多任务学习在生物医药发现上研究

马腾飞

S191000907

2019 年 12 月 30 日

目录

摘要

深度学习模型通常需要大量有标签数据才能训练出一个优良的分类器。但是，包括医学图像分析在内的一些应用无法满足这种数据要求，因为标注数据需要很多人力劳动。在这些情况下，多任务学习（MTL）可以通过使用来自其它相关学习任务的有用信息来帮助缓解这种数据稀疏问题。 MTL被分类成多任务监督学习、多任务无监督学习、多任务主动学习、多任务强化学习、多任务在线学习和多任务多视角学习。因此多任务可以用于多个应用场景，通过学习改善无监督学习过程的效果以及在共享参数模型基础上的多任务学习通过任务之间的相互促进增强算法性能。

关键字：多任务学习、生物医药发现、无监督学习、参数共享

Abstract

Deep learning models usually require a large amount of labeled data to train a good classifier. However, some applications, including medical image analysis, cannot meet this data requirement because labeling data requires a lot of human labor. In these cases, multi-task learning (MTL) can help alleviate this data sparse problem by using useful information from other related learning tasks. MTL is classified into multi-task supervised learning, multi-task unsupervised learning, multi-task active learning, multi-task reinforcement learning, multi-task online learning and multi-task multi-view learning. Therefore multitasking can be used in multiple application scenarios. Learning improves the effects of the unsupervised learning process and multi-task learning based on shared parameter models enhances the performance of algorithms through mutual promotion between tasks.

Keywords: multi-task learning, biomedical recovery, unsupervised learning, parameter shared

1 引言

1.1 多任务学习简介

多任务学习方法与以往我们使用的单任务学习方法存在较大的差异，单任务的学习方法仅仅研究一个特定的问题，在以数据驱动的学习任务中泛化能力就显的不够，而多任务学习的模型能够很好的考虑到多个问题之间的关联，可以学习到不同任务数据中隐含的共同特征，这些共同特征往往从多任务模型的低层获取，具有普适性，能更好的表达数据潜在的特征。共享这些共同特征可以使我们的模型更好的概括原始任务。由于多组学数据的来源各自不同，有很强的异质性，并且带标签的样本数量比较少且伴随着数据维度高的问题，申请人将在多组学数据上针对特定的一些任务使用多任务学习的方法进行表示学习用于改善这些任务的性能。，高效且准确的获得疾病相关信息。使用多任务学习方法可以在多组学数据上分析从而获得多个任务之间共享的特征。

1.2 多任务学习在生物医药上的应用

药物相互作用（DDI），药物靶标相互作用（DTI）与化合物蛋白质相互作用（CPI）的发现是生物医药方向的三类相似任务，使用多任务学习方法可以很好的利用这三类任务之

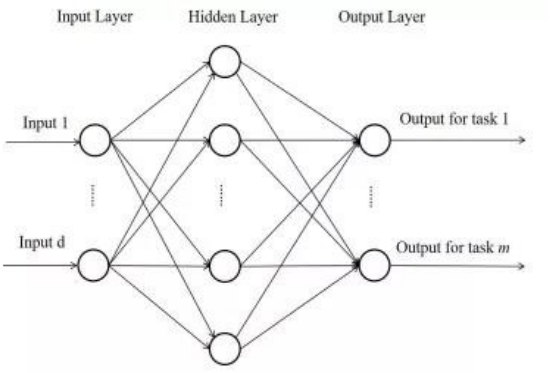

图 1: 多任务学习前馈网络模型

间的相似性。使用这类任务中具有共同特性的数据做为连接，使用多任务学习方法提取各个任务中的共同特征之后进行共享，每个任务使用共享的普适特征进行更抽象的学习。

由于在单一层次结构中选择的特征可能不是最优的分类，为了减轻不含信息或含少量信息的特征对特征选择过程的影响，因此使用一种加权的多任务学习方法获取特征，对于多任务的约束策略还需要深入研究。1）硬参数共享的方式是神经网络中常用的方法，在使用神经网络做药物的预测时候，由于样本数量少而单个样本特征多的特点，容易出现过拟合的现象，而使用硬参数共享的方式，对于多个相似任务，通过融合多个任务数据集，使用共享的表示学习模块等将学习到的特征传递到各个任务中，对每一个任务最终使用一个带加权的损失函数惩罚参数，大大降低了过拟合的风险。2）在使用软参数共享方式时，每个任务相互独立，拥有自己的完整训练模块，但是在每一个任务的低层做一些特征融合与提取等操作，达到共享参数的目的，通过这种方式自然的增加了我们模型中的隐式数据，由于所有任务不同程度的存在噪声，当某些任务在训练模型时，我们目标是学习到一个更好的表征，而通过这种参数的共享的方式恰好能够使学习到的数据具有一个更一般的表征，从而忽略了数据噪声造成的性能差异。下图以相互作用的发现为例，多任务学习可以将多个任务如药物相互作用预测，药物靶点相互作用预测，化合物蛋白质相互作用预测等任务统一到一个模型中去。可以学习到多个任务的共同特征，让多个任务可以相互促进。

2 国内外研究现状

现如今国内外许多研究者提出了多任务学习，他们的出发点都是基本一致的。一旦发现正在优化多于一个的目标函数，你就可以通过多任务学习来有效的求解，也即通过使用包含在相关任务的监督信号中的领域知识来改善泛化性能。从生物学的角度来看，将多任务学习视为对人类学习的一种模拟，通常从相关任务中获取知识用于自身任务。比如婴儿先学会识别脸，然后将学习到的这种技能运用到其他识别任务上去。在生物医药的DDI，

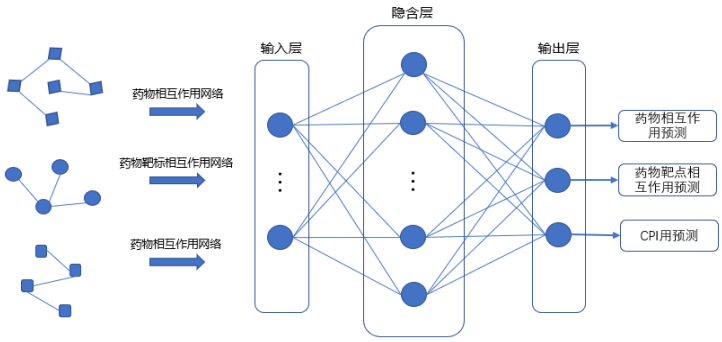

图 2: 多任务学习在药物发现上的应用

DTI等相关任务的发现上具有相似的原理，这些任务不光在数据层面上紧密相关，在算法上，每一层获取到的特征也具有更多的相似特征，这些共同特征也恰好满足了多任务学习的需要。对于多任务学习的模型中的参数共享也主要分为两种。

2.1 硬参数共享模型

参数的硬共享机制是神经网络的多任务学习中最常见的一种方式，这一点可以参考文献[1]，一般来说，它可以应用到所有任务的所有隐层上，而保留任务相关的输出层。这种机制也能很好的降低了过拟合的风险，在文献二[2]也证明了这些共享参数过拟合风险的阶数是N，其中N为任务的数量，比任务相关参数的过拟合风险要小。这一点在许多任务上具有非常明显的意义。越多的任务同时学习，我们的模型就能捕捉到越来越多任务的同一个表示，从而导致在我们原始任务上的过拟合风险越小。

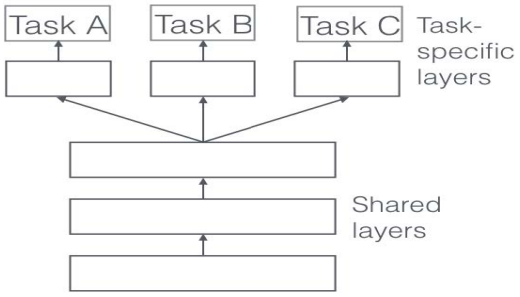

图 3: 在多任务深度神经网络中的硬参数共享机制

在这个模型中的低层用于这三个任务的共享层，该共享层将多个任务中使用的数据集融合，拿到具有更强泛化性的特征，增强每个独立任务的性能。在硬参数共享中，出现了多种的特征共享模式，通过在神经网络中不同层中添加特征共享用于增加模型的泛化性能。在图4中展示了三种不同特征共享方案的一种演化，在最左边的图中，参数共享的过程贯穿

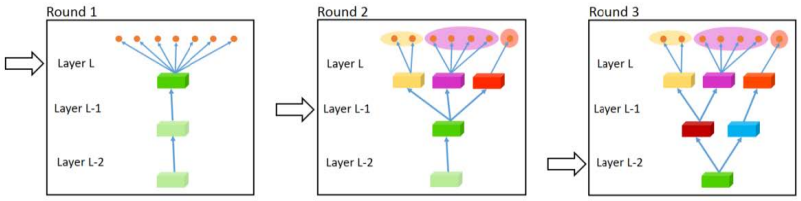

图 4: 在多任务深度神经网络中的特征共享方案

整个任务，通过多个约束条件共同惩罚参数。在中间的这种形式则通过在低层采用共享模式，在较高网络层中每个任务独立使用不同的网络结构，这样做的好处是对于特定的任务可以学习到倾向于该任务的特征而不会对其他任务产生过大的影响。在最右边的图中从底层就开始组件降低参数的共享程度，采用一种渐进的模式降低参数共享对各个任务的影响，同样可以达到能学习普适特征的目的。

2.2 软参数共享模型

在参数的软共享机制中，每个任务都有自己的模型，自己的参数，我们对模型参数的距离进行正则化来保障参数的相似。可以从图中观察到在使用这种方式的时候，每一个任务都有自己特定的模块，独立的数据集，只是在每个人物进行训练的时候，每一层的特征互相共享，通过这种方式获取对应于每个任务的普适特征。同样可以提高模型的整体泛化效果。

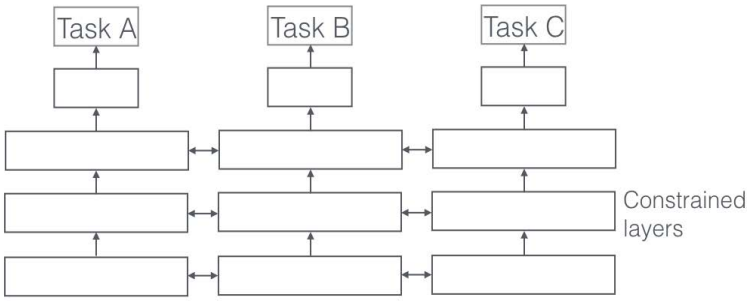

图 5: 在多任务深度神经网络中的软参数共享机制

在这个模型的基础上海交通大学的Wang Hongwei在一项推荐任务[4]中使用了这种参数共享机制。作者将商品推荐任务与知识图谱嵌入任务相互结合，在两个任务的低层加入了参数共享机制，通过两个任务的共同对象商品找到任务之间的共同关联。通过Cross操作将两个任务中有关商品的共同特征整合起来用于加强两个任务的内在关联，得到两个任务的内在关联之后使用Compress操作将这些关联的高维数据进行降维分发到不同的任务中去。这样做有多种好处，其一通过在低层加入这种特征融合与分解机制会更容易学习到一些普

适规律不会对顶层的任务的学习造成影响。其二，通过这种方式将不同任务的数据中可能包含的噪声分摊到不同的任务中去，自然的降低了过拟合的可能。下图展示了参数共享的过程。

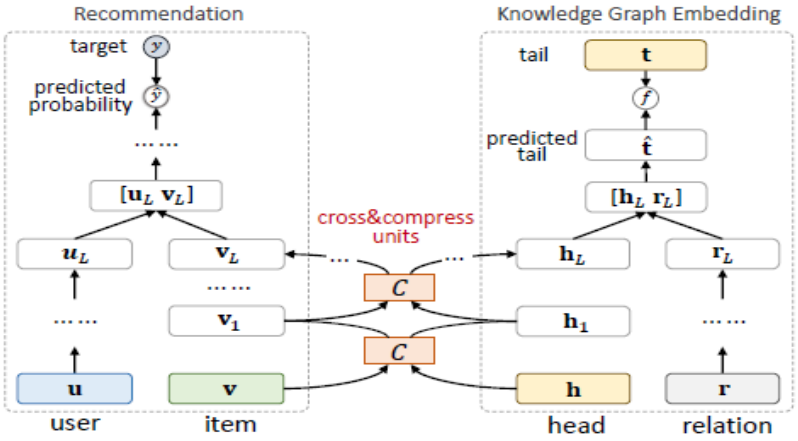

图 6: 使用软参数共享方式进行知识图谱嵌入推荐

2.3 多任务学习在药物发现上的应用

2.3.1 多任务网络应用于药物发现

斯坦福大学的Bharath Ramsundar在2015年时发表了一篇多任务学习网络在药物发现上的一篇技术文章[5]。在这篇文章中，作者通过整合多个公开的生物数据源于一个网络中，所有的任务基于这个特征网络进行学习，通过使用不同数量的任务数探测不同的任务数量对最终实验性能的影响。作者在实验中证实了通过增加任务数量，数据集的总数量也相应的增加，在一定程度上可以增强整体模型的性能，而且大大减少了过拟合的可能，这对于药物发现这种有效数据比较少的任务来说预防过拟合也同样改善了整体模型的性能。

在该模型的基础上，作者做了另外一个实验，通过不断修改模型中的任务数，判断在不同的数据集上，任务数量对最终性能的影响。通过实验的结果可以得知在任务数量到达某个值之前，任务的数量越多越能增强模型的性能，但是超过这个阈值之后性能急速下降。同样也可佐证参数的共享数量会对最终模型的预测性能有影响，一方面可能是减少了过拟合的作用，另一方面是由于任务数量过多的时候对于共享的参数所学习到的特征过于表层出现欠拟合的情况，影响模型的最终性能。

2.3.2 多任务学习在DDI预测上的使用

Xu Chu等人在IJCAI上的一篇2019年的工作[6]，使用了多任务学习的模型为无监督数据做标签预测，同时将有标签数据与无标签数据进行融合聚类，采用软参数共享的模式用

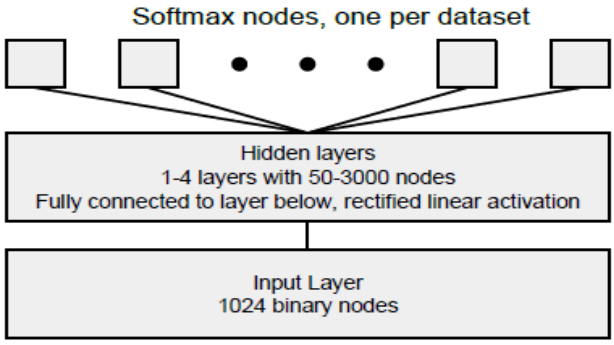

图 7: Massively networks in Drug Discovery

于drug-drug interaction的预测。作者在使用多任务模型做数据的标签预测的时候通过将有标签与无标签数据进行融合得到新的特征，一方面通过这种模型扩大了数据的容量，弥补了在做DDI预测的问题中数据量不足的缺点。作者提出将无监督任务与有监督任务相互结合得到，模型中将n个药物的单特征表示模块组合，将在该模块中学习到的特征用于最终DDI的预测。

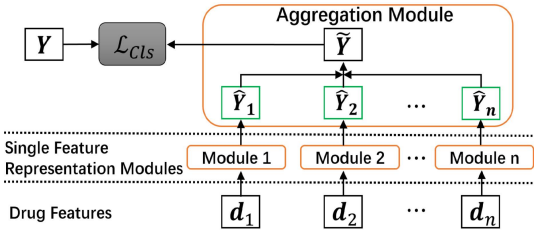

图 8: DDI Prediction

在药物单特征表示模块中作者定义了一个自编码器的网络结构，在编码器阶段编码药物的输入，作者使用了一个CuXCov损失函数对学习到的中间特征进行约束，同时对于预测来说，作者使用了一个Cls损失来约束drug-drug interaction的预测性能。同时对于自编码器的本身结构，作者也加入了一个Rcnst（重构损失），对于整个模型，作者将这些定义的损失函数进行加权求和。在多任务学习中，对于每个任务的的损失的权重的选取是一个比较复杂的工程，现在也有了许多这方面的研究。

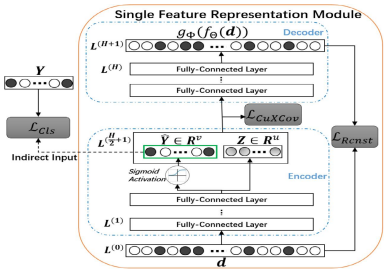

图 9: The architecture of SFRM

3 核心技术及创新点

3.1 多任务优化

在多任务的学习中，多任务的优化也就是对于每一个任务在整个模型中的重要程度以及如何优化多个惩罚函数。对于药物发现来说，比如DDI（drug-drug interaction），DTI（drug-target interaction）与CPI（compound-protein interaction）这些任务都有很多的相似之处。为了做到任务间的良好协作，需要使用一些策略来组合不同任务之间的惩罚权重。这个也是多任务学习中的核心内容，对于加权损失一般都有有如下形式。

$$L_{total} = \sum_i w_i L_i.$$

图 10: 加权损失

在这个公式中，ω_i表示针对于任务i的损失L_i的权重，最后再进行加权求和，因此对于这种形式，也预示着参数的选取对最终结果的影响也会非常的大，通过手工的方式进行选取的话将会耗费大量的人力。对于这种参数选择的特征，可以对这些参数进行训练，通过最小化损失L_total以及加上需要倾向的任务的限制条件，通过这种方式可以学习到参数可以不错的应用于模型中。这也是现在所采用的大多数方式[7]。

3.2 参数共享

多任务学习中另一个核心问题就是共享参数的选取以及何时进行共享。不同的参数共享策略会给模型的好坏带来非常大的影响，因此很多研究都在参数共享上花费了很大的功夫。在上文中也提到过，现在主要使用的两种——硬参数共享，软参数共享。对于医药发现上，大多采用了软参数共享的模式，软参数共享模式下共享低层的特征参数不会使某一种特征对整体模型的影响过大。因此共享参数的选择也成为了使用该模型的重要任务。

3.3 创新点

现在使用多任务学习做药物发现的实验室还不多，一方面多任务学习在不同的领域中间有很多的不确定性，比如在推荐任务，或者自然语言处理上可能会出现负增强的现象，这种情况往往跟多个任务之间惩罚的程度有关系，另一方面跟任务的相关性也很有关联。在药物发现中由于多个任务之间相互关联，使用的数据集关联性也很大，尽管已经标记的数据不多，但是对于多个任务来说间接的增强了可使用数据的数量，并且对于数据量比较少的任务，这种机制也恰恰通过数据增强了任务的性能。另一方面，在药物发现领域存在很多未标记的但是与标记数据相关联的数据，通过监督学习来促进无监督学习也是在多任务学习在医药发现领域中常常使用的一个创新点。

参考文献

[1] Caruana, R. Multitask learning: A knowledge-based source of inductive bias. In Proceedings of the Tenth International Conference on Machine Learning.

[2] Baxter, J. A Bayesian/information theoretic model of learning to learn via multiple task sampling. Machine Learning, 28:7–39.

[3] Aylien Ltd., Dublin. An Overview of Multi-Task Learning in Deep Neural Networks.

[4] Hongwei Wang, Fuzheng Zhang3. Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation.

[5] Bharath Ramsundar*, J. (2015). Massively Multitask Networks for Drug Discovery.

[6] Xu Chu et.c, R. (2019). MLRDA: A Multi-Task Semi-Supervised Learning Framework for Drug-Drug Interaction Prediction.

[7] Alex Kendall et.c, R. (2018). Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.

[8] Zhibo Wang., R. (2019). Network-based multi-task learning models for biomarker selection and cancer outcome prediction.

[9] Subhojeet Pramanik., R. (2019). OmniNet: A unified architecture for multi-modal multi-task learning.

[10] Diogo M. Camacho., J. (2018) Next-Generation Machine Learning for Biological Networks.

[11] Victor Sanh1., (2018). A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks.

[12] Y-h. Taguchi., (2019). Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data.

[13] Wen Zhang1,2*, Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data.

[14] Antonio de la Vega de León., (2018) Effect of missing data on multitask prediction methods.

[15] Bryan Perozzi., DeepWalk: Online Learning of Social Representations.

[16] Shikun Liu., End-to-End Multi-Task Learning with Attention.