多任务学习

张 钰 刘建伟 左 信
(中国石油大学(北京)自动化系 北京 102249)

摘 要 随着图像处理,语音识别等人工智能技术的发展,很多学习方法尤其是采用深度学习框架的方法取得了优异的性能,在精度和速度方面有了很大的提升,但随之带来的问题也很明显,这些学习方法如果要获得稳定的学习效果,往往需要使用数量庞大的标注数据进行充分训练,否则就会出现欠拟合的情况而导致学习性能的下降.因此,随着任务复杂程度和数据规模的增加,对人工标注数据的数量和质量也提出了更高的要求,造成了标注成本和难度的增大.同时,单一任务的独立学习往往忽略了来自其它任务的经验信息,致使训练冗余重复和学习资源的浪费,也限制了其性能的提升.为了缓解这些问题,属于迁移学习范畴的多任务学习方法逐渐引起了研究者的重视.与单任务学习只使用单个任务的样本信息不同,多任务学习假设不同任务数据分布之间存在一定的相似性,在此基础上通过共同训练和优化建立任务之间的联系.这种训练模式充分促进任务之间的信息交换并达到了相互学习的目的,尤其是在各自任务样本容量有限的条件下,各个任务可以从其它任务获得一定的启发,借助于学习过程中的信息迁移能间接利用其它任务的数据,从而缓解了对大量标注数据的依赖,也达到了提升各自任务学习性能的目的.在此背景之下,本文首先介绍了相关任务的概念,并按照功能的不同对相关任务的类型进行划分,之后对它们的特点进行了逐一描述.然后,本文按照数据的处理模式和任务关系的建模过程不同将当前的主流算法划分为两大类:结构化多任务学习算法和深度多任务学习算法.其中,结构化多任务学习算法采用线性模型,可以直接针对数据进行结构假设并且使用原有标注特征表述任务关系,同时,又可根据学习对象的不同将其细分为基于任务层面和基于特征层面两种不同结构,每种结构有判别式方法和生成式方法两种实现手段.与结构化多任务学习算法的建模过程不同,深度多任务学习算法利用经过多层特征抽象后的深层次信息进行任务关系描述,通过处理特定网络层中的参数达到信息共享的目的.紧接着,以两大类算法作为主线,本文详细分析了不同建模方法中对任务关系的结构假设、实现途径、各自的优缺点以及方法之间的联系.最后,本文总结了任务之间相似性及其紧密程度的判别依据,并且分析了多任务作用机制的有效性和内在成因,从归纳偏置和动态求解等角度阐述了多任务信息迁移的特点.

关键词 多任务学习;信息迁移;任务相似性;贝叶斯生成式模型多任务学习;判别式多任务学习;深度多任务学习
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.01340

Survey of Multi-Task Learning

ZHANG Yu LIU Jian-Wei ZUO Xin
(*Department of Automation, China University of Petroleum, Beijing* 102249)

Abstract With the development of artificial intelligence technology such as image processing and speech recognition, many learning methods, especially those using deep learning frameworks, have achieved excellent performance and greatly improved accuracy and speed, but the problems are also obvious, if these learning methods want to achieve a stable learning effect, they often need to use a large number of labeled data to train adequately. Otherwise, there will be an under-fitting situation which will lead to the decline of learning performance. Therefore, with the

收稿日期:2018-04-16;在线发布日期:2019-04-17.本课题得到国家重点研发计划项目(2016YFC0303703-03)、中国石油大学(北京)年度前瞻导向及培育项目(2462018QZDX02)资助.**张 钰**,博士研究生,主要研究方向为机器学习.E-mail:623242954@qq.com.**刘建伟**,博士,副研究员,主要研究方向为机器学习、模式识别与智能系统、复杂系统的分析、预测与控制、算法分析与设计.**左 信**,教授,主要从事油田生产、管道运输和石油化工过程的测量、控制与优化方面的理论研究和工业应用工作,目前主要研究领域为先进控制理论与应用、安全保护控制系统和深海油田自动化.

increase of task complexity and data scale, higher requirements are put forward for the quantity and quality of manual labeling data, resulting in the increase of labeling cost and difficulty. At the same time, the independent learning of single task often ignores the experience information from other tasks, which leads to redundant training and waste of learning resources, and also limits the improvement of its performance. In order to alleviate these problems, the multi-task learning method, which belongs to the category of transfer learning, has gradually attracted the attention of researchers. Unlike single-task learning, which only uses sample information of a single task, multi-task learning assumes that there is a certain similarity between the data distribution of different tasks. On this basis, the relationship between tasks is established through joint training and optimization. This training mode fully promotes information exchange between tasks and achieves the goal of mutual learning. Especially under the condition that the sample size of each task is limited, each task can get some inspiration from other tasks. With the help of information transfer in the learning process, the data of other tasks can be indirectly utilized. Thus, the dependence on a large number of labeled data is alleviated, and the goal of improving the performance of task learning is also achieved. Under this background, this paper first introduces the concept of related tasks, and describes their characteristics one by one after classifying the types of related tasks according to their functions. Then, according to the data processing mode and task relationship modeling process, the current mainstream algorithms are divided into two categories: structured multi-task learning algorithm and deep multi-task learning algorithm. The structured multi-task learning algorithm adopts linear model, which can directly assume the structure of the data and express the task relationship with the original annotation features. At the same time, it can be subdivided into two different structures based on task level and feature level according to the different learning objects. Each structure has two implementation means: discriminant method and generative method. Different from the modeling process of structured multi-task learning algorithm, deep multi-task learning algorithm uses the deep information abstracted by multi-layer features to describe the task relationship, and achieves the goal of information sharing by processing the parameters in the specific network layer. Then, taking two kinds of algorithms as the main line, this paper analyzed the structural assumptions, implementation approaches, advantages and disadvantages of different modeling methods and the relationship between them in detail. Finally, this paper summarizes the criteria for identifying the similarity and compactness between tasks, and the effectiveness and intrinsic causes of multi-task mechanism are also analyzed, then the characteristics of multi-task information migration are expounded from the perspectives of inductive bias and dynamic solution.

Keywords　multi-task learning; information transfer; similarity of tasks; Bayesian generative model of multi-task learning; discriminant approach of multi-task learning; deep multi-task learning via deep neural network

1　多任务学习背景

多任务学习(Multitask Learning, MTL)是同时考虑多个相关任务的学习过程,目的是利用任务间的内在关系来提高单个任务学习的泛化性能. 1994 年,Caruana 提出了多任务学习的概念[1]:多任务学习是一种归纳迁移机制,基本目标是提高泛化性能. 多任务学习通过同时训练多个相关任务,学习到任务之间的一些共享表示,并进一步地挖掘训练信号中的特定域信息来提高每个任务泛化能力,在数据挖掘、计算机视觉、语音识别、生物医疗、社交网络等领域有着广泛的现实应用.

多任务学习的出发点是在解决新问题时利用

的知识会习惯地受到已有相关问题的启发,借助于以往的经验可以提高学习效果.以往单任务学习(Single-task learning)是指每次只学习一个分类任务,并且只使用对应任务本身的数据集,在单任务学习中经常假定训练样本之间是独立并且同分布的.但是现实世界中存在很多相关的数据集,例如,在遥感图像处理问题中,每组图像数据都是在特定的地理位置收集的,此时数据源来自于同一设备不同视角,存在高度相关性,但是并不为每个感测任务设计单独的分类器,而是希望在任务之间共享数据,以提高整体感知性能[2-4].因此在很多情况下,虽然任务数据采集的来源和分布是相似的,即可能存在共同的归纳偏置,但是由于学习的目的不完全相同,不能简单地将它们合并为一个任务,此时可以将它们看作是由多个相关的任务组成,选择多个任务联合学习,从而获得一些潜在信息以提高各自任务的学习效果.

而且,对于一些训练样本个数少且特征维数高的任务,单任务学习出现秩亏并且有过拟合的风险,通过在一定结构的共享空间之内并行学习,当前任务可以接收其它辅助任务传入的特征信息,这样在交互过程中间接增加了单个任务样本空间的大小,另外多任务之间平均了各自的噪声差异,得到了更一般的表示模型,这样其它任务可以为相关的特征提供额外的参考信息,可以有效降低单个任务过拟合和泛化能力差的风险.

多任务学习虽然是一种迁移学习方法,但是不同于其它种类迁移学习,多任务学习并不注重源领域和未知领域的知识迁移,它主要利用域之间相似的知识信息,提升特定任务的学习效果,注重领域知识的共享性.两者特点的不同决定了学习过程的差别,迁移学习的目的是通过从源任务中转移知识来提升目标任务中的性能,而多任务学习则试图同时学习目标任务和源任务.从这个层面上来说,其它迁移学习方法应该侧重于归纳转移,而多任务学习侧重于共享.

1.1 多任务的类型分类

1.1.1 相关任务的分类

多任务学习有很多形式,联合学习(joint learning)、自主学习(learning to learn)和带有辅助任务的学习(learning with auxiliary task)等都可以称为多任务学习.

联合学习又称为对称多任务学习,试图同时执行所有任务以便提高单个任务的学习性能,由于时间、地点、设备、人工标注差异或其它变化因素,各个任务数据的统计特性可能有所不同,但是任务之间肯定存在高度相似,联合学习多个分类任务有助于减少任务之间概率分布差异,因此,在这个意义上,联合学习认为模型是对称的,不区分主任务和辅助任务.一般的多任务应用都是指对称多任务学习,出发点是将几个类似的学习任务同时进行训练,通过任务之间的特征信息迁移共同地提升所有任务的学习效率,本文中绝大部分算法也都是针对对称多任务学习,彼此没有主次之分.

自主学习又称为非对称多任务学习,目标是利用源任务的信息来改进某些目标任务的学习性能,通常在源任务被学习后使用[5],和迁移学习不同的是,自主学习仍是建立在共同学习基础上,并不强调源域和目标域分布的差异性,如果分布相似性条件不成立,使用非对称多任务学习是不合理的,只能考虑迁移学习.

针对辅助任务的用处不同,又可以将多任务学习划分为输入变输出逆多任务学习,对抗性多任务学习,辅助任务提供注意力特征的多任务学习和附加预测性辅助任务的多任务学习.

1.1.2 将输入变输出的逆多任务学习

一般在有监督的学习任务中,输入和输出之间有很明显的区分,观测值是输入,要预测的值是输出,也称为监督信号.有监督的学习模式中,任务利用训练集的样例学习相关特征表示用于测试集的预测,而在无监督学习中并没有监督信号,因此解决的办法是将特征同时作为输入和输出,利用不同的无监督样例的特征信息为彼此提供监督信号,此时允许测试样例的特征作为输入,将其它样例上的特征作为输出,进而达到不同的学习效果,最大地发挥这些特征的作用.

借鉴无监督学习的模式,文献[6]提出在有监督的多任务学习中如果存在比作为输入更有价值的特征时,可以使用其它任务上的样例的特征作为监督信号,学习目标任务训练集上其它的输入特征到这部分特征的映射关系,学习映射关系的过程可以作为辅助任务.这些特征可以提供监督信号的根本原因是当特征中存在噪声时,附加的辅助输出中的噪声往往比附加的辅助输入中的噪声小.

1.1.3 对抗性多任务学习

对抗性多任务学习借鉴了生成对抗网络中的观点,该观点最初是由 Ganin 等人[7-8]在解决无监督域适应学习问题时提出的,文献[9]将它用在了有监督的 RNN 多任务学习中.受到生成对抗网络中生成式模型的启发,引入对抗任务是为了加强学习任务间不变性表示的能力.在标准的多任务学习中为了

最大限度地提高主要任务和次要任务的分类精度，学习表示的过程是共享的. 不同于标准的多任务学习，对抗性多任务学习要得到的是对主要任务有利而与次要任务对抗的表示. 文献[9]提出的多任务学习框架不断利用辅助任务包含的相反信息，消除主要任务的噪声，从而学习到接近底层数据真实表示的特征.

具有两个任务的对抗性网络如图 1 所示. 它由三个子网络组成，包括主要任务输出子网络，次要任务输出子网络以及主次任务共同的输入网络，其中两个子网络是独立的，输入层提取任务间共享表示. 次要任务在反向传播过程中经过一个反转层将梯度方向反转，弱化了对抗任务的分类精度，目的是学习次要任务的对抗表示，所以对主任务有损害且无关的域依赖信息将从表示中清除. 通过共同学习一个辅助任务域和主要任务域的不变表示，对抗性任务可以发现目标任务最本质的特征，训练完成后将对抗性任务的输出子网删除，最终经过对抗学习过程之后，主任务学习的是去除无关信息的特征表示，鲁棒性增加，进而学习效果得到提升. 更一般的，如果学习分类任务，各个任务彼此之间也可以看作是一种特殊的对抗任务，因为每个任务中含有直接影响

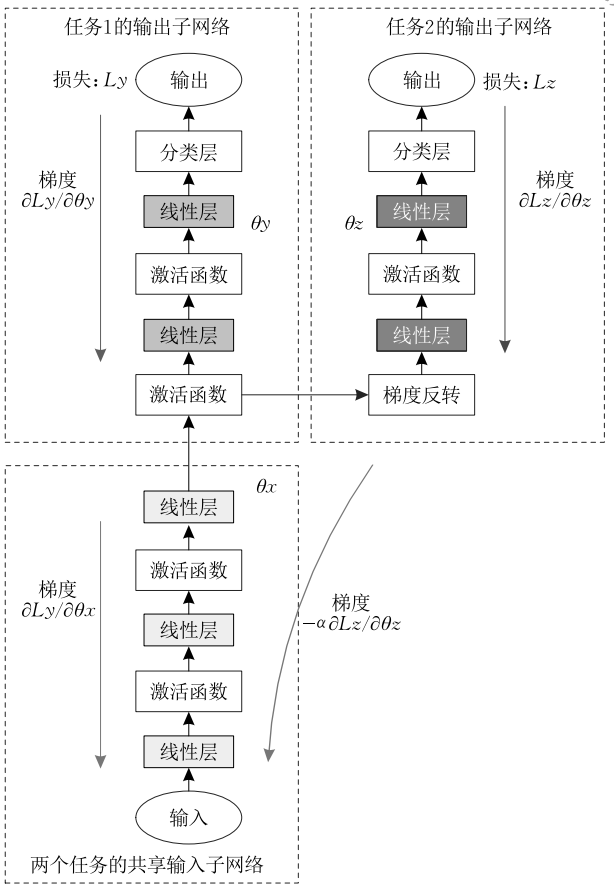


图 1　对抗性任务

分类结果的判别特征，需要最大程度地区别出各个分类的差异性.

1.1.4　辅助任务提供注意力特征的多任务学习

这里的辅助任务可以称为提示性任务，是一种为监督学习增加信息的方式. 文献[6]首先提出了这个概念，作者列举使用图像学习转向任务中检测路标的问题，转向问题本身不着重于路标检测，因为路标图像只占整个图像很小的一部分，但是自动驾驶问题中路标的检测是必要的，如果采用单任务学习，难以让这部分特征作为有效的输入，于是引入这些路标图像检测作为辅助任务，通过与转向任务学习相同的共享结构，在共享结构中添加上关于路标的特征. 此后，文献[10]将提示性任务用于带有生僻字检测的命名体识别系统，利用上下文提示信息作为辅助任务进行特征放大，为生僻词提供线索，避免了一些词语的词性标签对于词句的语义标注和外部词汇表的依赖性.

可以看到，在单任务学习过程中，一些显著特征对学习结果的影响较大，一些不常用特征往往被忽略，但是这类特征对于任务的某些功能是必要的，在一般的多任务学习中，这部分不常用的特征可以通过辅助任务单独引入，在共同学习过程中将其放大，平衡显著特征带来的学习不充分问题. 此类在目标任务中需要单独放大的特征一般称为需要注意力集中的特征.

1.1.5　附加预测性辅助任务的多任务学习

在线学习问题中包含很多对于主任务学习有价值的特征，但是有些特征不能作为输入，因为这些特征在学习过程中不能够被及时得到，而是在训练以及预测完成之后才能获得，例如在文献[6]中列举的自动驾驶问题中的路标检测，由于技术原因，在行驶过程中无法及时获得前方道路交通标示，只有当接近它的时候才能准确的测量，但是提前辨识路标对于适应路况必不可少，此时没有这部分特征的定性描述，也没有这部分特征的任何信息可以获取，这种情况下如果要识别这些特征，只能将这部分特征通过离线学习收集并添加到训练集作为实例样本. 与 1.1.2 节输入变输出的特征正好相反，这部分特征可以称作输出变输入的特征.

因此，当出现一些与主任务相关的未知特征时，往往可以将这些特征的学习作为辅助任务，在离线过程中收集，而在在线过程中为主任务提供额外的信息，帮助主任务学习更合理的归纳表示，这类辅助任务就称为预测性任务，这些额外的任务产生的未来特征测量值，可以应用于很多离线问题.

本节介绍了几种多任务的类型和非对称任务中的几种典型辅助任务，可以看到，引入辅助任务主要为目标任务增加有效的特征信息，为提高数据的使用效率提供了启发，一般旨在提升目标任务中不容易被使用特征的利用效率. 辅助任务与目标任务的特征相似度可以是局部的，只考虑主要特征的需求，例如对抗性任务只利用浅层的鲁棒不变表示；也可以是全局的，例如离线搜集数据当作辅助预测任务引导自动驾驶，当然，对于带有辅助任务的多任务学习，数据要进行预处理，转换到同一个特征空间下讨论，因为数据的来源往往都不相同. 本文中只有 2.7 节非监督多任务学习中涉及到了非对称任务，其余情形还是着重讨论对称多任务学习，因为实际应用中对称多任务学习的适用场景较为广泛，也更多的应用到了平均提升每个任务泛化性能的思想. 由于每个任务中特征的权重占比不同，如何选取合适的算法以高效利用数据找到适合于多任务关系的结构是一个难点，我们将在下一节多任务学习具体算法中介绍.

2 多任务学习算法

2.1 多任务学习的定义

在多任务学习中，给定 M 个任务的训练集 $\{T_m\}_{m=1}^{M}$，对于第 m 个任务 T_m，训练集 D_m 包含 n_m 个样例-标签对 $\{x_{m,j}, y_{m,j}\}_{j=1}^{n_m}$，$x_{m,j} \in R^D$ 为第 m 个任务的第 j 个样例，$y_{m,j} \in R$ 代表其对应的输出，n_m 是第 m 个任务的训练样本的个数，$\boldsymbol{W} \in R^{D \times M}$ 代表权值矩阵，即多任务模型参数矩阵，ε_m 代表任务下的噪声，则有线性模型：

$$y_{m,j} = \boldsymbol{w}_m^{\mathrm{T}} x_{m,j} + \varepsilon_m \tag{1}$$

其中第 m 个任务的模型向量 $\boldsymbol{w}_m$ 为 $\boldsymbol{W}$ 中的一列，大多数现有 MTL 算法的一个关键假设是，所有任务都通过某种结构相互关联，多任务中任务信息共享是通过特征的联系实现的，一般来说，多任务选取的特征属性都是相似的，而各个任务之间特征的重要性通过模型向量 $\boldsymbol{w}_m$ 反映，如果在模型向量中所占比重相似才能说明任务特征之间具有迁移性，因此多任务学习的目的是通过学习 $\boldsymbol{W}$ 的不同结构来反映任务之间的关系. 为了概念的统一，本文的多任务模型参数指的是模型参数矩阵.

2.2 多任务算法的分类

MTL 算法按照学习模式的不同，可以分为传统的结构化学习方法和深度多任务学习方法. 传统的结构化学习方法并不会像深度学习方法一样改变特征的表现形式，即不利用抽象后的特征，最终以结构约束的形式体现任务联系. 按照学习结构的不同，又可以分为基于任务层面的多任务学习和基于特征层面的多任务学习，其中基于任务层面的方法通常将大部分特征视为彼此相关的，并且任务的相关性是全局的，因此它注重总体特征的共享迁移，一般是同时考虑多个特征，而基于特征层面的学习方法是单独对各个任务中的特征进行建模，注重个体特征的共享迁移. 结构化学习方法从任务层面上可以分为模型参数共享方法，公共特征共享方法，多任务聚类结构方法和多任务子空间学习方法；从特征层面上可以分为鲁棒特征学习方法，联合低秩稀疏方法，脏模型方法，可变簇聚类方法，协同聚类方法. 并且，在结构化学习方法中，均可以采用基于块正则化的判别式方法和基于贝叶斯概率统计的生成式方法作为不同的实现手段.

有别于结构化学习方法，深度多任务方法对各个任务的特征逐层进行建模，任务关系的表述是通过改变层与层之间的连接方式. 深度多任务方法主要有硬参数共享、软参数共享、张量网络、自适应层连接和自适应层分堆. 在图 2 中，我们根据以上描述的学习模式对多任务学习进行了分类.

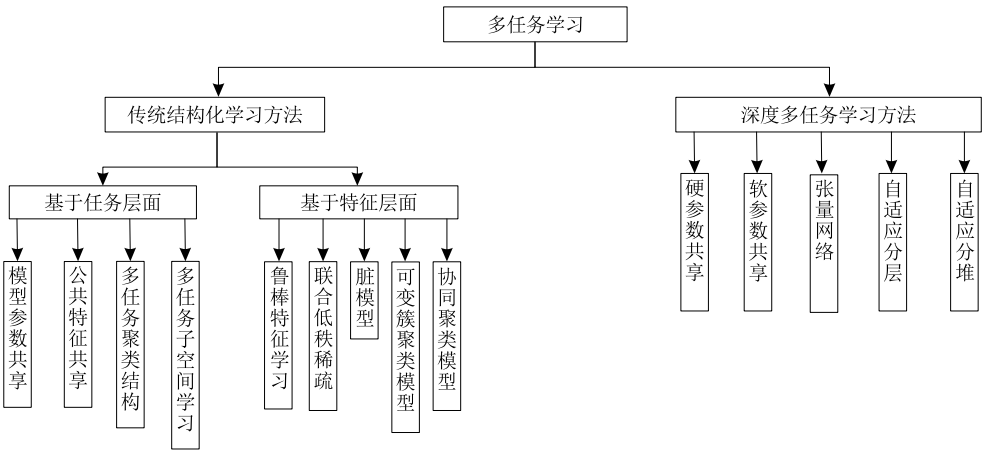


图 2 多任务学习分类图

2.3　基于任务层面的学习方法

2.3.1　模型参数共享

在早期阶段，许多多任务方法侧重于利用任务关系的先验信息，例如，基于所有任务的域知识彼此相似，Caruana 在文献[4]中提出了多层前馈神经网络，是最早的多任务统一共享结构学习模型之一. 在神经网络中，隐含层代表来自所有任务的共同特征，输出层中的每个神经元通常对应于一个任务的输出. 与神经网络相似，多任务支持向量机方法[11]是文献[4]的自然扩展，它假定存在一个通用的多任务模型，权衡每个任务到此通用模型的中心偏离度和模型参数平均值，使得总体损失函数达到最优. 每个支持向量机要学习的模型定义为 $W_t=W_0+V_t$，W_0 是任务集隐含的共同模型，V_t 表示各个任务和中心模型的距离，通过 L_2 范数正则化约束中心模型参数 W_0 和各自任务模型参数 V_t 的平均距离，使所有任务的模型参数接近此共同模型. 除了支持向量机，参数共享也可以通过鼓励所有任务参数相似实现，文献[12]根据给定的任务相似性图模型设计出一些基于拉普拉斯的正则化算子；文献[11]采用最小化参数差异的 Frobenius 范数；文献[13-14]将其转换到希尔伯特空间中讨论，其结果都保证了任务之间的总体相似度，因此，这类方法也可以称为平均约束学习. 总体来说，模型参数共享的出发点比较简单和直接，建立在任务相似性比较大的基础上，基本都是假设任务相互关联，学习目标是获得一个中心模型来描述这些任务的公共特征集，所以可以看到大多数约束都是使得各个任务模型足够靠近这个模型均值，而现实情况中大部分多任务之间很难满足非常紧密的相关性，于是出现了只挑选一些主要特征的公共特征共享方法.

2.3.2　公共特征共享

在公共特征共享的方法中，假设各种任务共享相同的稀疏模式，表现在参数矩阵上就是一些特征行整体被诱导为零，作用是强制所有任务共享一组公共特征来建模任务之间的相关性. 这类模型将联合损失定义为各任务损失和正则化项之和：

$$\min_{\boldsymbol{W}}\sum_{m=1}^{M}\left\|\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}^m-\boldsymbol{Y}^m\right\|_2^2+\lambda_1\boldsymbol{\Omega}(\boldsymbol{W}) \tag{2}$$

其中，λ_1 是权重参数，$\boldsymbol{\Omega}(\boldsymbol{W})$ 是正则化项，一般称为组套索(group-LASSO)结构，$\boldsymbol{\Omega}(\boldsymbol{W})$ 可以对比单任务回归中施加的正则化项，比如套索(LASSO)模型中的 L_1 范数和岭回归(Ridge)模型中的 L_2 范数，由于多任务的求解目标是参数矩阵，所以经常使用组合范数正则化进行约束，即可以防止过拟合又能满足通过稀疏挑选共享特征的要求，参数矩阵 $\boldsymbol{W}$ 的 (r,p) 范数定义为：$\|\boldsymbol{W}\|_{r,p}:=\left(\sum_{m=1}^{M}\|\boldsymbol{W}_m\|_r^p\right)^{\frac{1}{p}}$，常见的正则化是使用 $L_{2,1}$ 组合范数，通过统一计算参数矩阵 $\boldsymbol{W}$ 所有任务的第 d 行特征的 L_2 范数保证任务在这个特征属性上的相似度得到 $b(\boldsymbol{W})=(\|W_{1,:}\|_2,\cdots,\|W_{d,:}\|_2)$，然后再对 $b(\boldsymbol{W})$ 施加 L_1 范数，在一些特征上达到稀疏的效果和筛选的效果. $L_{2,1}$ 是多任务用来特征共享最经典而且使用最多的组合范数.

文献[15]直接对模型参数施加 $L_{2,1}$ 范数正则化约束的，$L_{2,1}$ 范数本身是不光滑的凸函数，所以在式(2)中，当损失函数选择凸函数(例如平方损失函数)时，整个目标函数就是可以找到全局最优解的凸函数. 以往优化通常使用的二阶锥规划和内点法，计算代价高且收敛慢，针对这个问题，文献[15]使用块提升算法进行优化，但是无法确定步长和收敛速度，文献[16-17]使用近端梯度加速法和一阶 nestrov 加速最优化黑箱方法，进一步确定了在 $L_{2,1}$ 范数优化问题中参数的学习步长，对迭代过程收敛速度进行了加速. 更近一步地，文献[18-19]考虑了组合范数的一般形式，文献[18]研究了组合范数的基数 L_0 稀疏，文献[19]详尽地介绍了混合范数 $L_{p,1}$，$p>1$ 的用法，将 $L_{p,1}$ 范数进行欧几里得投影并通过加速梯度算法求解优化，文献[20]提出了一个关于 $L_{p,1}$ 范数的统一解决方案，优化中结合了投影梯度法和拉格朗日乘子法，通过区间二分法找到了保证收敛的拉格朗日参数，使 $L_{p,1}$ 范数可以拓展到大规模数量的任务，并且作者通过理论分析证明，p 值选取的越大，使用组套索提取后的特征之间耦合程度越高. 因此，文献[21-25]使用了 $L_{\infty,1}$ 范数以后，保证了筛选特征之间的聚合程度更加紧密. 此外，受弹性网络的启发，文献[26-27]提出了一种包括平方范数正则化器的校准多任务特征学习公式，文献[27]对平方损失函数引入了校准砝码：$1/(\sigma_m\sqrt{n_m})$，其中，n_m 是第 m 个任务的样本容量，$\sigma_m=\dfrac{1}{\sqrt{n_m}}\|\boldsymbol{w}_m^{\mathrm{T}}x^m-y^m\|$ 是校准权重，不仅如此，文献[26]在引入校准砝码的基础上又在正则化项中加入了 $\|\cdot\|_F$ 范数来保证一个平滑的能对偶求解的问题. 文献[28]同时考虑了 $L_{\infty,1}$ 和 $L_{2,1}$ 两种范数，将组套索拓展到了既有分类任务又有回归任务异质任务集合中.

以上的多任务特征选择算法都假设任务之间具有正相关性，还有一类文献[29]考虑了另外一种情

况，即高度相关的输出子集可以共享一组公共的相关输入，而弱相关的输出不太可能受到相同输入的影响，比如考虑多个分类任务时，为了分类效果明显，每个分类任务之间有负相关性，因此显著特征也应该是不同的，作者通过引入负相关特征的假设，使用了正则化项 $\boldsymbol{\Omega}(\boldsymbol{W}):\sum_{d=1}^{D}\left(\sum_{m=1}^{M}|W_{d,m}|\right)^2$，其中 L_1 范数组合了任务之间的同一维特征，由于 L_1 范数的稀疏诱导作用，增强了同一维特征之间的竞争，这样任务之间同一维特征就出现了权重的分化，然后在此基础上再使用 L_2 范数对所有任务不同维度特征进行合并. 对比 $L_{2,1}$ 范数先使用 L_2 范数突出某一维特征再使用 L_1 范数诱导稀疏可知，文献[29]提出的正则化项是 $L_{2,1}$ 范数的反向过程.

上述方法是通过显式的正则化项挑选所有任务的共同特征来辨别任务的相关性，与上述方法不同，文献[30-31]先对原始特征进行线性变换，将多任务模型矩阵 $\boldsymbol{W}$ 分解为 $\boldsymbol{W}=\boldsymbol{U}A$，再通过对低维表示 A 施加组稀疏正则化来控制特征数目，进而使得 $\boldsymbol{W}$ 稀疏化挑选出跨任务的一些公共特征子集，这些特征由再生核希尔伯特空间中的正交基函数表示. 文献[32-33]进一步研究了一种通用的多任务核方法，用于学习非线性相关特征.

特征选择可以达到数据压缩的效果加速多任务的学习，并且有效地防止了在各个任务上可能存在的过拟合. 但是这种方法最大的弊端就是将多任务的关系描述得过于简单，只挑选出来了一些主要特征，排除了一些在整体任务框架之外的特征. 然而，由于特征具有依赖于任务特点的特异性，这部分特征在各自任务上的影响不应被消除.

2.3.3 多任务聚类结构

上述多任务学习模型都建立在假设单个任务的特征大范围相似的基础上，即所有任务以某种形式彼此相关. 虽然某些数据分布高度相关的子任务之间可以使用统一共享结构进行学习，但是更一般的情况是数据的分布具有更复杂的形式，如果只考虑特定的特征，目标任务为了适应其它任务会对自身的特征选择进行简化，容易造成学习不足. 同时，任务之间联系紧密程度也是不同的，即使集合中的所有任务都是相关的，某些任务之间紧密程度也可能比其它的更强，任务集上可能存在更相关的任务子集，这种想法促进了多任务聚类关系模型的产生[2,34-39]，聚类模型的主要思想是假设任务之间形成多个聚类簇，允许在一定数量的任务之间共享聚类子结构，并且只在组内学习共享结构，虽然多任务聚类模型也属于任务层面的多任务学习方法，但是它是参数共享和公共特征共享的特殊情况，只保证一个聚类簇内的任务之间的特征或模型参数是高度相似的.

文献[2]首先提出了多任务分类的概念，聚类方法首先确定几种最相关的任务类，对于一个特定类，需要通过推断任务关系，将不相关的任务从类中排除，因为不相关任务会带来干扰信息误导其它任务的学习，因此任务的学习只选择性地利用该类内相关任务的信息. 同时，作者指出，类层次结构在多任务空间被发现后，只需要判断与各个类中心的紧密程度，这不仅会提升任务集中的学习效果，还会增强对新任务的适应性，只需要判断与各个类中心的紧密程度，避免了重新训练整个任务集的繁琐过程. 判断类中任务关系的相关性是通过度量任务特征向量的距离，即用 K 近邻法对特征向量进行聚类，使得聚类损失在整个任务集上最小. 此后文献[38-40]提出了一种无先验类信息的方法，降低了对于聚类中心点初始化的要求，进一步提高了聚类收敛的稳定性和精度. 还有一种情况是任务分类和特征选择同时进行，但是这种情况下不同的聚类族中选择的特征子集往往有重叠(例如图 3 右第三个特征)，文献[41]扩展了组套索的结构，结合先验知识可以直接应用于聚类场景的特征选择，然而，由于在处理特征重叠问题时施加的正则化会使得大类和小类中特征选择的惩罚力度不平衡，针对这个问题，文献[42]使用了一种带权重的组套索：$\boldsymbol{\Omega}(\boldsymbol{W})=\sum_{g\in G}\left(\sum_{d\in g}(l_d^g)^2|W_d|^2\right)^{\frac{1}{2}}$，其中，对每个聚类组的权重进行了归一化，$l_d^g$ 是归一化之后的权重，如果第 d 个特征在第 g 个聚类族中，$l_d^g>0$，否则 $l_d^g=0$. 文献[43]提出了一种结构树形式的组稀疏，作者考虑了一种复杂的情况，当特征为具有多级聚类的树结构时，任务特征选择大部分相似但在某些维度有不关联的离散特征，这样离散特征位于单独的叶子节点，文中使用的组套索类似

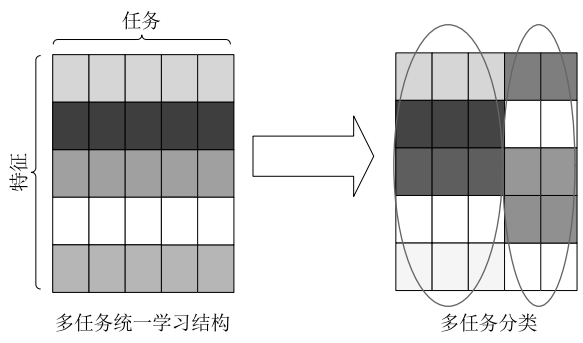


图 3 统一框架和分类结构

于文献[42]：$\boldsymbol{\Omega}(\boldsymbol{W})=\sum_{d\in D}\sum_{v\in V}l_v\|W^d_{G_v}\|_2$，$W^d_{G_v}$表明参数矩阵的第$d$个特征是在某个叶子节点也就是某个类中，并且权重$l_v$表明与此类关联的紧密程度.

更近一步，多任务的聚类方法不仅体现在直接比较模型参数的类似，还可以通过检测经过特征表示过后彼此之间的相似程度，而形成不同的任务类. 文献[34,44-46]提出了一种先映射到子空间，之后利用提取的主特征进行任务聚类的算法. 这几种方法出发点是一样的，即尽可能直接地根据特征向量的特点将任务集划分为最大相关的几个类.

任务聚类算法考虑了类中任务之间关联紧密程度，算法设计的关键是能根据特征相似性识别任务的类别，只允许组内任务共享信息和数据之间存在相似性，但是在不同的任务组之间无法彼此学习，在组间没有信息交换，这也是该方法的局限性.

2.3.4　多任务子空间学习

通过一些矩阵分解技术，例如非负矩阵分解，字典学习和稀疏编码，鲁棒主成分分析等方法，可以将原有的向量特征空间投影到矩阵空间中. 多任务学习的目标恰好满足在参数矩阵空间讨论的条件假设，并且它们之间满足相关性，证明多任务关系可以嵌入到低维子空间进行讨论，因此基于矩阵稀疏和低秩子空间的一些性质，一些算法开始在新的矩阵空间中研究多任务之间的关系.

子空间方法的假设是模型参数矩阵以线性低维特征映射的形式共享一个低秩特征子空间，这类模型可以表达为

$$\min_{\boldsymbol{W}}\sum_{m=1}^{M}L(\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}^m,\boldsymbol{Y}^m)+\lambda_1 rank(\boldsymbol{W}) \tag{3}$$

其中第一项$L(\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}^m,\boldsymbol{Y}^m)$为损失函数，第二项为秩函数. 早期工作[47-48]使用在非凸公式下的公共特征表示作为共享子空间对原有任务进行编码，通过秩函数$rank(\boldsymbol{W})$获得低秩子空间会产生优化问题，类似的，文献[49-50]使用最大限度地减少所有奇异值总和的迹范数正则化来学习低秩参数矩阵：

$$\min_{\boldsymbol{W}}\sum_{m=1}^{M}L(\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}^m,\boldsymbol{Y}^m)+\lambda_1\|\boldsymbol{W}\|_* \tag{4}$$

转换为式(4)后目标函数变为可求解的凸函数. 在此基础上，文献[51]不惩罚所有的奇异值，而使用一种带上限的迹范数正则化器，仅最小化小于某个阈值的奇异值，完整的恢复了底层低秩结构，进而扩展了迹范数低秩学习方法的适用范围，文献[52]通过凸松弛将迹范数的优化问题转换为近似凸函数，使其更容易求解.

文献[53]利用子空间的相似度，提出了基于低秩子空间的度量学习，除了学习适当的度量之外，该模型还直接在投影到子空间的低秩映射矩阵上进行优化，从而能有效地抑制噪声和防止过拟合. 在使用稀疏编码和字典学习背景下，文献[54]证明如果存在足够多的任务，在高位或无限维空间上可以通过字典原子的稀疏线性组合很好地近似任务参数.

同时，受到子空间方法的启发，类似于子空间的方法也得到了研究. 文献[55]提出了基于低秩属性嵌入的多任务学习，作者利用每对特征之间的相关关系将原始的二进制属性特征映射到连续的空间中，借助于其它任务的特征纠正恢复丢失的信息. 文献[56]提出了一种灵活的多任务聚类学习方法，在这种方法中，任务的子集被称为代表性任务，因此和子空间方法出发点类似，通过少量的一些代表性任务就可以编码和描述所有的任务. 不过，代表性任务的识别不同于通过投影学习到子空间方法，作者通过直接约束每个任务与代表性任务的距离，并且，使用不同的权重决定每个代表性任务贡献共享信息的程度：$\sum_{k=1}^{K}\sum_{m=1}^{M}Z_{mk}\|\boldsymbol{w}_m-\boldsymbol{w}_k\|_2^2$，可以看到不同于之前的聚类方法，通过最小化参数向量$\boldsymbol{w}_m$与多个代表任务$\boldsymbol{w}_k$的加权欧式距离，每个任务获得了不同程度的共享信息用于更精确地表示，此时，每个代表任务都被考虑作为一个基本类，当所有任务都选择同一个代表任务时，模型就退化为文献[11]所采用的方法. 此外，文献[57]使用核函数方法对基于迹范数的低秩方法进行非线性与线性扩展，使用非线性子空间表示任务的全局结构.

本节主要讨论了四种基于任务层面的学习方法，适用于相似度较大的任务集，侧重整体结构共享. 前两种方法假定了一个共享特征空间，参数共享方法的基准是在任务上寻找一个中心模型而公共特征共享方法的基准是找到一个经过特征选择后的共同结构，类似地，聚类结构在任务集上假定了几个特征共享空间，寻找几个类的中心模型，每个任务在学习过程中都可以参照其它任务彼此约束，达到特征选择和正则化的目的. 前三种方法都是简单适用范围有限的学习框架，过多地排除了共享结构之外的特征，导致了信息的丢失，而子空间方法更近一步，假设所有任务的区分信息保留在低维公共子空间中，开始考虑保留各自任务的区别信息，不必在多个任务的相同范围之内选择共享特征，或者在某些特

征上以同等重要性来共享信息，是一种隐式共享特征的方法. 同时，可以看到在矩阵分解方法中有一种不同于参数矩阵乘分解的鲁棒主成分分析方法，它考虑了通过拆分参数矩阵成叠加结构，正是基于这个思路，衍生出了从特征层面开始研究任务关系的方法.

2.4 基于特征层面的学习方法

基于任务层面的学习方法都是建立在所有任务或组内任务具有相关性的假设基础上，出发点是判断大部分特征是否具有总体相似度. 有时候任务关联性较强时，粗略共享是有效的，但是实际在很多任务中往往存在一些具有任务特异性的特征，这些区分度比较高的特征不能用简单的方式在任务间共享. 如果在包含这些特征的任务上强制共享这些特异特征可能会出现以下两种情况：在任务分类的时候由于其它特征的相似性使得这些特异特征被划分到不同的任务类中；或者在约束任务相关性的过程中被消除，其结果是导致负迁移和学习效果的下降. 一般来说，完善的多任务学习应该考虑任务层面和特征层面这两部分信息：在任务层面能够学习跨越任务的共享特征；在特征层面能够捕获共享结构之外的额外特异特征，即单独考虑只与各自任务相关的一些特征.

这类模型可定义为多任务间特征学习，一般在目标函数中增加多个正则化项，在此约束下学习模型参数，此时目标函数可以表示为

$$\min_{\boldsymbol{W},\boldsymbol{P},\boldsymbol{Q}} \sum_{m=1}^{M} \left\| \boldsymbol{W}_m^{\mathrm{T}} \boldsymbol{X}^m - \boldsymbol{Y}^m \right\|_2^2 + \lambda_1 \boldsymbol{\Omega}(\boldsymbol{P}) + \lambda_2 \boldsymbol{\Omega}(\boldsymbol{Q}) \quad (5)$$

$$\text{s.t. } \boldsymbol{W} = \boldsymbol{P} + \boldsymbol{Q}$$

其中，$\boldsymbol{P}$ 代表任务稀疏矩阵，挑选任务之间可以共享的特征，$\boldsymbol{Q}$ 代表特征稀疏矩阵，挑选各自任务上不能共享的特定特征，$\boldsymbol{\Omega}(\boldsymbol{P})$ 和 $\boldsymbol{\Omega}(\boldsymbol{Q})$ 为根据不同的要求可以选择的结构约束正则化项，λ_1，λ_2 是控制各项多任务特征结构权重关系的超参数. 值得注意的是，特征层面的稀疏特征和任务层面的稀疏特征行有本质不同，特征层面的稀疏代表单个特征或特征行的稀疏，没有固定结构；而任务层面的稀疏又称为组稀疏或联合稀疏，结果是不同特征行之间的稀疏（如图 3 左所示）. 下文中，在关于特征层面所述方法中的稀疏矩阵都是考虑单个特征的稀疏.

2.4.1 鲁棒特征学习

早期特征学习方法中，结构化鲁棒多任务特征学习[58]（Robust multitask feature learning Robust multitask feature learning）同时获取相关任务之间的一组共享特征矩阵 $\boldsymbol{P}$ 和识别特异特征任务矩阵 $\boldsymbol{Q}$，特异特征任务矩阵不与其它任务共同学习，因为大部分特征与其它任务都不相关. 这种方法可以看作是任务聚类结构的特殊形式，优点是离群任务或离群特征驻留在不影响其它任务的独立聚类簇中，不会因为离群任务包含的不相关信息影响其它任务，有利于提升整体多任务学习的鲁棒性. 在文章中，作者分别在 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 上使用了如下两种范数：

$$\begin{aligned} \boldsymbol{\Omega}(\boldsymbol{P}) &= \|\boldsymbol{P}\|_{1,2} \\ \boldsymbol{\Omega}(\boldsymbol{Q}) &= \|\boldsymbol{Q}^{\mathrm{T}}\|_{1,2} \end{aligned} \quad (6)$$

可以看到对 $\boldsymbol{P}$ 的约束是 $L_{2,1}$ 范数，目的就是剔除无关的离群特征并形成任务共享结构，与 $\boldsymbol{P}$ 类似，施加 $L_{2,1}$ 范数在转置的 $\boldsymbol{Q}$ 矩阵上，此时就可以孤立出离群的任务，正则化项 $\boldsymbol{\Omega}(\boldsymbol{P})$ 和 $\boldsymbol{\Omega}(\boldsymbol{Q})$ 的叠加是在特征和任务上的双向作用，总能筛选出离群变量，但是事实上，多任务问题不只有离群任务和离群特征，一些特征是以稀疏的形式存在的.

2.4.2 联合低秩项和稀疏项

文献[59-60]从多个相关任务中学习低秩子空间结构和非相关稀疏模式，其中使用低秩子空间捕捉所有任务之间的底层潜在相似性，使用稀疏判别特征表示任务之间的差异. 文献[59]提出了这类模型目标函数的一般形式为

$$\min_{\boldsymbol{W},\boldsymbol{P},\boldsymbol{Q}} \sum_{m=1}^{M} \left| L(\boldsymbol{W}_m^{\mathrm{T}} \boldsymbol{X}^m, \boldsymbol{Y}^m) + \lambda_1 \boldsymbol{\Omega}(\boldsymbol{P}) + \lambda_2 \boldsymbol{\Omega}(\boldsymbol{Q}) \right. \quad (7)$$

$$\text{s.t. } \boldsymbol{W} = \boldsymbol{P} + \boldsymbol{Q},\ rank(\boldsymbol{Q}) \leq \tau,\ \boldsymbol{\Omega}(\boldsymbol{P}) = \|\boldsymbol{P}\|_0$$

其中 $L(\cdot)$ 代表平方损失函数，是光滑且凸的，一般使用基数正则化 $\boldsymbol{\Omega}(\boldsymbol{P}) = \|\boldsymbol{P}\|_0$ 作为任务的非相关稀疏项的诱导范数，参数 τ 定义了低秩项 $\boldsymbol{Q}$ 的上界，由于秩函数和基数正则化项非凸且没有办法直接求解且 l_1 范数是 l_0 范数的凸包络，可将 l_1 范数缩放为 l_0 范数；同时借助于文献[61]中的结论可知迹范数是秩函数的凸包络表示，可将秩函数缩放为迹范数：$\|\boldsymbol{Q}\|_* \leq rank(\boldsymbol{Q})$，此时目标函数式(7)可以进一步表示为

$$\min_{\boldsymbol{W},\boldsymbol{P},\boldsymbol{Q}} \sum_{m=1}^{M} \left| L(\boldsymbol{W}_m^{\mathrm{T}} \boldsymbol{X}^m, \boldsymbol{Y}^m) + \lambda_1 \|\boldsymbol{P}\|_1 + \lambda_2 \|\boldsymbol{Q}\|_* \right. \quad (8)$$

$$\text{s.t. } \boldsymbol{W} = \boldsymbol{P} + \boldsymbol{Q},\ \|\boldsymbol{Q}\|_* \leq \tau$$

经过放缩之后式(8)为可处理的凸函数优化问题，也是联合稀疏项和低秩项学习方法的一般形式，文献[60]在保留了低秩结构 $\boldsymbol{\Omega}(\boldsymbol{Q}) = \|\boldsymbol{Q}\|_*$ 基础上，将 l_1 范数替换为组稀疏范数 $\boldsymbol{\Omega}(\boldsymbol{P}) = \|\boldsymbol{P}\|_{1,2}$ 用于分离离群任务，优化求解使用近端梯度方法进行加速，在文献[60]基础上，文献[62]选择平方铰链损失作

为损失函数，增强了鲁棒性和拓展性.

2.4.3　脏模型(Dirty)

文献[63]中，模型将所有可能与其它任务无法共享的特征都划分了出来，单独考虑稀疏特征的学习，其中在 $\boldsymbol{P}$ 的构造过程中创新性地使用了 $L_{\infty,1}$ 组范数正则化项，约束每行特征的总体相似度用来获得所有任务的共享结构. 文献[30]使用 $L_{2,1}$ 范数筛选任务之间的共享结构，使得在同一特征方向上尽可能的结构一致，与之不同的是，文献[63]提出的模型中，使用 $L_{\infty,1}$ 范数正则化项，不仅保证结构一致而且同一行特征的模型参数值也尽可能相似(由于 L_{∞} 范数的作用)，同时，L_1 范数约束诱导求解的 $\boldsymbol{P}$ 矩阵行具有稀疏特性，通过 $L_{\infty,1}$ 范数约束达到块稀疏化的效果，挑选出任务之间都能共享的一组通用特征. 矩阵 $\boldsymbol{Q}$ 表示不能被其它任务利用的特征，通过 $L_{1,1}$ 范数约束诱导行和列同时稀疏找出非共享特征之外的稀疏特征.

$\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 矩阵上正则化函数的组合可以学习到不同程度的共享特征，达到特征学习的目的. 文献[63]提出的模型虽然孤立提取出了稀疏特征项，但是并没有考虑这些稀疏特征之间的联系，特别是同一维特征上稀疏特征之间的相似性能否改善学习效果仍是一个有待研究的问题.

2.4.4　可变簇聚类模型

文献[64-65]推广了特征学习的形式，对特征的相似性进行了更一般的研究，通常的情况是在同一特征上任务之间可能具有相关性，而在另外一个特征上，这几个任务之间的相关性发生了其它变化，在多个任务之间同一维度的特征上，任务之间相关性都不是一致的. 作者提出在每一维特征上都应该达到任务聚类的效果，将文献[63]中整体相似结构矩阵 $\boldsymbol{P}$ 替换成离散关联的形式，从而在特征层面上保证任务的相关程度：

$$P_{clus}=\min\left(\sum_{d=1}^{D}\sum_{i<j}|P_{di}-P_{dj}|\right) \tag{9}$$

这种正则化约束打破了任务聚类只能根据多个特征进行任务分类的局限性，使得在同一维特征上任务之间的关联性更强.

在式(9)中令 $\lambda_1\to\infty$ 时，得 $P_{:,1}=P_{:,2}=,\cdots,=P_{:,m}$，也就是说每个任务模型向量接近于一个平均值并且近似相等，$\boldsymbol{P}$ 等价于一致共享结构，同时采用 $\|\boldsymbol{P}\|_F$ 约束复杂性. $\boldsymbol{Q}$ 矩阵此时依旧保留各自任务的特异特征，采用矩阵的 Frobenius-范数 $\|\boldsymbol{Q}\|_F$ 诱导 $\boldsymbol{Q}$ 从模型向量 $\boldsymbol{W}$ 中与 $\boldsymbol{P}$ 分离，与文献[13]提出的模型假设不同的是，这些特征完全没有联系，特征之间的相似性已经充分被 $\boldsymbol{P}$ 捕捉到.

2.4.5　协同聚类模型

文献[66-67]针对文本分类的无监督问题，通过非负矩阵分解和协同聚类的思想对任务和特征同时进行聚类. 在无监督场景下，文档实例和单词表对应任务和特征，并且任务和特征的关系矩阵不同于监督学习，是通过定义它们的联合分布矩阵 $\boldsymbol{P}$ 描述. 作者将 $\boldsymbol{P}$ 分解为：$\boldsymbol{P}\approx\boldsymbol{USV}^{\mathrm{T}}$，其中 $\boldsymbol{U}\in R^{D\times K_d}$ 代表将特征划分成 K_d 类，$\boldsymbol{S}\in R^{K_d\times K_m}$ 表示转置矩阵，$\boldsymbol{V}\in R^{M\times K_m}$ 代表将任务划分成 K_m 类，这样的分解满足任务和特征同时聚类的要求，得到了与任务特定信息相关的协同聚类，表达的是特征之间的私有信息. 同样的，在监督学习中，文献[68-69]也引入了协同聚类的思想，指出多任务的相关性更多地通过特征子集的相似程度来表示，比如在推荐系统中可以只根据几个关键因素判断用户喜好，即可以用特征的子集对任务进行聚类，并且每个任务分类依赖的子集是不重叠的，作者规定式(9)中的 $\boldsymbol{Q}$ 矩阵代表特征的局部相似性，与上述稀疏特征学习不同，此方法中 $\boldsymbol{Q}$ 是将挑选出的子集特征按照对角排列得到，从而达到在特征和任务方向同时聚类的效果，在文献[68]中，作者设计了一个正则化项 $\boldsymbol{\Omega}(\boldsymbol{Q})$ 描述局部相似性：

$$\|\boldsymbol{Q}\|_k^2=\sum_{i=k+1}^{\min(D,M)}\sigma_i^2(\boldsymbol{Q}) \tag{10}$$

其中 D 和 M 分别是特征和任务的个数，$\sigma_i(\boldsymbol{Q})$ 是 $\boldsymbol{Q}$ 的奇异值，作者指出协同聚类中任务和特征之间的映射关系是双射，并且能归为 k 类，$k\leq\min(D,M)$ 是类的个数，此时 $\boldsymbol{\Omega}(\boldsymbol{Q})$ 能取得的最小值就是式(10)，因此当 $k\geq rank(\boldsymbol{Q})$ 时，$\boldsymbol{\Omega}(\boldsymbol{Q})$ 为 0，相当于没有假设局部相似性.

特征子集是在属于一类的任务之间共享的，由于只注重学习这些子集会造成其它特征在任务之间共享的缺失(例如在复合范数正则化 $L_{2,1}$ 中的情况)，所以为了平衡 $\boldsymbol{Q}$ 矩阵只学习特征子集带来的影响，此时需要增加一个表示任务总体相似度的 $\boldsymbol{P}$ 矩阵上的正则化项：

$$\boldsymbol{\Omega}(\boldsymbol{P})=\sum_{m=1}^{M}\left\|p^i-\frac{1}{M}\sum_{j=1}^{M}p^j\right\|_2^2 \tag{11}$$

$\boldsymbol{P}=[p^1,p^2,\cdots,p^M]$ 代表各个任务的模型向量，所以式(11)的效果是约束每列向量靠近均值，保证总体相似度.

由于协同聚类模型虽然保留了任务的私有特征

信息,但是受制于正则化项式(21)的约束,有个明显的缺点就是最终形成的特征对角块结构没考虑特征的重叠,虽然分类任务区别特征的差异,但是彼此之间不可能没有重叠.

本节介绍了在多任务特征学习中的几种典型方法,主要技术手段是基于鲁棒主成分的矩阵分解.特征学习与以往方法不同的是,不再只考虑任务之间通用的结构,而将基于任务层面的整体特征学习转变为小范围之内单个或几个特征的学习,在保留了任务层面学习结构的同时,将学习的范围拓展到了特征层面,分理出了任务之间的共有特征和保留在各自任务的私有特征.其中孤立离群特征和离群任务方法、脏模型方法分离出了区别于其它任务的稀疏特征,增强了鲁棒性;可变簇聚类和协同聚类两种模型目的是针对稀疏特征的关系进行建模,学习稀疏特征的具体联系,形成一些特征簇,这两种方法与脏模型的稀疏矩阵对比如图 4 所示,几类典型方法的具体比较见表 1.

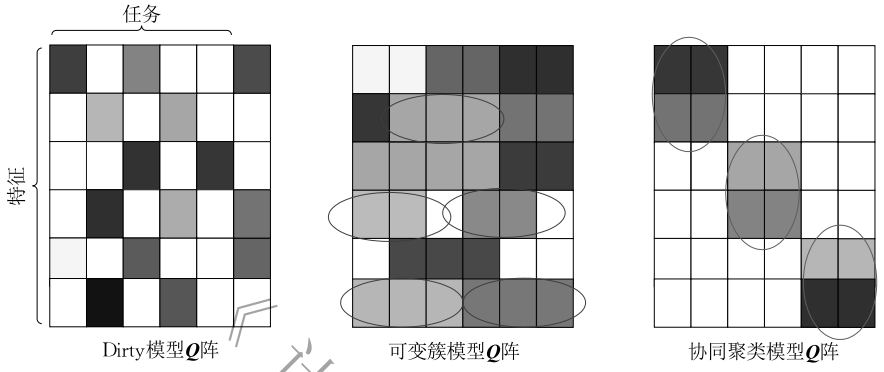


图 4 三种特征学习中的 $\boldsymbol{Q}$ 矩阵对比

表 1 几种典型多任务特征层面学习方法比较

特征学习方法	模型矩阵的形式和采用的正则化约束	达到的效果
鲁棒多任务特征学习[58](RMTL)	采用 $\boldsymbol{P}+\boldsymbol{Q}$ 的叠加形式,$\boldsymbol{\Omega}(\boldsymbol{P})=\|\boldsymbol{P}\|_{1,2}$ 并且 $\boldsymbol{\Omega}(\boldsymbol{Q})=\|\boldsymbol{Q}^{\mathrm{T}}\|_{1,2}$	挑选出离群的特征和变量
联合低秩项和稀疏项[59]	采用 $\boldsymbol{P}+\boldsymbol{Q}$ 的叠加形式 $\boldsymbol{\Omega}(\boldsymbol{P})=\|\boldsymbol{P}\|_1$,$\boldsymbol{\Omega}(\boldsymbol{Q})=\|\boldsymbol{Q}\|_*$	使用 $\boldsymbol{P}$ 捕捉与任务相关的判别特征同时使用 $\boldsymbol{Q}$ 学习低秩结构代表任务的底层共享性
脏模型(dirty)[63]	采用 $\boldsymbol{P}+\boldsymbol{Q}$ 的叠加形式,$\boldsymbol{\Omega}(\boldsymbol{P})=L_{\infty,1}$,$\boldsymbol{\Omega}(\boldsymbol{Q})=L_{1,1}$	使用 $\boldsymbol{P}$ 捕捉任务之间相似性的同时在 $\boldsymbol{Q}$ 中分离各个任务不能与其它任务共享的稀疏特征
可变簇模型[65]	采用 $\boldsymbol{P}+\boldsymbol{Q}$ 的叠加形式,$\boldsymbol{\Omega}(\boldsymbol{P})=\min\left(\sum_{d=1}^{p}\sum_{i<j}\|P_{di}-P_{dj}\|\right)$,且使用 $\|\boldsymbol{P}\|_F$ 惩罚复杂性,$\boldsymbol{\Omega}(\boldsymbol{Q})=\|\boldsymbol{Q}\|_F$	$\boldsymbol{P}$ 采用自定义范数度量每一维特征之间的相似性,$\boldsymbol{Q}$ 利用 Frobenius 范数诱导稀疏特征从 $\boldsymbol{W}$ 中与 $\boldsymbol{P}$ 分离
协同聚类模型[68]	采用 $\boldsymbol{P}+\boldsymbol{Q}$ 的叠加形式,$\boldsymbol{\Omega}(\boldsymbol{P})=\sum_{m=1}^{M}\left\|p^i-\frac{1}{M}\sum_{j=1}^{M}p^j\right\|_2^2$,$\boldsymbol{\Omega}(\boldsymbol{Q})=\sum_{i=k+1}^{\min(D,M)}\sigma_i^2(\boldsymbol{Q})$	$\boldsymbol{P}$ 保证了全局相似性,而 $\boldsymbol{Q}$ 在任务和特征同时聚类,形成了成对角排列的特征子集聚类簇,表现的是局部相似性

任务间的特征学习体现了任务在特征层面上的联系,对于任务完整性的描述是必要的,但是与向量空间相比,基于矩阵空间的学习技术缺少扩展性,假如样本维数过高时,会随着样本容量的大小在空间和时间复杂度上分别呈现二次方和三次方增长,单独分析这些特征的关联关系使得计算复杂性过高,如何有效地逼近目标矩阵使得方法更加鲁棒和精确是下一步应该考虑的问题.

2.5 贝叶斯生成式模型多任务学习

由于判别式方法引入的任务之间关系的假设比较固定,都是事先指定任务之间的关联方式,并以此设计目标函数,约束任务之间的模型或特征结构,一旦所采用的模型假设不合理就会对任务之间关系的推理产生副作用,引起任务之间的负迁移,同时,判别式方法没有考虑数据的生成过程,难以从数据中验证假设的正确性.因此,对于许多实际场景,描述任务关系的假设往往是不成立的.

为了克服这个问题,增加多任务学习结构的灵活性,基于贝叶斯方法的生成式模型的多任务学习也得到了广泛的研究,生成式方法能够针对任务相关性提供更灵活的建模手段,它只基于任务标识从数据样本中自动学习任务间的依赖关系,而不需要

预先定义任务关系结构，更贴近于数据本身的原有特性，同时，生成式模型具有灵活的框架和丰富的概率分布形式，除了能将大部分判别式的多任务学习方法转换为生成式表达之外，还有自己独特的多任务建模方法，例如层次贝叶斯方法，非参数聚类方法等，克服了很多结构通过判别式模型不能得到表达的缺点.

生成式方法假定要学习的模型参数 $\boldsymbol{W}$ 由一个潜在的概率模型生成，一般假定 $\boldsymbol{W}$ 满足一定的先验分布 $\boldsymbol{P}(\boldsymbol{W})$，此时模型参数矩阵 $\boldsymbol{W}$ 由参数的先验分布和给定模型参数的数据似然函数共同决定，即服从后验分布：

$$\boldsymbol{P}(\boldsymbol{W}|D)\propto\boldsymbol{P}(D|\boldsymbol{W})\boldsymbol{P}(\boldsymbol{W}) \tag{12}$$

其中在式(12)后两项中，先验分布提供了基本的关系假设，不同任务结构之间的关系由 $\boldsymbol{P}(\boldsymbol{W})$ 捕捉，似然函数 $\boldsymbol{P}(D|\boldsymbol{W})$ 能够利用数据本身的信息，最大化似然函数就是找到最符合各自任务数据特性的模型参数值. 在给定先验之前，单个任务学习彼此独立没有联系，在给定先验之后，通过和似然函数的共同作用不断调整关系假设，来自于所有任务的数据驱动使任务之间的信息传递成为可能.

2.5.1　层次贝叶斯框架的引入

为了摆脱结构的单一性，同时，考虑到生成式模型可以灵活地配置任务模型向量，更多的生成式模型假定任务的关系存在多样性. 一般将多个任务所服从概率分布的参数向量的关系构造成多个层次结构，而不直接从所需要得到的结果出发. 层次的概念指的是假设多任务模型参数 $\boldsymbol{W}$ 具有多个隐结构[70]，即可以对模型向量 $\boldsymbol{W}$ 引入多层概率假设，$\boldsymbol{W}$ 由超先验参数表示的概率分布生成，因此，层次贝叶斯模型假定模型参数形成层次依赖关系，对每一层参数都引入概率假设，层次结构的底层是特定于任务的单个参数模型，如高斯分布的均值和方差，任务之间传递信息是通过学习这些超参数. 在层的上面，子任务之间通过一个共同的概率分布参数连接在一起，即任务参数向量 $\boldsymbol{W}$，如图 5 所示.

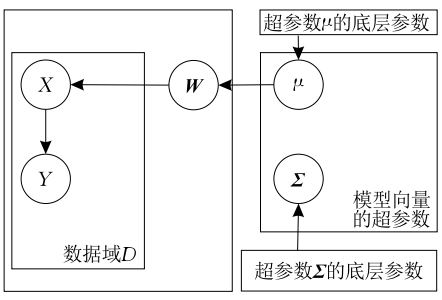


图 5　层次模型概率图

与式(12)中非层次贝叶斯模型可直接假定 $\boldsymbol{W}$ 的先验然后再求解的步骤不同，层次贝叶斯生成式模型的一般求解步骤是：首先假定模型参数 $\boldsymbol{W}$ 服从某种先验分布 $\boldsymbol{W}\sim f(\theta)$，然后将超参数的共轭先验分布设定为 $\theta\sim\boldsymbol{P}(\theta)$. 例如在文献[70-71]列举的高斯过程的求解方法中，要学习的任务模型参数 $\boldsymbol{W}$ 可以通过 EM 算法获得，将均值 μ 和协方差 $\boldsymbol{\Sigma}$ 的先验分布用逆 wishart 分布表示，再利用迭代吉布斯抽样方法从超参数的先验分布中抽样，抽出的模型参数样本服从均值和方差为设定值的高斯分布，然后依据全概率似然函数 $p(Y,X,\boldsymbol{W}|\Theta)$ 求出各自参数的边缘概率分布，逐个更新参数 $\boldsymbol{W}$，均值 μ 和协方差 $\boldsymbol{\Sigma}$，以及逆 wishart 分布中的超参数，这个过程反复执行，直到满足收敛准则.

在生成式的多任务学习中一般采用的都是层次贝叶斯框架，模型参数矩阵往往是层次结构，层次贝叶斯模型为类似任务提供了共享相同模型参数的机会，信息的表达会更丰富，又引入了一层底层概率分布，扩充了先验知识，各个任务不仅受其自身的训练数据影响，而且还通过不同层次的先验假设与其它相关任务建立起了深层联系，此时模型能够对任务的个体差异性和任务间的相关性进行建模，特征的构成兼具固定效应和随机效应，固定效应实现了参数的“硬共享”，而随机效应意味着通过对某些模型参数使用公有概率分布来实现“软共享”. 各个任务模型参数的求解是在与其它任务共享和保留当前任务的独立性之间权衡的结果，所以在多任务场景下，层次贝叶斯模型可以有效地实现任务之间的信息共享.

2.5.2　神经网络层次结构贝叶斯多任务学习

由于神经网络本身就是一种层次结构，贝叶斯层次结构可以由神经网络直接得到，所以文献[72-73]提出在特定任务的输出层参数上施加一个先验概率分布，通过计算似然函数，计算参数的最大后验概率. 如果任务数据之间联系比较紧密，那么后验概率分布就将任务分配给与这些数据概率分布最一致的那一类，具体生成式模型如下：

$$\begin{aligned} y_i^m &= \sum_{j=1}^{nhidden} A_{ij}\,h_{ij}^m + A_{i0} + noise \\ h_{ij}^m &= g\Big(\sum_{k=1}^{ninput} W_{ik}\,x_{ik}^m + W_{j0}\Big) \end{aligned} \tag{13}$$

此时全概率似然概率为

$$\boldsymbol{P}(D,A|\Lambda)=\prod_{m=1}^{M}\boldsymbol{P}(D_m|A_m,\boldsymbol{W},\sigma)\boldsymbol{P}(A_m|m,\boldsymbol{\Sigma}) \tag{14}$$

这种模型的层次性体现在任务分类层参数 A_m

上.在文献[72]中作者考虑了 A_m 的先验分布使用单一的高斯分布和混合高斯分布两种情况,当任务相似度比较高时,分类层参数使用单一高斯分布作为先验分布是有效的,因为一组均值和方差足以描述任务数据的特性,但是这种简单的先验假设并不符合一般情况,会丢失太多的任务特定信息,于是作者采用高斯混合分布作为先验假设用来扩充模型的表达能力,即 $A_i \sim \sum_{\alpha=1}^{cluster} q_\alpha N(A_m \mid m_\alpha, \Sigma_\alpha)$,可以看到每个任务的概率分布由多个单高斯分布经过加权混合产生,在每个单高斯分布上都会计算任务属于该分布的概率,然后再利用混合模型的软聚类特性进行类的划分,这种做法可以更充分地涵盖多个任务的数据信息,如图 6 左所示,当确定每个单高斯分布组分以后,通过一组共享的混合权重计算各自任务的分类概率.紧接着,文献[73]考虑高斯混合分量不变而为每个任务学习不同的混合权重,然后通过独立地计算输出层参数的后验分布并且估计在不同高斯混合分量上的软聚类概率,从而进行更精确地分类,如图 6 右所示.

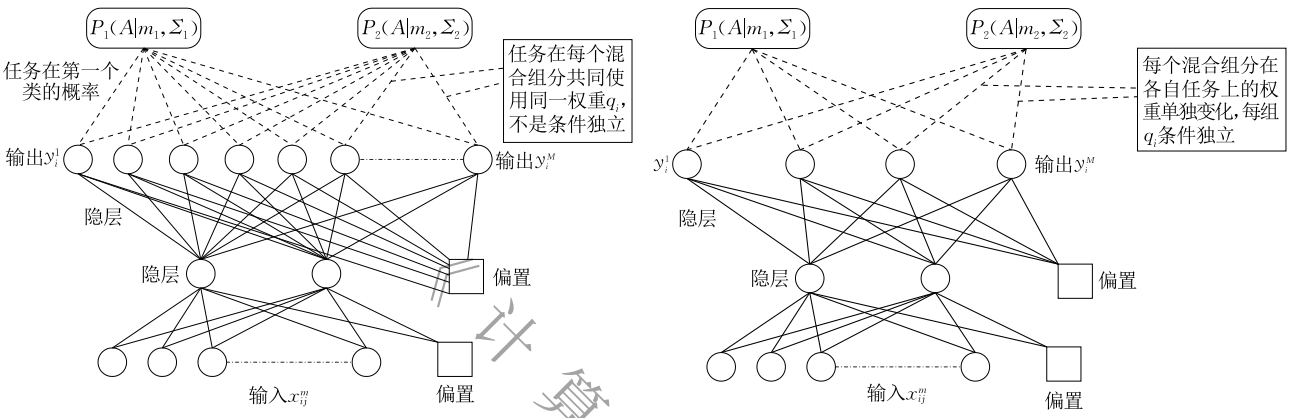


图 6 层次模型中均匀分类和混合分类

2.5.3 基于高斯过程的多任务关系学习

贝叶斯参数模型需要适当假设几个任务上的协方差函数的先验参数,才能从数据集中正确的学习任务之间的相关关系,因此要求任务模型参数 $\boldsymbol{W}$ 具有准确的先验假设,但是单个任务的模型参数可能很复杂,实际中也无法获得多任务关系背后真正的模型分布,分布函数的参数形式可能和正确的任务描述相差太大,无法模拟任务的相似性.一些学者致力于高斯过程的非参数多任务学习方法的研究[74-79],此处的非参数学习和非参数贝叶斯方法中非参数的意义不同,非参数贝叶斯方法不指定参数分布形式,而非参数学习是不指定输入和输出之间的具体函数关系,即将式(1)变为

$$\boldsymbol{Y}^m = f^m(\boldsymbol{X}^m) + \varepsilon^m, \varepsilon^m \sim N(0, \sigma_m^2) \tag{15}$$

通过对各个任务上 f^m 的求解描述,而且,由于核函数是对向量操作,所以将式(1)变为输入矩阵 $\boldsymbol{X}^m$ 和输出向量 $\boldsymbol{Y}^m$.模型的全概率似然函数为

$$p(\{\boldsymbol{Y}^m\} \mid \{\boldsymbol{X}^m\}, \theta) = \prod_m \int p(\boldsymbol{Y}^m \mid f^m, \boldsymbol{X}^m) p(f^m \mid \theta) \mathrm{d}f^m \tag{16}$$

在这种层次模型中,一般高斯过程的隐生成式模型,需要一个高斯先验施加到潜在的映射关系 $f: x \to y$ 上,即:$f^m \mid \theta \sim N(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$,其中 θ 为所有要求解的未知参数 $\theta = \{\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w, \sigma_m^2\}$,任务之间的相关性直接被 f^m 的高斯分布所捕获:方差越小,任务相关性越大,此时 $\boldsymbol{Y}^m \mid f^m, \boldsymbol{X}^m \sim N(f^m, \sigma_m^2)$ 代表输出在非参数映射先验假设下的条件概率分布,且满足高斯分布.而且这种非参数方法可以将线性模型扩展到非线性空间,超参数均值 $\boldsymbol{\mu}_w$ 和协方差矩阵 $\boldsymbol{\Sigma}_w$ 的先验分布一般从逆 wishart 分布中抽取(因为逆 wishart 分布与高斯分布互为共轭先验,有利于求解):

$$P(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) = N\left(\boldsymbol{\mu}_w \mid 0, \frac{1}{\pi}\boldsymbol{\Sigma}_w\right) \boldsymbol{IW}(\boldsymbol{\Sigma}_w \mid \tau, k) \tag{17}$$

此时多任务的模型为

$$\boldsymbol{Y}_i^m = f_i^m(\boldsymbol{X}_i^m) + \varepsilon^m,\ \varepsilon^m \sim N(0, \sigma_m^2) \tag{18}$$

这种层次模型也被称为直推式基于高斯过程的多任务模型,直推式方法存在一个显著的弊端就是,当加入新测试数据的时候,需要重新求解参数 $\theta = \{\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w, \sigma_m^2\}$.文献[75]提出了归纳式表示,将直推式方法变换到高斯核函数空间下求解,将 $N(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ 中超参数换成均值函数 $\mu_f(x) = \boldsymbol{\mu}_w^{\mathrm{T}} x$ 和协方差函数 $K(x_i, x_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\Sigma}_w x_j$,给出了高斯过程非参数方法转换到高斯核函数的一般过程,提出了利用多任务核函数方法求解任务关系,将其转换为核空间下等价的多任务模型:

$$\boldsymbol{Y}^m=\sum_{i=1}^{N_m}\alpha_i^m k(\boldsymbol{X}_i^m,x)+\varepsilon^m,\ \varepsilon^m\sim N(0,\sigma_m^2)\quad(19)$$

$$\alpha^m\sim N(\mu_\alpha,\boldsymbol{C}_\alpha)$$

$k(x_i,x)=\langle x_i,x\rangle$代表正定的核函数，$\mu_\alpha$和$\boldsymbol{C}_\alpha$的值是从先验分布中抽取得到的：

$$P(\mu_\alpha,\boldsymbol{C}_\alpha)=N\left(\mu_\alpha|0,\frac{1}{\pi}\boldsymbol{C}_\alpha\right)\boldsymbol{IW}(\boldsymbol{C}_\alpha|\tau,k^{-1})\quad(20)$$

归纳式模型对新来测试数据的适应性较强，不需要每次重新求解参数$\theta=\{\mu_f,K,\sigma_m^2\}$，模型具有良好的可扩展性.

上述模型中[75,78-79]，协方差矩阵$\boldsymbol{C}_\alpha$为块对角结构，每个任务块由同一核函数导出. 这种情况下，任务之间的输出都是条件独立的，任务间的相关关系仅仅通过任务间共享核函数实现. 与此相反，在文献[76,80-82]中对协方差的形式进行了修改，$\boldsymbol{C}_\alpha$不是块对角结构，因而，y的联合分布也不再是对角阵，非对角结构的协方差矩阵实现任务之间的关联，一个任务可以影响另一个任务的预测. 具体的方法引入了一个描述任务相似度的矩阵$\boldsymbol{\Sigma}_{mk}^f$，将式(18)等价为

$$\langle f^m(x)f^k(x')\rangle=\boldsymbol{\Sigma}_{mk}^f k^x(x,x')\quad(21)$$

这里，k为核函数，f是映射关系，假定输出满足正态分布：

$$\boldsymbol{Y}_i^m\sim N(f^m(\boldsymbol{X}_i^m),\sigma_m^2)\quad(22)$$

此模型将任务的协方差矩阵和高斯先验组合，Σ_{lk}^f代表第m个任务和第k个任务之间的关系，k^x是输入的协方差函数，$\boldsymbol{\delta}_l^2$是第l个任务噪声的协方差矩阵，为了避免参数冗余，令k^x，$\boldsymbol{\Sigma}^f$都为半正定矩阵. 例如，如果$\boldsymbol{\Sigma}^f$是块对角阵结构，那么任务的关系最后表示为分类形式，若要以任意的形式实现任务集关联，可以假定$\boldsymbol{\Sigma}^f$不具有特定的形式.

在整个构造过程中，仅用一个变量$\boldsymbol{\Sigma}_{mk}^f$去调整任务关系，这样就可以在输入特征$X$上使用一个共同的参数化协方差函数，减少了参数化的复杂度，因而可以比较灵活地学习任务关系.

虽然文献[39]提供了一种以任务协方差矩阵的形式来学习任务关系的全局方法，但是这个模型的一个缺点是，高斯核函数需要对任务间样本的协方差函数进行估计，由于对数似然函数是非凸的，作者使用了低秩近似，因此，任务协方差矩阵的学习对参数初始化很敏感，且不保证能找到最优解. 此外，由于该方法是基于高斯过程的，把它扩展到任务数目较多的大数据集上可能要考虑计算成本，此时采用低秩近似虽然降低了计算成本，但可能会限制任务协方差矩阵的表示能力.

文献[74,83-84]提出了一种凸多任务关系学习方法，以非参数形式为协方差矩阵建模，同时学习模型参数和任务关系. 该方法建立起了核空间和向量空间的联系，核函数表示的任务关系可以转换到权重向量空间中讨论，将原有的潜在映射关系f转换到新的向量空间讨论，并且用$\boldsymbol{w}_m^{\mathrm{T}}\phi(x_j^m)$代替潜在映射关系$f$，使用非线性特征映射$\phi(\cdot)$将原始向量空间$x_j^m\in R^d$映射到不同维度的向量空间$\phi(x_j^m)\in R^b$上，与式(21)和(22)相比可知：

$$f_j^m=\phi(x_j^m)^{\mathrm{T}}\boldsymbol{w}_m\sim N(0,\Sigma_{jj}k(x_j^m,x_j^m))\quad(23)$$

由此验证了两个空间在分布上是完全等价的，可以看到可以将隐关系函数变成一个非线性映射表达式，这里并不直接映射回原向量空间，是为了避免维度太高方差矩阵无法求解，此时将原来的非参数模型式(18)转换为

$$y_j^m=\boldsymbol{w}_m^{\mathrm{T}}\phi(x_j^m)+\varepsilon_j^m$$

$$\boldsymbol{W}=[\boldsymbol{w}_1,\cdots\boldsymbol{w}_m]\sim\left(\prod_{i=1}^{M}N(w_i|0_d,\varepsilon_i^2\boldsymbol{I}_d)\right)q(\boldsymbol{W})\quad(24)$$

$$\varepsilon_j^m\sim N(0,\sigma_m^2)$$

任务输出的条件后验概率分布为

$$y_j^i|x_j^i,\boldsymbol{w}_i,b_i,\varepsilon_i\sim N(\boldsymbol{w}_i^{\mathrm{T}}x_j^i+b_i,\varepsilon_i^2)\quad(25)$$

$q(\boldsymbol{W})=M\boldsymbol{N}_{d\times m}(\boldsymbol{W}|\boldsymbol{0}_{d\times m},\boldsymbol{I}_d\otimes\boldsymbol{\Sigma})$是对$\boldsymbol{W}$结构的建模，方差矩阵$\boldsymbol{I}_d\in R^{D\times D}$描述特征之间的关系，列方差矩阵$\boldsymbol{\Omega}\in R^{M\times M}$描述任务相关性.

以往的层次结构是将先验约束施加到参数向量$\boldsymbol{W}$上，而高斯核函数层次结构的先验约束变为施加到隐函数f上，假设每个任务的关系是从一个共享高斯先验分布中获得，摆脱了直接定义参数先验的局限性，但是为了更好地体现参数或者函数关系的复杂性，最好是不指定学习过程中的任何参数的具体形式.

2.5.4　非参数贝叶斯层次结构多任务关系学习

为了克服引入不适当先验假设对任务关系学习产生误导的问题，文献[85-86]提出了一种基于非参数先验的层次贝叶斯模型，与高斯核函数学习非参数形式的隐函数不同，此时非参数贝叶斯指的是参数的先验分布形式. 先验分布服从 Dirichlet 过程(DP)：

$$\begin{aligned}&\boldsymbol{w}_m|G\sim G\\&G\sim DP(\alpha,G_0)\end{aligned}\quad(26)$$

G_0是指定的基概率分布，不需要对先验分布的函数形式作任何假设，DP 可以逼近任意概率分布，因此可对任意复杂度多任务参数建模，同时 DP 是动态增量模型，可以在学习过程中自动调整模型复杂度即任务簇的个数，每个组共享相同的模型. 本质上来说，DP 模型是不断扩充类别的多任务分类过程，自

动根据训练数据的相似分布特点学习相关任务.

不过,由于 DP 产生相同的参数得到相关的任务,任务服从的概率分布形式是离散的,相同的任务才能组合到一组,因此作者提出可以将 DP 扩展到层次贝叶斯模型,克服离散特性带来的不足,即将狄利克雷混合模型(Dirichlet Process model)作为模型参数的先验.

2.5.5 判别式方法的生成式实现

判别式方法和生成式方法作为多任务学习的两种实现途径,处理手段有所不同,但有结构上的对应关系,本小节主要对应于判别式方法的学习结构给出在生成式框架下的实现.共享参数 $\boldsymbol{W}$ 是生成式模型最直接的假设,文献[87-88]利用不同任务之间的给定层次结构作为先验信息来帮助学习模型参数,其中假设任务权重参数 $\boldsymbol{W}$ 通过引入一个隐变量,用这个隐变量表示共享一个共同参数结构的概率分布,比如假设它们从同一个高斯过程分布中生成,此时生成式模型的目标是学习高斯过程的均值和协方差函数参数.由于隐变量服从概率分布底层参数的一致性,模型参数 $\boldsymbol{W}$ 建立在统一概率分布之上,由此而学习出来的结构类似于判别式模型中的模型参数共享方法.

对应于公共特征共享,生成式方法[89-90]提出了一个统一的概率框架,其中任务特征参数共享一个共同的结构,文献[91]在相关任务中引入元特征的特征选择,假设特征 d 可以由一组元特征 f_d 来表征,并定义元特征的先验为特征的相关性建模;文献[92]概率通过引入随着任务参数自动调整的概率框架,预先确定任务的相关性假设并调整参数的方差;文献[93]称为自适应组套索方法,类似于加权组稀疏的方法[42],作者将(2)中正则化项 $\boldsymbol{\Omega}(\boldsymbol{W})$ 和与各自任务特征 $f_{m,d}$ 有关的混合因子 $\theta=\Sigma_m v_m f_{m,d}$ 相乘作为参数矩阵 $\boldsymbol{W}$ 的超参数.以上生成式特征选择方法建立先验知识都是从特征入手,还有一类直接对参数矩阵 $\boldsymbol{W}$ 进行假设,文献[94-95]使用矩阵变量高斯先验进行特征选择,矩阵变量高斯分布可以表示为

$$vec(\boldsymbol{W}|D)\sim N(vec(\boldsymbol{M}),\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}) \tag{27}$$

通过将 $\boldsymbol{W}$ 向量化为 $vec(\boldsymbol{W})$,其中 $\boldsymbol{M}\in R^{D\times M}$ 代表 $\boldsymbol{W}$ 中每个元素的期望值,参数矩阵的协方差被分解为行协方差和列协方差的 Kronecker 积,$\boldsymbol{\Omega}$ 是 $D\times D$ 的列协方差矩阵,对应共享特征结构,$\boldsymbol{\Sigma}$ 是 $M\times M$ 的行协方差矩阵,对应任务相关性.文献[94]通过在极大似然函数中对 $\boldsymbol{\Omega}$ 和 $\boldsymbol{\Sigma}$ 施加 L_1 范数正则化进行共享特征选择;文献[95]利用高斯混合的矩阵变量先验捕获任务相关性,通过引入了一组广义稀疏诱导先验 γ_d 将某些特征块诱导为零从而达到组稀疏效果:

$$\boldsymbol{P}(W_{:,d})=\int_0^{\infty} N(0,\gamma_d^{-1}\Omega_d,\boldsymbol{\Sigma})\boldsymbol{P}(\gamma_d)\mathrm{d}\gamma_d \tag{28}$$

聚类结构可参照 2.5.4 节中的非参数方法生成聚类簇.关于低秩子空间结构,文献[96]提出了一种非参数贝叶斯模型,自动推断子任务空间的大小,过程是通过将参数矩阵分解为:$\boldsymbol{W}=\boldsymbol{Z}A+\xi$,其中 $\boldsymbol{Z}$ 作为新的子空间而 A 作为编码系数,并且添加噪声项 ξ.$\boldsymbol{Z}$ 定义的底层预测子空间的内在维数和稀疏度通过非参数的印度餐馆模型(Indian Buffet Process)加以实现.在文献[97]中作者进一步将线性子空间进行高斯混合用来逼近当模型参数是非线性流形的情形,避免来自不相关任务的负迁移.

针对特征层面的学习结构,文献[98]提出了脏模型[59]的生成式实现,通过引入一组关于二元隐变量的鲁棒先验分布用来识别稀疏特征,具体是通过假设参数矩阵中的每个元素 $W_{m,d}$ 满足一个离散混合 spike and slab 先验:

$$W_{m,d}\sim(1-\rho)\delta_0+\rho\pi(W_{m,d}) \tag{29}$$

其中 δ_0 判断该特征是否为零,而 $\pi(W_{m,d})$ 指定不为零特征的概率密度,简单的可以使用高斯分布.最近,文献[99]提出了联合低秩稀疏的贝叶斯方法,提供了数据对模型不确定性的概率解释,作者使用贝叶斯层次结构表示低秩分量 L 和稀疏分量 S.低秩结构是通过对 L 进行奇异值分解(SVD)并引入对角指示矩阵 $\boldsymbol{Z}$:$L=\boldsymbol{U}(\boldsymbol{Z}\Lambda)\boldsymbol{V}$ 得到的,诱导 $\boldsymbol{Z}$ 对角元素稀疏和通过奇异值矩阵收缩原理相同,都能得到低秩矩阵,每个对角元素 Z_{kk} 满足伯努利分布 $Z_{kk}\sim bernoulli(p_k)$,类似的,稀疏分量 S 分解为伯努利分布矩阵和实值矩阵的 Hadamard 积:$S=B\circ E$,其中 B 和 E 的每列分别服从伯努利分布和高斯分布:

$$\begin{aligned} B_m&\sim\prod_{d=1}^{D} Bernoulli(\pi_d)\\ E_m&\sim N(0,\gamma^{-1}\boldsymbol{I}_d)\end{aligned} \tag{30}$$

伯努利分布的参数 p_k,π_d 和高斯分布的参数 γ 分别服从 Beta 分布和 Gamma 分布构成层次结构.与传统的稀疏约束通过抑制非稀疏项使其近似趋于零不同,使用二项伯努利分布可以将取值固定到“0”和“1”,将 $\boldsymbol{Z}$ 和 B 中某些分量完全设置为零,达到了低秩约束和特征强稀疏的效果.

多任务生成式模型建立在先验假设的基础上.层次贝叶斯模型为参数矩阵引入了多层概率假设,将概率分布延伸为不同的层次,增强了对任务关系

的建模能力，为生成式方法提供了一个基本参数框架. 基于高斯过程的多任务学习方法将各个任务的输入和输出映射到高斯核函数空间，核函数服从多维高斯分布，任务之间的关系可以通过高斯分布中协方差矩阵不同的结构加以描述. 更一般地，非参数贝叶斯模型可以随着不同类型任务关系动态变化，将模型扩充到任意分布而不局限于单一概率假设，赋予了模型灵活度并且扩充了表达能力.

生成式模型增强了多任务关系的可解释性，摆脱了建模过程中人为指定共享结构的依赖，完全结合概率先验知识从数据中获得，但是生成式模型学习的结果受样本数据影响，对大规模特征的适应性较差，因此一些学者开始着重于研究对特征表征能力比较强的深度多任务学习.

2.6　深度多任务学习算法

深度神经网络能够学习到数据的多层抽象表示，对多任务之间内在的特征关系和任务关系有更强的解释能力，深度神经网络可以通过与其它网络共享参数来从相关任务中获益. 以卷积神经网络结构(Convolutional Neural Network，CNN)为基础的图像识别[100-102]和以长短记忆神经网络结构(Long Short-Term Memory，LSTM)为基础的语音识别[103]综合了多个识别过程，实现了效果的提升，也获得了越来越多的关注.

深度多任务学习可以分为硬参数共享神经网络和软参数共享神经网络，参数化处理的张量网络，以及近期出现的自适应层连接网络和自适应层分堆网络等形式.

2.6.1　硬参数共享结构神经网络

硬参数共享网络最早是 Caruana 在文献[6]中提出的，深度神经网络中硬参数共享的基本思想是不同的任务共享一些隐藏层，这样就可以学习多任务的联合表示. 在硬参数共享结构中，将所有任务输入公共隐层训练后得到一个通用的参数模型，这样的共享结构，可学习到多任务上的统一表示，任务之间只保留了一些有效信息，保证了多任务之间信息传递性的同时，降低了各自任务过拟合的风险. 硬参数共享可以看作是平均约束学习的一种非常粗糙的形式，其中单个任务模型隐藏层的参数被强制设定为与其它任务模型隐藏层的参数相同.

硬参数共享的优点是结构设计简单，只需设计隐藏层的结构，在不同的任务中共享网络参数，而不需要对这些层中的任务关系进行准确的建模. 缺点是建立在假定大多数任务有较高相关性的基础上，实际适用场景很有限，同时很多任务的特征关系受到强制约束而被丢弃，如果这种硬性结构在深层网络的关联中没有起作用，那么很可能会出现负迁移.

大多数硬参数共享网络方法[73,87,104-105]通过共享大部分隐层，学习特征的共享表示，只在输出端划分出根据任务特点单独设计的分类器预测输出层，因此在参数结构上，典型的设计是为底层使用相同参数，为顶层单独设计与各自任务相关的参数. 但是，由于各个任务在分类器层参数独立，信息在分类器层的自适应传播被阻断，导致任务关系没有得到建立. 而且有文献研究发现[106]，随着网络层数的增加，特征最终从一般到特定网络进行传递，而且特征传递在深层次网络中明显下降，因此发生负迁移的风险增加. 如何解决深层网络中特定任务层的关系是硬参数共享网络的研究重点.

2.6.2　软参数共享结构神经网络

很多场景中，硬参数结构往往依赖于人为预定义的共享结构，但是很多任务之间的联系并不是非常紧密，任务个数的增加使得任务之间的多样性也随之增加，只设计硬参数共享约束很难包含所有任务上独特的模型，因此很多学者提出了不指定共享结构的软参数共享神经网络，只在参数空间之内讨论哪些信息应该交互，软硬参数共享的对比如图 7 所示.

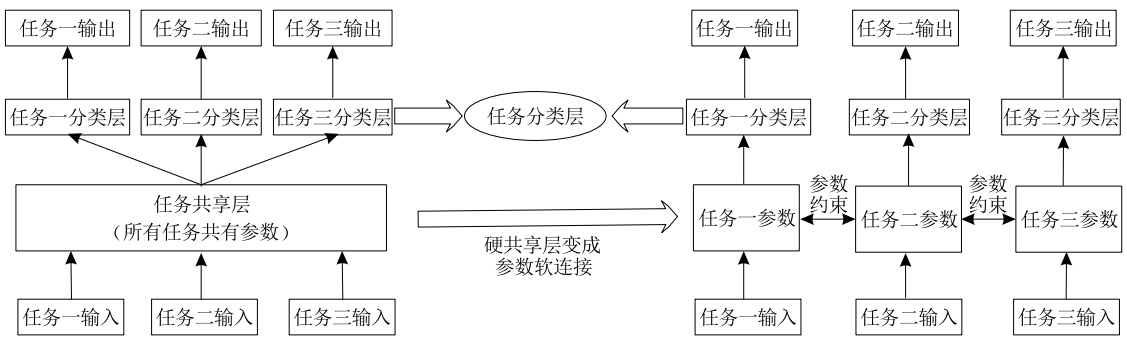


图 7　深度神经网络多任务学习的硬参数共享(左)和软参数共享(右)

深度神经网络结构中包含了大量的隐层，很难为一些不同任务之间划分相应的共享层，与硬参数共享结构不同，软共享结构保留了各自任务网络的结构，将所有层绑定同样的参数调整为假设所有的特征层存在于一个共享空间中，通过正则化方法约束模型参数的距离，保证任务之间是相似的而不是相同的，类似于多任务判别式模型中正则化约束下的统一特征结构. 软共享结构去除了硬参数共享中任务某些层之间不交换信息的限制，并且共享策略完全是在数据驱动的方式下实现的，是一种带参数约束的网络.

软参数共享方法具体的实现步骤是：为每个任务设计一样的网络结构，保证任务之间具有相似的结构关系. 然后主要是参数的学习过程，软参数共享在每一分层上对多任务参数堆叠，以层为单位形成所有层的张量集合，以参数比较复杂的卷积神经网络为例，卷积层的第 i 层张量 $w_{(i)}$ 维数为：$D_1 \times D_2 \times \cdots \times D_N$，任务数量为 M，在此层上可以施加范数 $\|\lambda w_{(i)}\|$ 约束任务的相似度，具体形式可以是文献[107-108]使用的 L_2 范数 $\|\lambda w_{(i)}\|_2$，也可以是文献[109]使用的迹范数 $\|\lambda w_{(i)}\|_*$.

软硬参数共享是建立在任务数据分布相似的基础上，但是不同的任务结构不可能完全相似，因此有学者开始研究深度多任务网络结构的划分，网络结构由人为指定转变为从数据中自主学习，同时将学习任务关系和学习特征迁移分离开来，使用部分参数共享或者部分结构共享代替单纯的全共享.

2.6.3 张量网络

张量方法是一种参数表示方法，可以看做子空间学习方法在深度网络下的自然推广. 由于深度网络的层数一般比较多，用张量方法可以代替传统的矩阵处理方法，寻求子结构之间的一些表示形式. 以往的一些多任务学习的深度学习方法[110-113]通过一些隐层学习共享特征表示，只在分类层分离任务，因为各任务的知识在分类层并没有共享，同时多个独立分类器并不能推断任务关系，导致多个任务上知识不能在不同的分类器之间自适应地迁移，可能导致在分类层上知识迁移不充分.

此外，标准的深度网络学习中，深层次的特征表示具有高度抽象性，文献[106]的最新研究表明：深层网络的特征会随着网络的深度加深，从一般特征表示转变为具体特征表示，这也是深度网络的特点. 但是在深度多任务学习中，以卷积网络为例，一般是共享全卷积层和第一个全连接层，之后再针对特定任务的结构分离为分类层[114]. 由于在图像处理中，这种设计已经完全可以满足要求，能够获得比较好的效果，但考虑神经网络的整体结构，随着任务之间差异的增加，特征的可迁移性在最后传递到最终分类层后会显著降低，全连接层到任务分类层并不能充分的过渡，因此，共享所有的特征层可能会增加负迁移风险，利用任务间的任务关系来解决深度网络中特定任务层的特征转移有待解决.

文献[115]论证了对输出端特定任务的分类层建模的必要性，并且提出了一种可以加强各个任务分类层之间联系的深度关系网络(Deep Relationship Network，简称 DRN). 在结构上，除了以往的硬参数共享层之外，在全连接层和任务特定分类层之间额外增加了描述多任务关系的全连接层，这个结构兼顾了两个因素，一是降低输出特征维数，二是和任务分类层结合，在任务之间建立深层联系，以此在多个任务的空间上，形成三阶张量网络.

由于在多任务学习中矩阵正态分布不能作为多层网络参数的先验分布，因此 DRN 引入了张量正态分布，在两个全连接层和任务特定分类层堆栈成的三阶张量网络参数上施加一个先验分布，从多层参数张量的后验分布中可以同时学习共享的深层特征和任务之间的关系，以此强化多任务特定层之间内在联系，克服了之前方法在任务特定层不能学习任务关系的局限，也避免了以往参数共享方法在特征层和分类层引起的负迁移.

由于应用的场景主要建立在图像分类上，通过同时学习特征迁移和任务关系，DRN 能够缓解在特征层中的负迁移和分类层中的信息迁移不足，此方法只针对全连接层之后的网络建模，大多数结构依旧依赖于卷积共享层，这是由于共享卷积层已经可以学习到大规模迁移的特征以及任务共享的参数，足够满足计算机视觉问题的要求，并不需要再对卷积层作更详细的区分. 但是 DRN 对于新任务的适应性较差，需要重新进行联合训练，也难以应对更复杂的多任务学习场景. 在此基础上，文献[113]从张量分解理论出发，将深度网络的参数纳入张量框架下进行统一表示，着手设计了深度多任务学习中全网络张量的共享结构，实现了端到端深度网络共享的主动学习，同时解决了非同构网络的输出问题. 这种方法能找到比原始张量更低阶的参数，对模型参数进行了压缩，因此张量模型参数的处理方法可以类比矩阵因式分解方法，训练模型分解成一个基矩阵和表示矩阵的内积形式.

基于整个网络的张量学习摆脱了个别层的任务依赖关系，整个任务网络建立在一些共享模型上，每个任务由基模型混合而成，一个任务的权值张量可以用任务基张量加权表示，通过调整各个组分之间不同的加权比例描述任务之间的关系，也就是说任务的关系建立在不同任务基组合的基础上. 这样做的好处是，当新数据出现时，可以选择重新训练参数，并保持基矩阵固定，这将大大减少学习参数的数量，从而减少所需的训练数据，克服了文献[115]对新任务适应性较差的障碍，与此同时全网络参数的张量分解完全由数据驱动自主学习，从而摆脱了需要人为定义多任务模型结构的不科学性和穷举搜索的繁琐性，也降低了模型结构选择的复杂性.

张量分解方法从软硬结构深度网络中脱离出来，将结构设计问题，变成了结构之间的表示形式学习问题. 由于不同任务的训练数据可以从不同概率分布中抽样得到，并由不同模型进行拟合，软硬结构网络不能确定任务之间的固有相关性，因此也无法验证预定义的参数共享策略的正确性.

尽管有以上优点，张量方法还是假设了任务之间存在整体相似度，即基于假设总体网络中至少能找到一些相似的子任务基网络. 但是，在很多任务中，并不是一些特征能自由地在隐特征层之间迁移，因此，出现了更具体地针对特征层的深度学习方法，这些方法主要是在各层之间设置一些连接结构，选择性地配置共享层，从而达到任务之间筛选共享特征的目的，因此他们被称为自适应动态层连接网络，包括 2.4.4 节的拆分缝合网络和 2.4.5 节的水闸网络.

2.6.4　自适应层连接拆分缝合网络

在多种类型的监督学习场景下，在多个任务之间和单个任务上用于获得更好学习性能的结构均是不同的. 通常多任务学习方法要求在任务之间共享一些特征，并保留一些与单个任务相关的自有特征. 由于结构选择的任务依赖性，多参数的硬共享和软共享为了获取适合各个任务的结构，需要针对不同任务进行改进，没有一般理论指导整个选择过程，从原理上设计共享表示和特定表示的规则. 因此，大部分基于软硬参数共享结构的神经网络不具有普遍适用性.

不同的任务集共享的结构不一样，并且每个任务都含有私有特征，通过枚举搜索并不能找到应该共享或者单独保留的网络层结构，同时如果在深层网络的训练中盲目搜索会造成算法复杂性非常高.

文献[116]设计了一种简单的拆分缝合网络，作为搜索和学习这种结构的原则性方法，针对特征的任务依赖性，可以为每层都设计一个拆分共享单元. 如图 8 所示，拆分共享单元在一些层或特征上有不同的组合，在任务之间起到连接或者阻断特征传递的作用，可以自动学习共享和任务特定表示的最佳组合. 该单元可在共享和特定任务表示之间自由移动，通过设置较低的层间连接边参数，决定哪些层是任务特定，或者通过分配较高的层间连接边参数选择任务之间特征共享方式. 各个任务在网络的每一层都学习到一个从输入到激活的线性映射作为共享表示，下一层在这个共享表示上执行非线性变换，最终在整个任务集上学习到共享和任务特定表示的最优线性组合.

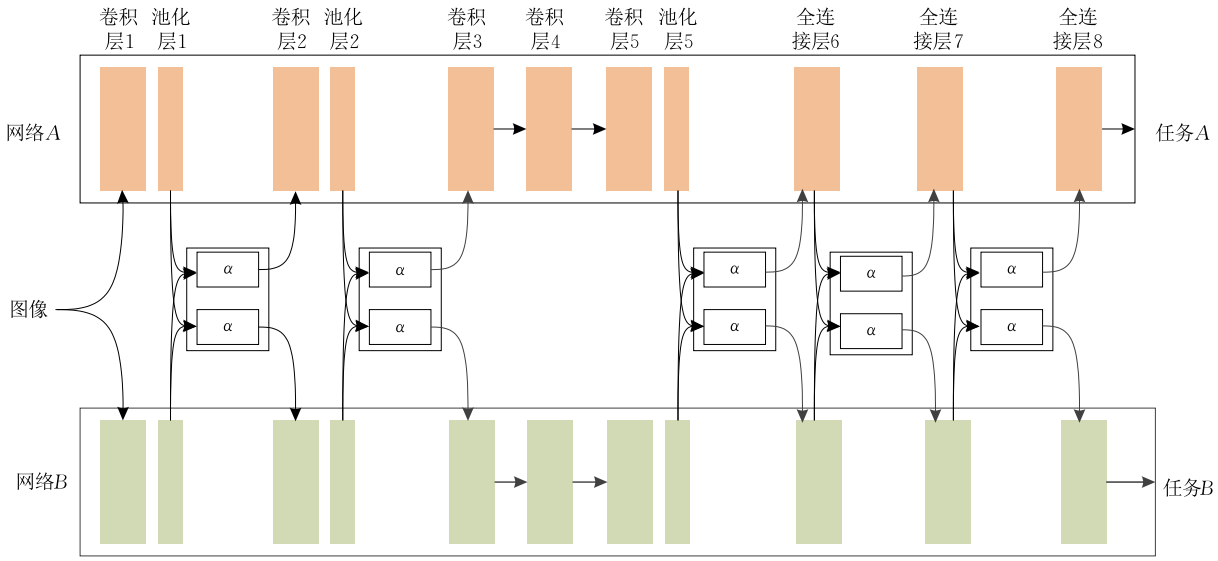


图 8　自适应层连接网络(拆分缝合网络)

此网络结构的另一个优势是属于增量网络,以往训练好的网络对于新任务的适应性较差,需重新探索可能的结构,但是重新枚举每一组任务所有可能的结构又是不切实际的,运算代价高昂的同时不一定寻找到最优解.拆分缝合网络中的优势在于最佳拆分结构只取决于手头的任务,设计共享特征和区别私有特征仅仅依赖于当前任务的特点,不需要再为每一个新任务增加拆分结构,而是在现有拆分缝合单元上进行新的特征组合,增强了拓展性.

2.6.5 自适应层连接水闸网络

虽然有一些理论提供了保证某些类型的多任务学习[2-3]的有效性,这些都不适用于松散相关的任务集,层内参数的软共享和硬共享很难确定是否能改进任务的泛化性能,特别是当任务只是松散相关的时候.为了克服这个问题,在此基础上,多任务关系假设的场景与判别式方法中的模型描述的任务松散相关假设一致[37],将性能增益与任务属性关联起来.

为了弥补在松散相关的任务的情况下深度学习理论的缺乏.文献[117]引入了水闸网络区分特征的私有和共享子空间,将每一层分为共享单元和松散属性特征单元(松散属性特征是与各自任务相关的,体现单个任务的特点,如图 9 所示).在结构上,水闸网络是拆分缝合网络以及硬参数共享网络的自然推广,虽然也是在层之间设计开关单元,但是从形式上相比拆分缝合网络增加了开关之间的连接,模型的目标是训练开关网络参数来控制多任务之间私有和共享空间的一般结构.水闸网络研究了有助于模型从辅助任务中获益的可能组合,以及各个任务之间应该共享的特征属性,从结构上统一了共享的方法,但是它并未指出属性的共享特征和私有特征的精确划分方法,而穷尽搜索多任务的共享子空间也是不现实的.

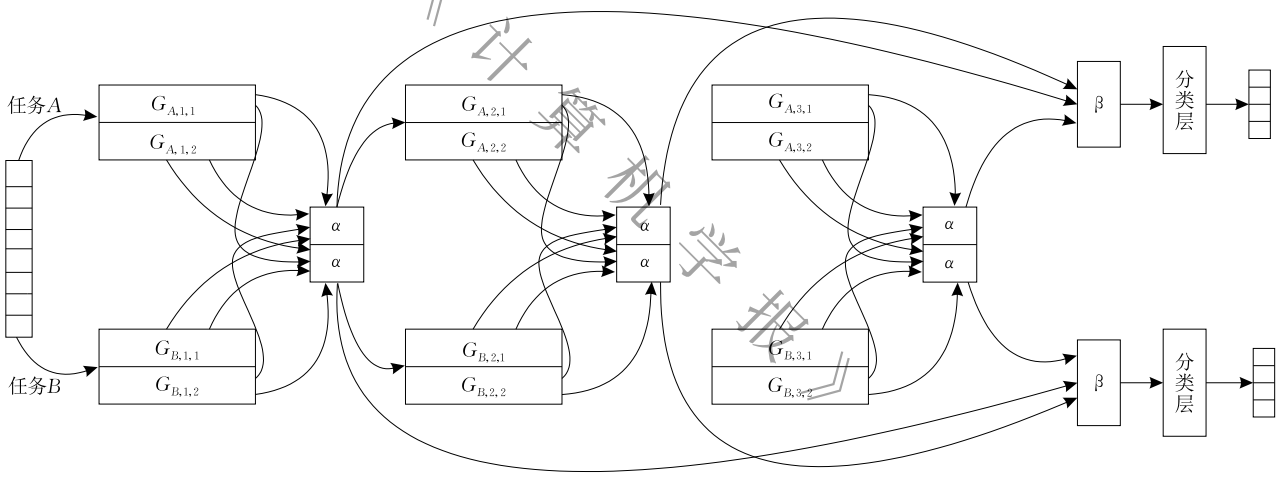


图 9 自适应层连接网络(水闸网络)

2.6.6 自适应任务层分堆

以上所述的深度网络都具有一个一般的结构:底层捕捉低层次的详细特征,这些特征可以在多个任务之间共享,而顶层是对应于任务的抽象特征.若出现一些不相关的任务,即使在同一个参数空间之内,也难以共享这些特征信息,因此自底向上不受限制的共享可能会遇到负迁移的问题,在两个不相关的任务中的信息共享,会添加干扰信息从而导致两者的性能恶化.为了避免这种情况,大多数多任务深度结构共享底层,直到共享阻塞后的某些层,再把层状结构变为树状结构,这样会导致描述任务特定性质的子网分支超出范围,如上述拆分缝合网络或者水闸网络,虽然连接任务特定子网的交叉缝合单元被设计来学习任务间的共享特征,寻找任务之间共享和特定表示的最佳组合,但网络的规模与任务数量成线性增长,从而导致可伸缩性问题.文献[118]提出了动态扩展多任务的组合结构,自动学习深度网络结构.如图 10 所示,每一层都进行任务之间的共享特性决策,自上而下贪婪地进行子分支任务分组,利用正交匹配的方法,由简单的初始化网络逐层扩展,动态地找到适合多任务分组的网络宽度,达到自适应模型加宽的效果,缺点也很明显,每层的可伸缩性较大,尤其是网络层数较深时,模型参数求解计算的难度加大.在深度多任务学习过程中,各个任务中不同网络层彼此之间的连接方式决定了任务联系的紧密程度.

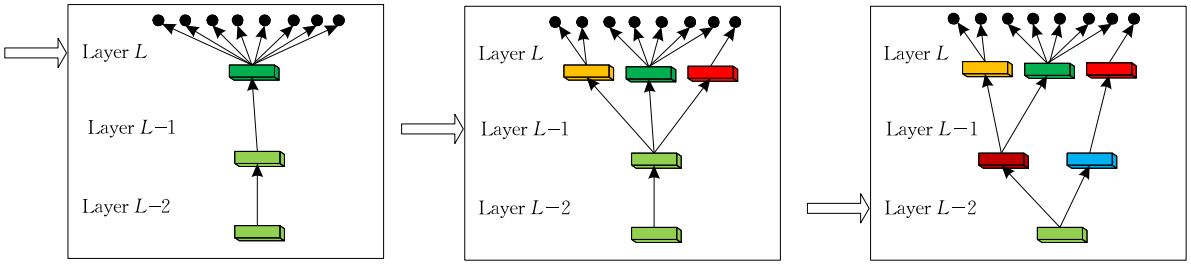


图 10　自适应层分堆网络

本节介绍了在深度多任务学习中几种典型结构，网络的复杂程度依次递增，学习的形式也由一般的层参数共享演变到具体层之间的连接和切换，由简单的整体相似性建模演变到共享和私有特征子空间的划分. 其中，软参数共享和硬参数共享在各个任务层结构之间是对称的，因此所有的层上特征参数都是在任务之间流动和共享的；张量网络是参数化表示的一种手段，通过学习一些层之间的张量表示，可以对各个任务网络层的差异性进行建模，但是在网络结构上和软硬参数共享是一致的，建立在任务层共享的基础上.

与上述三种网络不同，拆分缝合网络，水闸网络和自适应任务层分堆网络建立在改变层结构的基础上，划分了各自任务上共享和私有的特征. 拆分缝合网络和水闸网络是自适应层连接的两种表现形式，拆分缝合网络在两个任务网络对应的层之间设置了连接单元，控制特征参数信息的传递；水闸网络扩展了拆分缝合网络的结构，在每个任务网络输出的末端整合了所有的连接单元，完整地描述了层之间可能的信息传递；自适应网络分层是一种没有连接单元的学习方法，自顶向下逐层扩展，自动寻求各层之间可能的参数共享结构. 后三种方法都是通过改变网络结构细化了多任务学习的功能，区分了各个任务上的共享信息和私有信息，但是参数的增长规模和因此带来的运算代价是巨大的. 表 2 是深度多任务学习常见六种网络的比较.

表 2　深度多任务学习方法比较

深度多任务学习方法	改进类别	任务之间的关系描述	网络所使用参数的复杂程度	优缺点及意义
硬参数共享	训练算法的改进	基于任务层面，没有区分各自任务上的共享和私有特征子空间	低	实现了高效简洁的特征共享，但是人工判定的干预会削弱学习能力
软参数共享	训练算法的改进	基于任务层面，没有区分各自任务上的共享和私有特征子空间	高	扩展了硬参数共享中某些层之间信息不传递的缺陷
张量网络	训练算法的改进	基于任务层面，没有区分各自任务上的共享和私有特征子空间	高	将网络的设计问题转换为了参数表示问题，分离出的基张量对于新任务的适应性较强
自适应层连接拆分缝合网络	网络结构的改进	基于特征层面，区分了各自任务上的共享和私有特征子空间	高	通过层连接单元学习任务共享和特定表示的最佳组合，属于增量网络，减少了探索性
自适应层连接水闸网络	网络结构的改进	基于特征层面，区分了各自任务上的共享和私有特征子空间	高	从结构上扩展了拆分缝合网络，增加了连接单元之间的联系，但是有助于提升学习效果的特征仍然未知
自适应网络层分堆	网络结构的改进	基于特征层面，区分了各自任务上的共享和私有特征子空间	高	动态拓展每一层的共享结构，争取找到有效的学习组合，但是每层可伸缩性导致不容易收敛，计算难度也较大

2.7　非监督多任务学习算法

本节主要的算法都是在有监督学习的框架下讨论的，但实际应用中还存在一种情形是标签标注不足的无监督或半监督问题，因此本节讨论在此情形下的多任务学习算法.

早在文献[47]就设置了一个半监督二分类学习问题，通过将未标记数据的预测作为辅助任务引入，和目标任务中有标记的数据一起创建流形结构（或者叫图结构）. 作者假设数据分布是嵌入在低维流形中，可以在其上定义适当的平滑函数类，样本距离越接近，平滑度越高，分类效果也会趋于类似. 这种平滑结构可以通过结构学习来在低维预测空间中囊括，具体通过一个线性分类器实现，如下所示.

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\phi(x)+\boldsymbol{v}^{\mathrm{T}}\theta\varphi(x) \tag{31}$$

分类器中的 $\phi(x)$ 表示高维特征映射，$\theta\varphi(x)$ 是参数化的低维特征映射，通过 θ 可以约束所有任务

保持共享低维预测结构，最后使用权值 $\boldsymbol{w}$ 和 $\boldsymbol{v}$ 平衡这两个映射，作者指出引入辅助任务风险较小，即使任务不直接相关，分类预测器也会倾向于共享数据的相似平滑条件，因此也有发现潜在有益结构的可能性，这对有效利用无标记数据是至关重要的. 在此基础上，文献[119]采用此分类器提出了一种半监督学习框架下的多任务流型正则化方法，将二分类拓展到多标签，多个标签看作是多个相关任务，目标函数为

$$\min_{f\in\mathcal{H}_K}\frac{1}{N_m}\sum_{i=1}^{N_m}L(\boldsymbol{f}(x_{m,i}),y_{m,i})+\gamma_A\|\boldsymbol{f}\|_K^2+\gamma_A\|\boldsymbol{f}\|_I^2 \quad (32)$$

第三项 $\|\boldsymbol{f}\|_I^2$ 是流形正则化，用于进行流形空间内的平滑度约束，可以近似为：$1/(N_l+N_u)^2\times\boldsymbol{f}^{\mathrm{T}}\mathcal{L}\boldsymbol{f}$，$N_l$ 和 N_u 分别表示标记样本和未标记样本的个数，$\mathcal{L}$ 是拉普拉斯矩阵，$\boldsymbol{f}$ 是要学习的分类器：$\boldsymbol{f}=[f(x_1),\cdots,f(x_{N_l+N_u})]^{\mathrm{T}}$，分类器类似于式(31)的形式，但没有进行高维特征映射，只保留了第二项低维映射用于在流型结构中学习多标签的判别子空间. 流形正则化项保证了共享假设空间中的函数沿数据流形平滑，促进了未标记样本和标记样本的关系融合，对标记样本数量有限时是有帮助的.

针对无监督问题，文献[120]提出了一种域自适应多任务学习方法，通过衡量源域和目标域之间的散度以及利用目标域中未标记数据的内在结构来实现源域和目标域分类器的联合学习，目标函数包含四项：

$$\min_{f\in\mathcal{H}_K}\frac{1}{n_s}\sum_{i=1}^{n_s}L(f_s(x_i^s),y_i)+\gamma_I\|f_t\|_I^2+\gamma_A(\|f_s\|_K^2+\|f_t\|_K^2)+\gamma_M\boldsymbol{\Omega}(f_s,f_t) \quad (33)$$

作者将源域任务和目标域任务分别看作有监督的分类任务和无监督的聚类任务，因此目标函数中第一项代表源域中的经验损失，第二项是学习目标域中内在结构的流型正则化，第三项是正则化约束算法复杂度，第四项是约束源域和目标域之间散度足够小，通过考虑域之间的标签分布迁移，可以利用来自源域的类信息，将正确类标签分配给目标域.

无监督多任务学习也叫无监督多任务聚类算法，针对无监督任务，文献[121]提出一个新的多任务谱聚类模型，通过探索多个无监督任务的任务间相关性对它们进行聚类，同时学习预测样本和聚类标签之间的任务特定映射关系，具体实现是通过联合优化谱聚类和线性回归模型：

$$\min_{\boldsymbol{F}_m,\boldsymbol{W}_m} tr((\boldsymbol{F}_m)^{\mathrm{T}}\mathcal{L}_m\boldsymbol{F}_m)+\alpha\|\boldsymbol{F}_m-\boldsymbol{W}_m^{\mathrm{T}}x_m\|_F^2+\beta\boldsymbol{\Omega}(\boldsymbol{W}_m^{\mathrm{T}})$$
$$\text{s.t. } (\boldsymbol{F}_m)^{\mathrm{T}}\boldsymbol{F}_m=\boldsymbol{I} \quad (34)$$

其中 $\mathcal{L}_m$ 是关于第 m 个任务的拉普拉斯矩阵，第一项是关于利用谱聚类表示非监督任务之间的关系，并得到第 m 个所有样本的标签指示向量张成的矩阵 $\boldsymbol{F}_m$，然后使用 $\boldsymbol{F}_m$ 作为第二项经验损失中的标签学习各自任务的参数矩阵 $\boldsymbol{W}_m$，第三项正则化 $\boldsymbol{\Omega}(\boldsymbol{W}_m^{\mathrm{T}})$ 选取 2.3.2 节中的 $L_{2,1}$ 范数，用于筛选共同特征，这种方法的优点是谱聚类和特征选择的结合使学习聚类标签和各自任务映射的过程相互加强.

最近的文献[122]在文本分类问题中利用二分图和非负矩阵分解对多任务和多视图进行协同聚类，文献[123]提出了一种凸判别多任务特征聚类方法，利用高斯先验知识学习共享特征，进而对任务相关性进行建模，但它们是基于一个严格的假设，即所有任务具有相同的簇数，并且每个任务中的标签边际分布是均匀的，文献[124]中的方法学习由多个相关任务共享的子空间. 文献[125]中的算法学习核空间，其中相关任务的分布彼此接近，并且保留原始数据的几何结构，然而，它也假设相关任务共享同一组质心，并将所有任务的样本聚集在学习核空间中，这可能会干扰每个单个任务的标签边际分布.

以上多任务聚类方法假设任务之间的标签是相同的，是基于任务完全相关的理想假设. 然而将任务归类为完全相关或完全不相关太过于绝对，在许多实际应用中，任务通常是部分相关的，即任务之间只有部分标签是相同的，因此衍生出了基于局部相关的多任务聚类方法，它们适用于任务绝大部分数据分布相同或相似的情况. 在文献[126]提出的基于 Bregman 散度的多任务聚类(MBC)算法中，作者使用正则化项：$\frac{1}{M(M-1)}\sum_{m=1}^{M}\sum_{s=1,s\neq t}^{M}d(P^{(m)},P^{(s)})$ 约束源任务和目标任务之间的联系，$d(P^{(m)},P^{(s)})$ 代表任务之间的 Bregman 散度，使用此方法可以交替更新任务集群的中心并学习不同集群之间的关系，并且两部分相互促进，可以用于解决数据来自相同分布或相似分布时的多任务聚类问题，而文献[127-128]观察到在处理差异明显的任务之间往往会出现一些无法共享的数据点干扰聚类的进行，这种互相促进的提升模式算法容易很带来负迁移，作者提出了智能多任务 Bregman 聚类方法(S-MBC)用以识别和避免这种负迁移，在迭代过程中划分了不加任务关系约束的单任务 Bregman 聚类和施加关系约束的多任务 Bregman 聚类的两种情形，分别计算聚类的局部损失并且根据两者的较小值决定是否采用 Bregman 散度约束任务关系，进而避免负样本数据带来的影响，紧接着作者将其扩展到智能多任

务核聚类方法(S-MKC)，用以处理非线性可分数据.

不同于任务之间的分布约束，文献[128]提出了一种传递实例知识的多任务聚类方法，该方法引入任务间偏差，使用其它任务的相关样本进行重新加权来参与单个任务的聚类，并使用约束低秩分解并施加非负矩阵约束来维持每个任务的标签边际分布，然而此方法并没有明确考虑如何通过学习相关任务的信息来设置任务间偏差，并且只能处理二进制数据聚类问题. 文献[129]进一步使用共享子空间中的最近邻样本作为辅助数据来帮助单个任务进行聚类，作者指出由于传统聚类方法的基本假设是样本来自相同的分布，由于任务的分布不同，在原始空间中计算的不同任务中的任意两个样本之间的距离不能代表这两个样本的真实关系，即距离较小的不同任务中的样本可能不相关，因此不能在原始空间直接计算跨任务样本之间的最近邻，简单地将其它任务的相关样本作为辅助信息也可能会干扰原有任务的标签边际分布. 具体地，作者针对跨任务信息辅助单任务进行聚类问题同时使用两种最近邻的度量：在原始空间中使用最大均值差异(MMD)方法来使得相关任务的分布彼此接近，学习一个用于关联任务之间信息的共享子空间；此外将传统的维数约简引入到多任务设置中，在共享子空间中采用类似于拉普拉斯特征映射(LE)的方法保持原始数据的流形结构和特性，用于作为辅助数据对跨任务的信息进行扩展. 因此学习共享子空间的目标函数为

$$\min_{\boldsymbol{M}} \lambda tr(\boldsymbol{M}^{\mathrm{T}} XR(MX)^{\mathrm{T}}) + (1-\lambda) tr(\boldsymbol{M}^{\mathrm{T}} XL(MX)^{\mathrm{T}})$$
$$\text{s. t. } \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} = \boldsymbol{I} \tag{35}$$

其中，$\boldsymbol{M}$ 是子空间的正交基，MX 作为对原始特征 X 的子空间映射函数，式(35)上一行前项匹配与目标任务分布相关的任务，是最大均值差异的目标函数，后项是拉普拉斯特征映射目标函数，用以保留原有数据特性，通过在子空间中欧几里德距离最小找到原空间相邻的样本点. 然后，通过联合优化两组最近邻来学习每个任务中每对样本之间的共享最近邻(SNN)相似度，并应用传统的谱聚类方法对每个任务分别进行聚类.

近来对于部分相关的任务，文献[130-131]将其延伸到更一般的情况，能够自动识别任务间的关联实例并进行迁移，从而避免了任务部分关联时的负迁移. 文献[130]提出了自适应多任务聚类方法(SAMTC)，通过利用任务间聚类簇的正相关关系，对子任务进行重构. 具体步骤是使用 Jensen-Shannon 散度 $JSD(\boldsymbol{P}(X_i^t), \boldsymbol{P}(X_j^s))$ 衡量目标任务 X^t 中的第 i 个聚类簇和源任务 X^s 中的第 j 个聚类簇之间分布的差异，作用的效果是 SAMTC 学习到了在源任务中能被目标任务利用的相关实例，然后利用这些相关实例为每个目标任务构建相似矩阵进行谱聚类得到最终的结果.

在此基础上，文献[131]使用另一种方式从源任务中学习相关实例，进一步延伸出了基于流形正则化编码的多任务聚类方法(MRCMTC)，由于源任务和目标任务通常具有任务特定的特征，从而造成了相关数据点不能足够接近，而且这种域差异不利于使用源任务中的实例表示目标任务中的实例，因此，算法考虑交替地学习一个可以减少源任务和目标任务之间域散度的低维特征空间，并且将可以表示目标任务数据点的源任务实例进行编码，然后再利用谱聚类得到任务的聚类结果. 文献[130-131]为部分相关的任务开发一个更通用的多任务聚类框架，并且在实际中获得了较好的结果.

可以看到，无监督学习问题主要研究任务之间的聚类散度而半监督学习问题主要侧重于学习标记数据和无标记数据之间的流形结构，共同点是依赖结构学习来建立底层共享空间的固有联系，改善标记数据预测结构，因此，在非监督学习中对任务相关性的建模是至关重要的. 非监督任务优点突出，能提高数据使用效率并且提高标记数据有限情况下的预测效果，目前对非监督学习的研究比较少，部分相关的多任务聚类问题是值得考虑的问题，可以作为以后的研究方向.

3　多任务学习有效性的解释

3.1　多任务的作用机制

虽然多任务学习在很多场景下比单任务学习效果有所提升，但是其内部影响因素值得我们探究：与其它任务共同学习，单个任务学习效果提升的原因；适合共享学习的任务类型以及它们之间的联系；多任务中可以迁移的特征的特点以及多任务数据概率分布之间的关系等. 以往这些问题的解决主要依赖经验结果，一些研究也对多任务学习在不同条件下的有效性做了初步研究. 本节从将从多任务的工作机制、相关性判定，以及多任务的不同功能分类阐述多任务学习算法的有效性.

3.1.1　噪声平均

多个相关任务共同学习时，其它任务中包含与该任务相关的特征或子结构，但也有可能包含不相关的特征或子结构. 在文献[1]中作者指出在指定多任务学习过程中，不相关的部分相当于引入了更多

的随机噪声,而随机噪声的干扰能提高系统的稳定性,这是多任务本身特有的对抗特性带来的影响.

还有一方面原因是来自各任务数据本身的影响,由于各任务的训练信号中都存在独立的噪声,但是如果能学习到任务之间的共享特征,即辨别出了相似任务之间的不变的那部分,多任务的输出就等于是使用同样的真实信号加上独立噪声以后的效果[11].其它任务的加入,等于增加了这种原始真实数据样本个数,多个任务同时学习为平均噪声提供了更多的参考信号.

因此,适当地选取辅助任务可以实现平均噪声,提高主任务学习模型的泛化性能.

3.1.2 统计数据增广

多任务学习有效性可以从任务的数据分布解释,多任务学习的最终目标是学习多个任务上训练样本的概率分布迁移机制,从而提高分类器或者回归估计算法的泛化能力.

在多任务学习中,假设大部分任务的数据分布要满足总体相似条件并且假定任务之间包含一个共享空间,在联合学习中,它能间接地利用每个任务中都包含的那部分域共享信息,提高对隐藏在同一信息域中噪声的辨别能力.如果每个任务都可以学习到共享结构,那么在所有任务上关于这部分共享结构的样本都可用来帮助单个任务学习模型,所以利用这些相似数据学习任务之间不变子结构等于隐式增加了训练数据样本个数,间接扩大了样本空间,达到了统计数据增广的目的.

对于只能提供有限数量的训练数据的单任务学习场景,单任务学习中出现的秩亏和泛化能力不强问题,可以借助相关任务训练样本中的额外信息,学习多任务共享部分等于特征选择,从而达到了正则化的效果,因此通过多任务学习有助于最小化过拟合风险和提升泛化能力.

针对不同类型多任务数据对多任务学习算法有效性的影响,一些文献[132-133]将多任务的学习性能与数据属性进行比较,指出多任务学习与辅助任务标签分布的均匀程度有关,标签数量较少且结构相对紧凑的辅助任务有助于提供有效信息,从而帮助其它任务学习,但是各个任务样本容量的大小并不是主导因素,这种观点印证了文献[134]中的结论:每个任务需要多少信息来学习,而不是每个任务需要多少样本来学习,引出了多任务学习假设空间大小(归纳偏置)选取的问题.

3.1.3 表示偏置和特征选择

机器学习中的一个主要问题是归纳偏置,当只学习单一任务的时候,提取的偏置仅仅是针对该任务的,因此需要大量的训练样本提升学习模型的可靠性,多任务学习模型的核心假设是多个学习任务嵌入在相关学习任务的上下文环境中.在这样的环境中,学习多个任务的归纳偏置可以通过在不同任务上抽样来完成,从而放松了在单个任务上样本采样的负担.同时,它可以搜索某个假设空间,其中包含大多数环境中任务的良好解.因此,相关任务环境中同时学习多个任务可能比学习单个任务有更好的泛化能力.

当学习了足够多的训练任务后,学习模型可以适应来自相同环境下的新任务,也就是元泛化的概念,多任务的表示偏置机制使得任务能扩展到元泛化的形式,如果在足够多的训练任务下,学习到共享表示,任务之间不同的偏置通过学习任务之间的相似性得到缩减,以此产生的假设空间会减小.文献[135]证明了在对学习者可用的所有假设空间有一定限制下的情况下,在大量训练任务中表现良好的假设空间在同一环境中学习新任务时也能起到很好的作用.总的说来,多任务学习的优势依赖于合适的假设空间,既要包含所有任务的良好解,又要确保从有限大小的任务集中学习的模型有可靠的泛化能力.

文献[87]基于偏置学习的一般理论提出了多任务模型通过共享共同的最优假设类自动学习归纳偏置的概念,作者通过对假设类的 VC 维数和覆盖数的分析,给出了泛化界限.随后许多理论证明是利用特定的归纳偏置进行泛化误差上界推导,文献[54,132,136]用适当的公共线性算子来说明多任务学习的优点,通过利用 Rademacher 复杂度推导了泛化误差上界,用来度量假设空间相对于单任务学习的复杂性,文献[137]证明了迹范数约束的低秩子空间方法的超额风险边界,文献[138]根据一致稳定性给出了泛化误差界具体推导,并将其推广到了多任务经常采用的范数约束.作者首先给出了多任务环境下的一致稳定性的定义:

$$\left|\sum_{m=1}^{M}\left(L(f(x_m),y_m)-L(f(x_{m,-i}),y_{m,-i})\right)\right|\leq\tau \tag{36}$$

在一致稳定性基础上分别推导了一般多任务设置下的泛化误差界:

$$bound\leq\sum_{m=1}^{M}\frac{1}{n_s}\sum_{i=1}^{n_s}L(f_s(x_i^s),y_i)+2\tau+(4n_0\tau+MU)\sqrt{\frac{\ln(1/\delta)}{2n_0}} \tag{37}$$

其中 δ 和 n_0 是自选系数,U 是经验损失函数上边界,类似地,将式(37)中任务个数设置为 1 可以得到单

任务的一致稳定性，此时单任务泛化误差界为

$$bound \leq \sum_{m=1}^{M}\frac{1}{n_s}\sum_{i=1}^{n_s}L(f_s(x_i^s),y_i)+2\tau+(4n_0\sum_{m=1}^{M}\tau_m+MU)\sqrt{\frac{\ln(1/\delta)}{2n_0}} \quad (38)$$

式(37)和式(38)的差别在于不等式右边第一项经验风险和第三项稳定系数.得益于3.1.3节中介绍的数据增广机制，存在一些分布相似的数据为相关任务提供了更多的训练样本，可以假设第一项多任务经验风险更小，这样多任务约束和单任务约束的优劣可以通过比较第三项中单任务稳定系数之和与多任务稳定系数的大小决定，这项工作给出了一个是否采用多任务学习的判断标准.

最优的假设类为特征选择提供了可能，因为各种多任务学习方法都是假设整体结构或者部分结构有一定的关联关系，所以多任务学习方法的特征选择与假设空间的好坏有直接关系，文献[47]通过使用与文献[87]不同的覆盖数，指出当任务数 M 较大时可以可靠地选择共享参数作为共同的假设空间，降低了假设空间的复杂度.文献[18]进一步证明了在使用组稀疏的特征选择方法中，当维度 D 增长倍率小于与任务数量 M 有关的增长倍率 $\exp(\sqrt{M})$ 时，维度大小就不足以影响特征选择的结果，并且作者通过稀疏圆不等式(sparsity oracle inequalities)推导了严格的泛化误差上界，证明了特征选择的有效性.

3.1.4　量化平滑与提高动态优化特性

任务之间在共同学习过程中可以改变权值更新的动态特性，可能使网络更适合多任务学习，比如在神经网络中，多任务并行学习提升了浅共享层的学习率，可能较大程度上提升了学习的效果.单任务学习时，梯度的反向传播倾向于陷入局部最小值，往往单任务单独学习时有多个局部最小值，通过多个任务交互作用，可以帮助单个任务学习过程规避鞍点和局部最小点.一些文献对于动态特性改变的因素进行了探究，例如文献[139-140]验证了标签量化程度的不同对优化过程的影响.量化主要用于描述任务输出标签的数据特性且平滑程度根据标签离散程度划分，例如在分类任务中，分类的种类越粗略，量化的程度越低，平滑程度越高，平滑程度不同对于学习过程的影响不一，在文献[139]中提出具有较少量化的平滑任务在学习过程中更为容易，因为任务一般采用的目标函数都是连续的，更少量化的任务可以让输入和输出之间的关系更加具体，因此连续性的输出标签会更符合目标函数优化中的要求.

由于大部分任务使用的是离散类型的训练数据，文献[140]中指出如果有额外的训练信号比主要任务量化程度小，或者采用不同的量化级别，这些量化的训练信号可以作为额外任务提升主任务的学习效果.因为随着具有不同量化水平的任务加入，在主任务上更加容易地插入其它任务的粗略量化风险，每个任务都可以填补主要任务量化平滑所产生的空白，从而提升训练效果.

3.1.5　窃听机制

不同任务与特征的交互方式不同，一些任务很容易使用的特征或许在其它任务中不容易被学习到，因为各个任务对特征的放大程度不同，在其它任务上这些特征并不占有主导因素，占有主导因素的特征阻碍了它的学习，但是利用非主导特征可能会提升任务的效果，可以在共同学习的过程中能将这部分特征添加到共享结构中，通过交互方式简单的学习这些特征的信息，或者说是间接利用了另外任务中的有益特征.在多任务中这种借鉴其它任务学习特征的模式称之为窃听机制，最简单的方式是在文献[1]中，通过训练多任务模型直接预测一些重要特征，单任务学习中，这些特征并不太有重要性，这种情况在半监督学习中有更多地显现，标记数据由于信息不充分，需要与大量未标记数据投影到流形空间进行讨论，而流形空间通过等低维空间投影降维技术已改变了原有的特征结构，因此低维空间内的特征可能代表了原有空间不同特征之间的联系[47,119]，实现了间接地信息共享.

3.2　多任务的相关性判定

多任务学习之所以有效，是因为它是建立在假设任务之间含有相关信息，多个任务具有一些共享表示的基础之上.算法适用性的关键是判断任务之间的相关性，无关联的任务带来的干扰信息使其它任务的学习产生负迁移.虽然任务的相关性可以通过经验理论得到，但是一些研究者更倾向于通过理论系统解释多任务的相关性：在文献[87]中共同地学习任务之间的相似性仅仅通过模型选择的标准来体现，多任务学习的优势依赖于假设多个任务共享一个共同的最优假设类.文献[132]为相关任务学习的相似性问题提供了一个系统的数学模型，定义通过数据的产生机制描述任务的相关性.提出若两个任务中数据的概率分布都产生自同一类变换，那么两个任务是相关的.

作者假设各个任务的样本包含在同一给定的域中，假定在 $X\times\{0,1\}$ 上存在一个假设的概率分布 P，机器学习的本质就是发现一个假设 $h:x\rightarrow\{0,1\}$

去尽量接近 P,由于多任务学习是多个相似分布问题,不同任务的映射关系符合不同的分布,从不同的概率分布 $P_1,\cdots,P_n$ 中,抽取样本集 $S_1,\cdots,S_n$,给出了分布函数 F 相关的定义:

***F* 相关**. 如果 F 是一组映射 $f:X\to X$,P_1,P_2 是 $X\times\{0,1\}$ 上的两个概率分布,如果存在一个概率分布 $f\in F$ 对于任何 $T\in X\times\{0,1\}$,T 是 P_1 可测的当且仅当 $f[T]=\{(f(x),b)|(x,b)\in T\}$ 是 P_2 可测的,而且有 $P_1(T)=P_2(f[T])$,此时说 P_1,P_2 是 F 相关的. 提出若两个任务中的数据都产生自由同一类变换 F 得到的固定的概率分布,那么两个任务是 F 相关的.

当学习 F 相关的任务集时,定义一个假设空间 H 作为分布的最佳预测,假设每个能拟合各自任务概率分布 P 的最优假设空间为 H,在假设空间 H 上建立了一个由假设集合构成的假设空间,这相当于 F 的变换,H 在 f 的作用下为闭集. 基于最优假设空间能获得多个任务的泛化误差界,通过这个泛化误差界能够计算每个任务上的达到规定的置信度所需要的样本个数,并确定学习模型的泛化能力.

F 相关定义的优点是,这种相关性定义扩充了对函数映射集的要求,这些映射使得任务的关联关系变成了任意的形式,并不只局限于双射映射关系,甚至可以允许任务之间的实际转换是在已知的关系集合上近似得到.

但是缺点也很明显,此 F 相关定义的多任务关系很局限,通常只能针对处理属于同一分类问题的任务,不能用于处理不同问题的任务. 例如,考虑一组不同视角的相机,虽然很难确定它们各自的偏置,但是通过同一型号传感器收集的图像数据概率分布都应该是 F 相关的,显然这种情况下,任务相关性问题只是在相对的特定假设空间族内有意义.

文献[86]将这个假设族类型进一步丰富,作者更倾向于从个体任务向量的角度考虑相似性,而不是产生自同一个分布族;通过任务的权重参数向量的相近程度判断任务的相关性.

但是,相关性不能单单只用相似的权重来度量,这种只有当权重相近时才能确定任务是有联系的,使用相似特征的决策方法并不严谨,例如一些分类任务所使用的特征很可能是高度相似的,它们都是根据相同的特征来做决策,但是注意每个任务的分类结果不是相同的,也就是说特征在各自任务上权重并不相同,任务的相关性不能仅仅由相似性定义,这样会混淆具体的数据属性,只有每个特征的权重都足够相近的时候,才有理由将它们表示为高度相关. 对于不能整体近似的方法,深度学习提出了松散任务关系的概念,可以在网络中使用开关单元控制层之间参数和数据信息的传递[117],深度多任务学习方法主要解决了数据分布对任务关系的影响,例如在数据特性中发现标签数量适中并且分布紧凑的任务对提高其它任务的预测效果有辅助作用,在优化求解过程中,具有快速下降动态特性的任务能帮助其它任务避免局部最小点,提高学习效果[139-140].

4 多任务学习应用

多任务学习在计算机视觉、自然语言处理和语音识别等领域有着很好的应用前景,尤其得益于深度神经网络的发展,其强大的特征学习能力也对多个任务进行联合建模过程中的特征选取起到了辅助作用,一些领域中不同任务的复杂特征通过网络层共享建立了联系,通过利用其它任务的有用信息摆脱了外部信息的依赖,各个任务之间实现了需求互补. 由于深度网络的应用中涵盖了大量的背景知识而且结构复杂多变,所以本节按照应用领域首先介绍深度多任务学习的应用,紧接着介绍结构化多任务学习方法的应用.

4.1 深度多任务算法的应用

4.1.1 深度计算机视觉

在计算机视觉的图像处理领域,多任务深度学习已经广泛应用于人脸识别、细粒度车辆分类、面部关键点定位与属性分类等多个领域. 总体来说计算机视觉任务主要有五种:目标检测(detection),图像识别(recognition),物体定位(locallization),图像分类(classification),目标分割(semantic segmentation).

图标检测注重在图像中目标的搜索,目标需要有一定的形状或轮廓;图像识别寻找目标的特征;目标定位寻找目标物体在图像中的位置,也称为跟踪;图像分类是对图像进行全局描述然后给图像分配标签[101,141-142],与之不同的是图像目标分割,目标分割是对图像中的每个像素添加标签,具有相同标签的像素具有某种共同视觉特性.

目标分割、目标检测、目标识别、目标定位以及跟踪顺序为视觉场景理解中的一般流程.

视觉场景理解是机器自动分割并识别出图像中的内容,有两个核心问题:学习语义和目标的几何表示. 语义学习中包含语义分割和实例分割(最小边界框检测),语义分割赋予图像中每个像素一个语义标签,语义标签能在场景中分割语义连接的区域,将属于不同场景的对象分开,但是不区分单个类中的对

象实例，更具体地，实例分割不但要进行像素级别的分类，还需在具体类别基础上区分具体对象的实例，并通过包络框来描述它们.同时，如果要学习更困难的任务实例分割，还需要用到目标物体的空间几何关系，只有正确识别出图像中所有物体的方向，才能将不同物体精准区分开，一般通过表面法线估计，单目深度估计，或表面法线方向矢量坐标等方法[143]获取.

以往，这些任务都是利用各自的深层卷积网络模型单独学习出来的，如图 11 所示，一些近来的研究[144-147]将语义分割、实例分割和深度估计纳入了同一个多任务结构来创建场景理解系统，使用单目图像获得一个整体场景理解包括获得语义标签和实例分割，最后为每个实例估计三维深度，具体是通过共享卷积层作为编码器层，对图像的像素特征进行处理和提取，再为每个任务执行独立的解码层，拓展成特定于任务的输出通道以便准确地进行分割和检测.

图 11　多任务的实例分割、单目深度估计和语义分割学习

多任务学习在视觉理解中的大部分场景都有涉及，主要可以满足场景理解对于时效性的要求，只应用标准的视觉技术而不需要复杂的图形模型处理就能直接无间隔输出，具体做法是将像素语义与几何特征获取组合到一个通道中，通过定义一个联合交叉熵损失函数，来自不同任务的像素离散信号提供了不同的表示线索，通过共同训练提高了视觉图像处理的平滑性和准确性，为场景语义和几何坐标关系提供了更丰富的解释，使多个场景理解任务都能获得良好的性能.另外，一些研究者将语义分割、图像分类和目标检测等任务并行学习，省去目标检测使用卷积网络建议评分的步骤.

但是并不是所有的计算机视觉任务之间都能够相辅相成，大部分分类问题和目标分割问题都是要获取目标对象的不变表示来区分类别，诸如一些姿态估计和定位，则需要保留各自对象的几何和视觉特征[111,148-149]，两者的处理过程是互相矛盾的，但是此时也可将两者的学习过程联合起来，因为多任务联合卷积网络浅层的特征表示往往不具有特定性，所以在姿态估计或者对象定位任务中利用的是浅层特征包含的信息.

4.1.2　自然语言处理

虽然多任务学习在图像处理问题中取得了进展，但并不能在自然语言处理(NLP)中直接应用.图像信号通常是连续的，图像语义之间不会发生明显的跳变，与图像相比，自然语言标记是离散的：每个词都很好地反映了人类的思想，但相邻的词并不像图像中的像素那样共享多少信息，在自然语言处理中，神经网络是否可转移特征，很大程度上取决于任务的语义相似程度，这与图像处理中的特点不同.

序列标注是自然语言处理中的一个基本问题，属于深层次的语言语义学习任务，给定一系列单词，序列标记旨在预测每个单词的语言标签，大致的应用包括词性标注(Part-of-Speech tagging)、分块(chunking)、命名实体识别(Named Entity Recognition)、文本分类、浅层语法分析等.

在很多自然语言序列标注任务中，有助于训练过程的标签非常少，主要依赖人工标注的语义特征，比如需要形态信息和词性标签等，对特征的标记工作量庞大同时也不一定准确.而且在自然语言中深层次的语法语义分析通常是以词为单位，尤其是在中文中，需要把连续的汉字分割成语言学中有意义的词.分词是中文自然语言处理的上游任务，通常作为需要获取深层词语言语义序列标注任务的预处理过程.

以往的分词系统最常见的是词典匹配，但是由于依赖于词汇表的容量，很多词语识别的精度不够，因此有研究工作[104,150]开始将分词和其它更复杂序列标注任务一起建模.考虑到自然语言一个词句中单词属性容易被周围的单词影响，可以利用上下文信息为序列标注提供线索，利用这个特点，衍生出了多任务框架来解决多个序列标记问题[151-152]，除了对每个单词分配标签之外，加入预测周围单词的分词任务，激励系统学习通用的语义表征和句法结构，提高对不同序列标注任务的准确性，相当于在不需引入外界信息或者额外注释的情况下能为序列标注

训练提供额外信息，摆脱了对词典匹配中词汇表的依赖.文献[10]在命名体识别系统中将生僻字名称检测任务与语句层次的特征检测任务相结合，提高生僻字的识别效率；文献[153-154]引入辅助任务学习上下文信息为名称检测提供线索，通过一个双向LSTM同时训练根据前进和后向预测每个单词最可能的标签，文献[155-156]将中文分词任务与命名体任务共同训练，改善了中文媒体命名体识别的效果.此类方法也可推广到其它序列标记模型中，对文本分类、情感检测等场景[157-159]也适用.

联合建模的一大好处是分词系统可以与其它任务共享有用的信息[160]，分词的时候也会考虑到其它任务的要求，最新的分词系统能够预测边界信息表示，有助于为深层次的任务提供更丰富的信息，其它任务也会考虑各种分词的可能性，全局上可以取得最优解.但是，从一个域自动学习词汇上下文信息可能在另一个域中是无用的，上述对字符或者单词级别建模，必须保证数据来源的一致性，否则需要考虑模型的域适配问题.

跨语言学习是一种典型的域失配问题，通常，使用词对齐或平行语料库单独为不同语言提供单词级别的语义信息.文献[161]证明了多任务框架应用到跨语言处理的可行性，基于字符建模的多语种处理结构能够捕捉语言之间的相似形态，以统一的方式处理跨语言任务，在任务之间直接通过共享字符和单词级参数来学习语言特定的规律.在字符层面可以设计简单的网络共享层共享字符参数，因为在同一语言中，不同的序列标记任务共享语法规律，同时有些语言共享字符集形态，构成了任务之间的底层相似性.而在字符和单词上使用开关递归单元来编码上下文信息，实现字符和单词之间特征共享和分离，这样就避免了使用平行语料库和词对齐的繁琐过程，同时也提高了多语种问题序列标注的效果.

4.1.3 语音识别

LSTM递归神经网络在语音增强(Speech Enhancement，“SE”)和自动语音识(Automatic Speech Recognition，“ASR”)等方面取得了突出的性能.然而，噪声语音识别的一个问题是，语音增强和语音识别的标准并不相同.为了进一步提高语音识别中抗噪性能，相关研究[162-163]整合了这两个系统，提出了一种语音增强和语音识别相结合的LSTM网络架构，在一个统一的目标下，考虑语音信号的质量和语音识别的准确性，首先利用语音增强系统对含噪声的语音进行信号增强，然后利用增强语音对语音识别器进行训练，反过来SE使用ASR信息进一步增强去噪.

4.2 结构化算法的应用

4.2.1 生物医学

在生物医学方面，文献[164]找出不同数据模型参数之间的共性，研究了跨不同生物的蛋白质定位预测问题，针对基因的高度多态性，文献[165]通过度量等位基因之间的相似度并通过共享信息改进基因预测的性能，文献[93]利用基因表达性状之间的相关性和有关分子多态性的先验知识筛选基因对进行检测.针对疾病诊断研究中的高特征维、小样本问题，在临床应用中通常优先选用特征选择方法，因为可以方便地从神经影像学资料中获得并且易于解释，文献[166-168]通过结合多个临床认知测量之间的内在联系，使用了稀疏组套索在所有时间点从核磁影像中选择一组鉴别特征子集作为临床评分来预测疾病的进展情况，文献[169]将成像疾病预测问题扩展到非线性核函数，并且使用了L_p范数进行特征选择，此后，文献[170]利用子空间学习方法的特点，从原始特征集中选择类别判别和抗噪声特征应用在疾病影像学分类中.近来，文献[171]提出鲁棒低秩稀疏回归方法用来同时选择神经影像学标志物和与阿兹海默症认知测量有关的内在变量进行疾病预测，文献[172]结合了局部推理和全局推理，使用低秩稀疏分解去除由蛋白质接触预测中产生的残基耦合问题，文献[173]应用稀疏低秩分解对基因阵列数据高缺失的缺陷进行了补全.

4.2.2 目标跟踪和图像分类

在粒子滤波目标跟踪框架下，将低秩子空间和多任务学习引入到目标跟踪中，通过多任务学习挖掘粒子间的自相似性可以提高目标跟踪的速度和性能，文献[143,174]分别提出了联合稀疏多任务跟踪器和低秩稀疏跟踪器，将所有粒子仅用少量粒子表示，挖掘所有粒子间的相似性用于提高跟踪效果，文献[175]指出从一个大区域考虑细粒度取样时，有些粒子可能与其它粒子不同，使用联合稀疏或者低秩来共享相同结构可能会降低效果，它的做法是通过将稀疏系数分解为内点和离群点，并基于文献[58]的鲁棒分解理论提出了MTT(RMTT)目标跟踪，利用粒子的相似性进行联合稀疏，并且考虑粒子差异性进行离群粒子的稀疏；文献[176]将目标跟踪问题看作一个基于联合稀疏特征低秩表示的多任务特征学习过程.首先在多个初始帧中选择具有低秩表示的特征以获得子空间基.接下来使用改进的基于联合稀疏性的多任务特征学习框架来学习由低秩和稀疏属性表示的特征.文献[177]提出了一种具有结

构化和加权低秩正则化的多任务跟踪器，通过引入加权核范数，对多个子任务的不同秩分量自适应分配不同的跟踪重要性，将其组合来完成总跟踪任务.文献[178]提出了一种多任务联合稀疏表示和分类的正则化多任务学习(MTL)框架，用于视觉识别.文献[179]使用多个属性特征描述每个超像素点，将多个特征空间合并成一个亲和矩阵，然后使用稀疏和低秩分量在此亲和矩阵上寻求一致的图像分割算法.

文献[57,180-181]应用到了高分辨率卫星的遥感图像分类问题，文献[57]使用核方法将数据投影到非线性子空间处理二维图像的分类问题.文献[180-181]从面向像素的全局性方法发展到面向对象的方法，同时考虑全局像素和当前像素，均匀地对像素进行聚类，文献[180]针对提高分辨率造成分类方差过大或过小引起的光谱不确定问题，提出了一种联合稀疏和低秩结构的特征表示方法，采用 $L_{2,1}$ 范数正则化来增强模型参数中的组稀疏性，识别出图像分类的有效判别特征，同时由于卫星场景类同时具有多个特征依赖，应用迹范数低秩约束减少在分类任务中的冗余和相关性，基于同样的结构，文献[181]提出了一种带拉普拉斯类正则化的稀疏低秩联合 MTL 方法，使用拉普拉斯类正则化高效利用类标签信息，并且有别于文献[57]的二维目标，作者对三维图像形态轮廓进行特异性和相关性建模.

4.2.3　社交网络中的应用

在社交网络中亦有结构化算法的应用，文献[182]提出了一种分布式用户-服务器架构，使用核方法和正则化将多个用户端任务的信息进行融合提高访问效率，文献[183]使用多个标签进行训练，结合了多个情感分析和主题分类任务，文献[184]在社交网络中采用多时间段的文本，选择在线行为特征对用户心理进行推断，文献[185]对多个用户数据提取关键特征构成多用户推荐系统，文献[186]为多链路网络数据学习一个公共度量子空间，然后对每个链路进行独立编码提高传输效率，文献[187]利用贝叶斯层次结构对主题模型进行协同聚类.

5　实验比较

本节使用几种比较先进的多任务学习算法那与各自领域方法的对比实验.这几种比较有代表性的方法分别是基于特征层面的判别式典型算法可变簇聚类方法(Flex-clus)、基于特征层面的生成式典型算法联合稀疏低秩的概率方法(Probabilistic Low-Rank)，以及深度网络中的级联多任务语义分割算法.

首先介绍所使用的数据集.一般来说，如果要选用多任务方法学习模型，数据集的特点是任务之间既要存在相关性，又要存在单任务表述不了的偏置，尤其是在各个任务中一些特征的表达上存在差别.本节使用的多任务模型数据集的类型大概分为以下两种：第一种是回归问题，使用的数据集为成绩预测 school data 数据集(https://github.com/tjanez/PyMTL/tree/master/data/school)，包括 139 个学校 15362 个学生的成绩，每个学生包含 27 个特征属性，通常将每个学校的成绩预测作为不同任务，并且每个任务中的样本个数是不相同的，甚至有的任务中样本数量接近于特征个数，使用单任务回归容易产生过拟合，所以在以前的方法[31,61,73]中经常被用来作为多任务的数据集.

第二种是分类问题，可以利用多任务学习将多分类问题的数据集转换为二分类问题，本节使用了四个分类数据集，前两个是手写体识别数据集 MNIST(http://yann.lecun.com/exdb/mnist/)和 USPS(https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html)，为数字“0”到“9”的分类任务，由于图像差异可将一个十分类任务设置为十个二分类任务.MNIST 由 70000 个 28×28 的灰度图像构成，共 60000 个训练样本和 10000 个测试样本；USPS 中有 16×16 像素的灰度图像，共有 9298 个手写数字图像，其中 7291 个训练样本，2007 个测试样本，两个数据集的像素值均被归一化.第三个是 The extended Yale Face B 数据集，该数据集有 16128 张图像，共包含 28 个人的 9 种不同姿态的正面人脸，每种姿态又包含 64 种不同的光照情况，将图像大小归一化为 32×32，将每个人作为不同的分类任务.第四个是语义分割数据集(http://host.robots.ox.ac.uk/pascal/VOC/voc2012/VOCtrainval_11-May-2012.tar)PASCAL VOC 2012，包括 20 个对象类，共有 11530 个图片包括 27450 个标注物体和 6929 个语义分割.

5.1　判别式方法实验

本节对比了几种主要判别式方法的实验结果，其中包括单任务岭回归[188](Ridge)、正则化约束多任务方法[30](Regularized MTL)、脏模型[59](Dirty model MTL)、鲁棒多任务学习[61](Robust MTL)、稀疏鲁棒多任务学习[60](Sparse-Robust MTL)、聚类多任务模型[39](Clustered MTL)、利用协方差学习多任务关系的生成式模型[84](MTRL)、特征学习方法中的可变簇聚类模型[65](Flex-clus).这些是多

任务学习中最主要的几种学习方法，基本涵盖了不同类型的结构，Rigde 是单任务学习的基准方法，Regularized MTL 是任务层面的特征选择方法，此外，Robust MTL、Sparse-Robust MTL、dirty MTL 和 Flex-Clus 是特征层面的叠加模型；Cluster MTL 和 Flex-cluster 分别是任务层面聚类和特征层面聚类算法的代表.

回归问题实验参数的设置是从 school data 数据集的 139 项任务中各自随机挑选 10%，20% 和 30% 的样本作为训练集，另外 45% 用于测试，剩下的用于验证，为了减少统计变化带来的影响，结果重复 5 次取平均值，每次迭代的最大次数设为 5000，误差的容许度范围为 10^{-5}，评价指标选用归一化均方误差（$nMSE$），回归模型的 $nMSE$ 越低，空间和时间复杂度越小，效果也越好. 实验结果如表 3 所示，随着训练样本的增多，Ridge 效果仍然比较差因为每个任务样本容量过小甚至小于特征个数，容易产生过拟合的奉献，Regularized MTL 表现稳定是因为实际上此数据集的特征结构是一致的[11]，因此基于聚类思想的 Cluster MTL 方法和基于多种相关性的 MTRL 方法产生了负迁移，而基于特征层面的其它方法中只有 Flex-clus 方法捕捉了特征的联系，因为只有它对同一维特征施加了横向约束，正好满足此数据集各任务对应特征高度相似的特点.

表 3 回归数据集的均方根误差

方法	训练样本比 10%	训练样本比 20%	训练样本比 30%
岭回归	1.047±0.023[8]	0.908±0.015[8]	0.867±0.023[8]
正则化约束多任务方法	0.871±0.024[1]	0.784±0.019[2]	0.773±0.026[1]
脏模型	0.965±0.026[7]	0.842±0.017[7]	0.811±0.025[7]
鲁棒多任务学习	0.964±0.016[5]	0.820±0.008[3]	0.790±0.021[3]
稀疏鲁棒多任务学习	0.965±0.016[6]	0.820±0.008[4]	0.790±0.021[4]
聚类多任务模型	0.950±0.011[3]	0.820±0.011[5]	0.792±0.019[5]
协方差多任务关系模型	0.955±0.013[4]	0.823±0.009[6]	0.793±0.015[6]
可变簇聚类模型	0.875±0.021[2]	0.783±0.019[1]	0.774±0.026[2]

在分类问题中，从 USPS 和 MNIST 两个数据集中分别挑选 10，30，50 个样本用于训练（例如 USPS10 中数字代表样本个数），同时选 500 个验证集和 500 个测试集，并且使用主成分分析法在 USPS 和 MNIST 中分别挑选 64 个特征和 87 个特征. 结果如表 4 所示，实验结果表明随着样本数量的增多，大多数方法在两个数据集上排名基本一致，Ridge 方法的缺点类似于回归数据集的情况，在小样本设置下容易过拟合. 经过主成分分析，很多特征依旧存在关联并且更具有区分性，导致了更加多样化的簇结构，因此着重于任务差异的 Cluster MTL 方法和 MTRL 方法发挥了一定的效果，Flex-clus 方法能够共享灵活的特征簇因而表现最优. 在特征层面方法中，Dirty MTL 方法保留了过多的判别特征模糊了分类效果；由于此数据集各类之间的主要特征结构有很大相似性，Robust MTL 方法没有分离出离群的特征；反之，Sparse-Robust MTL 方法能够捕捉低秩空间因此性能良好.

表 4 手写体识别 MNIST 数据集分类误差

方法	USPS10	USPS30	USPS50	MNIST10	MNIST30	MNIST50
岭回归	0.754±0.055[2]	0.696±0.042[8]	0.613±0.052[8]	0.644±0.032[8]	0.421±0.080[8]	0.611±0.070[8]
正则化约束多任务方法	0.757±0.058[4]	0.415±0.042[2]	0.516±0.061[3]	0.530±0.035[6]	0.325±0.064[2]	0.400±0.086[2]
脏模型	0.819±0.052[8]	0.599±0.047[7]	0.573±0.060[7]	0.606±0.040[7]	0.373±0.080[7]	0.496±0.086[7]
鲁棒多任务学习	0.763±0.055[6]	0.459±0.044[5]	0.559±0.060[6]	0.466±0.044[1]	0.340±0.065[4]	0.413±0.080[4]
稀疏鲁棒多任务学习	0.790±0.053[7]	0.457±0.047[4]	0.475±0.057[2]	0.468±0.044[2]	0.334±0.060[3]	0.411±0.079[3]
聚类多任务模型	0.758±0.057[5]	0.461±0.046[6]	0.553±0.060[5]	0.470±0.046[4]	0.340±0.065[5]	0.414±0.079[5]
协方差多任务关系模型	0.752±0.050[1]	0.432±0.044[3]	0.552±0.059[4]	0.469±0.047[3]	0.342±0.064[6]	0.421±0.080[6]
可变簇聚类模型	0.756±0.055[3]	0.414±0.042[1]	0.445±0.057[1]	0.475±0.034[5]	0.285±0.056[1]	0.369±0.079[1]

5.2 贝叶斯生成式方法实验

此部分实验对比了当前较新的基于特征层面的概率稀疏低秩模型与结构化算法在回归问题和分类问题中的实验结果. 与单任务套索模型[189]（LASSO）、低秩迹范数正规化[49]（Trace）、聚类多任务模型[39]（Clustered MTL）、鲁棒多任务学习[61]（Robust MTL）以及稀疏鲁棒多任务学习[60]（Sparse-Robust MTL）进行比较. 在回归问题的实验中，训练样本设置中采用不同的训练样本比，从 school data 数据的 139 项任务的训练数据中分别随机挑选 10%、20% 和 30% 的样本组成一个小训练集，其余作为测试集，实验重复 20 次取均值，表 5 中显示了几

表 5　回归数据集的均方根误差

方法	训练样本比 10%	训练样本比 20%	训练样本比 30%
套索	1.587±0.023[6]	1.143±0.015[6]	1.021±0.023[6]
聚类多任务模型	0.969±0.024[4]	0.934±0.019[5]	0.919±0.026[5]
低秩迹范数方法	0.963±0.026[2]	0.928±0.017[4]	0.913±0.025[4]
稀疏鲁棒多任务学习	0.983±0.016[5]	0.909±0.008[2]	0.846±0.021[2]
鲁棒多任务学习	0.965±0.016[3]	0.917±0.008[3]	0.868±0.021[3]
联合稀疏低秩的概率方法	0.944±0.011[1]	0.855±0.011[1]	0.792±0.019[1]

种方法的 $nMSE$ 的平均值和标准偏差，可以看到当训练样本比变大即训练样本的增加时，所有竞争算法的 $nMSE$ 值减小，泛化性能得到提高.

后三种方法都采用了低秩加稀疏的结构组合，随着训练样本的增加，后三种方法也趋于占优，因为在前三种方法中，要么仅具有迹范数正则化（例如 Trace），要么仅有模型的稀疏性假设（LASSO 和 CMTL）. 这表明两点：(1) 使用低秩子空间假设对任务间的相关性进行建模是合理的，具有该假设的方法（Trace、Sparse-Robust MTL、Robust MTL、Probabilistic Low-Rank）比没有低秩假设的方法（CMTL 和 Lasso）取得了相对较好的性能；(2) 对比 Trace 与其它三种叠加模型可知，对底层低秩结构和稀疏结构进行组合建模可以极大地提高性能，低秩结构利用任务之间的关系，稀疏结构可以表达每个任务特定的信息.

并且，即使在基于同样假设的后三种叠加方法中，贝叶斯方法也有明显的优势，Probabilistic Low-Rank 也优于 Sparse-Robust MTL 和 Robust MTL 方法，因为这两种判别式方法近似凸松弛的优化方法会产生一定的精度损失，而考虑潜在的不确定性，对低秩稀疏分量的生成过程进行建模就会避免这种问题. 结果还表明，作为任务的数量增加，由于增加了更多任务的复杂性，大多数方法的误差趋于增加.

在分类任务的实验中，从 The extended Yale Face B 数据集中按每人为单位随机选择 30 个和 40 个训练样本，使用其余作为测试样本，在这两种不同的样本个数条件下重复 20 次实验取均值. 实验方法将回归数据集中的 LASSO 替换为线性支持向量机方法[190]，保留其余方法并且额外增加与深度自编码器方法边际堆栈去噪自编码器[191]（MSDA）、多任务目标跟踪自编码器[192]（MTAE）和去噪自编码器[193]（DAE）的对比，结果如表 6 所示. 同样的，Probabilistic Low-Rank 方法也优于深度学习方法 MSDA，MTAE 和 DAE，因为 MSDA 和 DAE 是多任务方法且不考虑任务之间的结构信息，而 MTAE 仅仅通过使用其它任务重构任务来捕获任务之间的结构信息，当各个任务类中的样本存在显著差异时，MTAE 的假设可能不成立.

表 6　The extended Yale Face B 数据集分类误差

方法	训练样本 30 个	训练样本 40 个
线性支持向量机方法	88.70±0.70%[3]	91.61±1.23%[3]
低秩迹范数方法	50.42±16.57%[8]	46.53±25.36%[8]
稀疏鲁棒多任务学习	82.50±1.28%[7]	81.99±1.82%[7]
鲁棒多任务学习	87.31±1.10%[6]	91.60±0.96%[4]
去噪自编码器	88.36±0.87%[4]	90.84±1.19%[6]
多任务目标跟踪自编码器	89.31±0.77%[2]	91.35±1.16%[5]
边际堆栈去噪自编码器	88.35±1.11%[5]	93.75±1.04%[2]
联合稀疏低秩的概率方法	91.17±0.96%[1]	96.41±0.61%[1]

由于 5.1 节和 5.2 节都是特征层面的方法，除去训练集划分不同的影响，通过横向比较，一般来说，特征层面的学习方法考虑的信息共享机制比较完善，可适用的范围广，但是在不同的数据集上，可以采用的信息迁移结构是不固定的，结构越完善的方法并不一定效果最好. 例如在表 4 手写体识别数据集的结果中可以看到，一些专注于特征层面的学习方法反而不如任务层面的学习方法，这种现象在 The extended Yale Face B 数据中不明显是因为采集样本较少，限制了该数据库的进一步应用，而在 MNIST 数据集中数据的维度过高同时样本容量充足，特征层面的学习方法优势会变弱，因为在很多数据集中样本的差异不是很大，随着样本信息的增多会使不同任务的特征结构趋于类似，单独考虑特征层面的稀疏结构约束基本上得不到体现. 因此，采用基于任务层面学习或者特征层面的方法，标准并不

是唯一的,没有统一的方法对于各个类型的数据集有普遍适用性,多任务方法的适用范围更多地需要考虑数据集的特点,这也是多任务学习的难点之一.

5.3 深度图像分割多任务方法实验

此部分实验是近年来在图像分割领域比较有影响力的工作[194],该方法使用一个级联式的多任务网络,用来输出任意大小图像感知实例的语义分割,分为三个子任务,分别为回归框、回归掩膜和实例分类,前两子任务都是类不可知的,分别输出目标预测得分和像素级分割掩膜,最后的子任务输出分类打分.所有网络共享卷积层结构,并且后一任务损失取决于前一任务的输出形成级联结构,这样也造成了前后任务类似于反向梯度的损失依赖关系,并且这种结构扩宽任务数,可以经由第三个任务的输出继续输入给第二个和第三个任务构成五级级联结构,为了研究此语义分割多任务方法的作用机制,作者在 PASCAL VOC 2012 数据集中采用了两种训练模式:直接采用 Fast R-CNN 结构[101]的预设值或者采用对感兴趣区域(RoI)经过反向梯度求解的端到端方法.评价指标分别采用 0.5 和 0.7 交并比下的平均精度均值 *mAP*(mean Average Precision),网络结构使用 VGG-16 net[195],表 7 展示了不同训练策略设置下的 *mAP* 多任务网络级联,结果表明在(e)三个任务不共享特征时已经超越了 ZF-net 最好效果(d),印证了将语义分割模型分解为级联网络的合理性,而在(f)中当三个网络共享卷积层权重时,结果 60.5%与不共享权重(e)中的基线 60.2%基本一致,没有得到提升的原因是因为三个任务呈现递进的关系,每个阶段学习的特征不应该完全相同,接下来,使用端到端训练方法(g)和扩展为五级级联(h)时,效果得到逐步提升,表明通过优化一个统一的损失函数,利用级联网络的反向传播训练网络可以自适应地共享特征,保持上述实验设置,在交并比 0.5 和 0.7 的条件下分别得到最好的结果 63.5%和 41.5%,与表 8 其它理想多任务方法[196-198]对比可知,MFC 实现了快速准确的实例分割,在处理每幅图的时间和精准度上均有较大提升,展示了多任务学习的优势.

表 7 PASCAL VOC 2012 数据集控制变量实验

训练策略	ZF net				VGG-16 net			
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
是否共享特征权重		√	√	√		√	√	√
是否采用端到端训练			√	√			√	√
是否采用五级级联				√				√
mAP@0.5(%)	51.8	52.2	53.5	54.0	60.2	60.5	62.6	63.5

表 8 PASCAL VOC 2012 数据集实例感知语义分割方法比较

方法	*mAP*@0.5(%)	*mAP*@0.7(%)	时间/每图(s)
SDS(AlexNet)[196]	49.7	25.3	48
Hypercolumn[197]	60.0	40.4	>80
CFM[198]	60.7	39.6	32
MNC[194]	63.5	41.5	0.36

6 未来研究方向

近年来,多任务学习虽然取得了一定程度的进展,但是仍然存在许多问题有待解决.以下我们尝试给出可能的研究问题:

(1)多任务学习的基准数据集的建立和维护,包括人工生成数据集和实际多任务数据集.多任务学习的学习效果常取决于多任务学习数据集所包含的信息.如何界定和划分多任务数据子集,多任务数据集的划分形式对各个任务学习效果影响的评判标准等,都是亟待解决的问题.构建大型多任务学习数据库,充分发挥多任务学习算法学习能力;

(2)提出多任务学习的完备的数学描述和理论体系.多任务学习理论体系的成熟,定能给多任务学习带来更多的实现手段和进步;

(3)分析多任务分类、回归、聚类、降维学习算法的统计特性,从理论上指导多任务学习过程,包括数据集划分、模型结构确定、模型参数选择等;

(4)如何保证多任务学习的性能总是比单个任务的学习性能好,探索是什么因素决定了学习效果,构造多任务学习模型和学习算法,实现一致地改进单个任务学习效果的通用多任务学习模型和范式;

(5)引进其它领域研究的理论和成果,探索新的多任务学习方式,对抗式多任务学习,恶意多任务学习,协作多任务学习,非监督多任务学习;

(6)多任务学习中,既要利用多个任务的共有相似特性,同时也要保留单个任务上特有的特性,如何权衡两者之间的关系,多任务学习中的各种结构的组合形式,具有人为选择任意性,没有一个统一的标准,无法事先判定这种组合形式的好坏.多任务学习的共有和私有表示学习也没有一个统一的标准,到底是怎样把共有和私有表示组合起来,是一个从理论到具体算法实践亟待解决的问题;

(7)多任务学习的目标函数通常为非凸多目标优化问题,目前的多任务学习训练算法不能避免鞍点问题,导致寻优过程失败,使得研究者无法知道到底是优化过程没有找到最优解使得预测结果不好,

还是其它的多任务表示和多任务组合有问题.应该尽快提出求解非凸多任务学习优化问题的优化求解算法,尝试使用多目标优化算法求解多任务学习问题.

7　结论与展望

多任务学习将一些相关过程纳入统一框架,在特征迁移和共享的过程中能发现其它任务中有益于自身的信息,通过任务之间的信息传递,大部分任务的需求都能被考虑到,所以理论上一个健壮的多任务模型全局上可以获得关于各个任务的最优解.随着机器学习对人工标注数据依赖的程度不断降低,多任务能显著减少对标记数据的需求,解决了一些任务样本欠缺造成的学习困难,在多种场景下显示出了强大的实用性,同时在很多应用领域中也获得了不错的效果,越来越受到学术界和工业界的关注.

早期的多任务矩阵模型通过直接划分权值矩阵描述任务之间的关系,是从一般共享结构到具体特征学习的演变,矩阵方法对小规模样本的处理取得了不错的效果,但随着任务复杂性的提高和数据量的增加,应用的重点转变为深度多任务学习,深层网络将任务的关系层次更加深入和具体化,其强大的特征学习能力也大大简化了特征选取方面的难度,同时也增加了对多任务关系的解释性,但是由于其网络结构的复杂性,搜索的难度往往比较大,尽管也可以增加开关单元控制信息的传递,因为运算代价的问题,仍然不能从根本上描述每个层上特征的具体结构.

对多任务的数据域的研究也是一个重点.当前的多任务学习主要建立在数据驱动的基础上,利用相似样本中的信息提升效果,但是数据的分布是提前预知的,例如是否来自同样的数据源,数据源的关联程度决定了任务的相似性,现有的多任务学习经常使用一些经验方法结合专业知识设计任务之间的信息交互迁移形式.广义的多任务应包含类型更丰富的数据源,在样本差异性很大的情况下仍能发现有益信息,因此自适应数据域有待研究.

在多任务学习中,大多数应用注重于结构的设计,虽然在不断的深化其中的具体联系,而缺少通用的选取信息共享和归纳偏置的方法和理论.因此选择合适的共享结构仍然是多任务学习的重点和难点,对于差异度较大的任务依然缺乏系统理论证明其多任务结构设计的合理性.综上所述,有关多任务学习的顶层设计方法仍是以后的重点.

参　考　文　献

[1] Caruana R. Learning many related tasks at the same time with backpropagation//Proceedings of the 8th International Conference on Neural Information Processing Systems. Denver, USA, 1994: 657-664

[2] Thrun S. Discovering structure in multiple learning tasks: The TC algorithm//Proceedings of the 13th International Conference on Machine Learning. Bari, Italy, 1996: 489-497

[3] Schlittgen R. Analysis of incomplete multivariate data. Computational Statistics & Data Analysis, 1999, 30(4): 478-479

[4] Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41-75

[5] Kienzle W, Chellapilla K. Personalized handwriting recognition via biased regularization//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 457-464

[6] Caruana R, Sa V R D. Promoting poor features to supervisors: Some inputs work better as outputs//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 1997: 389-395

[7] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016, 17(59): 1-35

[8] Ganin Y, Lempitsky V. Unsupervised domain adaptation by back propagation//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 1-9

[9] Shinohara Y. Adversarial multi-task learning of deep neural networks for robust speech recognition//Proceedings of the INTERSPEECH. San Francisco, USA, 2016: 2369-2372

[10] Cheng H, Fang H, Ostendorf M. Open-domain name error detection using a multi-task RNN//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 737-746

[11] Evgeniou T, Pontil M. Regularized multi-task learning//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 109-117

[12] Kato T, Kashima H, Sugiyama M, Asai K. Multi-task learning via conic programming//Proceedings of the 21st Advances in Neural Information Processing Systems. Vancouver, Canada, 2007: 737-744

[13] Evgeniou T, Micchelli C A, et al. Learning multiple tasks with kernel methods. Journal of Machine Learning Research, 2005, 6(6): 615-637

[14] Micchelli C A, Pontil M. Kernels for multi-task learning//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 921-928

[15] Obozinski G, Taskar B, Jordan M. Multi-task feature selection. Statistics Department, UC Berkeley, USA: Technical Report: 2, 2006

[16] Chen X, Lin Q, Kim S, et al. An efficient proximal-gradient method for single and multi-task regression with structured sparsity. Institute for Software Research, 2010, 26(5): 4717-4721

[17] Liu J, Ji S, Ye J. Multi-task feature learning via efficient $l_{2,1}$-norm minimization//Proceedings of the Conference on Uncertainty in Artificial Intelligence. Montreal, Canada, 2009: 339-348

[18] Lounici K, Pontil M, Tsybakov A B, et al. Taking advantage of sparsity in multi-task learning. arXiv preprint arXiv: 0903.1468, 2009

[19] Liu J, Ye J. Efficient l_1/l_q norm regularization. arXiv preprint arXiv:1009.4766, 2010

[20] Vogt J, Roth V. A complete analysis of the $l_{1,p}$ group-lasso. arXiv preprint arXiv:1206.4632, 2012

[21] Turlach B A, Venables W N, Wright S J. Simultaneous variable selection. Technometrics, 2005, 47(3): 349-363

[22] Schmidt M, Murphy K, Fung G, et al. Structure learning in random fields for heart motion abnormality detection// Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[23] Quattoni A, Carreras X, Collins M, et al. An efficient projection for $l_{1,\infty}$ regularization//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 857-864

[24] Vogt J E, Roth V. The group-lasso: $\ell_{1,\infty}$ regularization versus $\ell_{1,2}$ regularization//Proceedings of the 32nd Joint Pattern Recognition Symposium. Darmstadt, Germany, 2010: 252-261

[25] Negahban S, Wainwright M J. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization//Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1161-1168

[26] Gong P, Zhou J, Fan W, et al. Efficient multi-task feature learning with calibration//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2004: 761-770

[27] Nie F, Huang H, Cai X, et al. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization//Proceedings of the 24th Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 1813-1821

[28] Yang X, Kim S, Xing E P. Heterogeneous multitask learning with joint sparsity constraints//Proceedings of the 23rd Advances in Neural Information Processing Systems. Vancouver, Canada, 2009: 2151-2159

[29] Zhou Y, Jin R, Hoi S C H. Exclusive lasso for multi-task feature selection//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, 2010: 988-995

[30] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning//Proceedings of the 20th Advances in Neural Information Processing Systems. Vancouver, Canada, 2006: 41-48

[31] Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning. Machine Learning, 2008, 73(3): 243-272

[32] Caponnetto A, Micchelli C A, Pontil M, et al. Universal multi-task kernels. Journal of Machine Learning Research, 2008, 9(3): 1615-1646

[33] Jebara T. Multi-task feature and kernel selection for SVMs//Proceedings of the 21st International Conference on Machine Learning. Pittsburgh, USA, 2004: 55-61

[34] Ding C, He X. K-means clustering via principal component analysis//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 29

[35] Kang Z, Grauman K, Sha F. Learning with whom to share in multi-task feature learning//Proceedings of the International Conference on Machine Learning. Washington, USA, 2011: 521-528

[36] Wang Y, Khardon R, Protopapas P. Shift-invariant grouped multi-task learning for Gaussian processes//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany, 2010: 418-434

[37] Crammer K, Mansour Y. Learning multiple tasks using shared hypotheses//Proceedings of the International Conference on Neural Information Processing Systems. Hurrahs and Harleys, Lake Tahoe, USA, 2012: 1475-1483

[38] Kumar A, Iii H D. Learning task grouping and overlap in multi-task learning. arXiv preprint arXiv:1206.6417, 2012

[39] Jacob L, Bach F, Vert J P. Clustered multi-task learning: A convex formulation//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2008: 745-752

[40] Barzilai A, Crammer K. Convex multi-task learning by clustering//Proceedings of the International Conference on Artificial Intelligence and Statistics. San Diego, USA, 2015: 65-73

[41] Zhao P, Rocha G, Yu B. Grouped and hierarchical model selection through composite absolute penalties. Department of Statistics, UC Berkeley, USA, Technical Report: 703, 2006

[42] Jenatton R, Audibert J Y, Bach F. Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research, 2011, 12(10): 2777-2824

[43] Kim S, Xing E P. Tree-guided group lasso for multi-task regression with structured sparsity//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 543-555

[44] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix t-factorizations for clustering//Proceedings of the Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 126-135

[45] Li T, Ding C. The relationships among various nonnegative matrix factorization methods for clustering//Proceedings of the International Conference on Data Mining. Omaha, USA, 2007: 362-371

[46] Wang Y, Wipf D, Ling Q, et al. Multi-task learning for subspace segmentation//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France, 2015: 1209-1217

[47] Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 2005, 6(3): 1817-1853

[48] Negahban S, Wainwright M. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 2011, 39(2): 1069-1097

[49] Ji S, Ye J. An accelerated gradient method for trace norm minimization//Proceedings of the International Conference on Machine Learning. Montreal, Canada, 2009: 457-464

[50] Pong T K, Tseng P, Ji S, Ye J. Trace norm regularization: Reformulations, algorithms, and multi-task learning. SIAM Journal on Optimization, 2010, 20(6): 3465-3489

[51] Han L, Zhang Y. Multi-stage multi-task learning with reduced rank//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 1638-1644

[52] Chen J, Tang L, Liu J, et al. A convex formulation for learning shared structures from multiple tasks//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 137-144

[53] Yang P, Huang K, Liu C L. Multi-task low-rank metric learning based on common subspace//Proceedings of the International Conference on Neural Information Processing. Shanghai, China, 2011: 151-159

[54] Maurer A, Pontil M, Romera-Paredes B. Sparse coding for multitask and transfer learning//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 343-351

[55] Su C, Yang F, Zhang S, et al. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1167-1181

[56] Zhou Q, Zhao Q. Flexible clustered multi-task learning by learning representative tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 266-278

[57] He Z, Li J, Liu K, et al. Kernel low-rank multitask learning in variational mode decomposition domain for multi-hyperspectral classification. IEEE Transactions on Geoscience & Remote Sensing, 2018, PP(99): 1-16

[58] Gong P, Ye J, Zhang C. Robust multi-task feature learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 895

[59] Chen J, Liu J, Ye J. Learning incoherent sparse and low-rank patterns from multiple tasks//Proceedings of the International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1179-1188

[60] Chen J, Zhou J, Ye J. Integrating low-rank and group-sparse structures for robust multi-task learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 42-50

[61] Fazel M, Hindi H, Boyd S P. A rank minimization heuristic with application to minimum order system approximation//Proceedings of the 2001 American Control Conference. Montreal, Canada, 2001: 4734-4739

[62] Pan Y, Xia R, Yin J, et al. A divide-and-conquer method for scalable robust multitask learning. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(12): 3163-3175

[63] Jalali A, Ravikumar P D, Sanghavi S, et al. A dirty model for multi-task learning//Proceedings of the 23rd Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 964-972

[64] Shen X, Huang H C. Grouping pursuit through a regularization solution surface. Journal of the American Statistical Association, 2010, 105(490): 727-739

[65] Zhong W, Kwok J. Convex multitask learning with flexible task clusters. arXiv preprint arXiv:1206.4601, 2012

[66] Xie S, Lu H, He Y. Multi-task co-clustering via nonnegative matrix factorization//Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan, 2012: 2954-2958

[67] Gu Q, Zhou J. Co-clustering on manifolds//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 359-368

[68] Xu L, Huang A, Chen E. Exploiting task-feature co-clusters in multi-task learning//Proceedings of the American Association for Artificial Intelligence. Austin, USA, 2015: 1931-1937

[69] Murugesan K, Carbonell J, Yang Y. Co-clustering for multi-task learning. arXiv preprint arXiv:1703.00994, 2017

[70] Daumé Iii H. Bayesian multitask learning with latent hierarchies//Proceedings of the Conference on Uncertainty in Artificial Intelligence. Montreal, Canada, 2009: 135-142

[71] Schwaighofer A, Tresp V, Yu K. Learning Gaussian process kernels via hierarchical Bayes//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 1209-1216

[72] Heskes T. Empirical Bayes for learning to learn//Proceedings of the 17th International Conference on Machine Learning. Stanford, USA, 2000: 367-374

[73] Bakker B, Heskes T. Task clustering and gating for Bayesian multitask learning. Journal of Machine Learning Research, 2003, 4(5): 83-99

[74] Zhang Y, Yeung D Y. A regularization approach to learning task relationships in multitask learning. ACM Transactions on Knowledge Discovery from Data, 2014, 8(3): 1-31

[75] Yu K, Tresp V, Schwaighofer A. Learning Gaussian processes from multiple tasks//Proceedings of the International Conference on Machine Learning. Bonn, Germany, 2005: 1012-1019

[76] Bonilla E V, Chai K M A, Williams C K I. Multi-task Gaussian process prediction//Proceedings of the Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 153-160

[77] Ebden M. Gaussian processes: A quick introduction. arXiv preprint arXiv:1505.02965, 2015

[78] Raina R, Ng A Y, Koller D. Constructing informative priors using transfer learning//Proceedings of the International Conference on Machine Learning. Carnegie Mellon University, Pittsburgh, USA, 2006: 713-720

[79] Abernethy J, Bach F, Evgeniou T, et al. Low-rank matrix factorization with attributes. arXiv preprint cs/0611124, 2006

[80] Teh Y W, Seeger M, Jordan M. Semi-parametric latent factor models//Proceedings of the 14th International Conference on Artificial Intelligence & Statistics. Barbados, 2005: 565-568

[81] Yu K, Chu W, Yu S, et al. Stochastic relational models for discriminative link prediction//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2006: 1553-1560

[82] Bonilla E V, Agakov F V, Williams C K I. Kernel multi-task learning using task-specific features//Proceedings of the 16th International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Italy, 2007: 43-50

[83] Zhang Y, Yeung D Y. Multi-task learning using generalized t process//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, 2010: 964-971

[84] Zhang Y, Yeung D Y. A convex formulation for learning task relationships in multi-task learning//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA, 2010: 733-742

[85] Kai N, Carin L, Dunson D. Multi-task learning for sequential data via iHMMs and the nested Dirichlet process//Proceedings of the International Conference on Machine Learning. Oregon, USA, 2007: 689-696

[86] Xue Y, Liao X, Carin L, et al. Multi-task learning for classification with Dirichlet process priors. Journal of Machine Learning Research, 2007, 8(1): 35-63

[87] Baxter J. A model of inductive bias learning. Journal of Artificial Intelligence Research, 2000, 12: 149-198

[88] Han L, Zhang Y, Song G, Xie K. Encoding tree sparsity in multi-task learning: A probabilistic framework//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec, Canada, 2014: 1854-1860

[89] Zhang J, Ghahramani Z, Yang Y. Flexible latent variable models for multi-task learning. Machine Learning, 2008, 73(3): 221-242

[90] Zhang J, Ghahramani Z, Yang Y. Learning multiple related tasks using latent independent component analysis//Proceedings of the 24th Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 1585-1592

[91] Lee S I, Chatalbashev V, Vickrey D, et al. Learning a meta-level prior for feature relevance from multiple related tasks//Proceedings of the 24th International Conference on Machine Learning. Oregon, USA, 2007: 489-496

[92] Xiong T, Bi J, Rao B, et al. Abstract probabilistic joint feature selection for multi-task learning//Proceedings of the SIAM International Conference on Data Mining. Sparks, USA, 2009: 332-342

[93] Lee S, Zhu J, Xing E P. Adaptive multi-task lasso: With application to eQTL detection//Proceedings of the 24th Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 1306-1314

[94] Zhang Y, Schneider J G. Learning multiple tasks with a sparse matrix-normal penalty//Proceedings of the 24th Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 2550-2558

[95] Guo S, Zoeter O, Archambeau C. Sparse Bayesian multi-task learning//Proceedings of the 25th Advances in Neural Information Processing Systems. Granada, Spain, 2011: 1755-1763

[96] Rai P, Daume III H. Infinite predictor subspace models for multitask learning//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, 2010: 613-620

[97] Rai P, Daumé III H. Multitask learning using non-parametrically learned predictor subspaces//Proceedings of the 23rd Advances in Neural Information Processing Systems Workshop on Language Learning. Vancouver, Canada, 2009: 25-32

[98] Hernández-Lobato D, Hernández-Lobato J M, Ghahramani Z. A probabilistic model for dirty multi-task feature selection//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 1073-1082

[99] Yu K, Shao M, Li K, et al. Probabilistic low-rank multitask learning. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(3): 670-680

[100] Liu X, Gao J, He X, et al. Representation learning using multi-task deep neural networks for semantic classification and information retrieval//Proceedings of the North American Chapter of the ACL. Denver, USA, 2015: 912-921

[101] Girshick R. Fast R-CNN. arXiv preprint arXiv:1504.08083, 2015

[102] Zhang Z, Luo P, Chen C L, et al. Facial landmark detection by deep multi-task learning//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 94-108

[103] Arik S O, Chrzanowski M, Coates A, et al. Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv: 1702.07825, 2017

[104] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning//Proceedings of the International Conference on Machine Learning. Helsinki, Finland, 2008: 160-167

[105] Eaton E, Desjardins M, Lane T. Modeling transfer relationships between learning tasks for improved inductive transfer//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium, 2008: 317-332

[106] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3320-3328

[107] Liang Y, Liu L, Xu Y, et al. Multi-task GLOH feature selection for human age estimation//Proceedings of the 2011 18th IEEE International Conference on Image Processing. Brussels, Belgium, 2011: 565-568

[108] Long D, Cohn T, Bird S, et al. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser//Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. Beijing, China, 2015: 845-850

[109] Yang Y, Hospedales T M. Trace norm regularised deep multi-task learning. arXiv preprint arXiv: 1606.04038, 2016

[110] Chu X, Ouyang W, Yang W, et al. Multi-task recurrent neural network for immediacy prediction//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3352-3360

[111] Ouyang W, Chu X, Wang X. Multi-source deep learning for human pose estimation//Proceedings of the Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2337-2344

[112] Srivastava N, Salakhutdinov R. Discriminative transfer learning with tree-based priors//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2094-2102

[113] Yang Y, Hospedales T. Deep multi-task representation learning: A tensor factorisation approach. arXiv preprint arXiv: 1605.06391, 2016

[114] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

[115] Long M, Wang J, Yu P S. Learning multiple tasks with deep relationship networks. arXiv preprint arXiv: 1506.02117, 2015

[116] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning//Proceedings of the Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3994-4003

[117] Ruder S, Bingel J, Augenstein I, et al. Sluice networks: Learning what to share between loosely related tasks. arXiv preprint arXiv: 1705.08142, 2017

[118] Lu Y, Kumar A, Zhai S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5334-5343

[119] Luo Y, Tao D, Geng B, et al. Manifold regularized multitask learning for semi-supervised multilabel image classification. IEEE Transactions on Image Processing, 2013, 22(2): 523-536

[120] Zhang J, Li W, Ogunbona P. Unsupervised domain adaptation: A multi-task learning-based method. arXiv preprint arXiv: 1803.09208, 2018

[121] Yang Y, Ma Z, Yang Y, et al. Multitask spectral clustering by exploring intertask correlation. IEEE Transactions on Cybernetics, 2015, 45(5): 1083-1094

[122] Zhang X, Zhang X, Liu H, et al. Multi-task multi-view clustering. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3324-3338

[123] Zhang X L. Convex discriminative multitask clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 28-40

[124] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification//Proceedings of the 9th International Conference on Data Mining. Miami, USA, 2009: 159-168

[125] Gu Q, Li Z, Han J. Learning a kernel for multi-task clustering//Proceeding of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011

[126] Zhang J, Zhang C. Multitask Bregman clustering. Neurocomputing, 2010, 74(10): 1720-1734

[127] Zhang X, Zhang X, Liu H. Smart multitask Bregman clustering and multitask kernel clustering. ACM Transactions on Knowledge Discovery from Data, 2015, 10(1): 1-29

[128] Al-Stouhi S, Reddy C K. Multi-task clustering using constrained symmetric non-negative matrix factorization//Proceedings of the 2014 SIAM International Conference on Data Mining. Shenzhen, China, 2014: 785-793

[129] Zhang X, Zhang X, Liu H, et al. Multi-task clustering through instances transfer. Neurocomputing, 2017, 251: 145-155

[130] Zhang X, Zhang X, Liu H. Self-adapted multi-task clustering//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016

[131] Zhang X, Zhang X, Liu H, et al. Partially related multi-task clustering. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2367-2380

[132] Maurer A, Pontil M, Romera-Paredes B. The benefit of multitask representation learning. The Journal of Machine Learning Research, 2016, 17(1): 2853-2884

[133] Ben-David S, Schuller R. Exploiting task relatedness for multiple task learning//Proceedings of the Conference on Computational Learning Theory and Kernel Machines. Washington, USA, 2003: 567-580

[134] Baxter J. A Bayesian information theoretic model of learning to learn via multiple task sampling. Machine Learning, 1997, 28(1): 7-39

[135] Thrun S, Pratt L. Learning to Learn. Boston, MA, USA: Kluwer, 1998

[136] Maurer A. Bounds for linear multi-task learning. Journal of Machine Learning Research, 2006, 7(1): 117-139

[137] Maurer A, Pontil M. Excess risk bounds for multitask learning with trace norm regularization. Journal of Machine Learning Research, 2013, 30: 55-76

[138] Zhang Y. Multi-task learning and algorithmic stability// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2-6

[139] Alonso H M, Plank B. When is multitask learning effective? Semantic sequence prediction under varying data conditions// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Mediterranean City of Valencia, Spain, 2017: 149-154

[140] Bingel J, Søgaard A. Identifying beneficial task relations for multi-task learning in deep neural networks//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 164-169

[141] Sermanet P, Eigen D, Zhang X, et al. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013

[142] Liao Y, Kodagoda S, Wang Y, et al. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks//Proceedings of the 2016 IEEE International Conference on Robotics and Automation. Stockholm, Sweden, 2016: 2318-2325

[143] Zhang T, Ghanem B, Liu S, et al. Robust visual tracking via structured multi-task sparse learning. International Journal of Computer Vision, 2013, 101(2): 367-383

[144] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. arXiv preprint arXiv:1705.07115, 2017

[145] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2650-2658

[146] Teichmann M, Weber M, Zoellner M, et al. MultiNet: Real-time joint semantic reasoning for autonomous driving. arXiv preprint arXiv:1612.07695, 2016

[147] Uhrig J, Cordts M, Franke U, et al. Pixel-level encoding and depth layering for instance-level semantic labeling// Proceedings of the German Conference on Pattern Recognition. Cham, Germany, 2016: 14-25

[148] Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization// Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2938-2946

[149] Elhoseiny M, Elgaaly T, Bakry A, et al. Convolutional models for joint object categorization and pose estimation. arXiv preprint arXiv:1511.05175, 2015

[150] Collobert R, Weston J, Karlen M, et al. Natural language processing (Almost) from scratch. Journal of Machine Learning Research, 2011, 12(1): 2493-2537

[151] Yu J, Jiang J. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 236-246

[152] Rei M. Semi-supervised multitask learning for sequence labeling//Proceedings of the 55th Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 2121-2130

[153] Peng N, Dredze M. Multi-task domain adaptation for sequence tagging//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada, 2017: 91-100

[154] Søgaard A, Goldberg Y. Deep multi-task learning with low level tasks supervised at lower layers//Proceedings of the 54th Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 231-235

[155] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning//Proceedings of the 54th Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 149-155

[156] Peng N, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 548-554

[157] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning//Proceedings of the International Joint Conference on Artificial Intelligence. Phoenix, USA, 2016: 2873-2879

[158] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075, 2015

[159] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642

[160] Hashimoto K, Xiong C, Tsuruoka Y, et al. A joint many-task model: Growing a neural network for multiple NLP tasks. arXiv preprint arXiv:1611.01587, 2016

[161] Yang Z, Salakhutdinov R, Cohen W. Multi-task cross-lingual sequence tagging from scratch. arXiv preprint arXiv:1603.06270, 2016

[162] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR//Proceedings of the Latent Variable Analysis and Signal Separation. Liberec, Czech Republic, 2015: 91-99

[163] Chen Z, Watanabe S, Erdogan H, et al. Speech enhancement and recognition using multi-task learning of long short-term

memory recurrent neural networks//Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany, 2015: 3274-3278

[164] Xu Q, Pan S J, Xue H H, et al. Multitask learning for protein subcellular location prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(3): 748-759

[165] Jacob L, Vert J P. Efficient peptide-MHC-I binding prediction for alleles with few known binders. Bioinformatics, 2008, 24(3): 358-366

[166] Zhou J, Liu J, Narayan V A, et al. Modeling disease progression via multi-task learning. NeuroImage, 2013, 78(78): 233-248

[167] Jie B, Zhang D, Cheng B, et al. Manifold regularized multi-task feature learning for multimodality disease classification. Human Brain Mapping, 2015, 36(2): 489-507

[168] Suk H I, Wee C Y, Shen D. Discriminative group sparse representation for mild cognitive impairment classification// Proceedings of the 16th International Workshop on Machine Learning in Medical Imaging. Nagoya, Japan, 2013: 131-138

[169] Monajemi S, Eftaxias K, Sanei S, et al. An informed multi-task diffusion adaptation approach to study tremor in Parkinson's disease. IEEE Journal of Selected Topics in Signal Processing, 2016, 10(7): 1306-1314

[170] Zhu X, Suk H I, Lee S W, et al. Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. IEEE Transactions on Biomedical Engineering, 2016, 63(3): 607-618

[171] Xu J, Deng C, Gao X, et al. Predicting Alzheimer's disease cognitive assessment via robust low-rank structured sparse model//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3880-3886

[172] Zhang H, Gao Y, Deng M, et al. Improving residue-residue contact prediction via low-rank and sparse decomposition of residue correlation matrix. Biochemical & Biophysical Research Communications, 2016, 472(1): 217-222

[173] Wang Y, Yang D, Deng M. Low-rank and sparse matrix decomposition for genetic interaction data. BioMed Research International, 2015, 2015(1): 1-11

[174] Zhang T, Ghanem B, Liu S, et al. Low-rank sparse learning for robust visual tracking//Proceedings of the European Conference on Computer Vision. Berlin, Germany, 2012: 470-484

[175] Bai Y, Tang M. Object tracking via robust multitask sparse representation. IEEE Signal Processing Letters, 2014, 21(8): 909-913

[176] Kim H, Paik J. Low-rank representation-based object tracking using multitask feature learning with joint sparsity. Abstract and Applied Analysis, 2014, 2014(2): 1-12

[177] Fan B, Li X, Cong Y, et al. Structured and weighted multi-task low rank tracker. Pattern Recognition, 2018, 81: 1-12

[178] Yuan X T, Liu X, Yan S. Visual classification with multitask joint sparse representation. IEEE Transactions on Image Processing, 2012, 21(10): 4349-4360

[179] Cheng B, Liu G, Wang J, et al. Multi-task low-rank affinity pursuit for image segmentation//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2439-2446

[180] Qi K, Liu W, Yang C, et al. Multi-task joint sparse and low-rank representation for the scene classification of high-resolution remote sensing image. Remote Sensing, 2016, 9(1): 1-10

[181] He Z, Wang Y, Hu J. Joint sparse and low-rank multitask learning with Laplacian-like regularization for hyperspectral classification. Remote Sensing, 2018, 10(2): 299-322

[182] Dinuzzo F, Pillonetto G, De Nicolao G. Client-server multi-task learning from distributed datasets. IEEE Transactions on Neural Networks, 2011, 22(2): 290-303

[183] Huang S, Peng W, Li J, et al. Sentiment and topic analysis on social media: A multi-task multi-label classification approach//Proceeding of the ACM Web Science Conference. Pairs, France, 2013: 172-181

[184] Wang W, Li Y, Huang Y, et al. A method for identifying the mood states of social network users based on cyber psychometrics. Future Internet, 2017, 9(2): 9-22

[185] Krohn-Grimberghe A, Drumond L, Freudenthaler C, et al. Multi-relational matrix factorization using Bayesian personalized ranking for social network data//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 173-182

[186] Fang C, Rockmore D N. Multi-task metric learning on network data//Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Ho Chi Minh City, Vietnam, 2015: 317-329

[187] Yao W, He J, Wang H, et al. Collaborative topic ranking: Leveraging item meta-data for sparsity reduction//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 374-380

[188] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 1970, 12(1): 55-6

[189] Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society, 1996, 58(1): 267-288

[190] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297

[191] Chen M, Weinberger K Q, Xu Z, et al. Marginalizing stacked linear denoising autoencoders. Journal of Machine Learning Research, 2015, 16(1): 3849-3875

[192] Ghifary M, Bastiaan Kleijn W, Zhang M, et al. Domain generalization for object recognition with multi-task autoencoders //Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2551-2559

[193] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 2010, 11(12): 3371-3408

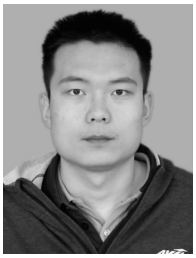
[194] Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3150-3158

[195] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[196] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 297-312

[197] Hariharan B, Arbelaez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 447-456

[198] Dai J, He K, Sun J. Convolutional feature masking for joint object and stuff segmentation//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern. Boston, USA, 2015: 3992-4000



ZHANG Yu, Ph. D. candidate. His main research interest is machine learning.

LIU Jian-Wei, Ph. D., associate professor. His main research interests include machine learning, pattern recognition and intelligent system, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing.

ZUO Xin, professor. His research interests include intelligent control, analysis and design of safety instrumented system and advanced process control.

Background

Multitask learning is a simulation of human learning mode. When we learn new knowledge, we will get inspiration by learning from past experience or by referring to similar process. In machine learning, we call multiple similar learning processes as tasks, and find that the related features between tasks are called information migration. The goal of multitasking is to use additional knowledge and information from other tasks to improve their own learning ability. Meanwhile, the multiple tasks are different but related, the training signals of each task contain specific domain information, so the task can be regarded as a kind of effective inductive bias method, the relationship between tasks is abstracted by sharing of some features, it can be considered that increasing the recognition of noise on each task, reducing the complexity of the model and the risk of over fitting, which can improve the generalization ability in the overall task sets.

At the same time, simultaneous prediction of multiple tasks can reduce the demand for data sources and sample size, accelerate the learning of overall model parameters, and enhance the comprehensibility of learning models. Therefore, in many applications, multi task learning can be used to improve the effect or performance, such as natural language processing, image recognition, speech recognition, etc.

Supported by the 2016 the National Key Research and Development Program (No. 2016YFC0303703-03) and the 2018 China University of Petroleum (Beijing) Prospective Orientation and Cultivation Project (No. 2462018QZDX02).