(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109214599 A
(43)申请公布日 2019.01.15

(21)申请号 201811253235.6

(22)申请日 2018.10.25

(71)申请人 北京师范大学
地址 100000 北京市海淀区新街口外大街19号
申请人 集智学园(北京)科技有限公司

(72)发明人 谷伟伟 高飞 张江

(74)专利代理机构 江苏爱信律师事务所 32241
代理人 唐小红

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 50/00(2012.01)
G16B 5/00(2019.01)

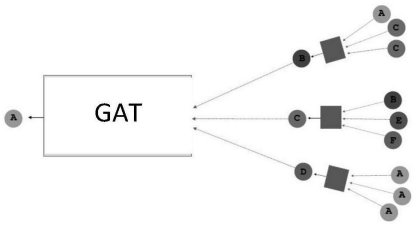
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种对复杂网络进行链路预测的方法

(57)摘要

本发明提供了一种对复杂网络进行链路预测的方法,基于图注意力网络(GAT)的端到端链路预测模型,以及该模型的分批训练方法。该模型的关键在于学习网络节点对周围邻居的注意力分布。模型的训练和利用模型预测的步骤包括:步骤一,输入无权无向同质网络的拓扑结构;步骤二,根据训练集的拓扑结构对所有节点进行一阶、二阶邻居采样,以便将网络分批;步骤三,将分批后的训练集输入上述模型训练出模型参数;步骤四,输入想要预测的点对,模型输出该点对之间有连边的概率。本发明所述模型具有端到端的特点。分批训练方法使得该模型对大规模复杂网络也适用。



CN 109214599 A

1.一种对复杂网络进行链路预测的方法，包括模型的构建和其分批训练方法，其特征在于，包括：对网络拓扑结构预处理，得到分批训练数据集；建立基于GAT的端到端链路预测模型；对模型进行分批训练，得出模型参数；利用训练好的模型对连边进行预测，所述模型包含训练好的GAT模型和其后的二分类器模型，其方法具体如下：

1).对需要处理的目标网络进行消除方向消除权重处理，得到网络无向无权的同质拓扑结构，该网络不能包含孤立节点；

2).上述网络中连边对应的点对作为训练集中的正例，同时随机采集与连边数等量且没有连边的点对，作为训练集中的负例；对正负例中出现的所有点进行固定数目一阶、二阶邻居采样，节点和其邻居看做整体，然后将训练集分批；

3).构建基于GAT的端到端链路预测模型，包含以下部分：

3.1).模型输入为点对和他们的一阶、二阶邻居，输出为该点对之间有连边的概率；

3.2).根据网络数据实际情况，初始化节点向量h_i^0为，其中i为节点下标；

3.3).节点向量在初始向量的基础上通过以下两层图注意力模型进行更新，第一层图注意力更新的公式具体为：

$$\alpha_{ij}=\frac{\exp\left(Leak\,\mathrm{Re}\,lu\left(a(Wh_i^0\,\|\,Wh_j^0)\right)\right)}{\sum_{k\in N(i)}\exp\left(Leak\,\mathrm{Re}\,lu\left(a(Wh_i^0\,\|\,Wh_k^0)\right)\right)}$$

$$h_i^1=\sigma\left(\sum_{j\in N(i)}\alpha_{ij}Wh_j^0\right)$$

其中α_{ij}表示节点i对节点j的注意力，h_i^1表示经过第一层GAT后节点的更新向量；节点向量更新的具体做法为，首先根据节点的二阶邻居和一阶邻居的初始向量信息，分别并行更新一阶邻居和该节点的向量，然后利用更新之后的向量，经过第二层GAT，再次更新该节点的向量；

3.4).经过上述3.3)步骤得到点对的更新向量h_i^2,h_j^2，将向量组合，得到点对之间连边的向量e_{ij}，组合方法如下：

$$e_{ij}=(h_{i1}^2h_{j1}^2,h_{i2}^2h_{j2}^2,\ldots,h_{i(d-1)}^2h_{j(d-2)}^2,h_{id}^2h_{jd}^2)$$

3.5).将上述连边向量输入逻辑回归分类器，得到该连边存在的概率值；

4).模型的训练方法为：每次输入训练集中的一批点对，由3)中的步骤计算点对之间连边存在的概率值，将各点对概率值与真实连边相比，得到该模型参数情况下的损失值，计算损失值的平均值作为这批数据的损失值，并利用梯度下降算法对模型参数进行更新；

5).利用模型训练好的参数，对新的连边进行预测，包括：对于要预测的连边，输入该连边对应的点对，输入训练好的模型中，得到该点对之间存在连边的概率值P，若$P\geq 0.5$，则预测该连边存在，否则预测为不存在；

6).在3.3)所述注意力模型中，并行计算多个注意力权重分布，在3.3)的基础上包含以下步骤：

6.1).第一层计算K_1个注意力分布，在此基础上采用平均的方式得到节点和其一阶邻居的更新向量，具体如下：

$$h_i^1 = \sigma\left(\sum_k^{K_1} \sum_{j \in N(i)} \alpha_{ij}^k W^k h_j^0\right)$$

6.2).第二层计算K_2个注意力分布,在此基础上采用拼接的方式得到节点的更新向量,具体如下:

$$h_i^2 = \Big\|_k^{K_2}\left(\sigma\left(\sum_{j \in N(i)} \alpha_{ij}^k W^k h_j^1\right)\right)$$ 。

一种对复杂网络进行链路预测的方法

技术领域

[0001] 本发明涉及深度学习与网络科学的交叉领域，具体涉及一种端到端的复杂网络链路预测模型和其分批训练方法。该模型利用注意力机制，结合网络拓扑结构，能表征网络连边。分批训练的方法使得该网络能处理大规模网络的链路预测问题。

技术背景

[0002] 大规模的复杂网络普遍存在于现实世界中，例如万维网、航空网络、在线社交网络和蛋白质网络等等。理解，预测和控制这些复杂网络是人类日益迫切的需求。复杂网络的研究属于交叉领域，即有从数学和物理角度的理论研究，也有结合计算机技术的算法研究，是当前科学领域的研究热点之一。一般情况下，复杂网络包含的连边繁多且不易被观察，人们收集的数据中不可避免的存在缺失和错误的连边；另外，限于人力物力，人们只能统计部分连边状况，不能遍历所有连边。链路预测是一种解决问题的技术，该技术使我们能在部分网络结构的基础上预测出隐藏的连边，并发现虚假的连边。在交通网络规划、在线社交、蛋白质功能等许多涉及复杂网络的领域中，链路预测技术都能带来很大的效益。传统的链路预测方法一般将网络各部分看作同质的，不区分各部分对目标节点影响力大小，这不符合实际情况，因而其预测效果也存在一定的瓶颈。

发明内容

[0003] 本发明目的在于利用注意力机制，克服上述提出的传统链路预测算法中的缺陷，提出一种基于GAT的端到端链路预测模型。该模型具有可学习的注意力权重，可以对网络不同部分分配不同的注意力大小。具体来说，本模型具有两层注意力模型，能在注意力的指导下聚合节点的一阶、二阶邻居信息，将聚合的信息组合成连边向量，再通过分类器判断该连边存在的概率值。利用训练集中的样本，指导本模型中各项参数通过梯度反传方法进行学习。用训练好的模型参数预测新节点对之间是否存在连边。另一方面，直接聚合节点所有邻居的向量需要将整个网络输入到模型中，当网络规模较大时，很难满足其所需计算机存储空间。针对于此，本发明通过对所有节点进行邻居采样，固定节点邻居数量，规避网络的幂率(power law)性质所带来的内存消耗，同时可以将单个大网络进行分批训练，提高收敛速度和GPU运算效率。

[0004] 为了实现上述目的，本发明提供了一种基于图注意力的端到端链路预测模型和其分批训练方法。所述端到端链路预测模型包括：双层图注意力模型和逻辑回归分类器；所述方法包括：对复杂网络各节点进行固定邻居采样；根据网络连边生成训练集并对其中的节点和邻居进行分批，每个节点赋予初始化向量，生成训练数据；将训练数据输入双层注意力模型，得到各点的更新向量，将点对的向量组合成连边的向量；将连边的向量通过逻辑回归得到该连边是否存在的概率值；根据损失函数对模型参数进行更新；所述链路预测模型包括训练好的双层注意力模型和逻辑回归分类器。

[0005] 上述技术方案中，所述方法具体包括：

[0006] 1).对需要处理的目标网络进行去方向去权重处理,得到网络无向无权的同质拓扑结构,该网络不能包含孤立节点。

[0007] 2).上述网络中连边对应的点对作为训练集中的正例,同时随机采集与连边数等量没有连边的点对,作为训练集中的负例。对正负例中出现的所有点进行固定数目一阶、二阶邻居采样,节点和其邻居看做整体,然后将训练集分批。

[0008] 3).构建基于GAT的端到端链路预测模型,包含以下部分:

[0009] 3.1).模型输入为点对和他们的一阶、二阶邻居,输出为该点对之间有连边的概率;

[0010] 3.2).根据网络数据实际情况,初始化节点向量为h_i^0,其中i为节点下标;

[0011] 3.3).节点向量在初始向量的基础上通过以下两层图注意力模型进行更新,第一层图注意力更新的公式具体为:

[0012] $$\alpha_{ij}=\frac{\exp\left(Leak\,\mathrm{Re}\,lu\left(a(Wh_i^0\|Wh_j^0)\right)\right)}{\sum_{k\in N(i)}\exp\left(Leak\,\mathrm{Re}\,lu\left(a(Wh_i^0\|Wh_k^0)\right)\right)}$$

[0013] $$h_i^1=\sigma\left(\sum_{j\in N(i)}\alpha_{ij}Wh_j^0\right)$$

[0014] 其中α_{ij}表示节点i对节点j的注意力,h_i^1表示经过第一层GAT后节点的更新向量。节点向量更新的具体做法为,首先根据节点的二阶邻居和一阶邻居的初始向量信息,分别并行更新一阶邻居和该节点的向量,然后利用更新之后的向量,经过第二层GAT,再次更新该节点的向量。

[0015] 3.4).经过上述步骤得到点对的更新向量h_i^2,h_j^2,将向量组合起来,得到点对之间连边的向量e_{ij},组合方法如下:

[0016] $$e_{ij}=(h_{i1}^2h_{j1}^2,h_{i2}^2h_{j2}^2,\ldots,h_{i(d-1)}^2h_{j(d-2)}^2,h_{id}^2h_{jd}^2)$$

[0017] 3.5).将上述连边向量输入逻辑回归分类器,得到该连边存在的概率值。其具体的计算过程为:

[0018] $$P_{ij}=\frac{1}{1+\exp\left(-We_{ij}+b\right)}$$

[0019] 4).模型的训练方法为:每次输入训练集中的一批点对,由3)的步骤计算概率值,将各点对概率值与真实连边相比,得到该模型参数情况下的损失值,计算损失值的平均值作为这批数据的损失值,并利用梯度下降算法对模型所有参数进行更新。

[0020] 5).所述注意力模型中,可以并行计算多个注意力权重分布,其特征在于,在3)的基础上包含以下步骤:

[0021] 5.1).第一层计算K_1个注意力分布,在此基础上采用平均的方式得到节点和其一阶邻居的更新向量,具体如下:

[0022] $$h_i^1=\sigma\left(\sum_{k}^{K_1}\sum_{j\in N(i)}\alpha_{ij}^kW^kh_j^0\right)$$

[0023] 5.2).第二层计算K_2个注意力分布,在此基础上采用拼接的方式得到节点的更新

向量,具体如下:

[0024] $$h_i^2 = \Big\|_k^{K_2}\left(\sigma\left(\sum_{j\in N(i)}\alpha_{ij}^k W^k h_j^1\right)\right)$$

[0025] 6).利用模型训练好的参数,对新的连边进行预测,其特征在于,具体包括:对于要预测的连边,输入该连边对应的点对,输入训练好的模型中,得到该点对之间存在连边的概率值P,若P＞0.5,则预测该连边存在,否则预测为不存在。

[0026] 有益效果

[0027] 1).本发明采用注意力模型对连边进行编码,使其能以一定的注意力分布整合邻居信息,克服了传统模型均匀处理网络的缺点;而且该模型是一个端到端的模型,能很方便的处理链路预测任务,较少了算法中人为干扰。

[0028] 2).本发明对网络进行固定邻居采样,从而可以将网络分批处理和训练,使得大规模网络也可以在有限的计算资源上得到处理。同时,该方法独立于网络的性质,所以具有算法层面的普适性。

[0029] 3).另外,该方法在具有上述优点的同时,在链路预测这个技术问题上取得了目前最好的精度。

附图说明

[0030] 图1为注意力机制图示,邻居向量经过GAT层得到目标节点新的向量;

[0031] 图2为节点邻居采样图示,在网络拓扑结构的基础上,对每个节点进行二阶固定邻居采样(图示为每阶邻居采样3个);

[0032] 图3为预测模型框架,首先由网络拓扑结构生成包含正负样本的训练数据,样本中的节点由图1所示的注意力机制,结合其采样邻居进行更新,然后将点对的向量组合成连边向量,最后又分类器根据连边向量判断该连边是否存在。

具体实施方式

[0033] 下面结合附图和在Cora网络上的具体实施过程对本发明做进一步阐述:

[0034] 本发明具体解决的问题是大规模复杂网络上的链接预测问题,以文献引文网络Cora数据集说明如下:

[0035] 将该数据集中的论文建模为网络上的节点,论文之间的引用关系建模为节点之间的连边,不考虑连边的方向和节点的类别,最后可以得到包含2708个节点,5429条连边的无权无向网络结构,而预测该网络中的连边对于科学学中的文献分析十分重要。本发明将网络中部分连边删除,作为要预测的连边,未删除的连边作为训练集。

[0036] 本发明采用一种基于图注意力端到端的链路预测模型和其分批训练方法,所述模型包括两层GAT(graph attention networks)模型和逻辑回归模型,所述训练方法包括节点固定邻居采样并分批以得到训练集,并分批对模型参数进行训练。

[0037] 利用分批数据训练基于图注意力的端到端的链路预测模型的具体步骤如下:

[0038] 1).如图3所示,将Cora数据集处理成一个无权无向的同质网路,并对其中每一个点的邻居进行固定数目采样,固定数目建议为15～25,若邻居总数多于固定数目,则随机采所需样本,否则可以重复采样。实际过程中,一阶邻居采样数目和二阶邻居采样数目可以不

一样；

[0039] 2).如图3所示，在上述Cora网络拓扑结构的基础上，有连边的点对作为正例，并随机采样等量的没有连边的点对作为负例，组成训练集。以Cora数据集为例，训练集包含约20,000个点对，将训练集分批，以待训练之用。每一批数据可以包含32～256个点对；

[0040] 3).对于一个点对，其中两个点有初始向量经过两次更新得到输出向量，分别对应于图3中的GAT1和GAT2，对于一批训练数据，上述过程可以并行运算；

[0041] 4).更新之后的点对的向量组合之后得到其连边的向量表示，然后该向量输入逻辑回归(Logistic Regression)得到其边存在的概率值，该概率值与真实连边做交叉熵可以得出预测的损失值。求出一批数据的平均损失值，并用梯度下降算法更新模型参数；

[0042] 5).一次循环中，针对步骤3～4，遍历所有分批数据来训练参数。整个训练过程循环多次。对于Cora数据集对应的网络，循环50次左右便可以训练完成。

[0043] 6).将未在训练集中出现的点对输入训练好的模型中进行预测，便可输出该点对有连边的预测概率值。对于Cora这个数据集，我们将数据预处理部分断开的连边，同时再采样等量的负例作为预测集，最后的再测试集上的连边预测准确度可以达到87％，在链路预测任务上是目前最好的方法。

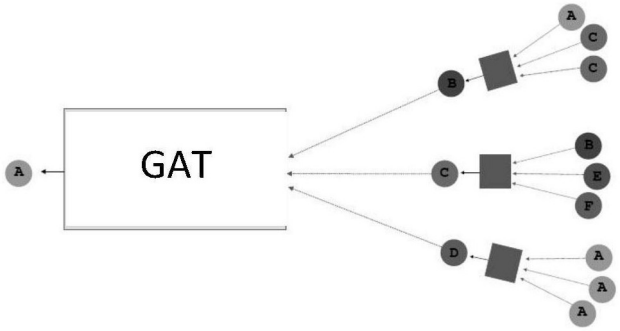


图1

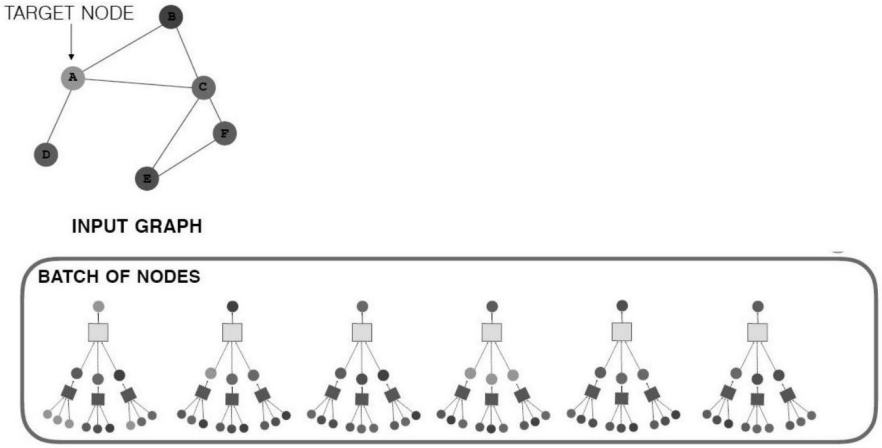


图2

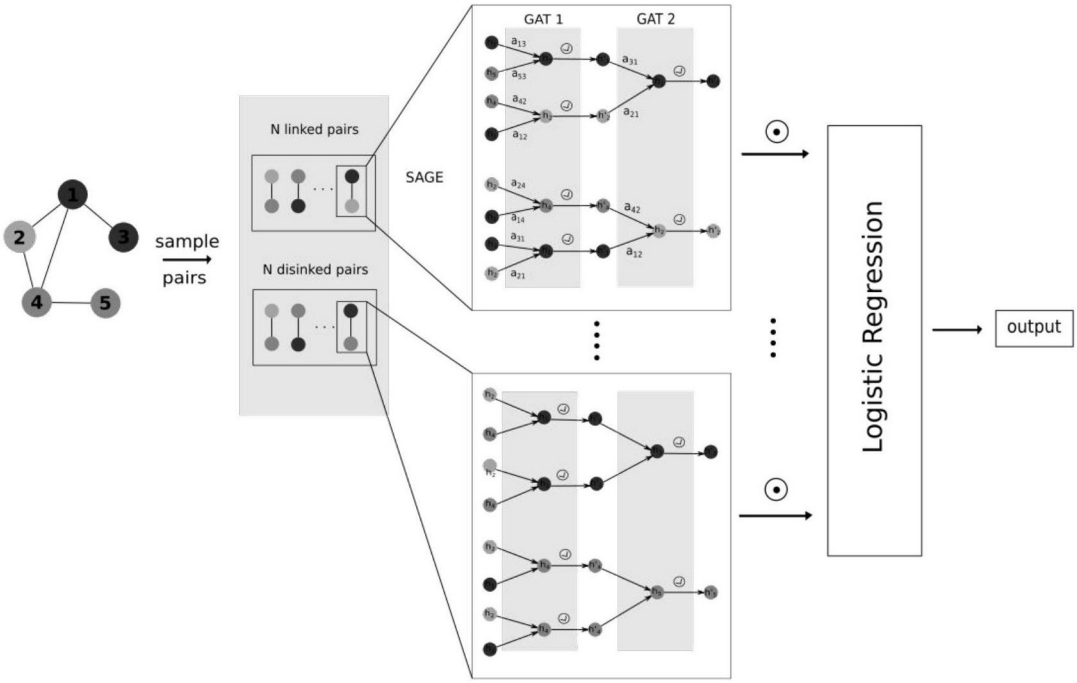

1
2
3
4
5
sample
pairs
N linked pairs
N disinked pairs
SAGE
GAT 1
GAT 2
Logistic Regression
output

图3