基于层次随机图模型的脑网络链路预测*

田 甜, 杨艳丽, 郭 浩, 陈俊杰
(太原理工大学 计算机科学与技术学院, 太原 030024)

摘 要: 针对一般链路预测算法在具有层次结构的脑网络中计算效率低且复杂度高的问题,提出了一种基于最大似然估计的层次随机图模型。该算法首先利用脑网络数据建立层次随机图,然后通过改进的马尔可夫蒙特卡罗算法采样树状图空间,最后计算脑网络边的平均连接概率,且通过评价指标对算法进行评价。实验结果表明,利用该算法对脑网络和三种不同的层次结构网络进行链路预测比较, 脑网络的预测结果最好。此外,所提出的算法较之传统的基于相似性的算法,该算法效果明显,且具有理想的计算复杂度。

关键词: 脑网络; 链路预测; 最大似然估计; 层次随机图模型; 马尔可夫蒙特卡罗算法
中图分类号: TP183 **文献标志码**: A **文章编号**: 1001-3695(2016)04-1066-04
doi:10.3969/j.issn.1001-3695.2016.04.023

Link prediction of brain networks based on hierarchical random graph model

Tian Tian, Yang yanli, Guo Hao, Chen Junjie
(*College of Computer Science & Technology*, *Taiyuan University of Technology*, *Taiyuan* 030024, *China*)

Abstract: Focusing on the problem of poor efficiency and high complexity in general link prediction algorithms applied in brain networks, this paper proposed a hierarchical random graph model based on maximum likelihood estimation. Firstly, this algorithm used brain networks data to create the hierarchical random graph model. Then, it sampled the space of all possible dendrograms using an improved Markov chain Monte Carlo algorithm. Finally, it calculated the average connection probability of brain network edges, and it also evaluated by the evaluation index. Experimental results show that the algorithm exhibits best result in brain network between different hierarchical networks. In addition, it obtains the good effect and reasonable computing complexity compared with the traditional algorithms based on similarity.

Key words: brain network; link prediction; maximum likelihood estimation method; hierarchical random graph; Markov chain Monte Carlo algorithm

0 引言

在复杂大脑神经网络[1]中,为了找到脑区之间潜在的关系,需要对各脑区节点之间的链路进行预测。脑网络链路预测[2]就是指在已有的脑区关系中寻找未知的链接关系,并通过已知的链接去探究未知的信息。此外,对于已知的脑网络,其数据可能模糊不清或者存在混乱。利用链路预测算法不仅可以甄别已观测到的链接,同时还可以根据算法结论,对错误的链接进行纠正。这对脑网络重组、结构优化和演化都有重要意义。

目前链路预测算法主要分为三类:基于概率模型算法、相似性定义算法、最大似然估计算法。早期链路预测的方法和研究思路主要是基于概率模型的。这类算法主要应用于机器学习和数据挖掘领域,且适用于大规模数据集的网络。Sarukkai[3]应用马尔可夫链模型(Markov)进行网络的路径分析与链路预测,主要解决万维网 Web 服务器的路径导航。此外,Madadhain 等人[4]利用已知的网络拓扑属性以及节点特性建立了条件概率模型来进行链路预测。近几年,Dong 等人[5~8]提出了基于网络拓扑结构的相似性方法[9]。这类方法主要用于具有相似性的简单无向网络。常见的局部相似性指标包括杰卡德系数、共同邻居和最短路径长度等。另外一类链路预测方法是基于网络结构的最大似然估计。该算法适用于具有组合结构的网络。2008 年,Claustet 等人[10]提出了利用网络具有明显的层次结构去建立层次随机图模型(hierarchical random graph,HRG)的链路预测方法。同时,Karrer 等人[11]将网络中的节点分组建立随机分块模型,去预测错误边和缺失边。

然而对于早期的基于概率模型的链路预测算法,主要用于大规模数据集网络,它计算复杂度太高及非普适性的参数使其应用范围受到了限制;对于基于节点相似性的算法,只适用于具有节点相似性的网络,在计算无节点相似性的真实网络时效率很低;而基于层次随机图的算法,在处理具有组合结构的网络时,具有很好的精确度。由于脑网络是一种典型的复杂系统,网络中的连接关系可以看做某种内在层次结构的反映。而建立和表达层次结构最有效、最直观的方法就是采用层次随机图描述。基于此,本文选用层次随机图模型来进行脑网络的层次构建和链路预测。该方法首先通过正常人脑网络的层次结构建立层次随机图模型,然后将马尔可夫蒙特卡罗算法[12,13]

收稿日期: 2014-11-27; **修回日期**: 2015-01-16 **基金项目**: 国家自然科学基金资助项目(61170136,61373101,61472270,61402318);太原理工大学青年基金资助项目(2012L014);太原理工大学青年团队启动项目(2013T047)

作者简介:田甜(1988-),女,山西忻州人,硕士,主要研究方向为功能性脑网络(363585954@qq.com);杨艳丽(1987-),女,山西太原人,博士研究生,主要研究方向为胞认知计算与建模分析;郭浩(1981-),男,山西太原人,讲师,博士,主要研究方向为智能信息处理与脑影像学;陈俊杰(1956-),男,河北定州人,教授,博导,主要研究方向为智能信息处理、数据挖掘的研究.

(Markov chain Monte Carlo,MCMC)应用到贝叶斯理论里,在马尔可夫链归于平稳后,采集最优层次随机图模型。以该最优模型为种子,采取与其似然成比例的一批树状图空间,最后将所有网络同一条连边的概率均值作为最终的结果。本算法克服了其他传统方法中处理组合结构网络导致效率低的问题,提高了建模的准确率,并且采用多样本树形图实现了求边概率均值。最后在真实的脑网络数据上进行实验,结果表明了算法的有效性。

1　层次随机图模型

复杂网络在很多情况下都具有层次结构。由 Claustet 等人在 2008 年提出的层次随机图模型,对来自不同领域且具有层次结构的食物链网、恐怖袭击网和梅毒螺旋体的代谢网络分别进行了链路预测。该层次结构模型目前还广泛应用于探究复杂网络的拓扑特性。本文将用层次随机图模型来构建大脑网络的层次结构。而最优层次随机图可以最好地描述脑网络的层次结构,其核心思想就是找到最佳描述网络的最优二叉树。

1.1　层次随机图模型的基本概念

网络层次随机图就是层次组织模型,其概念为:a)网络 G 是具有 n 个节点的简单无向图;b)树状图 D 是二叉树,它具有 n 片叶子对应于 G 中的 n 个节点,且它有 $n-1$ 个内部节点 r 对应于来自 D 中形成的节点对;c)给出 G 中的两个节点 i、j 和边连接的概率 p_{ij},其中 r 是节点 i、j 在 D 的最近共同祖先。概率 $p_r \in [0,1]$ 为每个内部节点 r 的概率值,且每个内部节点 r 独立于 p_r。那么边的连接强度 $p_{ij}=p_r$,即两个节点的连接概率等于距离它们最近的共同祖先节点所赋予的概率。所以,树状图 D 和概率集合 $\{p_r\}$ 共同定义了层次随机图 $(D,\{p_r\})$。

1.2　最优层次随机图的基本概念

找到最符合观察到真实网络 G 的层次随机图,即要找到最优层次随机图。首先,假设所有层次随机图是先验等可能的。然后,通过贝叶斯定理,给定正确解释数据的模型 $(D,\{p_r\})$ 的概率与生成可观测网络的似然估计值 L 成比例。则层次随机图对网络 G 的似然估计值表示为式(1),其中 r 为树状图 D 中节点对的最近共同祖先,L_r 和 R_r 分别代表以内部节点 r 为根的左子树和右子树的叶子节点数;E_r 代表以 r 为最近共同祖先的节点对在 G 中已形成连边的节点对数。最后,当树状图 D 确定后,使似然估计值 L 最大化的最优概率公式为 $\overline{p_r}=E_r/L_rR_r$;就将此概率公式带入式(1),得到最大似然 $L(D)$ 式(2)。利用式(2)可得到最优层次随机图。

$$L(D,\{p_r\})=\prod_{r\in D}p_r^{E_r}(1-p_r)^{L_rR_r-E_r} \tag{1}$$

$$L(D)=\Big[\prod_{r\in D}\overline{p_r}^{\overline{p_r}}(1-\overline{p_r})^{1-\overline{p_r}}\Big]^{L_rR_r} \tag{2}$$

2　MCMC 算法

为了抽样与似然估计值 $L(D)$ 成概率比的树状图模型空间,本文使用马尔可夫蒙特卡罗算法去解决这个高维运算问题。

2.1　MCMC 算法的基本思路

MCMC 基本思想是:构造一条 Markov 链,使其平稳分布为待估参数的后验概率,通过采集后验分布的样本对树状图模型空间采样。则基本思路可概括为:

a)构造 Markov 链。Markov 链的每一步包括:首先,随机均匀选择树状图的一个内部节点 r(非根)。然后,通过子树重排随机选择与该节点相关的可转换的子树。结果产生一个新的树状图 D'。它简单的表明,这种类型的转换是遍历的。

b)采用 Metropolis - Hastings 标准抽样树状图。在生成新的树状图 D' 后,根据 Metropolis - Hastings 标准[14]去接受或淘汰它。具体地说,如果 $\lg^{L(D')}\geqslant\lg^{L(D)}$,则接受认可新树状图 D';否则以 $L(D')/L(D)$ 的概率接受 D'。Metropolis - Hastings 标准结合了转换的遍历性,它不仅保证了细致平衡,也保证了和似然成比例的树状图的极限概率分布。

c)采集树状图空间。在大约 $O(n^2)$ 步骤后,马尔可夫链相对较快地收敛,似然值达到一个平稳状态,此时可以获得最优概率模型。利用该种子模型采样和它成概率比的模型空间。

2.2　子树重排

在创建马尔可夫链的过程里,需要在可能的树状图之间选择一组转换。而转换就是树状图的重排。即任何有限树状图节点对可以通过这样一系列有限的转换而关联起来。

子树重排的目的是获得新的树状图。它的方法是随机选择一个内部节点 r 并考虑以其儿女节点为根节点的子树集合 A、B 和以其兄弟节点为根节点的子树集合 C。通过交换子树 A 和 C 或子树 B 和 C,来获得新的树状图。这两种子树重排方法都可以保证不打乱它们的内部关系。

3　实验与分析

3.1　数据采集和预处理

实验对象是由山西医科大学第一临床附属医院招募的健康志愿者,共收集 28 例。所有被试者均无疾病史,无过敏史,酒精滥用史,视力或矫正视力正常且右利手。每位被试者在了解实验内容和目的后,均达成了书面协议。

本实验采用西门子公司的 3.0T 磁共振仪器来完成功能磁共振成像[15](functional magnetic resonance imaging,fMRI)数据的采集。采集工作是在山西医科大学第一临床附属医院完成的,所有的扫描工作由熟悉磁共振操作的放射科医生来完成。在扫描的过程中,要求被试者闭眼、放松、不去想特定的事情但要保持清醒不能睡着。利用统计参数映射 SPM8(http://www.fil.ion.ucl.ac.uk/spm/software/spm8/)软件对采集的 fMRI 数据进行处理,包括:层间时间较正、头动较正、联合配准、标准化、平滑处理。

3.2　构建功能性脑网络

定义功能性脑网络。首先,利用 AAL[16](anatomical automatic labeling)模板,将全脑划分成 90 个区域(左右半脑各 45 个),每个脑区代表脑网络中的一个节点。然后对脑区之间作 Pearson 相关。最后,共有 4 005 对节点组成边的全集,即有 4 005 个相关系数值 R。将 4 005 个 R 值按稀疏度 $a\%$ 划分,a 的取值为 5 ~ 40,步长为 5,即把所有的 R 值进行排序,排序之后取前 $a\%$,前 $a\%$ 代表节点对之间有边,而剩余的节点对之间代表无边。最终得出每个正常被试者共有八个的脑网络。如图 1 所示,描述了一个正常被试者的功能性脑网络(L 为左半脑,R 为右半脑)。

为了测试算法的准确性,对边集合 E 按照一定的规则划分为训练集 E_T 和测试集 E_P 两部分,其中 $E=E_T\cup E_P$ 且 $E_T\cap E_P=\varphi$。显然,E_T 作为已知信息和算法的输入,E_P 是用来验证算法的有效性。对之前选出的边按 10% ~ 90% 划分,其中随

机选取 90% 作为训练集 E_T，剩余 10% 作为测试集 E_P。

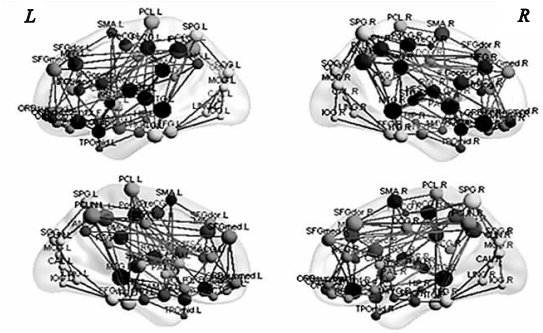

图1　功能性脑网络示意图

3.3　实验过程

使用层次随机图来预测缺失边的算法如下所示：

a)通过随机选择树状图，初始化马尔可夫链。首先，为树状图 D 和网络 G 分配内存空间，利用 G 中标记的非内部节点构建随机二叉树。其次，标签 D 中所有内部节点，最终将多种叶子排序结构转变为单一树状图结构。最后，初始化马尔可夫链，导入训练集 E_T，计算该树状图的最大似然值。

b)运行蒙特卡罗算法，直到达到平衡。将生成的树状图作为入口，在马尔可夫链没达到平稳之前，一直在作子树交换，利用 Metropolis - Hastings 标准判断是否接受该树状图，若接受，则替换。在达到归一化步数 N^2(N 为脑网络节点个数)时，即当时间复杂度达到 $O(n$^2)时，算法处于平衡。

c)利用 MCMC 算法抽样树状图模型空间。当马尔可夫链达到平衡后，采样一个最优树状图模型。以该最优模型为种子，采样和它成似然比的一批树形图，数量为 10 000。

d)排序。对于未连边每一对节点 i、j，通过平均在每一个抽样的树状图 D 的相应概率 p_{ij}，计算它们相互连接的平均概率。最后以均值降序排序 i、j 节点对，预测排名越靠前表示连边出现的概率越大。

3.4　评价指标

本文选用 AUC[17](area under the receiver operating characteristic curve)指标来量化预测算法的准确度。AUC 统计也就是指 ROC(receiver-operating characteristic)曲线下的面积。它可以被解释为每次从测试集中任选一条边，然后与随机选择的未连的边比较分数值。如果测试集中被选边的值大于不存在的边，数值加 1 分；如果两个数值相等，就加 0.5。独立比较 n 次后，若测试集中边得分高的次数为 n'，两次比较后数值相等的次数为 n''。则 AUC 定义为

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n} \tag{3}$$

显然，如果所有分数都是随机生成的，则 AUC = 0.5。而 AUC 大于 0.5 的程度表明预测算法在很大程度上要比随机选择的方法更好。

3.5　算法实现与结果分析

1)最优模型下的似然值统计

实验导入脑网络数据后，马尔可夫链在进行一系列的子树交换，似然值达到平衡。此时，时间达到 n^2(n 为 90 个脑节点数)，分布趋于平稳。如图 2 所示，X 轴代表稀疏度，Y 轴代表似然值。图中表示在八种不同的网络稀疏度规模下，Markor 链平稳时，算法采集最优种子模型所需的似然值。结果表明，随着网络边的增加，似然值呈下降趋势。图 3 中，X 轴代表时间，Y 轴代表似然值，该图表示在已知 40% 网络稀疏度规模下，一个正常被试者的似然值变化情况。如图 3 所示，Markov 链随着时间的增加，似然值在不断地增大，最终趋于平稳。

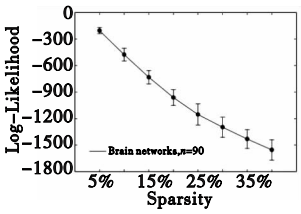

图2　八种不同稀疏度下的最大似然值

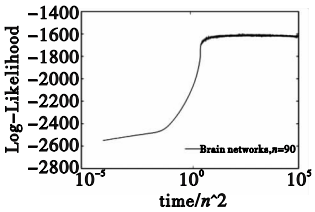

图3　40%稀疏度下的似然值

当 MarkoV 链达到平衡时，可利用 MCMC 算法采集最优种子模型，并通过 MATLAB 进行可视化。如图 4 所示，在已知 40% 的网络边下，采集到种子模型。其中，大脑网络共有 90 个脑区节点，每个脑区都定义了序号。图中数字表示节点号，方形表示脑网络节点，圆圈表示内部节点，连线表示节点之间有边。

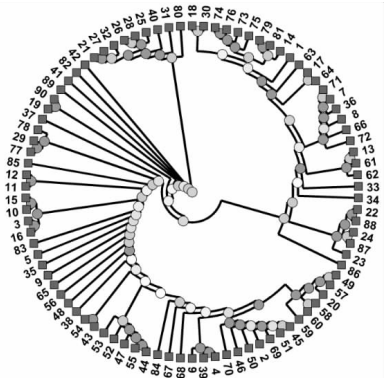

图4　最优层次图模型的MATLAB可视化

2)层次随机图模型在不同网络中的比较

利用层次随机图方法，将 30% 稀疏度下的大脑网络(Brain)与 Claustet 等人提出的三个网络进行链路预测比较(如图 5)。这三个网络分别为梅毒螺旋体代谢网络(T. pallidum)、恐怖分子关系网络(Terrorists)、草地物种食物网(Grassland)。对于每个网络，随机均匀移除一部分边，然后基于剩余的边，尝试预测网络连边概率值。然后进行 AUC 评分。图中的 X 轴代表导入实验的训练集占边集的比例，Y 轴代表评价指标 AUC。实验结果表明，利用层次随机图算法对所有的网络进行链路预测，Brain network 呈明显优势，而 T. pallidum 和 Grassland 网络的效果较差。并且，随着网络规模的增大，Brain network 的 AUC 值由上升逐渐趋于平稳。

3)不同的链路预测算法在脑网络中的比较

在大脑网络中，将层次随机图(HRG)模型与随机预测方法和基于相似性的链路预测方法进行比较。其中，采用基于相似性的方法包括局部相似性指标共同邻居(CN)和最短路径(LP)；采用全局相似性指标平均通勤时间(ACT)。CN 表示网络中两节点拥有更多的共同邻居时，节点相似性越大，且更易连边；LP 表示网络中两节点存在最短路径时，它们更易连边；ACT 表示网络中两节点有较短的通勤时间时，它们更易连边。

图 6 中，X 轴表示稀疏度，Y 轴表示评价指标 AUC，图 6 表示在 28 例正常人的八种不同网络稀疏度规模下，对各种算法作 AUC 统计，实验结果表明层次随机图算法优于随机预测和全局指标 ACT，而劣于局部指标 CN 和 LP。尽管层次随机图模型没有产生最优的结果，但其预测效果明显，具有较好的时间复杂度。

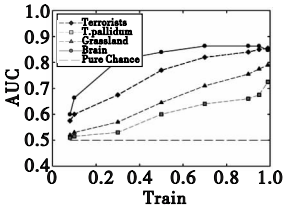

图5　层次随机图模型在不同网络中的比较

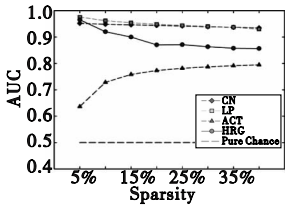

图6　不同算法在脑网络中的比较

事实上,在大部分真实的脑网络中,由于大量数据的模糊与丢失,很难找到脑网络节点之间的相似性。所以,鉴于脑网络具有明显的层次结构,非常适合用层次随机图来预测链路。

4　结束语

链路预测是研究复杂脑网络的关键问题。本文提出了将层次随机图应用到脑网络链路预测中的新方法,取得了较好的效果。主要分为三方面的工作。第一,应用层次随机图模型对脑网络进行层次构建,统计达到平衡时的最大似然估计,并将最优模型 MATLAB 可视化。第二,将层次随机图模型在不同网络中进行比较,脑网络表现最优。第三,不同算法在脑网络中进行比较,层次随机图算法的 AUC 评价指标较好。因此,本方法很适合用于具有明显层次结构,小规模的脑网络。不足是当脑网络划分成上千、上万个节点时,该算法的时间复杂度很高,计算需花费大量时间,这个问题也存在优化。

由于研究者发现大脑功能网络的异常拓扑变化和许多神经精神疾病(如重度抑郁症和精神分裂症等)有关,所以,下一步研究将探索病人的脑网络链路缺失情况以及复杂脑网络的拓扑特性。

参考文献:

[1] Zaleskya A, Fornito A, BullmorE E. Network-based statistic: identifying differences in brain networks[J]. Neuroimage, 2010, 53(4): 1197-1207.

[2] Lyu Linyuan, Zhou Tao. Link prediction in complex networks: a survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150- 1170.

[3] Sarukkai R R. Link prediction and path analysis using markov chains [J]. Computer Networks, 2000, 33(1-6): 377-386.

[4] O'Madadhain J, Hutchins J, Smyth P. Prediction and ranking algorithms for event-based network data[C] //Proc of ACM SIGKDD. New York: ACM Press, 2005: 23-30.

[5] Dong Enming, Li Jianping, Xie Zheng. Link prediction via convex nonnegative matrix factorization on multiscale blocks[J]. Journal of Applied Mathematics, 2014, 2014(2014): 1-9.

[6] Zhou Tao, Lyu Linyuan, Zhang Yicheng. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[7] Liu Haifeng, Hu Zheng, Tian Hui, *et al*. Hidden link prediction based on node centrality and weak ties[J]. Europhysics Letters, 2013, 101(1): 18004-18009.

[8] Cannistraci C V, Alanis-Lobato G, Ravasi T. From link prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports, 2013, 3(3): 1613-1626.

[9] 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.

[10] 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.

[11] Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks[J]. Physical Review E, 2011, 83(1): 16107-16117.

[12] 李成录, 张永仁. 基于贝叶斯与马尔可夫链蒙特卡洛融合算法的图像处理探究[J]. 电子测试, 2013, 9(16): 26-27.

[13] 朱嵩, 毛根海, 刘国华, 等. 改进的 MCMC 方法及其应用[J]. 水利学报, 2009, 40(8): 1019-1023.

[14] 鲁帆, 严登华. 基于广义极值分布和 Metropolis-Hastings 抽样算法的贝叶斯 MCMC 洪水频率分析方法[J]. 水利学报, 2013, 44(8): 942 - 949.

[15] 毛德旺, 袁建华, 丁忠祥, 等. 静息态功能磁共振成像在脑功能区定位中的初步应用[J]. 浙江医学教育, 2014, 13(1): 48-50.

[16] Yang Banghung, Tsai S Y, Chen Zhicheng, *et al*. Reliability evaluation of I-123 ADAM SPECT imaging using SPM software and AAL ROI methods[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2011, 648(1): 51-53.

[17] Xu Weichao, Dai Jisheng, Wang Qinruo, *et al*. Estimating the area under a receiver operating characteristic (ROC) curve: parametric and nonparametric ways[J]. Signal Processing, 2013, 93(11): 3111-3123.