NOTAM^2: Nonparametric Bayes Multi-Task Multi-View Learning

Hongxia Yang[a] and Jingrui He[b]
[a]Business Analytics and Math Sciences , IBM T.J. Watson Research Center, NY
[b] Department of Computer Science, Stevens Institute of Technology, NJ
email: yangho@us.ibm.com, jingrui.he@stevens.edu

KEYWORDS: Big Data; Gibbs Algorithm; Multi-Task Multi-View Learning; Nonparametric Bayes.

Abstract

Heterogeneous learning refers to addressing problems with multiple types of heterogeneity, e.g., task heterogeneity, view heterogeneity, etc. It finds abundant applications in cross-lingual document classification, cross-domain sentiment analysis, web image classification, etc. Traditional approaches handle different types of heterogeneity *separately* via multi-task learning, multi-view learning, etc. More recently, researchers start to *jointly* model different types of heterogeneity in order to improve the learning performance with limited training data. In this paper, we advance state-of-the-art in heterogeneous learning by jointly modeling task and view relatedness via nonparametric Bayes method. To be specific, we model task relatedness using normal penalty with sparse covariances to couple multiple tasks and view relatedness using matrix Dirichlet process. We also propose *NOTAM*^{2} algorithm, which is based on an efficient Gibbs algorithm. Experimental results demonstrate the effectiveness of *NOTAM*^{2}.

1 Introduction

Nowadays, we are facing *big data* in a variety of areas, such as social media, manufacturing, traffic analytics, etc. A common challenge in these *big data* areas is how to handle multiple types of data heterogeneity. For example, in social media, we may have micro-blogs coming from heterogeneous sources, such as Facebook and Twitter, and each micro-blog may be characterized by heterogeneous features, such as key words, hashtags, number of retweets, number of Facebook likes, etc; in manufacturing, we may have products from heterogeneous manufacturing lines, and each product may be characterized by heterogeneous environmental variables, such as temperature, pressure, etc; in traffic analytics, we can collect traffic information from heterogeneous geographic locations (e.g., different states), and for each location, we may have heterogeneous traffic indicators, such as volume, GPS positions, etc.

Recent years have seen growing interest in addressing problems with multiple types of data heterogeneity (Harel and Mannor, 2011; He and Lawrence, 2011; Han et al., 2012; Ding et al., 2012; Zhang and Huan, 2012). In particular, some problems have been formulated as multi-task multi-view learning, or M^2TV learning (He and Lawrence, 2011; Zhang and Huan, 2012), i.e., jointly learning in multiple tasks with partially overlapping or completely different feature spaces. Compared with traditional multi-task learning (O'Sullivan and Thrun, 1996; Yu et al., 2005; Chen et al., 2010, 2011; Zhou et al., 2011), where the feature space is *homogeneous* across different tasks, M^2TV learning is able to handle *heterogeneous* feature spaces; compared with traditional multi-view learning (Blum and Mitchell, 1998; Muslea et al., 2002; Farquhar et al., 2005; Kakade and Foster, 2007; Christoudias et al., 2008), where the examples come from a *homogeneous* task, M^2TV learning is able to leverage *heterogeneous* (related) tasks to improve the learning performance in each task.

A key question in M^2TV learning is how to model the relatedness among multiple tasks/views. Whereas existing work usually assumes that all the tasks/views are related, and mainly focuses on exploring various types of relatedness; in this work, we go one step further, and study (1) if all the tasks/views are related, and (2) how much they are related to each other. This is motivated by the fact that in many real applications, it is often not known a priori if all the tasks/views are equally related or not. In the adversarial cases where some tasks/views are *negatively* related to the others, simply applying the existing methods for M^2TV learning may even harm the performance. Although in traditional multi-task learning, there has already been some work testing the task relatedness (Chen et al., 2010; Zhou et al., 2011; Chen et al., 2011; Zhang and Yeung, 2012), to the best of our knowledge, our work is the first to study this problem in the context of M^2TV learning.

Motivated by the successful application of Bayesian hierarchical modeling in multi-task/multi-view learning (Bakker and Hesks, 2003; Archambeau et al., 2011; Han et al., 2012), we propose a nonparametric Bayes method for M^2TV learning, where task relatedness is modeled via a normal penalty that decomposes the full covariance of matrix elements into the Kronecker product, and view relatedness is modeled via a matrix Dirichlet process. Furthermore, we design the *NOTAM*^{2} algorithm, which stands for *NO*nparame*T*ric b*A*yes M^2tv learning. It is based on an efficient Gibbs algorithm scalable to relatively high dimensions.

The rest of the paper is organized as follows. In Section 2, we proposed the nonparametric Bayes method for M^2TV learning and conclude the paper in Section 3. Due to the page limit, we omit the algorithm and experimental results in this paper.

2 Nonparametric Bayes Method for M^2TV Learning

2.1 Notation

Suppose that we have T tasks and V views in total. For the v^{th} view, there are d^v features. For the t^{th} task ($t = 1, \ldots, T$), there are n^t examples $\mathcal{X}^t = \{\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_{n^t}^t\} \subset \mathbb{R}^{\sum_{v=1}^V d^v}$, and each example $\boldsymbol{x}_s^t = [(\boldsymbol{x}_s^{t1})', \ldots, (\boldsymbol{x}_s^{tV})']'$ with label y_s^t ($s = 1, \ldots, n^t$), where $\boldsymbol{x}_s^{tv} \in \mathbb{R}^{d^v}$ denotes the features from the v^{th} view ($v = 1, \ldots, V$), $()'$ denotes vector transpose, and y_s^t is either discrete for classification problems, or real-valued for regression problems. Notice that if a certain view is missing, the associated features will all be 0. Therefore, our problem setting is essentially the same as in He and Lawrence (2011) where some views are shared by multiple tasks, and some views are task specific. For the sake of clarity, we introduce an indicator matrix $\boldsymbol{I} \in \{1, 0\}^{T \times V \times n_t}$ to mark which view is missing from which task of which example, i.e., $\boldsymbol{I}_s^{tv} = 0$ if the vth view of tth task from sth example is missing and $\boldsymbol{I}_s^{tv} = 1$ otherwise. Throughout the paper we use subscripts to denote examples and superscripts to denote tasks and views.

2.2 Model Formulation

Our proposed model can be decomposed into multiple view models, where each view generates a predictor that can be used to make predictions on future data examples. To be specific, for the t^{th} task ($t = 1, \ldots, T$), we use a mixture linear regression model for the estimated output $\hat{y}_s^t$ ($s = 1, \ldots, n^t$) by averaging the prediction results from all view functions as follows:

$$\hat{y}_s^t = \sum_{v=1}^{V} \{(\boldsymbol{x}_s^{tv})' \boldsymbol{f}^{tv} + \epsilon_s^{tv}\},$$

where $\boldsymbol{f}^{tv} \in \mathbb{R}^{d^v}$ is the coefficient vector, and $\epsilon_s^{tv} \in \mathbb{R}$ is the observational error. Motivated by Yu and Chu (2007), we assume that $\boldsymbol{\epsilon} = \{\epsilon_s^{tv}\} \sim \text{N}(0, K \otimes \text{I}_V)$, where $K \otimes \text{I}_V$ is the kernel function of the Gaussian distribution and I_V is the V by V identity matrix. Here $K \in \mathbb{R}^{T \times T}$ models the task relatedness. To be specific, define a task graph as follows: the graph consists of T nodes with each node representing a single task; let $W \in \mathbb{R}^{T \times T}$ denote the adjacency matrix of the graph, whose element in the t^{th} row and $(t')^{\text{th}}$ column $B_{tt'} = \frac{1}{n^t n^{t'}} \sum_{s=1}^{n^t} \sum_{s'=1}^{n^{t'}} < \boldsymbol{x}_s^t, \boldsymbol{x}_{s'}^{t'} >$, where $t, t' = 1, \ldots, T$. For this graph, the Laplacian $\Delta = D - B$, where $D \in \mathbb{R}^{T \times T}$ is a diagonal matrix, with each diagonal element equal to the row sum of W. Using Δ, we define K as follows: $\forall t, t' = 1, \ldots, T$,

$$K_{tt'} = \left[\beta(\Delta + \frac{1}{\sigma^2} I)\right]^{-1}$$

where β is the positive parameter that controls the overall sharpness of the distributions: large values of β mean that the distribution is more peaked around its mean. For more flexibility, we let $\beta \sim \text{Ga}(a, b)$ and be adapted to the data through adjusting the distribution related parameters, where $\text{Ga}(a, b)$ stands for Gamma distribution with shape parameter a and scale parameter b. σ^2 controls the amount of regularization and we choose a proper prior $\sigma^2 \sim \text{IG}(c, d)$, where $\text{IG}(c, d)$ stands for Inverse-Gamma distribution with shape parameter c and scale parameter d.

We note several important aspects of the proposed Gaussian penalty. First, the task relatedness matrix K depends on the inverse of the regularized graph Laplacian Δ. Therefore, the relatedness between two tasks is global in the sense that it depends on all the tasks. Second, if we also have unlabeled data in addition to the labeled training data, all the unlabeled data can be used to define the adjacency matrix B (since it does not require label information), thus making it more reliable.

On the other hand, we model the coefficient vectors $\boldsymbol{f}^{tv}$ ($t = 1, \ldots, T$, $v = 1, \ldots, V$) through:

$$\begin{pmatrix} \boldsymbol{f}^{t1} \\ \vdots \\ \boldsymbol{f}^{tV} \end{pmatrix} \sim \text{N}\left(\mathbf{0}\ , \begin{bmatrix} \Psi^{11} & \Psi^{12} & \ldots & \Psi^{1V} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi^{V1} & \Psi^{V2} & \ldots & \Psi^{VV} \end{bmatrix}^{-1} \right).$$

Notice that we define the precision matrix instead of the covariance matrix here, which will be beneficial for the computation in the Gibbs steps. We extend the matrix DP prior (Dunson et al., 2008) to define view-specific covariance function $\Psi^{vv'} = \Psi^{v'v}$. In particular, we borrow information by incorporating dependence in the prior distributions for the coefficients $\{\Psi^{vv'}\}$. We start by assuming for $v \geq v' \geq 1$,

$$\Psi^{vv'} \overset{\text{ind}}{\sim} F^{vv'}, \quad \mathcal{F} \sim \mathcal{P},$$

where $\mathcal{F} = \{F_{vv'}, V \geq v \geq v' \geq 1\}$ is a matrix of random probability measures, and $\mathcal{P}$ is a probability measure on $(\Omega, \mathcal{F})$, with Ω being the space of symmetric $V \times V$ matrices, and $\mathcal{F}$ being a σ-algebra of subsets of Ω. The (v, v') element of Ω is a probability measure on $(\boldsymbol{\mathcal{X}^v}, \boldsymbol{\beta^v})$, where $\boldsymbol{\beta^v}$ is a Boreal σ-algebra of subsets of $\boldsymbol{\mathcal{X}^v}$.

Our focus is on the specification of $\mathcal{P}$. Assuming each element in $\mathcal{F}$ has a stick-breaking representation, we let

$$F^{vv'} = \sum_{h=1}^{\infty} \{W_h^{vv'} \prod_{l<h} (1 - W_l^{vv'})\} \delta_{\Theta^{vv'}}, \quad \Theta^{vv'} \overset{\text{ind}}{\sim} G,$$

where $\boldsymbol{W} = \{W^{vv'}, V \geq v \geq v' \geq 1\}$ is an array of random stick-breaking weights, and $\Theta = \{\Theta^{vv'}\}$ is a three dimensional triangular array of random atoms (for simplicity, we assume that $d^v = p$ for $v = 1, \ldots, V$, and if $d^v < p$ we fill in 0 values to make feature lengths equal). The third dimension (p) corresponds to the different predictors of the features, while the triangular matrix corresponds to different clusters.

Dependency within dimensions of $\mathcal{F}$ will be incorporated through dependent stick-breaking weights and a common parametric prior G. For the stick-breaking component, because of the symmetry ($\Psi^{vv'} = \Psi^{v'v}$), we let

$$W_h^{vv'} = \gamma_h^v \gamma_h^{v'}, \quad \gamma_h^v \sim \text{beta}(1, \alpha), \quad \alpha \overset{\text{ind}}{\sim} \text{Ga}(1, \alpha_0), \tag{1}$$

so that the probability $W_h^{vv'}$ is decomposed into the product of γ_h^v and $\gamma_h^{v'}$. In this way, we guarantee the symmetric property: $W_h^{vv'} = W_h^{v'v}$. The definition of γ_h^v ensures that the elements of $\boldsymbol{W} = \{W^{vv'}, V \geq v \geq v' \geq 1\}$ sum to one which makes (1) a valid probability measure. Figure 1 shows the graphical presentation of the proposed model.

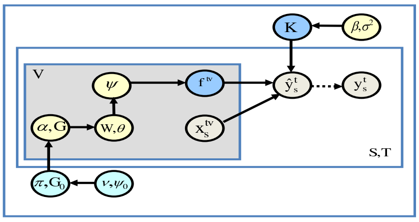

Figure 1: Graphical presentation for nonparametric Bayes multi-task multi-view learning model.

Similarly to Xue et al. (2007), we have the following relationship. For simplicity, here we assume that $V = 4$, and $V_1, \ldots, V_4$ stands for the four different views.

$$\Pr(\Theta^{V_1V_2} = \Theta^{V_1V_3}) = \frac{1}{(\alpha+1)(\alpha+2)-1},$$

$$\lim_{\alpha \to 0} \Pr(\Theta^{V_1V_2} = \Theta^{V_3V_4} | \Theta^{V_1V_4} = \Theta^{V_3V_4}) = \frac{1}{\alpha+1}.$$

The element G is a degenerate distribution:

$$G = \pi \boldsymbol{I}^{\infty} + (1-\pi) G_0, \quad G_0 \sim \text{Inverse-Wishart}(\nu, \Psi_0).$$

So when $\Psi^{vv'}$ falls into the $\boldsymbol{I}^{\infty}$ cluster, the corresponding covariance matrix will be $\boldsymbol{I}^0$(values all 0) and the nonsignificant $\boldsymbol{f}^{tv}$ will be set to 0.

3 Conclusion

In this paper, we propose a nonparametric Bayesian model for addressing problems with dual-heterogeneity, i.e., multiple tasks (task heterogeneity) and multiple views (view heterogeneity). Compared with state-of-the-art which assumes that the tasks/views are equally related, our main contribution is making use of normal penalty with sparse inverse covariances and matrix DP prior to learn from the data: (1) if the tasks/views are related; (2) how much they are related to each other. Furthermore, we design *NOTAM*^{2} algorithm based on an efficient Gibbs algorithm, which constructs predictors for all the tasks leveraging both the multi-task and multi-view nature. Experimental results on several real data sets show that *NOTAM*^{2} outperforms existing methods in M^2TV learning.

References

Archambeau, C., Guo, S., and Zoeter, O. (2011), “Sparse Bayesian Multi-Task Learning,” *NIPS*, .

Bakker, B., and Hesks, T. (2003), “Task clustering and gating for bayesian multitask learning,” *JMLR*, 4, 83–99.

Blum, A., and Mitchell, T. M. (1998), Combining Labeled and Unlabeled Sata with Co-Training,, in *COLT*.

Chen, J., Liu, J., and Ye, J. (2010), Learning incoherent sparse and low-rank patterns from multiple tasks,, in *KDD*, pp. 1179–1188.

Chen, J., Zhou, J., and Ye, J. (2011), Integrating low-rank and group-sparse structures for robust multi-task learning,, in *KDD*, pp. 42–50.

Christoudias, C., Urtasun, R., and Darrell, T. (2008), Multi-View Learning in the Presence of View Disagreement,, in *UAI*, pp. 88–96.

Ding, L., Yilmaz, A., and Yan, R. (2012), “Interactive Image Segmentation Using Dirichlet Process Multiple-View Learning,” *IEEE Transactions on Image Processing*, 21(4), 2119–2129.

Dunson, D., Xue, Y., and Carin, L. (2008), “The matrix stick- breaking process: flexible Bayes meta analysis,” *Journal of the American Statistical Association*, 103, 317C327.

Farquhar, J., Hardoon, D., Meng, H., Shawe-Taylor, J., and Szedmak, S. (2005), “Two view learning: SVM-2K, theory and practice,” *NIPS*, .

Han, S., Liao, X., and Carin, L. (2012), “Cross-Domain Multitask Learning with Latent Probit Models,” *NIPS*, .

Harel, M., and Mannor, S. (2011), Learning from Multiple Outlooks,, in *ICML*, pp. 401–408.

He, J., and Lawrence, R. (2011), A Graphbased Framework for Multi-Task Multi-View Learning,, in *ICML*, pp. 25–32.

Kakade, S. M., and Foster, D. P. (2007), Multi-view Regression Via Canonical Correlation Analysis,, in *COLT*, pp. 82–96.

Muslea, I., Minton, S., and Knoblock, C. A. (2002), Active + Semi-supervised Learning = Robust Multi-View Learning,, in *ICML*, pp. 435–442.

O'Sullivan, J., and Thrun, S. (1996), "Discovering structure in multiple learning tasks: The TC algorithm," *ICML*, pp. 489–497.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007), "Multi-Task Learning for Classification with Dirichlet Process Priors," *Journal of Machine Learning Research*, 8, 35–63.

Yu, K., and Chu, W. (2007), "Guassian Process Models for Link Analysis and Transfer Learning," *NIPS*, .

Yu, K., Schwaighofer, A., and Tresp, V. (2005), "Learning gaussian processes from multiple tasks," *ICML*, .

Zhang, J., and Huan, J. (2012), Inductive multi-task learning with multiple view data,, in *KDD*, pp. 543–551.

Zhang, Y., and Yeung, D.-Y. (2012), "A Convex Formulation for Learning Task Relationships in Multi-Task Learning," *CoRR*, abs/1203.3536.

Zhou, J., Chen, J., and Ye, J. (2011), Clustered Multi-Task Learning Via Alternating Structure Optimization,, in *NIPS*, pp. 702–710.