关于论文《Identification of Discriminative Subgraph Patterns in fMRI Brain Networks in Bipolar Affective Disorder》的学习报告

这篇paper是Philip S. Yu老师实验室做的对fMRI脑网络进行可区分的子图模式识别。Motivation是脑网络中的链接架构能够包含很多关于脑部结构的有用信息，而整个脑结构构成的网络图太庞大了，因此网络中的链接也变得非常复杂。而传统的对脑网络进行研究的算法一般都是选取特定的脑区域的局部模式或者成对链接，而这篇论文运用统计学和图形挖掘算法，以获得可区分的子图模式，这些子图模式是通过一系列的成对链接构成的，用以区分双向情感障碍疾病。

这篇paper提出的方法如下：

1. 二元图：G=(V,E)，就是无权图，任意节点之间的关系只为是否存在连接边，而这些连接边并不存在重要性区别，也就是不存在权重；
2. 大脑网络中的子图模式：表示大脑中某些区域的集合及它们之间的连接。目前已经有许多研究者研究如何从图数据中挖掘出有用的子图模式。而在挖掘子图模式时，需要指定挖掘准则，即我们需要挖掘出怎样的子图模式。一种较为典型的评估标准就是频率，即挖掘出经常出现的子图模式，而基于频率的子图模式挖掘一般是无监督的，从而导致对于解决分类问题没有很有效的区分性；这篇paper处理的是一个有监督任务，是在给定标签的情况下从健康人和患病者中识别出可用来区分这两种疾病的子图模式。Paper中采用的是一种G-test统计方法来评估某个子图的得分，定义如下：

$$t(g,\mathcal{D}) = 2m(p \cdot ln\frac{p}{q} + (1-p) \cdot ln\frac{1-p}{1-q})$$

其中p为子图模式g在正标签的个体中出现的频率而q为子图模式g在负标签的个体中痴线的频率。这个公式是基于一个假设：假定正标签和负标签的分布频率是相等的，而当子图模式g在两种标签的个体中的区分性越大，则G-test的得分就更高。

这里有一个问题就是如何得到要评估的子图模式g？因为一个巨大的脑网络中任意数量的节点都能构成一个子图模式，即使只考虑那些连接的节点，这个子图模式的数量规模也不小。Paper在实验部分提到关于选取候选的子图模式这个问题，是通过使用the BrainNet Viewer这个工具从脑网络中选取其中最重要的10个子图模式作为候选者。

3. 对fMRI脑网络进行子图模式挖掘：由于fMRI得到的是大脑的功能性结构图，因此这个网络图一般是带权重的，而以上提出的方法是针对二元图（无权图）。论文中采用的方法是把有权图转换为二元图，但并非采取设定权重阈值的方法，因为这种方法会丢失很多重要的信息。论文中采用的方法是通过将两个大脑区域之间的正相关关系（即原来的权重）作为对应边存在的概率来构成二元图。

定义$\tilde{G}$为有权图G（节点为n）得到的二元图，那么$\tilde{G}$的可能性就有2^n种，

而从G得到$\tilde{G}$的概率可以通过以下公式得到：

$$\Pr(\widetilde{G} \Rightarrow G) = \prod_{e \in E(G)} p(e) \prod_{e \in E(\widetilde{G}) - E(G)} (1 - p(e))$$

这里有一个问题，对于一个规模不小的网络，相应的某些子图模型应该也不小，而这里出现阶乘级的数量，这个算法复杂度会很大吧。
并且，这里的G只是一个个体的数据，我们要从全局考察某一个子图模式是否可以用来区分所有患病个体与正常个体，那么就需要对每个个体都要考虑，即

$$\Pr(\widetilde{\mathcal{D}} \Rightarrow \mathcal{D}) = \prod_{i=1}^{n} \Pr(\widetilde{G}_i \Rightarrow G_i)$$

从而得到全局测量准则为：

$$\textsc{Exp}\left(t(g, \widetilde{\mathcal{D}})\right) = \sum_{\mathcal{D} \in \mathcal{W}(\widetilde{\mathcal{D}})} \Pr(\widetilde{\mathcal{D}} \Rightarrow \mathcal{D}) \cdot t(g, \mathcal{D})$$

其中W($\tilde{D}$)为从包含有权图的数据集D可以得到的所有可能的转换为二元图的数据集。
所以这个算法最大的问题就是计算复杂度，但是论文中说他是通过某篇论文中的一种动态规划的方法可以有效地计算出以上的全局测量准则。