Grouped Multi-Task Learning with Hidden Tasks Enhancement

Jiachun Jin[1], Jiankun Wang[1], Lu Sun[1], Jie Zheng[1] and Mineichi Kudo[2]

[1]School of Information Science and Technology, ShanghaiTech University
[2]Graduate School of Information Science and Technology, Hokkaido University

October 4, 2023

Problem Setting

- We consider the problem of Grouped Multi-Task Learning (Grouped-MTL).

Problem Setting

- We consider the problem of Grouped Multi-Task Learning (Grouped-MTL).
- Hope: By correctly transferring information across the tasks, the generalization performance of each task can be improved.

Problem Setting

- We consider the problem of Grouped Multi-Task Learning (Grouped-MTL).
- Hope: By correctly transferring information across the tasks, the generalization performance of each task can be improved.
- Given T tasks, with training dataset $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$, where $\mathbf{X}_t \in \mathbb{R}^{d \times N_t}$ and $\mathbf{y}_t \in \mathbb{R}^{N_t}$. Suppose the linear model $\mathbf{y}_t = \mathbf{X}_t^\top \boldsymbol{w}_t$ is adopted. The task parameter matrix $\mathbf{W}$ is given by:

$$\mathbf{W} = \begin{bmatrix} | & & | \\ \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_T \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times T}$$

Problem Setting

- We consider the problem of Grouped Multi-Task Learning (Grouped-MTL).
- Hope: By correctly transferring information across the tasks, the generalization performance of each task can be improved.
- Given T tasks, with training dataset $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$, where $\mathbf{X}_t \in \mathbb{R}^{d \times N_t}$ and $\mathbf{y}_t \in \mathbb{R}^{N_t}$. Suppose the linear model $\mathbf{y}_t = \mathbf{X}_t^\top \boldsymbol{w}_t$ is adopted. The task parameter matrix $\mathbf{W}$ is given by:

$$\mathbf{W} = \begin{bmatrix} | & & | \\ \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_T \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times T}$$

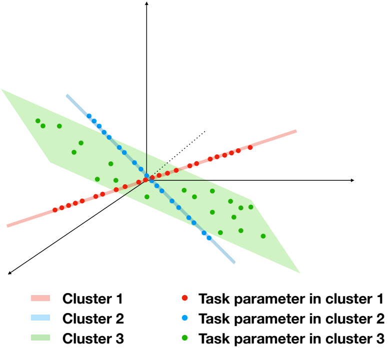
- Goal: Carry out task parameter learning and task clustering in a unified framework, and promote each other.

Subspace Structure of Task Parameters

- Task parameters in the same cluster lie in a shared low-rank subspace.

Subspace Structure of Task Parameters

- Task parameters in the same cluster lie in a shared low-rank subspace.
- Clusters ⇔ Subspaces, similar to the problem setting of **Subspace Clustering**, where we would like to cluster data points sample from a union of subspaces.



Subspace Clustering

- Self-expressiveness: A data point can be represented as a linear combination of the other vectors in the same subspace, i.e. $\boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is the representation of $\boldsymbol{x}_i$.

Subspace Clustering

- Self-expressiveness: A data point can be represented as a linear combination of the other vectors in the same subspace, i.e. $\boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is the representation of $\boldsymbol{x}_i$.
- Constraint is required to make the representation useful.

Subspace Clustering

- Self-expressiveness: A data point can be represented as a linear combination of the other vectors in the same subspace, i.e. $\boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is the representation of $\boldsymbol{x}_i$.
- Constraint is required to make the representation useful.
- Seeking a low-rank representation can be useful:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_*, \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XC}, \tag{1}$$

Subspace Clustering

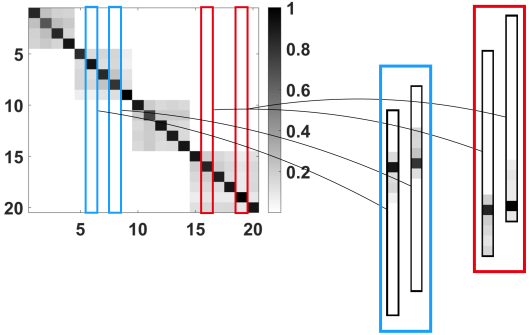
- Self-expressiveness: A data point can be represented as a linear combination of the other vectors in the same subspace, i.e. $\boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is the representation of $\boldsymbol{x}_i$.
- Constraint is required to make the representation useful.
- Seeking a low-rank representation can be useful:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_*, \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XC}, \tag{1}$$

- When the data $\mathbf{X}$ is noise-free, then the optimal solution to it is given by $\mathbf{C}^* = \mathbf{V}_0\mathbf{V}_0^\top$, here $\mathbf{X} = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^\top$ is the skinny SVD of $\mathbf{X}$ [Liu et al., 2012].

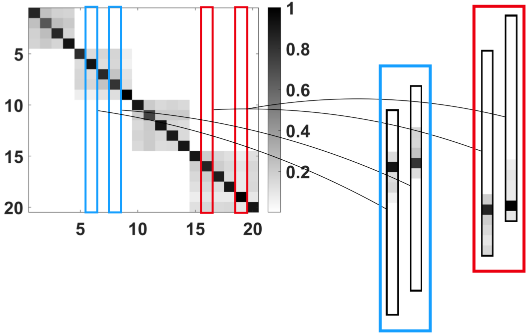
Subspace Clustering

- Produces representations that directly reveals the cluster structure: $\mathbf{C}^*$ must hold a block-diagonal structure, each block indicates a subspace cluster.



Subspace Clustering

- Produces representations that directly reveals the cluster structure: $\mathbf{C}^*$ must hold a block-diagonal structure, each block indicates a subspace cluster.



- This is representation learning.

Naive Task-Level Subspace Clustering

Replacing $\mathbf{X}$ with $\mathbf{W}$, and simultaneously fitting the data and enforcing task-level subspace structure, we reach the naive version of our objective function:

Naive Task-Level Subspace Clustering

Replacing $\mathbf{X}$ with $\mathbf{W}$, and simultaneously fitting the data and enforcing task-level subspace structure, we reach the naive version of our objective function:

$$\min_{\mathbf{W},\mathbf{C}} \mathcal{L}(\mathcal{D}, \mathbf{W}) + \frac{\lambda}{2}\|\mathbf{W} - \mathbf{WC}\|_F^2 + \gamma\|\mathbf{C}\|_* + \frac{\beta}{2}\|\mathbf{W}\|_F^2, \tag{2}$$

Naive Task-Level Subspace Clustering

Replacing $\mathbf{X}$ with $\mathbf{W}$, and simultaneously fitting the data and enforcing task-level subspace structure, we reach the naive version of our objective function:

$$\min_{\mathbf{W},\mathbf{C}} \mathcal{L}(\mathcal{D}, \mathbf{W}) + \frac{\lambda}{2}\|\mathbf{W} - \mathbf{W}\mathbf{C}\|_F^2 + \gamma\|\mathbf{C}\|_* + \frac{\beta}{2}\|\mathbf{W}\|_F^2, \tag{2}$$

Problem: Task parameters learned from data are not reliable, learning error may be amplified when used as a dictionary to represent themselves.

Hidden Tasks Enhanced Multi-Task Learning

- Extend the original dictionary by concatenating $\mathbf{W}$ with the hidden task parameters $\mathbf{H}$.

Hidden Tasks Enhanced Multi-Task Learning

- Extend the original dictionary by concatenating $\mathbf{W}$ with the hidden task parameters $\mathbf{H}$.
- Suppose both $\mathbf{W}$ and $\mathbf{H}$ are known and fixed, then the task-level subspace clustering problem can be formulated as:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_*, \quad \text{s.t.} \quad \mathbf{W} = [\mathbf{W}, \mathbf{H}]\mathbf{C}. \tag{3}$$

Hidden Tasks Enhanced Multi-Task Learning

- Extend the original dictionary by concatenating $\mathbf{W}$ with the hidden task parameters $\mathbf{H}$.
- Suppose both $\mathbf{W}$ and $\mathbf{H}$ are known and fixed, then the task-level subspace clustering problem can be formulated as:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_*, \quad \text{s.t.} \quad \mathbf{W} = [\mathbf{W}, \mathbf{H}]\mathbf{C}. \tag{3}$$

Theorem

When both $\mathbf{W}$ and $\mathbf{H}$ are known, the optimal solution is $\mathbf{C}^* = \mathbf{V}\mathbf{V}_W^\top = [\mathbf{V}_W; \mathbf{V}_H]\mathbf{V}_W^\top$, where $[\mathbf{W}, \mathbf{H}] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma}[\mathbf{V}_W; \mathbf{V}_H]^\top$ is the SVD of the concatenated matrix.

Hidden Tasks Enhanced Multi-Task Learning

Let's re-plug $\mathbf{C}^* = [\mathbf{V}_W; \mathbf{V}_H]\mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U\Sigma V}_H^\top$ into the original constraint:

Hidden Tasks Enhanced Multi-Task Learning

Let's re-plug $\mathbf{C}^* = [\mathbf{V}_W; \mathbf{V}_H]\mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}_H^\top$ into the original constraint:

$$\begin{aligned}
\mathbf{W} &= [\mathbf{W}, \mathbf{H}][\mathbf{V}_W\mathbf{V}_W^\top; \mathbf{V}_H\mathbf{V}_W^\top] \\
&= \mathbf{W}\overbrace{\mathbf{V}_W\mathbf{V}_W^\top}^{\mathbf{Z}} + \overbrace{\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}_H^\top\mathbf{V}_H\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top}^{\mathbf{L}}\mathbf{W} \\
&= \mathbf{W}\mathbf{Z} + \mathbf{L}\mathbf{W},
\end{aligned} \tag{4}$$

Hidden Tasks Enhanced Multi-Task Learning

Let's re-plug $\mathbf{C}^* = [\mathbf{V}_W; \mathbf{V}_H]\mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U\Sigma V}_H^\top$ into the original constraint:

$$\begin{aligned} \mathbf{W} &= [\mathbf{W}, \mathbf{H}][\mathbf{V}_W\mathbf{V}_W^\top; \mathbf{V}_H\mathbf{V}_W^\top] \\ &= \mathbf{W}\overbrace{\mathbf{V}_W\mathbf{V}_W^\top}^{\mathbf{Z}} + \overbrace{\mathbf{U\Sigma V}_H^\top\mathbf{V}_H\mathbf{\Sigma}^{-1}\mathbf{U}^\top}^{\mathbf{L}}\mathbf{W} \\ &= \mathbf{WZ} + \mathbf{LW}, \end{aligned} \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{T\times T}$ is the task correlation matrix, and $\mathbf{L} \in \mathbb{R}^{d\times d}$ is the feature correlation matrix.

Hidden Tasks Enhanced Multi-Task Learning

Let's re-plug $\mathbf{C}^* = [\mathbf{V}_W; \mathbf{V}_H]\mathbf{V}_W^\top$ and $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}_H^\top$ into the original constraint:

$$\begin{aligned}\mathbf{W} &= [\mathbf{W}, \mathbf{H}][\mathbf{V}_W\mathbf{V}_W^\top; \mathbf{V}_H\mathbf{V}_W^\top] \\ &= \mathbf{W}\overbrace{\mathbf{V}_W\mathbf{V}_W^\top}^{\mathbf{Z}} + \overbrace{\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}_H^\top\mathbf{V}_H\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top}^{\mathbf{L}}\mathbf{W} \\ &= \mathbf{W}\mathbf{Z} + \mathbf{L}\mathbf{W},\end{aligned} \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{T\times T}$ is the task correlation matrix, and $\mathbf{L} \in \mathbb{R}^{d\times d}$ is the feature correlation matrix.

The key

In reality, $\mathbf{H}$ is unreachable, so instead of exactly recovering $\mathbf{Z}$ and $\mathbf{L}$ from data, we take them as learnable parameters to enforce subspace structure with the effect of hidden tasks.

Hidden Tasks Enhanced Multi-Task Learning

To jointly carry out data fitting and hidden tasks enhanced subspace clustering, we reach our objective:

$$\min_{\mathbf{W},\mathbf{Z},\mathbf{L}} \mathcal{L}(\mathcal{D}, \mathbf{W}) + \frac{\lambda}{2}\|\mathbf{W} - \mathbf{W}\mathbf{Z} - \mathbf{L}\mathbf{W}\|_F^2 + \gamma(\|\mathbf{Z}\|_* + \|\mathbf{L}\|_*) + \frac{\beta}{2}\|\mathbf{W}\|_F^2. \tag{5}$$

Hidden Tasks Enhanced Self-Expressive Layer

Furthermore, we can extend our model from single layer to m layers, as the following:

$$\min_{\substack{\mathbf{W}_1,\\ \{\mathbf{Z}_k\},\{\mathbf{L}_k\}}} \mathcal{L}(\mathcal{D}, \mathbf{W}_m) + \sum_{k=1}^{m} \left(\frac{\lambda_k}{2} \|\mathbf{W}_k - \mathbf{W}_k \mathbf{Z}_k - \mathbf{L}_k \mathbf{W}_k\|_F^2 + \gamma_k (\|\mathbf{Z}_k\|_* + \|\mathbf{L}_k\|_*) \right) + \frac{\beta}{2} \|\mathbf{W}_1\|_F^2, \tag{6}$$

Hidden Tasks Enhanced Self-Expressive Layer

Furthermore, we can extend our model from single layer to m layers, as the following:

$$\min_{\substack{\mathbf{W}_1,\\ \{\mathbf{Z}_k\},\{\mathbf{L}_k\}}} \mathcal{L}(\mathcal{D}, \mathbf{W}_m) + \sum_{k=1}^{m} \left(\frac{\lambda_k}{2} \|\mathbf{W}_k - \mathbf{W}_k \mathbf{Z}_k - \mathbf{L}_k \mathbf{W}_k\|_F^2 + \gamma_k (\|\mathbf{Z}_k\|_* + \|\mathbf{L}_k\|_*) \right) + \frac{\beta}{2} \|\mathbf{W}_1\|_F^2, \tag{6}$$

the rationale is we can reformulate $\mathbf{W} = \mathbf{WZ} + \mathbf{LW}$ to:

$$\mathbf{w} = (\mathbf{Z} \otimes \mathbf{I}_d + \mathbf{I}_T \otimes \mathbf{L})\, \mathbf{w} = \mathbf{M}\mathbf{w}, \tag{7}$$

and we can extract deep hierarchical information, where $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ is the vectorization of $\mathbf{W}$:

$$\mathbf{w}_k = \mathbf{M}_{k-1}\mathbf{w}_{k-1} = \prod_{\ell=1}^{k-1} \mathbf{M}_\ell \mathbf{w}_1, \quad \mathbf{M}_\ell = \mathbf{Z}_\ell \otimes \mathbf{I}_d + \mathbf{I}_T \otimes \mathbf{L}_\ell. \tag{8}$$

Empirical Results

- We first generate a dataset that strictly follow our subspace assumption.

Empirical Results

- We first generate a dataset that strictly follow our subspace assumption.
- There are 4 task clusters contain 4, 5, 5 and 6 binary classification tasks, respectively.

Empirical Results

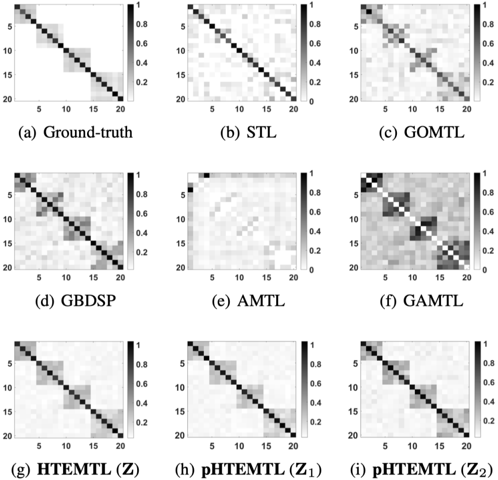
- We first generate a dataset that strictly follow our subspace assumption.
- There are 4 task clusters contain 4, 5, 5 and 6 binary classification tasks, respectively.
- Tasks within the same cluster share the same set of bases.

Empirical Results

- We first generate a dataset that strictly follow our subspace assumption.
- There are 4 task clusters contain 4, 5, 5 and 6 binary classification tasks, respectively.
- Tasks within the same cluster share the same set of bases.
- We generate the bases of each cluster by applying QR decomposition to a full-rank matrix.

Empirical Results

- We first generate a dataset that strictly follow our subspace assumption.
- There are 4 task clusters contain 4, 5, 5 and 6 binary classification tasks, respectively.
- Tasks within the same cluster share the same set of bases.
- We generate the bases of each cluster by applying QR decomposition to a full-rank matrix.



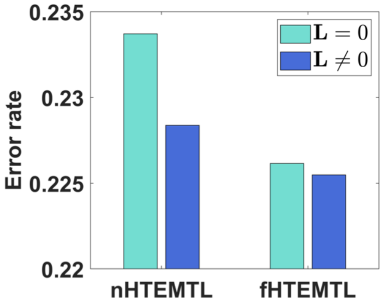
Empirical Results

Table 1: Experimental results (mean ± std) with different evaluation metrics. The best two results are highlighted in boldface.

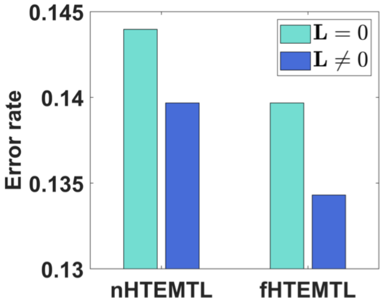
Dataset	Measure	STL	GOMTL	AMTL	GAMTL	GBDSP	KMSV	HTEMTL	pHTEMTL
Synthetic	ER↓	0.2392±0.0137	0.2271 ±0.0122	0.2377 ±0.0168	0.2233 ±0.0203	0.2179±0.0119	0.2340±0.0115	**0.2076±0.0128**	**0.2042±0.0044**
	AUC↑	0.8535±0.0415	0.8617±0.0464	0.8651±0.0371	0.8795±0.0348	0.8795±0.0313	0.8235±0.0281	**0.8906±0.0362**	**0.8868±0.0099**
Fashion-MNIST	ER↓	0.1097±0.0042	0.1017±0.0045	0.0751±0.0035	0.0717±0.0061	0.0847±0.0066	0.1012±0.0084	**0.0699±0.0130**	**0.0710±0.0051**
	AUC↑	0.9812±0.0107	**0.9937±0.0032**	0.9824±0.0086	0.9893±0.0078	0.9888±0.0052	0.9785±0.0072	0.9865±0.0163	**0.9905±0.0019**
CIFAR-10	ER↓	0.2880±0.0029	0.2393±0.0034	0.2387±0.0032	0.2361±0.0036	0.2359±0.0032	0.2525±0.0039	**0.2284±0.0041**	**0.2234±0.0011**
	AUC↑	0.8183±0.0089	0.8836±0.0094	0.8511±0.0139	0.8730±0.0099	0.8809±0.0118	0.8347±0.0168	**0.8878±0.0083**	**0.8880±0.0019**
AWA2-Attribute	ER↓	0.1784±0.0019	0.1753±0.0055	0.1493±0.0030	0.1550±0.0036	0.1789±0.0047	0.1794±0.0054	**0.1397±0.0031**	**0.1316±0.0009**
	AUC↑	0.7300±0.0421	0.7276±0.0595	0.7270±0.0595	0.7286±0.0725	0.7608±0.0622	**0.7917±0.0749**	0.7436±0.0584	**0.7619±0.0194**
School	rMSE↓	10.3127±0.0602	10.1606±0.0712	10.1604±0.0712	10.2398±0.0557	**10.1218±0.1035**	**10.1320±0.0711**	10.1806±0.0878	10.1769±0.0038
	MAE↓	8.1472±0.0379	8.1502±0.1764	**8.0321±0.0463**	8.0949±0.0305	**7.9732±0.0739**	8.0427±0.0668	8.0443±0.0480	8.0370±0.0168

Empirical Results

We further study the effect of hidden tasks by deactivating the effect of hidden tasks. This is equivalent to set the matrix $\mathbf{L}$ to $\mathbf{0}$.



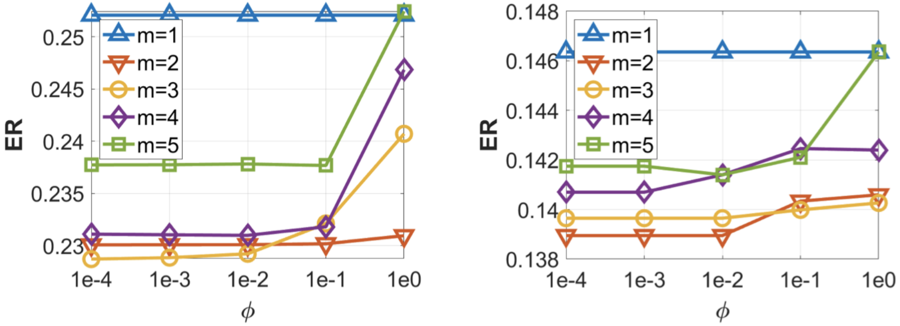
(a) CIFAR-10



(b) AWA2-Attribute

Empirical Results

We also study the effect of cascading HTE layers.



(a) CIFAR-10 (b) AWA2-Attribute

Thanks

Please refer to our paper for more details.