《复杂结构数据挖掘》第二次作业

191250026 丁云翔

我调查了在数据挖掘或机器学习社区发表的论文，找到了以下三种比较先进的图表示方法。

以编码器为基本单元的深度神经网络

特点

图自动编码器GraphEncoder在谱聚类的基础上进行改进，同样采用图嵌入的方法，更强调稀疏性。算法的流程如下：将标准化后的图相似矩阵送入一个以稀疏自动编码器为基本单元的深度神经网络（栈式自编码网络），通过一个贪婪的逐层预训练过程，寻找能通过重构逼近输入矩阵并满足期望的稀疏性的最佳非线性图表示。在堆叠若干层稀疏自动编码器之后，得到的最后一层的稀疏编码即为最终的在嵌入空间上的图表示，可以用它来进行聚类。

优点

1. 堆叠结构提供了消除边的平滑方法，随着训练从浅层到深层，对于聚类而言，图形表示将变得越来越清晰。
2. 实验表明，使用基于稀疏自动编码器的GraphEncoder进行聚类的性能优于直接在原始的标准化后的图相似矩阵上聚类，深层结构有助于获得更好的图表示。
3. 由于是稀疏表示，对于稀疏图的表达效果更好。

缺点

1. 需要遍历所有节点，并存储所有节点的邻接点信息，对于比较大的图计算开销较大。
2. 只利用了1-hop信息，图嵌入的效果可能不够好。

只采用attention mechanism的图神经网络模型

特点

占主导地位的图神经网络 (GNN) 过度依赖图链接，已经显露出了几个严重的性能问题，例如假死问题和过度平滑问题。更重要的是，固有的相互连接的性质使得图的并行处理变得不可能，这对于大型图的处理十分不利，因为内存限制限制了跨节点的批处理。所以提出了一种新的图神经网络模型GRAPH-BERT，只依赖attention机制，不涉及任何图卷积和聚合算子。GRAPH-BERT将原始图采样为多个子图，并且只利用attention机制在子图上进行表征学习，而不考虑子图中的边信息。GRAPH-BERT主要由如下四部分组成：

1. 将原始图分解为无边子图集合。对于每个目标节点，根据节点之间的相似度分数选取前K个节点作为该目标节点的上下文节点。
2. 节点输入特征的嵌入表示。考虑四种节点的特征：(1) raw feature vector embedding，(2) Weisfeiler-Lehman absolute role embedding，(3) intimacy based relative positional embedding，(4) hop based relative distance embedding。
3. 基于图transformer的节点表征学习编码器。编码器的输出即为最后学习到的节点特征表示（即图表示）。
4. 基于图transformer的解码器。主要考虑节点属性重建以及结构恢复任务。

优点

1. 解决了当前图神经网络（GNN）的主要方法过度依赖图中的连接关系造成的三大问题：模型假死、过度平滑、难以并行计算。
2. 处理大型图更加高效。
3. 基于具有节点属性重建和结构恢复任务的预训练的 GRAPH-BERT，我们在节点分类和图聚类任务上进一步微调 GRAPH-BERT，得到的模型在学习效果和效率上都优于现有的GNNs。

缺点

1. 由于不考虑图中的边信息，可能会忽略边所表示的一些重要信息。

图滤波

特点

当前主流的图表示技术是深度神经网络，通常能学习到很好的数据表示，但它涉及到的参数繁多、计算量大、容易过拟合、且难解释。所以提出一种浅层模型的实现——基于图滤波的图表示学习方法。从论文标题就可以看出，这是一种“对聚类友好的图表示方法”。该方法经实验验证，可用于子空间聚类任务，寻找一个对聚类友好的图表示，可以大大提高图聚类的性能，甚至接近深度学习方法的结果。该方法同时也具有一般意义，也可以推广到聚类以外的其他机器学习或数据挖掘任务上。基于图滤波的思想，还有其他一些衍生的图表示方法，如应用于多视图属性图数据上的聚类算法——多视图对比图聚类，其获得图表示的算法流程如下：首先使用了图滤波的方法从原始图数据中得到更加平滑的数据特征表示，滤除高频的噪声，使得到的特征表示更加有利于后继任务。接着利用数据的自表达性质以及一套自适应的权重分配机制，从原始的多图中学习得到一个高质量图。最后，受到自监督学习的启发，提出图上的对比学习正则项，拉进相似的数据点，提高图的聚类亲和性。

优点

1. 该方法有坚实的信号处理理论基础，模型简单，易于实现。
2. 使用传统浅层模型实现，参数更少，计算量大大减小、不易过拟合、可解释性强，效果与深度神经网络不相上下，但效率大大提升。

缺点

1. 在面对复杂的图结构时，表示能力可能不如深度神经网络。

课堂上学习的方法

特点

直接使用原始的图表示方法，包括顶点、边以及边上的权重。以聚类为例，对图结构进行聚类，可以视作对图进行划分。图聚类算法主要包括基于距离的方法和基于平凡子结构的方法。基于距离的方法的代表是K-medoids，该算法流程与K-means相似，唯一的不同在于K-mediods每次选取的簇中心点必须是样本点，而不能直接使用簇内样本均值。

优点

1. 方法最简单，不需要其他的模型对图数据进行表示学习或滤波等操作，可以直接在原始数据上进行聚类或分类。对于结构简单且在原始空间内能够有效区分不同类别样本点的图表现很好。

缺点

1. 基于原始的图表示进行分类或聚类，可能很难在原始空间内将不同类的样本点区分开。

对比

对比以上四种特征表示方式，可以发现：对于结构简单且在原始空间内能够有效区分不同类别样本点的图，可以直接采用原始的图表示方法；对于稀疏图，以编码器为基本单元的深度神经网络的图表示方法更加合适；对于大型图（尤其是需要并行处理的情况），只采用attention mechanism的图神经网络模型的图表示方法更加高效；对于较为复杂的图但计算和存储资源又有限（不能使用深度模型）的情况，图滤波的图表示方法最适合。

参考文献

1. Fei Tian，Bin Gao，Qing Cui，Enhong Chen，Tie-Yan Liu，《Learning Deep Representations for Graph Clustering》，Twenty-Eighth AAAI Conference on Artificial Intelligence
2. Ming Shao，Sheng Li，Zhengming Ding，Yun Fu，《Deep Linear Coding for Fast Graph Clustering》，Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)
3. Jiawei Zhang，Haopeng Zhang，Congying Xia，Li Sun，《GRAPH-BERT: Only Attention is Needed for Learning Graph Representations》，arXiv:2001.05140
4. Zhengrui Ma，Zhao Kang，Guangchun Luo，Ling Tian，Wenyu Chen，《Towards Clustering-friendly Representations: Subspace Clustering via Graph Filtering》，MM '20: Proceedings of the 28th ACM International Conference on MultimediaOctober 2020 Pages 3081–3089
5. Zhiping Lin，Zhao Kang，《Graph Filter-based Multi-view Attributed Graph Clustering》，Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
6. Erlin Pan，Zhao Kang，《Multi-view Contrastive Graph Clustering》，Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021)