

Multi-Task Learning via Structured Regularization: Formulations, Algorithms, and Applications

Jianhui Chen^{1,2}

¹Computer Science and Engineering, Arizona State University, Tempe, AZ 85287

²Center for Evolutionary Medicine and Informatics, the Biodesign Institute, Arizona State University, Tempe, AZ 85287

1. Introduction

Multi-task learning (MTL) [2] aims to enhance generalization performance by learning multiple tasks simultaneously. It has been applied successfully in many areas including bioinformatics, computer vision, and text mining. A common assumption in MTL is that all tasks are intrinsically related to each other. Under such an assumption, the informative domain knowledge is allowed to be shared across the tasks, implying what is learned from one task is beneficial to another. This is particularly desirable when there are a number of related tasks but only a limited amount of training data is available for learning each task.

2. Learning Multiple Tasks via Structured Regularization

Our research work on MTL is centered around using a shared low-rank structure to capture the intrinsic task relationship [1]. Our MTL formulations can be expressed in a generic form as

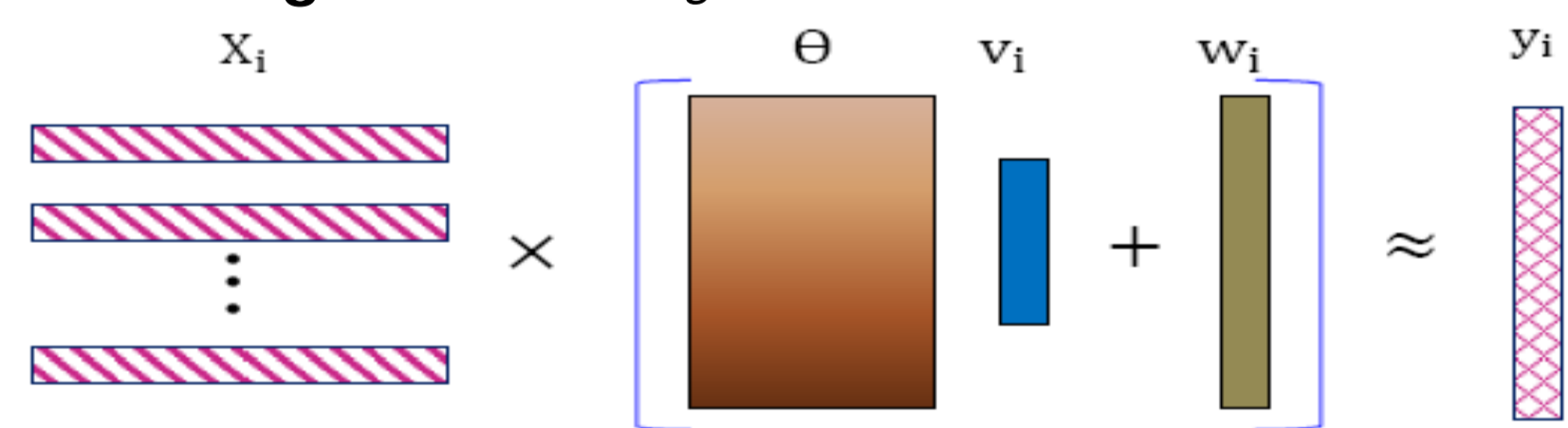
$$\begin{aligned} & \text{minimize} \quad \text{Empirical Loss} + \text{Structured Regularization} \\ & \text{subject to} \quad \text{Structured Constraint,} \end{aligned}$$

where Structured Regularization and Structured Constraint are employed to model the task relationship as well as induce the desirable structures among multiple tasks, and Empirical Loss measures how well the proposed MTL formulation can model the mapping from the samples to the associated targets.

2.1 Learning a Shared Low-Rank Structure

We learn a shared low-rank structure among all tasks and a task-specific feature mapping (for each task) simultaneously. The model structure is illustrated in Figure 2.1.

Figure 1: Learning a shared low-rank structure.



Mathematically this MTL approach (iASO) can be formulated as

$$\begin{aligned} & \text{minimize}_{\Theta, \{v_i, w_i\}} \sum_{i=1}^m \left\{ L(X_i(\Theta v_i + w_i), y_i) + \alpha \|\Theta v_i + w_i\|^2 + \beta \|w_i\|^2 \right\} \\ & \text{subject to} \quad \Theta^T \Theta = I_{h \times h}, \Theta \in \mathbb{R}^{d \times h}, d > h, \end{aligned} \quad (1)$$

where Θ , v_i , and w_i represent the shared low-dimensional structure parameter, the low-dimensional weight vector, and the high-dimensional weight vector, respectively, and the desirable low-rank structure is induced via an orthonormal constraint. The formulation in Eq. (1) is non-convex and its globally optimal solution can not be easily computed. Our analysis shows that iASO subsumes ASO [1], a classical multi-task learning algorithm, as a special case; it can also be naturally converted into a convex relaxation (cASO) as

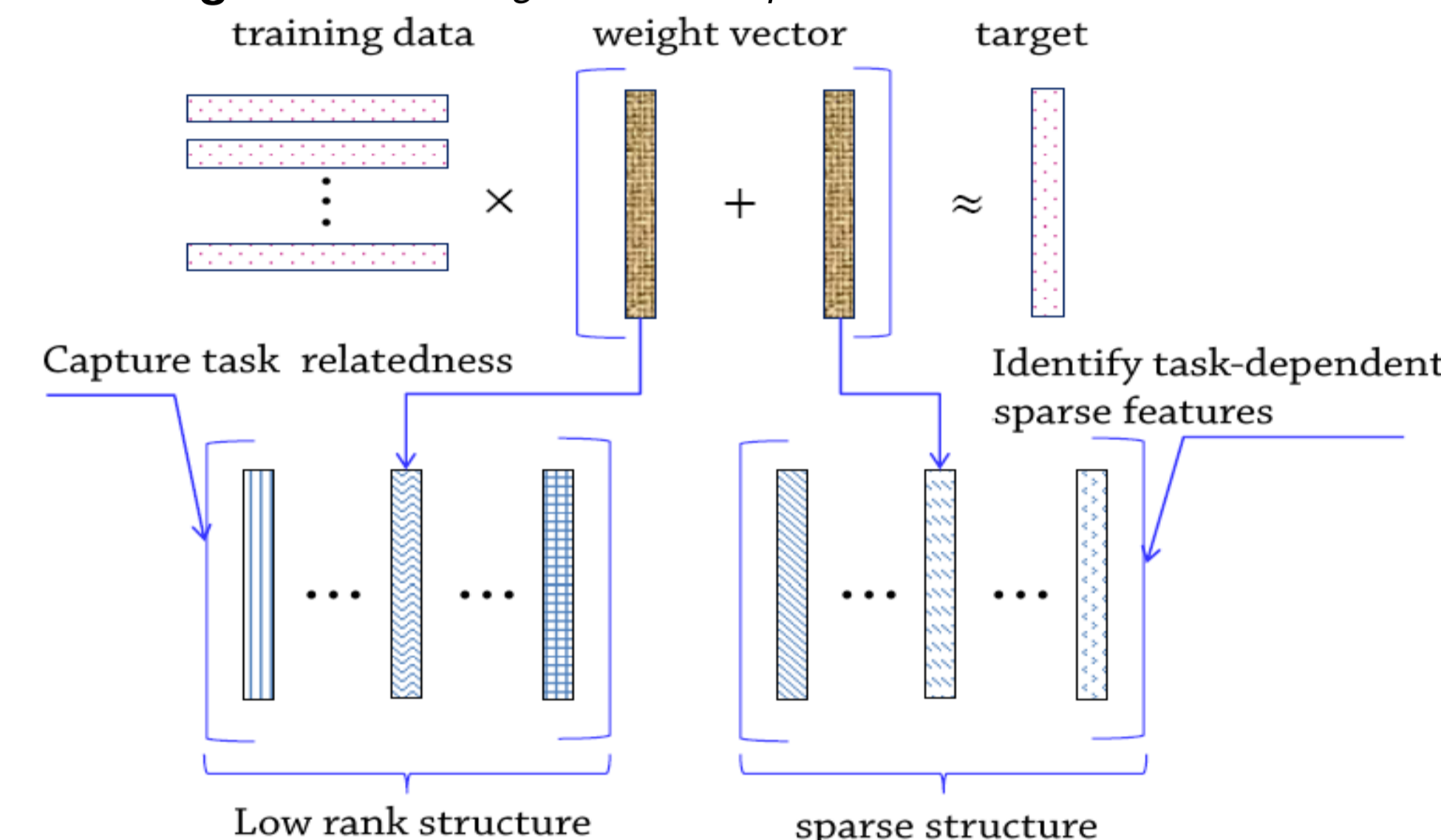
$$\begin{aligned} & \text{minimize}_{M, \{u_i\}} \sum_{i=1}^m \left\{ L(X_i u_i, y_i) + \hat{\alpha} u_i^T (\eta I + M)^{-1} u_i \right\} \\ & \text{subject to} \quad \text{tr}(M) = h, M \preceq I, M \in \mathbb{S}_+^d. \end{aligned} \quad (2)$$

The formulation in Eq. (2) is convex and its globally optimal solution can be efficiently obtained. Our theoretical analysis shows cASO finds a globally optimal solution for iASO under certain theoretical condition.

2.2 Learning Incoherent Sparse and Low-Rank Structures

We propose to identify the discriminative features for each task (via a sparse structure) and meanwhile capture the task relationship (via a shared low-rank structure). The proposed model structure is illustrated in Figure 2.

Figure 2: Learning incoherent sparse and low-rank structures.



Mathematically, this MTL approach (iSL) can be formulated as

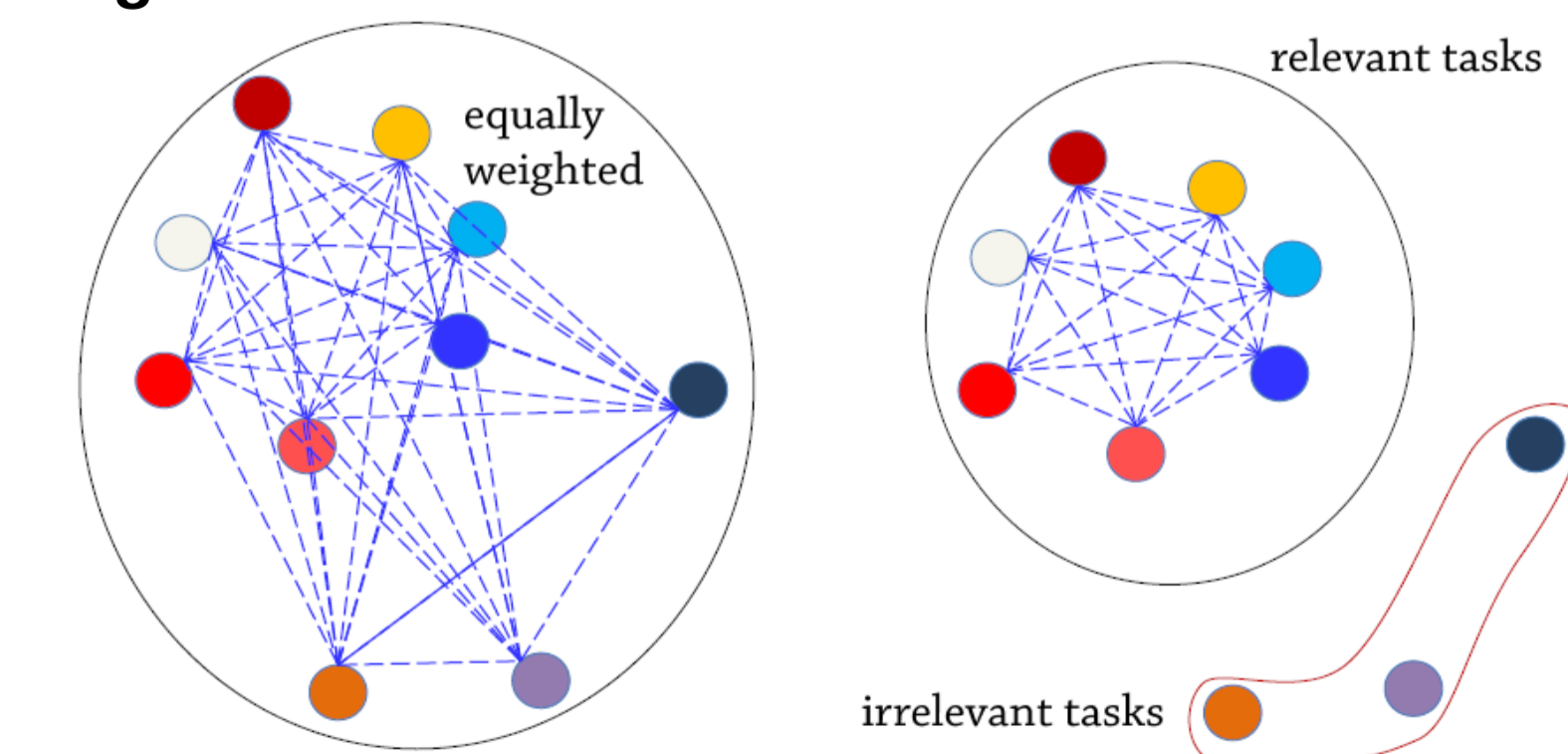
$$\begin{aligned} & \text{minimize}_{P, Q} \sum_{i=1}^m L(X_i(p_i + q_i), y_i) + \gamma \|P\|_1 \\ & \text{subject to} \quad \|Q\|_* \leq \tau, \end{aligned} \quad (3)$$

where the sparse structure is induced via an ℓ_1 -norm regularization, and the low-rank structure is induced via a trace norm constraint.

2.3 Learning Low-Rank and Group Sparse Structures

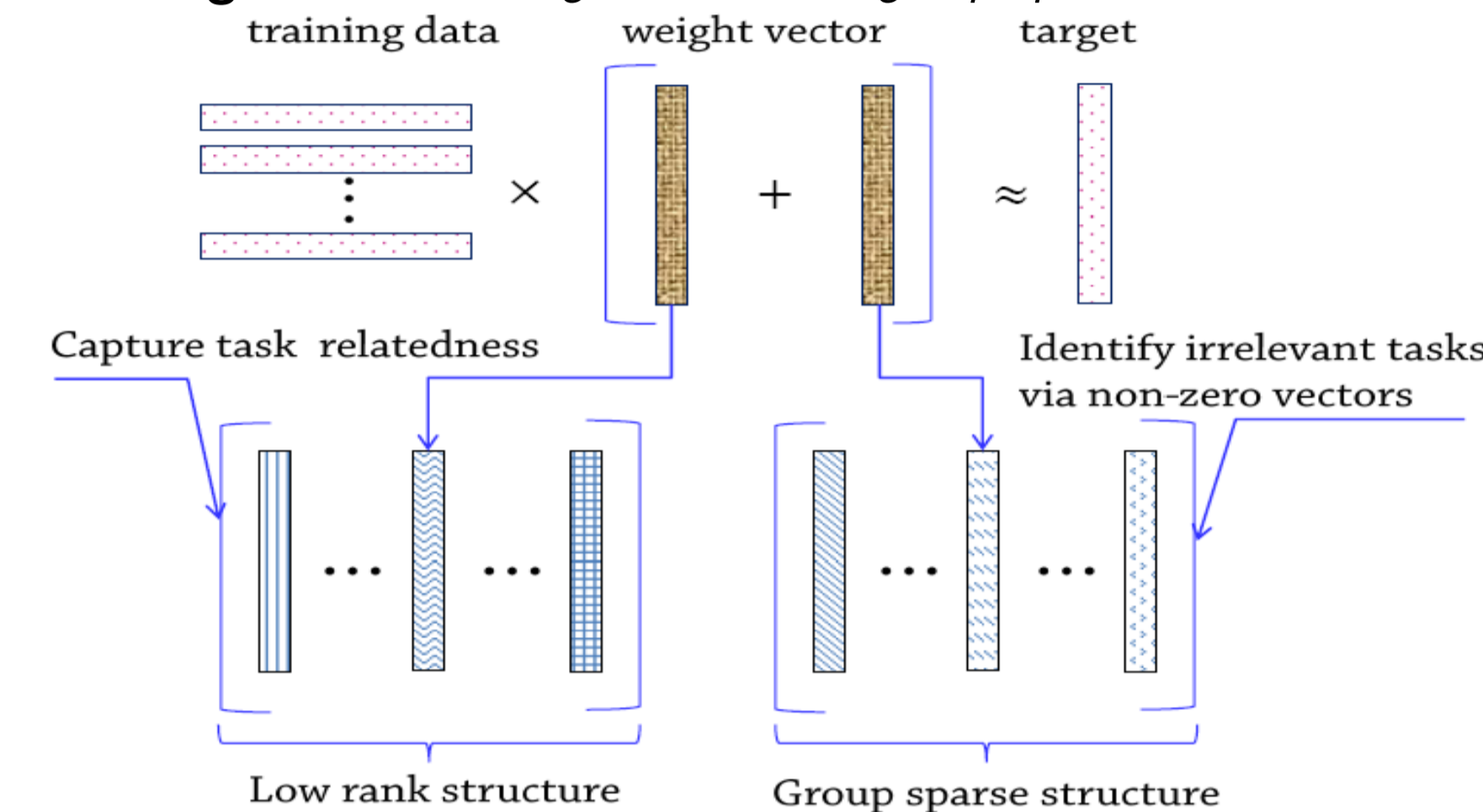
We consider the multi-task learning scenario where a group of learning tasks are related while other tasks are irrelevant to such a group, as illustrated in Figure 3.

Figure 3: Illustration of the relevant tasks and irrelevant tasks.



We propose a robust MTL approach in which the irrelevant tasks are identified via a group-sparse structure and the task relationship is captured via a low-rank structure. The proposed model structure is illustrated in Figure 4.

Figure 4: Learning low-rank and group-sparse structures.



Mathematically, this MTL approach (RMTL) is formulated as

$$\min_{L, S} \sum_{i=1}^m L(X_i(l_i + s_i), y_i) + \alpha \|L\|_* + \beta \|S\|_{1,2}, \quad (4)$$

where the low-rank structure is induced via a trace norm regularization, and the group-sparse structure is induced via an $\ell_{1,2}$ -norm regularization.

3. Optimization Algorithms

We employ two optimization algorithms for solving the proposed MTL formulations as briefly described below.

3.1 Alternating Optimization (AO) Algorithm

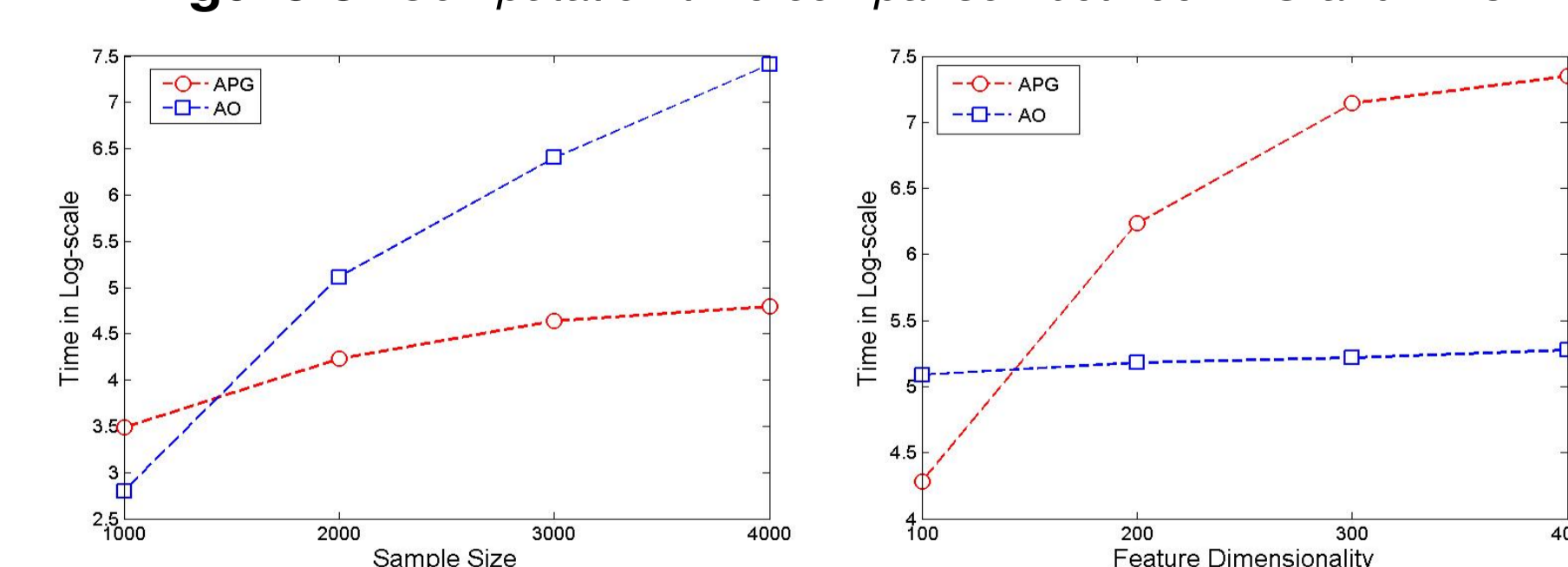
The AO algorithm is similar to the block coordinate descent method, in which the variables are optimized alternatively with the other variables fixed. AO is easy to implement and has fast practical convergence.

3.2 Accelerated Projected Gradient (APG) Algorithm

The APG algorithm is optimal among all first order techniques [3]. One key component in APG is the computation of the proximal mapping, which is involved in each iteration of APG. We develop efficient algorithms for solving the proximal mapping associated with each MTL formulation.

We compare AO and APG in terms of the required computation (in seconds) for solving the iSL formulation. The experimental results are summarized in Figure 5.

Figure 5: Computation time comparison between AO and APG.



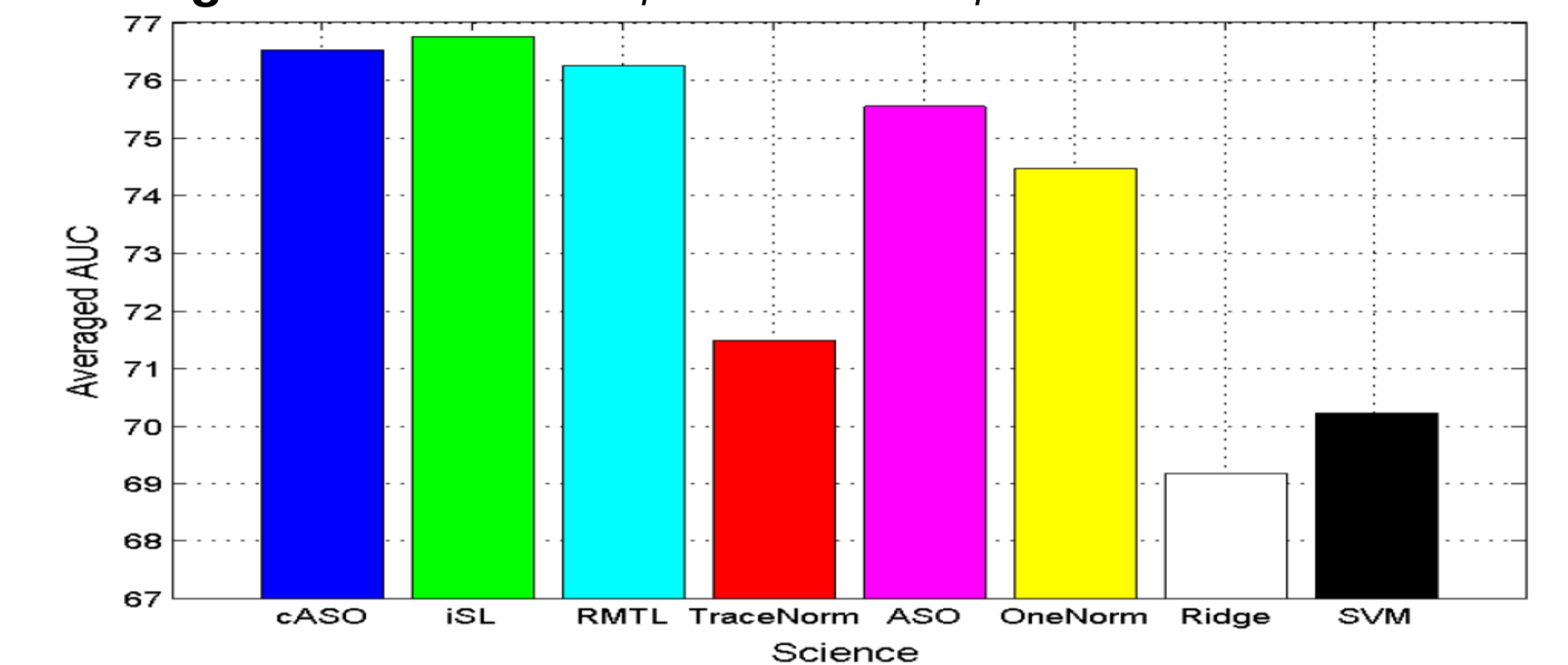
4. Applications

We present the empirical performance of the proposed MTL approaches on two real-world applications, as summarized below.

Application 1 - Topics categorization of the Yahoo web pages

The Yahoo data sets consist of 11 top-level categories, where each top-level category corresponds to one data set. Each top-level category is further divided into a set of second-level subcategories, where each second-level subcategory corresponds to a topic included in one data set (one top-level category). The performance comparison results are presented in Figure 6.

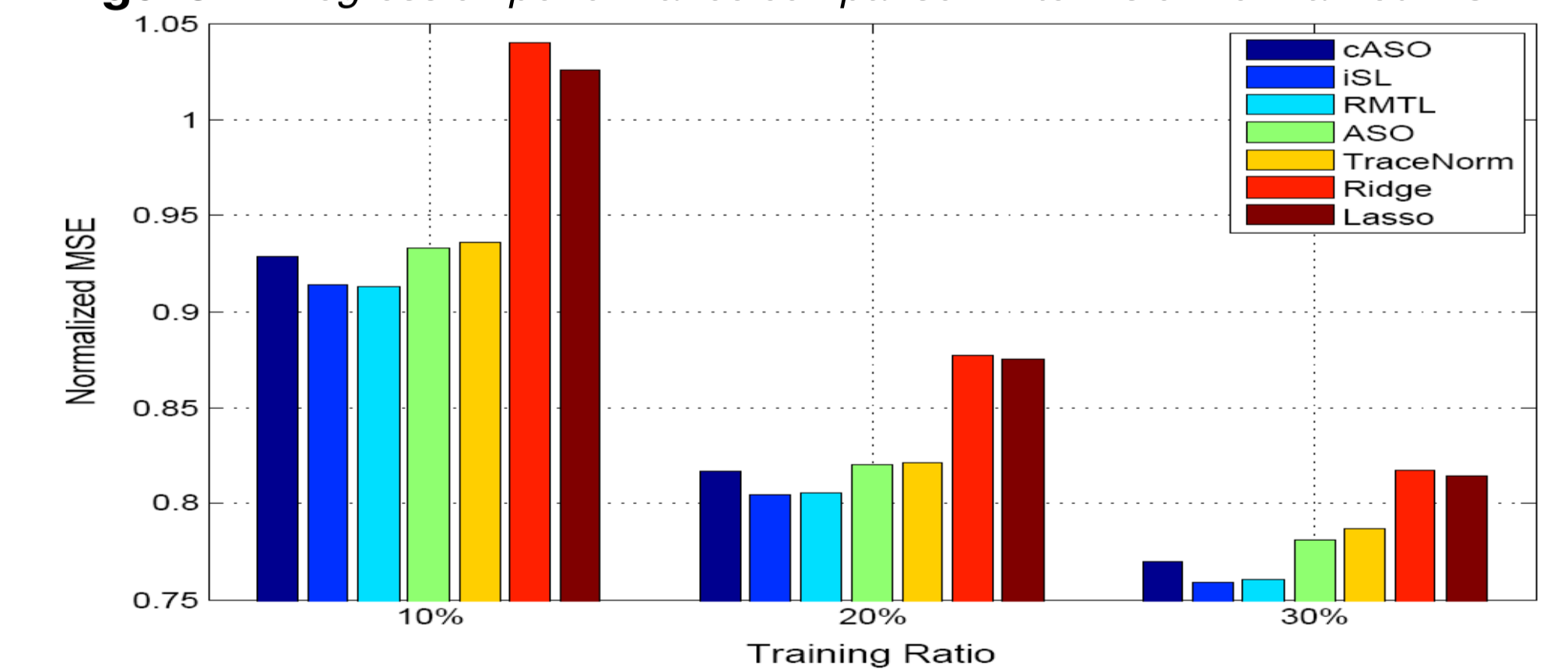
Figure 6: Classification performance comparison in terms of AUC.



Application 2 - Exam Score Prediction

The School data consists of 15362 student data from 139 school, where each student is represented by 27 features. We formulate the score prediction problem as a multi-task learning problem and the experimental results are presented in Figure 7.

Figure 7: Regression performance comparison in terms of Normalized MSE.



5. Conclusion

We proposed a series of MTL formulations in which the tasks relationship is captured by a shared low-rank structure. We further developed efficient optimization algorithms for solving the proposed MTL formulation. Experimental results on real-world applications demonstrate the effectiveness of the proposed MTL formulations and the efficiency of the proposed algorithms. We plan to conduct a theoretical analysis on the proposed MTL formulations and apply them to other real-world applications.

References

- [1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6, 2005.
- [2] R. Caruana. Multitask learning. *Machine Learning*, 28, 1997.
- [3] Y. Nesterov. *Introductory Lectures on Convex Programming*. Lecture Notes, 1998.