基于半监督学习的链接预测算法的研究 *

杨 [1,2], 杨炳儒[1], 唐志刚[1]
(1. 北京科技大学 信息工程学院, 北京 100083; 2. 江西农业大学 计算机与信息工程学院, 南昌 330045)

摘 要: 针对链接挖掘中网络的结构难以预测这个难点问题,提出了一个关于链接预测的新型半监督学习方法——基于快速共轭梯度方法和链接相似性传递增殖原理的链接预测算法,利用节点相似性等辅助信息去预测未知结构。该算法利用张量的形式去表示多维的复杂的多关系数据,利用克罗内克积与克罗内克和去计算张量之间的相似性,利用向量特技方法降低了算法的时间和空间复杂度。在社会网络和生物信息网络等环境下,通过实验验证了算法的有效性和健壮性。

关键词: 链接预测; 张量; 共轭梯度; 克罗内克积; 克罗内克和
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2010)08-2848-05
doi:10.3969/j.issn.1001-3695.2010.08.009

Research of link prediction algorithmic based on semi-supervisor learning

YANG Jun[1,2], YANG Bing-ru[1], TANG Zhi-gang[1]
(1. *School of Information Engineering, University of Science & Technology Beijing, Beijing* 100083, *China*; 2. *College of Computer & Information Engineering, Jiangxi Agricultural University, Nanchang* 330045, *China*)

Abstract: It is very hard to forecast about structure of network in link mining. To slove the problem, this paper proposed a new semi-supervisor learning algorithmic based on an accelerated conjugate gradient method and link similarity delivery proliferation, by using auxiliary information such as node similarity to predict the unknown structure. Used the tensor to represent the multidimensional complexity multi-relation data, calculated the similarity of tensors by Kronecker product and Kronecker sum, reduced the complexity of the compute time and RAM. The effectiveness and robustness of the algorithmic was tested in social networks and biological networks.

Key words: link prediction; tensor; conjugate gradient; Kronecker product; Kronecker sum

对 Internet、科学家合作网络、人类关系网络、蛋白质相互作用网络、语言学网络等复杂网络的研究才刚起步。链接,或者可以看做是关联关系,是复杂网络的基本结构。这些链接常常又蕴涵着实例数据中节点的如重要性、等级、所在类别等信息,对网络中任意两个节点链接关系的研究引起了人们越来越多的兴趣。实际上,链接预测[1]问题在信息科学中是一项长期的挑战,在计算机科学领域已经提出了许多基于马尔可夫链和机器学习过程的算法,然而,他们的工作没有跟上当前复杂网络的研究进步,对网络的结构特征缺乏相当的考虑,网络的结构特征可以为链接预测提供信息和观察。

1 背景

链接预测问题是基于观察到的链接和节点的属性去估计两个节点间的链接存在的可能性。它们可以分成两类:a)预测样本网络中缺失的链接,如食品网和 WWW 网[2];b)预测将来可能存在在进化网中的链接,如社会网络。另外,链接预测算法(或是基于相似技术的其他算法)也能应用于解决在部分标志网络[3,4]的链接预测问题,如蛋白质功能[3]、区分科学出版研究领域问题。

直到现在,大多数算法都是根据节点的相似性来设计的。节点相似性可以通过节点的重要属性来定义,也就是两个节点假如有许多共同的特征就被看做是相似的;另一种相似性是基于网络结构的,被称为结构相似性,并且可以进一步用做节点独立、路径独立和混合方案来分类。在文献[5]中作了相似性的介绍和比较。在 common neighbors(CN)中,Jaccard coefficient[6]、Adamic-Adar index[7]和 preferential attachment[8]是基于节点进行分类的,但 Katz index[9]、hitting time[10]、commute time[11]、root PageRank[12]、SimRank[13]、Blondel index[14]是基于路径索引分类的。除此之外,Leicht 等人[15]提出了一种测量节点相似性质量的方法,是基于两个在网络中的节点,如果它们的最邻近的节点是相似的,那么它们就是相似的。这导致一个惯矩阵相似性公式,它可以用邻接矩阵进行迭代评价。除了基于相似性预测算法,近年来还提出了一些更为复杂的方法。Clauset 等人[16]提出了一种基于网络层次结构的算法:首先使用一个层次随机图去统计一个实际网络的数据;然后层次网络中节点长度的潜在概率的依赖性可以被推断出来;最后可以根据潜在链接概率通过降序排列预测网络中缺失的链接,但该算法在推荐系统的设计上花费了许多的努力。实际上推荐一个条目给用户的过程可以被看做在用户条目对分网络中预测缺失的链接[16]。对于链接预测而言还有许多的困难:一个就是

收稿日期: 2010-01-15; **修回日期**: 2010-02-22 **基金项目**: 国家自然科学基金资助项目(60675030,60875029); 江西省教育厅科学技术研究项目(GJJ10422)

作者简介:杨 (1970-),江西南昌人,副教授,博士研究生,主要研究方向为数据挖掘、知识工程、柔性建模(ycjun515@163.com);杨炳儒(1943-),天津人,教授,博导,主要研究方向为知识工程与柔性建模;唐志刚(1976-),湖南永州人,博士研究生,主要研究方向为数据挖掘.

目标网络的稀疏[17],它导致的一系列严重问题就是一个链接的先验概率相当小,对建立统计模型带来很大难度;另一个是实际的系统非常巨大需要高效的算法。然而,计算时间和空间的复杂度是影响实际应用的一个关键因素,目前还没有系统的研究。通常讲的算法的精度与计算复杂度是紧密关联的,高精度通常隐含高复杂性。假如高精度算法的时间和空间是不可接受的,那么它就没有任何意义。因此,设计一个精确快速的算法是一个巨大的挑战,特别是对大而又稀疏的网络而言具有更大的意义。

2　链接预测

链接预测问题通常描述为预测在任意两个节点对之间存在一个链接的可能程度。在这里考虑一个更为一般的问题,即预测节点对之间多种类型的链接。

在一个表示网络的图中,有许多节点对之间存在着链接,如果是有向图,可以按照链接的方向将点对分成两个集合源集合和目的集合;如果是无向图,则可以随机将链接节点分成两个集合,这里表示为 $F=\langle f_1,f_2,\cdots,f_m\rangle$ 和 $T=\langle t_1,t_2,\cdots,t_n\rangle$。点对之间的链接类型为 $L=\langle l_1,l_2,\cdots,l_p\rangle$,$M=|F|$,$N=|T|$,$P=|L|$。例如,在一个参考文献网中,$s_i$可以是作者、引用者,$d_j$可以是原文、被引用文,$l_k$可以是撰写、引用链接。$F$ 和 T 是源节点和目的节点集合,这里假设 $M=N$。f_i、t_j和 l_k可以看成一个三元组$\langle s_i,d_j,l_k\rangle$,$l_k$是两个节点之间的一个单链接,节点间多连接的预测可以通过单链接预测实现。本文用 $M\times N\times P$ 三阶张量来表示节点与链接类型之间的关系。

$$[E]_{i,j,k}=e_{i,j,k} \tag{1}$$

其中:变量 $e_{i,j,k}$表示一个三元组$\langle s_i,d_j,l_k\rangle\in F\times T\times L$ 链接强度,变量的值越大,链接存在的可能性越大;值越小,缺失链接的可能性越大。

现在定义另一个 $M\times N\times P$ 三阶张量 E^* 来表示网络观察到的部分,E^* 是监督学习中训练数据集中的目标值。设 I 表示链接存在/缺失元组的索引。E^* 中的每一个元素定义如下:

$$[E^*]=\begin{cases} e^*_{i,j,k} & \text{if}(i,j,k)\in I \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

这里 $e^*_{i,j,k}$当有一个链接存在一个三元组中其值为正,反之为负。考虑使用基于节点的信息来进行链接预测,对集合 F、T 和 L 可以事先算出集合元素间的相似矩阵分别给定 S_F、S_T、S_L,这些矩阵中的值都是非负的、对称的。

2.1　半监督学习的链接预测方法

本文采用一种称为具有对象函数和智能三元组相似矩阵的链接繁殖方法进行链接预测。如上所述,链接预测问题可以看成一个半监督学习(一种更精确的转换学习方法)的问题。在半监督学习方法中,有一种最先进的被称为标志繁殖[18,19]的学习方法,该方法最初是利用标志繁殖原理来预测无标志节点的标志的。标志繁殖原理是:两个节点如果互相相似,则很有可能有一样的标志。这个原理可以被用做链接预测。这里把链接预测看做是预测一个三元组的标志,在此三元组是指链接强度。

将标志繁殖原理改做链接繁殖原理:两个三元组如果相似则很有可能有一样的链接。根据标志繁殖原理,定义最小化的目标函数如下:

$$\Gamma(\{e_{i,j,k}\})=\alpha\sum_{i,j,k,l,m,n}s_{ijk,lmn}(e_{ijk}-e_{lmn})^2+\sum_{(i,j,k)\in I}(e_{ijk}-e^*_{ijk})^2+\mu\sum_{(i,j,k)\notin I}e^2_{ijk} \tag{3}$$

其中:$s_{ijk,lmn}$表示的是两个三元组$\langle s_i,d_j,l_k\rangle$与$\langle s_l,d_m,l_n\rangle$之间对称的三元智能相似度。目标函数中第一项表示如果两者之间的相似度 $s_{ijk,lmn}$很大,那么两个三元组的两个链接强度的值 e_{ijk} 和 e_{lmn}应该是非常接近;第二项是损失函数,其针对 I 中的三元组调整预测结果适应它们的目标值;第三项是调整项以预防预测结果偏离 0 太远,也调整数值稳定性。αD 0 和 μD 0 是用来平衡整个等式的调整参数。

现在用张量来重写式(3),定义一个 $M\times N\times P$ 三阶张量如下:

$$[\delta]_{i,j,k}=\begin{cases}1 & \text{if}(i,j,k)\in I\\ \sqrt{\mu} & \text{otherwise}\end{cases} \tag{4}$$

设 Lap 是一个称为 Laplacian 拉普拉斯的 $MNP\times MNP$ 矩阵,定义如下:

$$Lap=D-S \tag{5}$$

其中:D 是一个对角矩阵,它的对角元素为

$$[D]_{i,i}=\sum_j[S]_{i,j}$$

S 是一个三元智能相似度矩阵,它的元素定义如下:

$$[S]_{MN(k-1)+M(j-1)+i,MN(n-1)+M(m-1)+l}=s_{ijk,lmn} \tag{6}$$

使用 Lap 和 δ 重写式(3)如下:

$$\Gamma(E)=\alpha\,\text{vec}(E)^{\mathrm{T}}L\,\text{vec}(E)+\|\text{vec}(E*\delta-\text{vec}(E^*))\|_2^2 \tag{7}$$

其中:$*$ 是 Hadamard 乘积(两个张量的元素乘积);vec(A)是通过对模型的维(如列)进行堆积构建而成的向量,定义如下:vec$(A)_{(k-1)NT+(j-1)N+i}=[A]_{i,j,k}$。当 $F=T$ 且没有链接方向,E^* 正面的切片是对称的,所以 E^* 的结果也是对称的。

为了得到使式(7)最小化的 E,对式(7)求 vec(E)的倒数,结果如下:

$$\partial\Gamma(E)/\partial\,\text{vec}(E)=2\alpha L\,\text{vec}(E)+\text{vec}(E*\delta)-\text{vec}(E^*) \tag{8}$$

当式(8) =0 时得到驻点,得到下列线性等式:

$$(2\alpha+\text{diag}(\text{vec}(\delta)))\,\text{vec}(E)=\text{vec}(E^*) \tag{9}$$

操作符 diag 产生一个对角元素由其参数向量给定的对角矩阵。

2.2　生成三重相似度矩阵

因为一个三重相似度矩阵 S 有 $M^2N^2P^2$ 个元素,给出所有的元素是不现实的。本文考虑用 S 中元素的相似矩阵 S_F、S_T、S_L 来系统构建 S 的相似矩阵。采用克罗内克积方法来构建三重相似度矩阵。

定义 1　给定实际矩阵 $A\in\mathrm{R}^{n\times m}$ 和 $B\in\mathrm{R}^{p\times q}$,克罗内克积 $A\otimes B\in\mathrm{R}^{np\times mq}$和列堆积操作 vec$(A)$ $\in\mathrm{R}^{nm}$定义如下:

$$A\otimes B=\begin{bmatrix}A_{11}B & A_{12}B & \cdots & A_{1m}B\\ \vdots & \vdots & & \vdots\\ A_{n1}B & A_{n2}B & \cdots & A_{nm}B\end{bmatrix},\text{vec}(A)=\begin{bmatrix}A_{*1}\\ \vdots\\ A_{*m}\end{bmatrix}$$

这里 A_{*j}表示 A 的第 j 列。克罗内克积有如下一性质:

$$\text{vec}(ABC)=(C^{\mathrm{T}}\otimes A)\,\text{vec}(B) \tag{10}$$

$$(A\otimes B)(C\otimes D)=AC\otimes BD \tag{11}$$

$$A\otimes B=A\otimes I_{pq}+I_{nm}\otimes B,A\in\mathrm{R}^{n\times m},B\in\mathrm{R}^{p\times q} \tag{12}$$

两个实际矩阵 $A,B\in\mathrm{R}^{n\times m}$ 的 Hadamard product(哈玛达积)是通过元素的乘法得到的,表示为 $A\odot B\in\mathrm{R}^{n\times m}$。哈玛达

积和克罗内克积有如下关系：

$$(A\otimes B)\odot(C\otimes D)=(A\odot B)\otimes(C\odot D)$$

首先，采用克罗内克积相似度方法，其思想是：如果两个三元组间三个交叉节点对彼此相似，那么两个三元组相似。克罗内克积相似矩阵可以定义如下：

$$S=S_F\otimes S_T\otimes S_L \tag{13}$$

其中：⊗表示克罗内克积。式(13)可以用元素积的方式表达：

$$s_{ijk,lmn}=[S_F]_{i,l}[S_T]_{j,m}[S_L]_{k,n} \tag{14}$$

这样，三元组之间的克罗内克积可以设计成每个集合相似性的积。这是使用在核方法[20~22]中点对相似性的扩展应用。如果 S_F、S_T和 S_L是作为定义在 S_F、S_T和 S_L特征空间的内积的核矩阵，那么克罗内克积相似性和三个特征空间的积空间间的内积相对应。使用克罗内克积相似性，可以对式(9)的拉普拉斯矩阵进行如下表示：

$$Lap=D_L\boxplus D_T\boxplus D_F-W_L\otimes W_T\otimes W_F \tag{15}$$

其中：D_F是一个对角矩阵，其对角元素定义为$[D_F]_{i,i}=\sum_j[S_F]_{i,j}$；$D_T$和 D_L有相似定义。

既然克罗内克积相似性的积空间有时变得非常复杂和极度的高维，本文考虑另一种具有更为紧致的特征空间(假如它是一个核函数)的相似性方法，称为克罗内克和相似性。克罗内克和相似性是基于如果两个三元组的三个交叉点对中的两个点对是相同的，而另外一个点对是相似的，则两个三元组是相似的。克罗内克和相似性定义如下：

$$\begin{aligned}S&=S_Z\otimes S_T\otimes S_F=\\&(S_Z\otimes I_N\otimes I_M)+(I_P\otimes S_T\otimes I_M)+\\&(I_P\otimes I_N\otimes S_F)\end{aligned} \tag{16}$$

其中：I_M是一个 $M\times M$ 大小的单位矩阵。这个等式同样可以用元素的形式来表示：

$$\begin{aligned}s_{ijk,lmn}=&[S_F]_{i,l}\gamma(j=m)\gamma(k=n)+\\&\gamma(i=l)[S_T]_{j,m}\gamma(k=n)+\\&\gamma(j=m)\gamma(i=l)[S_L]_{k,n}\end{aligned} \tag{17}$$

这里 γ 是一个函数，当它的参数为真返回 1，否则为 0。使用克罗内克和相似性方法，拉普拉斯矩阵可表示成如下形式：

$$Lap=Lap_L\boxplus Lap_T\boxplus Lap_F \tag{18}$$

这里 Lap_F 是一个拉普拉斯矩阵定义成 $Lap_X=D_X-S_X$。Lap_T和 Lap_Z有类似的定义。

从克罗内克积的定义可以得出，给定任意一对三元组，它们相识度的分值都大于 0，但是克罗内克和只有当三元组至少有两个元素是一样的才能得到正值。在使用克罗内克积时，即使元素相似矩阵不大，其克罗内克积也是非常巨大的($MNP\times MNP$)，要求的存储空间也非常大。尽管克罗内克和相似矩阵是相当稀疏的，但是仍然需要很大的空间。本文使用在核方法中使用了解决同样的相似矩阵空间扩展问题的方法——使用向量特技[23,24]的共轭梯度方法。

3　快速链接繁殖算法

3.1　用于链接繁殖的共轭梯度方法

共轭梯度方法 $Af=f^*$ 是解决系统线性问题的标准方法[25]。具体如下所示：

算法 1　共轭梯度( )

Input A, f^*, ε

Output f, p, r

f(0) := f^*

r(0) := f^* − A f(0); p(0) := r(0)

for t = 0, 1, 2, …, do

q(t) := Ap(t)

$\alpha(t):=\dfrac{\langle r(t),p(t)\rangle}{\langle p(t),q(t)\rangle}$

f(t+1) := f(t) + α(t)p(t)

r(t+1) := r(t) − α(t)p(t)

$\beta(t):=\dfrac{\|r(t+1)\|_2^2}{\|r(t)\|_2^2}$

if $\dfrac{\|r(t+1)\|_2^2}{\|r(0)\|_2^2}<\varepsilon^2$, return f(t+1)

p(t+1) := r(t+1) + β(t)p(t)

end

对共轭梯度方法进行改造，对式(9)使其线性化用于解决链接繁殖，A、f 和 f^* 分别用项 $2\alpha+\mathrm{diag}(\mathrm{vec}(\delta)$、$\mathrm{vec}(E)$ 和 $\mathrm{vec}(E^*)$，得到相应的共轭梯度算法——基于克罗内克积的算法(算法 2)和克罗内克和的算法(算法 3)。

算法 2　基于克罗内克积的链接繁殖算法

Input (S^*, Γ, S_F, S_T, S_L, α, ε)

Output S, P, R

S(0) := S^*

R(0) := $-\alpha(D_L\boxplus D_T\boxplus D_F-W_L\otimes W_T\otimes W_F)\mathrm{vec}(S(0))$; P(0) := R(0)

for t = 0, 1, 2, …, do

Q(t) := $-2\alpha(D_L\boxplus D_T\boxplus D_F-W_L\otimes W_T\otimes W_F)(P(t))+\Gamma*P(t)$

$\rho(t):=\dfrac{\langle R(t),P(t)\rangle}{\langle P(t),Q(t)\rangle}$

S(t+1) := S(t) + ρ(t)P(t)

R(t+1) := R(t) − ρ(t)P(t)

$\beta(t):=\dfrac{\|R(t+1)\|_2^2}{\|R(t)\|_2^2}$

if $\dfrac{\|R(t+1)\|_2^2}{\|R(0)\|_2^2}<\varepsilon^2$, return S(t+1)

P(t+1) := R(t+1) + β(t)P(t)

end

算法 3　基于克罗内克和的链接繁殖算法

Input (S^*, Γ, S_F, S_T, S_L, α, ε)

Output S, P, R

S(0) := S*

R(0) := $-\alpha(\mathrm{Lap}_L\boxplus \mathrm{Lap}_T\boxplus \mathrm{Lap}_F)\mathrm{vec}(S(0))$; P(0) := R(0)

for t = 0, 1, 2, …, do

Q(t) := $-2\alpha(D_L\boxplus D_T\boxplus D_F-W_L\otimes W_T\otimes W_F)(P(t))+\Gamma*P(t)$

$\rho(t):=\dfrac{\langle R(t),P(t)\rangle}{\langle P(t),Q(t)\rangle}$

S(t+1) := S(t) + ρ(t)P(t)

R(t+1) := R(t) − ρ(t)P(t)

$\beta(t):=\dfrac{\|R(t+1)\|_2^2}{\|R(t)\|_2^2}$

if $\dfrac{\|R(t+1)\|_2^2}{\|R(0)\|_2^2}<\varepsilon^2$, return S(t+1)

P(t+1) := R(t+1) + β(t)P(t)

end

在算法 1 中用向量来描述，而在算法 2 和 3 中用张量进行描述。但是在算法 2 和 3 中 $Lap_L\boxplus Lap_T\boxplus Lap_F$和 $D_L\boxplus D_T\boxplus D_F-W_L\otimes W_T\otimes W_F$仍然是非常大的矩阵，在运算上依然是一个瓶颈问题。

3.2　向量特技方法

为了解决算法 2 和 3 的运算瓶颈问题，本文采用一种向量特技的方法来有效解决运算问题，此方法可以加速克罗内克矩

阵和向量化的矩阵/张量的乘法运算。

设 A_F、A_T、A_L 分别表示 $M\times M$、$N\times N$、$P\times P$ 对称矩阵,B 表示一个 $M\times N$ 矩阵,向量特技方法如下所示:

$$(A_Y\otimes A_X)\,\mathrm{vec}(B)=\mathrm{vec}(A_XBA_Y) \tag{19}$$

式(20)的左边需要 $O(M^2N^2)$ 的时间和空间,右边则需要 $O(MN(M+N))$ 的时间和 $O(MN)$ 的空间,Vishwanathan[23] 使用此式加速图核的运算。在文献[26]中对张量提出了更为通用的方法,将应用扩展到三维乃至高维的张量中去。式(19)扩展到三维的形式如下:

$$(A_Z\otimes A_Y\otimes A_X)\,\mathrm{vec}(B)=\mathrm{vec}(B\times_1A_X\times_2A_Y\times_3A_Z) \tag{20}$$

式(20)的左边需要 $O(M^2N^2T^2)$ 的运算时间和空间,右边需要 $O(MNT(M+N+T))$ 的时间和 $O(MNT)$ 的空间。在克罗内克和的条件下有

$$(A_Y\,8\;\;A_X)\,\mathrm{vec}(B)=\mathrm{vec}(B\times_1A_X+B\times_2A_Y) \tag{21}$$

$$(A_Z\,8\;\;A_Y\,8\;\;A_X)\,\mathrm{vec}(B)=\mathrm{vec}(B\times_1A_X+B\times_2A_Y+B\times_3A_Z) \tag{22}$$

使用式(19)和(22)可以对算法 2 和 3 中克罗内克积和克罗内克和部分进行简化,如下所示:

$$D_L\,8\;\;D_T\,8\;\;D_F-W_L\otimes W_T\otimes W_F=B\times_1D_F\times_2D_T\times_3D_L-B\times_1W_F\times_2W_T\times_3W_L$$

$$Lap_L\,8\;\;Lap_T\,8\;\;Lap_F=B\times_1Lap_F\times_2Lap_T\times_3Lap_Z \tag{23}$$

3.3　算法的效率

链接相似性传递算法由于使用了向量特技方法而大大提高了内存使用效率,仅需要 $O(MNP+M^2+N^2+P^2)$ 空间,在计算复杂度上,理论上需要 $O(M^2N^2T^2(M+N+T))$,因为共轭远远小于梯度的算法每次迭代需执行 $O(MNT(M+N+T))$ 的时间且需要执行 $O(MNT)$ 次迭代。但是实际上迭代需要的次数 $O(MNT)$,笔者在实际的实验中发现,预测的性能在仅仅迭代了几次以后就不发生变化了。

4　实验与分析

在本章对多关系链接预测(三阶张量完成)设计了基于半监督学习方法的实验,在下列的实验中,设置实验参数为 $\alpha=0.001$,$\mu=1$。使用已经设计好的三元预测算法可以同时预测两个网络,换句话说,可以用点对链接预测方法分别预测两个网络,比较两种方法的链接预测性能。

本文选用两个生物网络数据集用于三元链接预测,每个数据集包含两个网络。第一个数据集是来源于两个不同实验室[27,28]的蛋白质相互作用网,在这个数据集中包含了历年来两个实验室的检测数据。其中一个网络(Med2Pub)有 3 879 个节点和 2 356 条边,另一个网络有(GOA)有 3 989 个节点和 2 104条边,这两个网络有 432 个链接是相同的。第二个数据集是一个物理性蛋白质相互作用网络和一个基因蛋白质相互作用网络,这两个网络数据都存放在 MIPS 数据库中。在物理网络中两个蛋白质如果通过实验确定有相互作用,那么就存在一个链接。在基因网络中两个蛋白质相应的基因如果同时出现导致细胞死亡的突变,那么就存在一个链接。物理性网络存在 1 425 个节点和 4 474 条边,基因网络存在 1 425 个节点和 1 476条边,两个网络有 234 条边是相同的。

在这两个数据集中,核矩阵和相似性矩阵由基因表达式、局部点和系统发育谱构成。本文随机选择 40% 的三元组数据作为训练集,其余的作为测试集,反复预测 10 次。

图 1 显示了具有克罗内克积相似性和克罗内克和相似性的两个蛋白质相互作用网络的平均 AUC 和标准偏差。单个(each)和同时(simultaneous)表示一个网络接一个网络进行预测和同时进行预测。分别显示了三个信息源的预测结果,从结果中可以看出,在大多数情况下两个网络同时进行预测能够提高预测的性能。

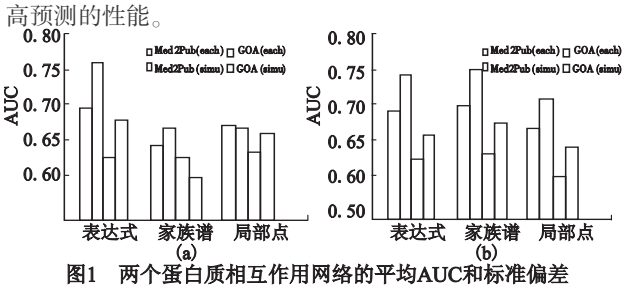

图1　两个蛋白质相互作用网络的平均AUC和标准偏差

图 2 显示了具有克罗内克积相似性和克罗内克和相似性的基因网络和物理性网络的运行结果。单个(each)和同时(simultaneous)表示一个网络接一个网络进行预测和同时进行预测。分别显示了三个信息源的预测结果。尽管这两个网络不如两个蛋白质相互作用网络性能提高得那么好,但特别在使用克罗内克和相似性时同时进行预测能够提高性能。同样地,在这两个数据集中,大多数情况下克罗内克和比克罗内克积要好。

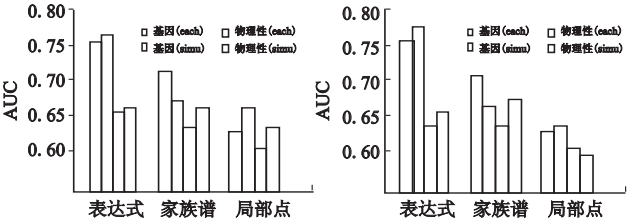

图2　基因网络和物理性网络的运行结果

5　结束语

本文将标志相似性传递技术运用到链接预测,提出了一种新的链接预测方法。在该方法中同时运用了节点间的链接信息和链接类型信息,使用了克罗内克和相似性矩阵和克罗内克积相似性矩阵作为相似性矩阵,共轭梯度方法作为有效的学习算法,使用了张量的向量特征技术,降低了普通标志相似性传递方法中扩展性难的问题,通过实验验证其有效性。

基于链接相似性传递算法还存在一些将来需要继续改进的工作:a)矩阵压缩方法的研究。因为即使相似性矩阵是稀疏的,结果也是稠密的,对于大尺度网络难以存储到内存中,必须寻求有效的压缩张量表示方法。b)拓扑信息的使用。笔者使用了节点对相似性的可见部分构建了相似性矩阵,但是网络拓扑信息应用中的矩阵因式分解和张量分解等方法也可以利用到相似性矩阵的应用中来。c)信息集成。要处理的数据源是复杂类型的,得到多种类型的相似性矩阵是非常关键的。要考虑一种集成的方法能够自动调整每种相似性矩阵的权重。d)一种快速的检索方法。因为要处理的对象是大尺度网络中的数据,要求能够有一种快速有效的检索方法,能够快速查找数据,提高处理的效率。

参考文献:

[1] GELOO L, DIEHL C. Link mining: a survey[J]. ACM SIGKDD

Explorations,2005,7(2):3-12.
[2] 郭景峰, 张健, 邹晓红. 基于链接的 Web 网页分类[J]. 计算机应用研究, 2008, 25(11): 3271-3274.
[3] HOLME P, HUSS M, SOCIAL J R. Role-similarity based functional prediction application to the yeast proteome[J]. Interface, 2005, 2(4):327-333.
[4] GALLAGHER B, TONG H, ELIASS-RAD T, *et al*. Using ghost edges for classification in sparsely labeled networks[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press, 2008.
[5] 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.
[6] JACCARD P. Distribution de la flore alpmedans le bassin des dranse-set dans quelques regions voisines. [J]. Bulletin de la Societe Vaudoise des Science Naturelles,1901,37:241-272.
[7] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3):211-230.
[8] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999,286(5439):509-512.
[9] KATZ L. A new status index derived from socio metric analysis[J]. Psychometrika,1953,18(1):39-43.
[10] GOBEL F, JAGERS A A. Random walks on graphs[J]. Stochastic Processes and Their Applications, 1974,2(4):311-336.
[11] FOUSS F, RENDERS J M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(3):355-369.
[12] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7):107-117.
[13] JEH G, WIDOM J. SimRank: a measure of structural-context similarity[C]//Proc of the 8th ACM SIGKDD Internatioanl Conference on Knowledge Discovery and Data Mining. 2002:538-543.
[14] BLONDEL V D, GAJARDO A, HEYMANS M, *et al*. A measure of similarity between graph vertices: applications to synonym extraction and Web searching[J]. SIAM Review, 2004, 46(4):647-666.
[15] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2):21-10.
[16] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191):98-101.
[17] ZHANG Yi-cheng, BLATTNER M, YU Y K. Heat conduction process on community networks as a recommendation model[J]. Physical Review, 2007, 99(15):154301.
[18] ZHU Xiao-jin, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions[C]//Proc of the 20th International Conference on Machine Learning. 2003:912-919.
[19] ZHU Xiao-jin, LAFFERTY J. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning[C]//Proc of the 22nd International Conference on Machine Learning. 2005:1052-1059.
[20] BASILICO J, HOFMANN T. Unifying collaborative and content-based filtering[C]//Proc of the 21st International Conference on Machine Learning. 2004.
[21] BEN-HUR A, NOBLE W S. Kernel methods for predicting protein-protein interactions[J]. Bioinformatics, 2005, 21(Suppl.1):i38-i46.
[22] OYAMA S, MANNING C D. Using feature conjunctions across examples for learning pair-wise classifiers[C]//Proc of the 15th European Conference on Machine Learning. 2004:322-333.
[23] LAUB A J. Matrix analysis for scientists and engineers[M]. [S.l.]:Society for Industrial and Applied Mathematics, 2005.
[24] VISHWANATHAN S V N, BORGWARDT K M, SCHRAUDOLPH N N. Fast computation of graph kernels[M]. Cambridge, MA: MIT Press, 2007:1449-1456.
[25] GOLUB G H, LOAN C F V. Matrix computations[M]. 3rd ed. [S.l.]:Johns Hopkins University Press, 1996.
[26] KOLDA T G, BADER B W. Tensor decompositions and applications Technical Report SAND2007[R]. [S.l.]:Sandia National Laboratories, 2007.
[27] ITO T, CHIBA T, OZAWA R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interaction 2001[C]//Proc of National Academy of Sciences. 2001.
[28] UETZ P, GIOT L, CAGNEY G, *et al*. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae[J]. Nature, 2000,403(6770):623-627.

(上接第 2847 页)

[4] BALABANOVIĆ M, SHOHAM Y. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3):66-72.
[5] SEO Y W, ZHANG B T. A reinforcement learning agent for personalized information filtering[C]// Proc of the 5th International Conference on Intelligent User Interfaces. New York: ACM Press, 2000: 248-251.
[6] BREESE J, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proc of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1998: 43-52.
[7] SARWAR B M, KARYPIS G, KOWSTAN J, *et al*. Item-based collaborative filtering recommendation algorithms[C]// Proc of the 10th International Conference on World Wide Web Conference. New York: ACM Press, 2001: 285-295.
[8] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: MIT Press, 1998.
[9] TEN HAGEN S H G, HAGEN S H G, KROSE B J A. Generalizing in TD(lambda) learning[C]//Proc of the 3rd Joint Conference on Information Science. 1997: 319-322.
[10] IRODOVA M, SLOAN R H. Reinforcement learning and function approximation[C]//Proc of the Florida AI Research Society Conference. 2005: 455-461.
[11] ZHEN Yi, LI Wu-jun, YEUNG D Y. TagiCoFi: tag informed collaborative filtering[C]// Proc of the 3rd ACM Conference on Recommender Systems. New York:ACM Press, 2009: 69-76.
[12] LIANG Hui-zhi, XU Yue, LI Yue-feng, *et al*. Collaborative filtering recommender systems using tag information[C]//Proc of International Conference on Web Intelligent and Intelligent Technology. Washington DC:IEEE Computer Society, 2008:59-62.
[13] PENG Jing, ZENG D. Topic-based Web page recommendation using tags[C]//Proc of IEEE International Conference on Intelligence and Security Informatics. Piscataway:IEEE Press, 2009:269-271.