ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.14, No.02, Jan. 2018
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于DBSCAN算法的复杂网络聚类

姜皓月,石梦彤,关童升,王思奇,陈嘉威,宁雪梅

(北京林业大学理学院,北京 100083)

摘要:复杂网络聚类方法可以挖掘复杂网络的结构,对复杂网络的研究具有重要意义。DBSCAN算法是一种基于密度的聚类算法,主要用于对传统数据点集进行聚类。由于复杂网络的特殊性质,对DBSCAN算法进行改进,采用相似度度量法代替传统算法中的欧式距离度量,对复杂网络进行聚类。其优点是聚类快速、可以发现任意形状的聚类、自动确定聚类数以及有效剔除噪声点。

关键词:复杂网络;网络聚类;密度聚类

中图分类号:TP301 文献标识码:A 文章编号:1009-3044(2018)02-0141-03

DOI:10.14004/j.cnki.ckt.2018.0062

Complex Network Clustering Based on DBSCAN Algorithm

JIANG Hao-yue, SHI Meng-tong, GUAN Tong-sheng, WANG Si-qi, CHEN Jia-wei, NING Xue-mei

(Beijing Forestry University College of Science, Beijing 100083, China)

Abstract: The method of complex network clustering can excavate the structure of complex network, which is of great significance to the research of complex network.DBSCAN algorithm is a density clustering algorithm, which is used to cluster traditional data points.Due to the special nature of complex network, to improve the DBSCAN algorithm,adopt the method of similarity measure to replace the Euclidean distance measurement in the traditional DBSCAN algorithm to cluster the complex network. . The advantages of this method are clustering fast, finding the clustering of arbitrary shapes, automatically determining the clustering number, and effectively eliminating the noise points.

Key words: complex network; network clustering; density clustering

现实世界中的许多复杂系统直接或间接地以复杂网络的形式存在[1],如社交网络、生物网络。研究者们通过对网络性质的深入研究,发现复杂网络具有集团化的特性。也就是说,整个网络是由若干个“类”构成的[2]。聚类算法把一组结构未知的数据进行分类,使每一类之间的相似性尽可能小,每一类之内的相似性尽可能大,其目的是寻找数据中有效的结构。因此,利用聚类算法可揭示出复杂网络中存在的网络社团结构、发现复杂网络中隐藏的规律。

DBSCAN是一种基于密度的聚类算法,要求聚类空间中的一定区域内所包含的对象的数目不小于某一给定阈值[3]。DBSCAN算法的优势是可以发现任意形状的聚类、自动确定聚类数以及有效剔除噪声点。因此本文使用DBSCAN算法对复杂网络进行聚类。由于网络与数据点集对距离的定义不同,本文用相似度度量代替传统DBSCAN算法中的距离度量。测试结果表明该算法对复杂网络的聚类是可行的。

1 算法介绍

DBSCAN算法是一种基于密度的空间数据聚类方法,其中心思想是:对于某一聚类中的每个对象,在给定半径 (文中用Eps表示)的邻域内数据对象个数必须大于某个给定值,也就是说,邻域密度必须超过某一阀值 (文中用MinPts表示)[4]。

为使用此算法进行复杂网络聚类,在一个网络D中,进行如下定义:

定义1(相似度S_{ij})S_{ij}代表网络中的节点i和j的连接程度,与节点i,j间的距离成反比,具体定义如下[5]:

首先,对于一个无向无权的网络G=(V,E),G的拉普拉斯算子是矩阵:

$$L_{i,j}=\begin{cases}1, \text{fori}\sim j\\ -d_i, \text{fori}=j\\ 0, \text{o th erwise}\end{cases} \tag{1}$$

其中i~j表示第i个和第j个节点有边相连,di是节点的度。矩阵L的指数定义为:

$$K^{\beta}\equiv\exp(\beta L)=\lim_{n\to\infty}(I+\frac{\beta L}{n})^n \tag{2}$$

其中β是取值为正的常数,通常在0.1~0.5之间。而这个极限总是存在,将上式展开如下:

$$\exp(\beta L)=I+\beta L+\frac{\beta^2}{2}L^2+\frac{\beta^3}{3!}L^3+\cdots \tag{3}$$

得到的矩阵K^{β}是对称和正定的。利用Pade逼近方法计算矩阵指数[6]。通过归一化核心矩阵K^{β},相似度矩阵S^{β}可以定

收稿日期:2017-12-09
基金项目:北京林业大学大学生科研训练计划(项目号:X201710022145)、国家自然基金项目资助(基金号:11501032)
作者简介:姜皓月(1996—),女,辽宁人,研究方向为数据分析;石梦彤,女;关童升,男;王思奇,女;陈嘉威,男;宁雪梅,通讯作者,讲师,博士研究生,研究方向为运筹学与控制论。

义为：

$$S_{ij}^{\beta}=\frac{K_{ij}^{\beta}}{\sqrt{K_{ii}^{\beta}K_{jj}^{\beta}}} \tag{4}$$

定义2(邻域N(p))：点p的邻域为：

$N(p)=\{q|dist(p,q\leqslant Eps)\}$)(Eps为给定的相似度$S_{ij}$的倒数)

定义3(邻域密度Dens(p))：点p的邻域密度是N(p)所包含的点的数目。

定义4(核心点Core Points)网络中，邻域密度大于某一给定阈值MinPts的点。

定义5(边界点Border Points)落在核心点的邻域内且邻域密度小于某一给定MinPts的点。

定义6(直接密度可达)若p在q的邻域内，且q是核心点，则称p从q直接密度可达。

定义7(密度可达)若有点$p_1,p_2,\cdots,p_n$，且p_i从p_{i+1}直接密度可达，则称点p_1从p_n密度可达。

定义8(密度连接)若有点o，且p、q都是从o关于同一 Eps 和MinPts密度可达的，则p和q是密度连接的。

定义9(类Cluster)若p为一核心点，D中所有从p密度可达的节点和p构成的集合称为一个类。

定义10(噪声点Noise Points)D中不属于任何一类的点。

算法描述如下：

访问一个出发点p，若p为核心点，找出所有密度可达的点形成一个类C，并将p标记为已处理。若p为非核心点，暂时将p标记为噪声点。

找到第一个类C后，重复步骤1，处理C中所有的节点，继续将C进行扩展[7]。

C中的节点全部访问过后，用同样的方法访问C以外节点。直到所有节点都归入某个类中或被标记为噪声点。

算法实现的实例如图1，图中八个节点被分为两类，并以不同颜色标记。

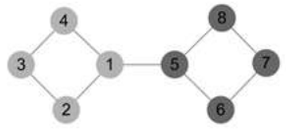

图1　算法示意图

2 实例验证

2.1 模拟数据

为检验算法的准确性与实用性，本文生成1000个包含30个节点的随机网络样本，并将坐标点进行编号。设定点1-10为第Ⅰ类，点11-20为第Ⅱ类，点21-30为噪声点。同一类内节点有边相连的概率P_1=80%，噪声点与任意类有边相连的概率P_2=20%，对1000个网络样本进行聚类，结果如图2。

分类错误的节点出现的频率如图3所示，聚类精度为96.167%。

调整P_2=30%，再次进行测试，结果如图4，聚类精度为95.3%。

2.2 真实数据

我们利用该算法测试了一些具有已知类结构的网络，并且可以检测到这些网络中的类。

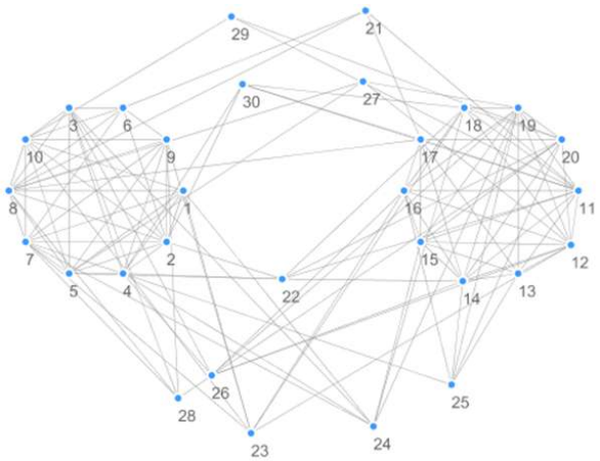

图2　包含30个节点的随机网络聚类结果

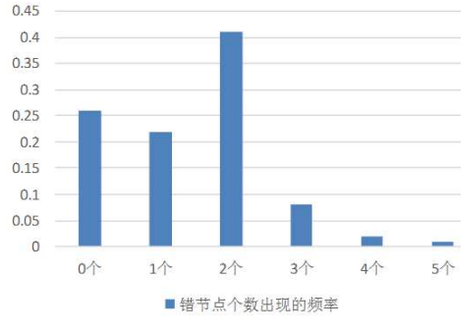

图3　错误出现的频率(P_2=20%)

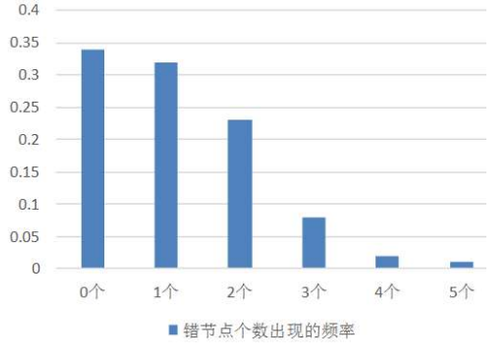

图4　错误出现的频率(P_2=30%)

首先测试了具有34个节点的Zachary研究的空手道俱乐部内部成员的关系网络，结果如图5，方形和圆形的节点代表已知的两个类，不同颜色的节点代表新划分的类。有三个节点判断错误，聚类精度为91.176%，节点3、14、20处于两个社团的交界处，本身具有一定歧义性[8]。

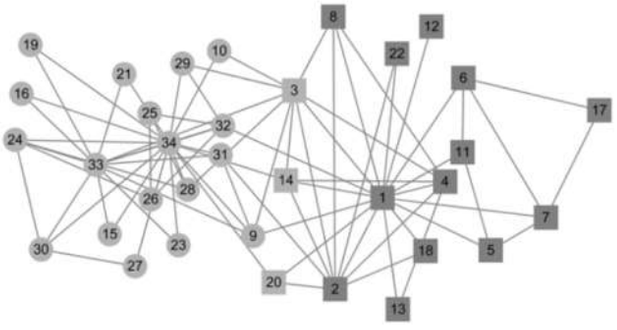

图5　空手道俱乐部网络聚类结果

接着我们测试了具有115个节点的足球俱乐部成员关系

网络，结果如图6：

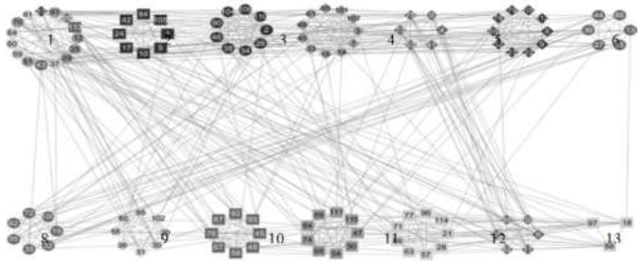

图6　足球俱乐部网络聚类结果

我们试着将足球俱乐部网络计算的模块与实验确定的聚类相匹配。使用超几何测量法作为最佳匹配标准，通过最小化计算组和实验组之间的随机重叠概率Polof，我们可以确定模块的最佳匹配实验复合体。

Pol定义为[9]：

$$P_{ol}=\frac{\binom{n_2}{k}\binom{N-n_2}{n_1-k}}{\binom{N}{n_1}} \tag{5}$$

其中n1是新划分的聚类，n2是已知的聚类结果，k是匹配的节点的数量，N是网络的大小聚类结果越准确，log(Pol)值越小。最筛选确定终结果较准确的类为：

表1　结果准确的类及log(Pol)值

Cluster	2	3	4	8
log (Pol)	-27.1059	-29.5815	-34.1984	-29.5445
Cluster	9	10	11	12
log (Pol)	-24.9087	-24.9087	-29.5815	-24.1879

3 算法评价

本文使用DBSCAN算法的原理对复杂网络进行聚类。针对复杂网络的特性，将传统DBSCAN算法使用的欧式距离度量改为相似度度量。

由于复杂网络具有小世界性，即网络间的平均路径长很小，所以本文的算法的一个优势是可以很好确定邻域半径范围；与谱聚类方法等算法相比，本算法可以自动确定聚类数；并且还具有可以有效剔除噪声点、发现任意形状的聚类的优点。

由于算法对输入参数较为敏感，不同的参数对结果的影响较大，所以需要对网络的相似度矩阵有所观察后方能得到较准确的结果。并且由于算法是对密度进行划分的，当空间密度分布不均匀时，聚类结果较差且参数较难选择。

参考文献：

[1] 李建, 郑晓艳. 复杂网络算法聚类综述[J]. 电脑知识与技术, 2009, 11(5):37-41.
[2] 汪小帆, 李翔, 陈关荣. 复杂网络的理论及其应用[M]. 北京：清华大学出版社, 2006：162.
[3] 王伟东, 芦金掸, 张讲社. 基于视觉原理的密度聚类算法[J]. 工程数学学报, 2005, 22(2):349-352.
[4] 周水庚, 周傲英, 曹晶. 基于数据分区的DBSCAN算法[J]. 计算机研究与发展, 2000, 37(10):1153-1159.
[5] Zhang S,Ning X M, Zhang X S. Graph kernels, hierarchical clustering, and network community structure：experiments and comparative analysis[J]. Eur. Phys. J. B, 2007：57, 67-74
[6] mathworks[EB/OL].http://www.mathworks.com/.
[7] 杨芳勋. DBSCAN 算法在电子邮件网络社团发现中的应用[J]. 计算机科学, 2017, 44(6A):591-593.
[8] 汪小帆, 李翔, 陈关荣. 复杂网络的理论及其应用[M]. 北京：清华大学出版社, 2006：166.
[9] Shihua Zhang, Xuemei Ning, Xiangsun Zhang. Identification of functional modules in a PPI network by clique percolation clustering[J]. Computational Biology and Chemistry, 2006(30): 445-451.