◎数据库、信号与信息处理◎

基于蛋白质相互作用网络的聚类算法研究

刘 昊,廖 波,彭利红
LIU Hao,LIAO Bo,PENG Li-hong

湖南大学 计算机与通信学院,长沙 410082
School of Computer & Communication,Hunan University,Changsha 410082,China
E-mail:4719999@qq.com

LIU Hao,LIAO Bo,PENG Li-hong.Research of clustering algorithms in protein-protein interaction network.Computer Engineering and Applications,2008,44(30):142-144.

Abstract: Protein-protein interaction network refers a new research area of computer science.The distance metric in such setting is redefined by the network distance,which has to be computed by the expensive shortest path distance over the network.The existing methods are not applicable to such cases.Therefore,by exploiting unique features of networks,a new clustering algorithm is presented,which uses the information of nodes and edges in the network to prune the search space and avoid some unnecessary distance computations.The experimental results indicate that the algorithm achieve high efficiency for clustering nodes in real protein-protein interaction network.

Key words: data mining;protein-protein interaction networks;clustering algorithm;network distance;shortest path

摘 要: 蛋白质相互作用网络是计算机科学技术的一个新研究领域。蛋白质相互作用网络中结点之间的距离度量需要通过基于网络的最短路径距离来重新定义,其计算代价高,这使得已有的基于欧几何距离的聚类算法不能直接运用到这种环境中。因此,通过蛋白质相互作用网络的特征提出了一种新的聚类算法。算法使用网络中的边和结点信息来缩减搜索空间,避免了一些不必要的距离计算。实验结果表明,算法对于真实的蛋白质相互作用网络中的结点聚类是高效的。

关键词: 数据挖掘;蛋白质相互作用网络;聚类算法;网络距离;最短路径
DOI: 10.3778/j.issn.1002-8331.2008.30.043 **文章编号:** 1002-8331(2008)30-0142-03 **文献标识码:** A **中图分类号:** TP311

随着生物信息学技术和海量生物数据技术的发展,蛋白质相互作用网络开始提出,并因为其应用得广泛性而得到越来越多地高度重视。聚类算法的研究在蛋白质相互作用网络这一复杂网络中挖掘其潜藏的具有生物学意义的信息,以及在注释未知蛋白质的生物学功能,了解生命活动的机制,并为新药靶点的发现和药物设计提供理论基础等许多方面都具有广泛的应用前景。可以说蛋白质相互作用网络是计算机科学和计算分子生物学的一场革命,是21世纪最重要的研究之一。

蛋白质相互作用网络与万维网,人类社会关系网都隶属于复杂网络,其数据量十分庞大,而且蛋白质相互作用网络通常由大量高密度的蛋白质结点构成,但是其中也不缺乏稀疏连接结点。这就决定了聚类是在蛋白质相互作用网络中进行数据挖掘的首要工具。对于聚类,研究热点和难点在于聚类的设计,使其聚类的过程具有高效性。数据挖掘中的聚类算法有很多,但是适合蛋白质相互作用网络的确不多,最近一些适合蛋白质相互作用网络的聚类得到提出,它的研究已经成为蛋白质相互作用网络研究中的热点。

根据适合蛋白质相互作用网络的聚类算法的特点,可将它们大致分为4类:划分的方法(partitioning method)、基于密度的方法(density-based method)、基于模型的方法(model-based method)、层次的方法(hierarchical method)。

划分的方法研究网络中包括所有的孤立点的每个部分。The Restricted Neighborhood Search Clustering(RNSC)[1]是通过一个函数值来发现最佳的划分部分。它是随机划分一个网络,然后迭代地移动类边缘上的一些点到这个类的邻接类中去,目的是为了能降低这两个类的函数值之和。这个算法最终要找到一个最佳的划分使得所用的类的功能函数值之和最小。这个方法的最大的缺点就是要事先知道要划分的目标类的确定个数。

基于密度的方法是在网络中寻找高密度连接的子网络。为寻找完全连通子图的最大团算法(maximum clique algorithm)[2],它能从蛋白质相互作用网络中检测出那些高度连接的蛋白质,但是从整体来看,它不能分类存在大量稀疏结点的网络。因为稀疏的连接性将会降低类的密度,这种基于密度的算法产生的

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.10571019)。
作者简介: 刘昊(1984-),男,硕士研究生,主要研究领域为数据挖掘、智能信息处理;廖波(1974-),男,博士,教授,主要研究领域为生物信息学、智能信息处理;彭利红(1978-),女,硕士研究生,主要研究领域为数据挖掘、智能信息处理。
收稿日期: 2008-04-16 **修回日期:** 2008-07-07

类中将不会包含那些稀疏结点。最近有研究人员尝试用基于密度的算法去发现蛋白质相互作用网络中的两者共用类（overlapping clusters）[3]。

基于模型的方法，基于马尔可夫模型的聚类算法（MCL）是 van Dongen[4]开发的用于图形聚类的算法，并成功应用于蛋白质家族分析等领域。该算法是基于随机流的一种方法，不需预先设定聚类数目。其关键点在于“直到一个紧密连接的模块中多数节点已被访问过，随机流才会离开这个模块”，并通过概率改变和反复修改矩阵以实现随机流模拟[5]。

层次聚类的方法有其独特的研究优势，因为生物系统的层次组织结构和层次聚类本身不需要事前知道网络中目标类的确定个数。它是迭代地将节点合并或着是递归地将网络划分成多个子网络。它的终止条件是当被类之间相似度或着相似距离达到预定阈值。The Super Paramagnetic Clustering（SPC）method[6]这个算法是迭代地将节点合并。而递归地将网络划分成多个子网络的层次方法是找到关键的边，将它删去。比如[7-8]这两个算法都是通过找到网络中节点间最短路径通过最多的边设为关键边，递归地完成网络的划分，形成各个类。层次聚类的算法唯一的缺陷是它对噪声数据的鲁棒性比较差。

1　基本定义

本章首先引入文献[9]中的网络关键点定义和加权的蛋白质相互作用网络定义，然后根据蛋白质相互作用网络的聚类特点，增加了聚类间的网络距离定义。最后，通过提出一个新的聚类块的概念来定义蛋白质相互作用网络中的结点聚类问题。

定义 1　一个网络可以表示为一个无向待权图 $G=(V,E,W)$，其中：V 是结点的集合；E 是边的集合；W 为正的实数集合（$W:E\rightarrow IR^+$），表示边所在对应的权值。

当聚类蛋白质相互作用网路中的结点时，需要定义结点之间的距离相似性为网络距离，假设 p 和 q 为网络中的两个结点，定义结点间网路距离度量如下：

定义 2　结点之间的网络距离。

（1）两个结点的最短路径边上的结点的距离：$W(p,q)$。

（2）结点 p 与关键结点 A 之间的距离：

$$1\times W(A,p_1)+\frac{1}{2}\times W(p_1,p_2)+\cdots+\frac{1}{2^n}\times W(p_{n-1},p_n) \tag{1}$$

（3）关键结点之间的网路距离：$Din(n_i,n_j)$为从关键结点 n_i 到关键结点 n_j 的最短路径距离。

定义 3　聚类块 M_s 与聚类块 M_t 的相似度函数

$$S(M_s,M_t)=\frac{\sum_{v_i\in M_s,v_j\in M_t}C(v_i,v_j)}{\min(|M_s|,|M_t|)} \tag{2}$$

$$C(v_i,v_j)=\begin{cases}1 & \text{if } v_i=v_j\\ w_{i,j} & \text{if } v_i\neq v_j \text{and} <v_i,v_j>\in E\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

定义 4　聚类块为根据阈值 ε 构建的某个最短路径上的一个微小聚类。网络中的聚类由一个或多个相邻的聚类块组成，给定一个网络中的 N 个结点的集合，对这些结点进行聚类的过程转换为根据它们的距离形成聚类块，并根据聚类块之间的网络距离合并这些聚类块的过程。

2　基于边的聚类算法

一个类是一个具有某种共同特征的对象的集合。聚类是将数据对象分组的过程，它的结果是使得相同组内对象之间的相似程度比不同组之间的对象间的相似程度要大。聚类实际上是一种无人监督的分类方法，这就意味着它不依靠于事前确定类里面的训练数据对象。在蛋白质相互作用网络中，类分为两个模块：蛋白质联合体和功能模块。蛋白质联合体是相互作用在同一时间，同一地点，构成单独的多分子机构的一组蛋白质。功能模块是产生于不同时间，不同地点的一组蛋白质，这些蛋白质相互作用完成某个特殊的细胞过程。

聚类能识别蛋白质相互作用网络中的功能模块和蛋白质联合体，这一过程有以下好处：

（1）充分了解蛋白质相互作用网络的体系结构和它的成分之间的关系。

（2）根据每个类中的蛋白质功能推测每个类的特殊功能。

（3）通过对比每个类中已知蛋白质的功能去阐明这个类中未知蛋白质的功能。

2.1　算法的思想

考虑利用结点包含的边信息，从中间某个层次将对象划分为初始聚类，再分别向底层和顶层进行层次的分裂和合并，调整聚类的结果。基于这种思想，提出了基于边的聚类方法，它实际上也是一种在层次聚类中采用动态模型的聚类算法。

2.1.1　算法说明

基于边的聚类算法分为 3 个阶段：初始化阶段、分类阶段和合并阶段。

（1）初始化阶段：根据用户输入的阈值确定关键结点，即一个关键结点和其周围的结点被指定到同一个聚类中，初始聚类的个数为关键结点的个数。关键结点的选取将在算法关键技术部分予以阐述。

（2）分裂阶段：根据结点之间的距离相似度函数，依次分裂大的聚类为细粒度的聚类块，得到聚类的中间结果，具体说来，对于每个初始的聚类，若其中某两个相邻结点之间的网络距离大于预先给定的阈值 ε，则在这两个相邻的结点之间将聚类划分开来，分裂为两个细粒度更小的聚类块；否则，可将其作为一个单独的聚类块。这个过程保证了聚类块内部的紧凑性，即聚类块内部任意两个相邻结点之间的网络距离均小于。

（3）合并阶段：根据聚类块之间的距离相似度循环合并相邻的聚类，以形成最终的聚类结果。

2.1.2　算法的执行流程

（1）根据用户输入的阈值 ε 从 n 个结点中选取关键结点作为初始的聚类中心。

（2）由结点间的网络距离计算方法，得出每个数据到聚类中心的距离，并得到初始聚类。

（3）根据结点之间的距离相似度函数，依次分裂大的聚类为细粒度的聚类块，得到聚类的中间结果。对于每个初始的聚类，若其中某两个相邻结点之间的网络距离大于预先给定的阈值，则在这两个相邻的结点之间将聚类划分开来，分裂为两个细粒度更小的聚类块。

（4）通过聚类块与聚类块之间的相似度函数的比较，将最大的相似度函数值与合并阈值比较，来决定是否合并这两个模块。在这个合并的过程中，发现类与类之间有存在重复的结点，当这些结点有边连接时，它们就是共享模块（overlapping structure）。

2.1.3　算法的关键技术

在非加权的图中，关键点的选取主要是考虑度较大的节

点,即度是一个重要的因素。有研究表明[10]高度连接的蛋白质比稀疏连接的蛋白质拥有更多的功能意义。

关键蛋白质 V_i 的选取是基于与节点 V_i 相连的边的权之和，边之间的权重 $W_{i,j}$ 是通过计算蛋白质 i 与蛋白质 j 的基因表达谱之间的关联系数得到[8]。

$$d_i=\sum_{v_j\in N(v_j)} w_{i,j} \tag{4}$$

3　性能分析与实验仿真

3.1　算法分析

基于边的聚类算法的代价都线性于结点数目 N,其最差情况下需要遍历整个网络图,因此,算法的时间复杂度为 $O(|V|\log|V|+N)$,其中,V 为网络中结点的个数,而 N 为网络中关键结点的个数。

分析基于边的聚类方法,其高效性主要表现在以下 3 点:

(1)由于算法从某个合适的层次分组对象进行聚类的初始化,使得仅需少数的分裂和合并就能得到聚类的结果,从而减少了将每个结点作为聚类的代价高的初始化以及中间合并过程。先分裂后合并的方法也避免了传统的层次算法合并后难以调整结果的聚类准确度问题。

(2)通过引入参数,不仅控制了聚类内对象之间的相似度,而且控制了聚类之间的相似度,使得能够在聚类的合并过程中尽早地发现合适的聚类结果。

(3)利用网络中聚类的特征,引入聚类块的概念来简化网络中聚类的表示和计算,并且在合并时只对结点相邻的聚类合并,减少了大量不同类之间的结点的网络距离的计算,提高了聚类算法性能。

3.2　实验内容和结果分析

为了评估聚类算法的有效性和性能,将基于边的聚类算法应用于人类有关 AD 蛋白质相互作用网络图[11](图 1),并分别与传统的基于欧几何距离的 Marland Bridge 算法[10]和 Korbel算法[12]进行比较。

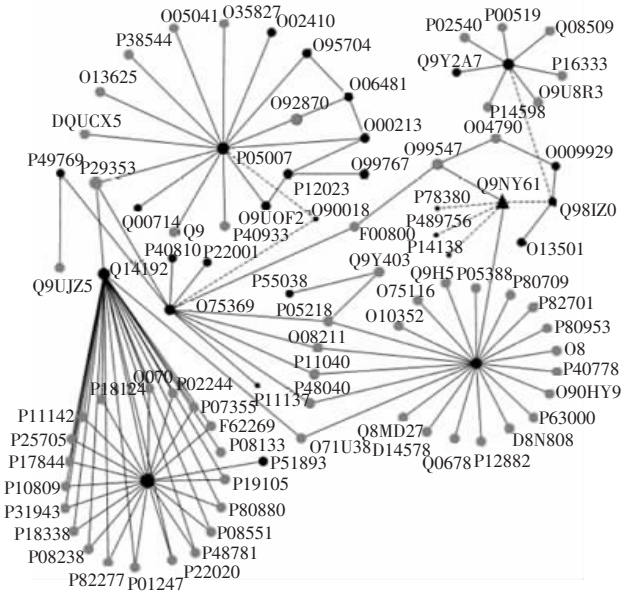

图 1　人类 AD 相关蛋白质相互作用图

统计出算法得到的聚类总数以及每个聚类中的结点,比较不同算法的聚类效果。计算出结点数平均绝对误差百分比[13](记作 MAPE(Mean Absolute Percent Error))。这个度量值可衡量聚类效果。$|nodes|$为结点总数,$olratio$ 为孤立点的比例,设为 1%,K 为产生的聚类个数,$|M_i|$为聚类 i 的结点个数。

$$MAPE=\frac{1}{k}\sum_{i=1}^{k}\left|1-\frac{|M_i|}{|nodes|\times(1-olratio)/K}\right| \tag{5}$$

实验统计结果见表 1。调整算法的参数使得边聚类算法产生三种聚类效果，聚类结果的平均相对误差均小于 Marland Bridge 算法和 Korbel 算法,聚类效果更好。

表 1　不同算法的聚类效果比较

	Marland Bridge	Korbel	基于边的算法
K	6	7	6
$MAPE$	0.051 5	0.049 7	0.033 9
K	6	7	7
$MAPE$	0.051 5	0.049 7	0.025 2
K	6	7	8
$MAPE$	0.051 5	0.049 7	0.041 1

在提出的算法中,ε 参数是一个用户定义的阈值，其取值决定了聚类结果。在最后的实验中改变参数 ε 取值,测试其对边聚类效率和聚类结果的影响,表 2 显示了算法随参数 ε 变化的聚类效果,不同 ε 得到的聚类结果精度不同,对于本实验,但 ε 取 1.7 时,算法得到较好的聚类结果见表 2。

表 2　参数 ε 对聚类效果的影响

ε	基于边的算法	
1.0	K	6
	$MAPE$	0.033 9
1.7	K	7
	$MAPE$	0.025 2
2.0	K	8
	$MAPE$	0.041 1

通过上述比较可以看出:对于蛋白质相互作用网络中的结点聚类,基于边的聚类方法可以有效地找出聚类的结果,而且算法的效率较高,伸缩性良好。

4　结论

蛋白质相互作用网络的研究是离不开聚类算法的研究,由于传统的生物学方法和拓扑学方法不能适用于蛋白质相互作用网络,所以近年来新的适用于蛋白质相互作用网络的聚类算法以成为蛋白质相互作用网络研究中的热点。

本文主要对蛋白质相互作用网络聚类方法进行研究,根据网络距离重新定义蛋白质相互作用网络中的聚类问题。利用出网络的特征,提出新的基于网络距离的聚类方法一基于边的聚类方法。实验表明,新的聚类方法是可行且高效的。聚类蛋白质相互作用网络中的聚类结果可以应用于检测蛋白质相互作用网络中的功能模块。

参考文献:

[1] King A D,Przulj N,Jurisica I.Protein complex prediction via cost-based clustering[J].Bioinformatics,2004,20(17):3013-3020.

(下转 245 页)