计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700020

基于特征学习的链路预测模型 TNTlink

王 慧[1,2] 乐孜纯[3] 龚 轩[1] 左 浩[1] 武玉坤[1]
1 浙江工业大学计算机科学与技术学院 杭州 310023
2 江西理工大学应用科学学院 江西 赣州 341000
3 浙江工业大学理学院 杭州 310023

摘 要 在合作作者网络中,链路预测可以预测当前网络中缺失的链接,以及新的或已解散的链接,根据网络中观测到的信息来推断两位作者在不久的将来是否会产生合作,对于挖掘和分析网络的演化、重塑网络模型具有重要意义。链路预测是计算机科学和物理学的重要研究方向,对此已有较深入的研究,其主要研究思路是基于马尔可夫链、机器学习和无监督的学习。然而,这些工作大多只使用单一的特征,即基于网络拓扑特征或者属性特征进行预测,很少将这些跨学科的特征组合考虑,结合多学科特征进行链路预测的研究非常少。文中设计开发了 TNTlink 模型,该模型结合网络拓扑特征、基本特征和附加特征,并结合物理学和计算机科学的领域知识,利用深度神经网络将这些特征集成到一个深度学习框架中,其在解决链路预测问题时取得了不错的效果。文中使用了 5 个数据集(ca-AstroPh,ca-CondMat,ca-GrQc,ca-HepPh 和 ca-HepTh),包含 69032 个节点和 450617 条边,从捕获的信息中利用二进制相似度和模糊余弦相似度计算和识别特征。如果节点在这些特征中表现出更多的相似性(如相似的节点、相同的关键字或彼此之间密切的关系),则两个节点间更有可能生成链接。除了考虑节点的特征外,还考虑了节点重要性对链路形成的影响,进而提出了一种新的链路预测指标 MI,以区分强影响和弱影响,对节点的重要影响进行建模。将所提模型与主流分类器在 5 个数据集上进行比较,结果表明 MI 和 TNTlink 有效地提高了链路预测的 AUC 值。

关键词:链路预测;拓扑特征;模糊余弦相似性;深度学习;基本特征;附加特征
中图法分类号 TP391

TNTlink Prediction Model Based on Feature Learning

WANG Hui[1,2],LE Zi-chun[3],GONG Xuan[1],ZUO Hao[1] and WU Yu-kun[1]
1 College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China
2 College of Applied Science,Jiangxi University of Science and Technology,Ganzhou,Jiangxi 341000,China
3 College of Science,Zhejiang University of Technology,Hangzhou 310023,China

Abstract In the co-author network,link prediction can predict the missing links in the current network and the new or disbanded links. It is of great significance for mining and analyzing the evolution of the network and remaking the network model to infer whether the two authors will cooperate in the near future according to the observed information in the network. As an important research direction of computer science and physics,link prediction has been studied in depth up to now. Their main research idea is based on the markov chain,machine learning and unsupervised learning. However,most of these work use only a single feature,namely the network topology features or attribute features to predict,few will consider these interdisciplinary features,and papers combined with multidisciplinary on link prediction are fewer. This paper designed and developed the TNTlink model. This model combines the network topology features,basic featues and the additional features,combines physics and computer science domain knowledge,and uses the depth of neural network to integrate these features into a deep learning framework dealing with the problem of link prediction,and good results have been achieved. Five data sets (ca-astroph,ca-condmat,ca-grqc,ca-hepph and ca-hepth) were used in this paper,containing 69032 nodes and 450617 edges. Binary similarity and fuzzy cosine similarity were used to calculate and identify these features from captured information. If nodes show more similarity in these features (for example,similar nodes,the same keywords,or a close relationship between them),the two nodes are more likely to generate links. Besides the features of nodes,the influence of node importance on link formation was also considered. A new link prediction index MI was proposed to distinguish strong effects from weak effects and to model the important effects of nodes. The proposed model was

到稿日期:2019-07-02 返修日期:2019-12-07 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:浙江省“一带一路”国际合作专项资金(2015C04005)
This work was supported by the Special Funding of “the Belt and Road” International Cooperation of Zhejiang Province (2015C04005).
通信作者:王慧(540168713@qq.com)

compared with mainstream classifiers on five datasets. The results show that MI and TNTlink can effectively improve link prediction AUC value.

Keywords Link prediction, Topological features, Fuzzy cosine similarity, Deep learning, Basic features, Additional features

1 引言

协作无处不在，如学术研究、产品开发和决策，96%的学术研究是合作的结果。在合作网络中，如何探索和分析合作关系是非常重要的。我们可以将其可视化为一个图，其中节点表示作者，边(link)对应作者之间的关系[1]。因此，链路预测问题涉及到根据观察到的网络信息推断两位作者是否会在不久的将来建立合作作者关系。

随着协作的加强，其因广泛的适用性而需要一个强大而通用的框架，我们提倡深度学习方法，它在链路预测的研究中得到了最广泛的关注，能从根本上找到有效的特征和良好的分类算法。目前，已有研究讨论了基于特征分类的链路预测问题[2-3]，如使用节点和链接的属性(如用户的年龄、兴趣和朋友)可以大大提高链路预测的性能。Li 等[3]提出了一种基于图的学习方法，该方法利用受教育程度、年龄、关键词、书名、引言等特征来预测二部图网络中的用户-条目链接。Scellato 等[4]和 Wang 等[5]在基于位置的网络中考虑了位置特征、社交特征和全局特征，利用监督学习框架进行链路预测。Pavlov 等[6]和 Ichise 等[7]提出了利用抽象信息、研究标题和事件信息来改进联合作者网络中链路预测的语义方法。Hansan 等[8]引入了邻近特征、聚合特征、拓扑特征，并使用不同类别的监督学习算法(决策树、KNN、多层感知器、SVM、RBF 网络)进行链路预测，其中 SVM 以较小的优势取胜。Wang 等[9]利用网络的拓扑特征，提出了一种基于邻域节点对边的贡献的相似性指标，该指标对于聚类系数较低的网络具有较高的效率。Adity 等[10]引入节点特征、边缘特征，提出了一种用于网络中可伸缩特征学习的半监督算法 Node2vec。Feng 等[11]提出了基于受限玻尔兹曼机(Restricted Boltzmann Machine)的深度信念网络来抽象和表示特征。2019 年，Yang 等[12]利用拓扑特征和社区理论提出了一种基于局部社团和节点相似性的链路预测算法(High-level Community and Robust Pearson's)，并将该算法与 11 种知名算法进行了对比，取得了优良的链路预测性能。

通过对上述基于特征的学习方法的比较，可以得到以下观察结果。1)链路预测方法通常直接使用基本的单个特征模型(如节点特征、拓扑特征、网络特征和社会特征)或修改其他知名模型。这些特性很少被同时考虑，如果只考虑单个因素，则会影响链路预测的性能、精度和实际应用。2)特征选择或模型构造是链路预测的核心任务，也是这些链路预测学习方法的主要区别。传统的学习框架主要站在物理学的视角研究链路预测问题，多数采用拓扑特征、节点特征和社区理论；而计算机领域的专家站在计算机的视角研究链路预测问题，主要使用属性特征。因此，利用跨学科特征进行链路预测的研究较少。前文提到的链路预测方法虽然结合了几种特征，但采用的特征都是基于物理学的领域知识。为了处理特定领域的链路预测问题，我们需要构造和使用领域特定的特征[5,8]；同时，充分挖掘各领域的优势，采用跨领域知识进行链路预测，改善链路预测性能是目前研究的重中之重。3)传统的浅层学习方法在有效表达非线性高维空间特征方面存在局限性。通过比较 RBM, RF, Propflow 和 Node2vec，深度学习链路预测方法允许计算模型学习多层次抽象的数据表示，可以提高预测性能，尤其是预测精度。因此，本文充分利用可获得的信息，利用深度学习提取有效特征，并将这些特征结合起来，建立一个有效的链路预测模型。在深度学习框架下提取的特征如下。

(1)拓扑特征[13-15]：包括 Degree, Volume, Common Neighbors, Jaccardcoefficient, Adamic/ Adar, Katz, Maxflow, ShortPath, ProFlow, PageRank, PA，这些特征大多与网络拓扑有关，可以显著提高分类器的性能。

(2)基本特征：包括关键字匹配、分类码、邮政编码、地理距离，采用二进制相似度进行计算。

(3)附加特征：包括作者隶属度和论文标题相似度，采用模糊余弦相似度进行计算。

本文结合拓扑特征、基本特征和附加特征，提出了一种新的基于特征学习的链路预测模型 TNTlink，该模型可在深度学习框架下预测未来的链路。仿真实验的结果表明了所提模型的有效性，当预测具有相似特征的作者之间的链接时，AUC 值高达 90.7%。特别地，本文认为基于基本特征和附加特征的链路预测可以大大提高链路预测的最终性能。此外，本文假设增加节点的影响力有助于未连接的节点在未来相互连接。通过对显著影响(MI)进行建模，提出了一种新的链路预测指标。将所提模型与 5 个基准数据集的主流分类器进行比较，结果表明链路预测的 AUC 值有明显的提高。

本文第 2 节介绍了理论基础，以及本文使用的方法和技术，如模糊余弦相似度、节点重要性的定义、DNN；第 3 节提出了一种新的链路预测模型 TNTlink，该模型在深度学习框架下结合了拓扑特征、基本特征和附加特征，还介绍了实验数据集和评价指标；第 4 节给出了仿真结果及 5 个数据集下的 AUC 值和 ROC 曲线，并讨论了去除特征对结果的影响；最后总结全文。

2 理论基础

2.1 模糊余弦相似度

余弦相似度度量字符级的语义相似度。为了分析合作作者网络，本文使用隶属关系作为两个作者之间的相似度指标。如果只使用节点相似性，则会有很多相同的相似性，这在复杂网络中是不兼容的。

将两位作者的隶属关系转化为词频向量 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_j$ 后[16]，定义两位作者之间的相似余弦 $s_{i,j}^{cs}$ 为：

$$S_{i,j}^{cs}(t)=\frac{(c_i,c_j)}{\|c_i\|\times\|c_j\|} \tag{1}$$

本文提出的模糊余弦相似度，首先计算两个向量的余弦相似度，然后根据计算结果划分模糊域，最后确定相似度值是不相似、相对相似还是完全相似，具体分类如图 1 所示。

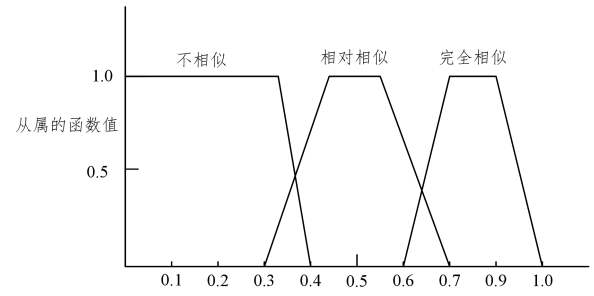


图 1　模糊余弦相似度的分类

Fig. 1　Fuzzy cosine similarity classification

2.2　节点重要性的定义

目前，在评估网络中节点的重要性方面已有大量的研究，常用的衡量节点重要性的指标是节点的度，即节点更倾向于与度大的节点连接。然而，在研究网络结构之后发现，这种影响是通过端点之间的路径构建的关系，短路径比长路径对节点连接的影响更大，而不仅仅依靠节点的度，特别是通过短路径连接另一个端点能够产生强关系，可以带来更大的影响，同理，通过长路径产生的关系带来的影响更弱。因此，本文提出了链路预测指标 MI，并对其影响进行了建模，区分了强影响和弱影响。MI 模型的定义如下：

在无向和无权网络 $G(V,E)$ 中，端点之间的关系是通过它们之间的路径建立的，在端点 v_i 和 v_j 之间，短路径特别是两跳路径表示强关系，而弱关系发生在 v_i 和 v_j 通过长路径连接时。强关系的定义为：

$$S_{v_iv_j}^{strong}(t)=|\Gamma(v_i)\bigcap\Gamma(v_j)|*\left[\sum_{\tau=1}^{t}\pi_{v_iv_j}(\tau)+\sum_{\tau=1}^{t}\pi_{v_jv_i}(\tau)\right] \tag{2}$$

其中，$\pi_{v_iv_j}(\tau)$ 表示 v_i 到 v_j 的转移概率，即从 v_i 到 v_j 游走了 τ 步的转移概率；t 表示经过了 t 步从 v_i 到 v_j，$|\Gamma(v_i)\bigcap\Gamma(v_j)|$ 表示节点 v_i 和 v_j 共同邻居的数目。弱关系产生的影响为弱影响，可以定义为：

$$S_{v_iv_j}^{weak}(t)=\left(\sum_{l=3}^{\infty}|paths^{\langle l\rangle}{}_{v_i,v_j}|\right)*\left[\sum_{\tau=1}^{t}\pi_{v_iv_j}(\tau)+\sum_{\tau=1}^{t}\pi_{v_j,v_i}(\tau)\right] \tag{3}$$

其中，$|paths_{v_i,v_j}^{\langle l\rangle}|$ 表示 v_i 和 v_j 之间长度为 l 的路径数目。本文认为路径为三跳以上的路径为弱关系，路径的长度从 $l=3$ 到无穷大。综上所述，重要性的影响 $S_{v_iv_j}^{MI}(t)$ 的定义如式(4)所示，应强调节点 v_i 和 v_j 的主要影响，并惩罚它们之间的弱影响。

$$\begin{aligned}S_{v_iv_j}^{MI}(t)&=S_{v_iv_j}^{strong}(t)+S_{v_iv_j}^{weak}(t)\\&=|\Gamma(v_i)\bigcap\Gamma(v_j)|*\left[\sum_{\tau=1}^{t}\pi_{v_iv_j}(\tau)+\sum_{\tau=1}^{t}\pi_{v_jv_i}(\tau)\right]+\\&\left(\sum_{l=3}^{\infty}|paths_{v_i,v_j}^{\langle l\rangle}|\right)\beta\left[\sum_{\tau=1}^{t}\pi_{v_iv_j}(\tau)+\sum_{\tau=1}^{t}\pi_{v_jv_i}(\tau)\right]\end{aligned} \tag{4}$$

MI 指数强调来自共同邻居中的强关系，通过应用惩罚参数 β 来限制三跳以上路径的弱影响。β 在 MI 中扮演着惩罚的角色，它的取值为 $(-\infty,1]$。

2.3　DNN

深度神经网络(Deep Neural Networks，DNN)在各种机器学习任务中获得了良好的准确性，它正被广泛应用于各种场合。DNN 的体系结构(即超参数)在很大程度上决定了 DNN 的精度和性能。目前，有两种流行的 DNN 分别是全连接神经网络(或多层感知器)和卷积神经网络(CNN)。本文采用全连通神经网络，采用两层 DNN 结构[17]。

DNN 是一个函数[18]，用 $F(X)$ 表示，其输入为 $x\in R^n$，输出为 $y\in R^m$。$F(x)$ 是一个由 K 个参数 $f_i(i\in\{1,k\})$ 组成的层次模型，每一层都是一层神经元，将激活功能应用于前一层 f_{i-1} 的加权输出。每一层是由参数化权重矩阵 $\boldsymbol{w}_i$、偏置值 b_i 和激活函数 ϑ_i：$f_i(x)=\vartheta_i(\boldsymbol{w}_i\cdot x+b_i)$ 组成，如图 2 所示。一个 DNN 模型的描述如下：

$$F(x)=f_k\cdot f_{k-1}\cdot f_{k-2}\cdot\cdots\cdot f_2\cdot f_1\cdot x \tag{5}$$

$$F(x)=\vartheta_k(\boldsymbol{w}_k\cdot\vartheta_{k-1}(\boldsymbol{w}_{k-1}\cdots\vartheta_1(\boldsymbol{w}_1\cdot x+b_1)\cdots+b_{k-1})+b_k) \tag{6}$$

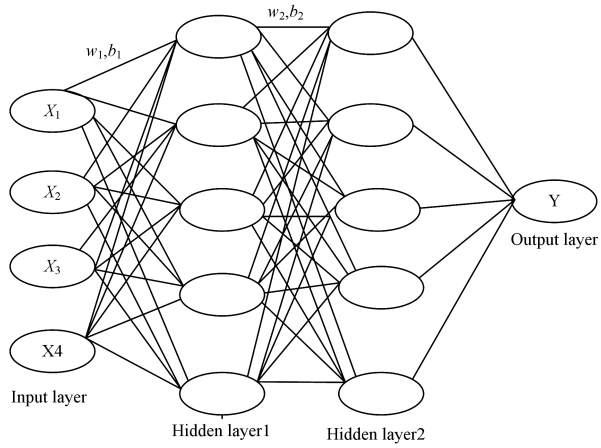


图 2　二层 DNN 模型

Fig. 2　Two-layer DNN model

2.4　比较基线

为便于比较，下面介绍 4 个经典分类器。

(1)RBM。RBM 由可见层和隐藏层组成[19]。通常用可见层单元来描述采集数据的特征，而隐藏层单元通常被认为是特征提取层。RBMs 通常使用对比发散(contrastive divergence)来进行学习。

(2)RandomForest。随机森林是一种重要的基于 bagging 的综合学习方法，可用于分类、回归等问题[20]。

(3)PropFlow。PropFlow 通过一个本地路径计算节点之间的信息，以评估两个节点之间的关系。PropFlow 值越大，未来连接的概率越大。Thi 等[21]证明了 PropFlow 优于 Common Neighbors，Adamic/Adar，Jaccard's Coefficient 和 Preferential Attachment。

(4)Node2Vec。Node2vec 是网络中学习节点连续表示的算法框架。它将节点映射到一个低维特征空间，以最大限度地保护节点的网络社区。Aditya 等认为 Node2vec 在链路预测方面优于 Common Neighbors，Jaccard's Coefficient 和 Adamic-Adar[10]。

3　链路预测模型

3.1　数据描述

为了建立一个有效的链路预测模型，必须从合作作者网络中定义和提取一组合适的特征。传统的基于拓扑和基于节点的特征在分类学习模型中很受重视。例如，可以将 VCP 度

量看作描述本地拓扑信息[22]的一个特别的特征。此外，许多研究表明，使用节点和链接的属性可以大大提高链路预测的性能。一般的链路预测模型只考虑了一般的特征，如节点和拓扑特征，但是针对某个领域的链路预测模型，也应该考虑非拓扑特征。

为了提高链路预测的性能、精度和实用性，本文根据链路预测的拓扑特征、基本特征和附加特征，提出了一种新的TNTLink模型，如图3所示。该TNTlink模型不仅包含了网络的拓扑特征，也包含了关于网络非拓扑特征的隐式信息，这些信息对于链路预测非常重要。

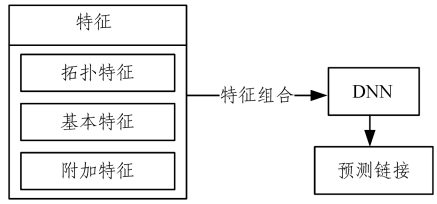


图3 TNTlink模型

Fig. 3 TNTlink model

为了清晰地表述TNTlink模型，表1对该模型所提取的具体特征进行了分析。

表1 TNTlink模型使用的特征分析

Table 1 Feature analysis used in TNTlink model

Composition	Feature	Description	remarks
Topology-Based features	In-degree	$d_{in}(v_i), d_{in}(v_j)$	
	Out-degree	$d_{out}(v_i), d_{out}(v_j)$	
	In-volume	$Vol_{in}(v_i), Vol_{in}(v_j)$	
	Out-volume	$Vol_{out}(v_i), Vol_{out}(v_j)$	
	Common Neighbors	$\lvert\varphi(v_i)\bigcap\varphi(v_j)\rvert$, $\varphi(v_i)$ is the neithbors of v_i	
	Jaccard Coefficient	$\frac{\lvert\varphi(v_i)\bigcap\varphi(v_j)\rvert}{\lvert\varphi(v_i)\bigcup\varphi(v_j)\rvert}$	
	Adamic/Adar	$\sum_{v_k\in\varphi(v_i)\cap\varphi(v_j)}\frac{1}{\log\lvert\varphi(v_k)\rvert}$	$l=5$ $\beta=0.005$
	Katz	$\sum_{l=1}^{\infty}\beta^l\cdot\lvert path^{\langle l\rangle}_{v_i,v_j}\rvert$, $path^{\langle l\rangle}_{v_i,v_j}$ is the set of all length-l paths from v_i to v_j	$l=5$
	Maxflow	Max f_{ab}, treating v_a as the source and v_b as the sink	$l=5$
	ShortPath	All shortest paths between $\varphi(v_i)$ and $\varphi(v_j)$	$l=5$
	ProFlow	If v_i and v_j are directly linked, $PF(v_i,v_j)=PF(a,v_i)\frac{w_{v_iv_j}}{\sum_{k\in\varphi(v_i)}w_{v_ik}}$	$l=5$
	PageRank	Random walk with resets in G_{ab}	
	Preferential Attachment	$\lvert\varphi(v_i)\rvert\cdot\lvert\varphi(v_j)\rvert$	
Basic features	Keyword match count	Keywork match count(v_i,v_j)	
	Classification codessimilarity	Classification codes similarity(v_i,v_j)	
	Geographical distance similarity	Geographical distance similarity(v_i,v_j)	
additional features	Author affiliation	Author affiliation similarity(v_i,v_j)	
	Papertitle similarity	Paper title similarity(v_i,v_j)	

(1)基于拓扑的特征(topology-based features)

任何网络，不管其类型或来源，必须支持基本拓扑特征，例如v_i和v_j的In-degree和Out-degree、In-volume和Out-volume、CommonNeighbors、JaccardCoefficient、Adamic、Katz、Maxflow、ShortPath、ProFlow、Preferential Attachment、PageRank。本文将用户分为两组，即精英用户和普通用户，通过使用PageRank算法来评估用户的重要性，选择前1%的用户作为精英用户，其他作为普通用户。由于这些特性在文献[23-30]中得到了广泛的研究和应用，因此本文只对这些特征进行了简单介绍。

(2)基本特征(Basic features)

根据领域知识，探索合适的领域特征是至关重要和具有挑战性的，因为它直接影响了预测模型的预测能力。本文根据作者合作网络的特点，提出了利用基本特征和附加特征进行链路预测。

TNTlink模型包括的基本特征如下。

1)关键词匹配数：keyword match count(v_i,v_j)被定义为在两个节点v_i和v_j之间关键词匹配的总数。该特征直接度量了作者之间的相似程度。这里列出了作者在论文中所介绍的所有关键词及关键词之间的相似重叠区域。重叠程度越高，它们在相关领域工作的可能性就越大，因此更有可能成为未来的合作者。

2)分类码相似性：Classification codes similarity(v_i,v_j)表示节点v_i与v_j之间分类码匹配的程度。通常，如果分类码相同，证明作者可能在相同或相似的领域研究，这些领域的研究人员通常更有可能合作。

3)地理距离相似性：Geographical distance similarity(v_i,v_j)定义了节点v_i与v_j之间的地理位置相似性。地理位置越近，作者合作的可能性越大，产生新链接的可能性就越大。

本文用两种方式描述基本特征：二进制相似性和公共属性的总数。二进制相似性是通过比较两个用户s和u的每个属性(如邮政编码、地理位置信息)来计算的。公共属性的总数是它们之间相同属性的总数。

例如：如果一个作者s声称他在浙江工作，他可能认识在浙江工作的其他人。因此，如果存在一个作者u在浙江，s更可能与u合作，而不是和其他地方的人合作，故s和u之间的

二进制相似度为1，否则为0。除此之外，通过统计特征中二进制相似度的总数得到公共属性的数目，并将其作为另一个基本特征，如果两个用户共享更多的公共属性，那么他们更可能成为朋友。

(3)附加特征(additional features)

我们还可以通过其他方式来描述两个用户属性之间的关系，例如作者从属关系或论文标题相似性。

1)作者关联相似性：Author affiliation similarity(v_i，v_j)定义了节点 v_i 与 v_j 之间的作者关联相似性。最常用的特征度量方法是二进制相似性，但是，其会导致潜在有用信息的丢失，例如，作者 u 是浙江工业大学计算机科学与技术学院的学生，而作者 s 是浙江工业大学理学院的学生，通过二进制相似性会将二者视为完全不同的隶属关系，这意味着 u 和 s 的二进制相似度为零。然而，来自同一所学校的两个不同学院的作者有可能会合作，通过二进制相似度进行判断是不合理的。基于这些原因，本文采用模糊余弦相似度来评价合作作者网络中作者之间的相似度。在上面的例子中，u 和 s 的余弦相似度约为0.503，在模糊集中，0.503的计算结果属于比较相似。这意味着模糊余弦相似度可以有效地捕获从属关系中隐藏的信息。

2)论文标题相似度：Paper title similarity(v_i，v_j)表示节点 v_i 与 v_j 之间的论文标题相似度。利用模糊余弦相似度估计论文标题相似度的接近度。题目越相似，作者的研究方向越相似，两个相似的作者合作的可能性就越大。

3.2　仿真模型的建立

在合作作者网络 $G=\langle V,E,T\rangle$ 中，每条边 $E=\langle u,v\rangle\in E$ 表示在时刻 t，u 与 v 之间的关系。在本文实验中，关系定义为共同创作一篇论文。边 E 分为训练集 E^P 和测试集 E^T，$E=E^P\cup E^T$，$E^P\cap E^T=\varnothing$。

本文选取了 ca-AstroPh[1)]，ca-CondMat[1)]，ca-GrQc[1)]，ca-HepPh[1)] 和 ca-HepTh[1)] 5个数据集组成的联合作者网络进行实验(见表2)。

表2　作者合作网络数据集

Table 2　Co-authorship network datasets

Dataset	Type	Nodes	Edges
ca-AstroPh	Undirected	18772	198110
ca-CondMat	Undirected	23133	93492
ca-GrQc	Undirected	5242	14496
ca-HepPh	Undirected	12008	118521
ca-HepTh	Undirected	9877	25998

本文使用了从1993年到2003年共11年的数据集，其中前9年数据作为训练集，后2年数据作为测试集。

3.3　评价指标

接收者操作特征(receiver operating characteristic)以ROC曲线来评价预测性能。针对一个二分类问题，将实例分成正类(postive)或者负类(negative)。但是，在实际分类时，会出现4种情况：若一个实例是正类并且被预测为正类，即为真正类(True Postive TP)；若一个实例是正类，但是被预测成为负类，即为假负类(False Negative FN)；若一个实例是负类，但是被预测成为正类，即为假正类(False Postive FP)；若一个实例是负类，但是被预测成为负类，即为真负类(True Negative TN)。ROC曲线下方的面积(AUC)通常用来衡量链路预测的整体性能。

4　实验数据和结果分析

针对网络中节点对链路预测的重要影响，本文提出了一种新的节点重要性指标 MI，即假设两跳路径的影响显著，三跳以上的路径影响较弱。通过节点重要性 MI 强调两跳路径，对三跳或多跳路径施加弱影响，从而取得了较好的 AUC 值。AUC 值的比较如表3所列。实验结果如图4所示。

本文从两个方面对所提出的 TNTlink 模型进行分析和评价。

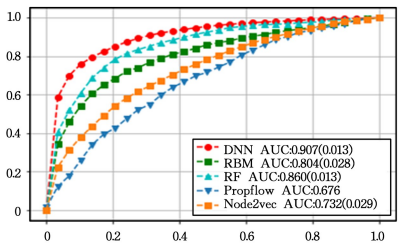
(1)表3和图4分别给出了 RBM，RF，Propflow，Node2vec 和 DNN 预测器在作者合作网络中的 ROC 曲线和 AUC 值，各数据集的最佳性能以黑体显示。从表3可以看到，在 ca-AstroPh 网络中，采用 DNN 时 *AUC* 值达到了90.7%，而采用 RBM，RF，Propflow 和 node2vec 时的 *AUC* 值分别是80.4%，86.0%，67.6%和73.2%。实验结果表明，结合拓扑特性、基本特征和附加特征，使用深度学习框架可以显著提高链路预测的性能。

(2)为了进一步证明 TNTlink 模型中的基本特征和附加特征会提高预测性能，本文从所有特性中删除一个特性，即没有拓扑模型、没有基本特征模型、没有附加特征模型，来训练3种“去掉一个特征”的预测模型，并对这些模型进行比较，以研究基本特征和附加特征对链路预测的影响。表4列出了3种“去掉一个特征”模型和 TNTlink 模型的 AUC 值的比较结果。可以发现，TNTlink 模型优于其他所有模型，这意味着去掉使用过的基本特征和附加特征都会降低其预测能力。

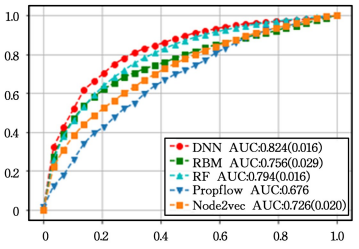
表3　*AUC* 值比较

Table 3　*AUC* value comparsion

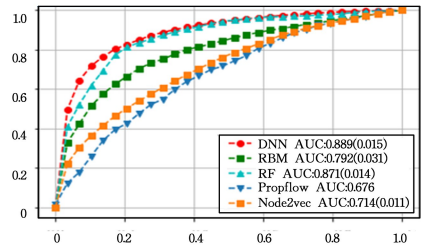
Dataset	RBM	RF	Propflow	Node2vec	DNN
ca-AstroPh(strong node)	0.804(0.028)	0.860(0.013)	0.676	0.732(0.029)	**0.907(0.013)**
ca-AstroPh(Weak node)	0.756(0.029)	0.794(0.016)	0.676	0.726(0.020)	**0.824(0.016)**
ca-CondMat(strong node)	0.792(0.031)	0.871(0.014)	0.676	0.714(0.011)	**0.889(0.015)**
ca-CondMat(Weak node)	0.764(0.013)	0.781(0.015)	0.711	0.750(0.010)	**0.801(0.011)**
ca-GrQc(strong node)	0.745(0.028)	0.791(0.019)	0.676	0.721(0.020)	**0.803(0.023)**
ca-GrQc(Weak node)	0.772(0.010)	0.784(0.014)	0.711	0.744(0.016)	**0.792(0.017)**
ca-HepPh(strong node)	0.757(0.029)	0.795(0.018)	0.676	0.726(0.020)	**0.826(0.017)**
ca-HepPh(Weaknode)	0.750(0.035)	0.795(0.017)	0.676	0.726(0.027)	**0.809(0.016)**
ca-HepTh (strong node)	0.770(0.024)	0.800(0.024)	0.676	0.716(0.022)	**0.817(0.016)**
ca-HepTh (Weak node)	0.746(0.026)	0.781(0.022)	0.669	0.740(0.019)	**0.783(0.017)**



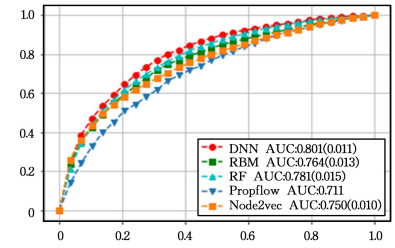
(a)ca-AstroPh 强节点的 ROC 曲线



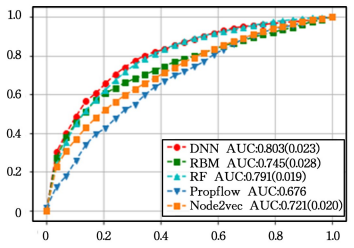
(b)ca-AstroPh 弱节点的 ROC 曲线



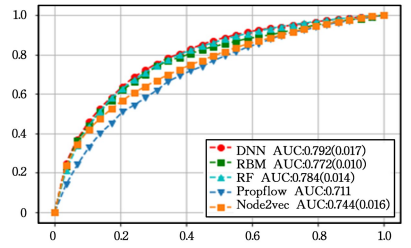
(c)ca-CondMat 强节点的 ROC 曲线



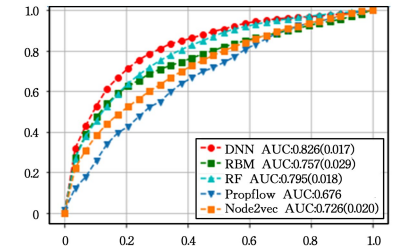
(d)ca-CondMat 弱节点的 ROC 曲线



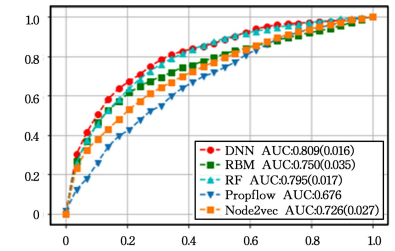
(e)ca-GrQc 强节点的 ROC 曲线



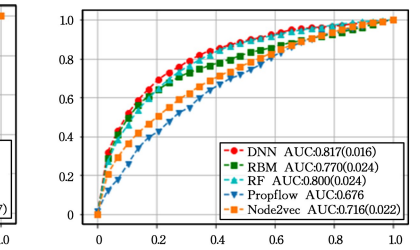
(f)ca-GrQc 弱节点的 ROC 曲线



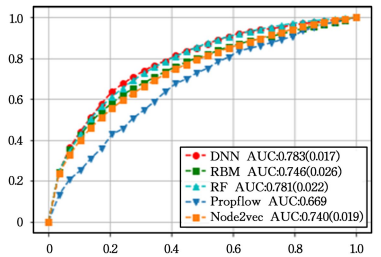
(g)ca-HepPh 强节点的 ROC



(h)ca-HepPh 弱节点的 ROC



(i)ca-HepTh 强节点的 ROC 曲线



(j)ca-HepTh 弱节点的 ROC

图 4 作者合作网的 ROC 曲线

Fig. 4 ROC curve of co-authorship networks

表 4 “去除一个特征”模型的比较

Table 4 Comparison of getting rid of a feature model

Type of Model	AUC
No Topolopy features	0.704
No basic features	0.752
No additional features	0.784
TNTlink model	0.907

结束语 本文采用深度学习的方法，通过对拓扑特征、基本特征和附加特征进行分析，建立了一种新的链路预测模型——TNTlink 模型。该模型结合二进制相似度和模糊余弦相似度来描述基本特征和附加特征。同时，通过强调强关系和惩罚弱关系，本文提出了一种新的预测指标 MI，即假设两跳路径的影响显著，而三跳以上路径的影响较弱。为了验证所提模型，在 5 个真实数据集上进行了实验，并将实验结果与 RBM，RF，Prowflow 和 Node2vec 这 4 个分类器进行比较。仿真结果表明，MI 和 TNTlink 模型能够有效地弥补传统预测方法的不足，具有更好的预测性能。虽然本文提出的 TNTlink 模型取得了不错的预测效果，但是该方法只适于静态网络，而实际的作者合作网络是动态的，下一步工作将考虑以时间为输入特征的动态网络链路预测。

参 考 文 献

[1] BLAGUS N, ŠUBELJ L, BAJEC M. Self-similar scaling of density in complex real-world networks[J]. Physica A, 2012, 391(8): 2794-2802.

[2] LICHTENWALTER R N, CHAWLA N V. Vertex collocation profiles: subgraph counting for link analysis and prediction [C]// Proceedings of the 21st World Wide Web Conference (WWW'12). ACM, 2012: 1019-1028.

[3] LI X, CHEN H. Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach[J]. Decis Support Syst, 2013, 54(3): 880-890.

[4] SCELLATO S, NOULAS A, MASCOLO C. Exploiting place features in link prediction on location-based social networks

[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, 2011:1046-1054.

[5] WANG P,XU B,WU Y,et al. Link prediction in socialnetworks: the state-of-the-art [J]. Science China Information Sciences,2015,45(9):1-38.

[6] PAVLOV M,ICHISE R. Finding experts by link prediction in co-authorship networks [C] // Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics (FEWS). Busan,2007:42-55.

[7] ICHISE R,WOHLFARTH T. Semantic and event-based approach for link prediction[C]// Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management (PAKM'08). Yokohama,2008:50-61.

[8] HASAN M I,CHAOJI V,SALEM S,et al. Link prediction using supervised learning [J]. Counterterrorismand Security, 2006,10(6):121-136.

[9] WANG J,RONG L L. Similarity index based on the information of neighbor nodes for link prediction of complex network[J]. Modern physics letters B,2013,27(6):1350039-1350049.

[10] ADITY G,LESKOVEC J. node2vec:Scalable feature learning for networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2016:855-864.

[11] FENG L,LIU B Q,SUN C J,et al. Deep belief network based approaches for link prediction in signed social networks[J]. Entropy,2015,17(4):2140-2169.

[12] YANG X H,YU J,ZHANG D. Link prediction method based on local community and nodes' relativity[J]. Computer Science, 2019,46(1):155-161.

[13] BARABSI A L,JEONG H,NEDA Z,et al. Evolution of the social network of scientific collaborations[J]. Physica A, 2002, 311(7):590-614.

[14] ZHANG C,OSMAR R,ZAIAN E. Neighbor-based link prediction with edge uncertainty[J]. Advances in Knowledge Discovery and Data,2019,36(12):462-474.

[15] YANG X H,YANG X H,LING F. Link prediction based on local major path degree [J]. Modern Physics Letter B, 2018, 32(1):29-35.

[16] GUNAWAN D,SEMBIRING C A,BUDIMAN M A. The implementation of cosine similarity to calculate text relevance between two documents[J]. Journal of Physics Conference Series, 2018,978(1):1-7.

[17] LE C Y,BENGIO Y,HINTON G. Deep learning[J]. Nature, 2015,7553(521):436-444.

[18] CHAN W,KE N R,LANE L. Transferring knowledge from a RNN to a DNN[J]. Interspeech,2015,10(6):3264-3268.

[19] LUO D S,WANG Y,HAN X Q. A cyclic contrastive divergence learning algorithm for high-order RBMS[J]. IEEE,2014,18(10): 3-6.

[20] AOUAY S,JAMOUSSI S,GARGOURI F,et al. Feature based link prediction[C]// 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications. IEEE, 2014: 10-13.

[21] THI D B,ICHISE R,LE B. Link Prediction in Social Networks Based on Local Weighted Paths[J]. Future Data and Security Engineering,2014,21(19):151-163.

[22] DONG Y X,TANG J,WU S. Link prediction and recommendation across heterogeneous social networks[C]// IEEE International Conference on Data Mining. IEEE Computer Society, 2012:181-190.

[23] NOWELL D L,KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology,2007,58(7):1019-1031.

[24] ZENG S. Link prediction based on local information considering preferential attachment[J]. Physica A,2016,443(2):537-542.

[25] LICHTENWALTER R N,LUSSIER J T,CHAWLA N V. New perspectives and methods in link prediction[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2010:243-252.

[26] LU L,ZHOU T. Link prediction in complex networks:A survey [J]. Phys. A,2011,28(6):1150-1170.

[27] ZHANG J. Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction [J]. Information Proc, 2017,53(1):42-51.

[28] ZHU Y X,LU L Y,ZHANG Q M,et al. Uncovering missing links with cold ends[J]. Physica A,2012,369(5):57-69.

[29] WU S,SUN J,TANG J. Patent partner recommendation in enterprise social networks[C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. ACM, 2013:43-52.

[30] WANG H,LE Z C,GONG X,et al. Link predicton of complex network is analyzed from the perspective of informatics [J]. Journal of Chinese Computer Systems,2020,41(2):316-326.

WANG Hui, born in 1983, Ph.D student, lecturer, is a member of China Computer Federation. Her main research interests include link prediction, deep learning, AI and big data.