Simultaneous Learning of Related Tasks

Delaram Motamedvaziri, Venkatesh Saligrama, David Castanon
Department of Electrical Engineering
Boston University — The Center for Information and Systems Engineering (CISE)

Abstract

We present a method for jointly learning $r > 1$ similar classification tasks. Our method builds upon a regularization problem using two regularizers which control the underlying structure of the model from completely unrelated tasks to practically the same tasks. We show that this problem is equivalent to a convex optimization problem.

Introduction

- In real life many learning tasks are related
- Similarity among the tasks can improve learning efficiency and accuracy
- Example: Many personality disorders share almost the same set of symptoms with different severity

	PTSD	SAD	PPD	AvPD	PMD
Irritability	⋆	⋆	★	⋆	⋆
Anhedonia	0	0	0	0	★
Hyper/Insomnia	★	★	⋆	⋆	⋆
Change of appetitie	0	⋆	0	0	0

Table: Mental Disorders and 3-level Symptoms Importance , **PTSD** Posttraumatic Stress Disorder **SAD**: Seasonal Affective Disorder, **PPD**:Paranoid Personality Disorder, **AvPD**: Avoidant Personality Disorder, and **PMD**: Psychotic Major Depression

Problem Setup

- Consider the problem of learning r related classification tasks.
- $\mathbf{x_i} \in \mathbf{R}^p$ is the vector of features, and $\mathbf{y_i} \in \mathbf{R}^r$ is the vector of labels corresponding to each task.
- We are given a set of n data samples $(\mathbf{x_i}, \mathbf{y_i})$ represented as:

$$X = \begin{pmatrix} \mathbf{x}_1{}^T \\ \vdots \\ \mathbf{x}_n{}^T \end{pmatrix} \ , \ Y = \begin{pmatrix} \mathbf{y}_1{}^T \\ \vdots \\ \mathbf{y}_n{}^T \end{pmatrix}$$

- Our goal is to find a mapping $\mathbf{f} : \mathbf{R}^p \mapsto \{0,1\}^r$ that classifies the data with the least possible cost
- We focus our attention on linear mapping of the form:

$$\mathbf{f}(\mathbf{x}, \Theta) = sign((x)^T.\Theta)$$

Previous Works

Previous works [3,4] assume that all tasks share the same set of features and the weights of features are close for all tasks.

However, these methods penalize the "too large", or "too small" elements and smooth out the structure:

$$\Theta = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \bullet & \cdot & \cdot & \cdot \end{pmatrix} \simeq \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

New Approach

Our method relies on the fact that there exists a decomposition of the form $\Theta = B + S$:

$$\Theta = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \bullet & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} + \begin{pmatrix} \cdot & \circ & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \bullet \\ \bullet & \bullet & \cdot & \cdot & \cdot \\ \cdot & \bullet & \cdot & \cdot & \cdot \end{pmatrix}$$

Although Θ does not have a simple structure, with this decomposition, matrix S is sparse and matrix B is block sparse.

Method

To recover the weight matrix Θ more accurately, from less data samples, we need to exploit its underlying structure:

- Element-wise sparsity in S is encouraged by applying l_1 norm regularization term.
- Block sparsity in B is imposed by l_1/l_∞ norm regularization.

Hence, our estimation $\widehat{\Theta} = \widehat{S} + \widehat{B}$ is the solution to the following optimization problem:

$$(\widehat{S}, \widehat{B}) \in argmin\{\mathcal{L}(S, B) + \lambda_b \|B\|_{1,\infty} + \lambda_s \|S\|_{1,1}\}$$

Where $\mathcal{L}(S, B)$ can be any misclassification loss function (e.g. Logistic Regression or Hinge loss).

Theory

Theorem

Suppose the estimated value of parameters $\widehat{S}$, $\widehat{B}$ are obtained from our method where

$$1 < \frac{\lambda_s}{\lambda_b} < r$$

$$\lambda_s > s\sqrt{\frac{\log pr}{n}}\frac{\alpha_s}{2-\alpha_s} \ , \ \lambda_b > s\sqrt{\frac{\log pr}{n}}\frac{\alpha_b}{2-\alpha_b}$$

and

$$n > max\{\frac{10s^3 log(pr)}{C_{min}\alpha_s^2}, \frac{10s^2 r(r\log 10 + logp)}{C_{min}\alpha_b^2}\}$$

Then, with probability

$$1 - 2\exp(-c_1 n) \to 1$$

we have:

(I) $Supp(\widehat{\Theta}) \subseteq Supp(\Theta^*)$

(II) $\|\widehat{\Theta} - \Theta^*\|_{\infty,\infty} \leq \underbrace{\{\sqrt{\frac{s\log(pr)}{C_{min}n}} + \lambda_b D_{max}\}}_{\gamma}$

(III) $Supp(\widehat{\Theta}) = Supp(\Theta^*)$, if $min|\Theta^*_U| > \gamma$

Numeric Results

Synthetic Dataset: This is a 2-task problem, with features being Gaussian random variables and data is generated by the model:

$$y^j = sign(\langle \theta^j, x \rangle) \ j = 1, 2$$

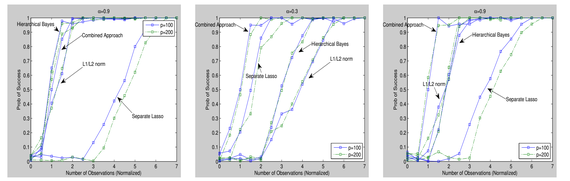

Figure: Comparision of our method with others when: overlap fraction is 0.9, 0.3, and when weights are not necessarily close

SRBCT Dataset: This dataset contains 63 training samples pf patients with 4 tumor classes: the Ewing tumors (EWS), Burkitt lymphoma (BL), neuroblastoma (NB) and rhabdomyosarcoma (RMS). It is tested on 25 test samples.

n	Combined Approach	Separate Lasso
Average Accuracy		
20	0.804	0.730
40	0.926	0.832
63	0.947	0.881
Variance of error		
20	0.521%	0.532%
40	0.553%	0.647%
63	0.562%	0.877%

Yeast Dataset: is formed by micro-array expression data and phylogenetic profiles. There are 1500 genes in the training data set and 917 in the testing dataset. The number of associated genes is 103 with $r = 14$ classification tasks.

n	Combined Approach	Separate Lasso
Average Accuracy		
100	0.553	0.488
500	0.627	0.585
1500	0.698	0.673
support size		
100	193	125
500	231	189
1500	287	244

References

- A. Jalali, P. Ravikumar, S. Sanghavi, C. Ruan, (2010) Dirty Model for Multi-task Learning,
- P. Ravinkumar, & M.J.Wainwright, & J.D.Lafferty, (2010) High-dimensional Ising model selection using l_1 -regularized logistic regression,
- T.Evgeniou, M.Pontil (2004) "Regularized Multi-task Learning",
- M.E., Schnell, E. & Barkai, E. (2009) Simultaneous support recovery in high dimensions: Benefits and perils of block l_1/l_{infty}-regularization