

基于结构深度网络嵌入模型的节点标签分类算法

陈世聪¹ 袁得崙^{1,2} 黄淑华^{1,2} 杨 明^{1,2}

1 中国人民公安大学信息网络安全学院 北京 100038

2 安全防范与风险评估公安部重点实验室 北京 100038

(chenshicong24@163.com)

摘 要 在海量数据呈现爆炸增长态势的互联网时代,传统算法已无法满足处理大规模、多类型数据的需求。近年来最新的图嵌入算法通过学习图网络特征,在链路预测、网络重构和节点分类实践中普遍取得了极佳的效果。文中基于传统自动编码器模型,创新地提出了一种融合 Sdne 算法与链路预测相似度矩阵的新算法,通过在反向传播过程中引入高阶损失函数,依据自编码器的新特征调整性能,改进传统算法中以单一方式判定节点相似度这一方法存在的弊端,并建立简易模型分析证明优化的合理性。对比最新研究中效果最好的 Sdne 算法,该算法在 Micro-F1 和 Macro-F1 两种评价指标上的提升效果均接近 1%,可视化分类效果表现良好。与此同时,研究发现高阶损失函数超参的最优值大致处于 1~10 范围内,数值的变化依旧能够基本稳定维持整体网络的鲁棒性。

关键词: 网络嵌入;深度学习;节点分类;自动编码器;复杂网络

中图法分类号 TP311

Node Label Classification Algorithm Based on Structural Depth Network Embedding Model

CHEN Shi-cong¹, YUAN De-yu^{1,2}, HUANG Shu-hua^{1,2} and YANG Ming^{1,2}

1 School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

2 Key Laboratory of Safety Precautions and Risk Assessment, Ministry of Public Security, Beijing 100038, China

Abstract In the era of Internet, where massive data is growing explosively, traditional algorithms have been unable to meet the needs of processing large-scale and multi type data. In recent years, the latest graph embedding algorithm has achieved excellent results in link prediction, network reconstruction and node classification by learning graph network characteristics. Based on the traditional automatic encoder model, a new algorithm combining Sdne algorithm and link prediction similarity matrix is proposed. By introducing a high-order loss function in the process of back-propagation, the performance is adjusted according to the new characteristics of the auto-encoder. The disadvantages of traditional algorithm in determining node similarity in a single way are improved. A simple model is established to analyze and prove the rationality of the optimization. Compared with the most effective Sdne algorithm in the latest research, the improvement effect of this algorithm on Micro-F1 and Macro-F1 two evaluation indicators is close to 1%, and the visual classification effect is good. At the same time, it is found that the optimal value of the hyperparameter of the higher-order loss function is approximately in the range of 1~10, and the change of the numerical value can basically maintain the robustness of the whole network.

Keywords Network embedding, Deep learning, Node classification, Auto-encoder, Complex network

1 引言

随着互联网的普及以及大数据时代的来临,规模庞大、类型复杂的图网络在众多领域得到了广泛应用。作为近年来深度学习领域发展的分支,从链路预测^[1-2]、网络重构到社区划

分^[3],图表示学习^[4-5]逐渐成为业界的热点研究方向。其最基本的思维是以有向有权图的形式描述数据关系,该方法将网络实体建模为顶点,将实体间的关系建模为边,把杂乱零碎的实际问题转化为图挖掘问题,从图结构角度研究现实网络的潜在价值。

到稿日期:2020-10-29 返修日期:2021-05-23

基金项目:国家社会科学基金重点项目(20AZD114);中国人民公安大学基本科研业务费项目(2021JKF215);中国人民公安大学公共安全行为科学实验室开放课题(2020SYS03);警务物联网应用技术公安部重点实验室开放课题

This work was supported by the National Social Science Foundation of China(20AZD114), Fundamental Research Funds for the Central Universities(2021JKF215), Open Research Fund of the Public Security Behavioral Science Laboratory, People's Public Security University of China(2020SYS03) and Open Research Fund of Key Laboratory of the Police Internet of Things Application Technology.

通信作者:袁得崙(yuandeyu@ppsuc.edu.cn)

现阶段图表示学习的发展衍生出解决各类实际问题的高效算法,典型代表有 Deepwalk, Node2vec, Line 和 Sdne (Structural Deep Network Embedding) 等算法。Deepwalk^[6]在词嵌入技术 Word2vec^[7]的基础上进行改进,应用随机游走^[8]机制,深度优先遍历所有网络节点,在给定起始节点和游走步长的条件下,将随机采样邻居节点作为下一个访问节点,按序组成规定长度的节点访问序列,以此表达不同节点之间的共现关系,在获得局部关联的采样数据后采用 skip-gram 模型对其进行训练,最终以数字向量形式表示复杂网络中离散节点的邻域结构。Node2vec^[9-10]在 Deepwalk 随机游走模型的基础上添加偏置参数,控制生成节点序列中对 BFS(Breadth First Search)邻域和 DFS(Depth First Search)邻域的倾向程度,同时创新使用 Alias 算法构造采样表,按概率计算公式抽取邻接点。Line^[11]算法提出了两种新的相似度定义,一阶相似度观察两个节点之间是否有连边,保留网络的局部特征;二阶相似度观察两个节点之间的共同邻居,保留网络的全局特征,两种相似度概念从不同的角度刻画出网络图的特征。Line 构建超参并采用梯度下降算法对一阶和二阶相似度进行分离优化,在训练模型的过程中运用负采样方法降低时间复杂度。

最新的 Sdne 算法^[12]通过构造由一阶相似度、二阶相似度和防止过拟合的 l_2 范数正则化项组成的整体损失函数,建立自动编码器模型,将节点向量编码降维,输入到高度非线性的潜在空间,保留节点局部和整体特性,捕获复杂网络的结构,再对低维向量进行解码升维处理,通过反复优化原始向量和重构向量间的损失差,来达到训练神经网络的目的。Li 等^[13]提出了一种粒子竞争模型 DPP,设计了独特的粒子竞争框架,根据模型的动态特性引入源节点的自动更新策略,进而完成更有效的动态节点分类,其在社区检测层面有广阔的应用前景。Sanjay 等^[14]针对当前流行的随机游走等多种网络嵌入算法,在不同数据集上进行分类精确度的实验数据对比,并可视化分类结果。

本文突破已有思维模式中单方面运用深度学习自编码器的禁锢,加入浅层嵌入模型中基于节点距离的思想,将传统链路预测领域使用的相似度指标处理为高阶损失函数,与最新的结构深度网络嵌入算法相融合,完善原有模型在处理节点对相似性过程中忽视网络结构潜在关联性的弊端,从细微处循环优化实验参数,同时重新选择合适的优化器和激活函数,降低因引入损失函数而增加的时间复杂度,在不同数据集中对 Micro-F1 值和 Macro-F1 值的提升均接近 1%。本文还对新引入调权参数的最优值和变化特征进行了探索,并与原有二阶损失函数的超参数进行比较,划分最优值。

2 基本模型描述

2.1 自动编码器

自动编码器^[15-18]作为前馈神经网络的一种特殊形式,在图嵌入领域中常用于降维特征学习。设 $\mathbf{R}^N$ 是将图网络表示为输入的 N 维邻接矩阵, $\mathbf{x}_i \in \mathbf{R}^N$ 是包含第 i 个节点局部邻域结构信息的邻接向量。自动编码器由编码 $g(\mathbf{x}_i): \mathbf{R}^N \rightarrow \mathbf{R}^D$ 和解码 $f(\mathbf{y}_i): \mathbf{R}^D \rightarrow \mathbf{R}^N$ 两部分组成。首先将邻接向量 $\mathbf{x}_i$ 映射到由多个非线性函数组成的低维表示空间 $\mathbf{R}^D$ 中,有效压缩数据

得到隐层表示向量,再将低维向量解码映射到重构空间,表示为重构向量 $\hat{\mathbf{x}}_i$,通过反向传播算法循环调整训练过程中的权重矩阵,以达到循环缩减输入值和目标值之间差值的效果。对输入向量进行 k 层编码后的每一层向量结果 $\mathbf{y}_i^{(k)}$ 为:

$$\mathbf{y}_i^{(k)} = \sigma(\mathbf{W}^{(k)} \mathbf{y}_i^{(k-1)} + \mathbf{b}^{(k)}), k=2,3,4,\dots,K \quad (1)$$

其中, $\mathbf{W}^{(k)}$ 表示第 k 层编码的权重矩阵, $\mathbf{y}_i^{(k-1)}$ 的含义是进行 $k-1$ 层编码后的节点向量, $\mathbf{b}^{(k)}$ 表示第 k 层的偏置向量, $\sigma(\cdot)$ 表示 sigmoid 非线性函数。

2.2 结构深度网络嵌入模型

与 Line 相似度定义相同, first-order 相似性由编码器拉普拉斯矩阵监督学习^[19]网络的局部特征, second-order 相似性由解码器无监督学习网络的整体结构,原有 Sdne 模型对一阶和二阶相似度损失函数赋予不同权重并进行迭代联合优化,模型结构如图 1 所示。

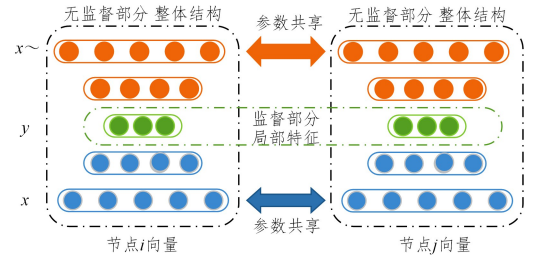


图 1 结构深度网络嵌入模型

Fig. 1 Structural deep network embedding model

一阶相似度损失函数使相邻节点对 (v_i, v_j) 对应的嵌入编码向量 $\mathbf{y}_i^{(k)}$ 和 $\mathbf{y}_j^{(k)}$ 在隐藏层无限接近,一阶损失函数的定义如下:

$$Loss_{1st} = \sum_{i,j=1}^n s_{i,j} \|\mathbf{y}_i^{(k)} - \mathbf{y}_j^{(k)}\|_2^2 = 2\text{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y}) \quad (2)$$

其中, $s_{i,j}$ 为邻接矩阵中的元素值,当节点对间存在直连边时, $s_{i,j}=1$,否则 $s_{i,j}=0$; $\text{tr}(\cdot)$ 表示矩阵的迹; $\mathbf{L}$ 为 Laplace 向量矩阵; $\mathbf{Y}$ 为编码器处理后的隐藏层空间向量矩阵。

在网络拓扑结构中,节点对之间存在连接表示关联程度较高,但是没有连接并不代表不存在相似性,其可能具备潜在关联并拥有众多共同邻居,却丢失了直观的连接表示,因此提出二阶损失函数的定义为:

$$Loss_{2nd} = \sum_{i=1}^n \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 = \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 \odot \mathbf{B} \quad (3)$$

其中, $\hat{\mathbf{x}}_i - \mathbf{x}_i$ 表示输出重构向量与输入原始向量之间的损失差, $\odot$ 表示 Hadamard 乘积,对 $\odot$ 两边的矩阵或向量进行逐元素乘积, b_i 表示惩罚系数的超参数。算法输入邻接矩阵中的每个节点向量,其中蕴含着中心节点的邻域结构信息,即优化的过程就是对结构相似的节点重构相似嵌入向量的过程。然而,由于图的稀疏特性,邻接矩阵向量中零元素的数量远高于非零元素的数量,若直接将邻接矩阵作为算法的输入,则零元素重构也更容易,但这不符合预期取得的理想结果。合理的解决方法是对非零元素的重构误差施加更高的惩罚系数 β 。当 $a_{ij}=0$ 时, $b_{ij}=1$; 当 $a_{ij}=1$ 时, $b_{ij}=\beta$ 。

Sdne 模型的最终目标是将两类损失函数线性相加进行优化,混合损失函数如式(4)所示, $Loss_{reg}$ 是一项规格化模型参数。基于稀疏网络的鲁棒性,整体优化效果对超参数 α 和 β 的敏感程度较小。

$$Loss_{mix} = \alpha Loss_{1st} + Loss_{2nd} + Loss_{reg} \quad (4)$$

3 算法实现

3.1 构建基于路径的 Katz 相似度矩阵

本文模型依据网络结构特征将邻接矩阵重构为相似度矩阵,直观表示节点对的相似度,重构矩阵按照数值大小区分邻域节点对中心节点的影响程度。Katz 指标^[20]从节点之间路径距离的角度出发,对不同路径赋予不同参数,短路径所占权重偏大,长路径所占权重偏小。Katz 指标权衡网络的所有路径,其相似度定义如式(5)所示,相似度(Similarity)矩阵 S 的计算方法如式(6)所示。

$$s(x, y) = \alpha A_{xy} + \alpha^2 (A^2)_{xy} + \alpha^3 (A^3)_{xy} + \dots \quad (5)$$

$$S = \beta A + \beta^2 A^2 + \dots + \beta^n A^n = (I - \alpha A)^{-1} - I \quad (6)$$

其中, $0 < \alpha < 1$ 是控制不同长度路径权重的参数; A 为邻接矩阵,其 n 次幂的计算结果为对应节点间路径长度为 n 的路径数量; I 为单位矩阵;随着路径长度的增加,可调参数 α^n 的大小也随之衰减,削弱了高阶路径的权重。

网络结构中任意节点对之间都存在着多条连接路径,中间存在的媒介节点越少,路径距离越短,联系就越紧密,因此低阶路径对节点相似度的贡献更大,设置参数 α 控制路径权重就显得尤为重要。在构造矩阵的过程中,依据高低阶路径对相似度的不同影响程度,通过动态微调参数 α 来达到理想的预测效果。一般情况下,参数 α 需要小于邻接矩阵 A 最大特征根的倒数。

3.2 模型设计

本文从算法结构角度引入高阶损失函数,提出了一种新的相似度结构深度网络嵌入(Similarity Structural Deep Network Embedding, SSdne)算法,用于优化传统算法,提升模型在节点标签分类实践中的预测精度,实验的基本流程如图2所示。

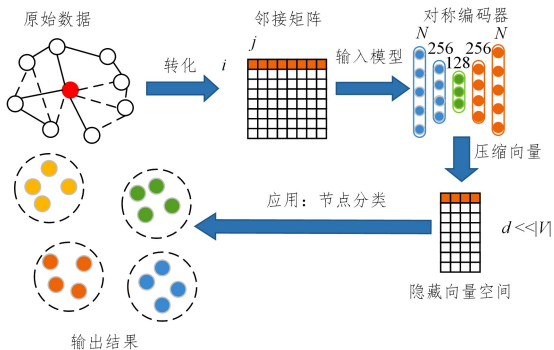


图2 实验流程

Fig. 2 Experimental process

当只考虑二阶路径时,Katz 指标就演变为基于共同邻居(Common Neighbors, CN)^[21-22]的相似指标,相比 CN 指标,Katz 指标按照比例引入高阶路径权重,思维考虑得更全面严谨,可以理解为是一种将共同邻居扩展到高阶邻居的算法。类似地,Katz 指标相比基于随机游走方式的 Cos+ 等相似度指标,算法设计更具备规律性和有序性,可减小游走过程中偶然性带来的预测误差。

传统的 Sdne 模型对基于节点间直连关系的一阶相似度和基于节点间共同邻居的二阶相似度做出了定义,有效捕获

了局部和整体特征。该模型是深度学习领域最新的且分类效果较好的图嵌入算法,虽然近年来各种分类算法层出不穷,但鲜有从网络嵌入的结构角度引入高阶损失函数来优化预测精度的研究。本文采取基于路径的 Katz 相似性指标算法,将高阶损失函数引入 Sdne 深度学习模型架构。高阶损失函数的定义如下:

$$Loss_{Katz} = \sum_{i=1}^n \|(\hat{x}_i - x_i) \odot S * \gamma\|_2^2 \quad (7)$$

其中, S 表示相似度矩阵, γ 是调节相似度矩阵中元素规模大小的参数,该高阶损失函数从一定层面上解决相似度判定的片面性问题。传统 Sdne 模型二阶损失函数的很大缺陷在于:算法仅根据中心节点与邻居节点是否存在直连边,就将 $B \in \{1, \beta\}$ 作为参数矩阵对损失差施加惩罚系数,却忽略了节点对存在潜在关联的概率,亦淡化了区分中心节点与邻居节点相似度的意义。倘若仅凭直接连接就对不同节点对的相似性一概而论,那么图网络存在的隐藏细节特征就无法被完整地表示,因此引入高阶相似度损失函数具有重要参考价值。

邻接矩阵描绘给定输入数据的实际情况,却没有挖掘网络拓扑结构存在的潜在信息,因此深层神经网络模型只处理输入邻接矩阵显然不完整。高阶损失函数从路径角度将邻接矩阵重构处理为 S_{katz} 矩阵,在自编码器的迭代训练过程中融入了对节点内在结构相似成分的考虑,且对相似度高的节点对根据评分施加相应程度的惩罚系数。二阶与高阶损失函数从节点间直接联系和全局路径两个角度分别对网络结构的直观相似度和隐藏相似度进行迭代累加运算,综合比较两个节点的周围邻居集合的相似度差异,使网络嵌入模型更细致紧密地保留网络结构的局部和整体特征,优化的完整算法如算法1所示。

算法1 相似性结构深度网络嵌入(SSdne)

输入:网络 $G=(V, E)$ 的邻接矩阵 A ,拉普拉斯矩阵 L ,节点标签 L ,超参 α, β, γ

输出:网络节点向量表示 Y 和更新参数 θ

1. 通过深度信念网络对深度自编码器进行预训练,获得初始化参数 θ
2. 构造节点间基于路径的相似度矩阵 S
3. $X = A$
4. Repeat
5. 基于 X 和 θ ,应用式(1)得到 $\hat{X}$ 和 $Y = Y^K$
6. $Loss_{Katz} = \|(\hat{X} - X) \odot S * \gamma\|_F^2$
7. $Loss_{2nd} = \|(\hat{X} - X) \odot B\|_F^2$
8. $Loss_{1st} = 2tr(Y^T LY)$
9. $Loss_{mix} = \alpha Loss_{1st} + Loss_{2nd} + Loss_{Katz} + Loss_{reg}$
10. 利用 $\partial L / \partial \theta$ 反向传播整个网络以获取更新的参数 θ
11. Until converge
12. 获得重构网络节点向量表示 Y

3.3 相似度定理分析与证明

Sdne 算法本质上采用邻接矩阵训练损失函数,而优化 SSdne 算法将邻接矩阵与 Katz 相似度矩阵进行联合处理。本文抽取算法细节,建立如图3所示的简易模型,通过对比原始算法的单一邻接矩阵与优化算法的整合矩阵,来分析节点相似性之间存在的差异,阐释核心优化处的优越性与合理性。

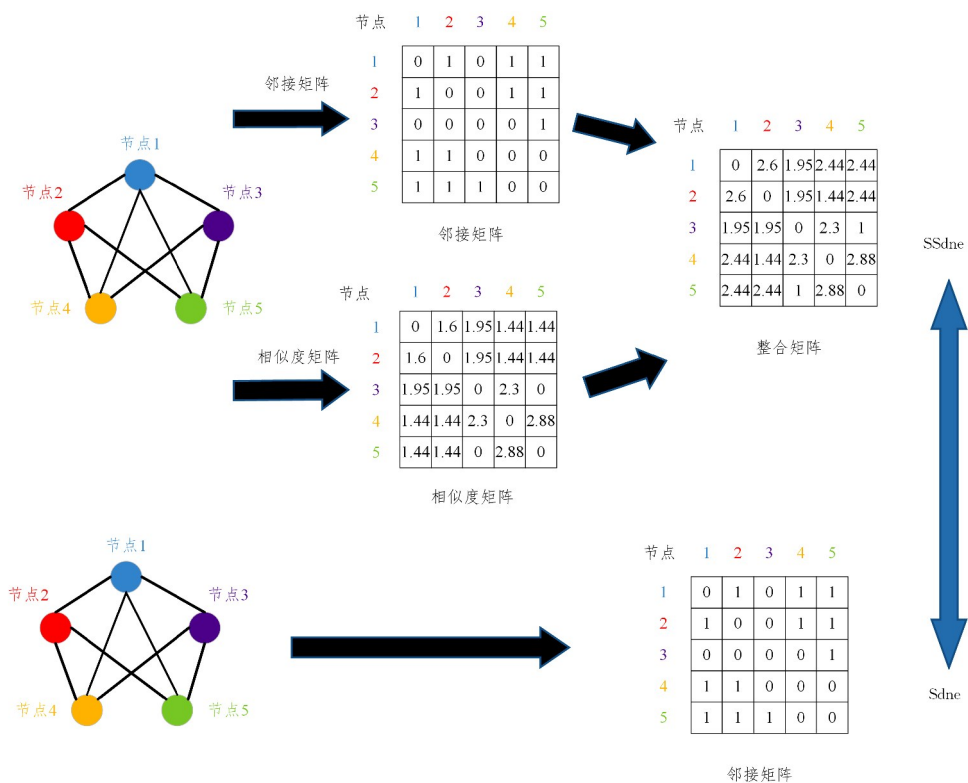


图3 简易分析模型

Fig. 3 Simple analysis model

在由5个节点构成的无向网络结构图中,我们假设任意路径权重为1,取衰减参数 $\alpha=0.8$,受网络直径限制,仅存在四阶路径,且计算路径时不重复到达同一个节点。优化算法计算相似性的方式如式(8)所示:

$$S_{xy} = \begin{cases} 0 + \alpha(A^2)_{xy} + \alpha^2(A^3)_{xy} + \alpha^3(A^4)_{xy} \\ 1 + \alpha(A^2)_{xy} + \alpha^2(A^3)_{xy} + \alpha^3(A^4)_{xy} \end{cases} \quad (8)$$

以节点3和节点4的计算为例:无共同邻居即二阶路径,则 $(A^2)_{xy}=0$;2条三阶路径为4-2-5-3和4-1-5-3,则 $(A^3)_{xy}=2$;2条四阶路径为4-1-2-5-3和4-2-1-5-3,则 $\alpha^3(A^4)_{xy}=2$ 。对比整合矩阵的数据计算结果:横向观察,同样无直连边的节点对2-3和4-5以及同样有直连边的节点对1-2和1-4,其相似度数数值截然不同;纵向观察,无直连边的4-5节点对相似度甚至超过网络结构中所有存在直连边的节点对,原因是节点的周围节点在结构上更趋于相似,此简易模型自然证明了节点对直连边是评判相似度的重要组成元素之一,却不是唯一依据,共同邻居的结构性质以及路径衰减参数的大小都会影响训练效果。总而言之,传统算法应用邻接矩阵的思维方式过于单一片面,其细节弊端显而易见。

3.4 实验细节

本文对3种拥有连接和标签信息的不同网络数据集进行节点分类,研究比较优化算法的提升效果,并且对调节权重的超参数进行敏感性测试,以寻找相对最优值区间。实验使用Keras在启用CPU的Tensorflow后端构造编码器和解码器各两层,将自动编码器结构的隐藏层维度固定为N-256-128-256-N。

对比传统Sdne模型中易陷入鞍点的SGD优化器,实验创新地使用Adam优化器进行梯度下降优化,在加速神经网络收敛速度和稳定性方面取得了极佳的效果,该优化器新一轮

梯度下降的计算方法是将目前真实梯度与前一轮次中用于梯度下降的梯度进行线性加权,其自适应的学习效率有显著优势,能够有效弥补稀疏网络存在的弊端。

同时将sigmoid函数替换为relu激活函数,一方面避免了损失函数反向传播时因梯度消失导致的信息丢失状况,且不需要使用一贯使用的Dropout函数,降低了过拟合现象发生的概率;另一方面加快神经网络的收敛速度,优化训练过程的时间复杂度。

4 实验结果

4.1 运行环境与数据集

本文的实验环境为64位mac操作系统,Intel Core i5的CPU处理器,使用的软件是2019.2版本的pycharm,python3.5以及tensorflow1.4。实验数据集选取Wiki网页关系、Usair航班连线、Cora引文网络3种,数据内容包括不同节点间存在的连接以及全部节点的所属类别。Wiki数据集中存在2405个节点和17981条边,所有标签共分为17类。Usair数据集中存在1190个节点和13599条边,所有标签共分为4类。Cora数据集中存在2708个节点和10858条边,所有标签共分为7类。数据集网络属性信息如表1所列,Gephi展示网络整体结构如图4—图6所示。

表1 数据集的网络属性

Table 1 Dataset network properties

数据集	节点	边	标签	平均度	平均 路径长度
Wiki	2405	17981	17	14.953	5.274
Usair	1190	13599	4	22.855	2.864
Cora	2708	10858	7	8.019	6.310

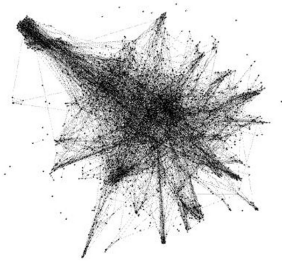


图 4 Wiki 网络结构
Fig. 4 Wiki network structure

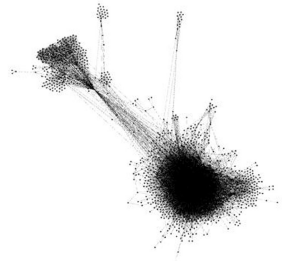


图 5 Usair 网络结构
Fig. 5 Usair network structure

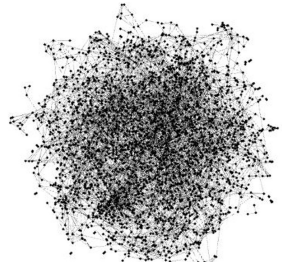


图 6 Cora 网络结构
Fig. 6 Cora network structure

4.2 实验结果分析

为评估优化模型的性能,使用节点标签分类任务验证图嵌入学习效果,采用流行 F1 分数评价指标,计算公式如下:

Precision = \frac{\sum_{A \in C} TP(A)}{\sum_{A \in C} (TP(A) + FP(A))} \tag{9}

Recall = \frac{\sum_{A \in C} TP(A)}{\sum_{A \in C} (TP(A) + FN(A))} \tag{10}

F1-measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}

其中,TP(A),FP(A),FN(A)分别表示将样本预测为 A 标签的真正例、假正例和假反例。Micro-F1 和 Macro-F1 是衡量多标签分类任务效果的重要指标,两类指标相似但又存在差异。Macro 指标平等对待所有样例,计算总体 Precision 和 Recall,再计算 F1 分数。Micro 对同一标签的样例给予相同权重,分别计算每一类别的 Precision 和 Recall 并对 F1 求值,最后求所有类别的平均 F1 分数。基于最新的研究分析发现,微平均和宏平均衡量指标能从多角度综合评判预测的准确率和召回率,微小的数值改变会影响模型性能,两类指标在节点分类领域具备极高的可信度。

微平均 Micro 指标的优点在于对采样数量少的类别关注

得更多,宏平均 Macro 指标的优点在于更关注采样数量多的类别。然而,Macro 指标平等对待所有标签类别,易受到稀有类别样本的影响;Micro 指标对数据集中的每一个样本都赋予相同的权重,因此受到普通类别样本的影响更大。

实验设置模型训练迭代轮次 epoch=40,维持 \alpha 和 \beta 参数值恒定,一阶相似度超参 \alpha=0.0001,二阶相似度超参 \beta=5。为减少过拟合的负面效应,编码器模型在全连接神经网络 Dense 层选择使用 kernel_regularizer, l1_l2 的正则化方式来限制权值的上下限,设置 l_1=0.00001, l_2=0.0001。为了确保优化算法精确度提升的普遍性,以及减少由数据集划分造成的实验误差,实验分别按照 10% 递增到 90% 的训练集比例划分数数据集,每轮结果为运行 20 次结果的平均值,基于平均值对比 SSdne 算法与传统 Sdne 算法在实际分类效果中的差异,同时选取 Line 算法和 Struc2vec 作为基线算法的比较,使结果更具说服力和可信度。3 类数据集上的节点分类效果数据如表 2—表 4 所列,算法的对比效果如图 7—图 12 所示。

表 2 Wiki 数据集的结果对比

Table 2 Comparison of Wiki datasets results

训练集	Micro (Sdne)	Micro (SSdne)	Macro (Sdne)	Macro (SSdne)
0.1	0.5002	0.5053	0.3536	0.3615
0.2	0.5596	0.5643	0.4145	0.4180
0.3	0.5785	0.5856	0.4312	0.4370
0.4	0.5984	0.6007	0.4534	0.4675
0.5	0.6134	0.6157	0.4719	0.4780
0.6	0.6146	0.6204	0.4721	0.4794
0.7	0.6109	0.6170	0.4695	0.4777
0.8	0.6178	0.6213	0.4740	0.4805
0.9	0.5937	0.6062	0.4652	0.4780

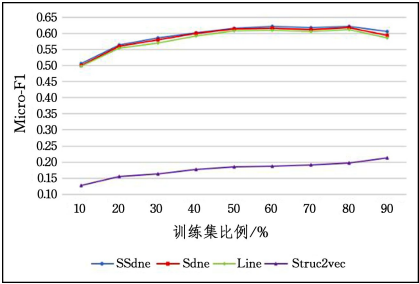


图 7 不同算法在 Wiki 数据集上 Micro 值的对比

Fig. 7 Comparison of Micro values on Wiki dataset of different algorithms

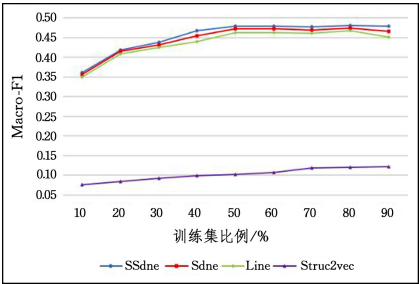


图 8 不同算法在 Wiki 数据集上 Macro 值的对比

Fig. 8 Comparison of Macro values on Wiki dataset of different algorithms

表 3 Usair 数据集的结果对比

Table 3 Comparison of Usair datasets results				
Training set	Micro (Sdne)	Micro (SSdne)	Macro (Sdne)	Macro (SSdne)
0.1	0.5539	0.5607	0.35634	0.36158
0.2	0.5724	0.5782	0.41459	0.41808
0.3	0.5797	0.5865	0.43129	0.43708
0.4	0.5850	0.5876	0.45344	0.46753
0.5	0.5894	0.5988	0.47194	0.47802
0.6	0.5987	0.6092	0.47215	0.47943
0.7	0.6188	0.6305	0.46950	0.47772
0.8	0.6134	0.6281	0.47406	0.48051
0.9	0.6342	0.6537	0.46528	0.47806

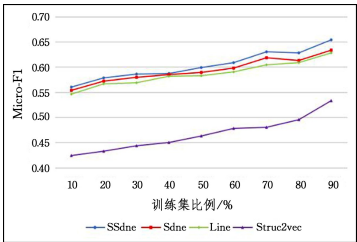


图 9 不同算法在 Usair 数据集上 Micro 值的对比
Fig. 9 Comparison of Micro values on Usair dataset of different algorithms

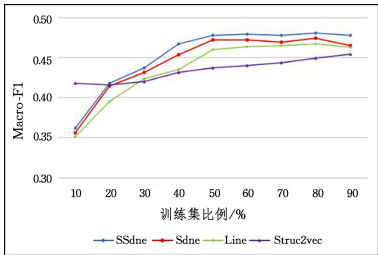


图 10 不同算法在 Usair 数据集上 Macro 值的对比
Fig. 10 Comparison of Macro values on Usair dataset of different algorithms

表 4 Cora 数据集的结果对比

Table 4 Comparison of Cora datasets results				
训练集	Micro (Sdne)	Micro (SSdne)	Macro (Sdne)	Macro (SSdne)
0.1	0.5340	0.5364	0.4777	0.4829
0.2	0.5822	0.5934	0.5522	0.5646
0.3	0.6163	0.6232	0.5879	0.5932
0.4	0.6299	0.6444	0.6021	0.6184
0.5	0.6448	0.6533	0.6253	0.6290
0.6	0.6511	0.6580	0.6261	0.6353
0.7	0.6640	0.6729	0.6378	0.6483
0.8	0.6638	0.6669	0.6437	0.6474
0.9	0.6685	0.6763	0.6388	0.6469

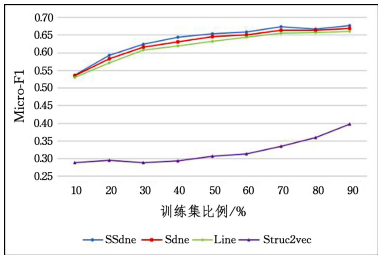


图 11 不同算法在 Cora 数据集上 Micro-F1 值的对比
Fig. 11 Comparison of Micro values on Cora dataset of different algorithms

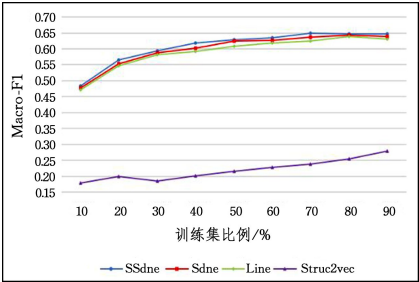


图 12 不同算法在 Cora 数据集上 Macro-F1 值的对比
Fig. 12 Comparison of Macro values on Cora dataset of different algorithms

实验数据显示,优化 SSdne 算法相比传统 Sdne 算法应用于 Wiki,Usair 和 Cora 网络中时,在不同比例训练集下, Micro 指标分别平均提升了 0.55%,0.98%,0.78%,Macro 指标分别平均提升了 0.77%,0.93%,0.83%。从折线图易观察出,无论实验按何种比例划分训练集,本文算法的微平均和宏平均指标值都普遍高于原始 Sdne 算法,其精确度略微优于 Line 算法,相比 Struc2vec 算法,本文算法的优化程度的效果存在显著提升。上文阐述的衡量指标存在特殊性,且优化算法能够从宏平均和微平均两个方面同时提高精确度,意味着优化的 SSdne 模型取得了理想的分类效果。

新算法在相对非稀疏的 Usair 网络中取得的精确度提升更大,说明引入的高阶损失函数在内部结构信息相对丰富的网络中表现更好。网络结构越稳定、饱满,每个节点的邻域结构信息就越完善,新型算法引入的损失函数分析中心节点邻域结构的精确性也更高,有助于保留自编码器输出节点向量的邻域特性,进而优化分类效果。

与此同时,Usair 数据集中网络节点的数量较少,而节点之间的连边数量较多,即如表 2 所列平均度值更高,网络结构信息更全面,因此在节点与边相对密集的网络图中:一方面,Struc2vec 算法从多角度计算节点之间的相似性,随机游走采样获取的顶点序列更完整,游走捕获的图网络空间结构的相似性也更丰富,因此节点分类的精确度有所提升;另一方面,Line 算法和 Sdne 算法引入的一阶和二阶相似度,在网络结构信息相对完整的训练过程中存在过拟合现象,预测精确度有所下降,最终导致 usair 数据集中评判指标的差别并不明显。

本文使用 TSNE 技术对高维数据进行降维可视化,属于不同类别的节点由不同颜色表示,可视化结果如图 13 一图 15 所示。最合理的可视化效果应为不同颜色类别的节点能够存在清晰的边界划分。本文中 Wiki 数据集的标签为 1,2,7,13 的 4 类样本的预测效果较好,其余个别属于不同标签的节点存在杂糅现象。Usair 数据集中标签为 0,1,2 的 3 类样本的预测效果较好,然而上述 3 类标签中的某些节点也存在相互掺杂的情况。Cora 数据集中标签为 1,2,3,4 的 4 类样本的预测效果较好,同时某些属于同一标签的少量节点也能够准确分类。

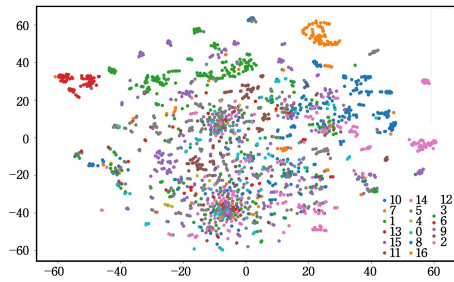


图 13 Wiki 节点分类效果的可视化

Fig. 13 Visualization of Wiki node classification effect

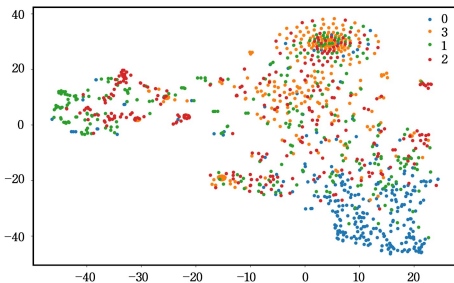


图 14 Usair 节点分类效果的可视化

Fig. 14 Visualization of Usair node classification effect

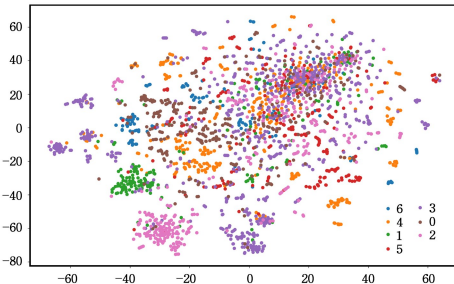


图 15 Cora 节点分类效果的可视化

Fig. 15 Visualization of Cora node classification effect

4.3 敏感度分析

从不同角度出发的相似性计算方法,获取到的相似性矩阵中的元素大小阈值自然也不同,因此需要设置超参数 γ 来调整高阶损失函数值,并且与二阶损失函数值保持基本一致。本文选择 Cora 数据集,为 γ 设定不同数值并分别在不同比例的训练集中进行实验测试,寻找最优的参数值范围,评判引入的高阶损失函数对稀疏网络整体性质规律的敏感程度,实验结果如图 16 和图 17 所示。

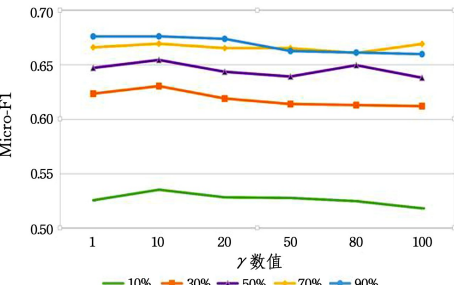


图 16 Micro 指标对超参数 γ 的敏感度

Fig. 16 Sensitivity of Micro index to hyper parameter γ

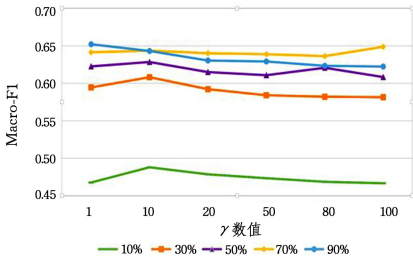


图 17 Macro 指标对超参数 γ 的敏感度

Fig. 17 Sensitivity of Macro index to hyper parameter γ

分析数据可知,调权参数 γ 在 1~10 范围内取值,节点分类的效果最佳,随着 γ 值的不断增长,优化算法的分类效果整体呈现下降趋势,但是曲线依旧能够基本保持平稳,即衡量指标对参数 γ 值的波动不敏感,说明引入的高阶损失函数超参数与一阶、二阶损失函数超参数同样都能够较好地维持稀疏网络的鲁棒性,不会对优化模型的性能造成实质性的改变。

通过实验发现,Katz 损失函数的调权参数与二阶损失函数的惩罚系数需要保持基本一致,调参直接决定着节点相似度波动范围的上下阈值,若存在重大偏差,则整体损失函数会表现出极度不平衡,进而导致分类性能下降,即引入高阶损失函数调参具有必要性。

结束语 本文在最新 Sdne 模型的基础上,创新地提出了一种结合传统链路预测相似度的算法,对原始模型存在的欠缺进行了完善补充,取得了良好的节点分类效果。实验还探索出了新引入调权参数的最优值,总结分析了超参对稀疏网络鲁棒性敏感度影响的规律特征,此类算法未来在社区划分、打击网络犯罪团伙等领域具有广泛的应用前景。

实验仅从路径角度构造高阶相似度损失函数,然而还存在各类基于共同邻居和随机游走等的相似度算法,如何多角度综合相似度算法,构造自定义损失函数,将其合理嵌入到优化模型中,是我们下一步研究的方向。另外,嵌入节点的附加信息属性也是扩展模型适用范围的一种重要方式,后续会对其做出补充。

参考文献

[1] LIBENNOWELL D,KLEINBERY J. The link-prediction problem for social networks[J]. Journal of the Association for Information Science and Technology,2007,58(7):1019-1031.

[2] LV L Y, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex network[J]. Phys Reue, 2009,80(4):211-223.

[3] SHEIKH N, KEFATO Z T, MONTRESOR A. Semi-Supervised Heterogeneous Information Network Embedding for Node Classification Using 1D-CNN[C]// 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE,2018:177-181.

[4] ZHANG D, YIN J, ZHU X, et al. Network representation learning: A survey [J/OL]. IEEE Transactions on Big Data. <https://ieeexplore.ieee.org/document/8395024>.

[5] CAI L J, XU Y B, HE T Q, et al. A New Algorithm of Deep-

Walk Based On Probability[J]. Journal of Physics: Conference Series, 2019, 1069(1): 130-135.

[6] PEROZZI B, AIRFOU R, SKIENA S. Deepwalk: Online learning of social representations[C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.

[7] TANG M, ZHU L, ZOU X C. Document Vector Representation Based on Word2Vec[J]. Computer Science, 2016, 43(6): 214-217.

[8] WANG H W, WANG J L, XIE X, et al. GraphGAN: Graph representation learning with generative adversarial nets[J]. arXiv, 2017, 30(22): 11-19.

[9] ABU-EL-HAIJA S, PEROZZI B, AL-RFOU R. Learning Edge Representations via Low-Rank Asymmetric Projections[C] // Proceedings of the 2017 ACM on Conference on Information and Knowledge Management(CIKM 17). 2017: 1787-1796.

[10] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 855-864.

[11] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C] // Proceedings of the 24th International Conference on World Wide Web. 2015: 1067-1077.

[12] WANG D, CUI P, ZHU W. Structural deep network embedding [C] // Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1225-1234.

[13] LI W Z, YIN D C, YUAN D Y, et al. Particle Propagation Model for Dynamic Node Classification[J]. IEEE Access, 2020(8): 140205-140215.

[14] SANJAY K, SANIDHYA C, SAKSHAM K, et al. Node Classification in Complex Networks Using Network Embedding Techniques[C] // Proceedings of the Fifth International Conference on Communication and Electronics Systems. 2020: 369-374.

[15] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Networks[J]. Advances in Neural Infor-

mation Processing Systems, 2014, 3: 2672-2680.

[16] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layerwise training of deep networks[C] // International Conference on Neural Information Processing Systems. MIT Press, 2006: 153-160.

[17] RANZATO M, BOUREAU Y L, LECUN Y. Sparse feature learning for deep belief networks[C] // Advances in Neural Information Processing Systems. 2007: 1185-1192.

[18] TOMAS N K, MAX W. Variational Graph Auto-Encoders[J]. Springer, 2016, 28(3): 61-63.

[19] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing Systems, 2002, 46(2): 585-591.

[20] LEO K. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 50(18): 39-43.

[21] SALTON G, MCGILL M. Introduction to Modern Information Retrieval[J]. McGraw-Hill, 1983, 32(6): 528-536.

[22] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Phys Rev E, 2001, 64(10): 251-264.



CHEN Shi-cong, born in 1997, master. His main research interests include cyberspace security and law enforcement technology.



YUAN De-yu, born in 1986, Ph.D, lecturer, Ph.D supervisor. His main research interests include cyber security and complex networks.

(责任编辑:柯颖)