文章编号:1004-9037(2014)05-0833-07

基于小波去噪的有向加权社团发现研究

张梁梁[1] 潘志松[2] 李国鹏[2] 胡谷雨[2]

(1.解放军理工大学气象海洋学院,南京,211101;2.解放军理工大学指挥信息系统学院,南京,210007)

摘要:目前大部分社团发现方法都是针对无向无权图,但实际的社会媒体中的社团内部个体交互过程可以抽象为一个有向加权图,并且权重中含有大量的噪声。为解决有向加权社团的划分问题,本文提出一种基于非负矩阵分解(Nonnegative matrix factorization, NMF)可去噪声的社团发现方法。该方法通过小波阈值去噪对社会网络数据进行去噪处理,结合有向加权的非负矩阵分解算法对去噪后的数据集进行社团发现,准确找出社团结构。在社会媒体的实验数据集和标准数据集上的实验结果表明,该算法针对带噪声的有向加权图社团发现问题具有良好划分性能,SNR 为 15 时,在 Lesmis 数据集上的社团划分准确率达到 96%,划分模块度值提高了 29%。本文为解决带噪的有向加权的社会网络数据提供了切实有效的处理方法。

关键词:小波去噪;非负矩阵分解;社团发现;有向加权社团

中图分类号:TP391 **文献标志码**:A

Weighted Directed Community Detection Method Based on Wavelet Denoising

Zhang Liangliang[1], *Pan Zhisong*[2], *Li Guopeng*[2], *Hu Guyu*[2]

(1. College of Meteorology and Oceanography, PLA University of Science and Technology, Nanjing, 211101, China;
2. College of Command Information Systems, PLA University of Science and Technology, Nanjing, 210007, China)

Abstract: Most community detection methods are aiming at solving undirected and unweighted datasets. However, datasets are often directed and weighted with noise in real world. In order to process noisy and directed weighted community detection, a method based on nonnegative matrix factorization (NMF) is proposed. In the algorithm, wavelet threshold denoising is used to denoise the social network datasets. And the community structure is abtained by community detection through NMF. Simulations show the proposed method is more effective, i. e. for Lesmis dataset when SNR is 15, the accuracy of dividing community is 96% and the modularity of the method is improved by 29%. The proposed method is more applicable than other community detection methods for directed weighted datasets with noise.

Key words: wavelet denoising; non-negative matrix factorization; community detection; weighted directed community

引 言

近年来,由于社会媒体研究的兴起,人们对社会网络的研究已不再局限于网络结构的基本理论,而是转向研究连接结构复杂的实际网络。实际网络通常由若干个社团组成,会伴随一定的噪声影响。通过剖析社会网络的社团结构,可以进一步挖掘网络的隐含信息,但噪声对社团结构的分析有很大的影响,会湮没网络的正确特性分布,所以去噪有助于分析网络特性,准确理解社会网络的组织方式,对改善现有网络性能和预测网络行为有重要意义。加权网络比无权网络更能反映网络的特性,例如:人和人之间关系的紧密程度,微博中互相转发

基金项目:国家高技术研究发展("八六三"计划)(2012AA01A510)资助项目;国家自然科学基金面上(41275002)资助项目。

收稿日期:2014-01-07;**修订日期**:2014-03-09

的次数等[1]。常见的高斯噪声对边权重的影响会增加社团发现的难度,导致各个节点的关系发生改变,不利于发现真实的社团关系。例如:某个垃圾用户对全网用户发送垃圾信息,对此时的网络数据进行社团发现的正确率会降低。

目前已有一些算法能从不带噪声的社会网络中正确提取社团结构,这些算法大致可分为 3 类[2]:基于图论的方法,如 Girvan 和 Newman 提出分裂方法,GN 算法和快速 GN 算法(Fast-GN);基于矩阵分解的方法,如对称的非负矩阵分解算法(Symmetric nonnegative matrix factorization, SymNMF)算法;基于优化的划分方法,如 N-Cut 算法和 A-Cut 算法等;还有近几年很流行的大规模社团的快速展开算法(Fast unfolding of communities in large networks, FUCL)等[3]。这些算法在不含噪声的数据集上,社团发现效率和正确率较高,但如果网络中含有噪声,它们的效率明显下降[4],算法错误率伴随网络复杂度的增加而增长。本文提出一种基于非负矩阵分解的社团发现方法来解决含有噪声的网络社团发现问题。

1 基于 NMF 的社团发现方法

非负矩阵分解(Nonnegative matrix factorization,NMF)是一种聚类和降维的技术,在很多领域中都有应用。其基本思想是将一个矩阵 $\boldsymbol{G}$ 近似分解成两个非负矩阵乘积的形式,形如:$\boldsymbol{G}\approx\boldsymbol{WH}$,$\boldsymbol{G}$ 中的列向量 $\boldsymbol{g}_j$ 用 $\boldsymbol{W}$ 中的行向量 $\boldsymbol{w}_j$ 乘以矩阵 $\boldsymbol{H}$ 来表示,即 $\boldsymbol{g}_j=\boldsymbol{w}_j\boldsymbol{H}$。其含义是矩阵 $\boldsymbol{G}$ 用低维空间的一组基 $\boldsymbol{H}$ 来表示,而 $\boldsymbol{W}$ 可以看作是权重系数矩阵。从物理意义上讲,由于 $\boldsymbol{W}$ 是 $n\times k$ 维,$\boldsymbol{H}$ 是 $k\times n$ 维,可以根据权重 $\boldsymbol{W}$ 将原来矩阵 $\boldsymbol{G}$ 中的 n 个节点聚成 k 个类,$\boldsymbol{W}$ 中每行的值表示节点对 k 个社团的隶属度。

文献[5]首先将 NMF 拓展,可以发现重叠社团结构。主要思想是利用扩散核,把邻接矩阵正规化后再分解,同时在求解 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的迭代过程用梯度下降法。该方法能够发现重叠社团,但算法复杂度高,只适合规模较小的网络。

文献[6]利用 Frobenius 范数分别提出了基于有向图和无向图的社团组织发现方法。表达式可以这样描述为 3 层分解

$$\| \boldsymbol{G}-\boldsymbol{XSX}^{\mathrm{T}} \|_{\mathrm{F}}^{2} \tag{1}$$

式中:$\boldsymbol{X}$ 是 $n\times k$ 维,$\boldsymbol{S}$ 是 $k\times k$ 维矩阵。假设 $x_{i,j}$ 是 $\boldsymbol{X}_{n\times k}$ 中的元素,它表示节点 i 属于第 j 个社团的隶属度,$\boldsymbol{S}_{k\times k}$ 表示各个社团之间的相互关系,它们给出了求解式(1)的交替迭代算法,并证明了算法的收敛性。文献[7]中提到针对加权的有向图。文献[6]关于有向加权图的工作,性能不是很好,其利用 KL(Kullback Leibler)散度距离作为度量,分别提出了有向加权图和无向图的社团发现算法,该算法利用乘法更新规则进行求解。但是文献[5,6]都没有考虑矩阵稀疏性的特性。文献[8]认为,只有非负的约束并不能完全刻画社团结构,利用 l_1 范数描述矩阵的稀疏性,并利用坐标下降法求解式(1),该算法对有向加权图求解性能较好,可以发现重叠的社团结构。上述方法虽然都对社团的发现算法进行了详细研究,但对于真实的社团发现,除了连接关系,没有考虑利用诸如标签、留言等其他的相互关系矩阵来实现社团发现。文献[9]针对社会网络,给出式(2)的二层分解来描述图中的社团结构

$$\min \| \boldsymbol{G}-\boldsymbol{WH}^{\mathrm{T}} \|_{\mathrm{F}}^{2} \tag{2}$$

同时还利用标签等其他的相互关系来构建综合的社团发现结构模型,并利用积极集和交替迭代方法进行求解。但是文献[9]方法对于加权的矩阵,性能不是很好。目前大部分算法都是针对在理想情况下获得通联矩阵的条件下进行,没有考虑噪声对最后结果的影响。本文就是期望揭示实际的社团发现问题中的通联矩阵含噪声对社团划分的影响,并且通过基于小波阈值去噪的 NMF 方法进行有向加权社团发现。

2 噪声对社团划分的影响

为了验证噪声对社团发现的影响,本文采用一个含有噪声的有向加权模拟数据的实例,然后对比两种流行的算法性能。

首先定义有向加权社团的模块度[10]

$$\boldsymbol{Q}=\frac{1}{\boldsymbol{W}}\sum_{ij}\left(W_{ij}-\frac{w_i^{\mathrm{out}}w_j^{\mathrm{in}}}{\boldsymbol{W}}\right)\delta(\boldsymbol{C}_i,\boldsymbol{C}_j) \tag{3}$$

式中:$\boldsymbol{W}$ 为整个网络中所有边的权重之和,W_{ij} 为节点 i 和节点 j 之间的边的权重,w_i^{out} 为节点 i 出度边的权重之和,w_j^{in} 为节点 j 入度边的权重之和,C_i 为节点 i 所属的社团,C_j 为节点 j 所属的社团。对于有向无权图,图中所有边的权重可以看作 1。C_i 与 C_j 相同,则 $\delta(C_i,C_j)=1$,否则为 0。

该实例是随机产生一个含有噪声的有向加权数据集。该数据可以很明确地划分成 5 个社团,社团中节点个数是递减的,社团聚集程度 $C_1>C_2>C_3>C_4>C_5$。节点个数如表 1 所示。

表 1　节点分布情况

Table 1　Distribution of nodes

社团	C_1	C_2	C_3	C_4	C_5
节点数	250	200	150	100	50

图 1 是产生的模拟数据的邻接矩阵 $\boldsymbol{G}$，它是社会网络的可视化结果，图中各点颜色深浅表示每条边权值的大小，该数据是原始数据，不带噪声。用 FUCL 算法[3]和 Wei F 算法[6]划分社团，测试结果是：5 个社团的模块度平均值是 0.423 和 0.402 5，划分结果理想。

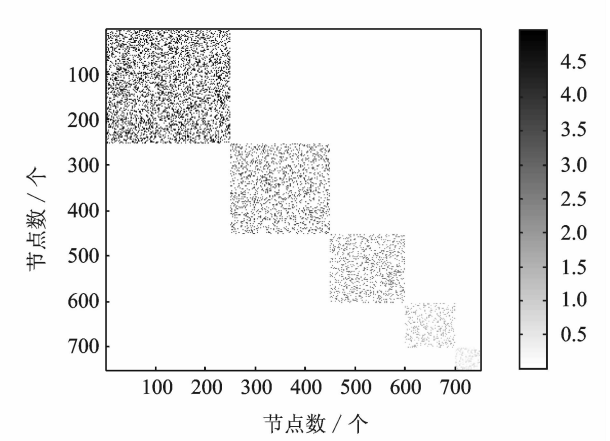

图 1　原始模拟数据

Fig. 1　Original simulated data

图 2 是在原始数据上加了噪声后的可视化结果，可以看出数据受到噪声影响，但仍然能够看出是 5 个社团，社团结构明显。通常情况下，噪音的比重越大，模块度函数值 Q 应该越小，划分效果越不理想。用 FUCL 算法和 Wei F 算法划分社团结果是 0.182 3 和 0.171 2，两种算法的 Q 都没有超过 0.3，社团发现的效果并不理想。对比的结果如表 2 所示。

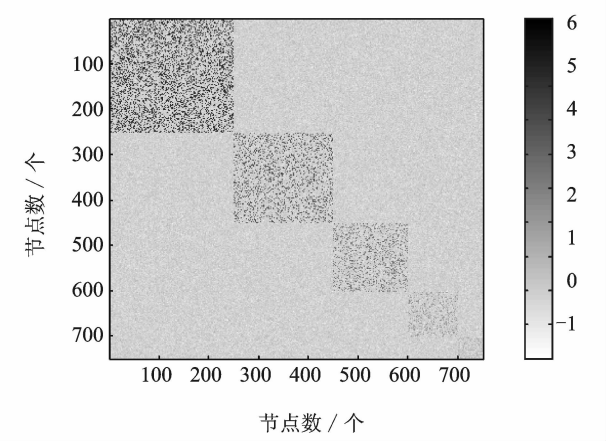

图 2　加噪模拟数据

Fig. 2　Simulated data with noise

表 2　FUCL 和 Wei F 算法的模块度值

Table 2　Modularity of FUCL and Wei F algorithm

模块度	原始数据	带噪数据
FUCL	0.423 0	0.182 3
Wei F	0.402 5	0.171 2

该实例说明含噪声的数据集会影响社团划分的结果，导致划分的正确率下降。

3　小波阈值去噪的有向加权图社团发现算法

本文算法的基本思想是针对有向加权图模型，重新定义目标函数的有效约束，赋予其物理意义，设计相应算法来求解目标函数以提高社团发现的效率和稳定性；其次针对带高斯噪声的邻接矩阵数据进行去除噪声，分析社团中节点的信息，重构关系矩阵，用新定义目标函数的算法进行社团发现。

3.1　小波阈值去噪

小波去噪有 4 种方法：模极大值去噪、小波系数相关性去噪、阈值去噪、平移不变量。阈值去噪法适合处理权重带高斯白噪声的邻接矩阵。

阈值去噪中要选择小波基、分解层数、阈值函数和阈值。在对真实数据的实验中发现硬阈值函数（Hard thresholding，HT）对邻接矩阵处理效果优于软阈值函数[11]。对阈值产生一般采用启发式（Heursure）方法。小波去噪中常用的小波函数有 Daubechies（dbN）小波和 Symlets（SymN）小波等。对于邻接矩阵这样的数据，可以近似地看成是块信号的类型，用 SymN 小波去噪的效果没有用 db 类小波去噪效果好。SymN 小波去噪后，邻接矩阵的噪声点还很多。

分解层数 j 可以决定噪声和信号分离的程度，一般来说 j 越大，分离效果越好，但是 j 大于一个阈值就会使得重构信号的精度下降，所以必须选取合适的 j 值。然而 j 的最大值与带噪信号的信噪比（Signal to noise ratio，SNR）相关，SNR 越大，则噪声越小，j 取较小值就可以很好地分离信号和噪声；如果 SNR 很小，说明数据中含有大量的噪声，j 必须取较大的值才能将信号和噪声分离。一般来说，要根据 SNR 的不同来选取不同的 j。若 SNR 大于 20，j 取 3 合适；SNR 小于 20，j 可以取 4 或者更大，去噪效果更好[11]。

3.2　可去噪的有向加权图社团发现算法

对于邻接矩阵 $\boldsymbol{G}$ 用非负矩阵的 3 项分解来近

似 $\boldsymbol{G}\approx\boldsymbol{XSX}^{\mathrm{T}}$，转置形式为 $\boldsymbol{G}^{\mathrm{T}}\approx(\boldsymbol{XSX}^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{XS}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$。

令：$\boldsymbol{A}=\boldsymbol{G}+\boldsymbol{G}^{\mathrm{T}}$，则 $\boldsymbol{B}=\boldsymbol{X}(\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}})\boldsymbol{X}^{\mathrm{T}}$。

其中矩阵 $\boldsymbol{S}$ 表示社团联系关系，令 $\boldsymbol{W}=\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}}$，定义如下的目标函数

$$\min L(\boldsymbol{X},\boldsymbol{S})=\|\boldsymbol{A}-\boldsymbol{B}\|_{\mathrm{F}}^{2}+\lambda\sum_{j=1}^{n}\|\boldsymbol{X}(j,:)\|_{1}^{2}=\left\|\frac{\boldsymbol{G}-\boldsymbol{XSX}^{\mathrm{T}}+\boldsymbol{G}^{\mathrm{T}}-\boldsymbol{XS}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}}{2}\right\|_{\mathrm{F}}^{2}+\lambda\sum_{j=1}^{n}\|\boldsymbol{X}(j,:)\|_{1}^{2}=\|\boldsymbol{A}-\boldsymbol{XWX}^{\mathrm{T}}\|_{\mathrm{F}}^{2}+\lambda\sum_{j=1}^{n}\|X(j,:)\|_{1}^{2}\quad(3)$$

$$\text{s. t.}\quad\begin{matrix}0\leqslant\boldsymbol{X}\leqslant\boldsymbol{I},\boldsymbol{W}\geqslant 0\\(\boldsymbol{WK})\otimes\boldsymbol{I}\geqslant 0\end{matrix}$$

其中 $\|\ \|_{\mathrm{F}}$ 是 F 范数。$\|\ \|_{1}$ 是 l_1 范数，用来控制模型的稀疏度，让节点尽可能被清晰地划分。λ 是用来平衡矩阵 $\boldsymbol{X}$ 和近似精度。$\boldsymbol{I}$ 是单位矩阵，其中元素都为 1，$\otimes$ 表示哈达玛(Hadamard)矩阵乘积。本文将不带去噪过程的算法称为非去噪的有向权社团发现算法(No-denoising weighted directed commity detection algorithm based on NMF，DNMF)。

3.3 更新原则

为了求解基于 NMF 的算法的最优解，一些优化的方法已经被广泛研究，例如：辅助函数法、梯度下降法、活动集法、坐标下降法等。本文将利用文献[8]中的有界非负矩阵 3 次分解算法(Bounded non-negative matrix tri-factorization，BNMTF)模型的坐标下降法对提出的算法进行优化。

定义矩阵 $\boldsymbol{N}_{k\times k}$ 中元素全为 1，函数可以写成

$$\boldsymbol{L}=\|\boldsymbol{A}-\boldsymbol{XWX}^{\mathrm{T}}\|_{\mathrm{F}}^{2}+\lambda\sum_{j=1}^{n}\|\boldsymbol{X}(j,:)\|_{1}^{2}=\mathrm{tr}[(\boldsymbol{A}-\boldsymbol{XWX}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{A}-\boldsymbol{XWX}^{\mathrm{T}})]+\lambda\mathrm{tr}(\boldsymbol{XNX}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{XS}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XSX}^{\mathrm{T}})-2\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})+\lambda\mathrm{tr}(\boldsymbol{XNX}^{\mathrm{T}})$$

$$\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{X}}=2\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}+2\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}^{\mathrm{T}}-2(\boldsymbol{AXW}^{\mathrm{T}}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{XW})+2\boldsymbol{XN}=4\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}-4\boldsymbol{AXW}+2\boldsymbol{XN}$$

$$\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{W}}=2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{X}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}$$

采用如下的更新规则[12]

$$\boldsymbol{X}\leftarrow\boldsymbol{X}.*\frac{\boldsymbol{AXW}-\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}}{2\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}+\lambda\boldsymbol{XN}}\quad(4)$$

$$\boldsymbol{W}\leftarrow\boldsymbol{W}.*\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{X}}{2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{X}}\quad(5)$$

对于带噪的有向加权邻接矩阵 $\boldsymbol{G}_{n\times n}$，当 SNR>20时，噪声对社团发现几乎没有影响；当SNR≤20 时，噪声会直接影响社团发现的结果，所以本文只考虑 SNR≤20 的情况。分解层数 j 在 SNR>20 时取 3 较合适，SNR≤20 时取 4 合适[11]，本文 j 取 4。小波阈值去噪的参数设置设置如下：硬阈值函数 HT，阈值产生用启发式方法，小波函数为 db5，$j=4$。对于阈值去噪后的有向加权图邻接矩阵 $\hat{\boldsymbol{G}}_{n\times n}$，SNR=10 是一个分界线，当 SNR≥10 时，需分两步进行数据的处理；当 SNR<10 时，完成处理第 1 步可以较高正确率划分社团。

第 1 步：将 $\hat{\boldsymbol{G}}_{n\times n}$ 中小于 0.5 的值置 0；

第 2 步：按原邻接矩阵比例还原 $\hat{\boldsymbol{G}}_{n\times n}$。

图 3 给出了带噪的有向加权图社团发现流程，整个去噪的有向加权图团发现算法(Denoising weighted directed community detection algorithm based on NMF，WDNMF)算法包含 3 个部分：小波阈值去噪预处理，对预处理后的数据进一步处理，用 DNMF 算法。

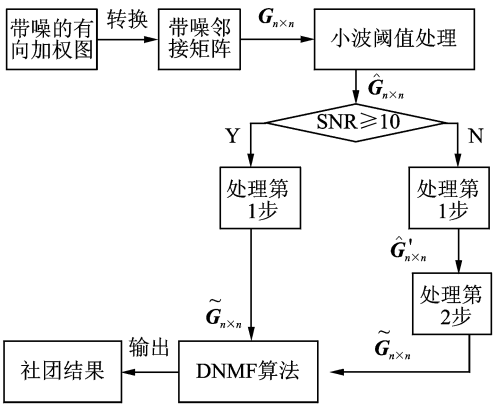

图 3 带噪的有向加权图社团发现流程

Fig. 3 Process of community detection on weighted directed graph with noise

下面给出详细的 WDNMF 算法：

输入：带噪声有向加权邻接矩阵 $\boldsymbol{G}_{n\times n}$，初始化矩阵 $\boldsymbol{X}_{n\times k}$，对称矩阵 $\boldsymbol{W}_{k\times k}$，HT，Heursure，db5，$j=4$

输出：社团结构

(1)if 10≤SNR≤20，小波去噪得到 $\hat{\boldsymbol{G}}_{n\times n}$

then 第 1 步，得到 $\tilde{\boldsymbol{G}}'_{n\times n}$，第 2 步：还原得到 $\tilde{\boldsymbol{G}}_{n\times n}$；

else if SNR<10，小波去噪得到 $\hat{\boldsymbol{G}}_{n\times n}$ then 第 1 步得到 $\tilde{\boldsymbol{G}}_{n\times n}$；

(2)初始化 $\boldsymbol{X}$ 和 $\boldsymbol{W}$，且 $\boldsymbol{A}=\tilde{\boldsymbol{G}}_{n\times n}+\tilde{\boldsymbol{G}}_{n\times n}^{\mathrm{T}}$；

(3)while $\boldsymbol{X}$ 和 $\boldsymbol{W}$ 不收敛 do

更新 $\boldsymbol{X}$：

令：$\text{temp}=\dfrac{\boldsymbol{AXW}-\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}}{2\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{XW}+\boldsymbol{XN}}$，$\Delta x=x^{j}.*\text{temp}$

if $x^{j}+\Delta x>1$ then $x^{j+1}=1$

else：$x^{j+1}=x^{j}+\Delta x$

更新 $\boldsymbol{W}$

令 $\text{temp}=\dfrac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{AX}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{X}}{2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWX}^{\mathrm{T}}\boldsymbol{X}}$，$\Delta \boldsymbol{w}=\boldsymbol{w}^{j}.*\text{temp}$

if $\boldsymbol{w}^{j}+\Delta w>\text{diag}(w_{ii})$

then $\boldsymbol{w}^{j+1}=\min(\text{diag}(w_{ii}))$

else $\boldsymbol{w}^{j+1}=\boldsymbol{w}^{j}+\Delta \boldsymbol{w}$

end while

(4)获得社团结构

文献[12]中证明了其收敛性。

4　实验结果分析

4.1　实验环境和真实数据集

实验环境：操作系统为 Windows 8，CPU 为 I5 处理器，工具软件为 Matlab。

本文在两个真实数据集上对比了 DNMF 算法性能，数据集如表 3 所示，对两个真实数据集上加上不同程度的噪声，验证了 WDNMF 算法的处理性能。

表 3　真实数据集

Table 3　Real datasets

数据集	结点数	边数	属性
Lesmis	77	254	有向加权
Neural network	297	234 5	有向加权

Neural network 来自 C. Elegans 研究的真实数据，是一种生物学中神经网络数据，每个节点代表一个神经元，边表示神经元之间的连接关系。

Lesmis 来自维克多雨果的小说《悲惨世界》。每一个节点有标签，表示一个人物；节点间连线表示人物在同一个场景中出现。

4.2　结果分析

表 4～6 统计了 10 次实验的带噪数据、基础小波去噪、处理第 1 步和第 2 步后的社团划分性能的平均值，带噪数据是指不做任何去噪处理，直接用本文的 DNMF 算法划分社团。

表 4 中统计社团错分率，处理第 2 步在 SNR≥15 时，错分率很低，SNR＝15 时两个数据的社团划分准确度最低为 96%和 97%；当 SNR＜15 时，完成小波去噪后，仅用第 1 步处理在 SNR＝3 时的社团划分准确度最低为 90%和 88%，高于处理第 2 步。

表 4　算法在 Lesmis 和 Neural network 数据集上的社团错分率　%

Table 4　Misclassification rate of algorithms on Lesmis and neural network dataset

SNR	Lesmis				Neural network			
	未去噪	小波去噪	第 1 步	第 2 步	未去噪	小波去噪	第 1 步	第 2 步
20	64	45	5	0	48	23	7	0
15	70	56	8	4	75	30	6	3
10	58	45	7	18	77	40	6	8
5	68	35	6	16	92	38	4	13
4	65	36	7	15	86	47	9	16
3	70	43	10	18	77	40	12	19

表 5　算法在 Lesmis 和 Neural network 数据集上的模块度值(Q)

Table 5　Modularity of algorithms on Lesmis and Neural network dataset

SNR	Lesmis				Neural Network			
	未去噪	小波去噪	第 1 步	第 2 步	未去噪	小波去噪	第 1 步	第 2 步
20	0.298 2	0.258 1	0.301 6	0.484 7	0.273 9	0.274 3	0.303 6	0.452 5
15	0.241 7	0.253 3	0.308 7	0.312 0	0.262 7	0.271 1	0.270 0	0.347 8
10	0.257 4	0.259 2	0.301 2	0.292 3	0.212 3	0.237 1	0.372 8	0.302 2
5	0.277 6	0.287 5	0.332 1	0.312 3	0.252 3	0.263 4	0.362 9	0.284 6
4	0.245 4	0.267 8	0.311 5	0.245 4	0.285 4	0.277 0	0.301 3	0.245 4
3	0.219 3	0.227 4	0.307 3	0.242 1	0.229 3	0.237 4	0.317 7	0.251 2

表 6 算法在 Lesmis 和 Neural network 数据集上的运行时间 s

Table 6 Running time of algorithms on Lesmis and Neural network dataset

SNR	Lesmis				Neural network			
	未去噪	小波去噪	第 1 步	第 2 步	未去噪	小波去噪	第 1 步	第 2 步
20	0.174 9	0.200 7	0.224 2	0.294 3	2.382 1	2.411 1	2.929 0	3.126 2
15	0.193 2	0.231 2	0.287 2	0.379 8	2.392 2	2.462 3	2.807 8	3.879 4
10	0.182 3	0.215 9	0.351 9	0.424 1	2.218 0	2.719 4	3.376 4	4.016 3
5	0.262 8	0.302 3	0.367 2	0.391 0	2.710 0	2.990 0	3.278 0	4.184 7
4	0.250 0	0.273 3	0.342 2	0.398 7	2.829 2	2.983 5	3.422 0	4.387 0
3	0.314 6	0.389 9	0.403 4	0.426 5	3.019 1	3.174 6	3.532 4	4.526 5

表 5 中统计了模块度值(Q),在 SNR=3 时,处理第 2 步的 Q 分别为 0.484 7 和 0.452 5;SNR=3时,处理第 1 步的 Q 分别为 0.307 3 和 0.297 7;可以看出保证了社团划分的正确率,Q 也相应地提高。

表 6 中统计了运行时间,处理第二步在算法运行过程中,耗费的时间最多,但在噪声较小的情况下能过获得高准确率的社团划分结果。

综合来讲,在 SNR=15 时,Lesmis 数据集耗时增加 48%,错分率降低 66%,模块度提高 29%;Neural network 数据集在耗时增加 45%,错分率降低 72%,模块度提高 32%。在 SNR=3 时,Lesmis 数据集耗时增加 28%,错分率降低 60%,模块度提高 40%;Neural network 数据集耗时增加 17%,错分率降低 65%,模块度提高 38%。

5 结束语

本文提出了一种新型可去噪的社团发现算法,该算法采用小波去噪和非负矩阵的混合模型,通过对去噪后的矩阵重构,得到近似邻接矩阵,然后进行基于非负矩阵分解的社团发现,不仅可以发现无向图的社团结构也可以用在有向加权图上。该算法能以较高的正确率划分社团结构,去噪后的划分有效地克服了大部分算法无法划分有向加权图的缺陷。在模拟数据集上实验验证了一般算法对带噪数据处理的局限性;在真实数据集上实验表明本文算法性能优越,具有较好的实际应用价值。

参考文献:

[1] 吕天阳,谢文艳,郑纬民,等. 加权复杂网络社团的评价指标及其发现算法分析[J]. 物理学报, 2012, 61(21): 210511-210511-3.
Lü Tianyang, Xie Wenyan, Zheng Weiming, et al. Analysis of community evaluation criterion and discovery algorithm of weighted complex network [J]. Acta Physica Sinica, 2012, 61(21): 210511-1-210511-3.

[2] 骆志刚,丁凡,蒋晓舟,等. 复杂网络社团发现算法研究新进展[J]. 国防科技大学学报, 2011, 33(1): 47-52.
Luo Zhigang, Ding Fan, Jiang Xiaozhou, et al. New progress on community detection in complex networks[J]. Journal of National University of Defense Technology, 2011, 33(1): 47-52.

[3] Vincent D B, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10(8):1-12.

[4] Danielle S B, Mason A P, Nicholas F W, et al. Robust detection of dynamic community structure in networks[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2013, 23(1): 013142-1-01342-5.

[5] Zhang S, Wang R S, Zhang X S. Uncovering fuzzy community structure in complex networks [J]. Physical Review E, 2007, 76(4): 046103-1-046103-10.

[6] Fei W, Tao L, Xin W, et al. Community discovery using nonegative matrix factorization[J]. Data Mining and Knowedge Discovery, 2011, 22: 493-521.

[7] Nguyen N P, Thai M T. Finding overlapped communities in online social networks with nonnegative matrix factorization [C] // MILCOM 2012: Military Communications Conference. Orlando, USA: IEEE, 2012:1-6.

[8] Zhang Y, Yeung D Y. Overlapping community detection via bounded nonnegative matrix tri-factorization[C]// KDD12: Proceedings of the 18rd International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM, 2012:563-570.

[9] Tang J, Xufei Wang, Huan Liu, et al, Integrating social media data for community detection[J]. Lecture Notes in Computer Science: Modeling and Mining Ubiquitous Social Media, 2012, 7472:1-20.

[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2):69-84.

[11] 杨智,罗国,袁芳芳. 基于平稳小波变换的膈肌肌电信号降噪 [J]. 数据采集与处理,2013,28(5):546-552.
Yang Zhi, Luo Guo, Yuan Fangfang. EMGdi denoising based on stationary wavelet transform [J]. Journal of Data Acquisition and Processing, 2013,28(5): 546-552.

[12] Wang Dingding,Li Tao, Zhu Shenghuo, et al. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization [C] // Proceeding of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. USA:ACM,2008:307-314.

作者简介:张梁梁(1989-),女,硕士研究生,研究方向:智能信息处理,E-mail:vermouthlove@hotmail.com;潘志松(1976-),男,教授,研究方向:模式识别、机器学习;李国鹏(1983-),男,博士生,研究方向:模式识别、机器学习;胡谷雨(1963-),男,教授,研究方向:智能信息处理、云计算、网络管理。