文章编号：2095-2163(2022)02-0083-08 中图分类号：TP183;O157.5 文献标志码：A

重叠社区检测及其果蝇视觉进化神经网络

罗 兰，张著洪
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：针对社区发现中，部分节点划分难的问题，探讨重叠社区检测的优化模型和求解的视觉进化神经网络。模型通过设计节点隶属度矩阵和节点分割规则，建立以模糊分割阈值为变量，且能评估社区划分效果的改进型模块度函数；算法设计中，以候选解构成的状态矩阵对应函数值矩阵作为输入，依据果蝇视觉系统的信息处理机制，建立以输出作为状态学习率的果蝇视觉前馈神经网络，进而借助灰狼优化的位置更新规则，设计状态更新策略，获得基于重叠社区检测的果蝇视觉进化神经网络及其算法。该神经网络的计算复杂度，由状态矩阵的大小及社区网络的节点数确定。比较性的数值实验显示，该求解重叠社区检测问题具有明显优势，有较好的应用潜力。

关键词：模糊聚类；重叠社区检测；果蝇视觉神经网络；灰狼优化；状态更新

Overlapping community detection and fly visual evolutionary neural network

LUO Lan, ZHANG Zhuhong
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Aiming at the difficulty of division of some nodes in community discovery, this work probes into both the overlapping community detection optimization model and the related fly visual evolutionary neural network. In the design of the model, an improved modularity function, which takes fuzzy segmentation thresholds as variables, is proposed to evaluate the divisional effect of community division, relying upon a membership matrix and a segmentation rule. In the design of the algorithm, a fly visual evolutionary neural network (FVENN) is developed to solve the model. Therein, the input is a function-valued matrix matched with a state or candidate matrix, and meanwhile an improved fly visual feedforward neural network is designed to output the learning rate of each element in the state matrix by means of the information processing mechanism of the fly's visual system. Hereafter, each state is updated based on its learning rate and a position update rule of grey wolf optimization. FVENN's computational complexity is determined by its input size and the number of nodes in the community network. Comparative experiments validate that FVENN outperforms the compared approaches and is of great potential to solving the problem of overlapping community detection.

【Key words】 fuzzy clustering; overlapping community detection; fly visual neural network; grey wolf optimization; state update

0 引 言

随着科技的快速发展，诸如社交网络、移动通信、博客、微信等已成为社会进步不可或缺的部分。如何从庞大网络中有效发现节点间的关系模式和社区结构，使得网络健康运行，已成为公安、网监等部门关注的现实问题，也是社区检测或社区发现研究所关注的重要科学问题。社区检测是一种将复杂网络分割为若干社区网络的聚类方法，即将具有相同特性的网络节点归并入同一社区，各社区之间连接稀疏[1]，其中包含非重叠和重叠两种类型社区检测。前者的每个节点仅属于一个社区，而后者的部分节点可属于多个社区。由于现实网络中同一节点通常具有多重属性，各节点间的连接也具有多样性，因此重叠是大多数真实网络的重要特征。由于节点对于不同社区的隶属程度存在较大差异，导致重叠节点与社区之间的隶属关系具有模糊性，加之真实网络中并没有严格的社区划分界限，使得研究社区网络结构的划分变得较为困难。2002 年，Newman[1]等针对非重叠社区检测问题，首次提出社区检测 Girvan - Newman (GN) 算法；2005 年，Pallaet[2]等针对重叠社区检测，提出派系过滤算法(Cluster Percolation method, CPM)；2011 年 Gregory 等[3]首次提出模糊重叠划分的概念等等。此后，一些学者就重叠社区检测展开一系列研究，获得的算法可概括为两种类型，即模糊聚类法[4-5]及基于进

基金项目：国家自然科学基金(62063002;61563009)。

作者简介：罗 兰(1996 -)，女，硕士研究生，主要研究方向：智能优化；张著洪(1966 -)，男，博士，教授，博士生导师，主要研究方向：数据科学与计算智能、深度学习等。

通讯作者：张著洪 Email：zhzhang@ gzu.edu.cn

收稿日期：2021-09-28

化计算的模糊聚类法[6-8]。前者通过构建距离公式来度量节点属于某社区的程度,最后依据相似度进行模糊 C 均值聚类,得到重叠社区结构的划分。但该方法需通过人为设定阈值来确定重叠社区。后者针对原始模糊聚类 FCM (Fuzzy c-means),因初始聚类中心随机确定而导致算法易陷入局部搜索的缺陷,可将进化算法与模糊聚类相结合,设计改进型优化算法。如,孙延维[7]等提出基于粒子群优化的模糊聚类社区检测方法,利用粒子群优化确定最优聚类中心,再利用 FCM 进行社区划分。Wang[8]等提出一种基于粒子群优化的重叠社区检测方法,将 FCM 嵌入粒子群优化算法的粒子选择中,通过最小化模糊 J_m 指数来优化聚类中心,但聚类数需人为指定。

综上,重叠社区检测因网络节点连接的多样性,加之常规模糊聚类法应用于社区划分时,分类阈值不具自适应性,导致算法的自适应能力弱、社区检测的精度低。为此,本文以节点分割的模糊阈值为变量,以社区划分的模块度函数为性能指标,建立重叠社区检测优化模型,进而将果蝇视觉信息处理机制与灰狼优化中位置更新策略结合,提出了求解重叠社区检测问题的果蝇视觉进化神经网络(Fly Visual Evolutionary Neural Network, FVENN)。

1 重叠社区检测模型

1.1 模糊隶属度

将社区网络 G 视为一个无向图,即 $G=(V,E)$。其中,V 是 G 的顶点(节点)构成的集合,即 $V=\{v_1,v_2,\cdots,v_n\}$;E 是 G 的所有边构成的集合,即 $E=\{(v_i,v_j)\mid v_i\in V, v_j\in V, i\neq j\}$,$|E|=m$。$G$ 的邻接矩阵是对称的 n 阶布尔矩阵,即 $\boldsymbol{A}=(A_{ij})^{n\times n}$,$A_{ij}=A_{ji}$,且 $A_{ij}\in\{0,1\}$。节点 i 的度数为 d_i,$1\leqslant i\leqslant n$。社区网络中,任意节点可以隶属于多个社区,利用 $[0,1]$ 连续区间内分布的模糊隶属度,量化重叠节点对不同社区的隶属程度,同一节点对所有社区的隶属度总和为 1。当某节点与某社区的隶属度为 0,则该节点完全不隶属此社区;当隶属度为 1 时,则该节点完全隶属此社区;当隶属度取值介于 0 ~ 1 之间时,隶属度值越大,则该节点属于该社区的隶属程度越高;反之,则越低。在给定图 G 的划分下,用 NC 表示 G 中所有社区的非中心节点构成的集合,即 $NC=\{NC_1,NC_2,\cdots\}$,G 中所有中心节点构成的集合表示为 $CN=\{CN_1,CN_2,\cdots\}$。另外,距离度量是无向图及社区检测中衡量节点之间连接强度的重要指标,在此利用扩散核相似性度量[9-10]、图 G 的连接矩阵和节点度数,设计节点间距离计算模型。具体而言,引入如下关于节点 NC_i 与 CN_j 之间扩散核相似性度量:

$$k_{ij}=e^{-\beta\times L_{ij}} \tag{1}$$

其中,β 为常数;$\boldsymbol{L}$ 为拉普拉斯矩阵,即 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$,$\boldsymbol{L}=(L_{ij})_{|NC|\times|CN|}$;$\boldsymbol{D}$ 是由图 G 中节点的度数构成的 n 阶对角矩阵,其主对角线上位置 (i,i) 处的值是节点 i 的度数,非对角线位置处的元素为 0。由于节点间的扩散核相似性与距离具有负相关关系,因此非中心节点 NC_i 到中心节点 CN_j 之间的距离 $dis(NC_i,CN_j)$ 由式(2)确定:

$$dis(NC_i,CN_j)=\max\{k_{ij},\ 1\leqslant i\leqslant|NC|,\ 1\leqslant j\leqslant|CN|\}-k_{ij} \tag{2}$$

于是,NC_i 节点属于以中心节点 CN_j 为社区 C_j 的模糊隶属度由式(3)计算:

$$U_{ij}=\frac{1}{\sum_{l}^{|CN|}\dfrac{dis(NC_i,CN_j)}{dis(NC_i,CN_l)}^{\frac{2}{x-1}}} \tag{3}$$

其中,x 是模糊加权指数,通常取值为 2。若 x 越大,则算法的聚类效果较差,反之则算法退化为硬 C 均值聚类(HCM)算法。U_{ij} 表示节点 NC_i 属于类别 CN_j 的程度,U_{ij} 较大,则节点 NC_i 属于类别 CN_j 的程度高。由非中心点与社区模糊集确定的隶属度值 U_{ij} 构成的隶属度矩阵表示为 $\boldsymbol{U}=(U_{ij})_{|NC|'\times|CN|}$。依据此隶属度矩阵和非中心节点 NC_i 的分割阈值 $r\text{¢}_{NC_i}$ 进行社区划分,划分规则如下:

规则 I 若 $U_{ij}\geqslant r'_{NC_i}$,则节点 NC_i 属于以 CN_j 为中心的社区 C_j。在此,$r\text{¢}_{NC_i}$ 作为分割值由式(4)确定:

$$r'_{NC_i}=\min_l U_{il}+r_{NC_i}\times(\max_l U_{il}-\min_l U_{il}) \tag{4}$$

其中,r_{NC_i} 为待定的分割阈值参数。

由以上规则可知,若 r_{NC_i} 越小,则节点 NC_i 属于多个社区的几率越大;反之,则仅属于一个社区。

以开源 ENZYMES_g163 网络图为例,取编号为 6、8、10 的节点为中心节点,如图 1 所示。依据式(3)获得各非中心节点与社区模糊集的隶属度矩阵,进而在给定非中心节点的待定参数值 r_{NC_i} 下,由式(4)获得非中心节点的分割值;随后,依据以上规则确定非中心节点所属社区。由该图获知,由于图 G 是一个连通图,加之非中心节点的分割阈值是通过随机方式生成,导致社区分割出现重叠现象。由此可见,分割阈值在社区检测中起到至关重要的作用。

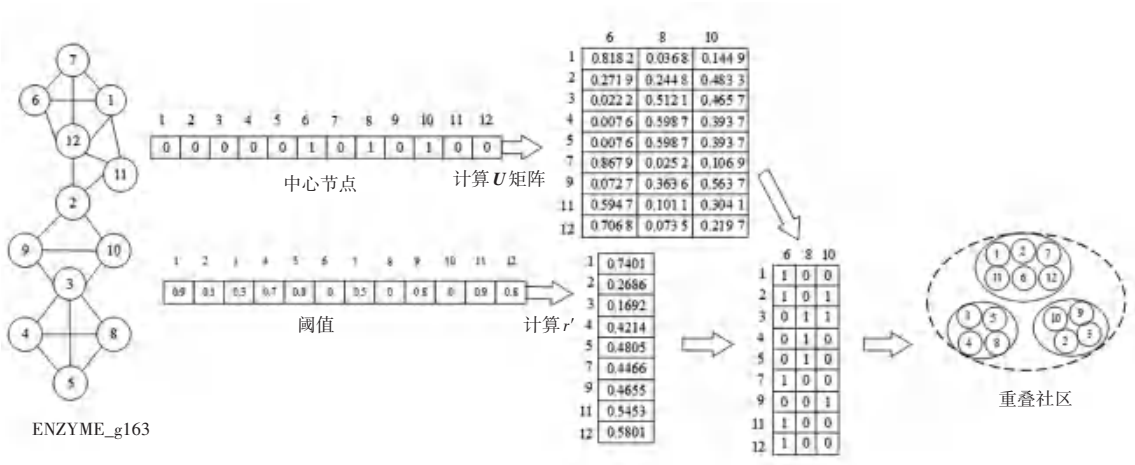

图 1　以 ENZYMES_g163 网络的社区检测事例

Fig. 1　Example of the community detection of the ENZYMES_g163 network

1.2　重叠社区检测优化模型

Newman[11]等提出一种仅适用于非重叠社区检测的模块度函数，其表示为社区网络图的实际边数与节点随机连接下的期望边数之差。随后，Shen[12]等针对重叠社区检测问题，将 Newman 的模块度函数扩展为如下函数 EQ：

$$EQ=\frac{1}{2m}\sum_{j=1}^{K}\sum_{v,w\in C_j}\frac{1}{O_vO_w}\left(A_{vw}-\frac{d_vd_w}{2m}\right) \tag{5}$$

其中，m 是社区网络无向图的边数；K 为社区网络划分的社区数；O_v 是节点 v 所属的社区数；d_v 是节点 v 的度数。

式(5) 表明，若 EQ 值越大，则重叠社区检测的质量越好，即属于多个社区的节点数较少；反之，若 EQ 值较小，则社区网络中出现多个节点属于多个社区，导致不同社区重叠现象严重。可是，一旦 O_v 和 O_w 均大于 1 时，O_vO_w 偏大，进而节点 v 和 w 对 EQ 值的贡献小，因此在社区重叠现象严重情形下，会导致社区检测效果较差。为此，借助 *Sigmoid* 函数，获得如下改进型模块度(Improved EQ，IEQ)函数：

$$IEQ(r_{NC_1},r_{NC_2},\cdots,r_{NC_{|NC|}})=\frac{1}{2m}\sum_{j=1}^{K}\sum_{k,l\in C_j}\delta(r_{NC_k},r_{NC_l})\left(A_{kl}-\frac{d_kd_l}{2m}\right) \tag{6}$$

其中，$\delta(r_{NC_k},r_{NC_l})$ 是节点 k、l 均属于中心节点为 CN_j 的社区 C_j 的权值，其由式(7) 确定：

$$\delta(r_{NC_k},r_{NC_l})=\begin{cases}1 & O_{NC_k}=1,O_{NC_l}=1\\ g(r_{NC_k}) & O_{NC_k}>1,O_{NC_l}=1\\ g(r_{NC_l}), & O_{NC_k}=1,O_{NC_l}>1\\ g(r_{NC_k})g(r_{NC_l}) & O_{NC_k}>1,O_{NC_l}>1,\text{且 } NC_k,NC_l \text{ 无重叠社区}\\ \dfrac{g(r_{NC_k})g(r_{NC_l})}{1+M_{kl}} & O_{NC_k}>1,O(r_{NC_l})>1,\text{且 } NC_k,NC_l \text{ 含重叠社区}\end{cases} \tag{7}$$

其中，M_{kl} 为节点 k，l 属于相同社区的社区数，O_{NC_k} 是在模糊分割阈值 r_{NC_k} 下非中心节点 NC_k 所属的社区数。

式(7) 表明，若节点 NC_k 或 NC_l 所属社区越多，则 $\delta(r_{NC_k},r_{NC_l})$ 越小；特别是，当此节点属于相同社区，则 $\delta(r_{NC_k},r_{NC_l})$ 更小。与式(5) 相比，式(6) 更能刻画含重叠社区的社区划分效果。于是，为了将图 G 分割为多个社区，使得每个社区内节点之间的连接强度大，不同社区之间包含公共节点的数目尽可能小。在此，可将重叠社区检测问题(P)用如下模型加以描述：

$$(\mathrm{P})\ \max_x f(x)=IEQ(x),x=(r_{NC_1},r_{NC_2},\cdots,r_{NC_{|NC|}})$$

$$\text{s.t.},0\leqslant NC_i\leqslant 1,1\leqslant i\leqslant|NC|$$

2　果蝇视觉进化神经网络

给定大小为 M×N 的状态矩阵 $\boldsymbol{A}=(x_{ij})_{M\times N}$，每个状态是问题(P)的候选解，状态 x_{ij} 对应的目标函数值 $f(x_{ij})$ 被视为灰度值，如此状态的灰度值构成大小为 $M\times N$ 的灰度图 $f(\boldsymbol{A})$。第 t 时刻的状态矩阵表示为状态矩阵 $\boldsymbol{A}^{(t)}=(x_{ij}^{(t)})_{M\times N}$，相应的灰度图表示为 $f(\boldsymbol{A}^{(t)})$。FVENN 作为一种求解优化问题的循环神

经网络,以果蝇视觉前馈神经网络在 t 时刻的输出 $\rho(t)$ 作为全局学习率,经由下式更新状态矩阵 $\boldsymbol{A}^{(t)}$ 中的状态:

$$x_{ij}^{(t+1)} = x_{ij}^{(t)} + \Delta(x_{ij}^{(t)}, \rho(t))$$
$$1 \leqslant i \leqslant M, 1 \leqslant j \leqslant N \quad (8)$$

其中,$\Delta(.,.)$ 是经由状态更新策略产生的状态转移增量。

2.1 改进型果蝇视觉神经网络

果蝇视觉系统具有独特的视觉信息处理机制,其主要由光感受器(Photoreceptor)、视网膜(Retina)、薄膜(Lamina)、髓质(Medulla)和小叶(Lobula)5 层构成,各层的信息处理机制存在显著的差异性。基于果蝇视觉系统的信息处理机制,文献[13]针对视觉场景下的碰撞检测与预警问题,获得一种人工果蝇视觉神经网络(Artificial Fly Visual Neural Network, AFVNN),其由 4 个神经层构成。在此,对该神经网络作适当改进后,得到改进型果蝇视觉神经网络(Improved Fly Visual Neural Network, IFVNN)。二者的主要区别在于:

(1)AFVNN 的各层大小依次递减,其 Lamina 层不仅涉及节点的投影和侧抑制操作,也涉及节点的去极化处理。同时 Lobula 层在汇集 Medulla 层中各节点处的运动方向量基础上,基于分流抑制模型产生模型的输出。

(2)IFVNN 的每个神经层均由 $M \times N$ 个节点构成,Lamina 层仅涉及节点的投影和侧抑制操作。IFVNN 的完整设计描述如下,改进型果蝇视觉神经网络结构如图 2 所示。

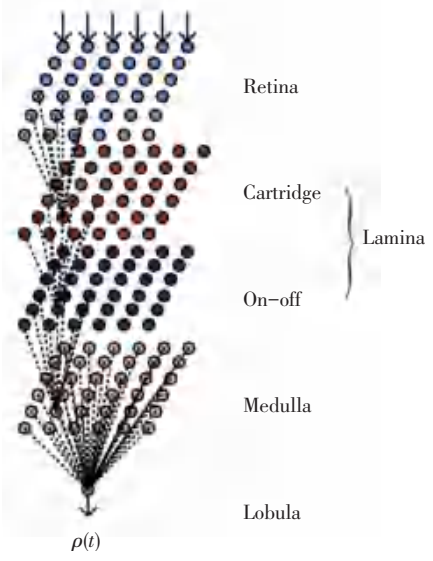

图 2 改进型果蝇视觉神经网络

Fig. 2 Improved fly visual neural network

(1)Retina:在第 t 时刻,节点(i, j) 接收灰度图$f(A^{(t)})$ 中像素点(i, j) 处的光亮强度(即灰度值),并经由式(9)输出光亮强度的变化量。

$$e_{ij}(t) = f(x_{ij}^{(t-1)}) - f(x_{ij}^{(t)}) \quad (9)$$

(2)Lamina:由 cartridge、on-off 两个子层构成,每个子层的节点数均为 $M \times N$。cartridge 层在 Retina 完成扩边处理后,*cart* 节点(i, j) 首先接收 Lamina 层中对应节点的 3×3 邻域节点的膜电位,并经由如下高斯卷积获得输出的膜电位,即

$$cart_{ij}(t) = \frac{1}{w}\sum_{0\leqslant k,l\leqslant 2} w_{kl} e_{i+k-1,\, j+l-1}(t), k, l \neq 1, \quad (10)$$

其中,w 是下列权重矩阵的元数之和,即

$$W = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

w_{kl} 是 W 位置(k, l) 处的元素;cartridge 层作扩边处理后,on-off 层中 *oo* 节点 (i, j) 首先接收 cartridge 层中对应节点的 3×3 邻域内节点的膜电位,进而经由以下侧抑制策略,产生其输出膜电位:

$$oo_{ij}(t) = oo_{ij}(t) - \frac{1}{8}\sum_{-1\leqslant u,v\leqslant 1} oo_{i+u,j+v}(t)$$
$$u + v \neq 0 \quad (11)$$

(3)Medulla:此层由 m_1 节点子层和 m_2 节点子层构成,且各层的节点规模均为 $M \times N$。m_1 节点(i, j) 首先接收 Lamina 层的 *oo* 节点层中节点(i, j) 的 3 × 3 邻域节点的输出,然后基于对称 EMD 运动方向检测器,获得 m_1 节点(i, j) 的输出膜电压,进而经由 AFVNN 的设计,获得 m_2 节点(i, j) 在水平、竖直方向的方向量 $m_{2h}(i,j)$ 和 $m_{2v}(i,j)$。

(4)Lobula:此层首先接收 Medulla 层中所有 m_2 节点在水平、竖直方向的方向量,并经由下式输出其膜电位:

$$\rho(t) = \frac{1}{M+N}\sum_{i=1}^{M}\sum_{j=1}^{N}(m_{2h}(i,j) + m_{2h}(i,j)) \quad (12)$$

2.2 状态更新

基本灰狼优化算法(Grey Wolf Optimization Algorithm, GWO),是由 Mirjalili[14] 等受灰狼捕食行为特性的启发,而提出的启发式随机搜索算法。该算法将优化问题的候选解视为一匹狼,通过狼群捕食猎物过程中不断进行信息交互的行为特性,更新匹狼的位置。在此,将式(12)的输出作为全局学习率,并利用改进型基本灰狼优化中灰狼的位置更新策略,建立 FVENN 的状态矩阵 $\boldsymbol{A}^{(t)}$ 中各状态的更新策略,具体如下:

将 $\boldsymbol{A}^{(t)}$ 中每个状态 $x^{(t)}$ 视为灰狼，$x^{(t)}$ 依据自身历史最好位置 x_{pbest}、全局最优状态 x_{gb} 以及以上 IFVNN 输出的学习率 $\rho(t)$ 更新为 $x^{(t+1)}$，即

$$x_k^{(t+1)} = \begin{cases} y_k + \rho(t) \cdot \lambda(t) \cdot (x_{m,k} - x_k^{(t)}) \cdot \ln\zeta, & \zeta \geqslant 0.5 \\ y_k - \rho(t) \cdot \lambda(t) \cdot (x_{m,k} - x_k^{(t)}) \cdot \ln\zeta, & \text{else} \end{cases} \tag{13}$$

其中，ζ 是 0~1 之间均匀分布的随机数，$\lambda(t)$ 是当前代的自适应扩张系数，即

$$\lambda(t) = (\lambda_{\max} - \lambda_{\min}) \frac{t_{\max} - t}{t_{\max}} + \lambda_{\min} \tag{14}$$

$t_{\max}$ 为最大迭代次数；$\lambda_{\max}$ 和 $\lambda_{\min}$ 为预设的自适应扩张系数的边界值；y 是 x_{pbest} 与 x_{gb} 的线性加权，即

$$y_k = \mu x_{pbest,k} + (1 - \mu) \cdot x_{gb,k} \tag{15}$$

其中，μ 是 $[0,1]$ 上服从均匀分布的随机数，x_m 是 $\boldsymbol{A}^{(t)}$ 中适应度最大的 3 个状态（x_1，x_2 和 x_3）产生的平均状态，即

$$x_m = \frac{x_1 + x_2 + x_3}{3} \tag{16}$$

从智能优化角度，在式（13）中，学习率 $\rho(t)$ 起到全局调节灰狼位置转移幅度的作用；$\lambda(t)$ 引导灰狼逐渐向食物靠近；式（15）使灰狼的位置逐渐介于自身最好位置和食物之间；式（16）使灰狼群中前 3 匹狼始终处于引领作用，引导其它灰狼进行位置更新，且通过灰狼位置信息交互，确保种群位置的多样性。

2.3　FVENN 算法描述

FVENN 由 IFVNN 及状态更新模块构成，以 $M \times N$ 状态矩阵对应的灰度图作为输入，产生当前状态矩阵中所有状态转移的全局学习率。在此学习率引导下，借助以上状态更新策略，使状态矩阵中的状态不断被更新。FVENN 的算法描述如下：

Step 1　参数设置：灰度图大小为 $M \times N$，图 $G=(V,E)$，最大迭代次数为 $t_{\max}$，自适应扩张系数的边界值为 $\lambda_{\max}$ 和 $\lambda_{\min}$；

Step 2　确定中心节点集 K

（**1**）初始化中心节点集 $K \neg \mathrm{f}$；

（**2**）将节点集 V 中节点，按其度数降序排列，首个节点 v 放入 K 中；

（**3**）从 V 中删除 v 及其邻接的节点；

（**4**）返回步（**2**），直到 V 为空；

Step 3　置 $t \neg 1$，随机初始化 $M \acute{} N$ 个候选解（即 $M \acute{} N$ 个模糊阈值向量）构成的初始状态矩阵 $\boldsymbol{A}^{(t)}$，将全局最好状态记作 x_{gb}；

Step 4　在 $[1, |K|]$ 内取随机整数 m，从集合 K 中随机选择 m 个节点作为中心节点，根据式（3）计算隶属度矩阵，依据式（6）计算 $\boldsymbol{A}^{(t)}$ 中各状态的灰度值 $f(x^{(t)})$；

Step 5　根据式（9）~式（12）计算 IFVNN 输出的学习率 $\rho(t)$；

Step 6　$\boldsymbol{A}^{(t)}$ 中所有状态依据式（13）执行状态更新，获得状态矩阵 $\boldsymbol{A}^{(t+1)}$；

Step 7　计算 $\boldsymbol{A}^{(t+1)}$ 对应的灰度图 $f(\boldsymbol{A}^{(t+1)})$，并借助 $\boldsymbol{A}^{(t+1)}$ 中所有状态，更新 x_{gb} 及各自的 x_{pbest}；

Step 8　依据式（4）及规则 I，计算 $\boldsymbol{A}^{(t+1)}$ 对应的灰度图 $f(A^{(t+1)})$；

Step 9　置 $t \neg t+1$，若 $t < G_{\max}$，则转步骤 4；否则，依据式（4）、规则 I 及 x_{gb}，输出图 G 的划分及 $f(x_{gb})$。

经由以上 FVENN 的描述获知，在状态矩阵从 $\boldsymbol{A}^{(t)} \sim \boldsymbol{A}^{(t+1)}$ 转移过程中，$\rho(t)$ 控制状态的变化幅度，引导状态向最优状态转移，提升获得最优社区划分的能力。同时利用对数函数，对状态分量进行随机扰动，以及引入自适应扩张系数 $\lambda(t)$，增强算法的局部勘测能力。

FVENN 的计算复杂度由 IFVNN、状态更新模块以及模块度函数决定。在 IFVNN 中，Retina 层含 MN 次加减法；Lamina 层含 $38MN$ 次运算；Medulla 层含 $34MN$ 次四则运算；Lobula 层共需 $4MN$ 次加减运算；模块度函数需运算 n^2 次（n 为网络节点数）。另外，状态更新模块共需 $5MNn^2$ 次乘除、$4MNn^2$ 次加减、MNn^2 次逻辑运算和 $MN\,n^2 m_p$（m_p 为优化问题中模块度函数的计算次数）次目标函数值运算。因此，FVENN 在一个周期内的运算次数为 $MN((m_p + 10)n^2 + 77)$。由此可知，FVENN 的计算复杂度由 M、N、n、m_p 确定。

3　数值实验

数值实验环境配置在 Windows10（CPU/3.7 GHz，RAM/4.0 GB）/Visual Studio 2013 下展开。为测试 FVENN 求解重叠社区检测问题的性能，选取两种经典重叠社区检测算法 SLPA[15]、NMF[16] 以及基于智能优化的重叠社区检测算法（灰狼优化算法 GWO[14] 及遗传算法 GA[17]）参与比较。测试事例包括真实世界网络和人工合成网络社区检测。各算法均求解每种事例 20 次，最大迭代数为 100。经由参数调试，FVENN 的参数设置为 $M = N = 10$，$\lambda_{\max} = 0.8$，$\lambda_{\min} = 0.6$；参与比较的方法参数设置与

其文献中的参数设置相同。各算法获得的最大模块度以及信息指标值 *ENMI*[18] 被用于比较分析。具体而言,给定网络 G,某给定算法 A 得到的社区划分为 X,Y 是网络 G 的实际划分。于是,$ENMI(X|Y)$ 定义为:

$$ENMI(X|Y) = 1 - \frac{1}{2}[H(X|Y) + H(Y|X)] \quad (17)$$

其中,

$$H(X|Y) = \frac{1}{|X|}\sum_i \frac{\min_j H(X_i|Y_j)}{H(X_i)} \quad (18)$$

式中,X_i 表示社区划分方式 X 下得到的第 i 个社区;$H(X_i)$ 是社区 X_i 的熵;$H(X_i|Y_j)$ 是社区 X_i 相对于社区 Y_j 的条件熵。

ENMI 的值域为[0,1],其值越大,说明算法 A 得到的社区划分还原真实社区划分的程度越高,反之则越低。

事例 1 真实网络

选取开源的 5 种真实世界网络[8](Karate、Dolphin、Polbooks、Football、SFI)作为社区检测事例。这些网络具有不同规模和分布特征(见表 1),可用于测试社区检测方法的性能。以上 5 种算法作用于表 1 中每种真实世界网络后,获得的统计结果见表 2~表 3(注:由于表 1 中 SFI 的社团数量未知,因此在表 3 中各算法不能获得关于 SFI 的 *ENMI* 值;另外,由于 NMF 的算法参数设置的特殊性,执行每个网络的社区检测时,每次运行中获得的结果相同)。

表 1 真实世界网络数据

Tab. 1 Real world network data

网络	节点数	边数	平均度	社团数量
Karate	34	78	4.75	2
Dolphin	62	159	5.13	2
Polbooks	105	441	8.4	3
Football	115	613	10.66	12
SFI	118	200	1.69	–

表 2 算法作用于真实世界网络得到的 *IEQ* 统计值比较

Tab. 2 Comparison of *IEQ* statistics obtained by the algorithm on real world networks

网络	Metric	SLPA	NMF	GA	GWO	FVENN
Karate	max	0.198 184	0.208 109	0.187 962	0.215 069	**0.256 451**
	mean	0.187 845	0.208 109	0.180 872	0.213 922	**0.256 10**
Dolphin	max	0.182 874	0.263 654	0.248 742	0.271 789	**0.342 767**
	mean	0.169 479	0.263 654	0.230 078	0.259 303	**0.339 94**
Polbooks	max	0.263 713	0.259 118	0.325 849	0.232 724	**0.419 771**
	mean	0.257 184	0.259 118	0.314 568	0.229 010	**0.401 849**
Football	max	0.219 027	0.304 927	0.278 456	0.273 991	**0.305 605**
	mean	0.207 584	**0.304 927**	0.270 368	0.267 949	0.291 219
SFI	max	0.332 731	0.376 532	0.359 614	0.341 89	**0.479 94**
	mean	0.327 846	0.376 532	0.347 819	0.341 89	**0.468 769**

表 3 算法作用于真实世界网络得到的 *ENMI* 值比较

Tab. 3 Comparison of *ENMI* statistics obtained by the algorithm on real-world networks

网络	SLPA	NMF	GA	GWO	FVENN
Karate	0.544 838	0.440 456	0.438 975	0.704 481	**1**
Dolphin	0.276 501	0.466 435	0.448 759	0.573 93	**0.770 714**
Polbooks	0.235 809	0.388 018	0.504 318	0.252 308	**0.655 481**
Football	0.406 431	**0.803 261**	0.607 813	0.455 327	0.748 902
SFI	–	–	–	–	–

由表 2~表 3 获知,对于 Karate、Dolphin、Polbooks 及 SFI 每种网络中,FVENN 获得的 *IEQ* 均值、最大值及 *ENMI* 值,整体上均大于其它算法得到的相应值。因此该方法的社区检测效果具有明显优势,表明本文的模块度函数作为重叠社区划分指标,能使不同社区之间包含的公共节点较少。特别是,对于 SFI 网络,其网络节点链接十分稀疏,平均节点度数仅为 1.69,在此情形下,FVENN 的社区检测优势更加突出。此外,对于 Football 网络,FVENN 的 *IEQ* 均值比 *NMF* 的值略小,但前者得到的 *IEQ* 最大

值比后者明显较大，表明 FVENN 经多次运行能得到较好的社区检测方案。其次，参与比较的 4 种算法的社区检测效果并不存在显著差异。相对而言，*NMF* 的社区检测效果较好；GWO 次之；GA 比 SLPA 能获得更好的社区检测效果。因此，表 2 说明，智能优化算法应用于社区检测具有较好的潜力。表 3 进一步证实，FVENN 求解 Karate、Dolphin 及 Polbooks 能获得最好的社区检测效果；但在求解 Football 时，所获效果略逊色于 NMF，其主要原因在于 Football 网络的实际社区数较大（12 个社区），以及 FVENN 运行中，初始中心节点的选取具有随机性。

事例 2　人工合成网络

人工合成网络经由 LFR 基准生成，更贴近真实的社交网络，涉及的参数设置与文献[19]中 LFR 网络的参数设置相同，即节点数为 100；混合参数 μ 取 0.1、0.3、0.5；重叠成员数 O_m 取 2、4、6、8；重叠节点数 O_n 取 $0.1N$、$0.3N$；每个社区的最小节点数 $c_{\min}$ 为 5，最大节点数 $c_{\max}$ 取 12；网络节点的平均度数 k、最大度数 $k_{\max}$ 分别取 10、50；τ_1 和 τ_2 分别为 2 和 1。因此，该人工合成网络由 24 个不同规模和特征的 LFR 网络构成。

将以上 5 种算法作用于这 24 个 LFR 网络，进而依据指标 *ENMI* 获得的 *ENMI* 值与重叠成员数 O_m 的关系，如图 3 所示。

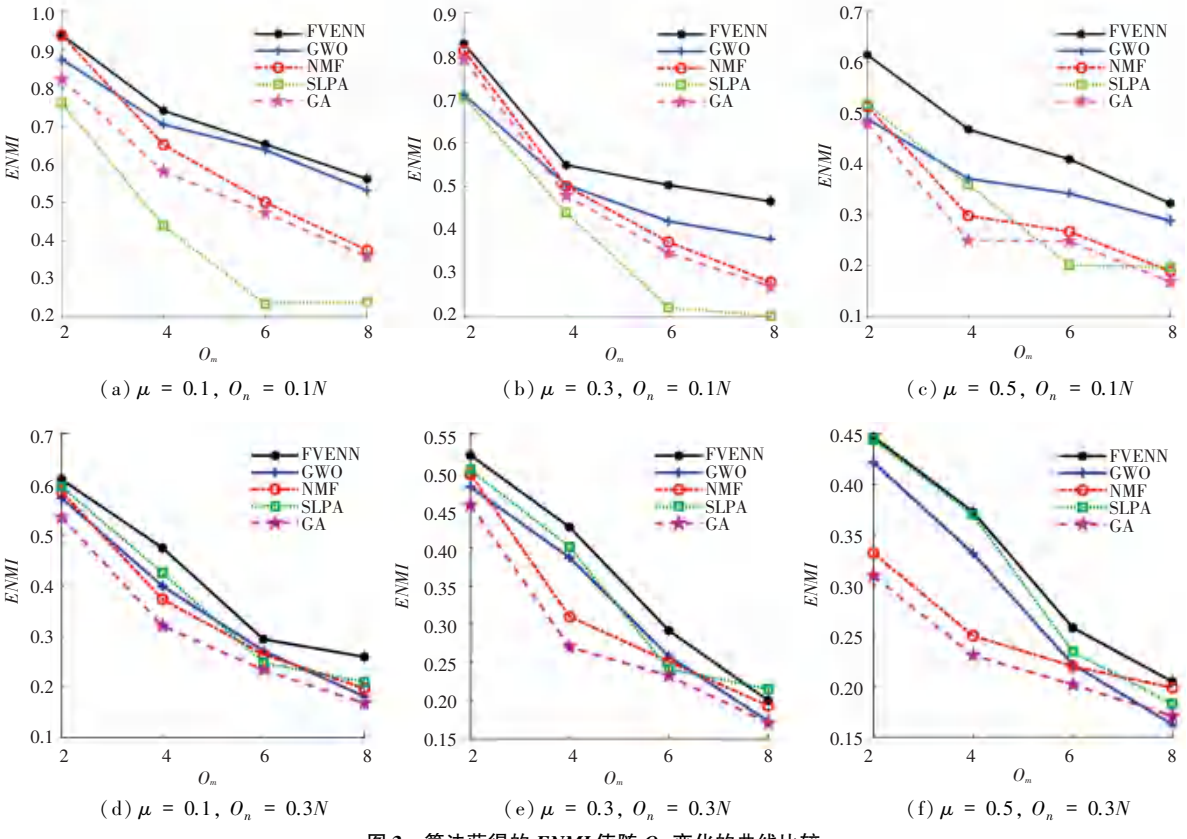

图 3　算法获得的 *ENMI* 值随 O_m 变化的曲线比较

Fig. 3　Curve comparison of *ENMI* value obtained by algorithms with O_m

由图 3 获知，在给定的 O_m 值下，对于此 24 个 LFR 网络中的每个网络，FVENN 获得的 *ENMI* 值，在大多数情况下均比其它算法得到的相应值大，说明该方法的社区检测效果具有明显优势。尽管在图 3(a)和图 3(b)中，当 O_m = 2 时，NMF 与 FVENN 获得的 *ENMI* 非常接近，但随着 O_m 的增大，FVENN 的社区检测效果逐渐突出，得到的 *ENMI* 值比其它算法的值要大。此外，随着 μ 值逐渐增大，各算法得到的 *ENMI* 值相应变小，但 FVENN 的 *ENMI* 值下降较平缓，且得到社区划分效果的优势更明显。由此表明，随着 LFR 网络的重叠节点数增加，FVENN 的社区检测优势更加突出。

4　结束语

重叠社区检测一直是极为困难的科学与工程问题，探讨如何能恰当刻画重叠社区检测特征的性能指标，以及设计快速求解的优化算法，是社区发现研究中关注的重要科技问题。本文在设计社区划分的模糊隶属度矩阵基础上，借助模糊分割阈值参数建立能凸显重叠节点对社区检测效果影响的改进型模

块度函数,进而将重叠社区检测问题描述为含连续变量的函数优化问题。随后,利用果蝇视觉系统的信息处理机制,获得改进型果蝇视觉神经网络,并将其输出与灰狼位置更新规则结合,获得果蝇视觉进化神经网络(FVENN)。比较性实验分析表明,FVENN 求解重叠社区检测问题具有明显优势,搜索效果稳定。另外,虽然 FVENN 检测小规模重叠社区的优势较为突出,但尚未应用于大规模重叠社区的检测问题。未来研究的重点之一将集中大规模情形下的社区检测模型设计,探讨高效求解的算法。

参考文献

[1] GIRVAN M, NEWMAN M E. Community structure in social and biological networks[J]. Proc. Natl Acad. Sci. 2002, 99(12): 7821-7826.

[2] PALLA G, DERANYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[3] GREGORY S. Fuzzy overlapping communities in networks[J]. Journal of Statistical Mechanics Theory and Experiment, 2010, 2(2): 2-17.

[4] ZHANG S, WANG R S, ZHANG X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering[J]. Physica A: Statistical Mechanics & Its Applications, 2007, 374(1): 483-490.

[5] WANG W, LIU D, LIU X, et al. Fuzzy overlapping community detection based on local random walk and multidimensional scaling[J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392(24): 6578-6586.

[6] YI D, XIAN F. Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm[J]. Neurocomputing, 2016, 188(5): 233-238.

[7] 孙延维, 彭智明, 李健波. 基于粒子群优化与模糊聚类的社区发现算法[J]. 重庆邮电大学学报(自然科学版), 2015,27(5): 660-666.

[8] WANG X, LIU G, PAN L, et al. Uncovering fuzzy communities in networks with structural similarity[J]. Neurocomputing, 2016, 210(19): 26-33.

[9] TIAN Y, YANG S, ZHANG X. An evolutionary multi-objective optimization based fuzzy method for overlapping community detection[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(11): 2841-2855.

[10] BINESH N, REZGHI M. Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria[J]. Applied Soft Computing, 2018, 69(8): 689-703.

[11] NEWMAN M E. Modularity and community structure in networks[J]. Proc Nat. Acad. Sci., 2006: 8577-8582.

[12] SHEN H, CHENG X, CAI K, et al. Detect overlapping and hierarchical community structure in networks[J].Physica A, 2009, 388(8):1706-1712.

[13] ZHANG Z H, YUE S G, ZHANG G P. Fly visual system inspired artificial neural network for collision detection [J]. Neurocomputing, 2015, 153(4): 221-234.

[14] Mirjalili S., Mirjalili S. M., Lewis A. Grey Wolf Optimizer[J]. Advances in Engineering Software, 2014:46-61.

[15] XIE J, SZYMANSKI B K, LIU X. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[J], IEEE computer society, 2011(1): 344-349.

[16] YANG J, LESKOVEC J. Overlapping community detection at scale: A nonnegative matrix factorization approach[C]// ACM: International Conference on Web Search & Data Mining, 2013: 587-596.

[17] TSUNG C K, HO H J, CHEN C Y, et al. Detecting overlapping communities in modularity optimization by reweighting vertices [J]. Entropy, 2020, 22(8):819-838.

[18] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2009, 80(1): 16-118.

[19] LANCICHINETTI A, FORTUNATO S. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2012,11(3):19-44.

(上接第 82 页)

4 结束语

本文通过对某石化企业原始数据进行处理,将得到预处理后的数据降维,建立基于随机森林的 RON 损失预测模型,对 RON 损失及其指标进行预测,通过预测值曲线与真实值曲线的对比,发现其预测结果接近于真实值,说明预测模型有效。

运用遗传算法优化主要变量,经过多次迭代优化后,最终完成了降幅超过 15%的优化目标。本文基于随机森林的汽油精制过程中辛烷值损失模型为中国车用汽油质量升级的关键技术及其深度开发提供了可靠依据。

参考文献

[1] 杜明洋,张甜甜,薄其高,等. 汽油精制过程中的辛烷值损失预测模型[J]. 齐鲁工业大学学报,2021,35(1):73-80.

[2] 徐心怡,贺兴,艾芊,等. 基于随机矩阵理论的配电网运行状态相关性分析方法[J]. 电网技术,2016,40(3):781-790.

[3] 王伟同,范海东,梁成思,等. 基于随机森林算法的对冲锅炉出口 NO_(x)排放量预测模型研究[J/OL]. 热力发电:1-9[2022-01-11].

[4] 李超,王杰,史运涛,等. 基于遗传算法的汽油调和优化系统[J]. 工业控制计算机,2018,31(10):79-81.