Identification of Negative Transfers in Multitask Learning Using Surrogate Models

Dongyue Li, Huy L. Nguyen, and **Hongyang R. Zhang**

Assistant Professor of Computer Science, Northeastern University, Boston

Background and Problem Statement

Problem setup: Suppose we would like to make predictions on a primary target task of interest. We have:

- A fully-labeled training and validation set from the target task.
- k source tasks, each with a fully-labeled training set.

Motivating questions. What is the best use of data from the source tasks? More specifically, how would selecting the best subset out of the k source tasks help with predicting the target task?

- A naive baseline is to enumerate all possible combinations of the source tasks. But this requires evaluating 2^k combinations.
- Another baseline is to enumerate every source task, combined with the target task. This reduces the number of evaluations to k but ignores higher-order correlations among the tasks.

Would it be possible to simultaneously model higher-order task correlations while allowing efficient computation?

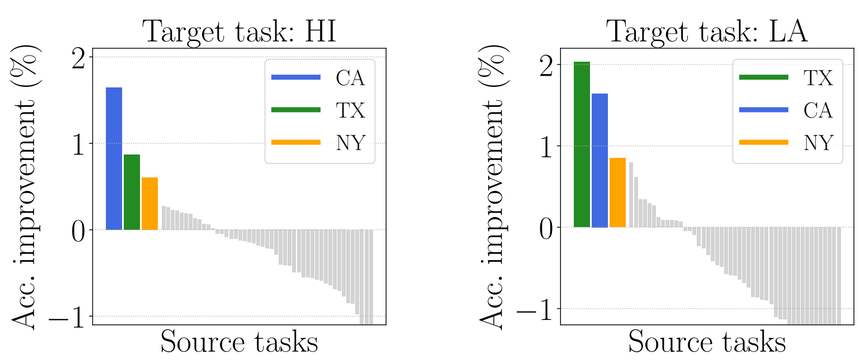
Technical Challenges

The problem of negative transfers: Due to the heterogeneity of the data across different tasks, negative transfers appear frequently when we combine several tasks to learn a joint model.

Illustration of positive transfers vs. negative transfers.

- **Dataset and setup:** We consider a binary classification task of predicting the income of individuals from US Census. We view the prediction task for each state as one task and use the rest of the fifty states as source tasks.
- **MTL vs. STL:** We randomly sample subsets of source tasks and perform MTL to combine the sampled source and target tasks.

Result: y-axis shows the difference between MTL (with one source task) and STL. The colored bars represent a few source tasks with notable positive transfers to the target task.



Predicting Transfer using Surrogate Models

Predicting positive vs. negative transfers: Given a subset of source tasks $S \subseteq \{1, \ldots, k\}$, is there a way to predict whether S provides a positive transfer to the target task in MTL?

Task modeling: We apply surrogate models to answer this question. Taking inspiration from recent work by Illyas et al. (2022), we apply linear surrogate models to approximate the multitask prediction loss of combining S with the target task in MTL. We estimate the surrogate model, which is now a function from subsets of source tasks to a loss value—$g(S) : \{0,1\}^k \to \mathbb{R}$, using random sampling.

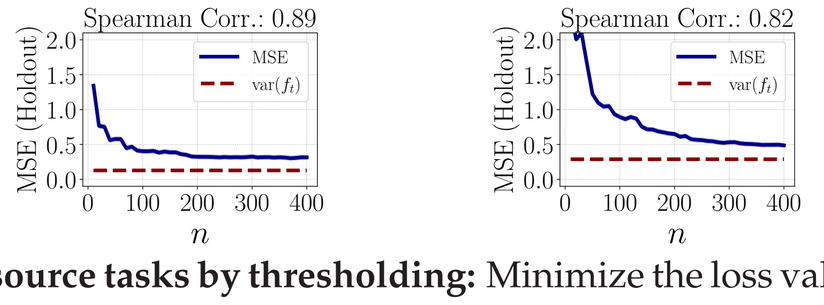
- **Constructing training data:** Sample n subsets of source tasks $\{S \subseteq \{1, \ldots, k\} : |S| = \alpha k\}$. Train a multitask model on S_i with target task $\mathcal{D}_t$ and evaluate on the validation loss for the target task:

$$\{(S_1, f_{\mathcal{A}}(t; S_1)), \ldots, (S_n, f_{\mathcal{A}}(t; S_n))\}.$$

- **Estimating surrogate models:** Let $\mathbb{1}_{S_i}$ denote a characteristic vector for each subset S_i. We estimate linear surrogate models: $g(S) = \theta^T \mathbb{1}_S$,

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \left(\theta^\top \mathbb{1}_{S_i} - f_{\mathcal{A}}(t; S_i)\right)^2.$$

Empirical results: Linear surrogate models accurately extrapolate, e.g.,



Selecting source tasks by thresholding: Minimize the loss value of $g(S)$, resulting in selecting source tasks with positive coefficients in $\hat{\theta}_n$:

$$S^\star = \arg\min_{S \subseteq \{1,2,\ldots,k\}} \mathbb{1}_S^\top \hat{\theta}_n = \left\{i : \hat{\theta}_n(i) < \gamma\right\}.$$

Analysis of Task Modeling

We provide two theoretical results to rigorously analyze our task modeling procedure.

- **Sample complexity:** We prove that by sampling $O(k\alpha^4 \log^2 k)$ subsets of size α, one can get an accurate estimate of the task model coefficients $\hat{\theta}_n$, over the uniform distribution of subsets of size α.
- **Task selection:** We prove that in a linear data generating model, there exists a threshold γ that is guaranteed to separate *related* source tasks from *unrelated* source tasks, compared with the target task.

In the first result, we only need the scoring function $f_{\mathcal{A}}$ to be bounded from above by a fixed constant C.

Theorem 1. Given $n = O(k\alpha^4 \log^2(k))$ samples from a distribution over $\{1, 2, \ldots, k\}$, w.h.p. the estimated task model $\hat{\theta}_n$ satisfies,

$$\left\|\hat{\theta}_n - \theta^\star\right\| \lesssim \frac{C\alpha^2 \log(k)\sqrt{k}}{\sqrt{n}} + \frac{C\alpha\sqrt{k}}{\sqrt{n}},$$

where $\theta^\star$ is the minimizer of the population loss.

In the second result, we analyze our thresholding procedure for selecting source tasks in a linear data-generating model. Each task i follows a linear model specified by a parameter vector $\theta^{(i)}$: $y = x^\top \theta^{(i)} + \epsilon$. Let a and b be two fixed values so that $b > a > 0$, a task is *related* if $\theta^{(i)} = \theta^{(t)} + z$, where z has mean zero and variance a^2; *unrelated* if $\theta^{(i)} = \theta^{(t)} + z$, where z has mean zero and variance b^2.

Theorem 2. Within the above setting, provided with sufficiently many samples of each task, there exists a threshold γ such that w.h.p.,

- $\hat{\theta}_n(i) < \gamma$ for any *related* task $i \in \{1, 2, \ldots, k\}$;
- $\hat{\theta}_n(j) > \gamma$ for any *unrelated* task $j \in \{1, 2 \ldots, k\}$.

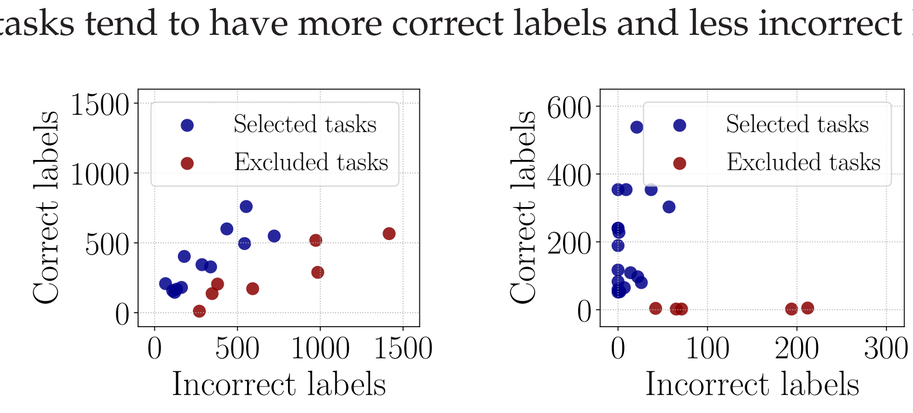
The analysis uses the fact that the subset distribution is uniform with fixed sizes. Under this distribution, the covariance structure in the task indices is approximately an identity matrix plus a constant term for every task. This covariance structure allows the task model coefficients to separate the related tasks from the unrelated tasks (i.e., amplifying the gap between a and b).

Experimental Results

We demonstrate three applications of task modeling, showing improved results for all.

- **Multitask weak supervision:** We consider text classification tasks with noisy supervision sources. Each source generates noisy labels for a subset of training samples and is viewed as a source task. Compared w/ WS baselines, achieve up to **3.6%** accuracy improvement.
- **MTL for NLP tasks:** Next, we consider 25 NLP tasks from GLUE, SuperGLUE, TweetEval, and ANLI benchmarks. Each task has 24 source tasks. Compared w/ MTL baselines, we outperform by **1.8%**.
- **Multi-group learning:** We consider binary classification tasks derived from the US Census, which uses ten tabular features to predict an individual's income. Each task is a dataset from one state and has 50 source tasks, with nine racial groups. Compared w/ GroupDRO, improve worst-group accuracy and fairness metrics by **1.29%**.

Illustration of separation in selected vs. not-selected source tasks: Selected tasks tend to have more correct labels and less incorrect labels.



Conclusion

- We propose a surrogate modeling method to identify negative transfers in multitask learning. Inspired by Datamodels (Illyas et al. (2022)), we find that linear task models work well for several applications.
- We provide a rigorous theoretical analysis of task modeling, including a sample complexity guarantee and a provable guarantee of thresholding in a linear data-generating model.
- We apply our methods to several applications and show significant improvement over prior approaches, including multitask weak supervision, MTL for NLP tasks, and robust multi-group learning.