基于嵌套生成对抗学习的网络嵌入

沈鹏飞[1],徐 臻[1],王 英[2]
(1. 中国电子科技南湖研究院,浙江嘉兴 314001; 2. 吉林大学计算机科学与技术学院,吉林长春 130012)

摘 要: 当前网络嵌入研究更多关注信息网络结构和结点之间一阶或高阶近似关系,对于网络结点自身属性考虑较少. 本文提出一种嵌套的生成对抗网络模型 N-GAN(Nesting Generative Adversarial Networks for Network Embedding),实现了网络结构和节点属性同时嵌入到低维向量,从而最大程度保存原始高维信息网络特征. N-GAN 模型设计灵活,具有很好的延伸性和扩张性,并在真实数据上验证了 N-GAN 的性能及其稳定性,其嵌入的低维表示在不同应用中表现出不错的性能.

关键词: 数据挖掘; 网络嵌入; 生成对抗学习; 信息网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)09-2155-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210592

Network Embedding Based on Nested Generative Adversarial Networks

SHEN Peng-fei[1], XU Zhen[1], WANG Ying[2]
(1. *China Nanhu Academy of Electronics and Information Technology*, *Jiaxing*, *Zhejiang* 314001, *China*;
2. *College of Computer Science and Technology*, *Jilin University*, *Changchun*, *Jilin* 130012, *China*)

Abstract: The current network embedding researches focus more on the information network structure and first-order or higher-order approximation of nodes, but less on the attributes of network nodes. This paper proposes a nested generative adversarial network model N-GAN(Nesting Generative Adversarial Networks for Network Embedding), which embeds the network structure and nodes' attributes into the low-dimensional vector at the same time, so as to preserve the feature of the original high-dimensional information network maximumly. N-GAN model is flexible in design and has good extensibility and expansibility. The performance and stability of N-GAN model are verified on real datasets. The embedded low-dimensional representation of N-GAN model shows good performance in different tasks.

Key words: data mining; network embedding; generative adversarial learning; information network

1 引言

信息实体数量不断增加导致信息数据极速增长,从如此海量数据发掘有价值信息,则需对数据进行合理整合才能高效分析[1~3]. 由于网络结构具有融合不同信息实体和关系的天然优势,已成为大数据分析的重要数据结构,且在数据挖掘和人工智能的相关研究中,无不对数据特征精确提取和分析,但随着信息维度不断增加,数据网络结构也会变得异常复杂,网络嵌入将很好为这两项研究做好基础工作[4~6]. 网络嵌入的研究目标就是将高维复杂的信息网络用低维向量表示,且低维表示包含的有效信息越多,则嵌入越成功,可应用于的任务也更多[7~9].

在当前网络嵌入的研究中,大部分工作都是更多考虑网络结构特征和结点之间的近似关系. 但信息网络中不只表达结点结构,还包含了大量潜在信息,如结点局部关系和自身属性等. 针对以上情况,本文借鉴生成对抗学习思想提出了一种嵌套的生成对抗网络模型 N-GAN(Nesting Generative Adversarial Networks for Network Embedding),将网络结构、结点对近似关系以及结点属性信息,通过一个三级对抗模型逐级学习,将不同特征信息嵌入在一个低维的表示中. 由此不仅将网络结构和结点对的近似关系嵌入在低维表示中,还将结点属性特征也嵌入到表示向量中,从而保留了原始网

收稿日期:2021-05-10;修回日期:2021-08-23;责任编辑:梅志强
基金项目:国家自然科学基金(No.61872161)

络更丰富的数据信息[10~12]. N-GAN 模型中第一级生成对抗模型学习保存网络的结构特征,在第一级模型基础上第二级对抗网络学习结点对属性信息,第三级网络在前两级基础上学习结点局部特征信息,每一级都是确保上一级特征充分学习的基础上再学习新的特征信息,从而使得不同特征信息平滑融入低维表示向量中. 本文主要做了以下三方面的工作:

(1) 根据信息网络的结构数据以及结点的属性,依次计算了结点对的三级特征,作为嵌入模型学习数据.

(2) 借鉴生成对抗网络的思想,通过嵌套的方式提出一个三级组合的生成对抗网络结构 N-GAN,并且详细给出了模型的运算理论和各个子模型的网络结构.

(3) 在真实数据集上设计了多组实验,与相关模型进行了比较分析,验证了 N-GAN 算法性能,并讨论了算法稳定性及效率.

2 相关研究

当前网络嵌入的研究大致可以划分为两类,分别为基于关系矩阵分解[13]和基于深度神经网络方法. 关系矩阵一般是网络结点的邻接矩阵、拉普拉斯阵、PPMI (Positive Pointwise Mutual Information)矩阵,基于关系矩阵的网络嵌入的研究应该可以追溯到很早,从谱聚类算法到矩阵的非负分解都是将高维的网络用低维向量表示. 矩阵分解主要是针对邻接矩阵和 PPMI 矩阵,文献[14]通过对社交网络中结点间邻接关系矩阵的非负分解得到网络的低维向量表示,从而预测用户之间的信任关系. Cheng 等人[15]借鉴矩阵分解的思想对符号社交网络提出一种非监督的特征提取方法,实现了在符号网络中结点低维表示之间的近似关系与高维网络中一阶近似和二阶近似相一致. Wang 等人[16]提出通过最优化矩阵的非负分解模型同时实现网络结点的社区发现和网络表示,实现结点的低维表示可以同时保存网络的局部结构和社区结构. Qiu 等人[17]发现了 Deepwalk[18]产生的隐式矩阵和 Laplace 图之间新的理论联系,并且指出 Deepwalk, LINE (Large-Scale Information Network Embedding)[19], PTE (Predictive text embedding)[20]和 node2vec[21]本质上是隐式的矩阵分解. 但是随着网络规模的增加,矩阵分解的时空复杂性会限制算法效率.

在以往的研究中,神经网络模型大致可以分为生成模型和辨别模型. Goodfellow 等人[22]创造性地提出了生成对抗网络 GANs(Generative Adversarial Nets),将生成模型和辨别模型结合起来,形成一个统一的对抗学习模型. Mirza 等人[23]在 GANs 的基础上对目标函数做了改进,提出了一个基于条件概率的生成对抗网络模型 (Conditional Generative Adversarial Nets, CGANs). Tolstikhin 等人[24]通过训练权值和弱生成器的方式提出一种新的生成对抗模型 AdaGAN(Adaptive GAN),AdaGAN 模型每次都是由上一次的弱生成器和训练好的权值以加权的方式产生新的生成器. Arjovsky 等人[25]用 Wasserstein 距离代替传统生成对抗网络中的 KL 散度,提出了 WGAN(Wasserstein GAN),WGAN 成功解决了传统生成对抗网络梯度消失、训练不稳定、模式崩溃等缺点. Mao 等人[26]通过在对抗训练的目标函数中使用最小二乘法提出了 LSGAN(Least Squares Generative Adversarial Networks),LSGAN 不仅克服了训练过程不稳定,梯度消失的缺点,且 LSGAN 收敛速度更快.

3 数据分析

本文收集了三个真实数据集:arXiV-AstroPh、arXiv-GrQc、Cora. arXiV-AstroPh 是在 arXiv 在线电子版上一个关于论文作者之间的科研合作关系的网络. arXiv-GrQc 是 arXiv 上关于广义相对论和量子宇宙学范畴的论文作者之间的科研合作关系网络. Cora 数据集包括两部分:一部分是论文之间的引用关系,另一部分是论文的类别数据,每一篇论文都有一个类别标签. arXiV-AstroPh 和 arXiv-GrQc 的相关统计信息如表 1 所示,Cora 相关统计信息如表 2 所示.

表 1 数据集 arXiV-AstroPh 和 arXiv-GrQc 信息统计

数据集	arXiV-AstroPh	arXiv-GrQc
结点个数	18772	5242
链路个数	198110	14496
结点最大度数	504	81
平均聚类系数	0.6306	0.5296
网络密度	0.0011	0.0010

表 2 数据集 Cora 信息统计

数据集	Cora
结点个数	2708
链路个数	5278
结点类别数	7
结点最大度数	168
同一类别结点最大个数	818
同一类别结点最小个数	180
网络密度	0.0014

本文对数据集做了简单的预处理,首先只关注不同结点之间的联系因此删除了有自环的链接,其次为了确保结点在网络中有足够的结构信息,本文删除了没有链接和只有一条链接的结点. 由表 1 和表 2 可以计算出 arXiV-AstroPh 和 arXiv-GrQc 中平均每个结点分别有 10.55 和 2.77 个链接,在 Cora 中平均每个结点有 1.95 个链接. 三个数据集的网络密度分别是 0.0011、0.0010、0.0014.

4　问题定义

网络结构的组成元素就是结点和边,本文用 $G=\{V,E\}$ 表示网络,其中 $V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3,\cdots,\boldsymbol{v}_n\}$ 是结点的集合,$E=\{\boldsymbol{e}_{12},\boldsymbol{e}_{13},\cdots,\boldsymbol{e}_{ij},\cdots,\boldsymbol{e}_{nx}\}$ 是网络中所有边的集合. 网络嵌入的目标就是通过嵌入算法将网络信息投影在一个低维的空间,网络嵌入的研究问题可以如式(1)所示:

$$G=\{V,E\}\xrightarrow{\text{N-GAN}}\boldsymbol{U} \tag{1}$$

其中,$\boldsymbol{U}\in\mathbb{R}^{n\times d}$,$d$ 是嵌入维度且 $d<n$,$\boldsymbol{U}$ 表示网络 G 的低维表示.

5　N-GAN模型

为了能对N-GAN模型有一个直观的了解,N-GAN模型的架构如图1所示.

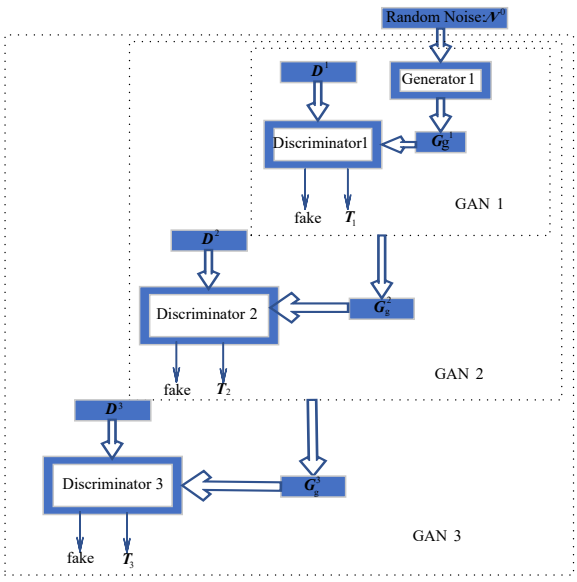

图1　N-GAN模型架构说明图

由图1可知,N-GAN模型通过嵌套组合的方式构成一个大的生成对抗网络,并且其中的每级内部也是一个生成对抗结构. 整个模型中只有一个生成器,三个辨别器分属于三个对抗网络,其中GAN1、GAN2、GAN3分别表示三个生成对抗网络. 三个辨别器分别学习不同的网络特征,其中 $\boldsymbol{D}^1$、$\boldsymbol{D}^2$、$\boldsymbol{D}^3$ 表示三个辨别器的真实输入数据,$\boldsymbol{G}_{\mathrm{g}}^{\ 1}$、$\boldsymbol{G}_{\mathrm{g}}^{\ 2}$、$\boldsymbol{G}_{\mathrm{g}}^{\ 3}$ 表示生成器网络在三次递进学习过程中的生成数据.

5.1　N-GAN模型工作原理

N-GAN模型中生成器网络逐级学习,每一级的辨别器只学习信息网络的一个特征. 生成器首先和第一级的辨别器对抗学习,生成器不断调整使得生成数据的分布接近网络的第一特征,然后再和第二级对抗网络对抗学习,以此使得生成器逐步融合网络结点对不同属性特征.

首先介绍GAN1网络的结构和工作原理. 网络结构是网络嵌入首要考虑的重要特征,因此将N-GAN的第一级生成对抗网络用于学习网络的结构特征. 为了表示网络结点对之间的结构信息,本文通过结点对之间的最短路径表示了结点对关系矩阵 $\boldsymbol{X}^{\mathrm{S}}$ 如下所示:

$$\boldsymbol{X}^{\mathrm{S}}=\begin{pmatrix}\delta_{11}&\delta_{12}&\cdots&\delta_{1n}\\ \delta_{21}&\delta_{22}&\cdots&\delta_{2n}\\ \vdots&\vdots&\ddots&\vdots\\ \delta_{n1}&\delta_{n2}&\cdots&\delta_{nn}\end{pmatrix} \tag{2}$$

$$\delta_{ij}=\frac{1}{s_{ij}} \tag{3}$$

其中,s_{ij} 表示网络中结点 i 和结点 j 之间的最短路径. 很明显结点对之间最短路径越小,$\boldsymbol{X}^{\mathrm{S}}$ 矩阵中对应的元素值越大. 生成器网络中,本文首先从网络全局角度计算结点对余弦相似度表示低维空间中的邻近程度. 生成器目标就是将生成结点的低维表示之间的相似度分布无限接近 $\boldsymbol{X}^{\mathrm{S}}$. 在GAN1中比较特殊的是 $\boldsymbol{D}^1$ 和 $\boldsymbol{T}_1$ 都等于 $\boldsymbol{X}^{\mathrm{S}}$. 为了避免传统生成对抗网络训练不稳定导致的模式崩溃,本文选择了Wasserstein距离计算数据分布差异. 如果将生成器和辨别器分别表示为参数化函数 $G(\cdot)$ 和 $D_1(\cdot)$,则GAN1的目标函数可定义如下:

$$\min_{G}\max_{D_1}L_{\mathrm{GAN1}}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}[D_1(\boldsymbol{x})]-\mathbb{E}_{z\sim p_z(z)}[D_1(G(\boldsymbol{z}))] \tag{4}$$

其中,$\boldsymbol{x}$ 表示真实数据,$\boldsymbol{z}$ 表示生成数据样本,初始为随机数据. 由此可见随着 $G(\cdot)$ 和 $D_1(\cdot)$ 之间不断的对抗学习,生成器生成结点的表示数据基本保留了结点对在网络中的邻近关系,即N-GAN模型通过GAN1网络保存了网络的结构信息.

本文基于结点关系矩阵 $\boldsymbol{X}^{\mathrm{S}}$ 和结点对相似性定义了一个结点对近似关系矩阵 $\boldsymbol{N}^{\mathrm{P}}$,计算如式(5)~(7)所示:

$$\boldsymbol{N}^{\mathrm{P}}=\boldsymbol{X}^{\mathrm{S}}+\lambda\times\boldsymbol{X} \tag{5}$$

$$\boldsymbol{X}=\begin{pmatrix}\eta_{11}&\eta_{12}&\cdots&\eta_{1n}\\ \eta_{21}&\eta_{22}&\cdots&\eta_{2n}\\ \vdots&\vdots&\ddots&\vdots\\ \eta_{n1}&\eta_{n2}&\cdots&\eta_{nn}\end{pmatrix} \tag{6}$$

$$\eta_{ij}=\frac{|\boldsymbol{M}(i,:)\cap\boldsymbol{M}(j,:)|}{|\boldsymbol{M}(i,:)\cup\boldsymbol{M}(j,:)|} \tag{7}$$

其中,$\boldsymbol{M}$ 为结点邻接矩阵,式(7)的分子部分表示结点 $\boldsymbol{V}_i$ 和 $\boldsymbol{V}_j$ 的共同邻居的数目,分母部分表示两结点所有邻居的数目,λ 是一个超参数取0.25,用于调和 $\boldsymbol{X}^{\mathrm{S}}$ 和 $\boldsymbol{X}$ 之间关系,希望能更合理反映结点对之间的近似关系. 相关研究发现网络结点的度是一个重要属性,对于反映结点相关特征具有积极作用[27],本文在 $\boldsymbol{N}^{\mathrm{P}}$ 基础上加入结点的属性定义了 $\boldsymbol{X}^{\mathrm{Sim}}$,如下所示:

$$\boldsymbol{X}^{\mathrm{Sim}}(i,j)=\frac{1}{2}\left(\frac{\boldsymbol{N}^{\mathrm{P}}(i,j)}{\boldsymbol{d}(i)}+\frac{\boldsymbol{N}^{\mathrm{P}}(j,i)}{\boldsymbol{d}(j)}\right) \tag{8}$$

其中$\boldsymbol{d}(i)$表示结点$\boldsymbol{V}_i$的度. 由此可见$\boldsymbol{X}^{\mathrm{Sim}}$中元素值是在对应的结点对直接近似和间接近似的基础上加入了结点度的影响,通过结点度数调节结点对之间的近似关系.

随着生成器在GAN1中训练,逐渐学习了网络的结构特征,N-GAN模型希望在保证生成数据保留结构特征的前提下继续学习新的特征. 因此将$\boldsymbol{X}^{\mathrm{Sim}}$作为网络结点对的新的特征,通过GAN2将生成器和第二级辨别器对抗学习,其中第二级辨别器的输入$\boldsymbol{D}^2$为$\boldsymbol{X}^{\mathrm{S}}$,输出$\boldsymbol{T}_2$为$\boldsymbol{X}^{\mathrm{Sim}}$. 这是因为经过GAN1的训练之后,生成器生成的数据分布接近于$\boldsymbol{X}^{\mathrm{S}}$,所以第二级辨别器目的就是学习从$\boldsymbol{X}^{\mathrm{S}}$到$\boldsymbol{X}^{\mathrm{Sim}}$的映射,从而在对抗学习的过程中促进生成器在保证满足上一级网络的基础上生成具有$\boldsymbol{X}^{\mathrm{Sim}}$特征的表示. 如果将经过GAN1训练好的生成器函数表示为$G_1(\cdot)$,第二级辨别器函数表示为$D_2(\cdot)$,则GAN2的目标函数可表示为:

$$\min_{G_1}\max_{D_2} L_{\mathrm{GAN2}}=\mathbb{E}_{\boldsymbol{y}\sim p_{\mathrm{data}}(\boldsymbol{y})}[D_2(\boldsymbol{y})]-\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[D_2(G_1(\boldsymbol{z}))] \tag{9}$$

其中,$\boldsymbol{y}$表示真实数据$\boldsymbol{X}^{\mathrm{S}}$,$D_2(\boldsymbol{y})$是辨别器学习到的数据,$\boldsymbol{z}$表示生成数据样本.

生成器经过前两级对抗训练,逐步学习了网络结构特征和具有结点属性的近似关系. 但结点之间的关系衡量不能单纯的依赖结点对之间的最短路径,根据物以类聚,人以群分的特点,考虑这样一种情况,如果结点$\boldsymbol{V}_i$有六个邻居结点,其中五个邻居结点聚集更密集,另外一个结点相对疏远,则结点$\boldsymbol{V}_i$向五个密集邻居结点靠拢的可能性更大. 本文计算了每个结点被其邻居结点的吸引程度,吸引力越强,则两个结点之间的关系越亲密. 为定义这种邻居结点的吸引力,假设任意结点$\boldsymbol{V}_i$的邻居结点表示为$N_{V_i}=\{\boldsymbol{V}_{ia},\boldsymbol{V}_{ib},\boldsymbol{V}_{ic},\boldsymbol{V}_{id}\}$. 为了计算各个结点对结点$\boldsymbol{V}_i$的吸引程度,本文首先计算每一个邻居结点与其他所有邻居的结点的邻近程度,以结点$\boldsymbol{V}_{ib}$为例,假设用p_b表示$\boldsymbol{V}_{ib}$与其他所有邻居结点的邻近量,计算方法如下:

$$p_b=\boldsymbol{X}^{\mathrm{S}}(b,a)+\boldsymbol{X}^{\mathrm{S}}(b,\mathrm{c})+\boldsymbol{X}^{\mathrm{S}}(b,d) \tag{10}$$

计算p_a,p_c,p_d的方法和式(10)类似. 由此可得结点$\boldsymbol{V}_i$的邻居结点与其余邻居的邻近量表示为$\boldsymbol{P}_{V_i}=(p_a,p_b,p_c,p_d)$,根据该向量本文定义了不同邻居结点对结点$\boldsymbol{V}_i$的吸引程度,假设$\mathrm{grav}(i,b)$表示结点$\boldsymbol{V}_b$对结点$\boldsymbol{V}_i$的吸引程度,则计算如下:

$$\mathrm{grav}(i,b)=\frac{p_b}{\min(\boldsymbol{P}_{V_i})} \tag{11}$$

计算其他邻居结点对结点$\boldsymbol{V}_i$的吸引程度与此类似. 由此本文结合网络结点的邻接矩阵$\boldsymbol{X}^{\mathrm{D}}$定义了一个邻居结点吸引力矩阵$\boldsymbol{Y}$,具体如下:

$$\boldsymbol{Y}=\begin{pmatrix}\theta_{11}&\theta_{12}&\cdots&\theta_{1n}\\ \theta_{21}&\theta_{22}&\cdots&\theta_{2n}\\ \vdots&\vdots&\ddots&\vdots\\ \theta_{n1}&\theta_{n2}&\cdots&\theta_{nn}\end{pmatrix} \tag{12}$$

$$\theta_{ij}=\begin{cases}\phi, & \text{if } \boldsymbol{X}^{\mathrm{D}}(i,j)=0\\ \mathrm{grav}(i,j), & \text{if } \boldsymbol{X}^{\mathrm{D}}(i,j)=1\end{cases} \tag{13}$$

其中ϕ是超参数,为0.5,本文结合矩阵$\boldsymbol{Y}$和$\boldsymbol{X}^{\mathrm{Sim}}$定义了网络中结点之间的聚集程度,用矩阵$\boldsymbol{X}^{\mathrm{J}}$表示,其中$\boldsymbol{X}^{\mathrm{J}}$的计算如下,$\odot$表示矩阵的Hadamard积.

$$\boldsymbol{X}^{\mathrm{J}}=\boldsymbol{X}^{\mathrm{Sim}}\odot\boldsymbol{Y} \tag{14}$$

在第三级生成对抗网络中,第三级辨别器的输入$\boldsymbol{D}^3$为$\boldsymbol{X}^{\mathrm{Sim}}$,输出$\boldsymbol{T}^2$为$\boldsymbol{X}^{\mathrm{J}}$. 通过生成器和第三级辨别器的对抗学习,使得第三级辨别器学习从$\boldsymbol{X}^{\mathrm{Sim}}$到$\boldsymbol{X}^{\mathrm{J}}$的映射,从而使得生成器生成结点数据分布在满足前两级特征的基础上不断接近$\boldsymbol{X}^{\mathrm{J}}$. 如果将经过GAN2训练的生成器表示为$G_2(\cdot)$,三级辨别器表示为$D_3(\cdot)$,则GAN3的目标函数可表示为:

$$\min_{G_2}\max_{D_3} L_{\mathrm{GAN3}}=\mathbb{E}_{\boldsymbol{w}\sim p_{\mathrm{data}}(\boldsymbol{w})}[D_3(\boldsymbol{w})]-\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[D_3(G_2(\boldsymbol{z}))] \tag{15}$$

其中,$\boldsymbol{w}$表示真实二级特征数据$\boldsymbol{X}^{\mathrm{Sim}}$,$D_3(\boldsymbol{w})$是辨别器学习到的数据.

根据以上过程可知,随着每一级的生成对抗网络的嵌套,生成器学习能力逐级递增,最终生成的低维表示融合了更丰富的网络特征数据.

5.2 N-GAN算法描述

上文对N-GAN设计结构和运行原理作了详细介绍,本节将对N-GAN算法运行过程进行描述,具体如算法1所示.

从算法过程可以看出N-GAN模型将生成器嵌套在三个生成对抗网络中,逐级学习,并且每一级学习新的特征的过程总是一个平滑过渡,最终将三个特征矩阵的信息融合在低维表示中.

5.3 实验验证

为验证N-GAN模型的有效性,本文选择了两个应用:链路预测和网络可视化,同时选取七个方法模型作为对比. 本文将对比模型和N-GAN对同一数据集进行嵌入计算,然后将不同模型计算的低维表示应用在实验任务中,通过实验结果分析算法性能.

Deepwalk:首个在网络嵌入时将网络结构通过随机游走转换成结点序列,并且利用Skip-gram学习结点的低维表示.

LINE-O1:提出了计算结点对之间相关近似的方法. LINE-O1是基于结点之间的一阶近似的LINE.

算法 1　N-GAN

输入:λ,超参数,lr,学习率,k,嵌入维度,h,批量数, 结点对关系矩阵 $\boldsymbol{X}^{\mathrm{S}}$, $\boldsymbol{X}^{\mathrm{Sim}}$ 和 $\boldsymbol{X}^{\mathrm{J}}$, $\boldsymbol{N}^{0}$,随机数据,T,辨别器网络的训练次数. m, 模型训练迭代数(epochs)

输出:生成器 $G(\theta_G)$,辨别器 $D_1(\theta_{D_1})$,$D_2(\theta_{D_2})$,$D_3(\theta_{D_3})$

1: 随机数据初始化参数 $\theta_G,\theta_{D_1},\theta_{D_2},\theta_{D_3}$

2:for epoch = 1:m do

3:　生成器生成数据↦ $\boldsymbol{G}_{\mathrm{g}}^{1}$

4:　$\boldsymbol{X}^{\mathrm{S}}$中按批量抽取结点的关系数据

5:　$\boldsymbol{G}_{\mathrm{g}}^{1}$中按批量抽取结点的生成数据

6:　极小极大化 $\mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[D_1(\boldsymbol{x})\right]-\mathbb{E}_{z\sim p_z(z)}\left[D_1(G(\boldsymbol{z}))\right]$更新参数 θ_G 和 θ_{D_1}

7:　更新后的生成器,生成数据↦ $\boldsymbol{G}_{\mathrm{g}}^{2}$

8:　$\boldsymbol{X}^{\mathrm{S}}$中按批量抽取结点的关系数据

9:　$\boldsymbol{X}^{\mathrm{Sim}}$中按批量抽取结点的关系数据

10:　$\boldsymbol{G}_{\mathrm{g}}^{2}$中按批量抽取结点的关系数据

11:　极小极大化 $\mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[D_2(\boldsymbol{y})\right]-\mathbb{E}_{z\sim p_z(z)}\left[D_2(G_1(\boldsymbol{z}))\right]$更新参数 θ_G 和 θ_{D_2}

12:　更新后的生成器,生成数据↦ $\boldsymbol{G}_{\mathrm{g}}^{3}$

13:　$\boldsymbol{X}^{\mathrm{Sim}}$中按批量抽取结点的关系数据

14:　$\boldsymbol{X}^{\mathrm{J}}$中按批量抽取结点的关系数据

15:　$\boldsymbol{G}_{\mathrm{g}}^{3}$中按批量抽取结点的关系数据

16:　极小极大化 $\mathbb{E}_{w\sim p_{\mathrm{data}}(w)}\left[D_3(\boldsymbol{w})\right]-\mathbb{E}_{z\sim p_z(z)}\left[D_3(G_2(\boldsymbol{z}))\right]$更新参数 θ_G 和 θ_{D_3}

17:结束 for 循环

18:生成低维表示

LINE-O2：类似于 LINE-O1,通过保存结点之间的二阶近似实现网络嵌入.

struc2vec[28]：通过计算网络中结点之间的空间结构相似性实现网络结点的低维嵌入.

GAN1：N-GAN 嵌套模型中第一级生成对抗网络.

GAN2_a：N-GAN 嵌套模型中二级嵌套生成对抗网络,其中第一级辨别器学习特征 $\boldsymbol{X}^{\mathrm{S}}$,第二级辨别器学习特征 $\boldsymbol{X}^{\mathrm{Sim}}$.

GAN2_b：N-GAN 嵌套模型中二级嵌套生成对抗网络,其中第一级辨别器学习特征 $\boldsymbol{X}^{\mathrm{S}}$,第二级辨别器学习特征 $\boldsymbol{X}^{\mathrm{J}}$.

需要说明的是所有网络层均选用 leaky ReLU 激活函数,其中 leak 值为 0.2,嵌入维度 d 设置为 128,所有网络模型的优化器都为“Adam”, 学习率 lr 设置为 0.001, 超参数 λ 和 ϕ 分别设置为 0.25 和 0.5. 为了保证结果的可靠性,实验重复 5 次取平均值.

5.3.1　链路预测

链路预测是信息网络研究中的一个重要应用,通过预测结点之间可能发生的关系链路,对舆情监测、社区发现和精准推荐等任务具有很大帮助. 本文将数据集 arXiV-AstroPh 和 arXiv-GrQc 作为链路预测实验数据. 实验准备时,将实验数据网络中 20% 的边隐藏起来作为需要预测的链路,所有模型根据已知的 80% 的链路信息将网络嵌入在低维表示空间. 在求得网络低维表示后,本文选用最简单分类器 KNN(k-Nearest Neighbor)作为链路预测工具. 在链路预测前将所有结点对的低维表示相减取绝对值,作为 KNN 分类器的特征输入,KNN 的输出就是对应结点对的标签数据,其中 KNN 参数设置邻居个数为 2. 为了使预测结果更可靠,本部分选用了 10 折交叉验证,每次对 10 折验证结果求平均,并通过准确率和 F1-macro 两个指标定量分析不同模型的嵌入效果. N-GAN 和各个对比模型根据上述方法,在两个数据集上链路预测结果如表 3 和图 2 所示.

表 3　数据集 arXiV-AstroPh 和 arXiv-GrQc 上链路预测结果

Model	arXiV-AstroPh		arXiv-GrQc	
	Acc	Macro-F1	Acc	Macro-F1
Deepwalk	0.9579	0.9306	0.9389	0.9149
LINE-O1	0.9250	0.8980	0.9120	0.8932
LINE-O2	0.9143	0.8712	0.9008	0.8655
struc2vec	0.9184	0.8393	0.9140	0.8352
N-GAN	0.9411	0.9158	0.9251	0.9031
GAN1	0.9307	0.9037	0.9137	0.8902
GAN2_a	0.9389	0.9134	0.9191	0.8970
GAN2_b	0.9382	0.9122	0.9202	0.8988

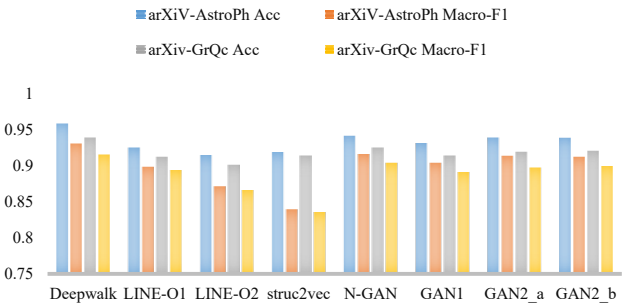

图 2　不同模型嵌入向量的链路预测分布

根据以上结果可以清楚的观察到,N-GAN 模型计算的低维表示向量在链路预测的任务表现出了相对不错的性能. 在两个数据集中预测准确率分别到 0.9411 和 0.9251,F1-macro 分别达到 0.9158 和 0.9031. N-GAN 计算的低维表示在链路预测时的准确率平均比 GAN1 高 0.0109,比 GAN2_a 高 0.0041,比 GAN2_b 高 0.0039,比 LINE-O1 高 0.0146,比 LINE-O2 高 0.0256,比 struc2vec 高 0.0169;F1-macro 指标平均比 GAN1 高 0.0125,比 GAN2_a 高 0.0043,比 GAN2_b 高 0.0040,比 LINE-O1 高 0.0139,比 LINE-O2 高 0.0411,比 struc2vec 高 0.0722. 同时可知 N-GAN 结果优于 GAN2_a、GAN2_b 两个二级嵌套模型,GAN2_a、GAN2_b 的结果又优于

GAN1,因此可以证明经过嵌套学习三级特征保存了相对较多的有效信息.

5.3.2 网络可视化

网络可视化是检验网络嵌入算法有效性的另一个重要应用. 本文在Cora数据集中选择了三个类结点作为本部分实验的验证数据,且选用PCA作为降维工具. 可视化结果如图3所示,其中图中红色菱形表示"强化学习",绿色三角形表示"基于案例",蓝色星号表示:"理论".

由图3可直观的观察到三类结点不同模型计算的低维表示的分布,很明显GAN1、GAN2、N-GAN三个模型计算的结点的表示向量具有相对更好的质量,但图3(*d*)、(*e*)两个子图在同类节点的集中程度和不同类别结点的疏远程度都没有N-GAN效果好. 从图3(*f*)子图可以看出,三类结点分布在三个区域,并且同一类别相对集中,不同类别的结点在显示空间中相对较远. 在这项任务中. 从图3(*a*)子图可以看到Deepwalk基本将不同类别的结点集中在一起,但是"基于案例"和"理论"两个类型的结点重合较多. LINE-O1计算的表示向量相比Deepwalk有较好的表现,将三类结点很好的分布在三个区域,但是三个区域的接壤部分重合太多,各类别结点也不够集中.

5.3.3 参数设置

深度网络模型在训练过程中,参数选择会成为影响模型性能的关键因素. 为解释N-GAN中相关参数的选择过程,本部分以链路预测为例,设置了多组实验验证不同参数取值对模型性能的影响. N-GAN模型中主要有ϕ和λ两个超参数和迭代轮次epoch. 因篇幅有限,此部分以ϕ的选择过程为例,λ过程类似. 为了探究不同的ϕ值在不同训练次数下对N-GAN的性能影响,本文将N-GAN中epoch分别设置为$\{50,150,250,500\}$,ϕ分别设置为$\{0.01,0.1,0.3,0.5,0.7,1.5\}$,在两个数据集上链路预测结果如图4所示. 当epoch从50增加到150时,不论ϕ取何值,N-GAN模型的性能都明显变好,这说明随着训练深入模型得到了充分的学习. 当epoch继续增加,N-GAN嵌入能力也在逐步改善,但幅度不大. 不考虑epoch影响,当ϕ从0.01增加到0.5时,N-GAN模型的性能在逐渐变好,但ϕ大于0.5之后,模型的性能有所下降,因此综合两个数据集考虑运行时间和算法性能,本文将迭代轮次epoch和ϕ分别取值250和0.5,使模型效率达到相对最优.

5.3.4 算法收敛性及效率分析

由上文可知N-GAN模型是一个嵌套的生成对抗网络,随着三级网络的嵌套,很明显增加了N-GAN网络模型的深度,本节通过分别分析三级网络的收敛性,讨论N-GAN算法整体的收敛性. 为了直观的展示

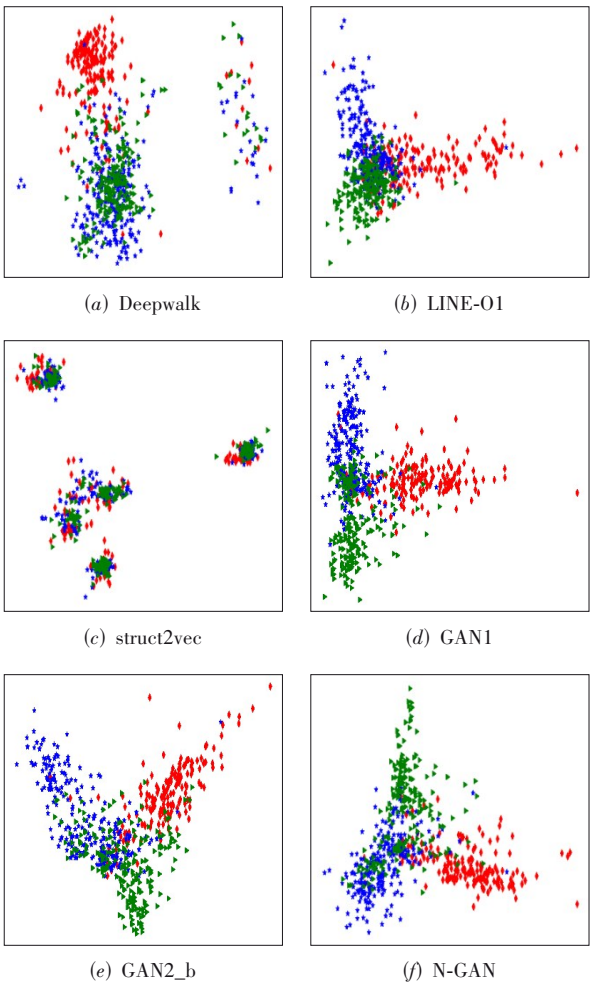

图3 可视化结果

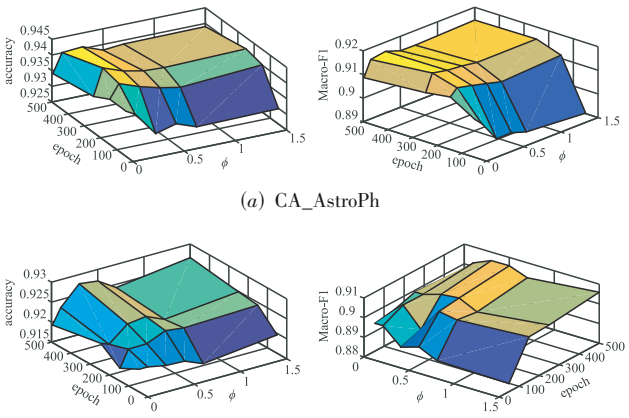

图4 超参数ϕ性能影响分析

各个生成对抗网络的收敛性,以迭代次数为横轴,三级网络的目标函数误差损失值为纵轴,分别对三级网络作图,如图5所示. GAN1随着迭代次数不断增加,函数损失值不断降低,收敛性并不是很好,这是为了

同时观察三级网络的收敛情况,此部分迭代次数设置较少,在实际实验过程中,当迭代次数超过200后GAN1误差损失的变化范围将维持在一个非常小的范围. GAN2当迭代次数从1增加到20时函数损失值缓慢降低,当迭代次数超过20后GAN2的误差损失基本趋于稳定. GAN3(相当于N-GAN)在迭代次数从1增加到25时函数损失值降低较快,当迭代超过25次后GAN3的误差损失趋于稳定. 综上所述,本文提出模型N-GAN在总体上保持了良好的收敛性,确保了算法运行的稳定.

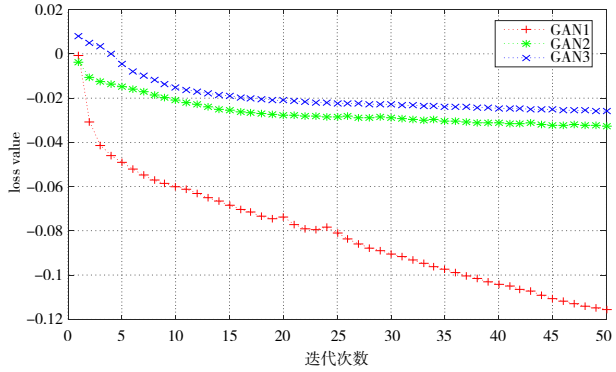

图5　N-GAN算法收敛性分析

由于神经网路训练参数较多导致训练时间较长,本文模型通过一种嵌套得方式增加了训练深度,导致N-GAN在运行效率上有所下降. 为探究嵌套模型运行效率,本文统计了GAN1、GAN2、GAN3多组迭代轮次的训练时间,如图6所示. 从图6可知,随着嵌套加深,模型时间复杂度在逐级增加,因此嵌套级数的选择不能无限制叠加,需要综合考虑性能和效率,增强模型实用性.

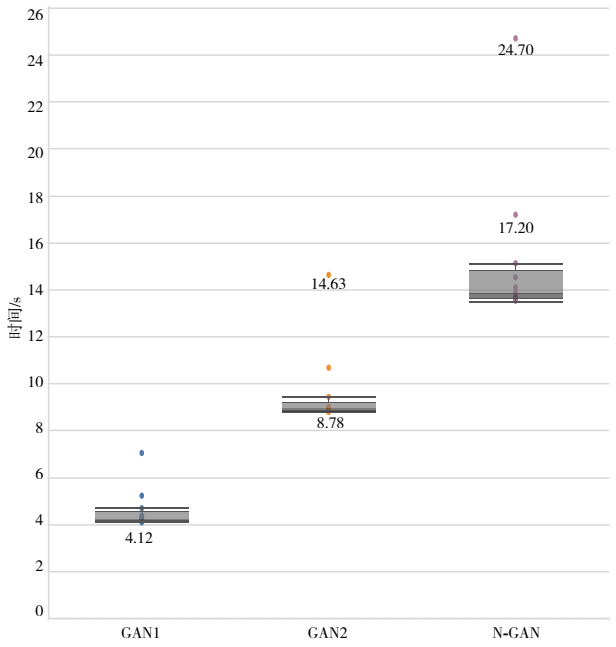

图6　模型运行时间分析

6　结束语

本文提出一种嵌套的生成对抗结构,通过嵌套的方式将三个生成对抗网络组合起来,形成一种新的生成对抗学习结构N-GAN. 在N-GAN中每一个子模型内部是对抗学习,模型与模型之间也是对抗学习,由此实现了逐级学习网络不同特征信息,最终将不同特征信息平滑融合在低维表示中. 本文在真实数据上根据两个应用任务设计了多组实验,验证了N-GAN算法的有效性. 在下一步研究中,将结合生成对抗学习模型和强化学习,进一步提高网络嵌入模型效率.

参考文献

[1] TANG J. Computational models for social network analysis: A brief survey[C]//Proceedings of the 26th International Conference on World Wide Web Companion- WWW' 17 Companion. New York: ACM Press, 2017: 921-925.

[2] CEN Y K, ZOU X, ZHANG J W, et al. Representation learning for attributed multiplex heterogeneous network [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 1358-1368.

[3] SUN G L. New progress in research and application of machine learning[J]. Chinese Journal of Electronics, 2020, 29 (6): 991-991.

[4] SUN X, SONG Z H, DONG J Y, et al. Network structure and transfer behaviors embedding via deep prediction model[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 5041-5048.

[5] 刘露, 胡封晔, 牛亮, 等. 异质网络中基于节点影响力的相似度度量方法[J]. 电子学报, 2019, 47(9): 1929-1936.
LIU L, HU F Y, NIU L, et al. Node influence based similarity measure method in heterogeneous network[J]. Acta Electronica Sinica, 2019, 47(9): 1929-1936. (in Chinese)

[6] WANG Z, YE X, WANG C, et al. RSDNE: Exploring relaxed similarity and dissimilarity from completely-imbalanced labels for network embedding[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 475-482.

[7] LAI Y Y, NEVILLE J, GOLDWASSER D. TransConv: relationship embedding in social networks[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelli-

gence. Palo Alto, CA: AAAI Press, 2019: 4130-4138.

[8] 国琳, 左万利. 基于兴趣图谱的用户兴趣分布分析及专家发现[J]. 电子学报, 2015, 43(8): 1561-1567.
GUO L, ZUO W L. Analysis of user interest distribution and expert finding based on interest graphs[J]. Acta Electronica Sinica, 2015, 43(8): 1561-1567. (in Chinese)

[9] ASSUNÇÃO F, LOURENÇO N, MACHADO P, et al. DENSER: deep evolutionary network structured representation[J]. Genetic Programming and Evolvable Machines, 2019, 20(1): 5-35.

[10] ZHANG P Z, GONG M G, ZHANG H, et al. DRLnet: deep difference representation learning network and an unsupervised optimization framework[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 3413-3419.

[11] ZHANG D K, YIN J, ZHU X Q, et al. Homophily, structure, and content augmented network representation learning[C]//2016 IEEE 16th International Conference on Data Mining. Piscataway: IEEE, 2016: 609-618.

[12] CAO S, LU W, XU Q. Deep neural networks for learning graph representations[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 1145-1152.

[13] 张昱, 刘开峰, 张全新, 等. 基于组合-卷积神经网络的中文新闻文本分类[J]. 电子学报, 2021, 49(6): 1059-1067.
ZHANG Y, LIU K F, ZHANG Q X, et al. A combined-convolutional neural network for Chinese news text classification[J]. Acta Electronica Sinica, 2021, 49(6): 1059-1067. (in Chinese)

[14] TANG J L, GAO H J, HU X, et al. Exploiting homophily effect for trust prediction[C]//Proceedings of the sixth ACM International Conference on Web Search And Data Mining-WSDM'13. New York: ACM Press, 2013: 53-62.

[15] CHENG K W, LI J D, LIU H. Unsupervised feature selection in signed social networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 777-786.

[16] WANG X, Cui P, WANG J, et al. Community preserving network embedding [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, Palo Alto, CA: AAAI Press ,2017: 203-209.

[17] QIU J Z, DONG Y X, MA H, et al. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec[C]//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. New York: ACM, 2018: 459-467.

[18] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710.

[19] TANG J, QU M, WANG M Z, et al. LINE: large-scale information network embedding[EB/OL]. (2015-05-12). https://arxiv.org/abs/ 1503.03578.

[20] TANG J, QU M, MEI Q Z. PTE: predictive text embedding through large-scale heterogeneous text networks[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1165-1174.

[21] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]//Proceedings. International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2016: 855-864.

[22] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Cambridge, USA: MIT Press, 2014: 2672-2680.

[23] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06). https://arxiv. org/abs/ 1411.1784.

[24] Tolstikhin I, Gelly S, Bousquet O. AdaGAN: Boosting generative models[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook,NY:Curran Associates Inc, 2017: 5430-5439.

[25] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein Gan[EB/OL]. (2017-01-26). https://arxiv. org/abs/ 1701.07875.

[26] MAO X D, LI Q, XIE H R, et al. Least Squares generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2813-2821.

[27] 权宇, 李志欣, 张灿龙, 等. 融合深度扩张网络和轻量化网络的目标检测模型[J]. 电子学报, 2020, 48(2): 390-397.
QUAN Y, LI Z X, ZHANG C L, et al. Fusing deep dilated convolutions network and light-weight network for object detection[J]. Acta Electronica Sinica, 2020, 48(2): 390-397. (in Chinese)

[28] RIBEIRO L F R, SAVERESE P H P, FIGUEIREDO D R. struc2vec: learning node representations from structural identity[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 385-394.

作者简介

沈鹏飞　男,1988年生,甘肃武威人,中国电子科技南湖研究院高级工程师,研究方向:机器学习、深度学习、图神经网络、信息网络嵌入和认知智能.
E-mail: shen_pf@qq.com

徐　臻　男,1989年生,浙江衢州人,中国电子科技南湖研究院高级工程师,研究方向:认知智能、知识图谱、多智能协同和博弈.
E-mail: xuzhen@cnaeit.com

王　英(通讯作者)　女,1981年4月生,黑龙江省阿城市人,现为吉林大学计算机科学与技术学院教授,博士生导师,研究方向为人工智能、机器学习、社会计算.
E-mail: wangying2010@jlu.edu.cn