Active Online Multitask Learning

Avishek Saha AVISHEK@CS.UTAH.EDU
Piyush Rai PIYUSH@CS.UTAH.EDU
Hal Daumé III HAL@CS.UTAH.EDU
Suresh Venkatasubramanian SURESH@CS.UTAH.EDU
School of Computing, University of Utah, Salt Lake City, UT, USA

Abstract

In this paper, we propose an online multitask learning framework where the weight vectors are updated in an *adaptive* fashion based on inter-task relatedness. Our work is in contrast with the earlier work on online multitask learning (Cavallanti et al., 2008) where the authors use a *fixed* interaction matrix of tasks to derive (fixed) update rules for all the tasks. In this work, we propose to update this interaction matrix itself in an adaptive fashion so that the weight vector updates are no longer fixed but are instead adaptive. Our framework can be extended to an active learning setting where the informativeness of an incoming instance across all the tasks can be evaluated using this adaptive interaction matrix. Empirical results on standardized datasets show improved performance in terms of accuracy, label complexity and number of mistakes made.

1. Introduction

Multitask learning (Evgeniou et al., 2005) refers to the setting where a set of related tasks are learned together with the goal of improved generalization across all tasks. It becomes especially important if there is a scarcity of labeled examples per task. An interesting case of multitask learning is when examples for various tasks arrive one-at-a-time, and the sequence of examples and the corresponding task index (the task which an incoming example belongs to) is chosen adversarially. Furthermore, this setting poses another challenge if one wants to do active learning (Settles, 2009) in this setting due to the presence of multiple related tasks. In this paper, we propose a framework to address these issues.

Presented at the ICML 2010 *Budgeted Learning Workshop*, Haifa, Israel, 2010.

2. Background

We assume the online multitask learning setting of (Cavallanti et al., 2008) in which, at each time step, the multitask learner receives an example that belongs to one of the K tasks. More specifically, the learner receives $\{(x_t, y_t), i_t\}$ where $x_t \in R^d$ is the example, $y_t \in \{-1, +1\}$ the label, and i_t is the task index for the round t. In this paper, we build on the algorithm proposed in (Cavallanti et al., 2008) (henceforth referred to as CMTL). In CMTL's proposed multitask Perceptron, *multitask instance* $\phi_t(x) \in R^{Kd}$ is:

$$\phi_t(x) = (\underbrace{0, \ldots, 0}_{d(i_t-1)times} \quad x_t \quad \underbrace{0, \ldots, 0}_{d(K-i_t)times})$$

and the different tasks are updated using rules which are derived from a pre-defined (fixed) task *interaction matrix*. This interaction matrix defines the different learning rates (η) to be used in the updates rules for different tasks. The weights of K Pereptrons are stored in a compound weight vector $w_s^T = (w_{1,s}^T, \ldots, w_{K,s}^T) \in \mathbb{R}^{Kd}$, where $w_{j,s} \in \mathbb{R}^d$ $\forall j \in \{1, \ldots, K\}$, and s denotes the number of mistakes made by the learner so far. The update rules are as follows:

$$\begin{aligned} w_s &= w_{s-1} + y_t (A \otimes I_d)^{-1} \phi_t \\ w_{j,s} &= w_{j,s-1} + y_t A^{-1}_{j,i_t} x_t \end{aligned} \tag{2.1}$$

where, $\otimes$ denotes the Kronecker product and

$$A^{-1} = \frac{1}{K+1} \begin{bmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 2 \end{bmatrix}$$

From the above $K \times K$ *interaction matrix* (A^{-1}), it follows that for $j = i_t, \eta = \frac{2}{K+1}$ whereas for tasks $j \neq i_t$, $\eta = \frac{1}{K+1}$. This updates scheme makes sense since it basically does a fixed, constant update for the current task i_t but at the same time also does "half-updates" for the remaining $K-1$ tasks, since they are expected to be related to the current task. The matrix A can be seen as enforcing co-regularization in the presence of multiple related learning tasks, an idea also pioneered in the literature on multitask learning for both batch (Evgeniou et al., 2005) as well as online setting (Agarwal et al., 2008).

3. Our Approach

Our first contribution is to learn this interaction matrix in an adaptive manner. At each round, our *adaptive* interaction matrix is derived by weighing each entry of the fixed interaction matrix A^{-1} by the corresponding entry of the following matrix:

$$U = \begin{bmatrix} 1 & \frac{1}{e^{||w_1 - w_2||^2}} & \dots & \frac{1}{e^{||w_1 - w_K||^2}} \\ \frac{1}{e^{||w_2 - w_1||^2}} & 1 & \dots & \frac{1}{e^{||w_2 - w_K||^2}} \\ \dots & \dots & \dots & \dots \\ \frac{1}{e^{||w_K - w_1||^2}} & \frac{1}{e^{||w_K - w_2||^2}} & \dots & 1 \end{bmatrix}$$

and our modified update rule becomes,

$$w_{j,s} = w_{j,s-1} + y_t(A^{-1}_{j,i_t} \times U_{j,i_t})x_t \qquad (3.1)$$

Since the interaction matrix is adaptive in our setting, it can also be seen as doing a kind of *adaptive co-regularization*, based on the current similarities of the tasks being learned.

3.1. An Active Learning Extension

Our framework can be easily extended to an active learning setting that takes into account the task relatedness. A naïve active learning strategy could be to use the margin based sampling for active learning using the randomized technique of (Cesa-Bianchi et al., 2006). More specifically, the approach proposed in (Cesa-Bianchi et al., 2006) uses a sampling probability term $p = b/(b + |r_{i_t}|)$ to decide whether to query the label of an incoming example i_t, where r_{i_t} is the signed margin of this example on the hypothesis being learned. The parameter b is set to a fixed value and dictates how aggressive the sampling is done. However, this approach does not exploit the inter-task relatedness in the presence of multiple tasks. We therefore propose to use matrix U of task similarity coefficients to set the sampling parameter b. One way of doing this would be to set $b = \sum_j e^{-||w_{i_t} - w_j||^2}$ which is nothing but the sum of the $i_t^t h$ row (or column) of the matrix U. It is easy to see that the expression for b would take a large value (meaning more aggressive sampling) if the tasks are highly similar, whereas b will have a small value (moderately aggressive sampling) if the tasks are not that highly related. Our experiments with this setting are shown in Figure 1(b).

4. Theoretical Results

In this section, we analyze the mistake bound of our approach and show how the task interaction matrix leads to a mistake bound that is better than the case of independently trained Perceptrons. The statement of the theorem below is similar to Theorem 1 of (Cavallanti et al., 2008) with the difference being the fact that the task interaction matrix now depends on the weights $w = [w_1^T w_2^T \dots w_K^T]$.

Theorem 4.1. *The number of mistakes m made by our algorithm, run with the interaction matrix B on any finite sequence $(\phi_1, y_1), (\phi_2, y_2), \dots \in \mathbb{R}^{Kd} \times \{-1, 1\}$, satisfies*

$$m \leq \inf_{w \in \mathbb{R}^{Kd}} \left(\sum_{t \in \mathcal{M}} l_t(w) + \frac{2(w^T B_\otimes w)}{K+1} + \sqrt{\frac{2(w^T B_\otimes w)}{K+1} \sum_{t \in \mathcal{M}} l_t(w)} \right),$$

where $\mathcal{M}$ is the set of mistaken trial indices, and $B_\otimes = B \otimes I$ is such that B^{-1} is an elementwise multiplication of the fixed matrix A^{-1} and the weighing matrix U. .

Proof. Follows along the same lines as in (Cavallanti et al., 2008) □

To show that the above mistake bound is provably better than independentaly trained Perceptron, we show the case of $K = 2$ and $d = 2$. This is just for the clarity of exposition. The analysis of cases of $K > 2$ are significantly more involved but the idea remains the same. For $K = 2$ and $d = 2$, the following result holds for $\frac{w^T B_\otimes w}{K+1}$:

$$\frac{w^T B_\otimes w}{3} = \frac{2}{4 - p^2} \left[||w_1||^2 + ||w_2||^2 - p w_1^T w_2 \right] \qquad (4.1)$$

where $p = e^{-||w_2 - w_1||^2}$.

Please see the appendix for the proof of the above result. It is easy to see that, when all tasks are equal (which for $K = 2$ means $w_1 = w_2$ and

$p = 1)$, $2/(4-p^2)\left[||w_1||^2 + ||w_2||^2 - pw_1^T w_2\right] < \left[||w_1||^2 + ||w_2||^2\right]$ which implies that our mistake bound for $K = 2$ is better than the sum of squares $(||w_1||^2 + ||w_2||^2)$ based mistake bound for 2 independent Perceptrons.

5. Experiments

We compare the following approaches: (a) multitask Perceptron (CMTL) (Cavallanti et al., 2008), (b) multitask Perceptron with active learning (CMTL-AL), (c) proposed adaptive multitask Perceptron (AMTL) with an adaptive *interaction matrix*, (d) our adaptive multitask Perceptron with active learning (AMTL-AL), (e) single task Perceptron for K independent tasks (STL), based on, (a) classification accuracy, (b) number of labels queried, and (c) total number of mistakes.

Our preliminary experiments have been conducted on 20-newsgroups dataset constructed akin to the way in (Daumé III, 2009) and (Raina et al., 2006)). We report experimental results on varying proportion of the training data (Fig. 1(a)), with each experiment averaged over 20 runs for random permutations of the training data order, and standard deviations also reported. Results with full training data are reported in Table 1.

As we see in Fig. 1(a), AMTL-AL achieves the best classification accuracy followed by AMTL, both of which are better classification accuracies reported by CMTL or CMTL-AL. The results on the number of mistakes made also depict a similar behavior.

For both CMTL and AMTL, the active learning based strategies have similar classification accuracies but smaller label complexity as compared to their passive counterparts. Nonetheless, in between the active strategies CMTL-AL and AMTL-AL, AMTL-AL makes a smaller number of mistakes.

Method	Acc (Std)	Labels (Std)	Mistakes (Std)
STL	56.65 (±3.70)	10142 (±0)	4817 (±58)
CMTL	73.42 (±4.42)	10142 (±0)	3233 (±29)
CMTL-AL	73.59 (±3.28)	9157 (±35)	3050 (±27)
AMTL	**75.47 (±2.31)**	10142 (±0)	3051 (±28)
AMTL-AL	**74.77 (±2.17)**	**9125 (±38)**	**2893 (±35)**
Acc: Accuracy | **Std:** Standard Deviation			

Table 1. Accuracy, label complexity and mistakes of *20-newsgroups with full training data*. Results are averaged over 20 runs with random data order permutations.

6. Discussion and Future Work

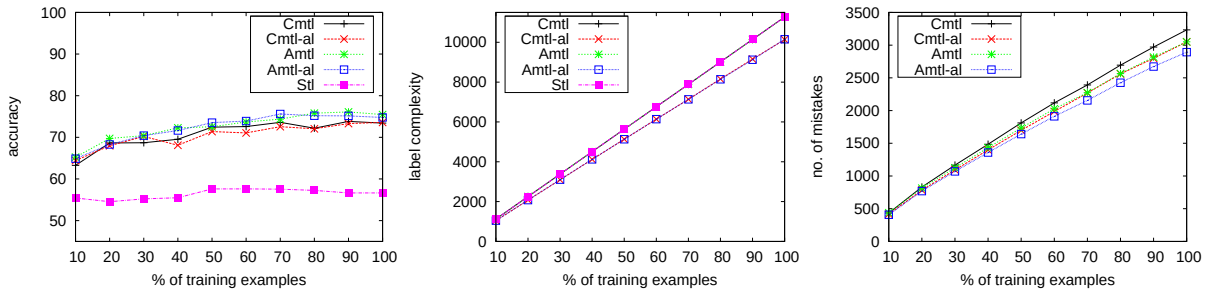
In this paper, we have proposed an adaptive strategy for online multitask learning. Our approach constructs an adaptive *interaction matrix* which quantifies the relatedness among the multiple tasks and uses this matrix to derive update rules for the various tasks. Subsequently, we augment the proposed adaptive online multitask learning with an active learning strategy which reduces the label complexity with marginal loss in accuracy.

Despite their simplicity and empirical success it is not theoretically apparent why the proposed algorithms perform well. For instance, we note that that the mistake bounds obtained in our case are weaker as compared to those of (Cavallanti et al., 2008). However, empirically we observe smaller number of mistakes. Theoretically analyzing the superior performance of the proposed adaptive and active online multitask learning strategies and providing mistake and regret bounds should be an interesting line of future work. Moreover, it would be interesting to investigate alternate forms of the *interaction matrix* that are more amenable to analyze and lead to tighter mistake bounds. One point to be noted here is that we have used the notion of similarity between tasks as the Euclidean distance between their weight vectors. This may not always be a correct notion of task relatedness: for example, two weight vectors pointing in exact opposite directions would have a large Euclidean distance whereas they clearly are very related. Perhaps, a more natural measure of relatedness would be the *complexity of transformation* between the two weight vectors. We are currently exploring ways of learning such transformations along with learning the weights, and using them as the notion of relatedness in the interaction matrix. The hope is that it would result in improved performance and easier analysis.

In addition, it can be seen that the current active learning strategy uses only margin information for selective sampling. However, the *interaction matrix* which relates the different tasks provides useful insights into which instances are more useful than the others. It would be interesting to design active learning strategies that extract useful information from the *interaction matrix* in order to query the labels of the most informative samples.

References

Agarwal, Alekh, Rakhlin, Alexander, and Bartlett, Peter. Matrix regularization techniques for online multitask learning. *Technical report, EECS Department,*



(a) (b) (c)

Figure 1. (a) Accuracy, (b) Label-complexity, (c) Number of mistakes for *20-newsgroups* with varying training data.

University of California, Berkeley, 2008.

Cavallanti, Giovanni, Cesa-Bianchi, Nicolò, and Gentile, Claudio. Linear algorithms for online multitask classification. In *COLT*, 2008.

Cesa-Bianchi, Nicolò, Gentile, Claudio, and Zaniboni, Luca. Worst-case analysis of selective sampling for linear classification. *J. Mach. Learn. Res.*, 2006.

Daumé III, Hal. Bayesian multitask learning with latent hierarchies. In *UAI*, Montreal, Canada, 2009.

Evgeniou, Theodoros, Micchelli, Charles A., and Pontil, Massimiliano. Learning multiple tasks with kernel methods. *JMLR*, 6:615–637, 2005.

Raina, Rajat, Ng, Andrew Y., and Koller, Daphne. Constructing informative priors using transfer learning. In *ICML*, pp. 713–720, 2006.

Settles, Burr. Active learning literature survey. In *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison, 2009.

Appendix

Proof. [Result 4.1] We derive the expression of $w^T B_\otimes w$ for the simple case of $K = 2$ and $d = 2$.

Our proposed matrices are,

$$B^{-1} = A^{-1} \odot U = \begin{bmatrix} 2 & \frac{1}{e^{||w_1 - w_2||^2}} \\ \frac{1}{e^{||w_2 - w_1||^2}} & 2 \end{bmatrix}$$

which when inverted would give

$$B = \frac{3}{4 - e^{-2||w_2 - w_1||^2}} \begin{bmatrix} 2 & -\frac{1}{e^{||w_1 - w_2||^2}} \\ -\frac{1}{e^{||w_2 - w_1||^2}} & 2 \end{bmatrix}$$

Susbtituting, $e^{-||w_2 - w_1||^2} = p$, we have:

$$B = \frac{3}{4 - p^2} \begin{bmatrix} 2 & -p \\ -p & 2 \end{bmatrix}$$

where, $w_1 = [w_{11} w_{12}]^T$, $w_2 = [w_{21} w_{22}]^T$ and $w^T = [w_1^T w_2^T]$ (for $K = d = 2$). Now, using:

$$B_\otimes = \begin{bmatrix} 2 & 0 & -p & 0 \\ 0 & 2 & 0 & -p \\ -p & 0 & 2 & 0 \\ 0 & p & 0 & 2 \end{bmatrix}$$

Now, we compute,

$$\begin{aligned} & w^T B_\otimes w \\ &= \frac{3}{4 - p^2} \begin{bmatrix} w_{11} & w_{12} & w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} 2w_{11} - pw_{21} \\ 2w_{12} - pw_{22} \\ -pw_{11} + 2w_{21} \\ -pw_{12} + 2w_{22} \end{bmatrix} \\ &= \frac{3}{4 - p^2} \Big[w_{11}(2w_{11} - pw_{21}) + w_{12}(-pw_{12} + 2w_{22}) \\ &+ w_{21}(2w_{11} - pw_{21}) + w_{22}(-pw_{12} + 2w_{22}) \Big] \\ &= \frac{3}{4 - p^2} \Big[||w_1||^2 + ||w_2||^2 \\ &+ w_{11}(w_{11} - pw_{21}) + w_{12}(-pw_{12} + w_{22}) \\ &+ w_{21}(w_{11} - pw_{21}) + w_{22}(-pw_{12} + w_{22}) \Big] \\ &= \frac{3}{4 - p^2} \Big[\sum_1^2 ||w_i||^2 + (w_{11}^2 - 2pw_{11}w_{21} + w_{21}^2) \\ &+ (w_{12}^2 - 2pw_{12}w_{22} + w_{22}^2) \Big] \\ &= \frac{3}{4 - p^2} \Big[2||w_1||^2 + 2||w_2||^2 - 2pw_1^T w_2 \Big] \end{aligned}$$

We have,

$$\begin{aligned} & w^T B_\otimes w = \frac{3}{4 - p^2} \Big[2||w_1||^2 + 2||w_2||^2 - 2pw_1^T w_2 \Big] \\ \implies & \frac{w^T B_\otimes w}{3} = \frac{2}{4 - p^2} \Big[||w_1||^2 + ||w_2||^2 - pw_1^T w_2 \Big] \end{aligned} \tag{6.1}$$

The result follows. □

For $K = 2$, the value of $w^T A_\otimes w/(K + 1)$ from (Cavallanti et al., 2008) can also be expressed as,

$$\frac{w^T A_\otimes w}{3} = \frac{2}{3} \Big[||w_1||^2 + ||w_2||^2 - w_1 w_2 \Big] \tag{6.2}$$

Comparing Eq. 6.1 with Eq. 6.2, we notice that both are essentially the same for $p = 1$.

Similar analysis hold for higher values of K and d but much more involved. We skip the details for simplicity.