

基于 Katz 自动编码器的城市路网链路预测模型

盛津芳, 刘家广, 王 斌

SHENG Jinfang, LIU Jiaguang, WANG Bin

中南大学 信息科学与工程学院, 长沙 410083

School of Information Science and Engineering, Central South University, Changsha 410083, China

SHENG Jinfang, LIU Jiaguang, WANG Bin. Katz Auto Encoder for Urban Road Network Link Prediction Model. Computer Engineering and Applications

Abstract: Urban road network is a special complex network. The link prediction of road network has important application value in urban planning and urban structure evolution. As the road network is highly sparse and nonlinear, this paper proposes a road network link prediction model based on the Katz Auto Encoder Network Embedding (hereafter as KAENE). This prediction model is a deep learning network embedding model based on the Auto Encoder with the following characteristics. First, it uses Katz similarity matrix to preserve the structural characteristics of the road network. Second, the multi-layer non-linear Auto Encoder is used to learn network characterization of the road network. Third, during the model training period, it uses L2 loss function to preserve the local characteristics of the road network. Fourth, it also uses L2-norm to prevent overfitting. Finally, it combines with the directional characteristics of the road network to improve the link prediction accuracy of the road network. Through experiments, this paper compared the different performance between KAENE model and other link prediction models using different urban road network data, at the same time, it also identified the influence of different embedded dimensions on the prediction accuracy of KAENE model. At last, we know the network characterization learning process of this model through visualization. The experimental results show that KAENE achieves a good performance in the link prediction tasks of six representative road network datasets at home and abroad.

Key words: Complex Networks; Link Prediction; Network Embedding; Auto Encoder; Urban Road Network

摘 要: 城市交通道路网络 (以下简称“路网”) 是一种特殊的复杂网络, 对路网进行链路预测在城市规划与城市结构演化方面有着重要的应用价值。针对路网的高度稀疏性、高度非线性特点, 提出了一种基于 Katz 相似度自动编码器 (Katz Auto Encoder Network Embedding, KAENE) 的路网链路预测模型, 它是一种基于自动编码器的深度学习网络嵌入模型, 使用 Katz 相似度矩阵保存路网的特征, 利用多层非线性自动编码器对路网进行网络表征学习, 在模型训练阶段通过局部线性嵌入损失函数保存路网的局部特征, 在此基础上引入 L2 范数来提高模型的泛化能力, 最后结合路网的方向性特征提高路网的链路预测精确度。通过实验对比了 KAENE 模型与其它链路预测模型在国内外的不同城市路网数据上的表现以及不同嵌入维度对 KAENE 模型预测精度的影响, 最后通过可视化了解了模型的网络表征学习过程。实验结果表明, KAENE 在国内 6 个具有代表性的路网数据集的链路预测任务中取得了良好的表现。

关键词: 复杂网络; 链路预测; 网络嵌入; 自动编码器; 城市路网

文献标志码: A **中图分类号:** TP391.9 **doi:** 10.3778/j.issn.1002-8331.1801-0002

基金项目: 国家科技重大专项课题资助(No.2017ZX06002005)。

作者简介: 盛津芳 (1971-), 女, 博士, 副教授, 主要研究领域为大数据、云计算、复杂网络, E-mail: jfsheng@csu.edu.cn; 刘家广 (1992—), 男, 硕士研究生, 主要研究领域为复杂网络、深度学习; 王斌 (1973—), 男, 博士, 教授, 主要研究领域为软件工程、云计算、复杂网络。

1 引言

随着城市化的发展,城市交通路网结构因交通模式、自然条件、人口区域分布、演化阶段及发展水平的不同会产生明显差异,经济技术的发展时刻改变着路网结构^[1],城市道路交通路网演化会变得越来越复杂。路网的复杂演化加大了城市道路规划的决策难度,如果道路规划不合理,就会导致路网资源无法合理分配,甚至有可能阻碍城市化发展。

20 世纪以来,随着“小世界网络”与“无标度网络”研究的兴起,复杂网络可以作为一种网络建模工具,进而对路网进行建模,复杂网络的链路预测则为研究路网的演化方向提供了可行的解决方案。目前,链路预测在朋友圈好友推荐以及探究蛋白质相互作用实验领域取得了成功^[2]。以路网为例,对未知链路的预测或者对未来链路的预测本质上是对路网的演化方向的预测,也是对路网拓扑结构的数据挖掘过程。

应用复杂网络理论进行城市路网拓扑结构描述和路网网络演化行为理解已成为交通地理信息系统重点研究方向之一。宋等^[3]利用复杂网络理论,构建了以连接度、介中心和接近度为度量指标定义道路重要度评价模型。赵等^[4]利用复杂网络理论分析了路网的复杂特性及鲁棒性。越来越多的复杂网络模型被用来分析路网结构特征。但是,利用复杂网络链路预测加深对路网演化行为理解,目前处于空白研究领域。因此,利用复杂网络的链路预测理论对路网进行数据挖掘,进而揭示路网的演化本质具有非常重要的意义。

近年来,网络嵌入技术^[5]成为研究者的热点方向,网络嵌入技术是指将网络中的节点或者边投影到低维向量空间中,再用于后续复杂网络的数据挖掘任务。网络嵌入技术将网络中的节点或者边投影到低维向量空间中,而不损失网络的结构特征,网络中两个节点的相似度信息可以用低维空间中节点间的向量空间距离表示。这种技术不但可以用于复杂网

络的链路预测,还可以用于网络中节点分类聚类任务,因此具有非常好的扩展性。目前,将网络嵌入技术用于路网的链路预测面临着三个主要困难:(1)路网的高度非线性。路网节点之间的链接关系非常复杂,捕获路网的非线性特征非常困难。(2)路网的局部性与全局性特征。模型在进行训练时学习到的局部特征与全局特征很难得到兼顾。(3)路网的高度稀疏性。相对于科学家论文合作网络、Facebook 朋友网络、互联网网络等网络来说,路网具有高度稀疏性。

目前,深度学习在语音识别、计算机视觉等多类应用中取得了突破性的进展,深度学习所学得的模型中,非线性操作的层级数更多。基于深度学习的链路预测,可以使模型学习到复杂网络中更深层次的网络拓扑特征,为路网的高度非线性难题提供了解决方案。

基于以上应用背景以及路网这类“特殊”的复杂网络,本文提出了基于 KAENE 的城市路网链路预测模型,它是一种基于自动编码器的深度学习网络嵌入模型,使用 Katz 相似度矩阵保存路网的结构特征,利用多层非线性自动编码器对路网进行网络表征学习,通过局部线性嵌入损失函数保存路网的局部特征,最后结合路网的方向性特征提高路网的链路预测精确度。KAENE 将会对于城市路网复杂演化的科学管理与规划、提高路网的资源利用率以及增强交通网络的平衡性和可靠性具有重要的现实意义。

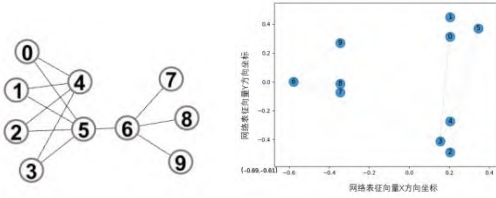
2 相关工作

2.1 网络嵌入技术

复杂网络的链路预测受到来自不同领域科学家的广泛关注,各种模型不断不提出。吕等^[6]在链路预测的研究综述中总结了基于相似性指标、机器学习以及概率模型的链路预测方法。近几年来,基于网络嵌入技术的链路预测开始受到研究者的广泛关注。网络嵌入技术将图的节点或者边投影到低维向量空间中,实现节点的向量化的同时可以保持网络结构的不变性。如图 1 所示,图 1(a)为一个简单

的复杂网络图,图 1(b) 经过 LAP^[7] 将图 1(a) 表征映射到二维平面空间。从图 1(b) 中可以看出, LAP 将整个网络划分成了 3 个集合: $U_1=[0,1,5]$, $U_2=[2,3,4]$, $U_3=[6,7,8,9]$, 从横坐标轴进行切割, 将其划分为 2 个集合 $[U_1, U_2]$ 与 U_3 , 与图 1(a) 相比, 可以认为其表征结果在拓扑结构上具有一定的结构相似性, 但是仍有改进的空间。由于二维空间的表征能力有限, 我们通常关注的是网络在高维空间的向量表示。

目前基于网络嵌入技术的链路预测相关研究成果主要有 LAP^[7]、Node2vec^[8]、SDNE^[9]、GF^[10]、LLE^[11]、NMF^[12]、SVD^[13]、HOPE^[14]等。



(a) 复杂网络图 (b) LAP 表征学习结果

图 1 网络嵌入技术表征模型

2.2 深度学习

传统的机器学习通过手工来提取原始数据的特征, 深度学习的出现, 避免了这一繁琐的过程。在语音识别、计算机视觉等领域, 深度学习模型已经变得越来越重要。在过去, 训练两个以上隐藏层的神经网络非常困难, 其主要难点在于高维参数空间和高非凸目标函数以及基于梯度下降的方法很容易陷入局部最小。随着计算能力的提升以及无监督和有监督的训练算法的进展^[15,16,17], 这些问题都得到了有效解决。

目前, 无监督学习方法, 如自动编码器^[18]与 RBMs^[19]等, 这类算法采用逐层初始化神经网络的方式, 自动实现了对未标注数据的特征提取。在深度学习框架下, Chang 等^[20]使用了多层卷积神经网络实现了对异构数据的网络表征学习。Wang 等^[9]提出 SDNE 模型, 使用了自动编码器对复杂网络进行表征学习, 通过在损失函数中引入“一阶近似”与

“二阶近似”来保存网络的结构特征。然而, 在路网的链路预测任务中, SDNE 模型的预测效果有待提高。本文提出了一种 KAENE 链路预测模型, 与 SDNE 不同之处在于: (1) 使用 Katz 相似度矩阵保存路网的全局特征 (2) 在模型训练阶段引入局部线性嵌入损失函数保存路网的局部特征, 在此基础上引入 L2 范数来提高模型的泛化能力 (3) 结合路网的方向性特征提高路网的链路预测精确度。

3 模型的设计与实现

3.1 相关定义

定义图 $G=(V,E)$ 为一个无向连通图, V 为 G 中节点的集合, E 为边集合。定义 $N=|V|$, N 为图 G 的节点数, $M=|E|$ 为图 G 的边数, 则该网络中一共有 $N(N-1)/2$ 节点对。给定时间 δ 时图 G 中各节点对状态, 链路预测问题可以形式地描述为推断当前状态中或将在时间 $\delta+t$ 中形成的缺失链接的子集。

定义 A 为图 G 的邻接矩阵, A 满足以下条件,

$$A = (a_{ij})_{n \times n}, \quad a_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & \text{else} \end{cases} \quad (1)$$

A 是一个对称矩阵, 对角线元素均为 0, 矩阵的各个行和 (列和) 是各个顶点的度。所有元素相加和为边数的二倍。

定义网络 G 的平均度 Avd ,

$$Avd = \frac{|E|}{|V|} \quad (2)$$

即边的数量与节点数量的比值, 平均度表达了网络的密度。一般来说, 平均度越小网络越稀疏。

定义 A^n 为邻接矩阵的 n 阶幂,

$$A^n = A^n \quad (3)$$

$A^n_{(i,j)} = k$ 表示的是从节点 i 出发走到节点 j 走 n 步, 有 k 种路径可达。如果 k 的值为 0, 表示节点 i 与节点 j 在 n 步内不可达。

3.2 Katz 相似矩阵

Katz 相似度^[21]由 KatzL 在 1953 年提出, 它考虑了节点间的所有路径数, 对于短路径之间的节点赋予了较大的权重值, 对于长路径之间的节点对赋予

了较小的权重值。已知邻接矩阵 A ，则 Katz 相似性矩阵的定义如下，

$$S^{Katz} = \alpha A + \alpha^2 A^2 + \alpha^3 A^3 \dots \quad (4)$$

其中 $\alpha \in (0,1)$ ， α 是控制路径权重的调节参数，通过它来调节远距离路径对相似度的影响，当网络节点数量很大时，Katz 系数在有限时间内很难被计算出。但是当 α 小于邻接矩阵 A 最大特征值的倒数时， S^{Katz} 可以进行收敛， S^{Katz} 收敛后的表达形式如下，

$$S^{Katz} = (I - \alpha A)^{-1} - I \quad (5)$$

I 为单位矩阵， $(I - \alpha A)^{-1}$ 表示 $(I-A)$ 的逆矩阵。在对路网进行网络训练时， α 表达了长距离路径对预测结果的贡献度。

为了将 Katz 相似度指标系数映射到 $[0,1]$ 之间，需要对数据进行标准化，常用的标准化方法有 Min-Max 标准化以及 Z-score 标准化，本文中使用了 Min-Max 标准化，其计算方式如下，

$$x^* = \frac{x - \min}{\max - \min} \quad (6)$$

其中 $\min$ 为 S^{Katz} 中的最小值， $\max$ 为最大值。

3.3 自动编码器与损失函数

自动编码器是一种无监督机器学习技术，利用神经网络来产生一个高维输入的低维表示，从而达到网络嵌入的目的。传统的维度下降依赖于线性方法，如 PCA^[22]，通过找出高维数据中最大的方差的方向进行降维。然而，PCA 方法的线性也是导致自身可以抽取出的特征维度类型上的很大限制。自动编码器使用大量的非线性函数克服了这些限制，从而可以体现数据的天然非线性。

单层自动编码器的基本原理如下，假设输入一个 n 维向量 $X_i^{(n)}$ ，通过自动编码神经元映射到一个 k 维向量 $Y_i^{(k)}$ ，其中 $k < n$ ，有

$$Y_i^{(k)} = \sigma(W^{(e)} X_i^{(n)} + b^{(k)}) \quad (7)$$

$Y_i^{(k)}$ 为 $X_i^{(n)}$ 在 k 维空间的向量表示， σ 为编码神经元激活函数，通常使用的是 Sigmoid 函数，本文所使用的激活函数都为 Sigmoid。 $W^{(e)}$ 为编码权重矩阵， $b^{(k)}$

为 k 维编码偏差向量。

为了得到训练误差，需要将 $Y_i^{(k)}$ 作为输入，通过解码神经元解码后得到 $Z_i^{(n)}$ ，解码过程如下，

$$Z_i^{(n)} = \sigma(W^{(d)} Y_i^{(k)} + b^{(n)}) \quad (8)$$

σ 为解码神经元激活函数， $W^{(d)}$ 为解码权重矩阵， $b^{(n)}$ 为 n 维解码偏差向量。

评估训练模型的好坏主要由目标损失函数决定。目标损失函数经常用方差代价函数，

$$L_{MSE} = \frac{\|X_i^{(n)} - Z_i^{(n)}\|^2}{2} \quad (9)$$

整个模型的训练过程中，应用梯度下降算法对权重矩阵与偏差进行调整，

$$\begin{aligned} W^{(d)} &= W^{(d)} - \eta \frac{\partial L_{MSE}}{\partial W^{(d)}} \\ &= W^{(d)} - \eta Z_i^{(n)} \sigma'(W^{(d)} Y_i^{(k)} + b^{(n)}) \end{aligned} \quad (10)$$

$$\begin{aligned} b^{(n)} &= b^{(n)} - \eta \frac{\partial L_{MSE}}{\partial b^{(n)}} \\ &= b^{(n)} - \eta Z_i^{(n)} \sigma'(W^{(d)} Y_i^{(k)} + b^{(n)}) \end{aligned} \quad (11)$$

其中 η 为学习速率，由于 Sigmoid 函数的性质， σ 的导函数为 $\sigma'(x)$ ，当 x 取 Sigmoid 函数两端的值（很大或者很小的值）时，Sigmoid 函数的斜率很小，即 $\sigma'(x)$ 的值会很小，因此导致 $W^{(d)}$ 与 $b^{(n)}$ 的权值下降的很慢，针对这个问题，我们引入了交叉熵损失函数，函数定义如下，

$$\begin{aligned} loss^{entropy} &= -\frac{1}{n} \sum_x \left[X_i^{(n)} \lg Z_i^{(n)} + (I - X_i^{(n)}) \lg (I - Z_i^{(n)}) \right] \end{aligned} \quad (12)$$

其中 I 为单位向量， $loss^{entropy}$ 对权重 $W^{(d)}$ 与偏差 $b^{(n)}$ 导数为

$$\frac{\partial L_H}{\partial W^{(d)}} = \frac{1}{n} \sum_x x_i (\sigma(W^{(d)} Y_i^{(k)} + b^{(n)}) - X_i^{(n)}) \quad (13)$$

$$\frac{\partial L_H}{\partial b^{(n)}} = \frac{1}{n} \sum_x (\sigma(W^{(d)} Y_i^{(k)} + b^{(n)}) - X_i^{(n)}) \quad (14)$$

可以看出公式中没有 σ 的导函数 $\sigma'(x)$ ，因此避免了损失函数更新过慢的问题。而权值更新取决于 $(Z_i^{(n)} -$

$X_i^{(n)}$)即误差,误差越大更新越快,误差越小更新越慢。为了使模型学习到局部特征,引入LLE中的局部线性嵌入损失函数,

$$loss^{LLE} = \sum_i |Y_i^{(k)} - \sum_j A_{ij} Y_j^{(k)}| \quad (15)$$

其中A为邻接矩阵,通过加入局部线性嵌入的限制,使得网络在进行表征学习后,节点间保持嵌入之前的邻接关系。为了防止模型的过拟合,常用的做法是在损失函数中加入L2范数规则限制,

$$L2 = \frac{1}{2} \sum_{k=1}^K (||W^{(e)}||_F^2 + ||W^{(d)}||_F^2) \quad (16)$$

最后得到的损失函数为

$$L_H(X, Z) = loss^{entropy} + \beta loss^{LLE} + L2 \quad (17)$$

单层的自动编码器是利用单层非线性函数学习网络的结构特征,只含有一个隐藏层,是一种浅层的学习模型。浅层学习模型的局限性在于有限样本和计算单元情况下对复杂函数的表示能力有限,针对复杂问题其泛化能力受到一定制约。自动编码器可以使用一种堆叠技术,实现更深层次的网络学习。堆叠的自动编码器的训练过程如图2所示。 $X_i^{(n)}$ 作为第一层编码输入后输出 $Y_i^{(n-1)}$,再将输出的 $Y_i^{(n-1)}$ 作为新的输入进入到第二层自动编码器,堆叠自动编码器由多个自动编码器通过这样的方式堆叠,逐层进行非监督预训练,最终保留 $Y_i^{(k)}$ 作为节点i的表征向量。

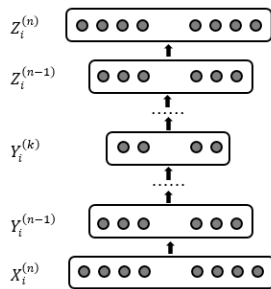


图2 堆叠自动编码器无监督训练过程

3.4 路网的方向性特征

由于城市路网道路有方向性特征,横向道路往往会与纵向道路连接,为了引入这种特征,定义集

合C表示横向道路节点, $n=|C|$ 为横向节点个数,集合D表示纵向道路节点, $m=|D|$ 为纵向节点个数,定义图G邻接限制矩阵L有以下形式,

$$L = \begin{pmatrix} n \times n & n \times m \\ m \times n & m \times m \end{pmatrix} \quad (18)$$

其中 $n \times n$ 与 $m \times m$ 矩阵部分为全零矩阵, $n \times m$ 与 $m \times n$ 为全1矩阵,定义矩阵S中的元素值 $S_{(i,j)}$ 与 $L_{(i,j)}$ 的&计算,则其结果 $R_{(i,j)}$,

$$R_{(i,j)} = \begin{cases} 0, & L_{(i,j)} = 0 \\ S_{(i,j)}, & L_{(i,j)} = 1 \end{cases} \quad (19)$$

$R=S \& L$ 保存了路网的方向性特征对结果的影响,通过添加L限制提高预测的准确度。

3.5 KAENE 模型

整个模型的框架描述如下,首先,在进行网络表征学习之前需要计算路网的Katz相似矩阵 S^{Katz} ,然后标准化后得到X,利用自动编码器对X矩阵进行网络嵌入学习得到k维网络表征矩阵 $Y^{(k)}$,利用 $Y^{(k)}$ 表征矩阵重构邻接矩阵后得到Z表征重构邻接矩阵,对Z矩阵添加L限制最后得到R预测结果矩阵。模型的预测过程如算法1所示。

算法1 KADNE 路网链路预测算法

输入: 路网 $G=(V,E)$ 、Katz 参数 α 、局部线性嵌入系数 β 、表征维度 k 、迭代次数 n 、学习速率 η

输出: R表征预测结果矩阵

①输入G根据式(1)得到邻接矩阵A

②输入A、 α ,根据式(5)计算 S^{Katz}

③对 S^{Katz} 根据式(6)进行标准化后得到X

④for $i \leftarrow 1$ to n do :

⑤输入X、k,根据式(7)计算 $Y = Y^{(k)}$

⑥根据式(8)计算 $Z = Z^{(n)}$

⑦根据式(17)计算 $Loss = L_H(X, Z)$

⑧输入 η 、 β ,利用用 Adma 梯度下降算法和反向传播算法最

小化 $Loss$ 为目标,以学习速率 η 更新模型权重以及偏差参数

⑨endfor

⑩得到网络表征向量 $Y = Y^{(k)}$

⑪根据式(8)重建邻接矩阵,得到表征重构矩阵 Z

⑫根据式(18)得到 L

⑬根据式(19)计算预测邻接矩阵 $R = Z \& L$

4 数据准备与处理

4.1 路网建模与数据处理

OpenStreetMap^[23]是由 ODbL (Open Database License)授权的共享开放地图数据库,OSM 是 OpenStreetMap 定义的地图 XML 格式文件,其主要包含了三种数据基本单元即节点(nodes)、路(ways)以及关系(relations),nodes 定义了地图中点的坐标,ways 定义了路段,一条 way 由多个 node 节点组成,relations 则定义了基本单元之间的关系。

在路网建模方面,主要有两种拓扑建模方法,分别为主方法与对偶拓扑网络方法^[24]。主方法是直接将道路抽象为边、交叉口抽象为节点,而对偶拓扑方法则恰恰相反,它将道路按路名映射为节点、交叉口映射为边。如图 3 所示,可以看出采用对偶方法不但使用了更少的节点与边,而且在功能的表达上更能体现其结构特点。

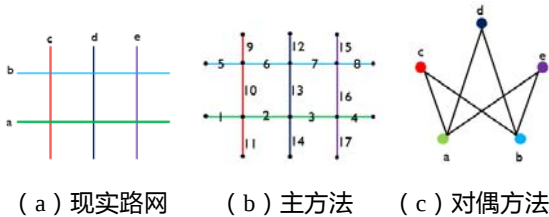


图 3 路网拓扑模型

原始采集的路网数据具有边缘节点独立性(边缘节点度为 1)以及网络非联通性等问题,可采用节点删除和保留最大联通子图两个方法解决。

4.2 数据集

本文采集了国内外具有代表性的几个城市路网数据集,经过对偶拓扑建模以及数据处理后,形成的数据集信息如表 1 所示:

表 1 国内外具有代表性的城市路网数据集信息表

城市名称	节点	边	平均度
HengYang.China	726	1430	1.97
ChangSha.China	5018	10497	2.09
HangZhou.China	11299	23772	2.10
Rockford.USA	4767	10559	2.22
Seattle.USA	8191	17732	2.16
SanFrancisco.USA	11669	24067	2.06

从表中数据我们可以看出路网的平均度均在 2.0 左右,相比于科学家论文合作网络(平均度 21.10)、Facebook 朋友网络(平均度 25.64)、Internet 拓扑网络(平均度 9.86)等^[25]网络来说,路网则显得更稀疏,而网络的稀疏性也是复杂网络链路预测需要解决的一个难题。

4.3 训练集与测试集划分

在进行链路预测之前,我们需要划分训练集与测试集,通过随机采样的方式去掉 20%的边,这些被去掉的边会成为实验的真实测试数据。如图 4 所示,实线与虚线共同构成了整个网络结构,虚线部分的边集合是通过随机采样选择的测试数据,实线部分的边集合是训练数据集,首先将训练数据放入模型中训练,然后将训练好的模型对测试集进行预测并计算 MAP 指标。

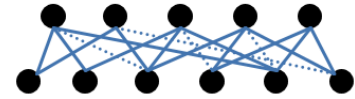


图 4 训练集与测试集的划分

4.4 评价方法

评价复杂网络链路预测模型的常用方法是 MAP^[5]。MAP 估计每个节点的预测精度值,并计算所有节点的平均预测精度,计算方法如下:

$$Pre@k(i) = \frac{|\{j|i,j \in V, index(j) \leq k, \Delta_i(j)=1\}|}{k} \quad (20)$$

$Pre@k$ 是 $topk$ 条边的预测精确度 其中 $index(j)$ 表示经排序的第 j 个节点, $\Delta_i(j) = 1$ 表示节点 i 与节点 j 存在一条边。定义节点 i 的预测精确度为

$$AP(i) = \frac{\sum_j Pre@j(i) * \Delta_i(j)}{|\{\Delta_i(j)=1\}|} \quad (21)$$

则 MAP 的计算方式为

$$MAP = \frac{\sum_i AP(i)}{|V|} \quad (22)$$

其中, V 是网络中的节点集合。

5 实验结果与分析

5.1 实验内容

本文实施以下几个对比实验：

(1) 与常用的链路预测算法在时间耗费与预测精度上的对比。一般来说预测精度是首要评价指标, 其次是时间耗费。通过该实验可以直接评价模型的好坏。

(2) 对比不同的表征维度对不同路网预测精度的影响。不同的表征维度对实验结果也会产生影响, 我们希望增加该实验来找出合适的表征维度。

(3) 不同梯度下降算法对 $loss$ 函数的影响。梯度下降算法对找到 $loss$ 函数的极值以及在优化模型的效率上有着重要的意义, 我们希望能够找到合适的梯度下降算法对模型进行训练。

(4) 自动编码器的层数对预测精度的影响。通过构建多层网络模型, 研究网络的层数对预测结果的影响, 便于最优的网络层次选择。

(5) 路网网络嵌入的二维可视化效果。我们希望了解模型在进行表征学习时, 模型学到了什么样特征, 通过将节点映射到平面空间, 从而可以直观

地展示出来。

5.2 时间耗费与预测精度对比

本文对比了 18 种链路预测算法, 其中包括了 10 种相似度算法以及 8 种网络嵌入算法, 10 种基于相似度算法中又包括了 5 种局部相似度指标算法 [CN, Jaccard, PA, AA, RA]^[6], 2 种路径相似度指标算法 [LP, Katz]^[6], 3 种随机游走相似度指标算法 [RWR, ACT, Cos]^[6]; 8 种网络嵌入算法中包括了 6 种基于矩阵分解的算法 [NMF, SVD, LAP, HOPE, LLE], 一种随机游走算法 Node2Vec 以及一种深度学习算法 SDNE。各算法预测 MAP 指标与时间消耗如表 2 所示 (时间单位: 秒)。基本参数设置如下, 网络嵌入算法表征维度都为 128, 迭代次数均为 50 次。

图 5 为各算法的平均 MAP 预测指标, 图 6 为各算法的平均时间耗费, 从图 5 和图 6 中我们可以看出在基于相似度算法中, 基于路径相似度的 Katz 算法与基于随机游走相似度的 ACT、Cos 算法取得了不错的效果, 但是在时间耗费上, ACT 与 Cos 的时间耗费较多。在基于网络嵌入技术的算法中, 深度学习模型如 SDNE、KAENE 所耗费的时间是其他网络表征模型的几倍或几十倍, 其原因主要在于深度学习的迭代机制以及多层神经网络的参数过多导致的。在预测精确度上, KAENE 取得了最高的平均 MAP 指标, 同时它也是耗时最多算法。SVD 在小数据集中取得了不错的效果, 同时它是耗时最少的网络表征模型, 全局精确度最差的算法是 Jaccard, 全局耗时最少的算法是 PA (优先连接指标) 算法。

表 2 链路预测算法时间耗费与预测精度表

s

Method	HengYang		ChangSha		HangZhou		Rockford		Seattle		SanFrancisco	
	Time	MAP	Time	MAP	Time	MAP	Time	MAP	Time	MAP	Time	MAP
CN	0.598	0.150	28.216	0.158	211.551	0.155	22.232	0.161	103.216	0.166	229.172	0.151
Jaccard	0.393	0.153	27.649	0.156	321.805	0.154	22.944	0.160	86.054	0.165	239.904	0.147
PA	0.669	0.253	27.053	0.245	138.337	0.240	22.588	0.230	73.224	0.239	136.183	0.216
AA	0.446	0.153	28.376	0.161	329.088	0.158	22.646	0.165	108.973	0.167	303.026	0.153
RA	0.409	0.155	27.081	0.162	304.336	0.159	21.718	0.166	107.413	0.167	296.361	0.154
LP	1.289	0.224	91.129	0.224	1299.226	0.228	74.183	0.233	386.289	0.222	1130.770	0.201

Katz	5.457	0.272	53.008	0.264	500.587	0.262	51.473	0.256	233.878	0.263	514.554	0.230
RWR	7.674	0.198	73.553	0.211	1017.762	0.211	62.314	0.215	351.615	0.197	744.623	0.173
ACT	7.803	0.272	759.828	0.271	10577.082	0.267	534.125	0.257	2779.629	0.264	9434.607	0.251
Cos	4.584	0.268	794.242	0.250	9705.852	0.248	517.142	0.269	3511.995	0.245	8761.833	0.243
NMF	32.639	0.361	1302.549	0.356	6500.033	0.350	1183.974	0.327	3450.774	0.350	7259.661	0.308
SVD	1.438	0.650	65.607	0.463	387.730	0.382	55.179	0.439	196.515	0.426	407.642	0.362
GF	8.662	0.333	132.432	0.314	539.370	0.308	118.519	0.291	324.221	0.310	581.470	0.265
HOPE	3.013	0.576	88.405	0.426	482.998	0.388	76.742	0.403	257.071	0.400	533.500	0.343
LAP	9.646	0.619	398.847	0.537	2089.411	0.495	420.597	0.484	1355.974	0.521	2365.255	0.428
LLE	7.368	0.638	414.252	0.586	3401.714	0.562	536.283	0.516	3297.191	0.572	3914.500	0.490
Node2vec	6.110	0.434	134.158	0.507	689.923	0.464	129.721	0.420	361.078	0.475	755.138	0.415
SDNE	2006.504	0.534	8008.310	0.508	20229.245	0.505	12027.505	0.523	13569.462	0.507	20355.337	0.408
KAENE	2966.766	0.643	9775.685	0.554	26665.995	0.583	13322.652	0.506	15224.112	0.577	26554.688	0.517

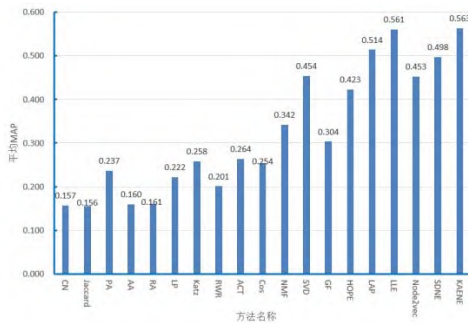


图5 各算法平均 MAP 指标

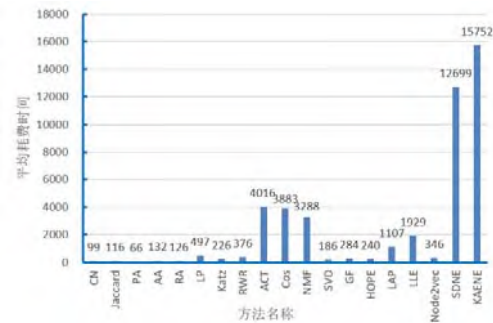


图6 各算法平均耗时

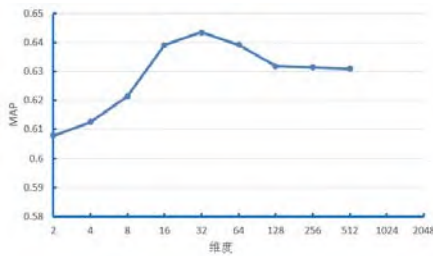


图7 HengYang 路网不同表征维度时预测精度

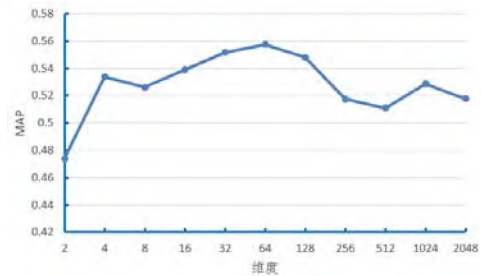


图8 ChangSha 路网不同表征维度时预测精度

5.3 不同表征维度对算法的影响

为了探索不同表征维度对预测结果的影响，本文选取了 ChangSha 与 HengYang 两个城市的路网数据集。一般来说，网络表征的维度通常为 2^n 且 $2^n < |V|$ ， n 为自然数且大于 0。

图 7 中由于节点数量为 726，因此只能嵌入到 512 维。对比图 7 与图 8 我们可以发现随着维度的不断增加，预测精度值形成了正态分布，因此，我们在选择表征维度参数时一般不会选择太小或太大。

另外一个发现是，随着数据集网络节点数量的不断增加，最优表征维度也在提高，比如 HengYang 数据集 ($|V|=726$) 的最优网络表征维度为 32，而 ChangSha 数据集 ($|V|=5018$) 的最优网络表征维度为 64。

5.4 梯度下降算法对 loss 的影响

对损失函数进行梯度下降时所使用梯度下降算法的选择也非常重要，本文对比了 SGD、Adam、Adadelata、RMSprop、Nadam、Adagrad、Adamax 7

种梯度下降算法在 ChangSha 路网数据集中的表现，梯度下降算法相关参数均使用默认参数，模型迭代次数为 50 次。实验结果如图 9 所示。从图中可以看出使用 Adam 算法取得了最快的下降效果及最小的 $loss$ 值，因此效果最好。而 SGD 算法刚开始虽然取得不错的下降效果，但是随着迭代次数的增加，其不稳定性开始体现出来，使得 $loss$ 值跳出了极小值范围，最终进入了 $loss$ 值很高的平稳状态，因此效果最差。其他 5 种算法刚开始都处于平稳下降状态，最后随着迭代次数的增加都逼近了极小值，因此效果一般。本文中，我们选用了 Adam 梯度下降算法。

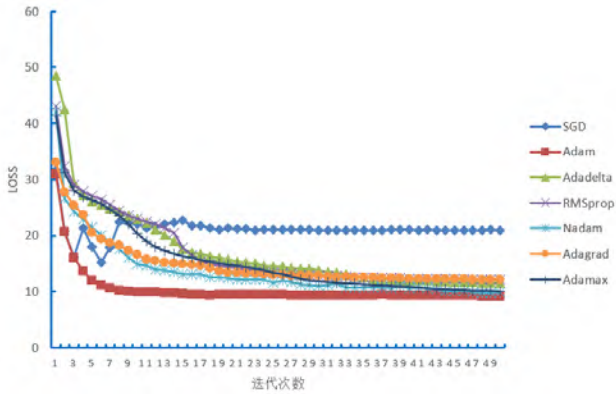


图 9 梯度下降算法对损失函数的影响

5.5 自动编码器的层数对预测精度的影响

为了探究自动编码器层数对预测精度的影响，本文设计了自动编码器层数与预测结果的对比实验，实验选取了 HengYang 作为实验数据集，相关参数设置如下：初始表征维度为 512，每层表征维度呈等差数列递减，最终表征维度为 128，迭代次数均为 50 次。实验结果如图 10 所示，从图中可以看出随着层数的增加，MAP 指标一开始整体呈现上升趋势，在达到 40 层以后，MAP 指标开始下降，说明此时模型泛化能力开始下降。

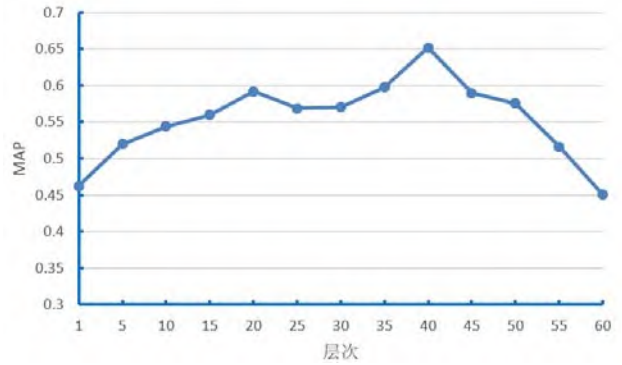
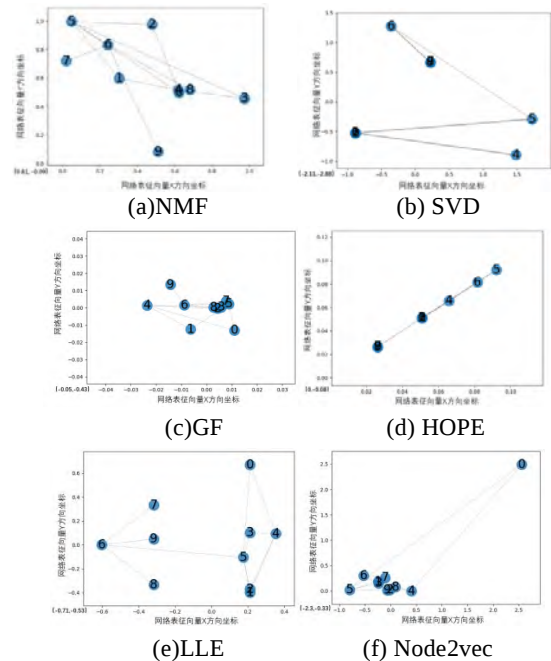


图 10 自动编码器层数对预测精确度的影响

5.6 模型可视化效果

由于路网的节点数量较多，在二维平面上展示时，导致大部分节点都重叠在了一起。因此，在本次实验中，我们仍选用了图 1(a) 中的复杂网络图。此外，我们对比了 [NMF, SVD, GF, HOPE, LLE, Node2vec, SDNE] 七种网络嵌入模型在该数据集上可视化的表现（注：本次实验 KAENE 无法添加 L 限制）。如图 11 所示，从网络拓扑结构上看 LLE 与 KAENE 与图 1(a) 更相近，而 SDNE 则更倾向于学习到了网络的“二阶近似”特性，即具有共同邻居的两个节点之间具有相似性，因为它将节点 [4,5] 聚集到了一起。



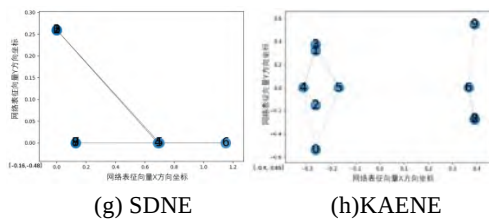


图 11 网络特征可视化效果

6 结束语

本文从路网链路预测的应用背景出发,结合深度学习的思想,提出了基于 KAENE 的路网链路预测模型。实验结果表明,KAENE 取得了良好的表现。虽然,KAENE 在预测精度上较高,但是其逐层训练机制与高迭代次数,往往使得模型需要耗费大量的训练时间,对于海量的路网数据来说,需要更多的计算资源和内存空间,对于一些 GIS 研究机构来说,是不可接受的。因此,寻求更加高效的链路预测算法或者实现模型的并行计算将会是本文后续要解决的问题。

参考文献:

- [1] Zhao F, Sun H, Wu J, et al. Analysis of road network pattern considering population distribution and central business district[J]. PloS One, 2016, 11(3): e0151676.
- [2] Zhao Y, Wu Y J, Levina E, et al. Link prediction for partially observed networks[J]. Journal of Computational and Graphical Statistics, 2017.
- [3] 宋海权, 郭进, 刘刚. 基于复杂网络的城市道路重要度评价及路网自动综合方法[J]. 测绘工程, 2017, 26(1): 8-12.
- [4] 赵国锋, 苑少伟, 慈玉生. 城市路网的复杂网络特性和鲁棒性研究[J]. 公路交通科技, 2016, 33(1):119-124.
- [5] Goyal P, Ferrara E. Graph Embedding Techniques, Applications, and Performance: A Survey[J]. Knowledge-Based Systems, 2018.
- [6] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651-661.
- [7] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C] Advances in Neural Information Processing Systems. 2002: 585-591.
- [8] 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.
- [9] Wang D, Cui P, Zhu W. Structural Deep Network Embedding[C] ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1225-1234.
- [10] Ahmed A, Shervashidze N, Narayanamurthy S, et al. Distributed large-scale natural graph factorization[C] Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 37-48.
- [11] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.
- [12] Lee D. Learning the parts of objects with nonnegative matrix factorization[J]. Nature, 1999, 401(6755):788.
- [13] Dan K. A Singularly Valuable Decomposition: The SVD of a Matrix[J]. College Mathematics Journal, 1996, 27(1):2-23.
- [14] Ou M, Cui P, Pei J, et al. Asymmetric Transitivity Preserving Graph Embedding[C] KDD. 2016: 1105-1114.
- [15] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
- [16] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):págs. 212-223.
- [17] Bengio Y, Lamblin P, Dan P, et al. Greedy layer-wise training of deep networks[C]// International Conference on Neural Information Processing Systems. MIT Press, 2007:153-160.
- [18] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533.
- [19] Hinton G E. A practical guide to training restricted Boltzmann machines[M] Neural Networks: Tricks of the Trade. Springer, Berlin, Heidelberg, 2012: 599-619.
- [20] Chang S, Han W, Tang J, et al. Heterogeneous Network Embedding via Deep Architectures[C] ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015:119-128.
- [21] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.
- [22] Zakzeski J, Bruijninx PCA, Jongerius AL, et al. The catalytic valorization of lignin for the production of renewable chemicals[J]. Chemical Reviews, 2010, 110(6): 3552-3599.
- [23] Haklay M, Weber P. OpenStreetMap: User-Generated Street Maps[J]. IEEE Pervasive Computing, 2008, 7(4):12-18.
- [24] 赵玲, 邓敏, 王佳璐, 等. 应用复杂网络理论的城市路网可靠性分析[J]. 测绘科学, 2013, 38(3):64-68.
- [25] Kunegis J. Konect: the koblenz network collection[C] Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 1343-1350.