DOMAIN INVARIANT TRANSFER KERNEL LEARNING

Guide: Prof. Harish Karnick

Md Enayat Ullah, 12407
Abheet Aggarwal, 12012

December 2, 2015

Indian Institute of Technology Kanpur

MOTIVATION

MOTIVATION

- Generalization Error Bound for identical probability distributions is guaranteed by Statistical Learning Theory.[1]
- Big data era resulted in proliferation of huge amount of hetrogenous data.
- Performance drops significantly when standard supervised classifiers are evaluated on datasets outside their domain.[2]

LITERATURE REVIEW

Distribution Discrepancy(parametric)

- Kullback-Leibler Divergence, Bregman divergence:

$$D_{KL}(P, Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

  Problem: Density-estimation: non-trivial.

Distribution Discrepancy(non-parametric)

- **Theorem:** Let p and q be probability measures and $\mathcal{H}$ be a universal RKHS, then $MMD(p, q) = 0$ iff $p = q$.
- $$MMD(p, q) \triangleq \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{z \sim q}[f(z)]\right)$$
  $$MMD(\mathcal{X}, \mathcal{Z}) \triangleq \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(z_j) \right\|_{\mathcal{H}}$$
- ϕ: Encapsulates the higher-order statistics.
- **Problem:** Involves an intermediate SDP step $\sim O(n^{6.5})$ Computationally Prohibitive.

Other Approaches

- Multiple Kernel Learning(MKL): Ensemble of pre-computed kernels.[3]
  Problem: Inadequate to fully encode the data distribution
- Surrogate Kernel Matching(SKM): Linearly transforms the source kernel onto the eigenspace of target kernel.[4]
  Problem: Cannot capture non-linear kernel maps.

APPROACH

Problem Definition: Domain

- **Domain**: A domain $\mathcal{D}$ is composed of an d-dimensional feature space $\mathcal{F}$ and a marginal probability distribution $P(x)$ i.e. $\mathcal{D} = \{\mathcal{F}, P(x)\}, x \in \mathcal{F}$.

Problem Definition: TKL

- **Transfer Kernel Learning**: Given a labeled source domain $\mathcal{Z} = \{(z_1, y_1), ...(z_m, y_m)\}$ and an unlabeled target domain $\mathcal{X} = \{x_1, ..., x_n\}$ with $\mathcal{F}_{\mathcal{Z}} = \mathcal{F}_{\mathcal{X}}, \mathcal{Y}_{\mathcal{Z}} = \mathcal{Y}_{\mathcal{X}}$, we learn a domain-invariant kernel $k(z, x) = \langle \phi(z), \phi(x) \rangle$ such that $P(\phi(z)) \simeq P(\phi(x))$.
- **Problem:** ϕ cannot be explicitly represented.

Overview

- $P(\phi(X)) \sim P(\phi(Z)) \implies K_{\mathcal{X}} \sim K_{\mathcal{Z}}$ [4]
  Problem: Empirical(Data-dependent) kernel matrices, different dimensions $K_{\mathcal{Z}} \in \mathbb{R}^{m\times m}$, $K_{\mathcal{X}} \in \mathbb{R}^{n\times n}$

- **Solution:**
  - Generate $\overline{K}_{\mathcal{Z}}$ extrapolated, using the eigensystem of $K_{\mathcal{X}}$(embodies the structure of X) - Nystrom Approximation
  - Match extrapolated $\overline{K}_{\mathcal{Z}}$ to ground truth $K_{\mathcal{Z}}$ to learn hyperparameters - Spectral Kernel design

Nystom Approximation

- **Mercers Theorem**: Let $k(z,x)$ be a continuous symmetric non-negative function which is positive semi-definite and square integrable w.r.t. distribution $p(x)$, then

$$k(z,x) = \sum_{i=1}^{\infty} \lambda_i \phi_i(z)\phi_i(x)$$

The eigenvalues λ_i's and orthonormal eigenfunctions ϕ_i's are the solutions of:

$$\int k(z,x)\phi_i(x)p(x)dx = \lambda_i \phi_i(z)$$

Assumption: $\mathcal{X}$ and $\mathcal{Z}$ are identically distributed.

Nystom Approximation

- Nystrom Approximation(Quadrature formulae)[5] :

$$\sum_{j=1}^{n} \frac{k(z, x_j)\phi_i(x_j)}{n} \simeq \lambda_i \phi_i(z)$$

- $\overline{K}_{\mathcal{Z}} = \overline{\Phi}_{\mathcal{Z}} \Lambda_{\mathcal{X}} \overline{\Phi}_{\mathcal{Z}}^T$
- However, Nystrom Approximation is only valid of identical distributions.
- Nystrom Approximation Error(NAE) essentially embodies MMD, and in case of different distributions, NAE tends to be very large.
- Minimizing NAE $\equiv K_{\mathcal{X}} = K_{\mathcal{Z}} \equiv P(\phi(x)) = P(\phi(z))$

Spectral Kernel Design

- **Theorem**:If a positive semi-definite kernel matrix $\mathbb{K} \in \mathbb{R}_{n\times n}$ has eigensystem $\{\gamma_i, \phi_i\}_{i=1}^{n}, \gamma_1 \geq ... \geq \gamma_n \geq 0$, then the family of matrices
$$K_\lambda = \sum_{i=1}^{n} \lambda_i \phi_i \phi_i^T, \lambda_1 \geq ... \geq \lambda_n \geq 0$$
will produce PSD kernels with K_λ as kernel matrices
- How is it different from MKL?

EigenSystem Relaxation

- Learnable parameters: $\overline{K}_{\mathcal{Z}} = \overline{\Phi}_{\mathcal{Z}} \Lambda \overline{\Phi}_{\mathcal{Z}}^{T}$
- Kernel matching across domains

$$\min_{\Lambda} \left\| \overline{K}_{\mathcal{Z}} - K_{\mathcal{Z}} \right\|_{\mathcal{F}}^{2} = \left\| \overline{\Phi}_{\mathcal{Z}} \Lambda \overline{\Phi}_{\mathcal{Z}}^{T} - K_{\mathcal{Z}} \right\|_{\mathcal{F}}^{2}$$

$$\lambda_i \geq \zeta \lambda_{i+1}, i = 1, \ldots n-1$$

$$\lambda_i \geq 0, i = 1, \ldots n$$

- $\zeta(DampingFactor) \geq 1$
  - Eigenspecturm of PSD follows Power law.
  - Larger eigenvectors contributes more to the knowledge transfer.

IMPLEMENTATION

QP

- QP problem:

$$\min_{\lambda}(\lambda^T \mathbf{Q}\lambda - 2\mathbf{r}^T\lambda)$$
$$\mathbf{C}\lambda \geq \mathbf{0}$$
$$\lambda \geq \mathbf{0}$$

- Where:

$$\mathbf{Q} = (\overline{\Phi}_{\mathcal{Z}}^T \overline{\Phi}_{\mathcal{Z}}) \circ (\overline{\Phi}_{\mathcal{Z}}^T \overline{\Phi}_{\mathcal{Z}}^T)$$
$$\mathbf{r} = diag(\overline{\Phi}_Z^T \overline{\Phi}_Z^T)$$
$$\mathbf{C} = \mathbf{I} - \zeta \bar{\mathbf{I}}$$

- **Improvement**: Real-world data usually exhibit the eigengap property
  $r = min(500, n)$. Take $\lambda \in \mathbb{R}^{r \times 1}$.

Scalable Implementation

- **Intuition:** Why not extrapolate the Kernel matrix $K_{\mathcal{X}}$ using a small sample of $\mathcal{X}$?
- $\Phi_{\mathcal{X}} \simeq K_{\mathcal{X}\hat{\mathcal{X}}} \Phi_{\hat{\mathcal{X}}} \lambda_{\hat{\mathcal{X}}}^{-1}$
  $\overline{\Phi}_{\hat{\mathcal{Z}}} \simeq K_{\hat{\mathcal{Z}}\mathcal{X}} \Phi_{\mathcal{X}} \lambda_{\hat{\mathcal{X}}}^{-1}$
- Cross domain Nystrom Approximation: $\overline{\Phi}_{\mathcal{Z}} \simeq K_{\mathcal{Z}\hat{\mathcal{Z}}} \overline{\Phi}_{\hat{\mathcal{Z}}} \Lambda_{\hat{\mathcal{X}}}^{-1}$

Support Vector Machines

- Trained on source kernel matrix $\overline{K}_{\mathcal{Z}} = \overline{\Phi}_{\mathcal{Z}} \Lambda \overline{\Phi}_{\mathcal{Z}}^{T}$
- Applied on the cross-domain kernel matrix $\overline{K}_{\mathcal{X}\mathcal{Z}} = \Phi_{\mathcal{X}} \Lambda \overline{\Phi}_{\mathcal{Z}}^{T}$

$$y_{\mathcal{X}} = \overline{K}_{\mathcal{X}\mathcal{Z}}(\alpha \circ y_{\mathcal{Z}}) + b$$

Computational Complexity

Algorithm: Transfer Kernel Learning	
Compute $\mathcal{K}_{\mathcal{Z}}, \mathcal{K}_{\mathcal{X}}, \mathcal{K}_{\mathcal{Z}\mathcal{X}}$ by kernel k	$O(d(m+n)^2)$
Eigendecompose $\mathcal{K}_{\mathcal{X}}$ for $\{\Lambda_{\mathcal{X}}, \Phi_{\mathcal{X}}\}$	$O(rn^2)$
Extrapolate for source eigensystem $\overline{\Phi}_{\mathcal{Z}}$	$O(rmn)$
Solve QP problem for eigenspectrum λ	$O(rn^2 + r^3)$

Overall Complexity: $O(d+r)(m+n)^2$

Approximation Error Analysis

- $\epsilon_{Nys} = \|K_{\mathcal{Z}} - K_{\mathcal{Z}\mathcal{X}} K_{\mathbb{X}}^{-1} K_{\mathcal{X}\mathcal{Z}}\|_{\mathcal{F}}$
- $\epsilon_{TKL} = \|\overline{\Phi}_{\mathcal{Z}} \Lambda \overline{\Phi}_{\mathcal{Z}}^{T} - K_{\mathcal{Z}}\|$
- $\epsilon_{TKL} \leq \epsilon_{Nys} \leq 4m\sqrt[2]{C_k mn\epsilon} + C_k mn\epsilon \|K_{\mathcal{X}}^{-1}\|_{\mathcal{F}}$

RESULTS

Results, Experiments

- Dataset
  - Text
    - 20-Newsgroups: 4 sub-categories
    - Reuters: 4 sub-categories
  - Images
    - Caltech+Amazon : 256 and 31 sub-categories

Results, Experiments

Tuned Parameters: $C = 10, \zeta = 2$

Dataset	SVM	TKL
orgs vs people	69.24	76.40
orgs vs place	63.71	75.11
comp vs rec	85.51	90.44
comp vs sci	74.23	83.42
Amazon vs Caltech	62.90	69.80

Table: Accuracies

Conclusion and Future Work

- Conclusions:
  - Domain-invariant kernel learned by directly matching source and target distributions in RKHS.
  - Learned a family of spectral kernals extrapolated by target eigenspace by minimizing the NAE.
  - Outperforms the standard SVM on benchmark datasets.
- Future Work:
  - Non power law damping constraints.
  - $r = min(500, n)$ eigenvector selection method can be improved.

References I

- V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.
- S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.
- L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 465–479, 2012.

References II

- K. Zhang, V. Zheng, Q. Wang, J. Kwok, Q. Yang, and I. Marsic, "Covariate shift in hilbert space: A solution via sorrogate kernels," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 388–395, 2013.
- K. Zhang, I. W. Tsang, and J. T. Kwok, "Improved nyström low-rank approximation and error analysis," in *Proceedings of the 25th international conference on Machine learning*, pp. 1232–1239, ACM, 2008.

Questions?