Multi-Task Model and Feature Joint Learning *

Ya Li, Xinmei Tian
University of Science and Technology of China
muziyiye@mail.ustc.edu.cn, xinmei@ustc.edu.cn

Tongliang Liu, Dacheng Tao
University of Technology, Sydney
tliang.liu@gmail.com, dacheng.tao@uts.edu.au

Abstract

Given several tasks, multi-task learning (MTL) learns multiple tasks jointly by exploring the interdependence between them. The basic assumption in MTL is that those tasks are indeed related. Existing MTL methods model the task relatedness/interdependence in two different ways, either common parameter-sharing or common feature-sharing across tasks. In this paper, we propose a novel multi-task learning method to jointly learn shared parameters and shared feature representation. Our objective is to learn a set of common features with which the tasks are related as closely as possible, therefore common parameters shared across tasks can be optimally learned. We present a detailed deviation of our multi-task learning method and propose an alternating algorithm to solve the non-convex optimization problem. We further present a theoretical bound which directly demonstrates that the proposed multi-task learning method can successfully model the relatedness via joint common parameter- and common feature-learning. Extensive experiments are conducted on several real world multi-task learning datasets. All results demonstrate the effectiveness of our multi-task model and feature joint learning method.

1 Introduction

Multi-task learning jointly learns multiple tasks by exploring the interdependence between them. Recent works have witnessed the fast development of multi-task learning in various research areas, such as web image and video search [Wang *et al.*, 2009], disease prediction [Zhang and Shen, 2012], relative attributes learning [Chen *et al.*, 2014], *etc.* The basic assumption in MTL is that tasks are related, so learning one task will benefit from learning other tasks. The key problem in MTL, therefore, is how to model the relatedness/interdependence across tasks. Existing multi-task learning algorithms have two principal ways to learn the relatedness: sharing common models/parameters [Evgeniou and Pontil, 2004; Xue *et al.*, 2007; Yu *et al.*, 2005; Rai and Daume, 2010], and sharing common features representations [Argyriou *et al.*, 2008; Jebara, 2011; Lapin *et al.*, 2014].

MTL in the category of sharing common models/parameters (multi-task model learning) assumes that the tasks are related in such a way that the true models have something in common in their parameters. For example, Xue et al. constructed a hierarchical Bayesian framework for learning task relatedness using the Dirichlet process and assumed that the Bayesian models shared a common prior [Xue *et al.*, 2007]. Evgeniou and Pontil developed a novel multi-task learning method based on the minimization of regularization functions, similar to support vector machines, and assumed that the hyperplanes of all tasks are close to a mean SVM hyperplane [Evgeniou and Pontil, 2004].

MTL in the category of sharing common feature representations (multi-task feature learning) assumes that the tasks are related in the sense that they all share a small set of features. For example, a framework was proposed for learning sparse representations shared across multiple tasks [Argyriou *et al.*, 2008]. It is based on the well-known L1-norm regularized single-task learning and controls the number of learned common features across tasks. Jebara gave a summary of feature selection and kernel selection in [Jebara, 2011]. In considering the effectiveness of multi-task learning for high dimensional feature space, Lapin et al. proposed a novel multi-task learning method to learn a low dimensional representation jointly with corresponding classifiers [Lapin *et al.*, 2014].

Neither multi-task model learning nor multi-task feature learning can model relatedness well. Recent works have attempted to simultaneously learn model relatedness and feature relatedness [Li *et al.*, 2014; Yang *et al.*, 2013]. Multi-task model learning directly mines relatedness in the original feature space. However, the performance of multi-task model learning may be degraded, as relatedness measured by the

*This work is supported by the NSFC 61201413 and 61390514, the Fundamental Research Funds for the Central Universities WK2100060011 and WK2100100021, Australian Research Council Projects: FT-130101457, DP-140102164, and LP-140100569. Y. Li and X. Tian are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China, China. T. Liu and D. Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia

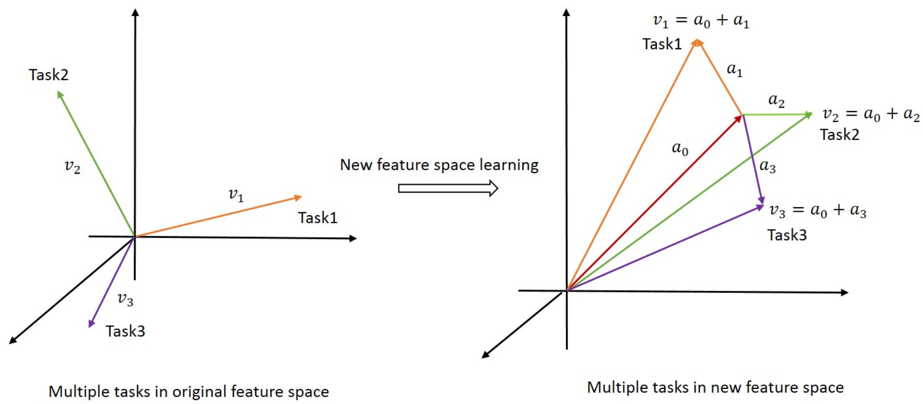

Figure 1: Illustration of our multi-task model and feature joint learning.

original features may not be obvious in a real world dataset. Multi-task feature learning solves this problem by mining potentially common feature representations, but ignores the model relatedness between tasks in the learned common feature space. In this paper we propose a multi-task learning method, which jointly learns shared model and shared feature representation. In our multi-task model-feature joint learning method, we learn a set of common features shared by multiple tasks to maximize task relatedness, therefore common models shared across tasks can be optimally learned simultaneously. The proposed method is formalized as a non-convex problem. We propose an alternating algorithm to solve this challenging problem. Theoretical analyses are also presented which prove that joint model and feature learning is able to model task relatedness well.

2 Multi-task model and feature joint learning

Our main idea is to learn a shared model and shared feature representations simultaneously, as illustrated in Figure 1. Multiple tasks in real world applications may not be closely related due to complexity and noise. In other words, they have weak interdependence and their models/hyperplanes may differ significantly in the original feature space. We hope to learn a feature mapping matrix $\mathbf{U}$, through which the hyperplanes of all tasks are closely related, enabling them to share a common hyperplane a_0. a_t is the offset of the t-th task, which compensates for the limitation of the study ability of the mapping matrix U and reflects its own unique characteristics.

2.1 The proposed formulation

Assume we are given T different learning tasks. Each task t is associated with a set of data:

$$D_t = \{(x_{t1}, y_{t1}), (x_{t2}, y_{t2}), ..., (x_{tm_t}, y_{tm_t})\},$$

where x_{ti} is the $i-$th input feature and y_{ti} is its corresponding output. $x_{ti} \in \mathbb{R}^d, y_{ti} \in \mathbb{R}, t \in \{1, 2, ..., T\}$, and $i \in \{1, 2, ..., m_t\}$. Our goal is to learn T different linear functions using the above T datasets $\{D_1, D_2, ..., D_T\}$ as follows:

$$f_t(x_{ti}) = v_t^T x_{ti} \approx y_{ti}. \tag{1}$$

Single-task learning methods learn the T different linear functions separately using their own data (such as linear regression, SVMs), while multi-task learning methods learn the T different functions jointly by mining the relationships between tasks.

Our objective is to learn an orthogonal feature mapping matrix U through which all tasks can share a central hyperplane a_0 but also preserve their unique model a_t,

$$f_t(x_{ti}) = \langle a_t + a_0, U^T x_{ti} \rangle. \tag{2}$$

The central hyperplane a_0 represents the interdependent information across tasks. The offset a_t captures the unique characteristic of each task. Both a_0 and a_t are learned in the new feature space. We give our proposed multi-task learning model as follows:

$$\min_{V, a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} l\left(y_{ti}, \langle v_t, U^T x_{ti} \rangle\right) + \tfrac{\gamma}{T} \|V - a_0 * \mathbf{1}\|_{2,1}^2 + \beta \|a_0\|_2^2, \tag{3}$$

where $V = [v_1, v_2, ..., v_T]$. $\mathbf{1}$ is a $1 \times T$ vector with all entries being 1. $\|a_0\|_2$ is the 2-norm of vector a_0, which can be formulated as $\|a_0\|_2 = (\sum_{i=1}^{d} |a_{0i}|^2)^{\frac{1}{2}}$. This regularization term is used to guarantee the smoothness of the central hyperplane a_0. $\|V - a_0 * \mathbf{1}\|_{2,1}$ represents the $(2,1)$-norm of matrix $(V - a_0 * \mathbf{1})$, which can be formulated as $\|V - a_0 * \mathbf{1}\|_{2,1} = (\sum_{i=1}^{d} \|v^i - a_{0i} * \mathbf{1}\|_2)$. v^i is the i-th row of matrix V. The $(2,1)$-norm regularization ensures that common features will be selected across all tasks. It encourages the group sparse property, which means that many rows of the learned matrix $(V - a_0 * \mathbf{1})$ are all zero.

Note $v_t = a_t + a_0$, problem (3) can be rewritten as

$$\min_{A, a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} l\left(y_{ti}, \langle a_t + a_0, U^T x_{ti} \rangle\right) + \tfrac{\gamma}{T} \|A\|_{2,1}^2 + \beta \|a_0\|_2^2, \tag{4}$$

where $A = [a_1, a_2, ..., a_T]$.

Our proposed formulation differs from the formulation proposed in [Argyriou *et al.*, 2008] in three main aspects. First, the method proposed in [Argyriou *et al.*, 2008] ignores the limitation of the learning ability of feature mapping matrix U. It may be more feasible to select the common features by regularizing V around $a_0 \in \mathbb{R}^d$ rather than the original point. Our proposed problem (3) learns the shared feature around point a_0 instead of the original point. Second, the method proposed in [Argyriou *et al.*, 2008] focused on learning shared features and did not consider the relationships of task models. After the shared feature was learned, they treated multiple tasks independently when learning their model parameters. From the formulation of problem (4), we can see that our method jointly learns the shared features and shared common parameters. Third, the minimization of our proposed objective function will be more difficult because we learn the shared features and shared common parameters simultaneously. This will be shown in the following section.

2.2 Equivalent convex optimization problem

Problem (4) is a non-convex problem. It is difficult to solve such a non-convex optimization problem directly. We will give an equivalent convex optimization problem of problem (4) in this section [Argyriou *et al.*, 2008].

Theorem 1. *Problem (4) is equivalent to the following convex optimization problem:*

$$\begin{aligned} \min_{W,w_0,D} \quad & \sum_{t=1}^{T}\sum_{i=1}^{m_t} l\left(y_{ti}, \langle w_t + w_0, x_{ti}\rangle\right) \\ & + \frac{\gamma}{T}\sum_{t=1}^{T}\langle w_t, D^{+}w_t\rangle + \beta\langle w_0, w_0\rangle, \\ \textit{s.t.} \quad & D \in S_+^d, trace(D) \leq 1, range(W) \subseteq range(D). \end{aligned} \tag{5}$$

In particular, if ($\hat{A}$,$\hat{a_0}$,$\hat{U}$) *is an optimal solution of problem (4), then* $\hat{W} = \hat{U}\hat{A}, \hat{w_0} = \hat{U}\hat{a_0}, \hat{D} = \hat{U}Diag(\frac{\|\hat{a}^i\|_2}{\|\hat{A}\|_{2,1}})_{i=1}^d \hat{U}^T$ *is an optimal solution of problem (5). Conversely, if* $(\hat{W}, \hat{w_0}, \hat{D})$ *is an optimal solution of problem (5) then any* $(\hat{A}, \hat{a_0}, \hat{U})$*, such that the columns of* $\hat{U}$ *form an orthonormal basis of eigenvectors of* $\hat{D}$ *and* $\hat{A} = \hat{U}^T\hat{W}, \hat{a_0} = \hat{U}^T\hat{w_0}$ *is an optimal solution of problem (4).*

Noting that S_+^d represents the set of positive semidefinite symmetric matrices and range(W) denotes the set $\{x \in \mathbb{R}^n : x = Wz, \text{for some } z \in \mathbb{R}^T\}$. $Diag(a_0)_{i=1}^d$ represents a diagonal matrix with the components of vector a_0 on the diagonal. D^+ is the pseudoinverse of matrix D.

2.3 An optimization algorithm

In this section, we propose an alternating algorithm to solve problem (5) by alternately minimizing it with respect to (W, w_0) and D, as presented in Algorithm 1. We can ultimately obtain the solution to problem (4) through the relationships between the optimal solution of problem (4) and problem (5) in Theorem 1.

In Algorithm 1, we first fix D and minimize the problem over (W, w_0). When D is fixed, the minimization over w_t cannot simply be separated into T independent problems because of the existence of w_0. Therefore, it is more difficult to solve our proposed problem. It is shown as follows:

Algorithm 1 Multi-task model and feature joint learning

Input: training data $\{(x_{ti}, y_{ti})\}_{i=1}^{m_t}, t \in \{1, 2, ..., T\}$
Output: W, w_0, D
Initialize $D = \frac{I}{d}$, d is the dimension of the data
while $\|W - W_{prev}\| > tol_1 \quad or \quad \|w_0 - w_{0prev}\| > tol_2$ **do**
  $\min_{W_1,W_0} \|Y - X^T(W_1 + W_0)\| + \frac{\gamma}{T}W_1^T D_0^+ W_1 + \beta w_0^T w_0$
  $D = \frac{(WW^T)^{\frac{1}{2}}}{trace(WW^T)^{\frac{1}{2}}}$
end while

$$\begin{aligned} \min_{W,w_0} \quad & \sum_{t=1}^{T}\sum_{i=1}^{m_t} l\left(y_{ti}, \langle w_t + w_0, x_{ti}\rangle\right) \\ & + \frac{\gamma}{T}\sum_{t=1}^{T}\langle w_t, D^{+}w_t\rangle + \beta\langle w_0, w_0\rangle, \\ \text{s.t.} \quad & D \in S_+^d, trace(D) \leq 1, range(W) \subseteq range(D). \end{aligned} \tag{6}$$

We consider the situation in which the loss function is a least squared loss and make changes to solve the above problem. Suppose $X_t = [x_{t1}, x_{t2}, ..., x_{tm_t}] \in \mathbb{R}^{d \times m_t}$ represents all data points in task t. $Y_t = [y_{t1}, y_{t2}, ..., y_{tm_t}]^T \in \mathbb{R}^{m_t}$ represents the outputs of the m_t data points in task t. M is the total number of data points of all T tasks:

$$M = m_1 + m_2 + ... + m_T.$$

Let $X = bdiag(X_1, X_2, ..., X_T) \in \mathbb{R}^{dT \times M}$ and $Y = [Y_1^T, Y_2^T, ..., Y_T^T]^T \in \mathbb{R}^M$, X represents a block diagonal matrix with the data of T different tasks as the diagonal elements. Y is the output vector of all data points in the T tasks by aligning the outputs of each of the tasks. Let $D_0 = bdiag(\underbrace{D, D, ..., D}_{T}) \in \mathbb{R}^{dT \times M}$, $W_0 = [\underbrace{w_0^T, w_0^T, ..., w_0^T}_{T}]^T \in \mathbb{R}^{dT}$ and $W_1 = [w_1^T, w_2^T, ..., w_T^T]^T \in \mathbb{R}^{dT}$.

Problem (6) can be reformulated as

$$\min_{W_1,W_0} \|Y - X^T(W_1 + W_0)\|_2^2 + \frac{\gamma}{T}W_1^T D_0^+ W_1 + \beta w_0^T w_0. \tag{7}$$

Let I be a $d \times d$ identity matrix and $I_0 = [\underbrace{I, I, ..., I}_{T}]^T \in \mathbb{R}^{dT \times d}$, then $W_0 = I_0 \times w_0$. In fact, problem (7) can be formulated as a standard 2-norm regularization problem if we introduce new variables. Let $Z_1 = \sqrt{\frac{\gamma}{T}}(D_0^+)^{\frac{1}{2}}W_1$, $Z_2 = \sqrt{\beta}w_0$. Then $W_1 = \sqrt{\frac{T}{\gamma}}(D_0^+)^{-\frac{1}{2}}Z_1$ and $W_0 = \sqrt{\frac{1}{\beta}}I_0 Z_2$. $(D_0^+)^{\frac{1}{2}} = bdiag(\underbrace{(D^+)^{\frac{1}{2}}, (D^+)^{\frac{1}{2}}, ..., (D^+)^{\frac{1}{2}}}_{T})$ and $(D_0^+)^{-\frac{1}{2}} = bdiag(\underbrace{(D^+)^{-\frac{1}{2}}, (D^+)^{-\frac{1}{2}}, ..., (D^+)^{-\frac{1}{2}}}_{T})$. We have

$$\frac{\gamma}{T} W_1^T D_0^+ W_1 + \beta w_0^T w_0 = [Z_1^T, Z_2^T][Z_1^T, Z_2^T]^T = Z^T Z$$
$$W_1 + W_0 = [\sqrt{\frac{T}{\gamma}} (D_0^+)^{-\frac{1}{2}}, \sqrt{\frac{1}{\beta}} I_0][Z_1^T, Z_2^T]^T = PZ \tag{8}$$

where $Z = [Z_1^T, Z_2^T]$ and $P = [\sqrt{\frac{T}{\gamma}} (D_0^+)^{-\frac{1}{2}}, \sqrt{\frac{1}{\beta}} I_0]$. Then problem (7) can be formulated as follows:

$$\min_{Z} \|Y - X^T P Z\|_2^2 + Z^T Z. \tag{9}$$

The above problem is a standard 2-norm regularization problem and has an explicit solution:

$$Z = (P^T X X^T P + I)^{-1} P^T X Y. \tag{10}$$

W and W_0 can be derived from Z, then problem (6) is solved.

The second step of Alogorithm 1 is to fix (W, w_0) and minimize problem (5) over D. We just need to solve the following problem for a fixed W and w_0:

$$\begin{aligned} \min_{D} \quad & \sum_{t=1}^{T} \langle w_t, D^+ w_t \rangle, \\ \text{s.t.} \quad & D \in S_+^d, trace(D) \leq 1, range(W) \subseteq range(D). \end{aligned} \tag{11}$$

The optimal solution is given as follows [Argyriou *et al.*, 2008]:

$$\hat{D} = \frac{(WW^T)^{\frac{1}{2}}}{trace(WW^T)^{\frac{1}{2}}}. \tag{12}$$

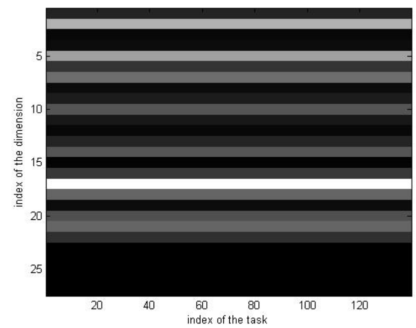

Figure 2: Absolute value of learned weight matrix A_0.

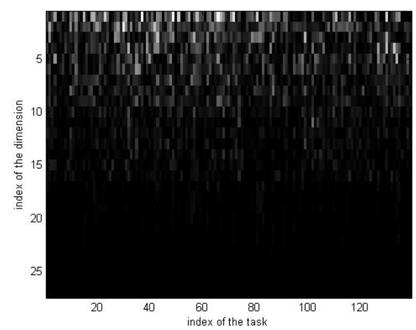

Figure 3: Absolute value of learned weight matrix A.

3 Theoretical Analysis

In this section, we derive a generalization bound for proposed problem (4). We change the soft constraints $\frac{\gamma}{T}\|A\|_{2,1}^2$ and $\beta\|a_0\|_2^2$ into hard constraints. Then, problem (4) becomes:

$$\begin{aligned} \min_{a_t, a_0, U, \varepsilon_1, \varepsilon_2} \quad & \sum_{t=1}^{T} \sum_{i=1}^{m_t} l\left(y_{ti}, \left\langle a_t + a_0, U^T x_{ti} \right\rangle\right) + \varepsilon_1 + \varepsilon_2, \\ \text{s.t.} \quad & \gamma \frac{1}{T} \|A\|_{2,1}^2 \leq \varepsilon_1, \\ & \beta \|a_0\|_2^2 \leq \varepsilon_2. \end{aligned} \tag{13}$$

Note that problem (13) is equal to problem (4) and that both ε_1 and ε_2 are of order $\mathcal{O}(1)$ (see [Vainsencher *et al.*, 2011]). Let $\varepsilon_1 = \varepsilon_2 = \mathcal{O}(1)$, problem (13) becomes:

$$\begin{aligned} \min_{a_t, a_0, U} \quad & \sum_{t=1}^{T} \sum_{i=1}^{m_t} l\left(y_{ti}, \left\langle a_t + a_0, U^T x_{ti} \right\rangle\right), \\ \text{s.t.} \quad & \|A\|_{2,1}^2 \leq \mathcal{O}\left(\frac{T}{\gamma}\right), \\ & \|a_0\|_2^2 \leq \mathcal{O}\left(\frac{1}{\beta}\right). \end{aligned} \tag{14}$$

Thus, a problem with soft constraints can be analyzed in the form of hard constraints. Mehta and Gray [Mehta and Gray, 2013] directly set $\varepsilon = 1$ to analyze a soft constraint problem by changing it into a hard constraint problem. We will analyze proposed problem (4) in the same way and provide a generalization bound to the following problem:

$$\begin{aligned} \min_{a_t, a_0, U} \quad & \sum_{t=1}^{T} \sum_{i=1}^{m_t} l\left(y_{ti}, \left\langle a_t + a_0, U^T x_{ti} \right\rangle\right), \\ \text{s.t.} \quad & \|A\|_{2,1}^2 \leq \frac{T}{\gamma}, \\ & \|a_0\|_2^2 \leq \frac{1}{\beta}. \end{aligned} \tag{15}$$

To upper bound the generalization error, we assume that the loss function l satisfies the following Lipschitz-like condition, which has been widely used (see [Mohri *et al.*, 2012]).

Definition 1. *A loss function l is c-admissible with respect to the hypothesis class H if there exists a $c \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative real numbers, such that for any two hypotheses $h, h' \in H$ and example $(x, y) \in \mathcal{X} \times \mathbb{R}$, the following inequality holds:*

$$|l(y, h(x)) - l(y, h'(x))| \leq c|h(x) - h'(x)|.$$

The result is as follows:

Theorem 2. *Let the loss function l be upper bounded by B, that is $l(y, f(x)) \leq B$, and be c-admissible with respect to the linear function class. For any A, a_0 and U learned by*

Table 1: Performance comparison between our proposed MFJL method and seven baseline methods on School dataset in terms of averaged nMSE and aMSE.

Measure	Training ratio	Ridge	Lasso	TraceNorm	Sparse-LowRank	CMTL	RMTL	DirtyMTL	MFJL
nMSE	10%	1.0398	1.0261	0.9359	0.9175	0.9413	0.9130	0.9543	**0.7783**
	20%	0.8773	0.8754	0.8211	0.8126	0.8327	0.8055	0.8396	**0.7432**
	30%	0.8171	0.8144	0.7870	0.7657	0.7922	0.7600	0.7985	**0.7299**
aMSE	10%	0.2713	0.2682	0.2504	0.2419	0.2552	0.2330	0.2327	**0.1898**
	20%	0.2303	0.2289	0.2156	0.2114	0.2131	0.2018	0.2048	**0.1813**
	30%	0.2156	0.2137	0.2089	0.2011	0.1922	0.1822	0.1943	**0.1776**

Table 2: Performance comparison of multi-task regression algorithms on SARCOS dataset in terms of averaged nMSE and aMSE.

Measure	Training size	Ridge	Lasso	TraceNorm	Sparse-LowRank	CMTL	RMTL	DirtyMTL	MFJL
nMSE	50	0.2454	0.2337	0.2257	0.2127	0.2192	0.2123	0.1742	**0.1640**
	100	0.1821	0.1616	0.1531	0.1495	0.1568	0.1456	0.1274	**0.1155**
	150	0.1501	0.1469	0.1318	0.1236	0.1301	0.1245	0.1129	**0.1057**
aMSE	50	0.1330	0.1228	0.1122	0.1073	0.1156	0.0982	0.0625	**0.0588**
	100	0.1053	0.0907	0.0805	0.0793	0.0852	0.0737	0.0458	**0.0415**
	150	0.0846	0.0822	0.0772	0.0661	0.0755	0.0674	0.0405	**0.0379**

problem (4) with the soft constraints about $\gamma\frac{1}{T}\|A\|_{2,1}^2$ and $\beta\|a_0\|_2^2$ being replaced by the hard constraints $\|A\|_{2,1}^2 \le \frac{T}{\gamma}$ and $\|a_0\|_2^2 \le \frac{1}{\beta}$, and for any $\delta > 0$, with probability at least $1-\delta$, we have

$$E_x \sum_{t=1}^{T}\sum_{i=1}^{m_t} l\left(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle\right) - \sum_{t=1}^{T}\sum_{i=1}^{m_t} l\left(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle\right) \le 2c\left(\sqrt{\frac{T}{\gamma}} + \sqrt{\frac{1}{\beta}}\right)\sqrt{\sum_{t=1}^{T} m_t S(X_t)} + 3B\sqrt{\frac{\sum_{t=1}^{T} m_t \ln(\frac{2}{\delta})}{2}},$$

where $S(X_t) = tr\left(\hat{\Sigma}(x_t)\right) = \frac{1}{m_t}\sum_{i=1}^{m_t}\|x_{ti}\|_2^2$ is the empirical covariance for the observations of the t-th task. Let $m_1 = \ldots = m_T = m$ and $\|x_t\|_2 \le r, t = 1, \ldots, T$, with probability at least $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T} E_x l\left(y_t, \langle a_t + a_0, U^T x_t\rangle\right) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m} l\left(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle\right) \le \frac{2cr}{\sqrt{\gamma m}} + \frac{2cr}{\sqrt{\beta m T}} + 3B\sqrt{\frac{\ln(2/\delta)}{2mT}}.$$

Remark 1. *According to Theorem 2, the two terms $\frac{2cr}{\sqrt{\gamma m}}$ and $\frac{2cr}{\sqrt{\beta m T}}$ are the generalization bounds with respect to the learning of $A = (a_1, \ldots, a_T)$ and a_0, respectively, where A corresponds to the specific task in the set of multiple tasks, and a_0 corresponds to the shared hyperplane of the multiple tasks. Our theoretical result shows that a_0 can be learned with the order of $\mathcal{O}(\sqrt{1/mT})$, which means the shared hyperplane can be successfully learned by increasing the number of tasks. We have therefore theoretically justified that the proposed multi-task learning is superior to single task learning. Additionally, the employed orthogonal operator U and regularization on $\|A\|_{2,1}$ will encourage a_0 to be as large as possible. The generalization bound of problem (4) will therefore converge faster than that of problem proposed in [Argyriou* et al.*, 2008], which means the proposed method is more efficient.*

4 Experiments

In this section, we present extensive experiments conducted on several real-world datasets including School, SARCOS, and Isolet. These datasets have been wildly used for evaluation in previous multi-task learning works, for example in [Argyriou *et al.*, 2008; Gong *et al.*, 2012b; Chen *et al.*, 2011; Kang *et al.*, 2011; Gong *et al.*, 2012a]. We compare the performance of our proposed multi-task model and feature joint learning (MFJL) method with two single-task learning methods and five state-of-the-art multi-task learning methods. The two single-task learning methods are ridge regression (Ridge) and least squares with L1-norm regularization (Lasso). The five multi-task learning methods are least squares with trace norm regularization (TraceNorm), least squares with low-rank and sparse structures regularization (Sparse-LowRank) [Chen *et al.*, 2012], convex multi-task feature learning (CMTL) [Argyriou *et al.*, 2008], robust multi-task learning with low-rank and group-sparse structures (RMTL) [Chen *et al.*, 2011] and dirty model multi-task regression learning (DirtyMTL) [Jalali *et al.*, 2013]. We choose

Table 3: Performance comparison of multi-task regression algorithms on Isolet dataset in terms of averaged nMSE and aMSE.

Measure	Training ratio	TraceNorm	Sparse-LowRank	CMTL	RMTL	DirtyMTL	MFJL
nMSE	15%	0.6044	0.6307	0.7000	0.5987	0.6764	**0.5691**
	20%	0.5705	0.6166	0.6491	0.5741	0.6344	**0.5526**
	25%	0.5622	0.6011	0.6288	0.5635	0.6212	**0.5498**
aMSE	15%	0.1424	0.1486	0.1650	0.1411	0.1594	**0.1314**
	20%	0.1343	0.1452	0.1528	0.1352	0.1494	**0.1301**
	25%	0.1321	0.1412	0.1477	0.1324	0.1459	**0.1292**

these five multi-task learning methods as our competitors because their objective formulations are similar to ours and they have achieved top-level performance on benchmark datasets. All these methods use a least square loss function.

4.1 School dataset

The School dataset is from the Inner London Education Authority. It consists of the examination scores of 15,362 students from 139 secondary schools in 1985, 1986 and 1987. There are 139 tasks in total, corresponding to examination scores prediction in each school. The input features include the year of the examination, 4 school dependent features and 3 student-dependent features. We follow the same setup as previous multi-task learning works and obtain a 27-dimensional binary variable for each example.

We randomly select 10%, 20% and 30% of the examples in each respective task as a training set, and the remaining examples are used for testing. For each training ratio, we repeat the random splits of the data 10 times and report the average performance. The parameters of all methods are tuned via cross-validation on the training set. We evaluate all these regression methods using normalized mean squared error (nMSE) and averaged mean squared error (aMSE) [Gong *et al.*, 2012b; Chen *et al.*, 2011].

The experimental results are shown in Table 1. From these results, we make the following observations. (1) All multi-task learning methods outperform single-task learning methods, which proves the effectiveness of multi-task learning. (2) Our proposed joint learning method significantly and consistently outperforms all other baseline MTL methods, especially when the training ratio is small. This demonstrates that our method can successfully learn the optimal feature space in which multiple tasks are closely related.

We also show the absolute values of learned weight $A_0 = [\underbrace{a_0, a_0, ..., a_0}_{T}]$ and A in Figure 2 and Figure 3, respectively. Here, the training ratio is 20%. The black areas in the figures denote zero value. From Figure 3, we can see that the learned weight matrix A is very sparse, and there are about 15 nonzero rows referring to the shared features across tasks. As for matrix A_0, we find that some features not shared in matrix A will be used in the central hyperplane a_0, which will increase the utilization of information in the features.

4.2 SARCOS dataset

The SARCOS dataset is related to an inverse dynamic problem for a seven degree-of-freedom SARCOS anthropomorphic robot arm. It consists of 48,933 observations corresponding to seven joint torques. Each observation is described by a 21-dimensional feature vector, including seven joint positions, seven joint velocities, and seven joint accelerations, therefore we have seven tasks in total. Our task is to map the 21-dimensional features to the seven joint torques. We randomly select 50, 100 and 150 examples to form three separate training sets respectively, and randomly select 5000 examples as test sets. All experiments are run 15 times to avoid randomness. The validation methods are the same as described in the experiments on the School dataset for all the multi-task learning methods.

The experimental results in terms of averaged nMSE and aMSE are given in Table 2. We make similar observations to those in the experiment on the School dataset. Our proposed joint learning method again achieves much better performance than other baseline algorithms, which demonstrates the effectiveness and robustness of our proposed multi-task model and feature joint learning method.

4.3 Isolet dataset

We have also tested our method on the Isolet dataset. This dataset is collected from 150 speakers, each of whom speaks all the English letter of the alphabet twice, i.e., each speaker provides 52 data examples. The speakers are grouped into five subsets of 30 similar speakers; thus we have five tasks corresponding to the five speaker groups. The five tasks have 1560, 1560, 1560, 1558 and 1559 corresponding samples. Each English letter corresponds to a label (1-26) and we treat the English letter labels as regression values following the same setup as [Gong *et al.*, 2012a]. We randomly select 15%, 20%, 25% of the samples to form three training sets and use the rest of the samples as test sets. We first preprocess the data with PCA by reducing the dimensionality to 100. Experiments are repeated 10 times.

Experiments on the School and SARCOS datasets have shown that Ridge and Lasso do not perform well for multi-task learning problem, therefore we only compare MFJL with five multi-task learning algorithms in this experiment. The experimental results on the Isolet dataset in terms of averaged nMSE and aMSE are given in Table 3. It is clear that our MFJL outperforms the other five multi-task learning methods stably, which demonstrates that our method is suitable for problems in a variety of applications.

5 Conclusion

In this paper, we propose a novel multi-task learning method by jointly learning shared parameters and shared feature representations. A detailed description of our proposed multi-task learning method is provided. Additionally, we present a theoretical bound which directly demonstrates that the proposed multi-task learning method performs better than single-task learning and is able to successfully model the relatedness between tasks. Various experiments are conducted on several landmark datasets and all results demonstrate the effectiveness of our proposed multi-task learning method.

References

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Chen *et al.*, 2011] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 42–50, 2011.

[Chen *et al.*, 2012] Jianhui Chen, Ji Liu, and Jieping Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):22, 2012.

[Chen *et al.*, 2014] Lin Chen, Qiang Zhang, and Baoxin Li. Predicting multiple attributes via relative multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*, pages 1027–1034. IEEE, 2014.

[Evgeniou and Pontil, 2004] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

[Gong *et al.*, 2012a] Pinghua Gong, Jieping Ye, and Changshui Zhang. Multi-stage multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2012.

[Gong *et al.*, 2012b] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Dining*, pages 895–903, 2012.

[Jalali *et al.*, 2013] Ali Jalali, Pradeep D. Ravikumar, and Sujay Sanghavi. A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory*, 59(12):7947–7968, 2013.

[Jebara, 2011] Tony Jebara. Multitask sparsity via maximum entropy discrimination. *The Journal of Machine Learning Research*, 12:75–110, 2011.

[Kang *et al.*, 2011] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

[Lapin *et al.*, 2014] Maksim Lapin, Bernt Schiele, and Matthias Hein. Scalable multitask representation learning for scene classification. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1434–1441, 2014.

[Li *et al.*, 2014] Chengtao Li, Jun Zhu, and Jianfei Chen. Bayesian max-margin multi-task learning with data augmentation. In *Proceedings of the 31st International Conference on Machine Learning*, pages 415–423, 2014.

[Mehta and Gray, 2013] Nishant Mehta and Alexander G Gray. Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 36–44, 2013.

[Mohri *et al.*, 2012] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

[Rai and Daume, 2010] Piyush Rai and Hal Daume. Infinite predictor subspace models for multitask learning. In *International Conference on Artificial Intelligence and Statistics*, pages 613–620, 2010.

[Vainsencher *et al.*, 2011] Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. The sample complexity of dictionary learning. *The Journal of Machine Learning Research*, 12:3259–3281, 2011.

[Wang *et al.*, 2009] Xiaogang Wang, Cha Zhang, and Zhengyou Zhang. Boosted multi-task learning for face verification with applications to web image and video search. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009.*, pages 142–149. IEEE, 2009.

[Xue *et al.*, 2007] Ya Xue, Xuejun Liao, Lawrence Carlin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:35–63, 2007.

[Yang *et al.*, 2013] Ming Yang, Yingming Li, et al. Multi-task learning with gaussian matrix generalized inverse gaussian model. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 423–431, 2013.

[Yu *et al.*, 2005] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. *Proceedings of the 22nd International Conference on Machine learning*, pages 1012–1019, 2005.

[Zhang and Shen, 2012] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *Neuroimage*, 59(2):895–907, 2012.