NIPS18一篇使用图神经网络来做图的边预测任务的论文

解决的问题

如何能自动而非人工定义的方式来学习图中的结构信息，从而进行边预测。

边预测任务就是预测图中的两个顶点是否有可能有边相连。一种常用的方法为启发式方法(heuristic)，它根据定义的顶点相似度来判断这条边存在的概率有多大。几种定义相似度的方法可以根据需要使用的邻居顶点的跳数来分类，例如common neighbors与preferential attachment是一阶的，因为它们只需要一跳邻居的信息，而Adamic-Adar和resource allocation为二阶，Katz、rooted PageRank与SimRank是更高阶的相似度。

这种启发式方法的缺点在于，边存在的概率很大程度依赖于定义的顶点相似度。例如选取common neighbors这个相似度，在社交网络可能是成立的，因为如果两个人有很多共同的朋友，他们两个确实更有可能认识，但是在蛋白质交互网络截然相反，有越多相同邻居顶点的蛋白质反而越不可能建立联系。所以，与其预先定义一种相似度，不如根据网络的特点自动的学习出来。

另一个挑战是，高阶的相似度相较于低阶相似度往往能带来更好的表现，但是随着阶数越高，每个顶点所形成的子图会越来越逼近完整的图，这样会带来过高的时间复杂度与空间复杂度。本文的另一个贡献就在于，定义了一种逼近的方式，不需要h阶的子图也能近似的获取h阶子图中包含的信息，之间的误差有理论上限。

做法及创新

同ICLR20的论文ICMC一样（毕竟是同一个作者），论文对子图的定义方式为，给定一对顶点(x, y)，它的子图为顶点x与y不高于h阶的邻域的一个并集，数学描述如下，也就是与顶点x或y的距离小于等于h所构成的点的集合：

> 给定一个图$G = (V, E)$，以及图上两个顶点x、y，它的h阶围绕子图(enclosing subgraph)$G^h_{x,y}$为图G的一个子图，满足$\{i|d(i,x) \leq h \ or \ d(i,y) \leq h\}$.

接下来是定义一个γ-decaying heuristic函数，它用来逼近h阶子图的信息而不需要实际计算h阶子图：

$$H(x,y) = \eta \sum_{l=1}^{\infty} \gamma^l f(x,y,l)$$

其中γ是一个位于$(0,1)$的衰减因子，η是一个正的常数或一个上界为常数的函数。因为这里的求和从1到∞，接下来的定理说明可以用有限项去逼近$H(x,y)$，误差随着h的增加而指数下降：

> 定理一：
>
> 如果函数$f(x,y,l)$满足：
>
> 1. $f(x,y,l) \leq \lambda^l$，其中$\lambda < \frac{1}{\gamma}$
> 2. 对于$l = 1, 2, \ldots, g(h)$，$f(x,y,l)$能够从h阶子图$G^h_{x,y}$中计算得到，其中$g(h) = ah + b$，$a, b \in \mathbb{N}, \ a > 0$

证明的方法很容易理解：

> 逼近项为：
>
> $$\tilde{H}(x,y) = \eta \sum_{l=1}^{g(h)} \gamma^l f(x,y,l)$$
>
> 计算差值可以得到：

$$
\begin{aligned}
|H(x,y)-\tilde{H}(x,y)| &= \eta \sum_{l=g(h)+1}^{\infty} \gamma^l f(x,y,l) \\
&\le \eta \sum_{l=ah+b+1}^{\infty} \gamma^l \lambda^l \\
&= \eta \frac{(\gamma\lambda)^{ah+b+1}}{1-\gamma\lambda}
\end{aligned}
$$

第一个不等式是根据定理一的第一个条件，最后一个等号是根据等比数列的求和公式，当项数$n \to \infty$且$q \in (0,1)$时，结果为$\frac{a_1}{1-q}$。

到这里可能还是不知道这个$H(x,y)$和图中h阶的信息有什么关系，下面就通过Katz、rooted PageRank和SimRank三个高阶相似度来具体说明怎么使用：

在说明之前，先介绍一个引理，接下来会用到，证明起来也很直观：

> 顶点x与y之间任意一条长度l满足$l \le 2h+1$的路径都被包含在子图$G^h_{x,y}$中

证明：

> 即证明给定一条长度为l的路径$w =< x, v_1, \dots, v_{l-1}, y >$中的每一个顶点都在子图中。取其中任意一个顶点$v_i$，满足$d(v_i,x) \ge h$且$d(v_i,y) \ge h$，根据子图$G^h_{x,y}$的定义它不在其中。那么有：
>
> $$2h+1 \ge l = |< x, v_1, \dots, v_i >| + |< v_i, \dots, v_{l-1}, y >| \ge d(v_i,x) + d(v_i,y) = 2h+2$$
>
> 矛盾，不等号是因为$d(x,y)$就是表示两个顶点之间的最短路径，所以有$d(v_i,x) < h$或$d(v_i,y) < h$，则顶点v_i在子图$G^h_{x,y}$中。

Katz index

给定一对顶点(x,y)，Katz index定义为：

$$
\mathrm{Katz}_{x,y} = \sum_{l=1}^{\infty} \beta^l |\mathrm{walks}^{<l>}(x,y)| = \sum_{l=1}^{\infty} \beta^l [A^l]_{x,y}
$$

其中$\mathrm{walk}^{<l>}(x,y)$是这两个顶点之间长度为l的路径构成的集合，A^l是邻接矩阵的l次幂。从表达式可以看到，长度越长的路径在计算时会被β^l衰减的越多$(0 < \beta < 1)$，短路径有更大的权重。

对比两式可以发现：

$$
\mathrm{Katz}_{x,y} = \sum_{l=1}^{\infty} \beta^l |\mathrm{walks}^{<l>}(x,y)| = \sum_{l=1}^{\infty} \beta^l [A^l]_{x,y}
$$

$$
H(x,y) = \eta \sum_{l=1}^{\infty} \gamma^l f(x,y,l)
$$

Katz index是论文中定义的γ-decaying heuristic函数的一种特殊形式，取$\eta = 1, \gamma = \beta$，$f(x,y,l) = |\mathrm{walks}^{<l>}(x,y)| = [A^l]_{x,y}$。根据引理，只要取长度小于2h+1的路径，其中的顶点就会全部被子图给包含，这也就满足了定理一的第2个“可计算”条件。对于第一个条件，可以通过数学归纳法说明Katz index的表达式同样满足：

> 给定任意的顶点i、j，$[A^l]_{i,j}$的上限为d^l，其中d是网络中的最大顶点度

数学归纳法证明：

> 当$l=1$时，$A_{i,j}$退化成了顶点的度，那显然有$A_{i,j} \le d$成立。假设$k=l$时也成立$[A^l]_{i,j} \le d^l$，当$k = l+1$时：
>
> $$
> [A^{l+1}]_{i,j} = \sum_{k=1}^{|V|} [A^l]_{i,k} A_{k,j} \le d^l \sum_{k=1}^{|V|} A_{k,j} \le d^l d = d^{l+1}
> $$

第一个等式就是矩阵乘法的定义，因为$[A^{l+1}]$的含义就是$l+1$个邻接矩阵A相乘。因此，对比定理一的第一个条件，我们只要取$\lambda=d$，d满足$d<\frac{1}{\beta}$就能够成立，这样一来两个条件都被满足了，这说明Katz index能够很好地从h阶子图中近似。

PageRank

rooted PageRank来源于这篇论文Topic-sensitive PageRank，它通过迭代计算PageRank向量π_x来得到某一点相对于其它顶点的相似度。具体来说，它计算一个从顶点x开始的随机漫步的平稳分布，这个随机漫步以概率α移动到任一邻居上或以概率$1-\alpha$回到顶点x。这个平稳分布满足：

$$\pi_x=\alpha P\pi_x+(1-\alpha)e_x$$

其中$[\pi_x]_i$表示在这个平稳分布下漫步到顶点i的概率，P为转移矩阵，其中$P_{i,j}=\frac{1}{|\Gamma(v_j)|}$，这里的$\Gamma(v_j)$表示顶点$v_j$的一跳邻居构成的集合。如果一个顶点与五个顶点相连，那它转移到其中任意一个顶点的概率就是$\frac{1}{5}$。

rooted PageRank应用于边预测任务时，用来得到一对顶点(x,y)的分数，以$[\pi_x]_y$或$[\pi_x]_y+[\pi_y]_x$（对称）表示，分数越高越有可能有边相连。

接下来就要说明rooted PageRank如何能够同样以论文中提出的γ-decaying heuristic函数进行表示。根据inverse P-distance理论，$[\pi_x]_y$能够等价地改写为：

$$[\pi_x]_y=(1-\alpha)\sum_{w:x\rightsquigarrow y}P[w]\alpha^{len(w)}$$

这里的求和范围$w:x\rightsquigarrow y$表示所有从x开始结束于y的路径，$P[w]$定义为$\prod_{i=0}^{k-1}\frac{1}{|\Gamma(v_i)|}$，$k$是路径长度，$v_i$是路径中的顶点，通过这条路径来从$x$到$y$的概率就是漫步到路径中每一个顶点的概率的连乘。

接下来就是证明这个形式满足定理一的两个条件：

> 首先进一步改写：
>
> $$[\pi_x]_y=(1-\alpha)\sum_{l=1}^{\infty}\sum_{\substack{W:x\rightsquigarrow y\\ len(w)=l}}P[w]\alpha^l$$
>
> $$H(x,y)=\eta\sum_{l=1}^{\infty}\gamma^l f(x,y,l)$$
>
> 对比：取$\gamma=\alpha,\eta=(1-\alpha),f(x,y,l)=\sum_{l=1}^{\infty}\sum_{\substack{W:x\rightsquigarrow y\\ len(w)=l}}P[w]$。因为这时候$f(x,y,l)$表示一个随机漫步恰好以$l$步从顶点$x$漫步到$y$的概率，有$\sum_{z\in V}f(x,z,l)=1$，则$f(x,y,l)\leq 1<\frac{1}{\alpha}$，这样就满足了定理一，而根据引理，只要取长度小于等于2h+1的路径，路径中的点就会被全部包含在子图中，也就满足了第二个"可计算"条件。

SimRank

SimRank的核心思想是，如果两个顶点的邻域相似，那它们也相似：

$$s(x,y)=\gamma\frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)}s(a,b)}{|\Gamma(x)|\cdot|\Gamma(y)|}$$

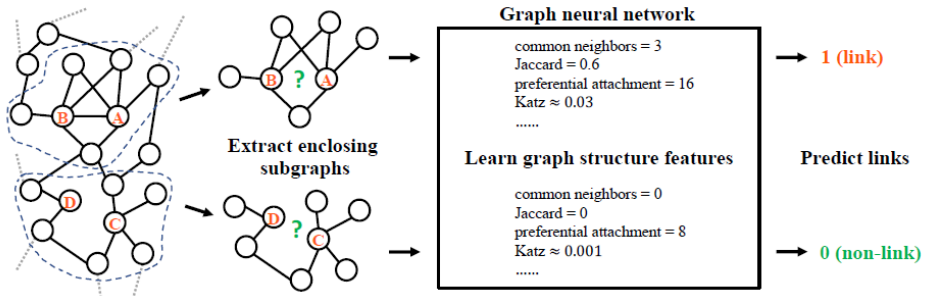
它有一个等价定义形式：

$$s(x,y)=\sum_{w:(x,y)\multimap(z,z)}P[w]\gamma^{len(w)}$$

其中$w:(x,y)\multimap(z,z)$表示从顶点x开始的随机漫步与从顶点y开始的随机漫步第一次相遇于顶点z。证明与rooted PageRank基本一致，可以见原论文。

总结来说，γ-decaying heuristic函数的思想是，对于远离目标顶点的结构信息通过指数衰减的方式给一个更小的权重，因为它们带来的信息十分有限。

SEAL框架



这一节就是根据上面的理论分析建立一个用于边预测任务的框架。一个图神经网络的典型输入形式是(A, X)，在本论文中，A自然地被定义为子图$G^h_{x,y}$的邻接矩阵，子图的获取即来自正样本（已知边）也来自负样本（未知边）。接下来的部分就是介绍论文怎么定义顶点的特征矩阵X，它包含三个部分：structural node labels、node embeddings和node attributes。

Node labeling

跟作者的另一篇论文ICMC一样，通过给顶点打标签的方式来区别顶点在子图中的不同角色，这么做的意义在另一篇博客说过了这里就不写了，具体打标签的方式为：

- 起始顶点x与目标顶点y的标签都为"1"
- 如果两个顶点i、j距离起始顶点与目标顶点的距离都相同，那么它们的标签一样
- $(d(i,x), d(i,y)) = (a,b) \rightarrow label: a+b$

将顶点的标签进行one-hot编码后作为结构特征。

Node embeddings + Node attributes

Node attributes一般数据集直接给定，而Node embeddings是通过一个GNN得到，具体做法是：给定正样本$E_p \in E$，负样本E_n，$E_p \wedge E_n = \emptyset$，在这么一个图$G'=(V,E\and E_n)$上生成embeddings，防止过拟合。

数据集

USAir、NS、PB、Yeast、C.ele、Power、Router、E.coli