta^\top = I$$

这个问题就很容易解了，直接对 $\beta \bar{K} - \alpha w^* w^{*\top}$ 进行特征分解

$$\beta \bar{K} - \alpha w^* w^{*\top} = H \Lambda H^\top,$$

然后取特征值最小的 r 个特征向量就是 Θ^* 了。

最后上算法：

Input: Data samples $\{(x_i, y_i)\}_{i=1}^{n_1}$ in source domain D_S ;
 Data samples $\{(x_i)\}_{i=n_1+1}^{n_1+n_2}$ in target domain D_T ;
 Linear or nonlinear feature map Φ .
Parameters: r and trade-off coefficients (α, β) .
Output: Optimal concept subspace projection Θ , the corresponding adaptive classifiers $\{v_l\}_{l=1}^m$ in the embedded space and the general classifiers $\{w_l\}_{l=1}^m$

Initialize Construct $\tilde{x}_i$ as in Equation (8) and $\bar{K}$ as in Equation (20), set $\Theta = I$, $v_l = 0, l = 1, \dots, m$.

$$\tilde{x} = \begin{cases} x, & \text{(linear)} \\ K(\cdot, x), & \text{(kernel)} \end{cases} \quad (8) \quad \bar{K} = \begin{cases} XDX^\top, & \text{(linear)} \\ KDK, & \text{(kernel)} \end{cases} \quad (20)$$

repeat
 1 for $l = 1$ to m do
 with fixed $(\Theta, \{v_l\})$, solve the optimization problem $\{w_l^*\} = \arg \min_{w_l} \sum_{i=1}^{n_1} L(w_l^\top \tilde{x}_i, Y_{il}) + \alpha \|w_l - \Theta^\top v_l\|^2$
 end
 2 Do the eigen-decomposition $\beta \bar{K} - \alpha w^* w^{*\top} = H \Lambda H^\top$ (with the diagonals of Λ in ascending order),
 and let the rows of Θ^* be the first r rows of H .
 3 Compute $v_l = \Theta w_l, l = 1, \dots, m$.
until convergence

嗯简单明了。

启发：

- 1: 分类的构造不一定非得在原始的数据空间，或者降维后的空间，可以两者兼而有之；
- 2: 多类分类的问题可以明确的提出来，忽悠一下；
- 3: 之间还讲过对不同的 source domain 的数据点不同的权重，来算两个 domain 的 match. 这里也可以用上。

4: 文章中没有提到 manifold, 也没提到 target label 的信息, 我们可以也加入进去。

最后上作者王道, 中科院毕业后去香港中文干了博士, 现在自己创业了



Bo Chen received the Ph.D. degree in Systems Engineering and Engineering Management from The Chinese University of Hong Kong in 2011, and M.S. degree in Applied Mathematics from the Graduate University of Chinese Academy of Science in 2007. His research interests include statistical machine learning, data mining, and optimization theory and applications.