1 稀疏学习在多任务学习中的应用

龚平华　张长水
清华大学 自动化系，北京 100084

1 引言

近年来，具有稀疏结构的机器学习问题成为机器学习领域一个很活跃的研究课题，尤其是在互联网数据呈现爆炸式增长的今天，它已经成为人们从海量数据中提取有用信息的重要工具。简单来讲，我们把具有稀疏结构的机器学习问题称为稀疏学习（sparse learning）。在稀疏学习中，我们往往会得到稀疏的解，即解向量或者矩阵中很多元素为零，因而稀疏学习自动具有特征选择的直观解释，即某个元素为零表示相应的特征没有选上，相反，则表示相应的特征被选中。稀疏学习具有特征选择的功能，使得它在机器学习领域有着广泛的应用，特别是在多任务学习问题中，稀疏学习得到了特别的关注。多任务学习[1]通过把相关的任务放到一起学习，并发掘任务之间的共享信息，从而达到提高推广性能的目的，它是近年来稀疏学习领域的一个十分重要的应用问题。随着稀疏学习研究的不断深入，多任务学习的各种应用成果层出不穷，在目标识别[1]，语音分类[2]、手写字符识别[3]、生物医药信息挖掘[4,5]等领域都有很成功的应用。在多任务学习中，一个关键的问题是如何实现任务之间信息的共享。在现有的一些多任务学习算法中，任务之间共享的信息有很多种，包括神经网络中的隐层单元[1,6]、贝叶斯模型中的先验[7-10]、高斯过程的参数[11]、特征映射矩阵[12]、分类权重向量[13]、相似度量矩阵[2,14]、低秩子空间[15,16]、一组相关的特征[17-24]等。多任务特征学习，旨在任务之间学习出相关的特征，最近引起了人们广泛的关注，特别是在各种稀疏学习模型在很多问题中取得成功的应用后，人们对稀疏模型在多任务特征学习的应用研究更是热情高涨，但现有的一些多任务特征学习模型要求任务之间要么同时共享某一个特征，要么同时不共享某一个特

征[19-21,24]，这个要求在实际问题中有些苛刻，因为实际问题中很可能只有部分任务共享某些特征，而且还可能存在异常的任务。为此，本文介绍两种多任务特征学习的稀疏模型，它们分别从不从的角度来放松“所有任务共享某些特征”的要求。具体说，第 2 节将介绍一个鲁棒的多任务特征学习算法，该算法能够检测异常任务。第 3 节将介绍一个非凸的多任务特征学习模型，该模型能够使得一些特征被某些任务共享而不是被所有的任务共享。第 4 节将对全文做一个总结。

2 鲁棒多任务特征学习

我们首先介绍一个鲁棒的多任务特征学习模型[25]，该模型在学习到共享特征的同时，还能够检测到异常任务。具体来说，我们将权重矩阵分解成两个矩阵之和。对第一个矩阵施加一个组稀疏惩罚，以达到学习共享特征的目的。同时，对第二个矩阵的转置也施加一个组稀疏惩罚，以达到检测异常任务的目的。我们利用加速梯度下降法来有效求解相应的优化问题，使得提出的算法能够适用于大规模的问题。此外，我们对提出的模型进行了理论上的分析。具体来说，我们给出了参数估计误差和预测误差的界。另外，在假设真实的权重幅度大于噪声这个基础上，理论分析指出，在一定条件下，我们的模型能够获得真实共享的特征和异常任务。

下面介绍鲁棒多任务特征学习模型。假定有 m 个任务，分别对应训练数据$\{(X_1,\boldsymbol{y}_1),\cdots,(X_m,\boldsymbol{y}_m)\}$，其中 $X_i\in\mathbb{R}^{n_i\times d}$ 是第 i 个任务的数据矩阵（每一行是一个样本）；$\boldsymbol{y}_i\in\mathbb{R}^{n_i}$ 是第 i 个任务的预测目标（在回归问题中，y_i 是连续的回归值，在分类问题中，y_i 是离散的分类标签）；d 是样本的维数；n_i 是第 i 个任务的样本个数。我们归一化数据，使得 X_i 每一列的长度为 1，即

$$\sum_{j=1}^{n_i}(x_{jk}^{(i)})^2=1,\quad \forall k\in\mathbb{N}_d \tag{1}$$

其中，$x_{jk}^{(i)}$ 表示 X_i 的第(j,k)个元素；$\mathbb{N}_d$表示集合$\{1,\cdots,d\}$。我们为每个任务学习一个线性函数

$$\boldsymbol{y}_i\approx\boldsymbol{f}_i(X_i)=X_i w_i,\quad i\in\mathbb{N}_m \tag{2}$$

其中 $W=[w_1,\cdots,w_m]\in\mathbb{R}^{d\times m}$是权重矩阵。我们将 W 分解成两个矩阵之和，即 $W=P+Q$（请参考图 1 所示的示意图），同时分别对 P 和 Q 施加不同的惩罚，以挖掘任务之间的关系。

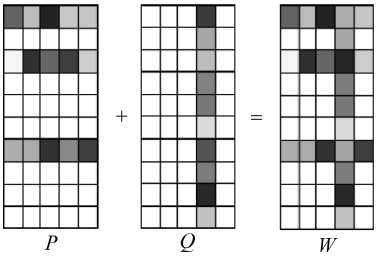

图 1 鲁棒多任务特征学习(rMTFL)权重矩阵分解示意图(该图来源于文献[25])
其中有色方块表示相应的特征被选中。这里共有 5 个任务,其中第 4 个任务是异常任务

具体来说,我们将鲁棒多任务特征学习 (robust multi-task feature learning,rMTFL) 建模成如下的优化问题:

$$\min_{W,P,Q}\sum_{i=1}^{m}\frac{1}{mn_i}\|X_i w_i-\boldsymbol{y}_i\|^2+\lambda_1\|P\|_{1,2}+\lambda_2\|Q^{\mathrm{T}}\|_{1,2},\quad \text{s.t. } W=P+Q \tag{3}$$

其中第一个对 P 的正则项用来获取任务之间的共享特征;第二个对 Q 的正则项用来发现异常的任务;λ_1 和 λ_2 是非负的正则项参数; P 的 $\ell_{1,2}$ 范数定义为: $\|P\|_{1,2}=\sum_{i=1}^{d}\|\boldsymbol{p}^i\|$,$\boldsymbol{p}^i$ 是 P 矩阵的第 i 行。具体来说,第一个正则项对 P 施加一个 $\ell_{1,2}$ 范数的惩罚,这使得问题(3)的最优解 P^* 的每一行要么都是零,要么都不是零[18]。这样,所有的任务要么同时选上某一组特征,要么同时不选某一组特征,然而,限制所有的任务同时共享一组共同的特征在实际的应用中可能太苛刻,因为异常的任务往往会存在。为了解决这一问题,我们引入了一个对 Q 的正则项,用来发现异常的任务。类似地,问题(3)的最优解 Q^* 的每一列要么都是零,要么都不是零,不是零的列就对应异常的任务。直观上来讲,如果 Q^* 的第 i 列的每个分量都是非零的,那么 W^* 的第 i 列的每个分量都是非零的,这样第 i 个任务与其他的任务就不再共享相同的特征,从而被检测为异常的任务,同时,对于其他的任务(对应于 Q^* 的非零列),则共享一组共同的特征(请参考图 1)。

接着我们将详细介绍如何有效求解问题(3)的鲁棒多任务特征学习优化问题。记

$$\begin{aligned} l(P,Q)&=\sum_{i=1}^{m}\frac{1}{mn_i}\|X_i(\boldsymbol{p}_i+\boldsymbol{q}_i)-\boldsymbol{y}_i\|^2\\ r(P,Q)&=\lambda_1\|P\|_{1,2}+\lambda_2\|Q^{\mathrm{T}}\|_{1,2}\end{aligned} \tag{4}$$

其中,$l(P,Q)$是经验损失函数;$r(P,Q)$是正则项;$\boldsymbol{p}_i$ 表示矩阵 P 的第 i 列。我们注意到问题(3)的目标函数是一个连续可微的函数 $l(P,Q)$与一个不可微的函数 $r(P,Q)$的和,

我们可以用加速的梯度下降法[26-28]来求解式(3)。记

$$T_{R,S,\eta}(P,Q) = l(R,S) + \left\langle \frac{\partial l(R,S)}{\partial R}, P-R \right\rangle + \frac{\eta}{2} \| P-R \|_F^2 + \left\langle \frac{\partial l(R,S)}{\partial S}, Q-S \right\rangle + \frac{\eta}{2} \| Q-S \|_F^2 \tag{5}$$

那么我们可以通过下列的迭代来求解问题(3)：

$$(P^{(k+1)},Q^{(k+1)}) = \arg\min_{P,Q} T_{R^{(k)},S^{(k)},\eta_k}(P,Q) + r(P,Q) \tag{6}$$

其中 $R^{(1)}=P^{(0)}$，$S^{(1)}=Q^{(0)}$，$R^{(k)}=P^{(k)}+\alpha_k(P^{(k)}-P^{(k-1)})$，$S^{(k)}=Q^{(k)}+\alpha_k(Q^{(k)}-Q^{(k-1)})$，$\forall k \geqslant 1$；$\eta_k$（$k \geqslant 1$）被设置成 $\eta_k = 2^{m_k}\eta_{k-1}$，其中 m_k 是使得下列线搜索条件成立的最小整数：

$$l(P^{(k+1)},Q^{(k+1)}) \leqslant T_{R^{(k)},S^{(k)},\eta_k}(P^{(k)},Q^{(k)}) \tag{7}$$

我们注意到$(R^{(k)},S^{(k)})$实际上是$(P^{(k)},Q^{(k)})$与$(P^{(k-1)},Q^{(k-1)})$的一个线性组合。这里，α_k对算法的收敛性能起到了很大的作用，根据 Beck 等人[28]的设定，我们取 $\alpha_k=(t_{k-1}-1)/t_k$，其中 $t_0=1$ 且对于所有 $k \geqslant 1$，$t_k=(1+\sqrt{t_{k-1}^2+1})/2$。具体算法流程如算法 1 所示。根据 Beck 等人[28]的理论分析，对于加速的梯度下降法，我们有以下的收敛定理：

定理 1 上述加速梯度下降法产生的序列$\{P^{(k)},Q^{(k)}\}$满足

$$f(P^{(k)},Q^{(k)}) - f(P^*,Q^*) = O\left(\frac{1}{k^2}\right) \tag{8}$$

其中 $f(\cdot,\cdot)$和(P^*,Q^*)分别表示问题(3)的目标函数和最优解。

算法 1 鲁棒多任务特征学习（rMTFL）

1. 初始化 $P^{(1)}=P^{(0)}$；$Q^{(1)}=Q^{(0)}$；$t_0=1$；
2. **For** $k=1,2,\cdots$ **do**
3. $\quad \alpha_k=(t_{k-1}-1)/t_k$；
4. $\quad R^{(k)}=P^{(k)}+\alpha_k(P^{(k)}-P^{(k-1)})$；$S^{(k)}=Q^{(k)}+\alpha_k(Q^{(k)}-Q^{(k-1)})$；
5. $\quad$ 根据式(6)计算$(P^{(k+1)},Q^{(k+1)})$直到线搜索条件(7)成立；
6. $\quad$ 更新 $t_k=(1+\sqrt{t_{k-1}^2+1})/2$
7. **End**

对于加速的梯度下降法，有两个事情需要得到妥善的解决：如何有效求解式(6)以及如何选择一个合适的初始值 η_0。由于式(6)的解耦性，我们将式(6)分解成下列两个独立的问题：

$$P^{(k+1)} = \arg\min_P \frac{1}{2}\left\| P - \left[R^{(k)} - \frac{1}{\eta_k}\nabla_R l(R^{(k)}, S^{(k)}) \right] \right\|_F^2 + \frac{\lambda_1}{\eta_k} \| P \|_{1,2}$$
$$Q^{(k+1)} = \arg\min_Q \frac{1}{2}\left\| Q - \left[S^{(k)} - \frac{1}{\eta_k}\nabla_S l(R^{(k)}, S^{(k)}) \right] \right\|_F^2 + \frac{\lambda_2}{\eta_k} \| Q^T \|_{1,2} \tag{9}$$

上述的两个问题都有如下的闭式解[20]，计算复杂度是 $O(dm)$：

$$(\boldsymbol{p}^{(k+1)})^i = \max\left(0, 1 - \frac{\lambda_1}{\eta_k \| (\boldsymbol{u}^{(k)})^i \|}\right)(\boldsymbol{u}^{(k)})^i, \quad \forall i \in \mathbb{N}_d$$
$$\boldsymbol{q}_j^{(k+1)} = \max\left(0, 1 - \frac{\lambda_2}{\eta_k \| v_j^{(k)} \|}\right) v_j^{(k)}, \quad \forall j \in \mathbb{N}_m \tag{10}$$

其中，$U^{(k)} = R^{(k)} - \frac{1}{\eta_k}\nabla_R l(R^{(k)}, S^{(k)})$，$V^{(k)} = S^{(k)} - \frac{1}{\eta_k}\nabla_S l(R^{(k)}, S^{(k)})$；$(\boldsymbol{u}^{(k)})^i$ 和 $v_j^{(k)}$ 分别表示 $U^{(k)}$ 的第 i 行和 $V^{(k)}$ 的第 j 列。

设置 η_0 为 $l(P,Q)$ 的梯度的 Lipschitz 常数 L 是一个比较好的步长初始化策略，但实际上，Lipschitz 常数 L 是未知的，而计算 L 所需要的计算量一般是很大的，虽然如此，我们能够给出 L 的上界与下界。记 $D \in \mathbb{R}^{\sum_{i=1}^m n_i \times dm}$ 为一个分块对角矩阵，其中 $\sqrt{\frac{2}{mn_i}} X_i (i \in \mathbb{N}_m)$ 是第 i 个分块矩阵，那么 Lipschitz 常数 L 就是 D 的最大奇异值的平方。根据矩阵范数的性质[29]，我们有：

$$\max\left\{ \frac{\| D \|_{\infty,1}^2}{dm}, \frac{\| D^T \|_{\infty,1}^2}{\sum_{i=1}^m n_i} \right\} \leqslant L \leqslant \| D \|_{\infty,1} \| D^{\mathrm{T}} \|_{\infty,1} \tag{11}$$

其中 $\| D \|_{\infty,1} = \max_{i=1}^{dm} \| \boldsymbol{d}^i \|_1$，$\boldsymbol{d}^i$ 是 D 矩阵的第 i 行。我们注意到 D 是一个分块对角矩阵，当 m（任务的个数）比较大时，D 是稀疏的，此时上述对 L 的上界与下界是比较紧的。如果我们设定 η_0 为 L 的上界，这时我们就不需要进行线搜索，因为当 $\eta_k \geqslant L$ 时，线搜索条件总是成立的[28]。虽然将 η_0 设置成 L 的上界能够不用进行线搜索，但这可能会由于每次的步长太小而导致外部迭代的次数很多。相反，如果我们设定 η_0 为 L 的下界，虽然我们需要进行线搜索，但这可能只需要比较少的外部迭代。在我们的实验中，我们用 L 的下界来初始化 η。

最后，我们对所提出的模型进行理论分析。为了简单起见，我们假设每个任务的训练样本个数都是 n。下面的定理描述了问题(3)最优解的一个十分重要的性质，它在后续的理论分析中起到了关键性的作用。

定理 2 令 $(\hat{P}, \hat{Q})$ 是(3)的一个最优解（其中 $m \geqslant 2, n, d \geqslant 1$）。假设 $\boldsymbol{y}_i = X_i w_i^* + \delta_i = \boldsymbol{f}_i^* + \delta_i (i \in \mathbb{N}_m)$，其中 $\delta_i \in \mathbb{R}^n$ 的每一个分量都是服从均值为 0，方差为 σ^2 的高斯分布；W^*

是可以被分解为两个矩阵之和的真实权重矩阵：$W^*=[w_1^*,\cdots,w_m^*]=P^*+Q^*\in\mathbb{R}^{d\times m}$。同时所有任务的数据 $X_i(i\in\mathbb{N}_m)$都按照式(1)进行归一化。选择正则化参数 λ_1 和 λ_2 为

$$\lambda_1,\lambda_2\geqslant\alpha,\quad \alpha=\frac{2\sigma}{mn}\sqrt{dm+t}, \tag{12}$$

其中 t 是一个正数。那么对于任意的 $P,Q\in\mathbb{R}^{d\times m}$，下列的不等式以不小于 $1-\exp\left(-\frac{1}{2}\left(t-dm\log\left(1+\frac{t}{dm}\right)\right)\right)$的概率成立：

$$\sum_{i=1}^{m}\frac{1}{mn}\left\|X_i(\hat{\boldsymbol{p}}_i+\hat{\boldsymbol{q}}_i)-\boldsymbol{f}_i^*\right\|^2\leqslant\sum_{i=1}^{m}\frac{1}{mn}\left\|X_i(\boldsymbol{p}_i+\boldsymbol{q}_i)-\boldsymbol{f}_i^*\right\|^2$$
$$+2\lambda_1\left\|(\hat{P}-P)^{\mathcal{J}(P)}\right\|_{1,2}+2\lambda_2\left\|(\hat{Q}^{\mathrm{T}}-Q^{\mathrm{T}})^{\mathcal{J}(Q^{\mathrm{T}})}\right\|_{1,2} \tag{13}$$

其中 $\mathcal{J}(P)$表示 P 的非零行的索引集合。

在定理 2 的基础上，我们给出几个鲁棒多任务特征学习的理论的界。首先，我们引入几个符号。令 $X\in\mathbb{R}^{mn\times dm}$ 为一个分块对角矩阵，其中 $X_i\in\mathbb{R}^{n\times d}(i\in\mathbb{N}_m)$是第 i 个分块矩阵；$F^*=[\boldsymbol{f}_1^*,\cdots,\boldsymbol{f}_m^*]$是真实预测矩阵。定义一个在任意矩阵 $A\in\mathbb{R}^{d\times m}$上的向量化操作运算符'vec'为 $\mathrm{vec}(A)=[\boldsymbol{a}_1^{\mathrm{T}},\cdots,\boldsymbol{a}_m^{\mathrm{T}}]^{\mathrm{T}}$。那么式(13)就可以写成下列等价的形式：

$$\frac{1}{mn}\left\|X\mathrm{vec}(\hat{P}+\hat{Q})-\mathrm{vec}(F^*)\right\|^2\leqslant\frac{1}{mn}\left\|X\mathrm{vec}(P+Q)-\mathrm{vec}(F^*)\right\|^2$$
$$+2\lambda_1\left\|(\hat{P}-P)^{\mathcal{J}(P)}\right\|_{1,2}+2\lambda_2\left\|(\hat{Q}^{\mathrm{T}}-Q^{\mathrm{T}})^{\mathcal{J}(Q^{\mathrm{T}})}\right\|_{1,2} \tag{14}$$

下面给出一个对训练数据矩阵的假设，这个假设是受限特征值假设[30]的一个推广。

假设 1 对于矩阵 $\Gamma_P\in\mathbb{R}^{d\times m}$和 $\Gamma_Q\in\mathbb{R}^{d\times m}$，令 r 和 $c(1\leqslant r\leqslant d,1\leqslant c\leqslant m)$分别是$|\mathcal{J}(P^*)|$和$|\mathcal{J}(Q^{*\mathrm{T}})|$的上界，$\beta_1$ 和 β_2 都是正数。假设存在两个正数 $\kappa_1(r)$和 $\kappa_2(c)$使得

$$\begin{aligned}\kappa_1(r)&=\min_{\Gamma_P,\Gamma_Q\in\mathcal{R}(r,c)}\frac{\left\|X\mathrm{vec}(\Gamma_P+\Gamma_Q)\right\|}{\sqrt{mn}\left\|(\Gamma_P)^{\mathcal{J}(P)}\right\|_F}\\ \kappa_2(c)&=\min_{\Gamma_P,\Gamma_Q\in\mathcal{R}(r,c)}\frac{\left\|X\mathrm{vec}(\Gamma_P+\Gamma_Q)\right\|}{\sqrt{mn}\left\|(\Gamma_Q^{\mathrm{T}})^{\mathcal{J}(Q^{\mathrm{T}})}\right\|_F}\end{aligned} \tag{15}$$

其中$\mathcal{R}(r,c)$定义为：

$$\mathcal{R}(r,c)=\{\Gamma_P,\Gamma_Q\in\mathbb{R}^{d\times m}\,|\,\Gamma_P\neq0,\Gamma_Q\neq0,|\,\mathcal{J}(P)|\leqslant r,|\,\mathcal{J}(Q^{\mathrm{T}})|\leqslant c,$$
$$\left\|(\Gamma_P)^{\mathcal{J}_\perp(P)}\right\|_{1,2}\leqslant\beta_1\left\|(\Gamma_P)^{\mathcal{J}(P)}\right\|_{1,2},\left\|(\Gamma_Q^{\mathrm{T}})^{\mathcal{J}_\perp(Q^{\mathrm{T}})}\right\|_{1,2}\leqslant\beta_2\left\|(\Gamma_Q^{\mathrm{T}})^{\mathcal{J}(Q^{\mathrm{T}})}\right\|_{1,2}\} \tag{16}$$

其中

$$\mathcal{J}(P)=\{i\,|\,\boldsymbol{p}^i\neq\boldsymbol{0}\},\quad \mathcal{J}_\perp(P)=\{i\,|\,\boldsymbol{p}^i=\boldsymbol{0}\} \tag{17}$$

分别表示 P 的非零行和零行的索引集合；$|\mathcal{J}|$表示集合$\mathcal{J}$中元素的个数。

我们要说明的是假设 1 是一个对受限特征假设[30]的一个推广，类似的假设请参考其他更多多任务学习的文献[21, 31]。基于上述的假设以及定理 2，我们有下列定理：

定理 3 令$(\hat{P},\hat{Q})$是问题(3)的一个最优解(其中$m\geqslant 2, n, d\geqslant 1$)，取正则化参数$\lambda_1$和$\lambda_2$使其满足式(12)，那么，在假设 1 以及定理 2 的条件下，下列的不等式以不小于$1-\exp\left(-\frac{1}{2}\left(t-dm\log\left(1+\frac{t}{dm}\right)\right)\right)(t>0)$的概率成立：

$$\begin{aligned}
&\frac{1}{mn}\left\|X\mathrm{vec}(\hat{P}+\hat{Q})-\mathrm{vec}(F^*)\right\|^2\leqslant\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)^2\\
&\left\|\hat{P}-P^*\right\|_{1,2}\leqslant\frac{(\beta_1+1)\sqrt{r}}{\kappa_1(r)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)\\
&\left\|\hat{Q}^T-Q^{*T}\right\|_{1,2}\leqslant\frac{(\beta_2+1)\sqrt{c}}{\kappa_2(c)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)
\end{aligned}\tag{18}$$

此外，如果下列的条件成立：

$$\begin{aligned}
&\min_{j\in\mathcal{J}(P^*)}\left\|(\boldsymbol{p}^*)^j\right\|>\frac{2(\beta_1+1)\sqrt{r}}{\kappa_1(r)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)\\
&\min_{j\in\mathcal{J}(Q^{*T})}\left\|(\boldsymbol{q}^*)_j\right\|>\frac{2(\beta_2+1)\sqrt{c}}{\kappa_2(c)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)
\end{aligned}\tag{19}$$

那么以相同的概率，下面的两个集合

$$\begin{aligned}
&\hat{\mathcal{J}}_1=\left\{j:\left\|\hat{\boldsymbol{p}}^j\right\|>\frac{(\beta_1+1)\sqrt{r}}{\kappa_1(r)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)\right\}\\
&\hat{\mathcal{J}}_2=\left\{j:\left\|\hat{\boldsymbol{q}}_j\right\|>\frac{(\beta_2+1)\sqrt{c}}{\kappa_2(c)}\left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)}+\frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)\right\}
\end{aligned}\tag{20}$$

分别表示出了P^*和Q^*的稀疏模式，即

$$\begin{aligned}
&\hat{\mathcal{J}}_1=\mathcal{J}(P^*)\\
&\hat{\mathcal{J}}_2=\mathcal{J}(Q^{*\mathrm{T}})
\end{aligned}\tag{21}$$

定理 3 为鲁棒多任务特征学习模型(rMTFL)提供了一个很重要的理论保证。具体来说，定理 3 给出了参数估计误差(鲁棒多任务特征学习模型获得的权重矩阵$\hat{P},\hat{Q}$与真实权重矩阵P^*,Q^*之间的误差)和预测误差(鲁棒多任务特征学习模型得到的预测值与真实值的误差)的界。另外，在假设真实的权重幅度大于噪声这个基础上，我们的理论分析指出，在一定条件下，我们的算法能够以大概率获得真实共享的特征和异常任务，即$\mathcal{J}(P^*)$和$\mathcal{J}(Q^{*\mathrm{T}})$。

下面在一个真实数据(MRI)上来验证我们提出的鲁棒多任务特征学习算法的有效性。MRI 数据来自 ANDI 数据库,它包含 675 个病人的 MRI 数据,这些数据用 FreeSurfer (www. loni. ucla. edu/ADNI/) 进行了预处理。MRI 数据的每个样本包含 306 个特征,这些特征被分成 5 个类别:皮质骨平均厚度,皮质骨厚度的标准差,皮质分割平均容量,皮质分割容量的标准差,白质分割平均容量以及表面积。数据的预测目标是分别来自不同时间段(M06,M12,M18,M24,M36 和 M48)的简易精神状态测试(mini mental state examination,MMSE)成绩。我们去掉了 MRI 质量不合格以及有缺失元素的数据,经过处理后,我们有 6 个任务(每个任务对应一个时间段),每个任务分别有 648,642,293,569,389 和 87 个样本。我们将鲁棒多任务特征学习(rMTFL)算法和 7 种具有代表性的多任务学习算法进行比较,它们是:岭回归多任务学习(ridge),ℓ_1 正则多任务学习(lasso),迹范数正则多任务学习(trace),$\ell_{1,2}$ 正则多任务学习(L1,2),脏模型多任务学习(DirtyMTL)[32],稀疏结构和低秩多任务学习(SLR)[33] 以及鲁棒多任务学习(RMTL)[31]。所有的 8 种算法都用二次损失函数。在我们的实验中,当连续两次迭代之间目标函数值的相对改变小于10^{-5}时,就停止算法。我们随机地将每个任务的样本以不同的比例划分成训练集和测试集。在训练集上学习到了模型参数后,用归一化的均方误差(nMSE)和平均均方误差(aMSE)[31,34,35]在测试集上评估 8 个多任务学习算法的性能。对于每一个训练比例,nMSE 和 aMSE 都是 10 次随机划分后计算得到的平均值,所有算法的参数都是通过 3 份交叉验证的方式来确定的。表 1 的结果表明,rMTFL 的性能在所有的算法中是最好的。

表 1 8 种多任务学习算法在 MRI 数据上的平均 nMSE 和平均 aMSE

度量	训练比例	ridge	lasso	trace	L1,2	DirtyMTL	SLR	RMTL	rMTFL
nMSE	0.15	0.9494	0.6469	0.6889	0.6445	0.6355	0.6905	0.6930	**0.5743**
	0.20	0.9355	0.6242	0.6629	0.6555	0.6231	0.6648	0.6557	**0.5700**
	0.25	0.9151	0.6015	0.6230	0.6446	0.6082	0.6244	0.6239	**0.5498**
aMSE	0.15	0.0270	0.0188	0.0195	0.0184	0.0177	0.0196	0.0196	**0.0168**
	0.20	0.0262	0.0177	0.0184	0.0184	0.0172	0.0185	0.0182	**0.0163**
	0.25	0.0255	0.0170	0.0171	0.0181	0.0167	0.0171	0.0171	**0.0157**

3 多阶段多任务特征学习

在第 2 节,我们通过检测异常任务的方法对“所有任务共享一组共同的特征”这个问题给出了详细的解决方案,在本节,我们通过另外一种方式来讨论这个问题。具体来说,

我们提出了一个截断-ℓ_1,ℓ_1 正则化的非凸多任务特征学习模型[36, 37],该模型使得一些特征可以被一些任务共享而不是被所有的任务共享。从直观上来看,由于截断-ℓ_1,ℓ_1 正则相对于基于 ℓ_1 范数类型的正则(凸稀疏正则)更加接近于 ℓ_0 类型的正则,且基于 ℓ_1 范数类型的正则会导致过惩罚的问题,我们提出的非凸正则模型应该能够得到比 ℓ_1 正则模型更好的性能。然而,非凸的模型使得求解相应的非凸优化问题变得困难,为了解决这一问题,我们提出了一个多阶段多任务特征学习(MSMTFL)算法来求解相应的非凸优化问题,同时为更好地理解提出的优化算法,我们还从两个不同的角度对提出的优化算法进行了直观的解释。最后我们对模型进行了详细的理论分析。虽然多阶段多任务特征学习算法获得的解可能不是全局最优解,但我们的分析指出,这个解能够保证有好的性能。具体地,我们对多阶段多任务特征学习算法所获得解的参数估计误差进行了分析,在稀疏特征值条件下(该条件比 Jalali 等人[32]中的一致性条件要弱),多阶段多任务特征学习算法得到的参数估计误差的界随着迭代的向前推进而不断变小,在迭代步数达到一定数量时,多阶段多任务特征学习算法得到的参数估计误差的界比基于 ℓ_1 范数类型的正则(凸稀疏正则)多任务特征学习算法要好。

我们介绍一下非凸多任务特征学习模型。假设有 m 个任务的数据$\{(X_1, \boldsymbol{y}_1), \cdots, (X_m, \boldsymbol{y}_m)\}$,其中 $X_i \in \mathbb{R}^{n_i \times d}$ 是第 i 个任务的数据矩(每行是一个样本);$\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ 是第 i 个任务的预测目标向量;d 是样本的维数;n_i 是第 i 个任务的样本个数。考虑学习一个权重矩阵 $W=[w_1, \cdots, w_m] \in \mathbb{R}^{d \times m}$,该权重矩阵包含有 m 个线性预测模型的权重向量:$\boldsymbol{y}_i \approx \boldsymbol{f}_i(X_i) = X_i w_i, i \in \mathbb{N}_m$。我们提出一个能够同时学习出 m 个任务的非凸多任务特征学习模型,该模型是一个基于截断-ℓ_1,ℓ_1 的非凸正则优化问题。具体地,我们首先给 W 的每一行施加一个 ℓ_1 范数惩罚,就能得到一个列向量,然后我们给这个列向量施加一个截断-ℓ_1 惩罚[38, 39],也即我们提出的非凸多任务特征学习模型可以写成下列的优化问题:

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ l(W) + \lambda \sum_{j=1}^{d} \min(\| w^j \|_1, \theta) \right\} \tag{22}$$

其中,$l(W)$是关于 W 的一个经验损失函数;$\lambda(>0)$是一个正则化参数;$\theta(>0)$是一个阈值参数;w^j表示矩阵 W 的第 j 行。这里,我们着重考虑损失函数是下列的二次函数的情况:

$$l(W) = \sum_{i=1}^{m} \frac{1}{m n_i} \| X_i w_i - \boldsymbol{y}_i \|^2 \tag{23}$$

直观来看,由于截断-ℓ_1,ℓ_1 惩罚的影响,问题(22)的最优解 W^* 包含有很多零行。同时,由于 W 的每一行施加了一个 ℓ_1 范数惩罚,那么对于非零行$(w^*)^k$,其中的一些分量

可以是零。这样,问题(22)的最优解 W^* 的某一行可能会出现有些元素是零,有些元素不是零的情况,即一些特征可以被一些任务共享而不是被所有的任务共享。

接着我们详细讨论如何求解相应的非凸优化问题。优化问题(22)是非凸的,这给求解优化问题带来了一定的困难。下面我们提出一个多阶段多任务特征学习(multi-stage multi-task feature learning,MSMTFL)算法来求解问题(22)(具体算法流程如算法 2 所示)。在这个算法中,一个关键的步骤是如何有效地求解问题(24):

算法 2 多阶段多任务特征学习(MSMTFL)

1. 初始化 $\lambda_j^{(0)}=\lambda$;
2. **For** $\ell=1,2,\cdots$ **do**
3. 令 $\hat{W}^{(\ell)}$ 是优化问题(24)的一个最优解;
4. 令 $\lambda_j^{(\ell)}=\lambda I(\|(\hat{w}^{(\ell)})^j\|_1<\theta)(j=1,\cdots,d)$,其中 $(\hat{w}^{(\ell)})^j$ 是 $\hat{W}^{(\ell)}$ 的第 j 行;$I(\cdot)$ 表示一个取值为 $\{0,1\}$ 的指示函数。
5. **End**

$$\min_{W\in\mathbf{R}^{d\times m}}\left\{l(W)+\sum_{j=1}^{d}\lambda_j^{(\ell-1)}\|w^j\|_1\right\} \tag{24}$$

我们注意到问题(24)的目标函数是一个可微的损失函数与一个不可微的正则项之和,所以我们可以用 FISTA 算法[28]来求解这个子问题。接下来,我们对提出的优化算法从不同角度进行一些直观的解释。

首先,我们从局部线性逼近的角度对算法进行直观的解释。为此,我们定义两个辅助函数:

$$\begin{aligned}&\boldsymbol{h}:\ \mathbb{R}^{d\times m}\mapsto\mathbb{R}_+^d,\quad \boldsymbol{h}(W)=[\|w^1\|_1,\cdots,\|w^d\|_1]^{\mathrm{T}}\\&g:\ \mathbb{R}_+^d\mapsto\mathbb{R}_+,\quad g(\boldsymbol{u})=\sum_{j=1}^{d}\min(u_j,\theta)\end{aligned} \tag{25}$$

利用上述定义的辅助函数,问题(22)可以写成下列等价的形式:

$$\min_{W\in\mathbb{R}^{d\times m}}\{l(W)+\lambda g(\boldsymbol{h}(W))\} \tag{26}$$

注意到 $g(\cdot)$ 是一个凹函数,我们说一个向量 $\boldsymbol{s}\in\mathbb{R}^d$ 是函数 g 在 $v\in\mathbb{R}_+^d$ 处的一个次梯度,如果对于所有的 $\boldsymbol{u}\in\mathbb{R}_+^d$,下列的不等式都成立(对于凸函数的次梯度定义,我们只需要在凹函数的前面加一个负号就能够得到):

$$g(\boldsymbol{u})\leqslant g(v)+\langle\boldsymbol{s},\boldsymbol{u}-v\rangle \tag{27}$$