云南民族大学学报(自然科学版),2024,33(3):359 - 367　　CN 53 - 1192/N　ISSN 1672 - 8513
doi:10.3969/j.issn.1672 - 8513.2024.03.011　　http://ynmz.cbpt.cnki.net

融合结构和聚类的对称非负矩阵分解链路预测

陈广福[1,2],陈　浩[3]
(1.武夷学院 数学与计算机学院,福建 武夷山 353400; 2.认知计算与智能信息处理福建省高校重点实验室,福建 武夷山 353400; 3.江苏大学 京江学院电子信息工程学院,江苏 镇江 212013)

摘要: 大部分链路预测算法仅单一考虑节点聚类或链接聚类而忽略网络结构与聚类内在关联性导致预测准确度下降.针对此问题,提出基于对称非负矩阵分解(SNMF)链路预测框架融合多类型结构和聚类信息捕获网络保持网络局部、全局以及节点和链接聚类.首先,融合节点和链接聚类系数(NEC)捕获节点邻域相关联程度,再将无向无权3个基于局部相似度方法共同邻居(CN)、资源分配(RA)和Adamic - Adar(AA)与聚类相融合同时保持结构和聚类;其次,将邻接矩阵映射到低维潜在空间,利用图正则化融合以上信息分别提出3个链路预测模型即SNMF - NEC - CN、SNMF - NEC - AA和SNMF - NEC - RA;此外,通过迭代更新规则学习所提模型参数,获得最优预测概率矩阵.在6个网络上与现有代表性方法比较,实验结果显示所提模型AUC和F1值分别提高了22%和11.4%.

关键词: 链路预测;对称非负矩阵分解;局部结构;节点和链接聚类
中图分类号:TP391　　**文献标志码:**A　　**文章编号:**1672 - 8513(2024)03 - 0359 - 09

当前,链路预测在不同领域中具有很大应用价值[1],例如预测药物 - 标靶交互[2];预测全球恐怖组织网络中各联盟关系[3]等.研究人员在不同学科提出不同链路预测算法可大致归纳为:基于结构相似度及降维模型.前者利用现存网络结构知识去计算未连接节点间预测分数,其中局部相似度最为简洁和预测效果良好,这类算法最典型代表包含:共同邻居(common neighbors,CN)[4]、资源分配(resource allocation,RA)[5]以及AA指标[6].然而,以上3个基于局部方法仅考虑局部结构,当网络十分稀疏时,上述方法获得低预测准确度.为弥补不足,一些研究人员将基于局部相似度与3阶路径、协同过滤框架和非负矩阵分解模型相融合.例如Muscoloni等[7]提出基于3阶路径(length three,L3)预测框架,该框架尽可能包含所有节点3阶路径信息并在蛋白质网和食物链网上预测准确度优于2阶路径相似度;Lee等[8]在文献[2]基础上融合局部相似度CN、AA和RA与3阶路径融合分别提出3阶路径共同邻居(common neighbors length three,CN - L3)、3阶路径Adamic - Adar(adamic - adar length three,AA - L3)和3阶路径资源分配(Resource Allocation Length three,RA - L3);Lee等[9]在协同过滤和自包含协同过滤框架融合局部CN、AA和RA,实验表明性能显著优于局部相似度;此外,Wang等[10]在非负矩阵分解预测模型融合网络内部和外部的辅助信息如局部相似度CN、AA和RA鲁棒稀疏网络和识别噪音问题.尽管以上方法在一定程度上鲁棒稀疏网,但在极度稀疏网络预测准确度较差.而本文基于非负矩阵分解(nonnegative matrix factorization,NMF)预测模型是基于降维模型典型方法,其核心是在NMF基础上附加各类息信提高预测精度.例如Chen等[13]提出在非负矩阵分解基础上融合节点属性去探索网络结构与节点属性关联性;Chen等[14]提出鲁棒非负矩阵分解融合网络局部和全局结构去讨论附加信息是否有助于预测精度提升;陈广福等[15]提出对称非负矩阵分解融合节点聚类信息并适用于无向和加权网络尽管上述基于NMF链路预测获得较好预测准确度,不足之处仅考虑网络局部结构信息.

本文解决以下2个问题:1)如何融合聚类信息与网络结构;2)如何在低维潜在空间同时保持网络局部和聚类信息.首先计算节点与链接聚类系数并聚合成节点和链接聚类(node and edge clustering,NEC)分数矩阵;其次,启用可调参数将3个经典基于局部相似度(CN、AA和RA)与节点和链接聚类分数矩阵相融合保持

收稿日期:2022 - 04 - 05.
基金项目:福建省自然科学基金(2021J011146);武夷学院引进人才科研启动基金(YJ202017).
作者简介:陈广福(1979 -),男,博士,讲师.主要从事链路预测和网络表示研究.

结构和聚类信息;最后,利用图正则化方法将上述信息与特征因子矩阵映射到低维潜在空间提出3个融合结构和聚类的对称非负矩阵分解链路预测模型即SNMF-NEC-CN、SNMF-NEC-AA和SNMF-NEC-RA.在6个无向无权网络上,实验结果表明本文3个模型预测精度显著优于现存代表性指标.

1 方法

1.1 局部相似度与聚类信息

网络结构是演绎网络演化过程的重要信息.同时,不同类型的网络结构信息为链路预测提供充足理论支持,常见基于局部结构相似度包含:CN、RA及AA指标,具体定义如表1所示.

表1 无向无权基于局部相似度

指标名称	公式
Common Neighbor(CN)	$s_{xy}^{\mathrm{CN}}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$
Adamic-Adar(AA)	$s_{xy}^{\mathrm{AA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\lg(k_z)}$
Resource Allocation(RA)	$s_{xy}^{\mathrm{RA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k_z}$

其中$\Gamma(x)$表示节点x的邻居数.

以上3个指标可有效捕获网络局部结构信息,然而真实网络是十分稀疏的,如电力网和层次路由器网等.为改善稀疏网络的预测精度,本文融合节点与链接聚类系数捕获整个网络节点链接和局部节点聚类信息.链接聚类系数(edge clustering coefficient,ECC)衡量任意节点两个端点与其邻居节点关联程度,其定义如式(1)所示.

$$\mathrm{ECC}(x,y)=\frac{\mathrm{CN}(x,y)}{\min(k_x-1,k_y-1)}. \tag{1}$$

其中$\mathrm{CN}(x,y)$表示节点x和y共同邻居数,k_x表示节点x度.

节点聚类系数(node clustering coefficient,NCC)衡量网络任意节点与该节点邻域关联程度,其定义如式(2)所示.

$$\mathrm{NCC}(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{2t_z}{k_z(k_z-1)}. \tag{2}$$

其中,$\Gamma(x)$表示节点x邻居数.

融合式(1)和(2)构建节点与链接聚类指标(node and edge clustering,NEC)衡量网络节点与链接紧密程度,定义如式(3)所示.

$$S_x^{\mathrm{NEC}}=\sum_{y\in\Gamma(x)}\mathrm{ECC}(x,y)\mathrm{NCC}(x,y). \tag{3}$$

1.2 链路预测模型

对称非负矩阵分解(symmetric nonnegative matrix factorization,SNMF)通过1阶相似度更有效捕获网络节点聚类信息[16].本文用$\boldsymbol{A}$表示网络邻接矩阵,SNMF目标将$\boldsymbol{A}$映射到低维潜在空间,目标损失函数:

$$\min_{\boldsymbol{W}\geqslant 0}\boldsymbol{J}_{\mathrm{SNMF}}=\frac{1}{2}\lVert\boldsymbol{A}-\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\rVert_F^2. \tag{4}$$

其中$\boldsymbol{W}\in\mathbf{R}_+^{n\times K}$表示基本矩阵,$\boldsymbol{K}$表示潜在空间维数.

然而式(4)仅考虑网络原始链接信息无法获取更多有效网络结构信息例如节点聚类或多类型局部结构等.为融合更多结构信息,将表1所列出3个经典局部结构与式(3)相融合共同保持网络结构与聚类信息,定义如下.

$$S_{xy}^{\mathrm{NEC-CN}}=\beta S_{xy}^{\mathrm{CN}}+S_x^{\mathrm{NEC}}. \tag{5}$$

$$S_{xy}^{\mathrm{NEC-AA}}=\beta S_{xy}^{\mathrm{AA}}+S_x^{\mathrm{NEC}}. \tag{6}$$

$$S_{xy}^{\mathrm{NEC-RA}}=\beta S_{xy}^{\mathrm{RA}}+S_x^{\mathrm{NEC}}. \tag{7}$$

其中,可调参数β作用是控制结构信息贡献.

再使用正则化技术将式(5)、(6)和(7)去保持更多类型的结构信息适用不同类型网络,定义如下.

$$o_1=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{w}_i-\boldsymbol{w}_j|^2\boldsymbol{S}_{ij}^{\mathrm{NEC-CN}}=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{NEC-CN}}\boldsymbol{W})-\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{NEC-CN}}\boldsymbol{W})=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{NEC-CN}}\boldsymbol{W}). \tag{8}$$

$$o_2=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{w}_i-\boldsymbol{w}_j|^2\boldsymbol{S}_{ij}^{\mathrm{NEC-AA}}=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{NEC-AA}}\boldsymbol{W})-\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{NEC-AA}}\boldsymbol{W})=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{NEC-AA}}\boldsymbol{W}). \tag{9}$$

$$o_3=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{w}_i-\boldsymbol{w}_j|^2\boldsymbol{S}_{ij}^{\mathrm{NEC-RA}}=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{NEC-RA}}\boldsymbol{W})-\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{NEC-RA}}\boldsymbol{W})=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{NEC-RA}}\boldsymbol{W}). \tag{10}$$

其中,$Tr(\cdot)$表示矩阵的迹,$\boldsymbol{D}^{\mathrm{NEC-CN}}$、$\boldsymbol{D}^{\mathrm{NEC-AA}}$及$\boldsymbol{D}^{\mathrm{NEC-RA}}$是对角阵,$\boldsymbol{L}_{\mathrm{NEC-CN}}=\boldsymbol{D}^{\mathrm{NEC-CN}}-\boldsymbol{S}^{\mathrm{NEC-CN}}$、$\boldsymbol{L}_{\mathrm{NEC-AA}}=\boldsymbol{D}^{\mathrm{NEC-AA}}-\boldsymbol{S}^{\mathrm{NEC-AA}}$及$\boldsymbol{L}_{\mathrm{NEC-RA}}=\boldsymbol{D}^{\mathrm{NEC-RA}}-\boldsymbol{S}^{\mathrm{NEC-RA}}$表示拉普拉斯矩阵.

通过融合式(4)、(8)、(9)和(10)共同构建统一 3 个链路预测模型,其 3 个目标损失函数为:

$$\min_{\boldsymbol{W}\geqslant 0}J_{\mathrm{SNMF-NEC-CN}}=\frac{1}{2}\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2+\alpha o_1. \tag{11}$$

$$\min_{\boldsymbol{W}\geqslant 0}J_{\mathrm{SNMF-NEC-AA}}=\frac{1}{2}\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2+\alpha o_2. \tag{12}$$

$$\min_{\boldsymbol{W}\geqslant 0}J_{\mathrm{SNMF-NEC-RA}}=\frac{1}{2}\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2+\alpha o_3. \tag{13}$$

为方便与统一学习所提 3 个目标损失函数参数,将上述 3 个损失函数归纳为一个统一的目标损失函数,有:

$$\min_{\boldsymbol{W}\geqslant 0}J=\frac{1}{2}\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{W}). \tag{14}$$

为学习所提模型参数,本文采用拉格朗日乘法法则求解最优解. 由矩阵迹性质有:$\mathrm{Tr}(\boldsymbol{AB})=\mathrm{Tr}(\boldsymbol{BA})$,再重写式(14),有:

$$J(\boldsymbol{W})=\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-2\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}})+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{W}). \tag{15}$$

引入拉格朗日乘子矩阵 $\Phi=[\varphi]_{\boldsymbol{nk}}$,再重写式(15),有:

$$J(\boldsymbol{W})=\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-2\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}})+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{W})+\mathrm{Tr}(\Phi\boldsymbol{W}^{\mathrm{T}}). \tag{16}$$

删除式(16)中与 $\boldsymbol{W}$ 无关项,有:

$$J(\boldsymbol{W})=\mathrm{Tr}(-2\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}})+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{W})+Tr(\Phi\boldsymbol{W}^{\mathrm{T}}).$$

对 $J(\boldsymbol{W})$求偏导,有:

$$\frac{\partial J(\boldsymbol{W})}{\partial\boldsymbol{W}}=-\boldsymbol{A}\boldsymbol{W}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}+\Phi.$$

由 KKT(Karush-Kuhn-Tucker)条件且 $\varphi_{nk}\boldsymbol{w}_{nk}=0$,有:$-\boldsymbol{A}\boldsymbol{W}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}=0$.

因此,更新规则如下.

$$\boldsymbol{W}_{nk}\leftarrow\boldsymbol{W}_{nk}\left(\frac{\boldsymbol{A}\boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}}\right)_{nk}. \tag{17}$$

2　模型收敛性

本节主要讨论所提模型是否具备收敛性,证明过程如下.

定理 1　目标损失函数(14)在迭代更新规则(17)下是单调递减的.

上述若定理成立,则需要以下辅助定理.

定义:对任意的 x 和 x',若 $H(x,x')$是 $F(x)$的辅助函数,那么必须满足以下条件:$H(x,x')\geqslant F(x)$,$H(x,x)=F(x)$

引理 1　若 $H(x,x')$是 $F(x)$的辅助函数, $F(x)$在更新规则下单调递减.

$$x^{(t+1)}=\arg\min_x H(x,x^{(t)}). \tag{18}$$

证明

$$F(x^{(t+1)})\leqslant H(x^{(t+1)},x^{(t)})\leqslant H(x^{(t)},x^{(t)})=F(x^{(t)}).$$

首先,将式(14)中删除与 $\boldsymbol{W}$ 无关项,有:$F(\boldsymbol{W})=\mathrm{Tr}(-2\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\alpha\boldsymbol{W}^{\mathrm{T}}LW)$.

求 $F(\boldsymbol{W})$对 $\boldsymbol{W}$ 的 1 次和 2 次偏导数:

$$F'(\boldsymbol{W})=\left[\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{W}}\right]_{nk}=[-2\boldsymbol{A}\boldsymbol{W}+2\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+2\alpha\boldsymbol{L}\boldsymbol{W}]_{nk}\cong[-\boldsymbol{A}\boldsymbol{W}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}]_{nk}.$$

$$F''(\boldsymbol{W})=\left[\frac{\partial^2 \boldsymbol{F}}{\partial \boldsymbol{W}^2}\right]_{nk}=[-2\boldsymbol{A}+2\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+2\alpha\boldsymbol{L}]_{nn}\cong[-\boldsymbol{A}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\alpha\boldsymbol{L}]_{nn}.$$

引理 2 $H(\boldsymbol{W}_{nk},\boldsymbol{W}_{nk}^{(t)})$是$F_{nk}$的辅助函数,有:

$$\mathrm{H}(\boldsymbol{W}_{nk},\boldsymbol{W}_{nk}^{(t)})=F_{nk}(\boldsymbol{W}_{nk}^{(t)})+F'_{nk}(\boldsymbol{W}_{nk}^{(t)})(\boldsymbol{W}_{nk}-\boldsymbol{W}_{nk}^{(t)})+\frac{[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{D}\boldsymbol{W}]_{nk}}{\boldsymbol{W}_{nk}^{(t)}}(\boldsymbol{W}_{nk}-\boldsymbol{W}_{nk}^{(t)})^2. \tag{19}$$

证明 $F_{nk}(\boldsymbol{W}_{nk})$泰勒展开式为:

$$F_{nk}(\boldsymbol{W}_{nk})=F_{nk}(\boldsymbol{W}_{nk}^{(t)})+F'_{nk}(\boldsymbol{W}_{nk}^{(t)})(\boldsymbol{W}_{nk}-\boldsymbol{W}_{nk}^{(t)})+[-\boldsymbol{A}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\alpha\boldsymbol{L}]_{nk}(\boldsymbol{W}_{nk}-\boldsymbol{W}_{nk}^{(t)})^2. \tag{20}$$

又因$H(W_{nk},W_{nk}^{(t)})>F_{nk}(W_{nk})$,式(19)与(20)比较有:

$$\frac{[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}]_{nk}}{\boldsymbol{W}_{nk}^{(t)}}\geqslant[-\boldsymbol{A}+\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\alpha\boldsymbol{L}]_{nk}. \tag{21}$$

又因$[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}]_{nk}=\sum_{l=1}^{n}\boldsymbol{W}_{nl}^{(t)}[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}]_{lk}\geqslant\boldsymbol{W}_{nk}^{(t)}[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}]_{nn}$和$\alpha\boldsymbol{L}[\boldsymbol{W}]_{nk}\geqslant(\alpha\boldsymbol{L}-\boldsymbol{A})\boldsymbol{W}_{nk}^{(t)}$.

综上所述,式(21)成立$H(W_{nk},W_{nk}^{(t)})>F_{nk}(W_{nk})$得证. 根据式(21)可证定理1成立.

证明 $x^{(t+1)}=\arg\min_x H(x,x^{(t)})\Rightarrow F'_{nk}(\boldsymbol{W}_{nk}^{(t)})+2\frac{[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{D}\boldsymbol{W}]_{nk}}{\boldsymbol{W}_{nk}^{(t)}}(\boldsymbol{W}_{nk}-\boldsymbol{W}_{nk}^{(t)})=0.$

$$\boldsymbol{W}_{nk}^{(t+1)}=\boldsymbol{W}_{nk}^{(t)}-\boldsymbol{W}_{nk}^{(t)}\frac{F'_{nk}(\boldsymbol{W}_{nk}^{(t)})}{2[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{D}\boldsymbol{W}]_{nk}}=\boldsymbol{W}_{nk}^{(t)}\left(\frac{\boldsymbol{A}\boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\alpha\boldsymbol{L}\boldsymbol{W}}\right)_{nk}.$$

定理1得证.

3 实验结果与分析

3.1 评价度量

使用AUC[17]和F1[18]度量衡量所提模型性能,其中F1是召回率与精度的综合性指标.

a) AUC(areas under ROC),测试集E^{P}中相似分数大与不存在集中链接分数的概率进行比较,若独立地比较n次,若有n_1次测试集中的链接的分数值大于不存在集中的链接的分数,有n_2次两分数值相等,AUC定义:

$$\mathrm{AUC}=\frac{n_1+0.5\times n_2}{n}$$

其中,n表示比较总数.

b) F1度量是召回率(Recall)和准确率(Precision)调和平均值,其定义如下.

$$\mathrm{F1}=\frac{2\times\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}.$$

3.2 数据集

本文采用6真实世无向无权网络:a)蛋白质交互网络(YEAst,YEA)[19],b) NIPS[19],c) 国际电子道路网络(euroroads,ER)[19],d) 空中交通管制网(air traffic control,ATC)[19],e)电力网(powergrid,PG)[19],f) Router是路由层次网[19],其6个网络特征结构统计如表1.

表1 6个真实世界无向网络拓扑特征统计

Network	$\lvert V\rvert$	$\lvert E\rvert$	$<k>$	C	Density
YEA	2 361	7 182	5.629 8	0.130 1	0.002 4
Router	5 022	6 258	2.490 0	0.033 0	0.000 5
NIPS	2 888	2 981	2.064 4	0.000 3	0.000 7
ER	1 174	1 417	2.413 9	0.016 7	0.002 1
ATC	1 226	2 615	2.131 3	0.057 6	0.001 7
PG	4 941	6 594	2.669 1	0.080 1	0.000 5

其中,$\lvert V\rvert$表示节点数量,$\lvert E\rvert$是边总数,$<k>$是平均度,C是节点聚类系数,Density表示网络稠密程度.

3.3　基准方法

a) 局部结构相似度算法 CN、AA 和 RA[4-6],定义如下.

$$S_{ij}^{\mathrm{CN}}=|\Gamma(i)\cap\Gamma(j)|,S_{ij}^{\mathrm{AA}}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\log k_z},S_{ij}^{\mathrm{RA}}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{k_z}.$$

b) 3 个基于 3 阶路径相似度(CN - L3、AA - L3 和 RA - L3)[8],定义如下.

$$S_{ij}^{\mathrm{CN-L3}}=\sum_{x\in\Gamma_i,y\in\Gamma_j}a_{xy},S_{ij}^{\mathrm{AA-L3}}=\sum_{x\in\Gamma_i,y\in\Gamma_j}\frac{a_{xy}}{\sqrt{\log(k_x)\log(k_y)}},S_{ij}^{\mathrm{RA-L3}}=\sum_{x\in\Gamma_i,y\in\Gamma_j}\frac{a_{xy}}{\sqrt{k_xk_y}}.$$

c) SCF - CN、SCF - AA 和 SCF - RA 预测指标[9],其定义如下:

$$S^{\mathrm{SCF}}=(\boldsymbol{A}+\boldsymbol{I})S+[(\boldsymbol{A}+\boldsymbol{I})S]^{\mathrm{T}}.$$

其中,A 是网络邻接矩阵,S 分别是 ***CN***、***AA*** 和 ***RA***.

d) 节点和链接聚类指标(node and link clustering,NLC)[20]该算法将节点和边聚类相组合,其定义如下.

$$S_{xy}^{\mathrm{NLC}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{\mathrm{CN}(x,z)}{k_z-1}\times C_z+\frac{\mathrm{CN}(y,z)}{k_z-1}\times C_z,\text{其中 }C_z=\frac{2t_z}{k_z(k_z-1)}.$$

e) 融合多类型结构以及节点属性相似度的非负矩阵分解链路预测模型(non - negative matrix factorization via structure and attribute similarity,SASNMF)[10]该框架是基于非负矩阵分解基础上融合 CN、AA 和 RA 分别构成 CN - SASNMF、AA - SASNMF 和 RA - SASNMF 模型同时保持网络局部和全局结构.

3.4　结果分析

本文所提模型涉及到的参数设置如表 2 所示.

表 2　所提模型主要参数设置

参数	YEA	NIPS	ER	ATC	Router	PG
α	0.1	0.1	0.1	0.1	0.1	0.1
β	0.5	0.5	35	0.5	0.5	20
K	70	70	70	70	70	70
迭代次数	70	70	70	70	70	70

从以下 3 个方面评价所提模型性能:a)采用 AUC 和 F1 度量全面评估所提模型性能;b)采用消融法评估所提模型增加附加聚类和结构信息是否有助于提高预测精度;c)评估所提模型鲁棒性.

第一个实验所提模型与现存代表性算法进行对比,实验结果报告表 3 中,可观察到以下 2 个现象:

a)NLC 指标在 6 个稀疏网络中均获得较低预测精度例如在 NIPS 网络中 F1 值为 0,其主要原因是该指标仅考虑局部节点和链接聚类,当网络十分稀疏时节点间共同节点和链接数量显著不足,导致预测精度下降. CN、AA 和 RA 3 个指标均考虑节点共同邻居数捕获网络局部结构,与 NLC 相比,性能获得明显提高. CN - L3、AA - L3 和 RA - L3 考虑 3 阶路径信息捕获网络更多节点信息,与 CN、AA 和 RA 3 个指标相比,在大部分网络中性能有所提高. 而 SCF - CN、SCF - AA 和 SCF - RA 3 个指标基于自包含协同过滤框架下保持局部结构,利用对称性更有效获得网络结构,与上述指标相比,性能有显著提高.

b) CN - SASNMF、AA - SASNMF 和 RA - SASNMF 与所提模型都是基于非负矩阵分解框架中融合多源结构,不同是前者仅附加网络局部结构而后者同时保持局部、全局链接聚类和局部节点聚类,从表 3 观察到所提模型在不同度量中性能均优于前者. 例如 AUC 度量,在 YEA、NIPS、ER、ATC、PG 和 Router 上,SNMF - NEC - CN 与 CN - SASNMF 相比,AUC 值分别提高了 5.4%、50.1%、22%、7.3%、27.6% 和 12.9%;F1 度量,在 YEA、NIPS、ER、ATC、PG 和 Router 上,SNMF - CN - KN 与 CN - SASNMF 相比,F1 值分别提高了 1.8%、21.7%、2.4%、2.4%、6.6% 和 5%. 表明所提模型融合局部、全局链接聚类和局部节点聚类显著改善稀疏网络预测准确度.

表 3　基准方法与所提模型在 6 个网络上 AUC 和 F1 值

指标	度量	YEA	NIPS	ER	ATC	PG	Router
CN	F1	0.194	0.141	0.077	0.264	0.123	0.107
	AUC	0.735	0.483	0.539	0.713	0.626	0.651

续表 3

指标	度量	YEA	NIPS	ER	ATC	PG	Router
AA	F1	0.353	0.169	0.072	0.377	0.186	0.204
	AUC	0.732	0.503	0.539	0.732	0.626	0.648
RA	F1	0.323	0.527	0.087	0.387	0.087	0.383
	AUC	0.883	0.773	0.557	0.836	0.632	0.934
CN – L3	F1	0.605	0.018	0.082	0.545	0.156	0.599
	AUC	0.855	0.511	0.541	0.821	0.595	0.906
AA – L3	F1	0.599	0.014	0.068	0.549	0.156	0.593
	AUC	0.852	0.509	0.532	0.825	0.592	0.899
RA – L3	F1	0.633	0.624	0.106	0.532	0.199	0.609
	AUC	0.888	0.859	0.553	0.823	0.631	0.929
SCF – CN	F1	0.340	0.809	0.180	0.459	0.166	0.410
	AUC	0.890	0.818	0.584	0.871	0.696	0.947
SCF – AA	F1	0.669	0.747	0.179	0.619	0.270	0.649
	AUC	0.892	0.880	0.587	0.877	0.693	0.946
SCF – RA	F1	0.673	0.781	0.167	0.600	0.284	0.656
	AUC	0.897	0.918	0.586	0.868	0.697	0.948
NLC	F1	0.352	0.000	0.002	0.177	0.076	0.184
	AUC	0.700	0.500	0.501	0.596	0.542	0.609
CN – SASNMF	F1	0.840	0.634	0.740	0.851	0.761	0.838
	AUC	0.861	0.317	0.585	0.876	0.633	0.843
AA – SASNMF	F1	0.844	0.563	0.735	0.851	0.857	0.763
	AUC	0.884	0.146	0.587	0.877	0.896	0.647
RA – SASNMF	F1	0.852	0.833	0.731	0.845	0.763	0.867
	AUC	0.896	0.819	0.560	0.864	0.640	0.920
SNMF – NEC – CN	F1	0.858	0.851	**0.854**	**0.875**	0.873	0.889
	AUC	0.914	0.818	**0.899**	**0.949**	0.944	0.975
SNMF – NEC – AA	F1	**0.858**	0.835	0.848	0.874	**0.869**	0.882
	AUC	**0.918**	0.795	0.879	0.942	**0.945**	**0.977**
SNMF – NEC – RA	F1	0.854	**0.853**	0.817	0.870	0.840	0.875
	AUC	0.906	0.818	0.853	0.935	0.927	0.977

其次,评估所提模型鲁棒性,通过改变划分原始网络比率改变网络稀疏性. 设划分比率为 30%、40%、…,90%,由于空间有限本实验仅选取 CN、CN – L3、SCFCN、CN – SASNMF 和 SNMF – NEC – CN 分析它们鲁棒性,在图 1 可观察到,SNMF – CN – KN 获得最优性能. 当训练集仅占 30% 时,CN、CN – L3、SCFCN 和 CN – SASNM 指标预测准确度显著下降主要原因是以上方法无法捕获更多网络结构信息鲁棒网络稀疏性,随机训练集比率增加 AUC 和 F1 值也开始逐渐上升. 然而,所提模型 SNMF – NEC – CN 在训练集 30% 时依然获得最优性能表明该模型鲁棒于稀疏网络.

4 参数敏感性分析

4.1 参数 α 变化

实验结果如图 2,当 $\alpha=0$ 时,所提模型退化为仅保持一阶、局部结构和聚类 3 类信息,此时预测准确度较低主要原因是所提模型在稀疏网络无法捕获更多结构信息. 当 α 逐渐上升,AUC 和 F1 值开始上升,直到当 $\alpha=0.1$ 时 AUC 和 F1 值最优. 当 $\alpha>0.1$ 时,AUC 和 F1 值又逐渐下降,主要原因是 α 增加导致所提模型损失函数误差增大. 因此,当 $\alpha=0.1$ 时, AUC 和 F1 值最优.

4.2 参数 β 变化

参数 β 是控制局部结构对所提模型性能影响,实验结果报告在图 3. 当 $\beta=0$ 时,所提模型退化为仅保持

一阶、聚类和全局结构信息,当β值逐渐增大,AUC 和 F1 值逐渐开始上升,对 YEA、NIPS、ATC 和 Router 网,β=0.5 时最优,对 ER 和 PG 网β分别为 20 和 35 时最优,主要原因是不同网络结构具有不同作用.

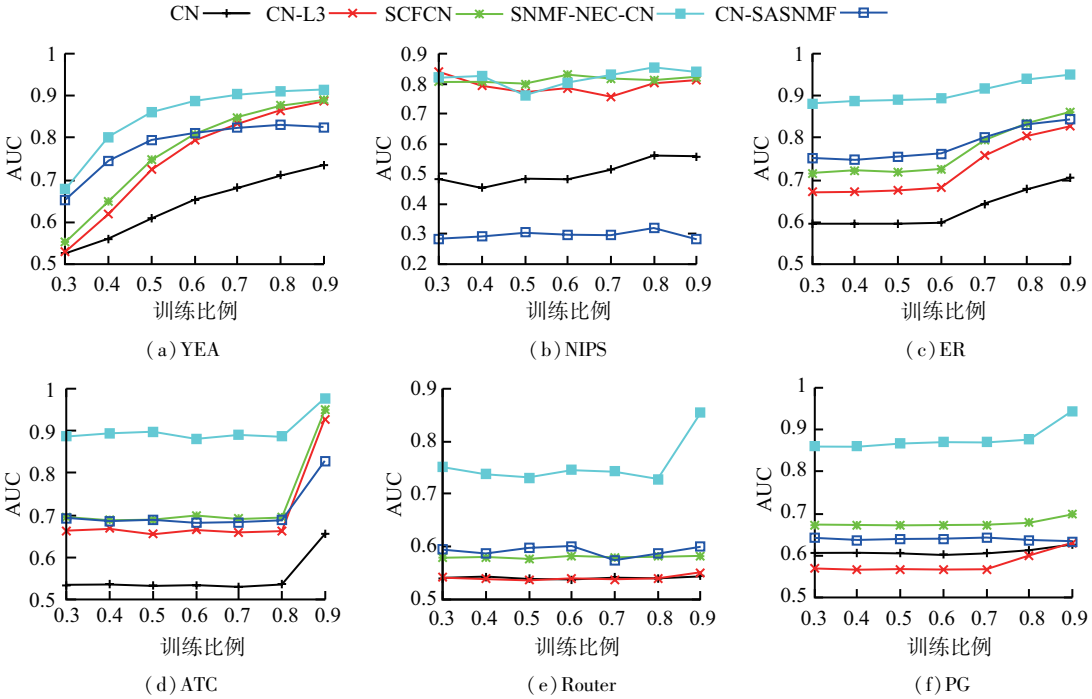

图 1　不同训练集下对应 AUC 变化

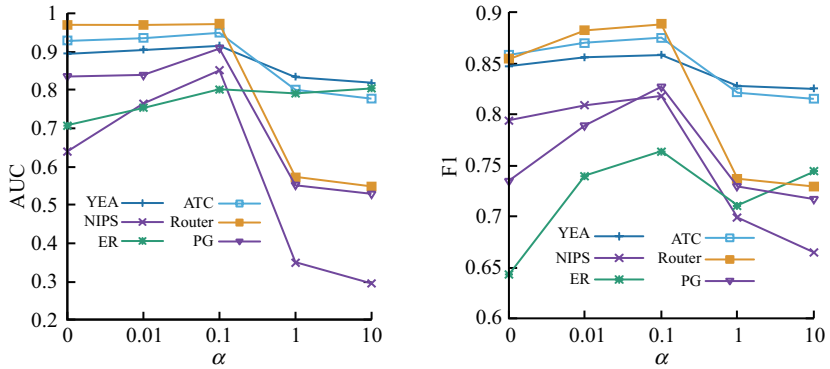

图 2　参数 α 对应不同 AUC 和 F1 值

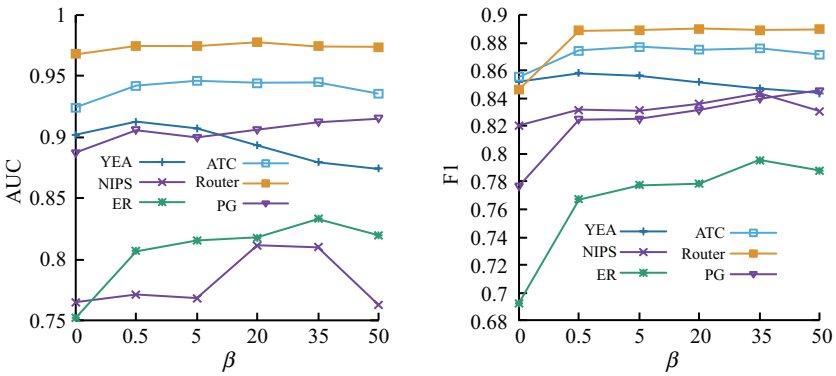

图 3　参数 β 对应不同 AUC 和 F1 值

4.3　参数 K 变化

参数 K 直接影响预测准确度,实验结果报告图 4,当 K=10 到 60 之间,AUC 和 F1 值逐渐增大. 当 K=10 时,AUC 和 F1 值最小主要原因是充分利用网络结构信息. 当 $K\geqslant$70 时,AUC 和 F1 值开始保持恒定,因此,当 K=70 时,预测准确度最优.

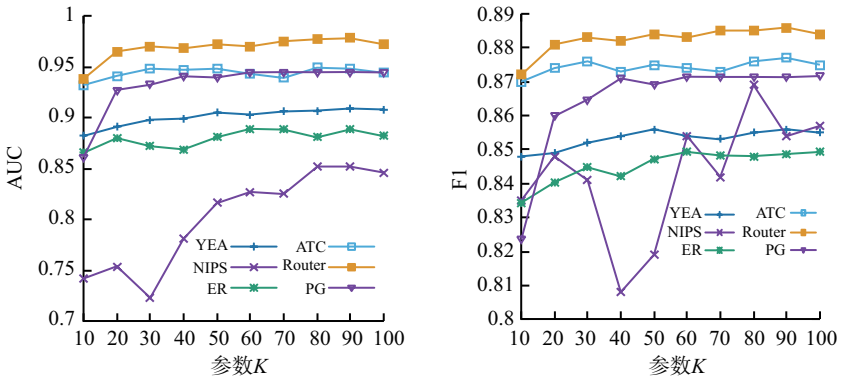

图 4　不同 *K* 在 6 个网络上对应不同 AUC 和 F1 值

4.4　迭代次数变化

迭代次数影响所提目标损失函数收敛速度,迭代次数越小表示收敛越快所提模型预测准确度就越高,设迭代次数变化范围{10,20,30,…,100},实验结果报告如图 5 所示,可观察到当迭代次数为 10 时,大部分网络预测准确度均较差,表明算法开始状态无法充分捕获网络局部、全局结构和聚类信息,当迭代次数逐渐增大,AUC 和 F1 值逐渐增大直到迭代次数大于等于 70 时,性能开始稳定. 因此,迭代次数为 70 时,预测准确度最优.

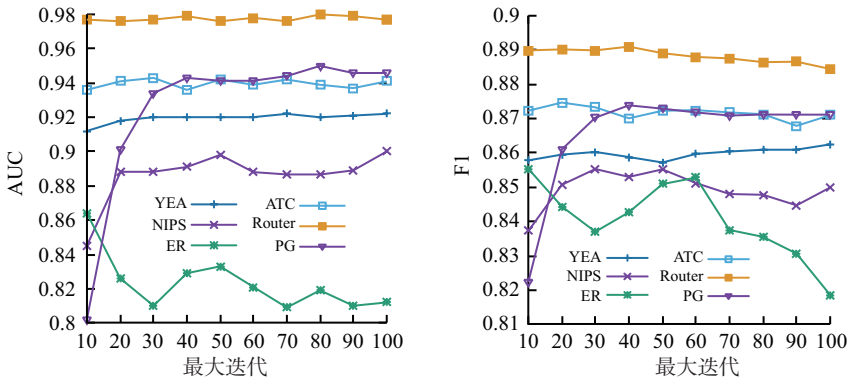

图 5　不同迭代次数在 6 个网络上对应不同 AUC 和 F1 值

5　结语

如何融合网络结构鲁棒稀疏网络是链路预测最重要的任务之一,本文提出一个融合 3 个经典局部相似度与聚类信息的非负矩阵分解链路预测模型. 该框架采用可调参数将节点和链接聚类系数与基于局部相似度相融合同时保持结构和聚类信息,然后启用图正则化将上述信息与特征因子矩阵相结合映射到低维潜空间,最后启用迭代更新规则优化所提模型. 实验表明,所提 3 个预测模型性能显著提高.

参考文献:

[1] 李艳丽,周涛. 链路预测中的局部相似性指标[J]. 电子科技大学学报. 2021,50(3):422 – 427.

[2] LU Y,GUO Y,KORHONEN A. Link prediction in drug – target interactions network using similarity indices[J]. BMC bioinformatics,2017,18(1): 1 – 9.

[3] FANG L,FANG H,TIAN Y,ET AL. The alliance relationship analysis of international terrorist organizations with link prediction [J]. Physica A: Statistical Mechanics and its Applications,2017,482: 573 – 584.

[4] LIBEN – NOWELL D,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.

[5] ADAMIC L A,ADAR E. Friends and neighbors on the web[J]. Social networks,2003,25(3): 211 – 230.

[6] ZHOU T,LÜ L,ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B,2009,71(4): 623 – 630.

[7] KOVÁCS I A,LUCK K,SPIROHN K,et al. Network – based prediction of protein interactions[J]. Nature communications,2019,10(1): 1 – 8.

[8] ZHOU T,LEE Y L,WANG G. Experimental analyses on 2 - hop - based and 3 - hop - based link prediction algorithms[J]. Physica A: Statistical Mechanics and its Applications,2021,564: 125532.
[9] LEE Y L,ZHOU T. Collaborative filtering approach to link prediction[J]. Physica A: Statistical Mechanics and its Applications,2021,578: 126107.
[10] WANG W,TANG M,JIAO P. A unified framework for link prediction based on non - negative matrix factorization with coupling multivariate information[J]. PloS one,2018,13(11): e0208185.
[11] 顾秋阳,吴宝,池仁勇. 基于高阶路径相似度的复杂网络链路预测方法[J]. 通信学报,2021,42(7):61 - 69.
[12] RAFIEE S,SALAVATI C,ABDOLLAHPOURI A. CNDP: Link prediction based on common neighbors degree penalization[J]. Physica A: Statistical Mechanics and its Applications,2020,539: 122950.
[13] Chen B,Li F,Chen S,et al. Link prediction based on non - negative matrix factorization[J]. PloS one,2017,12(8): e0182968.
[14] CHEN G,XU C,WANG J,et al. Robust non - negative matrix factorization for link prediction in complex networks using manifold regularization and sparse learning[J]. Physica A: Statistical Mechanics and its Applications,2020,539: 122882.
[15] 陈广福,王海波. 基于聚类信息和对称非负矩阵分解的链路预测模型研究[J]. 计算机应用研究,2021,38(12):3733 - 3738.
[16] HE Z,XIE S,ZDUNEK R,et al. Symmetric nonnegative matrix factorization: algorithms and applications to probabilistic clustering[J]. IEEE Transactions on Neural Networks,2011,22(12): 2117 - 2131.
[17] HANLEY J A,MCNELL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology,1982,143(1) : 29 - 36.
[18] YANG Y,LICHTENWALTER R N,CHAWLA N V. Evaluating link prediction methods[J]. Knowledge and Information Systems,2015,45(3): 751 - 782.
[19] ROSSI R A,AHMED N K. The network data repository with interactive graph analytics and visualization[C]//Proceedings of the Twenty - Ninth AAAI Conference on Artificial Intelligence. Texas,USA: AAAI Press,2015: 4292 - 4293.
[20] WU Z,LIN Y,WAN H,et al. Predicting top - L missing links with node and link clustering information in large - scale networks[J]. Journal of Statistical Mechanics: Theory and Experiment,2016,2016(8): 083202.

Link prediction based on symmetric nonnegative matrix factorization combining structure and clustering

CHEN Guang-fu[1,2],CHEN Hao[3]

(1. College of Mathematics and Computer science,Wuyi University,Wuyishan 354300,China;2. The key Laboratory of cognitive computing and intelligent information processing of Fujian education insitutions,Wuyishan 354300,China;
3. College of Electronical and Information Engineering,Jiangsu University Jingjiang college,Zhenjiang 354300,China)

Abstract: Most of the existing link prediction algorithms only consider node clustering or link clustering and ignore the internal correlation between network structure and clustering,which leads to a decrease in prediction accuracy. In view of the above shortcomings,we propose a link prediction framework based on symmetric non-negative matrix factorization (SNMF) to fuse structure and cluster information capture network to maintain network local,global and node and link clustering. Firstly,the fusion node and edge clustering coefficient (NEC) captures the degree of node neighborhood association,and then undirected and weighted three local similarity methods Common Neighbor (CN), Resource Allocation (RA) and Adamic-Adar (AA)) and clustering while maintaining structure and clustering; secondly,the adjacency matrix is mapped to a low-dimensional latent space,and the above information is fused using graph regularization to propose three link prediction models :SNMF-NEC-CN,SNMF-NEC-AA and SNMF-NEC-RA; Furthermore,the proposed model parameters are learnt by iteratively updating the rules to obtain the optimal prediction probability matrix. Comparing with the existing representative methods on six networks,the experimental results show that the AUC and F1 values of the proposed model are improved by 22% and 11.4%,respectively.

Key words: link prediction; symmetric nonnegative matrix factorization; local structure; node and edge clustering

(责任编辑　段　鹏)