基于结构与属性的社区划分方法

万新贵,李玲娟
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:目前通行的社区划分方法大多基于结构,但单纯基于结构的划分不能挖掘出社区对象的潜在关系,因而不能发现社区的变化趋势。为此,提出了基于结构的社区划分算法(Community Division based on Structure,CDS)。该算法利用度和节点欧氏距离对社会网络进行结构划分;同时针对经典 K-means 算法在社区划分中所存在的随机选取初始中心点以及 k 值选取不合理所导致的聚类结果不佳问题,提出了一种基于社区结构的非人为设定 k 值的 K-means 算法—NPCluster(Non Presetting Cluster)算法。该算法基于由 CDS 算法所提到的社区结构,依次选取度最大的节点作为聚类中心点,以小于平均特征欧氏距离为基准合并簇集,反复迭代直至聚类完成。理论分析和对比实验结果表明,CDS 算法能够有效划分出社区结构;相对于 K-means 算法,NPCluster 算法在已划分的社区结构上具有更高的聚类精度和更好的时效性;结构与属性相结合的社区划分方法是有效可行的。

关键词:社区划分;度;K-means;中心点;欧氏距离
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)08-0097-05
doi:10.3969/j.issn.1673-629X.2017.08.020

Community Division Method with Structure and Attribute

WAN Xin-gui,LI Ling-juan
(School of Computer,Nanjing University of Posts and Telecommunications,
Nanjing 210003,China)

Abstract:Most of the current methods of community division are based on structure,but the structure-based division cannot excavate the potential relationship of community objects,which is not to find the tendencies of community variations. Therefore a community-based partitioning algorithm (Community Division based on Structure,CDS) has been designed which applies degree and node-Euclidean distance to divide social network. Simultaneously,an algorithm by nonhuman (artificial) setting k-value—NPCluster algorithm (Non Presetting Cluster)—based on the community structure has been proposed,which is based on the community structures divided by CDS algorithm and has improved the unsatisfactory clustering outcomes caused by the inappropriateness of random selection of initial centers and that of K value. Thus the maximum degree nodes are chosen as a cluster center in turn and the data are merged and clustered until the average feature-Euclidean distance is less than a given threshold. Theoretical analyses and experimental results show that the proposed CDS algorithm can effectively divide the community structures;compared with K-means algorithm,NPCluster algorithm has higher clustering quality and better clustering timeliness on the divided community;the community division method based on structure and attribute is practical and effective.

Key words:community division;degree;K-means;center;Euclidean distance

0 引 言

在社会网络研究[1-3]中关心的两个方面是联系和结构。目前基于结构角度的社区划分研究比较充分,但是单纯基于结构的划分(称为硬划分)对社区内对象的潜在关系(比如兴趣的异同等)表现不够。而这种潜在关系的发现(称为软划分)对预测社会网络社区的变化趋势有着重要的参考价值。

数据挖掘(Data Mining)[4]一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。聚类挖掘[5]是数据挖掘的算法之一,它将大量的数据划分为性质相同的子类,以便于了解数据的分布情况,挖掘结

收稿日期:2016-08-04 **修回日期**:2016-11-10 **网络出版时间**:2017-06-05
基金项目:国家自然科学基金资助项目(61302158,61571238)
作者简介:万新贵(1991-),女,硕士研究生,CCF 会员(E200041361G),研究方向为流数据挖掘;李玲娟,教授,CCF 会员(E200015276M),研究方向为数据挖掘、信息安全、分布式计算。

果具有簇内高相似性和簇间低相似性,聚类挖掘的性质相当符合社会网络社区的特点。

聚类挖掘算法[6]主要分为四大类:基于划分的聚类算法、基于层次的聚类算法、基于密度的聚类算法以及基于网格的聚类算法。不同类别的聚类算法适用于不同的应用场景。K -means 算法[7]作为聚类算法中经典的划分算法,其最大优势在于简洁和快速,在实践中应用广泛。该算法简单,易于理解和可扩展,并且很容易修改以便处理不同的数据,例如无监督性学习或流数据。但 K -means 算法仍有值得改进的地方,随机选择中心点以及人为设定 k 值是 K -means 算法最大的缺陷,针对这些缺陷,提出了许多改进算法[8-12]。

为了将软划分与硬划分进行有效结合,设计了社区结构建立算法(Community Division based on Structure,CDS)。该算法利用节点度与节点欧氏距离实现了社区中心的选择,完成了社区结构的建立;并基于 CDS 算法,提出一种非人为预先设定 k 值的聚类算法—NPCluster(Non Presetting Cluster K -means)。该算法的原理与 CDS 算法类似,同样基于节点度设置聚类中心,避免了 k 值的不合理设置导致的聚类结果不理想;基于特征欧氏距离进行聚类,实现多维数据集的类别划分。两种算法总体实现了社会网络的硬划分与软划分的结合。

1 相关工作

1.1 基本定义

定义 1(特征欧氏距离):数据集中每个数据点可以定义为 $X_i = (X_{i1}, X_{i2}, \cdots, X_{id})$, d 为数据点的维度,即特征的个数,则特征欧几里得距离的计算公式为:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \tag{1}$$

D 值越小,代表两个数据点的特征相似度越大,从而可以将这两个数据点划分到一个簇集中。

定义 2(节点欧氏距离):假设有 M 个节点,可以把节点集定义为 $X = (X_1, X_2, \cdots, X_m)$, X_{ij} 表示节点 X_i 到 X_j 的最短距离(跳数),则节点欧氏距离定义为:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{M} (x_{ik} - x_{jk})^2} \tag{2}$$

D 值越小,代表两个节点的相似度越大,从而可以将这两个节点划入同一个社区结构中。

定义 3(欧氏距离矩阵及平均欧氏距离):在数据集中,数据点两两之间的特征欧氏距离形成特征欧氏距离矩阵;在无向图中的定义与在数据集中的定义大致相同,以下特征欧氏距离与节点欧氏距离统称欧氏距离。

平均欧氏距离是由欧氏距离矩阵决定的,将欧氏距离矩阵求和再除以数据点数,得到平均欧氏距离。随着迭代过程的进行,数据点逐渐减少,欧氏距离矩阵将随数据点的变动进行更新,平均欧氏距离也会随之改变。

定义 4(度及度的计算):在无向图中,每个节点连边的条数就是该节点的度数。

度的计算是根据节点的关系矩阵 $\boldsymbol{R}$ 来确定的,点 X_i 与 X_j 之间有边,则 $r_{ij} = 1$,反之 $r_{ij} = 0$。按行读取矩阵 $\boldsymbol{R}$,将 i 行中的值相加,即得到点 X_i 的度。

1.2 社区划分

现实世界中,社会网络[1,13]以各种关系网络存在于多个领域,社会网络分析已经成为数据挖掘中的一个研究热点,而社区划分则是社会网络分析中的一项重要内容。社区发现的研究已取得了很多成果,使用聚类算法进行社区划分的研究比较普遍。

大多数实际网络都有一个共同的性质,即社区结构。整个网络就是由若干个“社区”或“组”构成的,而这些社区则具有社区内部高内聚、社区之间低内聚的特性,所以揭示网络的社区结构对于深入了解网络结构与分析网络特性是很重要的。

在社区划分的研究中,社区划分的算法所要划分的网络大致分为两类:比较常见的网络(仅包含正联系的网络)和符号社会网络(网络中既包含正向联系的边,也包含负向联系的边)。

社区划分算法有很多种,比较经典的三种是[14]:Kernighan-Lin 算法、基于 Laplace 图特征值的谱二分法和 GN 算法。

1.3 *K*-means 算法

K -means 算法是一种基于样本间相似性度量的间接聚类方法,属于非监督学习方法。此算法以 k 为参数,将 n 个对象分为 k 个簇,以使簇内相似度较高,而簇间相似度较低。

K -means 算法描述如下:

算法 1: K -means 算法。

输入:数据集, k 值;

输出:簇集。

(1)适当选择 k 个类的初始中心;

(2)在第 m 次迭代中,对任意一个样本,求其到 k 的欧氏距离,将该样本归到距离最短的中心所在的类;

(3)利用均值更新该类的中心值;

(4)对于所有的 k 个聚类中心,如果利用步骤 2 和步骤 3 迭代更新后,值保持不变,则迭代结束,否则继续迭代。

K -means 算法需要人为确定 k 值的大小,并且算法随机选取初始簇心,对初始簇心非常敏感,因此针对 K -means 算法的改进主要从这两方面入手。

2　CDS 算法

2.1　算法基本思想

基于硬划分方法产生的社区结构是进行软划分的基础。因此,首先设计一种建立社区结构的算法-CDS 算法。该算法的基本思想是:首先计算各节点的度值,选取度值最大的节点作为中心节点;然后计算所有节点之间的欧氏距离,形成节点欧氏距离矩阵,计算得出平均节点欧氏距离;将除中心节点以外的节点与中心节点的节点欧氏距离进行比对,当节点欧氏距离小于平均节点欧氏距离时,将此节点纳入该中心节点所在的社区,算法迭代至所有节点都被纳入社区。

以节点度的大小作为中心点的选择依据,符合社会网络中度越大凝聚力越强的物理意义;以平均节点欧氏距离作为聚类的标准符合社会网络中,距离同一节点的跳数越接近,节点之间越接近的物理意义。

2.2　算法流程与描述

CDS 算法的主体流程如图 1 所示。

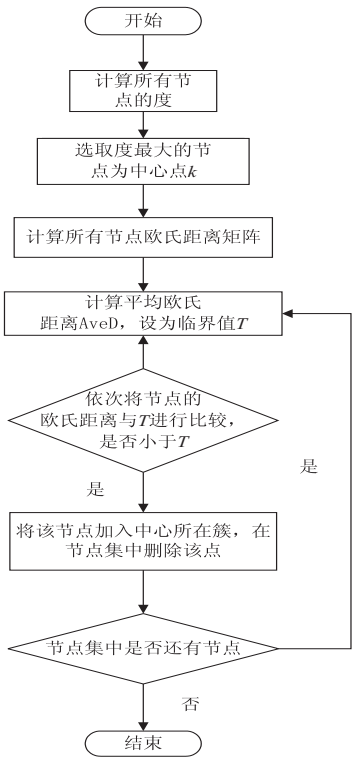

图 1　CDS 算法流程图

CDS 算法具体描述如下:

算法 2:社区结构建立算法。

输入:一个无向无权网络 $G = < V, E >$, V 是节点集合, E 是边的集合;

输出:网络的社区结构。

(1)计算所有节点的度 degree;

(2)找出现有节点中 degree 最大的节点,设置为中心节点 k ;

(3)计算所有节点的节点欧氏距离矩阵;

(4)计算节点集中节点的平均欧氏距离 AveD,将之设定为临界值 T ;

(5)将各个非中心节点与中心节点的欧氏距离与 T 进行比较,若小于 T ,则将该点加入社区结构,在节点集中删除该点;

(6)重复步骤 4 和 5,当节点集中没有节点时,社区划分完成。

3　NPCluster 算法

3.1　NPCluster 算法思想

K -means 算法的 k 值需事先确定,自主设定 k 值并不能保证合理性。假设被聚类的数据集中的数据点之间已经非常靠近,不需要进一步分类,但是由于人为设定的 k 值,数据集被聚集为 k 类。显然上述算法得到的结果是不合理的。针对这种情况,提出一种基于已建立的社区结构的非人为预先设定 k 值的聚类算法-NPCluster 算法。

在一个数据集(社区)中,某个数据点(节点)的度直接反映了这个点在该社区中的结构地位,即度越大,这个点越靠近所在社区的中心。若根据度来选择聚类中心,即可避免人为确定 k 值和随机选择聚类中心点。两个数据点之间的距离一般用特征欧氏距离表示,特征欧氏距离越小,表示这两个点越靠近。如果用数据点的多维特征作为数据点欧氏距离计算的基础,那么有理由相信,特征欧氏距离越小,两者的整体特征越接近。所以,以度作为中心点的选择,以特征欧氏距离作为聚类阈值的算法思想是可行的,并且同时兼顾了社会网络的结构和内在关系两个方面。

NPCluster 算法的基本思想是:首先计算数据点的特征欧氏距离矩阵,求得平均特征欧氏距离;然后选取度最大的数据点作为聚类中心点,将剩余数据点与中心点的特征欧氏距离与平均特征欧氏距离进行比对,若小于平均特征欧氏距离,则将该点划入中心点所在簇;迭代直至所有数据点都被划分,最后一个簇内的数据点在特定情况下可以视为孤立点。

该算法结合了数据点的结构关系和特征关系,基于这种聚类的社区划分是一种软硬划分的结合,聚类结果更符合社区的物理意义,也有其特殊的应用价值。由于 NPCluster 算法必须基于已建立的社区结构,所以算法的聚类结果会受到社区结构的影响,并且该算法不适用于一般无结构的数据集,仅适用于社会网络。

3.2　NPCluster 算法流程与描述

NPCluster 算法的主体流程如图 2 所示。

NPCluster 算法的具体描述如下:

算法 3:NPCluster 算法。

输入:数据集;

输出:簇集。

(1)计算所有数据点的度 degree;

(2)选取现有数据点中度值最大的数据点,设置为簇心 k;

(3)计算所有数据点的特征欧氏距离矩阵;

(4)求出数据集中所有数据点的平均特征欧氏距离 AveD,将之设定为临界值 T;

(5)将各非中心点与中心点的特征欧氏距离逐一与 T 进行比较,若小于 T,则将该点加入该簇,并将该点从数据集中删除;

(6)重复步骤 3 ~5,当数据集中没有数据点时,聚类结束。

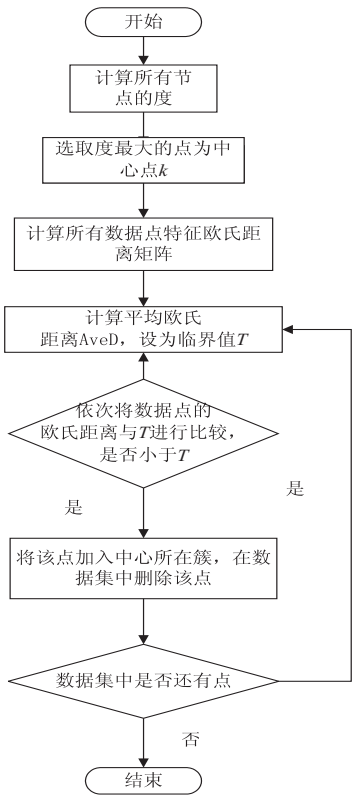

图 2 NPCluster 算法流程图

实现该算法的程序中涉及到的方法主要有:

(1)caldistance(String file)方法:根据多维特征计算所有数据点之间的特征欧氏距离,形成特征欧氏距离矩阵;

(2)avgdistance(double[ ][ ] a)方法:根据已得出的欧氏距离矩阵与数据点的数量,计算所有数据点的平均特征欧氏距离;

(3)Caldegree(String file)方法:根据度矩阵计算所有数据点的度;

(4)findmaxdegree(Map<Integer,Integer>m1)方法:根据已得出的数据点的度找出度最大的数据点。

通过这些方法,更易于理解 NPCluster 算法的核心思想。

4 实验结果及分析

4.1 CDS 算法验证

考虑到 NPCluster 算法基于已建立的社区结构,所以在 NPCluster 算法与 K -means 算法的对比实验前,需先验证 CDS 算法的正确性,以确保整体算法在无结构数据集上仍然具备有效性。

CDS 算法验证数据采用来自空手道俱乐部中的一个社区进行实验,社区结构图如图 3 所示,具体社区结构如表 1 所示。

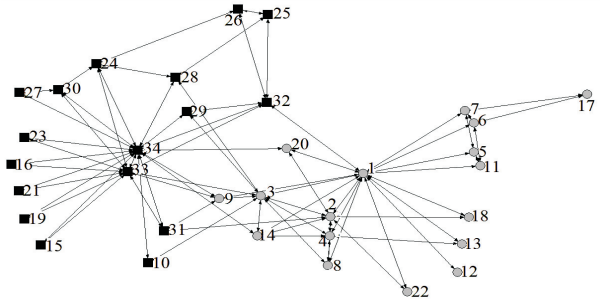

图 3 空手道社区结构图

表 1 空手道社区结构

社区	数字
社区 1	10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34
社区 2	1,2,3,4,5,6,7,8,9,11,12,13,14,17,18,20,22

由图 3 可以看出,实验数据基于结构共分为两个社区,图中数字代表社区中的各节点。

CDS 算法社区结构划分结果如图 4 所示。

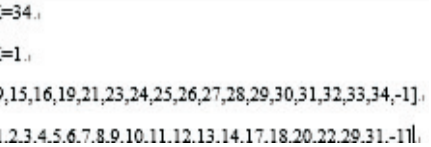

图 4 空手道社区结构划分图

由图 4 与表 1 的对比可以看出,节点 10 被划分错误,节点 29 和 31 被同时划分到两个社区中,经过计算得到 CDS 算法社区结构划分的正确率达到 97.06%,实验结果表明 CDS 算法是有效的。

4.2 NPCluster 算法性能验证

NPCluster 算法与 K -means 算法的对比实验采用校内网离散型数据集,该数据集共 7 000 条 6 维数据。使用 CDS 算法进行社区结构划分得到 3 个社区,选取人数最多的社区 2 作为对比数据集。社区 2 共 4 500 条数据,按兴趣团体共分为 3 大类。

4.2.1 精确度测试及分析

NPCluster 算法与 K -means 算法的精确度测试结果如图 5 所示。其中为确保实验不受人为因素影响,将 K -means 算法的 k 值设置为 3,与已知类别数目相同。

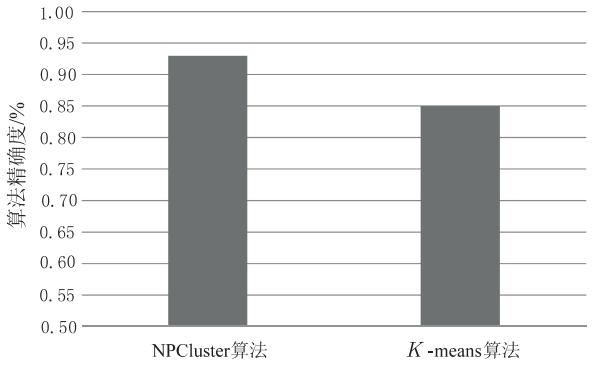

图 5　算法精确度测试结果

从图 5 可以看出,NPCluster 算法聚类的正确率明显高于 K -means 算法,并且 NPCluster 算法聚类得出的最后一个簇的成员可以看作一个孤立点。

由此可以得出,NPCluster 算法相对于基本 K -means 算法而言有着明显优势,既不用人为给定 k 值,又能找出数据集中的孤立点,且算法精确度更高。

4.2.2　效率测试及分析

NPCluster 算法效率测试结果如图 6 所示。

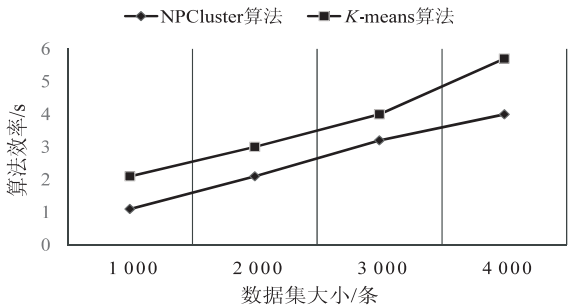

图 6　算法效率测试结果

由图 6 可以得出,NPCluster 算法效率要高于 K -means 算法。这是因为 K -means 算法在迭代选取中心点时消耗的时间较多,而 NPCluster 算法只需要比较特征欧氏距离即可。考虑到 NPCluster 算法是基于社区结构的,社区结构的建立过程也有时间消耗,由于算法原理相似,CDS 算法的时间消耗与 NPCluster 算法相当。因此,最终对比结果表明,基于社区结构的 NPCluster 算法的效率与 K -means 算法基本持平。

5　结束语

为了研究社会网络中的社区划分,并进一步挖掘社区对象的潜在关系,重点研究了基于结构与属性的社区划分方法。提出了基于结构的社区划分算法—CDS,以节点度和节点欧氏距离为社区结构划分准则,对社会网络进行结构划分,形成社区结构。基于上述社区结构,提出了一种非人为设定 k 值的聚类算法—NPCluster。该算法以节点度作为聚类中心选取依据,以多维特征的平均欧氏距离作为聚类阈值,聚类成功的点在数据集中被删除,经过多次迭代,直至数据集中不存在点,聚类结束,所产生的簇即对应于社区兴趣团体。由于 NPCluster 算法是基于社区结构来划分兴趣团体的,所以数据点之间基于特征的紧密性呈现会受到社区结构的影响。实验结果表明,CDS 算法能够以较高的准确度划分社区结构;NPCluster 算法在聚类效果上优于基本 K -means 算法,总体算法执行效率与 K -means 算法相当。

NPCluster 算法具备无需人为干扰的特性,在社会网络数据集上具备较好的适应性,能够发现结构社区下的兴趣团体划分,以及社会网络软硬结合的社区划分。

参考文献:

[1] 宗乾进,袁勤俭,沈洪洲. 国外社交网络研究热点与前沿[J]. 图书情报知识,2012(6):68-75.

[2] 王　亮. 基于局部聚类的复杂网络社区发现算法研究[D]. 大连:大连理工大学,2011.

[3] 张　鑫,刘秉权,王晓龙. 复杂网络中社区发现方法的研究[J]. 计算机工程与应用,2015,51(24):1-7.

[4] Wu X,Zhu X,Wu G Q,et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(1):97-107.

[5] 周　涛,陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用,2012,48(12):100-111.

[6] Verma M,Srivastava M,Chack N,et al. A comparative study of various clustering algorithms in data mining[J]. International Journal of Engineering Research and Applications,2012,2(3):1379-1384.

[7] Krishna K,Murty M N. Genetic K-means algorithm[J]. IEEE Transactions on Systems,Man,and Cybernetics,Part B (Cybernetics),1999,29(3):433-439.

[8] Celebi M E,Kingravi H A,Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm[J]. Expert Systems with Applications,2013,40(1):200-210.

[9] 王玉雷,李玲娟. 一种密度和划分结合的聚类算法[J]. 计算机技术与发展,2015,25(9):53-56.

[10] 尹成祥,张宏军,张　睿,等. 一种改进的 K-Means 算法[J]. 计算机技术与发展,2014,24(10):30-33.

[11] 刘莉莉,曹宝香. 基于差分进化算法的 K-Means 算法改进[J]. 计算机技术与发展,2015,25(10):88-92.

[12] 赵京胜,孙梦丹,张　丽. 一种有效的 K-means 初始中心优化算法[J]. 信息技术与信息化,2016(5):77-79.

[13] 朱　琪,于济坤,王明德,等. 社会网络数据的可视化[J]. 吉林大学学报:信息科学版,2015,33(5):584-587.

[14] 时京晶. 三种经典复杂网络社区结构划分算法研究[J]. 电脑与信息技术,2011,19(4):42-43.