Learning Task Relational Structure for Multi-Task Feature Learning

De Wang, Feiping Nie, Heng Huang
Dept. of Computer Science and Engineering
University of Texas at Arlington
Arlington, Texas, 76019, U.S.A.
Email: {wangdelp, feipingnie}@gmail.com, heng@uta.edu

***Abstract*—In multi-task learning, it is paramount to discover the relational structure of tasks and utilize the learned task structure. Previous works have been using the low-rank latent feature subspace to capture the task relations, and some of them aim to learn the group based relational structure of tasks. However, in many cases, the low-rank subspace may not exist for the specific group of tasks, thus using this paradigm would not work. To discover the task relational structures, we propose a novel multi-task learning method using the structured sparsity-inducing norms to automatically uncover the relations of tasks. Instead of imposing the low-rank constraint, our new model uses a more meaningful assumption, in which the tasks from the same relational group should share the common feature subspace. We can discover the group relational structure of tasks and learn the shared feature subspace for each task group, which help to improve the predictive performance. Our proposed algorithm avoids the high computational complexity of integer programming, thus it converges very fast. Empirical studies conducted on both synthetic and real-world data show that our method consistently outperforms related multi-task learning methods.**

***Keywords*-Multi-Task Learning; Task Relational Structure; Structured Sparsity-Inducing Norm**

I. Introduction

Multi-task learning (MTL) is an emerging machine learning paradigm that learns multiple tasks (in the image categorization example, each category of image is a task) jointly. In recent years, it has been observed by many researchers [1], [2], [3], [4] that if tasks are correlated, one task would benefit from others by jointly learning multiple tasks. Thus, multi-task learning has been widely researched and applied to many related areas such as computer vision [5], medical image analysis [6], bioinformatics [7], and natural language processing [8].

In order to exploit multi-task learning for vision tasks, one paramount issue is to discover the task relational structure and utilize the uncovered task relations to enhance the multi-task learning models. Only by successfully capturing the relational structure of tasks, we can benefit from learning these tasks jointly and increase the predictive performances. Otherwise, if dissimilar or non-related tasks are learned jointly, it may lead to the *negative effect on the knowledge transfer*, resulting that learning these tasks jointly is worse than learning them independently. Thus, discovering the correct task relational structure is the pillar for multi-task learning.

In current literature, the common way to utilize the task relations in the multi-task learning models is to assume that the related tasks share a common yet latent low-rank feature subspace [1], [8]. The trace norm (nuclear norm) regularization is often used to discover the low-rank subspace. For example, in [1], Argyriou *et. al.* proposed to minimize a loss function with the trace norm based regularization of the model parameter matrix for all tasks.

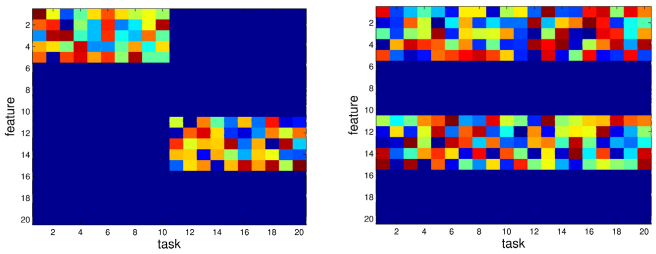
However, these methods only assume the task correlations and impose them in the learning model without explicitly discovering the task relational structure. In the other recent work [5], Kang *et. al.* presented a new method to discover the task group relations by combining the low-rank subspace and task group learning, and used the learned task correlations to enhance the predictive performance.

In general, the related tasks are expected to share a common subset of useful features. If related tasks are grouped together, tasks from the same group should share a common subspace consisting of the same subset of features. Based on this intuition, in this paper, we propose a new multi-task learning model, which can automatically discover the task group relational structure, and learn the feature subspace shared by each task group. Our method jointly optimizes the model parameter matrix of multi-task learning and the group indicator matrix of tasks, which groups tasks based on their relations. As a result, the uncovered task group relational structure and common feature subspaces can further enhance the multi-task learning model.

II. New Multi-Task Learning Model

Multi-task learning can utilize the interrelations between tasks by learning multiple tasks jointly. In [1], a multi-task feature learning model was introduced to find a common low-rank feature subspace for all tasks. Because the trace norm is often used to find a low-rank subspace, their objective function minimizes the loss function with the trace norm regularization on the model parameter matrix of entire tasks, which is as following:

$$\min_{W} \sum_{t=1}^{T} l(\mathcal{D}_t; w_t) + \gamma \|W\|_*^2 . \tag{1}$$



(a) (b)

Figure 1. The structures of model parameter matrix W discovered using different regularization terms: (a) applying task group relational structure regularization (imposing $\ell_{2,0}$-norm on each individual task group, which is discovered automatically) on the model parameter matrix; (b) applying the flat task regularization (imposing $\ell_{2,0}$-norm on all the tasks).

where $l(\cdot)$ is a loss function, T is the number of tasks,$\mathcal{D}_t$ consists of the data matrix (n_t data points) and labels (y_t) for the t-th task, $W \in \Re^{d\times T}$ is the model parameter matrix, w_t is the t-th column of W, and γ is the tradeoff parameter, $\|W\|_*$ denotes the trace norm of W.

This model assumes all tasks share the same common subspace without investigating the task relational structure. However, not all tasks are correlated to each other. If the dissimilar or non-correlated tasks are learned jointly, it may lead to negative effect on the knowledge transfer and reduce the predictive performance. Thus, it is crucial to discover the task relational structure and learn the correlated tasks together.

In this paper, we propose to learn the task group relational structure for multi-task learning. If the task group structure is learned, this information can be used to better regularize the feature space: by regularizing parameters in each group rather than regularizing the whole matrix across groups, we can obtain better knowledge transfer between similar tasks and avoid negative transfer between dissimilar tasks. However, learning task group structure is very difficult since partitioning of tasks into groups involves combinatorial optimization problem.

Motivated by the above observation, in this paper, we propose a new regularization formulation for multi-task learning. The underlying idea is: *If the tasks are grouped according to the task interrelations, the tasks from the same group are supposed to share a common feature subspace, whereas tasks in different groups are more likely to have different feature subspaces.* We implement this mechanism into a newly proposed regularization formulation based on structured sparsity-inducing norm, which is widely used in to achieve sparsity for feature selection [9], dimensionality reduction [10], and regression/classification [11] tasks. An efficient algorithm is proposed to learn the task group structure. The algorithm avoids the computational cost of solving combinatorial optimization problem.

When we propose the new objective and theoretical analysis, we will use the following example for illustration purpose, in which we have 20 tasks and 20 features: without loss of generality, we assume that the group structure of the tasks is as in Figure 1(a). In this figure, the first 10 tasks belong to one group, and the remaining tasks belong to another group. The feature subspace of Group 1 (Task 1 to Task 10) consists of features from 1 to 5, and the subspace of Group 2 (Task 11 to Task 20) consists of features from 11 to 15.

A. New Group Structured Multi-Task Learning Objective Function

The $\ell_{2,1}$-norm of a matrix M is defined as: $\|M\|_{2,1} = \sum_i \left\|M^i\right\|_2 = \sum_i \sqrt{\sum_j m_{ij}^2}$, where M^i indicates the i-th row of the matrix M. $\ell_{2,1}$-norm is a sparsity-inducing norm that enforces some rows of the matrix to be zero. $\ell_{2,0}$-norm a matrix M is defined as the non-zeros rows in M.

It is well known that the $\ell_{2,1}$-norm is a convex envelop of the $\ell_{2,0}$-norm and can lead to structured sparsity in the model parameter matrix, which has been successfully used for multi-task feature learning in previous research [3]. We propose to regularize the model parameter matrix by the $\ell_{2,1}$-norm to discover the task group relational structure. However, we cannot straightforwardly apply the $\ell_{2,1}$-norm or $\ell_{2,0}$-norm regularization on the model parameter matrix W. Because different task groups share different feature subspaces, the regularization should be applied within each task group. If the regularization is applied on the entire model parameter matrix, all tasks would share the same common feature subspace. Thus, the learned structure of model parameter matrix will be similar to Figure 1(b). However, this is not the case for those dissimilar and non-correlated tasks. We expect the task group regularization term can discover the real task relational structure such that the model parameter matrix has the proper structural patterns as shown in Figure 1(a). Therefore, the regularization on the entire model parameter matrix may lead to negative effect on the knowledge transfer and thus deteriorate classification performance.

If we regularize the model parameter matrix within each group using the structured sparsity-inducing norms, it naturally comes out to solve the following objective function:

$$\min_W \sum_{t=1}^{T} l(\mathcal{D}_t; w_t) + \gamma \sum_{g=1}^{G} \|W_g\|_{2,0}\,, \tag{2}$$

The loss function in our proposed model can be any type of loss functions. However, optimizing Eq. (2) would lead to trivial solution, *i.e.*, the solutions found by using $\ell_{2,0}$-norm regularization on grouped model parameter matrix ($\sum_{g=1}^{G} \|W_g\|_{2,0}$) and on one single non-grouped model parameter matrix ($\|W\|_{2,0}$) could be identical. As a result, the correct task group relational structure cannot be properly

discovered. This situation is similar to the trace norm regularization case we discussed in previous section (using trace norm to the parameter matrix of entire tasks may uncover the similar low-rank subspaces as using the trace norm to the parameter matrices of each individual task group).

We are going to explain the reason with using the illustration example in Figure 1. When the $\ell_{2,0}$-norm is used as regularization for (1) the parameter matrix of entire tasks and (2) the parameter matrices of each individual task group, the penalties will generate the same results in these two cases. The value of $\ell_{2,0}$-norm is the number of retained features. In the first case, when the $\ell_{2,0}$-norm regularization is imposed to the parameter matrix of the entire tasks, the learning results are shown in Figure 1(b). The penalty term value is 10 (the number of retained features). In the second case, when the $\ell_{2,0}$-norm regularization is imposed to the parameter matrices of each individual task group, we expect to discover the task group relational structure as shown in Figure 1(a). In this case, the value of the penalty term is the sum of number of features retained in every group, *i.e.*, $5 + 5 = 10$. However, the penalty term values in these two cases are the same, *i.e.* 10, such that the model has no motivation to discover the correct task group structure as shown in Figure 1(a). Thus, in the second case, the resulted task group relational structure can combine all tasks into one group and leave no tasks in the other group, which is identical to task relational structure as shown in Figure 1(b).

To tackle this issue, we propose to use the square of $\ell_{2,0}$-norm as regularization, and optimize the following objective:

$$\min_{W} \sum_{t=1}^{T} l(\mathcal{D}_t; w_t) + \gamma \sum_{g=1}^{G} \|W_g\|_{2,0}^2 \,. \tag{3}$$

In our new formulation, the penalty term values are different in the above two cases: for case in Figure 1(a), the value of the penalty term is $5^2 + 5^2 = 50$; for case in Figure 1(b), the value becomes $(5 + 5)^2 = 100$. Because we aim to minimize Eq. (3), the model will choose case (a), *i.e.* the correct group relational structure of tasks can be discovered. What's more, if the ground-truth structure of W is Figure 1(a), the minimum penalty is achieved by grouping the tasks as two groups. If $G = 3$, one of the two groups need to be split into two new groups. As a result, the penalty term value becomes $5^2 + 5^2 + 5^2 = 75$, which is greater than the penalty value of two groups. From the above analysis, we can see that using the square of $\ell_{2,0}$-norm as regularization the new model in (3) is optimized to discover the correct task group relational structure. Because solving the $\ell_{2,0}$-norm minimization requires integer programming optimization with high computational cost and the $\ell_{2,1}$-norm is a good convex approximation of the $\ell_{2,0}$-norm, we will utilize the squared $\ell_{2,1}$-norm in our new objective to discover the task group relational structure.

Let $Q \in \Re^{G \times T}$ be the binary group indicator matrix whose elements $q_{gt} \in \{0, 1\}$, and $q_{gt} = 1$ if the t-th task belongs to the g-th task group, otherwise $q_{gt} = 0$. Let $Q_g \in \Re^{T \times T}$ be a diagonal matrix whose elements are q_{gt}. Obviously, $Q_g(t, t) = 1$ if task t belongs to the g-th group, otherwise $Q_g(t, t) = 0$, and $\sum_g Q_g = \boldsymbol{I}$. Then W_g can be written as: $W_g = WQ_g$, which is actually a subset of columns of W that belong to the g-th group. Thus, our new group structural multi-task learning model solves:

$$J_{opt} = \min_{\substack{W, Q_g \\ Q_g(t,t) \in \{0,1\}}} \sum_{t=1}^{T} l(\mathcal{D}_t; w_t) + \gamma \sum_{g=1}^{G} \|WQ_g\|_{2,1}^2 \tag{4}$$

In our new objective, we simultaneously optimize the coefficient matrix W and the task group relational structure matrix Q_g. Therefore, the task group relational structure can be automatically learned. Meanwhile, the squared $\ell_{2,1}$-norm also imposes the structured sparsity to the tasks from the same group, such that they share the common feature subspace and jointly learn together. Please notice that although the $\ell_{2,1}$-norm was used in previous multi-task learning, the previous models cannot discover the task group relational structure. Our regularization term is new with combining both $\ell_{2,1}$-norm and group indicator matrix. Based on our above analysis, our new regularization can discover the task relational structure better than existing work using trace norm based regularization.

III. Optimization Algorithm

In this paper, we propose an iterative re-weighted strategy to optimize the proposed model (4), which converges very fast. Actually, the algorithm converges in less than 10 iterations on all the data sets in our experiments.

The loss function in our proposed model can be any type of loss functions. Since our focus is on the regularization term, the loss function is set as least square loss for simplicity.

Using the iterative re-weighted strategy, solving problem (4) is equivalent to solve the following problem:

$$\begin{aligned} &\min_{W, Q_g} \left\| Y - X^T W \right\|_F^2 + \gamma \sum_{g=1}^{G} Tr(Q_g W^T D_g W Q_g) \\ = \quad &\min_{W, Q_g} \left\| Y - X^T W \right\|_F^2 + \gamma \sum_{g=1}^{G} Tr(Q_g W^T D_g W) \,, \end{aligned} \tag{5}$$

where D_g is a diagonal matrix whose diagonal elements are: $D_g(t, t) = \frac{\|WQ_g\|_{2,1}}{\|(WQ_g)^t\|_2}$, where M^t represents the t_{th} row of M.

We adopt a two step procedure to alternatively optimize the objective function over W and Q.

When Q is fixed, taking derivative *w.r.t* W and set it as 0, problem (5) becomes:

$$XX^T W - XY + \gamma \sum_{g=1}^{G} D_g W Q_g = 0 \,. \tag{6}$$

Table I
RMSE OF OUR MODEL WHEN USING DIFFERENT NUMBER OF GROUPS G. THE SMALLER IS BETTER.

G	1	2	3	4	T
RMSE	0.54	0.32	**0.25**	0.53	0.29

It is easy to see that Eq. (6) can be decoupled between tasks. Thus Eq. (6) can be written as:

$$XX^T w_t - Xy_t + \gamma \sum_{g=1}^{G} Q_g(t,t) D_g W_t = 0\,. \tag{7}$$

Thus, we get the update equation for w_t as following:

$$w_t = (XX^T + \gamma \sum_{g=1}^{G} Q_g(t,t) D_g)^{-1} X y_t\,. \tag{8}$$

When W is fixed, problem (5) is reduced to:

$$\min_{Q_g} \sum_{g=1}^{G} Tr(Q_g W^T D_g W)\,. \tag{9}$$

It seems difficult to solve problem (9). However, after careful analysis, we find that problem (9) can be decoupled by tasks. Denotes $Q \in \Re^{T\times G}$, where $Q(:,g) = diag(Q_g)$ ($diag(M)$ is defined as taking out the diagonal elements of M as a vector), and $A \in \mathbb{R}^{T\times G}$, where $A(:,g) = diag(W^T D_g W)$. Then problem (9) is reduced to:

$$\min_{Q} Tr(QA^T) \Rightarrow \min_{Q} \sum_{t=1}^{T} Q(t,:) A(t,:)^T\,. \tag{10}$$

Because each task exactly belongs to one group, only one elements in $Q(t,:)$ can be 1 and all others should be 0. So the solution for problem (10) should be as following:

$$\begin{aligned} Q(t,g) &= 1 \quad if \quad g = g^* \\ Q(t,g) &= 0 \quad if \quad g \neq g^*\,, \end{aligned} \tag{11}$$

where $g^* = \arg\min_g A(t,g)$.

It deserves to be mentioned that: *although the model (4) contains integer variables Q_g which casts it as an integer programming problem and makes it very difficult to solve, our algorithm avoids the computational complexity of integer programming by decoupling the problem with tasks when solving Q. Hence, the updating process is very fast.* However, the algorithm proposed in [5] discretized the integer programming problem, and then used a nested loop in their algorithm to alternatively optimize model parameter W and group indicator matrix Q. Thus, their algorithm converges slowly. We summarize our algorithm in Alg. 1.

The algorithm will converge into a local minimum since each iteration is decreasing the objective function value. For space reason, we omit the proof of convergence.

Algorithm 1 Algorithm to solve the problem (4).

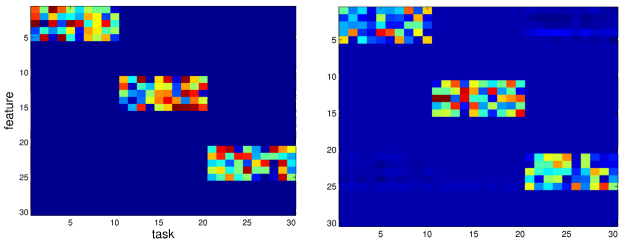
Initialize $D_g = I$.
Random initialize Q.
repeat
Update W by Eq. (8):

$$w_t = (XX^T + \gamma \sum_{g=1}^{G} Q_g(t,t) D_g)^{-1} X y_t$$

Update Q by Eq. (11):

$$Q(t,g) = 1 \text{ if } g = g^*;\ Q(t,g) = 0 \text{ if } g \neq g^*\,,$$

Compute D_g by $D_g(t,t) = \frac{\|WQ_g\|_{2,1}}{\|(WQ_g)^t\|_2}$.
until Converges



(a) Ground-truth structure (b) Task structure recovered by our model

Figure 2. The task relational structure in parameter matrix. Blue background means the value is zero.

IV. EXPERIMENTAL RESULTS

In this section, we will present our empirical results to evaluate the proposed multi-task group learning approach on vision tasks. The comparison methods used in this paper are as following:

(1) Non-Grouped MTL (NG): the resulted objective when we set $G = 1$ in our model (4), where all tasks indiscriminately assigned into one group and learnt jointly. This is equivalent to $\ell_{2,1}$-norm regularized classification.

(2) Single Task (Single): the resulted approach when we set $G = T$ in our model (4), where all tasks are learnt independently.

(3) Grouped Low-Rank MTL (GLR): the method described in the paper [5].

(4) Non-Grouped Low-Rank MTL (NGLR): the method described in the paper [1]. This method assumes all tasks share a common low-rank subspace.

Experiments are conducted on both synthetic data and six real world vision data. We follow the multi-task learning experimental set up in previous papers [12], [5]. When we use one-vs-rest mechanism to solve multi-class classification problems, we can jointly learn the c (c is the number of classes) one-vs-rest learning tasks together as a multi-task learning process. First, we use two handwritten digit recognition data sets: MNIST and USPS, which were used to evaluate the Grouped Low-Rank MTL method in the paper [5]. Moreover, four face recognition data sets are used to further validate the effectiveness of our method, which are

ORL, YaleB, PIE, and UMIST.

Five-fold cross validation is conducted to evaluate the performance of those methods. The group number G is chosen by performing cross validation. RMSE (root mean squared error, used for regression task) and error rate (used for classification task) are computed to evaluate the performance of these methods.

A. Synthetic Data

The synthetic data are important to help us check whether a method works as what we expected. The synthetic data used in our experiments is generated as following: the data set consists of 900 samples, 30 features, and 30 tasks. We first generate the model parameter matrix W with sparsity structure as in Figure 2(a). This means that the first 10 tasks, middle 10 tasks and the last 10 tasks come form 3 task groups, respectively. Tasks in each group share a common subspace consisting of 5 features (the non-zero parts in Figure 2(a)). Then the data matrix X is randomly generated. After that, we compute the target variable $Y = X^T W$.

Figure 2(b) shows the task relational structure in model parameter matrix recovered by our algorithm. We can see that the pattern (the retained features in different task groups) in Figure 2(b) is very similar to the ground truth structure of W in Figure 2(a). This means that our model can correctly uncover the task relational structure and find the proper common feature subspaces for different task groups. Using these recovered feature subspaces, the predictive performance of multi-task learning can be enhanced.

Figure 3 shows the group structure of tasks by visualizing the group indicator matrix Q. We see that when the group number G is set to be 3 (the ground truth group number), our model recovers the true group structure for all tasks except for one task.

Table I shows the RMSE for different numbers of groups. The RMSE is normalized by the standard deviation of target variable Y. The best RMSE is achieved by setting $G = 3$ (the ground truth number of task groups), and the result is significantly better than others.

B. Real World Data Sets

In this section, we evaluate the performance of our model on handwritten digit recognition data sets and face recognition data sets. Following previous multi-task learning experimental setup [12], [5], we treat multi-way classification as a multi-task learning problem, where each task is a classification task of one digit against all the others.

Two handwritten digit data sets are used: MNIST, USPS. Both of them contain 10 digits from 0 to 9, and one digit is one class. Multi-way classification problem can be treated as multi-task learning problems using the one-vs-rest paradigm, such that each class is a task.

Four image data sets are used for face recognition tasks: ORL, YaleB, PIE, and Umist. These data sets are all comprising of human faces, represented using gray scale pixel values. For the fairness of comparison, the regularization parameters γ are tuned with the same granularity in the same range for different methods, both from $\{10^{-4}, 10^{-3}, ..., 10^{3}, 10^{4}\}$.

Table II provides the error rates of different methods on the two digit data sets and four face data sets. On both handwritten digit data sets, our method (denoted by MTGL) outperforms other comparison methods. This is because our multi-task group learning approach is able to discover the group relational structure of tasks, and utilize different common feature subspaces for different task groups. MTGL also outperforms GLR, this shows that: with respect to discovering the correct group structure in the multi-task learning setting, identifying a subset of features as the subspace is better than finding a low-rank subspace. The performance of NG (non-grouped) method is worse than MTGL, and NGLR is also worse than GLR. The reason is that severe incorrect knowledge transfer occurs while indiscriminately requiring all tasks belong to one group. What's more, our method (MTGL) outperforms GLR and NGLR with statistical significance on three out of the four data sets: YaleB, PIE, and UMIST. On the MNIST data set, our method's improvement over other methods is not significant. Actually, on both handwritten digit data sets, the performance of Single ($G = T$, which is equivalent to *Lasso*) and MTGL is quite close. Sounds like on these two data sets, the traditional *Lasso* regularized classification performs pretty good. So the improvement of using multi-task learning is not obvious.

C. Running Time Comparison

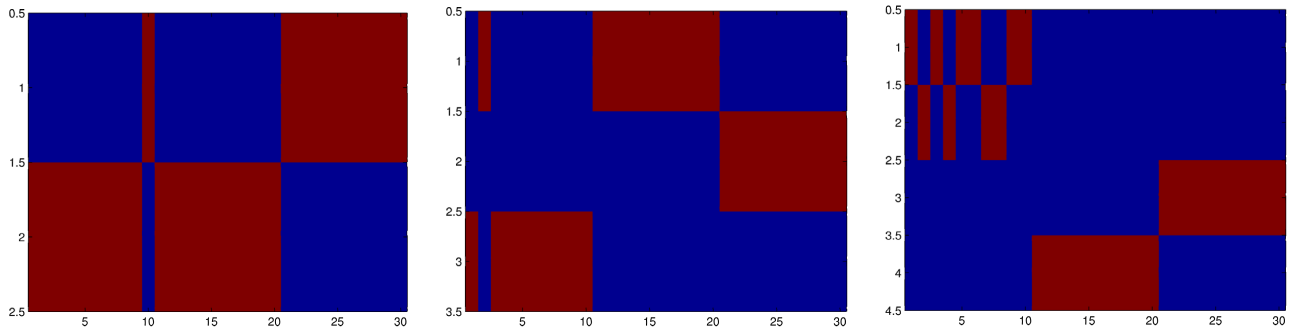
Table III shows the running time comparison for our MTGL and the GLR method. It is computed by running the algorithm 20 times on a double core, 3.40Ghz, 64 bit operating system with 16GB memory. The average time and standard deviation for running the algorithm one time is reported. Table III shows that our algorithm converges much faster than the GLR method.

V. Conclusion

In this paper, we propose a new group structured multi-task learning model with assuming the correlated tasks from the same group should share the common feature subspaces. Our method can discover the group relational structure of tasks and learn the shared common feature subspaces for each task group. We derive an efficient algorithm using the iterative re-weighted optimization strategy. The algorithm avoids the integer programming which has high computational cost, thus our new algorithm converges very fast.

ACKNOWLEDGEMENT

This work was partially supported by NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, NIH AG049371.



(a) G=2 (b) G=3 (c) G=4

Figure 3. Task group structure Q when group number G is varying. Horizontal axis is task, vertical axis is group. Red area means which group the task is assigned to.

Table II
MEAN ERROR RATES AND STANDARD DEVIATIONS OF DIFFERENT METHODS ON REAL WORLD DATA SETS.

	MNIST	USPS	ORL	YaleB	PIE	UMIST
MTGL	**15.5±0.5**	**8.6±0.4**	**3.7±0.2**	**0.3±0.07**	**3.9±0.3**	**13.0±1.2**
NG	16.1±0.4	9.4±0.6	4.3±0.3	0.4±0.06	**3.9±0.4**	13.5±1.2
GLR	16.2±0.7	10.9±0.6	3.9±0.2	3.9±0.1	7.1±0.6	21.6±1.7
NGLR	16.4±0.4	11.26±0.3	3.9±0.2	4.2±0.1	7.7±0.7	22.3±1.8
Single	16.1±0.5	9.2±0.4	5.0±0.4	0.6±0.07	4.5±0.4	15.5±1.4

Table III
AVERAGE TIME (SECONDS) AND STANDARD DEVIATION OF RUNNING THE ALGORITHM ONCE ON DIFFERENT DATA SETS.

	MTGL	GLR
MNIST	3.5±0.2	16.8±1.1
USPS	3.8±0.3	16.7±1.0
ORL	7.9±0.5	54.5±4.1
YaleB	9.9±0.5	53.9±4.3
PIE	18.1±0.7	123.2±7.7
UMIST	3.9±0.3	32.6±3.2

REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[2] S. Kim and E. Xing, "Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity," *ICML*, 2010.

[3] G. Obozinski, B. Taskar, and M. Jordan, "Multi-task feature selection," *Technical report, Department of Statistics, University of California, Berkeley*, 2006.

[4] J. Liu, S. Ji, and J. Ye., "Multi-task feature learning via efficient $l_{2,1}$-norm minimization," in *UAI2009*, 2009.

[5] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 521–528.

[6] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen, "High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1286–1294, 2012.

[7] S. Lee, J. Zhu, and E. P. Xing, "Adaptive multi-task lasso: with application to eqtl detection," in *NIPS*, 2010, pp. 1306–1314.

[8] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *The Journal of Machine Learning Research*, vol. 6, pp. 1817–1853, 2005.

[9] D. Wang, F. Nie, H. Huang, J. Yan, S. L. Risacher, A. J. Saykin, and L. Shen, "Structural brain network constrained neuroimaging marker identification for predicting cognitive functions," in *Information Processing in Medical Imaging*. Springer Berlin Heidelberg, 2013, pp. 536–547.

[10] D. Wang, F. Nie, and H. Huang, "Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track)," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 306–321.

[11] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1813–1821, 2010.

[12] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," *ICML*, 2007.