Co-regularization Based Analysis of Feature Sharing Algorithms

Abhishek Kumar, Avishek Saha, Hal Daumé III, Tom Fletcher, Suresh Venkatasubramanian
School Of Computing, University Of Utah
{abhik,avishek,hal,fletcher,suresh}@cs.utah.edu

1 Introduction

A common approach in domain adaptation (DA) [1] and multitask learning (MTL) [2] is to create an expanded feature representation by sharing features across domains (in DA) or across tasks (in MTL) and then learning a classifier over this expanded feature set. In this paper, we refer to such techniques as *feature sharing algorithms (FSA)*. One such FSA is EASYADAPT [1], which takes each feature in the original problem and replicates it three times: general, source-specific and target-specific. In this *extended* feature space the source data will contain general and source-specific features whereas the target data will contain general and target-specific features. EASYADAPT is simple, easy to implement as a preprocessing step and outperforms many existing techniques [1, 3], namely, SOURCEONLY (target hypothesis trained on labeled source data only), TARGETONLY (target hypothesis trained on labeled target data only), ALL (combination of source and target labeled data) and PRIOR (target hypothesis trained with SOURCEONLY as a *prior* on the weight vector) [4]. However, a theoretical analysis of why EASYADAPT performs better than the other aforementioned approaches is clearly missing.

In this abstract, we present such an analysis. In order to achieve our goal, we model EASYADAPT in terms of *co-regularization*. This is an idea that originated in the context of multiview learning and for which there exists some theoretical analysis [5].

We denote source and target empirical errors for some hypothesis h as $\hat{\epsilon}_s(h)$ and $\hat{\epsilon}_t(h)$ and the corresponding expected errors as $\epsilon_s(h)$ and $\epsilon_t(h)$. PRIOR and EASYADAPT optimize the following cost functions:

$$\mathcal{Q}_{PR} = \hat{\epsilon}_t(h) + \lambda_2||h||^2 + \lambda||h_s - h||^2 \quad \text{where,} \quad h_s = \arg\min_h\{\hat{\epsilon}_s(h) + \lambda_s||h||^2\} \tag{1.1}$$

$$\mathcal{Q}_{EA} = \alpha\hat{\epsilon}_s(h_1) + (1-\alpha)\hat{\epsilon}_t(h_2) + \lambda_1||h_1||^2 + \lambda_2||h_2||^2 + \lambda||h_1 - h_2||^2 \tag{1.2}$$

In the above, we assume that our hypothesis class is comprised of real-valued functions over an RKHS with reproducing kernel $k(\cdot,\cdot), k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Let us also define the kernel matrix and partition it corresponding to source labeled and target labeled data as $K = \begin{pmatrix} A_{s\times s} & C_{s\times t} \\ C'_{t\times s} & B_{t\times t} \end{pmatrix}$, where A,B,C are kernel submatrices and the subscripts denote the domains whose data have been utilized to construct these submatrices. Assume that the loss function is bounded by 1. Proceeding in a manner similar to [5] (cf. section 3.1) we substitute trivial hypotheses $h = h_1 = h_2 = 0$ in all the cost functions which makes all regularizers and co-regularizers 0. Assuming that the loss function is bounded by 1, we get $\mathcal{Q} \leq 1$ for all cost functions. So we can define the base hypothesis class as: $\mathcal{H} := \{(h_1, h_2) : \lambda_1||h_1||^2 + \lambda_2||h_2||^2 + \lambda||h_1 - h_2||^2 \leq 1\}$.

By fixing $h_1 = h_s$ a priori and making $\lambda_1 = 0$, we can define the target hypothesis classes for PRIOR as: $\mathcal{J}^t_{PR} := \{h_2 : \mathcal{X} \mapsto \mathbb{R}, (h_s, h_2) \in \mathcal{H}, \lambda_1 = 0\}$. Target hypothesis classes for EASYADAPT is given by: $\mathcal{J}^t_{EA} := \{h_2 : \mathcal{X} \mapsto \mathbb{R}, (h_1, h_2) \in \mathcal{H}\}$. The respective source hypotheses are defined as: $\mathcal{J}^s_{PR} := \{h : \lambda_s||h||^2 \leq 1\}$ and $\mathcal{J}^s_{EA} := \{h_1 : \mathcal{X} \mapsto \mathbb{R}, (h_1, h_2) \in \mathcal{H}\}$

2 Theoretical Results

We present the following two theorems (without proof) which provide upper and lower bounds on the complexity of the source and target hypothesis classes for PRIOR ($\mathcal{J}^t_{PR}$) and EASYADAPT ($\mathcal{J}^t_{EA}$).

Theorem 2.1. *For the hypothesis class $\mathcal{J}^t_{PR}$ if we define $\hat{R}_m(\mathcal{J}^t_{PR}) = E_\sigma \sup_{h_2 \in \mathcal{J}^t_{PR}} |\sum_i \sigma_i h_2(x)|$, then we have,* $\frac{1}{\sqrt[4]{2}} \frac{2C^t_{PR}}{N_t} \leq \hat{R}_m(\mathcal{J}^t_{PR}) \leq \frac{2C^t_{PR}}{N_t}$ *where,* $(C^t_{PR})^2 = \left[1 - ||h_s||^2 \left(\frac{1}{\lambda} + \frac{1}{\lambda_2}\right)^{-1}\right] \frac{1}{\lambda_2 + \lambda} tr(B)$ *and* $||h_s||^2 \leq \frac{1}{\lambda_s}$.

It can be observed that the complexity decreases with increasing norm of source hypothesis h_s. The complexity also decreases with increase in the value of hyperparameters λ and λ_2. The complexity of source hypothesis class for PRIOR has the same form as that of SOURCEONLY, which is given by

$$\hat{R}_m(\mathcal{J}^s_{PR}) \leq \frac{2}{N_s} (tr(A)/\lambda_s)^{1/2} \tag{2.1}$$

Theorem 2.2. *For the hypothesis class $\mathcal{J}^t_{EA}$ if we define $\hat{R}_m(\mathcal{J}^t_{EA}) = E_\sigma \sup_{h_2 \in \mathcal{J}^t_{EA}} |\sum_i \sigma_i h_2(x)|$, then we have,* $\frac{1}{\sqrt[4]{2}} \frac{2C^t_{EA}}{N_t} \leq \hat{R}_m(\mathcal{J}^t_{EA}) \leq \frac{2C^t_{EA}}{N_t}$ *where,* $(C^t_{EA})^2 = \left(\frac{1}{\lambda_2 + \left(\frac{1}{\lambda_1} + \frac{1}{\lambda}\right)^{-1}}\right) tr(B)$.

The complexity of EASYADAPT class also decreases with an increase in the values of hyperparameters. It decreases more rapidly with change in λ_2 compared to λ and λ_1. The kernel block submatrix corresponding to source samples does not appear in any of the target class bounds. Due to the symmetry of EASYADAPT cost function in source and target hypothesis, the complexity of source hypothesis class can be bounded by

$$\frac{1}{\sqrt[4]{2}} \frac{2C^s_{EA}}{N_s} \leq \hat{R}_m(\mathcal{J}^s_{EA}) \leq \frac{2C^s_{EA}}{N_s}, \quad \text{where} \quad (C^s_{EA})^2 = \left(\frac{1}{\lambda_1 + \left(\frac{1}{\lambda_2} + \frac{1}{\lambda}\right)^{-1}}\right) tr(A), \tag{2.2}$$

Th. 2.2 is required to obtain the bound on the target hypothesis class of EASYADAPT and subsequently bound the source hypothesis class using symmetry arguments. So, we can compare either of Th. 2.2 or Eq. 2.2 with Eq. 2.1. Assuming, $\lambda_s = \lambda_1$ and all other parameters remaining the same, the upper bound of Eq. 2.2 is smaller than that of Eq. 2.1. Hence, we claim that the complexity of EASYADAPT source class (Eq. 2.2) is less than the complexity of PRIOR source class (Eq. 2.1). Generalization bounds for EASYADAPT and PRIOR can be obtained by plugging in the source class complexities in the rademacher complexity based generalization bound expressions [6]. If we compare the source class complexities of PRIOR and EASYADAPT, it can be easily seen EASYADAPT provides better generalization performance on target. Hence, this framework nicely explains the superior performance of EASYADAPT compared to PRIOR.

A careful analysis of the above results also reveal a *new* notion of domain similarity in terms of the traces of the kernel submatrices A (constructed from source samples) and B (constructed from target samples). If source and target domains are similar then we have $tr(A) \approx tr(B)$, whereas these trace values are considerably different if the domains are far apart. In addition, this notion of domain similarity is computable from finite source and target samples.

3 Discussion

Empirical results in [1] showed that EASYADAPT outperforms PRIOR on a wide variety of tasks, but no theoretical justification was provided for this superior performance. We have theoretically analyzed EASYADAPT and explained its superior performance over PRIOR. It would be interesting to explore whether the obtained bounds can be further tightened by leveraging unlabeled data in EASYADAPT.

References

[1] H. Daumé III, "Frustratingly easy domain adaptation," in *ACL*, Prague, Czech Republic, 2007.

[2] T. Evgeniou and M. Pontil, "Regularized multitask learning," in *KDD*. New York, NY, USA: ACM, 2004, pp. 109–117.

[3] L. Duan, I. W. Tsang, D. Xu, and T. S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *ICML*. New York, NY, USA: ACM, 2009, pp. 289–296.

[4] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Computer Speech & Language*, vol. 20, no. 4, pp. 382–399, 2006.

[5] D. S. Rosenberg and P. L. Bartlett, "The Rademacher complexity of co-regularized kernel classes," in *AISTATS*, 2007.

[6] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2003.