doi:10.3969/j.issn.1000-436x.2014.z2.012

ESYN：基于动态模型的高效同步聚类算法

董学文[1,2]，杨超[1,2]，盛立杰[2]，马建峰[1,2]
(1. 西安电子科技大学 计算机网络与系统安全陕西省重点实验室，陕西 西安 710071)
2. 西安电子科技大学 计算机学院，陕西 西安 710071）

摘 要：基于动态同步模型，提出一种高效同步聚类 ESYN 算法。首先，根据非矢量网络的局部结构信息，提出节点相似度的定义，以准确描述节点间的链接密度；其次，利用 OPTICS 算法进行矢量化预处理，将非矢量网络转换为一维坐标序列；最后，在通用 Kuramoto 动态同步模型中，增加基于全局信息的耦合强度分析，同时不断增加同步半径，自动选取最优的聚类结果。在大量人工合成数据集和真实数据集上的实验结果表明算法聚类准确率较高。

关键词：聚类；同步模型；矢量化；模块度
中图分类号：TP181 **文献标识码**：A **文章编号**：1000-436X(2014)Z2-0086-08

ESYN: efficient synchronization clustering algorithm based on dynamic synchronization model

DONG Xue-wen[1,2], YANG Chao[1,2], SHENG Li-jie[2], MA Jian-feng[1,2]
(1. Shaanxi Key Laboratory of Network and System Security, Xidian University, Xi'an 710071, China;
2. School of Computer Science and Technology, Xidian University, Xi'an 710071, China)

Abstract: Clustering is an important research field in data mining. Based on dynamical synchronization model, an efficient synchronization clustering algorithm ESYN is proposed. Firstly, based on local structure information of a non-vector network, a new concept vertex similarity is brought up to describe the link density between vertices. Secondly, the network is vectoried by OPTICS algorithm and turned into one-dimensional coordination sequence. Finally, global coupling analysis is applied to generalized Kuramoto synchronization model, synchronization radius is increased and the optimal clustering result is automatically selected. The experimental results on a large number of synthetic and real-world networks show that proposed algorithm achieves high accuracy.

Key words: clustering; synchronization model; vectorization; modularity

1 引言

聚类是数据挖掘领域中一种重要的分析技术[1,2]，根据数据之间在预先制定的属性上的相似性聚集成簇。在过去 10 年中，数据聚类吸引了研究人员的广泛关注，并提出一系列的聚类算法。这些算法可以分为如下几类：基于密度的聚类算法、基于图论的聚类算法、基于模块度优化的聚类算法等。

DBSCAN[3]是一种代表性的基于密度的聚类算法，可在含噪声的空间数据集中快速发现密度超过

收稿日期：2014-06-25
基金项目：长江学者和创新团队发展计划基金资助项目（IRT1078）；国家自然基金委员会—广东联合基金重点基金资助项目（U1135002）；国家科技部重大专项基金资助项目（2011ZX03005-002）；国家自然科学基金青年基金资助项目（61303219）；陕西省自然科学基金资助项目（2014JQ8297, 2014JQ8295）；中央高校基本科研业务费专项基金资助项目（JY10000903006, K5051303007）
Foundation Items: The Program for Changjiang Scholars and Innovative Research Team in University (IRT1078); The Key Program of NSFC-Guangdong Union Foundation (U1135002); The Major National S & T Program (2011ZX03005-002); The National Natural Science Foundation of China(61303219); The Natural Science Basic Research Plan in Shaanxi Province(2014JQ8297, 2014JQ8295); The Fundamental Research Funds for the Central Universities (JY10000903006, K5051303007)

给定阈值的任意形状聚类。但是，它把参数 Eps 和 MinPts 的设置任务留给用户，且算法对参数 Eps 较为敏感。基于图论的 Chameleon[4]聚类算法将矢量数据建模为图，通过引入互连性和近似性 2 个指标来控制簇的分裂和合并，可以发现高质量的任意形状聚类。2004 年 Newman 和 Grivin 提出模块度函数[5]用以评估社团聚类质量。模块度定义为簇内实际连接数目与随机连接情况下簇内期望连接数目之差,用来定量地刻画网络簇结构的优劣。研究人员提出一些基于模块度优化的算法，如 FastModularity[6]和 LHLC[7]。

近年来，研究人员开始利用同步技术进行聚类算法的研究。同步是自然、社会、工程中普遍存在的现象，表现为不同的进程对于时间的一致性。例如，萤火虫的同步发光现象和心脏起搏细胞的同步收缩现象。研究人员提出一些能有效捕捉同步动力过程的模型，如广义 Kuramoto 模型[8,9]。受同步现象启发，Böhm 等[10]提出了一种基于同步原理的聚类算法 Sync，利用同步动力模型来探测数据集中的聚类。给定一个邻域半径，一个对象在以自身为圆心的一个超球形邻域内的所有邻居对象的同步作用下产生位移。在非线性作用力的影响下，相近的对象将会同步达到相同的相位并形成聚类。然而 Sync 算法仅能对矢量网络数据进行分析，并且算法运行时间比较长。黄建斌等针对非矢量网络数据，提出一种快速局部同步聚类算法 SYN[11]。首先使用 OPTICS 算法[12]，将网络中节点对象按照链接密度关系进行排序，从而对网络数据矢量化。通过不断扩大节点同步的邻域半径，可以得到不同分辨率的多种社团划分结果，同时结合社团模块度函数自动选择最佳聚类结果。

但是，分析发现 SYN 算法对于节点的相似性定义不够精确，并且在没有考虑节点间的耦合强度差异。据此，本文提出一种基于局部同步的高效聚类 ESYN 算法。

本文的创新点主要如下：1) 提出一种节点相似度的定义，利用局部结构信息，精确描述非矢量网络中的链接密度，提高网络矢量化结果的准确性；2) 在对节点相似度的分析过程中，引入熵值计算，并对比 SYN 算法中的结构相似度定义，指出节点相似度定义的有效性；3) 将节点间的全局耦合强度引入网络同步模型，提高社团划分的准确度，对社团本身结构特点不作要求，可以发现任意形状的社团。

2　基于广义 Kuramoto 同步模型的 SYN 聚类

本节介绍文献[11]中广义 Kuramoto 同步模型，以及基于该模型的 SYN 聚类算法。表 1 给出本文使用的符号说明。

表 1　符号说明

符号	说明
D	网络数据集
N	网络节点数
x	数据集 D 中一个节点
l_x	节点 x 映射后的坐标
$l_x(t)$	时间 t 时节点 x 的坐标
ε	同步半径
$\tau(x)$	节点 x 的邻域，包含 x 和 x 的邻居节点
$dist(x,y)$	节点 x,y 之间的距离
$degree(x)$	节点 x 的度
$N_\varepsilon(x)$	节点 x 的 ε-邻域
D_ε^t	当同步半径为 ε，时间为 t 时的数据集
C_ε^t	当同步半径为 ε，时间为 t 时的聚类结果
Q_ε^t	当同步半径为 ε，时间为 t 时的模块度

2.1　广义 Kuramoto 同步模型

定义 1　(结构相似度 S_{xy})给定无向网络 $G=\{V,E,w\}$，对于 G 中任意一对节点 (x,y) 定义它们之间的结构相似度 S_{xy} 如下。

$$S_{xy}=\frac{\sum_{z\in\tau(x)\cap\tau(x)}w(x,z)w(y,z)}{\sqrt{\sum_{z\in\tau(x)\cap\tau(y)}w(x,z)^2}\sqrt{\sum_{z\in\tau(x)\cap\tau(y)}w(y,z)^2}} \tag{1}$$

其中，$\tau(x)$ 是由节点 x 以及 x 的邻接节点构成的集合，$w(x,z)$ 是节点 (x,z) 之间边上的权值。若 G 为无权图，则 G 中所有边上的权值均默认为 1，则式(1)可简化为

$$S_{xy}=\frac{|\tau(x)\cap\tau(y)|}{\sqrt{|\tau(x)||\tau(y)|}} \tag{2}$$

结构相似度 S_{xy} 用来计算网络中任意 2 个相邻节点 (x,y) 间的链接密度。

定义 2　(ε–邻域)对象 x 的 ε–邻域是到对象 x 的距离小于等于 ε 的所有对象组成的集合 $N_\varepsilon(x)$。

$$N_\varepsilon(x)=\{y\in X\,|\,dist(y,x)\leqslant\varepsilon\} \tag{3}$$

其中，$dist(x,y)$是距离度量函数。

定义 3　(广义 Kuramoto 模型）广义 Kuramoto 模型描述了基于平均场耦合的大量周期振子的动力学：相位振子采用独立的频率运动，并在振子之间的非线性频率吸引下趋于耦合。n 个周期振子相互作用时的同步情况如式(4)所示。

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t}=\omega_i+\frac{C}{n}\sum_{j=1}^{n}\sin(\theta_j-\theta_i) \tag{4}$$

其中，i=1,⋯,n, ω_i, θ_i 分别为第 i 个节点的频率和相位，C 为是各振子之间的耦合强度。

广义 Kuramoto 模型将网络中每个对象都看作一个独立的相位振子，并在其 ε–邻域内进行同步。设 x 为网络中的一个数据对象，l_x 为对象 x 经过矢量化后的坐标值，l_x 的变化过程如下

$$\frac{\mathrm{d}l_x}{\mathrm{d}t}=\omega+\frac{C}{d}\sum_{y\in N_\varepsilon(x)}S_{xy}\sin(l_y-l_x) \tag{5}$$

其中，d 为 x 节点在 ε–邻域内邻居节点的个数，令 $\mathrm{d}t=\Delta t$，则

$$l_x(t+1)=l_x(t)+\Delta t\omega+\frac{\Delta tC}{d}\sum_{y\in N_\varepsilon(x)}S_{xy}\sin(l_y-l_x) \tag{6}$$

所有节点对象都有一个独立的频率 ω，由于不同对象的频率差异会干扰甚至阻止社团的形成，但对聚类的结果没有影响。所以 $\Delta t\omega$ 这一项可以忽略。ΔtC 是一个常数，为了简便将它设置为 1，得到式(7)。

$$l_x(t+1)=l_x(t)+\frac{1}{d}\sum_{y\in N_\varepsilon(x)}S_{xy}\sin(l_y-l_x) \tag{7}$$

2.2　SYN 算法

基于广义 Kuramoto 同步模型，黄建斌等在文献[11]中提出一种快速同步聚类算法 SYN。算法主要分为 2 步。

预处理　使用 OPTICS 算法，根据式(1)代表的节点链接密度关系，将各节点进行排序，排序结果保证链接密度大的节点距离较近。OPTICS 算法经常作为数据预处理算法，处理结果供其他算法使用。然后根据排序结果将各节点均匀映射到一维坐标区间[0,1)上，得到一个一维坐标序列。

同步聚类　根据式(7)计算每个对象与其–邻域内进行同步调整。对调整坐标后的所有节点重新进行社团划分，将距离小于 ε 的节点判定为同一个社团。得到社团划分结果后，计算其模块度。在不断增加邻域半径 ε 值的同步过程中，得到一系列聚类结果，选择其中模块度最大的作为最优聚类结果。

3　同步模型优化

本节首先指出 SYN 算法中不足之处，针对不足，制定相应改进方法，提出一种高效的同步算法 ESYN(efficient synchronization algorithm)。

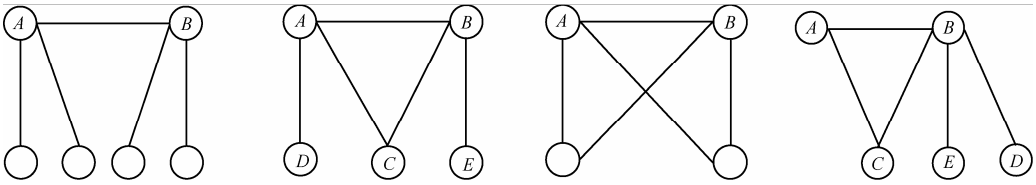
3.1　节点相似度

SYN 算法中提出结构相似度定义（见式(1)），通过网络局部结构信息描述网络间的链接密度关系。SYN 算法使用 OPTICS 算法和结构相似度差异对节点进行排序，并且在同步过程中使用结构相似度作为同步系数（见式(7)）。相似度定义准确与否，对排序结果和同步聚类过程产生重大影响。

分析发现，结构相似度的定义不够准确。举例说明如下：若节点 A 与节点 B 相连，并且 $|\tau(A)|=|\tau(B)|=4$ 的网络有 3 种情况，如图 1(a)、图 1(b)，图 1(c)所示。

根据式(1)，图 1(a)~图 1(d)中节点 A、B 间的结构相似度值分别是 1/2,3/4,1 和 $3/\sqrt{15}\approx 0.77$，这 4 个相似度值均在区间[1/2,1]。该区间相对来说是比较小的区间，2 个相似度值也比较接近，导致在排序和同步过程中，很难将链接密度区分开来，排序和同步效果就不明显。另外，当 $|\tau(A)|=|\tau(B)|=4$，且节点 A、B 相连时，仅有图 1(a)、图 1(b)和图 1(c) 3 种情况。这 3 种情况下，图 1(a)中链接密度最小，$|\tau(A)|$ 与 $|\tau(B)|$ 间仅存在一条边。可图 1(a)中 A、B 间结构相似度值为 1/2，而该值代表的链接密度显然偏大。

对相似度进行重新定义，以更好体现网络中的



(a) 网络 1　(b) 网络 2　(c) 网络 3　(d) 网络 4

图 1　4 个无向网络

链接密度。那么，相似度具体和哪些因子有关呢？

首先，显然图 1(c)中节点 A、B 间的链接密度比图 1(a)、图 1(b)要大，而 3 图 $|\tau(A)|=|\tau(B)|$ 均等于 4，3 图中 $|\tau(A)|$ 与 $|\tau(B)|$ 最大区别是间重复节点比例，即 $\dfrac{|\tau(x)\cap\tau(y)|}{|\tau(x)\cup\tau(y)|}$ 值。因此，$|\tau(A)|$ 与 $|\tau(B)|$ 重复节点比例直接影响链接密度大小。

定义相似度因子 1 如下。

对于带权图，$w(x,z)$ 节点 x,z 之间边上权值，则

$$Factor1(x,y)=\frac{\sum_{z\in\tau(x)\cap\tau(y)}w(x,z)w(y,z)}{\sum_{z\in\tau(x)\cup\tau(y)}w(x,z)w(y,z)} \tag{8}$$

对于无权图

$$Factor1(x,y)=\frac{|\tau(x)\cap\tau(y)|}{|\tau(x)\cup\tau(y)|} \tag{9}$$

其次，当相似度因子 $Factor1$ 相同时，如何进一步定义链接密度。如图 1(b)与图 1(d)，$|\tau(A)|\cap|\tau(B)|$ 值均等于 3，$|\tau(A)|\cup|\tau(B)|$ 均等于 5，$Factor1(A,B)$ 值均为 3/5。显然图 1(d)与图 1(b)节点 A、B 间链接密度不尽相同，但如何进一步区分链接密度呢？

图 1(b)与图 1(d)中，集合 $|\tau(A)|$ 与集合 $|\tau(B)|$ 的并集和交集均相同，不同的是非交集节点分布情况不同。图 1(b)中，非交集节点 D、E 分散分布，即分别与节点 A、节点 B 相连。图 1(d)中，非交集节点 D、E 集中分布，均与节点 B 相连。在定义相似度前，需要比较图 1(b)与图 1(d)哪个图中节点 A、B 链接密度较大。

引入熵值计算，比较图 1(b)与图 1(d)链接密度。熵（entropy）用于测量系统中的混乱程度，具体计算见式(10)。

$$H(X)=-\sum_{i=1}^{n}p(X_i)\ln p(X_i) \tag{10}$$

其中，$H(X)$ 表示熵，$p(X_i)$ 表示 X 系统中 X_i 出现的概率。

假设每个节点概率相同，图 1(b)中集合 $|\tau(A)|\cup|\tau(B)|$ 可分为 3 个部分 $part1=\{A,B,C\}$，$part2=\{D\}$，$part3=\{E\}$，则 $p(part1)=3/5$，$p(part2)=1/5$ 和 $p(part3)=1/5$，根据式(10)得出熵值为 0.95。同理，图 1(d)中集合 $|\tau(A)|\cup|\tau(B)|$ 可分为 2 个部分 $part1=\{A,B,C\}$，$part2=\{D,E\}$，则 $p(part1)=3/5$，$p(part2)=2/5$，根据式(10)得出熵值为 0.67。对比可知，图 1(b)中混乱程度较大，而混乱程度大则体现网络链接密度小，因此可知图 1(b)中节点 A、节点 B 间链接密度小于图 1(d)。

因此当集合 $|\tau(A)|$ 与集合 $|\tau(B)|$ 的并集和交集均相同时，非交集节点集中分布时，链接密度较大。

假设非交集节点数为 $|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|$，根据其分布是集中程度还是分散分布，总共有 $\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1$ 种情形，链接密度大小也分为 $\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1$ 个级别，具体链接密度级别数与节点度数有关。

具体相似度因子 2 如式(11)所示。

$$Factor2(x,y)=\frac{\min(\tau(x),\tau(y))-|\tau(x)\cap\tau(y)|}{\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1} \tag{11}$$

所有 $\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1$ 个链接密度级别间的差异应该较小，应小于相似度因子 $Factor1$ 的影响。如图 1(d)中节点 A、B 间的链接密度比图 1(b)中大，但应该小于图 1(c)中对应值，因为图 1(c)中 $|\tau(A)|$ 与 $|\tau(B)|$ 间重复节点重比例更大。

综上所述，可得出节点相似度(vertex similarity)的定义。

对于带权图

$$V_{xy}=\frac{\sum_{z\in\tau(x)\cap\tau(y)}w(x,z)w(y,z)}{\sum_{z\in\tau(x)\cup\tau(y)}w(x,z)w(y,z)}\cdot\left(1-\frac{1}{|\tau(x)\cup\tau(y)|}\cdot\frac{\min(\tau(x),\tau(y))-|\tau(x)\cap\tau(y)|}{\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1}\right) \tag{12}$$

对于无权图

$$V_{xy}=\frac{|\tau(x)\cap\tau(y)|}{|\tau(x)\cup\tau(y)|}-\frac{1}{|\tau(x)\cup\tau(y)|}\cdot\frac{\min(\tau(x),\tau(y))-|\tau(x)\cap\tau(y)|}{\left\lfloor\dfrac{|\tau(x)\cup\tau(y)|-|\tau(x)\cap\tau(y)|}{2}\right\rfloor+1} \tag{13}$$

根据式(13)，可计算出图 1(a)~1(d)中节点 A、B 间的节点相似度分别为 2/9,1/2,1 和 3/5，这几个值较均匀的分布在区间[2/9,1]中，相对于式(1)计算出

的结构相似度区间[1/2,1]要更合理，因而根据局部结构信息，节点相似度比结构相似度描述链接密度更准确。

节点相似度代替结构相似度描述链接密度，利用 OPTICS 算法得到的一维序列更合理，同时式(7)转换为

$$l_x(t+1)=l_x(t)+\frac{1}{d}\sum_{y\in N_\varepsilon(x)}V_{xy}\sin(l_y-l_x) \tag{14}$$

3.2 节点耦合强度分析

文献[11]中对Kuramoto模型将所有节点看成等价节点，节点间的耦合强度也相同，这样忽略了不同节点的重要程度差异。

例如图 2(a)中无向无权网络，根据式(13)，节点 V_4 与其邻居节点 V_3,V_5 间的节点相似度分别为 $V_{V_3V_4}=3/10$, $V_{V_4V_5}=3/32$ 。由式(14)可知，根据 OPITCS 排序结果得出 $l_{V_5}-l_{V_4}=l_{V_4}-l_{V_3}$，则进一步运算可得出 V_4 与 V_1,V_2,V_3 同属一个社区。实际上 V_4 与其他几个节点同属一个社区。

事实上，不同节点的重要性存在差异，对网络的影响也不尽相同。例如图 2(a)中，V_5 节点的度数最大，对网络的影响也最大。因而，需要对节点的全局重要性进行分析。对于 Kuramoto 同步模型来说，重要性体现在全局范围内不同节点间的耦合强度不同。

研究人员已经提出一些评价节点重要性的算法，例如 K-SHELL[13]，PageRank[14]等。但这些方法计算起来，相对复杂，效率较低。在本文中，使用最简单的节点重要性评价的方法：节点重要性与其度数紧密相关。而节点耦合强度显然与 2 个节点均有关，因此将 ΔtC 转换为 $\frac{degree(x)+degree(y)}{maxDegree}$，则式(6)进一步转换为

$$l_x(t+1)=l_x(t)+\frac{1}{d}\sum_{y\in N_\varepsilon(x)}V_{xy}\sin(l_y-l_x)\cdot\frac{degree(x)+degree(y)}{maxDegree} \tag{15}$$

其中，*maxDegree* 为网络中最大度数。

4 ESYN 算法

4.1 算法描述

节点相似度定义体现局部范围内节点的相似程度，耦合强度分析体现全局范围中节点对的重要程度，二者需结合起来才能将同步过程描述更完整、更准确。

文献[11]中提出 SYN 存在相似度定义不够准确，且忽视节点耦合强度差异的问题，本文重新定义节点相似度用于描述链接密度，同时在 Kuramoto 模型中增加耦合强度分析，提出 ESYN（efficient synchronization clustering）聚类算法。

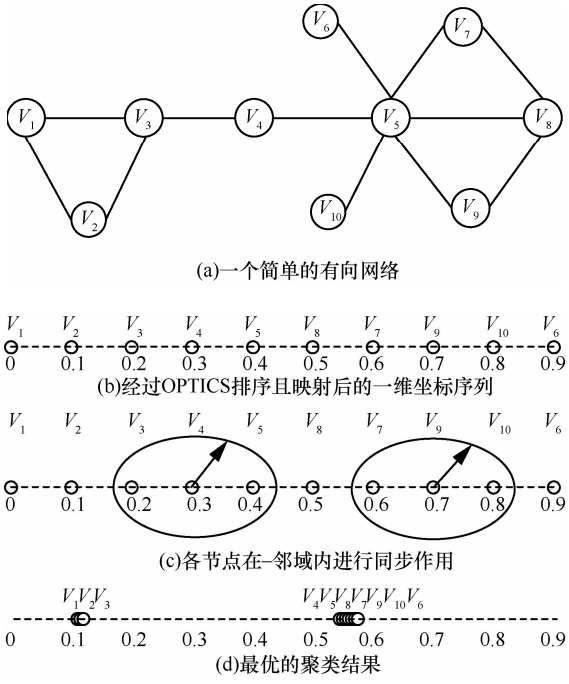
ESYN 算法步骤如下。

1) 矢量化预处理。使用 OPTICS 算法，利用公式(8)中节点相似度定义描述网络中节点链接密度关系，将各节点进行排序；然后根据排序结果将各节点均匀映射到一维坐标区间[0,1)上，得到一个一维坐标序列，从而将非矢量网络转化为一维矢量网络。图 2(a)中矢量化后结果如图 2(b)所示。

2) t=0 时，设定参数 ε 的初始值。

3) t 的值加 1，根据式(15)计算每个对象与其 ε–邻域内的邻居对象相互作用后的新坐标位置。对调整坐标后的所有节点重新进行社团划分，将距离小于 ε 的节点判定为同一个社团。得到社团划分结果后，计算其模块度。图 2(c)为一维矢量网络在 ε–邻域内同步作用示意图。

4) 增大 ε 值，重复执行步骤 3)，得到另一个聚类结果和模块度。在不断增加邻域半径 ε 值的同步过程中，得到一系列聚类结果，选择其中模块度最大的作为最优聚类结果。图 2(d)为网络的最佳聚类。



(a)一个简单的有向网络

(b)经过OPTICS排序且映射后的一维坐标序列

(c)各节点在ε–邻域内进行同步作用

(d)最优的聚类结果

图 2 ESYN 算法执行过程

4.2　算法时间复杂度分析

基于动态模型的局部同步聚类算法是由矢量化预处理、同步聚类和模块度计算 3 部分组成。对于矢量化预处理，需要计算所有相连节点相似度，此部分的时间复杂度为 $O(N\log N)$,其中 N 为网络中节点数。对于同步聚类过程，节点的邻域内的数据对象个数最多为 N 个。在一次同步运动中，每个点都进行同步计算，因而时间复杂度为 $O(N^2)$。一次同步运动后进行模块度计算，首先对同步过程所得到的社团两两进行一次模块度 Q 值计算，遍历网络中的所有边，时间复杂度为 $O(M)$,其中 M 是给定网络的边数。所以每一次同步聚类得到社团划分总的时间复杂度为 $O(M+N^2)$。重复同步运动过程，直至网络中所有节点都聚到一个社团。因此，ESYN 算法的时间复杂度为 $O(L(M+N^2))$，其中 L 为同步次数，ESYN 算法时间复杂度与 SYN 算法时间复杂度相同。

5　实验结果与分析

本节在多个人工合成和真实数据集上对提出的 ESYN 的性能进行实验评价。将与 SYN 算法以及基于模块度优化的 FastModularity 算法、LHLC 算法进行比较。ESYN 算法采用 ANSI C++编写，其余算法采用作者提供的源代码。所有实验均在 2.8GHz CPU, 2GB 内存的计算机上完成。

5.1　实验数据集

为了详细分析 ESYN 算法的性能，实验选取了不同规模的真实网络和人工网络，以比较相关算法在不同类型、不同规模和不同混合系数的网络上的运行结果。

真实数据集方面，本文采用了如下 2 种数据集。

1) College-Football Network[15]：该网络(http://networkdata.ics.uci.edu/data.php?id=5)是美国大学足球赛网络。美国大学共有 115 支大学生足球队，在 2000 年常规赛期间共打过 1 232 场比赛，因而 College-Football 网络包含 115 个节点，1 232 条边。所有球队根据比赛关系分为 12 个联盟，即网络的社团结构。

2) Zarchary's Karate Network[16]：该网络是美国一所大学中的空手道俱乐部成员间的关系网络（http://networkdata.ics.uci.edu/data.php?id=105）。俱乐部包含 34 个成员，即网络包含 34 个顶点，并包含 78 条边，代表俱乐部成员之间的人际关系。由于突发的原因，俱乐部管理者与主要教师之间针对是否提高收费这一问题，产生了激烈的争论并最终导致俱乐部分裂成 2 部分，即划分的网络社团。

人工数据集方面，本文采用 Lancichinetti 等人开发的工具 LFR-Benchmark 生成了节点数分别为 1 000、100 00 的人工网络。每种不同节点数的网络均包含混合系数从 0.1~0.8 间隔为 0.1 的 8 个不同网络。

5.2　质量评价标准

本文采用基于信息论的指标——规范化交互信息[17](NMI, normalized mutual information)——对聚类算法的结果进行评价。

设 D 是包含 n 个数据的集合$\{d_1,d_2,\cdots,d_n\}$。假设 U 和 V 为 D 的 2 种不同的聚类结果，分别包含 R、S 个社团：$U=\{U_1,U_2,\cdots,U_R\}$，$V=\{V_1,V_2,\cdots,V_S\}$，$\bigcup_{i=1}^{R}U_i=\bigcup_{j=1}^{S}V_j=D$，$\bigcap_{i=1}^{R}U_i=\bigcap_{j=1}^{S}V_j=\varnothing$。

聚类 U 的信息熵定义为 $H(U)=-\sum_{i=1}^{R}P(i)\log P(i)$，其中，$P(i)$ 为第 i 个社团 U_i 出现的概率 $P(i)=\dfrac{|U_i|}{n}$。2 个聚类 U 和 V 之间的互信息 $I(U,V)=\sum_{i=1}^{R}\sum_{j=1}^{S}P(i,j)\log\dfrac{P(i,j)}{P(i)P(j)}$。其中，$P(i,j)$ 为同一数据同时属于 U_i,V_j 的概率，$P(i,j)=\dfrac{|U_i\cap V_j|}{n}$。归一化的互信息 $NMI(U,V)$ 定义为

$$NMI(U,V)=\frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{16}$$

NMI 在区间[0,1]之间取值，*NMI* 值越高表示聚类效果越好。当 *NMI* 值为 1 时，表示算法的聚类结果与标准聚类结果完全相同；*NMI* 值为 0 时，算法聚类结果与标准聚类结果各自独立，彼此间没有共享的信息。

5.3　实验结果与分析

影响网络的社团结构划分的因素有很多,除了聚类算法、网络规模等因素外，更受网络本身社团结构复杂程度，即网络的混合系数的影响。对于同一聚类算法，网络的社团结构越模糊，即混合系数越高，网络社团结构划分越难，准确率 *NMI* 值也越低。对于 ESYN 算法，需要为 ε 设置初始值。ε 初始值只需是较小值即可，对聚类影响不是很大。设置初始参数 的初始值为 KNN(3)，并且增加同步半径时，参数 ε 每次增加 KNN(4)-KNN(3)，其中 KNN(m)函数为网络中的 m 个邻域的平均值。

5.3.1　真实网络实验结果

对于真实网络，其社团结构往往较人工网络复

杂。在不同规模的真实网络数据集上，对 ESYN, SYN, LHLC 和 FastModularity 等 4 种算法的社团结构检测结果进行了对比分析。社团检测时，注意消除网络中的重复边带来的影响。对于 College-Football 和 Zachary's Karate 2 种真实网络，4 种算法进行社团划分结果的 *NMI* 值如表 2 所示。

表 2　　真实网络上的聚类结果 *NMI* 值

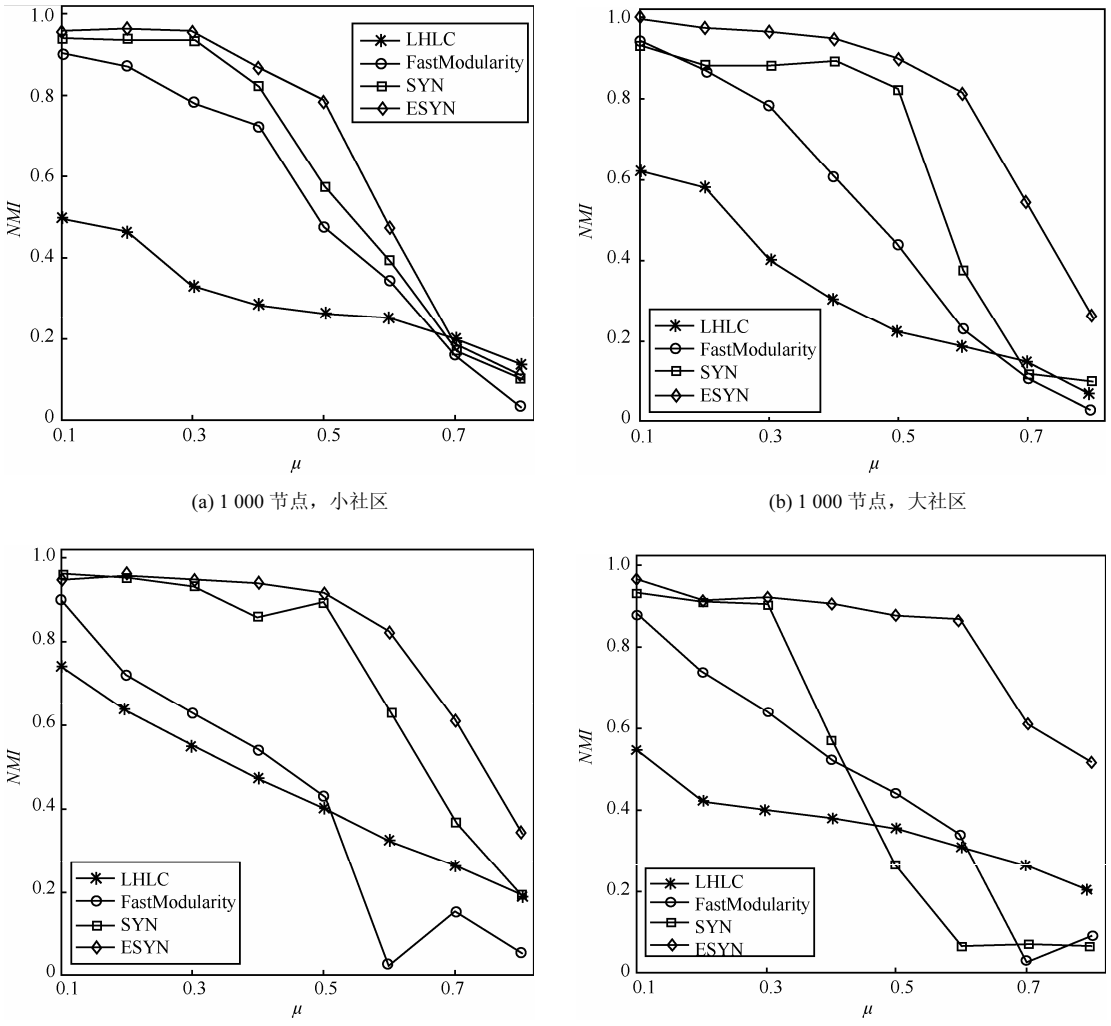
真实网络	ESYN	SYN	LHLC	FastModularity
College-Football	0.92	0.89	0.32	0.23
Zachary's Karate	1	1	0.42	0.75

其中，College-Football 源自 UCI 网络数据集（http://networkdata.ics.uci.edu/data.php?id=5），该网络存在重复边的情况。文献[11]未消除重复边带来的影响，所得出 *NMI* 值偏大。去掉 College-Football 网络的重复边后，SYN 算法处理后得出的 *NMI* 值为 0.89。

对比 4 种算法在两种真实网络中的 *NMI* 值，LHLC 算法对于 2 种真实网络，均不能有效划分其社团结构，FastModularity 算法可以识别出 Zachary's Karate 网络中社团结构，但准确率不高。SYN 和 ESYN 能较准确的识别 2 个网络中的社团结构，其中 ESYN 算法识别效果更好，准确率更高。

5.3.2　网络实验结果

本文采用 4 组人工网络数据集，其中 2 组包含 1 000 个节点，另 2 组包含 10 000 个节点，网络中混合系数从 0.1~0.8。在这些人工网络数据集上多次运行 ESYN, SYN, FastModularity, LHLC 算法，检测各网络中的社团结构，记录器划分结果的准确率(*NMI* 值)，如图 3 所示。总体上,随着混合系数的上升，4 种算法社团结构划分准确率都而呈下降趋



(a) 1 000 节点，小社区　　(b) 1 000 节点，大社区

(c) 10 000 节点，小社区　　(d) 10 000 节点，大社区

图 3　4 种算法在人工数据集上的社团划分 *NMI* 值曲线

势；对于同一网络数据集，ESYN 算法 *NMI* 值要高于其他 3 种算法；当混合系数较大($\mu > 0.4$)时，ESYN 算法对于数据集的识别效果更好，*NMI* 值有较大提升。

6　结束语

聚类是数据挖掘领域重要的技术之一。本文提出一种高效的局部同步聚类 ESYN 算法，算法使用节点相似度描述局部范围内的链接密度，相对于结构相似度更为准确；在 Kuramoto 模型中增加节点耦合强度分析，能有效体现全局范围内节点对的重要程度差异。算法执行过程中，首先使用 OPTICS 算法将非矢量网络进行矢量化，得到一维坐标序列；然后使用基于 ε –邻域的 Kuramoto 模型进行同步，并获得聚类结果。不断增加同步半径 ε 值，在获得的一系列聚类结果中，自动选择模块度最大的聚类结果作为最优聚类。实验结果显示，相对于 SYN 算法，本文提出的 ESYN 聚类算法在不增加算法时间复杂度的基础上，能提高算法的识别准确率，同时较同类算法有一定优势，因此更加高效，并能够很好地应用于实际网络数据分析系统中。

参考文献：

[1] GUAN J, GAN Y, WANG H. Discovering pattern-based subspace clusters by pattern tree[J]. Knowledge-Based Systems, 2009,22(8): 569-579.

[2] ZHU S, WANG D, LI T. Data clustering with size constraints[J]. Knowledge-Based Systems, 2010,23(8):883-889.

[3] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithmfor discovering clusters in large spatial databases with noise[A]. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining[C]. 1996.226-231.

[4] KARYPIS G, HAN E H, KUMAR V. Chameleon: hierarchical clusteringusing dynamic modeling[J]. Computer, 1999,32(8):68-75.

[5] NEWMAN M E, GIRVAN M. Finding and evaluating community structurein networks[J]. Physical review E, 2004, 69(2): 026113.

[6] CLAUSET A, NEWMAN M E, Moore C. Finding community structure invery large networks[J]. Physical review E, 2004, 70(6): 6111-6116.

[7] LEUNG I X, HUI P, LIO P, *et al*. Towards real-time communitydetection in large networks[J]. Physical Review E, 2009,79(6):6107-6117.

[8] NEWMAN M. Detecting community structure in networks[J]. The European Physical Journal B-Condensed Matter and Complex System, 2004,38(2):321-330.

[9] DEKKER A H, TAYLOR R. Synchronization properties of trees in the Kuramoto model[J]. SIAM Journal on Applied Dynamical Systems, 2013, 12(2):596-617.

[10] BÖHM C, PLANT C, SHAO J, *et al*. Clustering by synchronization[A].Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2010.583-592.

[11] 黄健斌, 白杨, 康剑梅等. 一种基于同步动力学模型的网络社团发现方法[J]. 计算机研究与发展, 2012,49(10):2198-2207.
HUANG J B, BAI Y, KANG J M, *et al*. A network community detection method based on dynamic model of synchronization[J]. Journal of Computer Research and Development, 2012, 49(10):2198-2207.

[12] ANKERST M, BREUNIG M M, KRIEGEL H P, *et al*. OPTICS: orderingpoints to identify the clustering structure[J]. ACM SIGMOD Record, 1999, 28(2): 49-60.

[13] KITSAK M, GALLOS L K, HAVLIN S, *et al*. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010,6(11): 888-893.

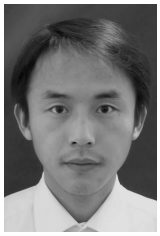
[14] BRYAN K, LEISE T. The $25,000,000,000 eigenvector: The linear algebrabehind Google[J]. Siam Review, 2006,48(3):569-581.

[15] GIRVAN M, NEWMAN E. Community structure in social and biologicalnetworks[J]. PNAS, 2002, 99(12):7821-7826.

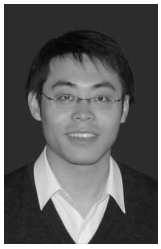
[16] ZACHARY W W. An informationflow model for conflict and fission in small groups[J]. Journal of anthropological research, 1977, 33(4): 452-473.

[17] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: is a correction for chance necessary[A]. Proceedings of the 26th Annual International Conference on Machine Learning[C]. ACM, 2009.1073-1080.

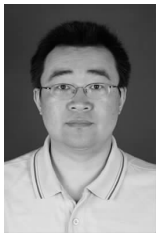
作者简介：



董学文（1981-），男，湖北黄冈人，博士，西安电子科技大学副教授，主要研究方向为 Web 数据挖掘、社交网络信息分析、无线网络安全。



杨超（1979-），男，陕西西安人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为移动互联网、无线网络安全。



盛立杰（1976-），男，山西大同人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为未来互联网体系结构、软件定义网络 SDN、移动互联网。

马建峰（1963-），男，陕西西安人，博士，西安电子科技大学教授、博士生导师，主要研究方向为移动互联网、网络安全、密码学。