

基于属性扩展图的聚类算法研究

任巍英, 高媛

(中北大学 电子与计算机科学技术学院, 山西 太原 030051)

摘 要: 提出一种新的图聚类算法, 结合结点的结构及属性特性, 使用统一的随机移动距离计算结点间的相似度, 在邻接随机移动距离矩阵的基础上进行聚类。实验结果表明, 基于属性扩展图的聚类算法在图拓扑结构的基础上, 充分考虑了各个结点所拥有的属性特点, 得到的聚类结果将更好的切合实际的应用。

关键词: 数据聚类; 结点相似度; 随机移动距离; 聚类目标函数

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2012)04-0110-04

Clustering Algorithm Based on Attribute Augmented Graph

REN Wei-ying, GAO Yuan

(School of Electronics and Computer Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: In this paper, we propose a new graph clustering algorithm, based on both structural and attribute characteristics, and use neighborhood random walk distances to measure vertex closeness, and make a clustering based on the transitional probability matrix. The experimental results show the effectiveness of the clustering algorithm that based on attribute augmented graph. The clustering algorithm has fully considered the attribute characteristics of every vertex based on the topological structure of graph. The result of clustering will better suit the practical application.

Key words: data clustering; vertex similarity; random walk distance; clustering objective function

1 引言

聚类是分析和探测数据的重要方法, 有着广泛的应用。近年来, 人们在图的相似度精确测量和匹配方面提出一些较好的方法, 但图的聚类问题仍是引起人们兴趣和挑战的课题之一。现有的许多图聚类方法主要集中于图的拓扑结构, 使每个分区达到一个有凝聚力的内部结构。但这些聚类方法往往忽略结点所具有的特有属性。本文提出一种新的结合结构相似性及属性相似性的聚类算法, 该算法在图拓扑结构的基础上充分考虑了各个结点所拥有的属性特点, 将更好的切合实际的应用^[1]。

本文所研究的目标是在结点结构及属性相似性的基础上对大型数据进行聚类, 将结点划分为 k 个不同的簇。一种可用的权衡结构相似性及属性相似性的方法为, 定义结点 v_i 与结点 v_j 间的距离为:

$d(v_i, v_j) = \alpha \cdot d_s(v_i, v_j) + \beta \cdot d_A(v_i, v_j)$, 其中 $d_s(v_i, v_j)$ 和 $d_A(v_i, v_j)$ 分别测量结点间的结构相似性和属性相似性, 而 α, β 分别表示两者的权重。这种看似简单的方法却很难设置/调整 α 与 β 的值以权衡结构及属性相似性的重要性。现提出一种新的方法对结点结构及属性相似性进行测量, 以权衡这两个看似独立, 甚至相互矛盾的性质, 并在此基础上提出了一种新的聚类方法。

2 相关定义

2.1 非加权无向图

一个非加权无向图 $G=(V, E)$ 是由结点的有穷非空集合 V 和边的集合 E 构成的二元组。其中 $V=(v_1, v_2, \dots, v_{|V|})$, 其中 $|V|$ 表示图中结点的个数, $E=\{(v_i, v_j) | v_i, v_j \in V\}$ 是边的集合, 即 V 中结点的偶对。

2.2 属性扩展图

为了表示结点的属性特性,在图 G 中增加一些新的结点以表示结点的属性.新结点 v_{jk} 代表了一个属性值对 (a_j, a_{jk}) ,如果结点 v_i 在属性 a_j 上具有属性值 a_{jk} ,则在 v_i 与 v_{jk} 之间增加一条新的边.

属性图 $G=(V, E, A)$,其中 $A=\{a_1, a_2, \dots, a_m\}$. 属性 a_i 能取得的所有属性的值域为 $Dom(a_i)=\{a_{i1}, \dots, a_{in_i}\}$,其中 $|Dom(a_i)|=n_i$. 属性扩展图定义为 $G_a=(V \cup V_a, E \cup E_a)$,其中, $V_a=\{v_{ij}\}_{i=1, j=1}^{m, n_i}$ 代表新添加结点的集合.新添加结点 v_{ij} 代表第 i 个属性取得了第 j 个属性值.如果存在 $a_j(v_i)=a_{jk}$,则边 $(v_i, v_{jk}) \in E_a$,表示结构结点 v_i 在新结点 a_j 上具有属性值 a_{jk} .存在边 $(v_i, v_j) \in E$ 则这条边为结构边,存在 $(v_i, v_{jk}) \in E_a$ 则这条边为新添加边.在定义属性扩展图后,结点间的属性相似性还可以表示为:在同一个属性结点上,两个结点具有相同的属性值.到此,在属性扩展图中,如果结构结点 v_i 与 v_j 通过其他的结构结点相连接,或者是它们拥有许多相同的属性值,或者两者都具备,则说明结点 v_i 与 v_j 是相邻的.

2.3 相似度矩阵

属性扩展图的相似度矩阵 P_A 是 $|V \cup V_a| \times |V \cup V_a|$ 维的矩阵,结点 v_i 与结点 v_j 的相似度计算方法如下:原有的边 $(v_i, v_j) \in E$ 与新添加边 $(v_i, v_{jk}) \in E_a$ 的权重具有不同的测定方法,同样的 m 个属性值的重要性也会不同,在计算相似度的过程中同样发挥不同的作用.定义图中原存在的边具有统一的权重 w_0 ,对于按不同属性值添加的边拥有不同的权重值,属性 $a_1, a_2, \dots, a_m$ 的权重依次为 $w_1, w_2, \dots, w_m$.如果结点 v_i 与 v_j 之间存在一条边,则它们之间的相似度为

$$P_{v_i v_j} = \frac{w_0}{|N(v_i)| + w_1 + w_2 + \dots + w_m},$$

$N(v_i)$ 为结点 v_i 的度;类似的,结点 v_i 与新添加结点 v_{jk} 的相似度为

$$P_{v_i v_j} = \frac{w_j}{|N(v_i)| + w_1 + w_2 + \dots + w_m};$$

新添加结点 v_{jk} 与结点 v_j 的相似度为

$$P_{v_{jk} v_j} = \frac{1}{|V(v_{jk})|};$$

最后,两个属性结点 v_{ip} 与 v_{jp} 间相似度为

$$P_{v_{ip}, v_{jp}} = 0.$$

2.4 邻接随机移动距离矩阵

图 G 中,两结点是否是相邻近的取决于这两结

点的连通性,如果结点 v_i 与结点 v_j 间有多条路径相连通则说明这两结点是相邻近的,反之,是不相邻近的.本文用邻近随机移动距离来测试结点是否是相邻近的^[2].

矩阵 P 为图 G 的 $N \times N$ 阶移动概率矩阵, l 为随机移动的长度, $c \in (0, 1)$ 为参数.则结点 v_i 向邻居结点 v_j 随机移动的距离 $d(v_i, v_j)$ 定义如下:

$$d(v_i, v_j) = \sum_{\substack{\tau: v_i \rightarrow v_j \\ \text{length}(\tau) \leq l}} p(\tau) c (1-c)^{\text{length}(\tau)} \quad (1)$$

式中, τ 为结点 v_i 到结点 v_j 的路径,该路径长度为 $\text{length}(\tau)$,其相似度为 $p(\tau)$.

邻接随机移动的距离矩阵为

$$R^l = \sum_{\gamma=1}^l c (1-c)^{\gamma} P^{\gamma} \quad (2)$$

在属性扩展图 G 中, P_A 为该图的随机移动概率矩阵,则图 G 的邻接随机移动距离矩阵为

$$R_A^l = \sum_{\gamma=1}^l c (1-c)^{\gamma} P_A^{\gamma} \quad (3)$$

3 聚类算法

3.1 聚类算法描述

给定属性扩展图 G ,随机移动的长度 l ,及参数 c ,及聚类个数 k .聚类算法的执行步骤如下:

(1) 初始化属性权重值 $w_1 = \dots = w_m = 1.0$ 及结构权重值 $w_0 = 1.0$.

(2) 计算随机移动距离矩阵 R_A^l .

(3) 扫描矩阵 R_A^l ,选择其中的较大值进行聚类,根据不同的结点可按以下四种情况进行:

① 该较大值所对应的两结点都不属于任何一个簇,则将这两个结点分为一个簇;

② 两结点中有一个结点已属于一个簇,则将另一个结点也归入该簇;

③ 两结点同时属于同一个簇,则返回;

④ 两结点分别属于不同的簇,分别计算这两个结点属于不同簇及将两个簇合并后的目标函数值,取目标函数值较大的一种情况.

不断的将结点划分到各个簇,并合并不同的簇,当簇的个数达到 k 时,结束本次迭代.

(4) 如果目标函数不在发生变化,则聚类结束,否则按 2.3 节中提出的方法更新属性权重值 $w_1, w_2, \dots, w_m$,返回(2).

(5) 输出聚类结果.

3.2 聚类目标函数

聚类的目的在于最大化簇内各结点的相似

性^[3],而最小化簇间各结点的相似性,根据这个原则进行聚类目标函数的设定.本文中采用邻近随机移动距离测量各结点间的距离,因此目标函数设定为最大化簇内的随机移动距离,即最大化簇内各结点的相似度.

结点集 V_1 和 V_2 为两个不同的结点集,它们之间的距离 $d(V_1, V_2)$ 定义如下:

$$d(V_1, V_2) = \sum_{v_i \in V_1, v_j \in V_2} \frac{d(v_i, v_j)}{|V_1| \times |V_2|} \quad (4)$$

式中, $d(v_i, v_j)$ 为结点 v_i 与结点 v_j 之间的随机移动距离.该公式可定量计算两个结点集之间的相似度.

给定属性扩展图 $G=(V, E, A)$, 各属性边的权重 $W=\{w_1, w_2, \dots, w_m\}$, 随机移动的长度 l 及参数值 c , 将图 G 划分为 k 个簇, 即 $\{V_i\}_{i=1}^k, V_i \cap V_j = \emptyset$ 且 $\bigcup_{i=1}^k V_i = V$, 聚类目标函数定义如下:

$$O(\{V_i\}_{i=1}^k, W) = \sum_{i=1}^k d(V_i, V_i) \quad (5)$$

当该目标函数取得最大值时, 将取得最好的聚类结果.

3.3 自动更新权重值

文中通过最大化目标函数得到最好的聚类结果, 现提出一种自动更新权重值的方法, 在迭代的过程中取得最大的目标函数, 从而得到最好的聚类结果.

设定 $w_0=1.0$ 为定值, 而在迭代的过程中不断更新结点属性的权重值 $\{w_1, w_2, \dots, w_m\}$. 设定 $W^t = \{w_1^t, \dots, w_m^t\}$ 为第 t 次迭代的权重值, 且设定 $w_1^0 = w_2^0 = \dots = w_m^0 = 1.0$. 通过属性 a_i 第 t 迭代所得的权重值与增量 Δw_i^t 来更新权重值 w_i^{t+1} . 属性 a_i 第 $t+1$ 迭代的权重值计算如下:

$$w_i^{t+1} = \frac{1}{2}(w_i^t + \Delta w_i^t) \quad (6)$$

如果一个簇中的多个结点在属性 a_i 上具有相同的属性值, 说明属性 a_i 具有好的聚类趋势, 属性 a_i 的权重值将增加; 如果一个簇中仅仅有几个结点在属性 a_i 上具有相同的属性值, 则说明属性 a_i 不是好的聚类属性, 该属性的权重值将减小. 现定义 $R(v_p, v_q)$ 来测试两个结点是否具有相同的属性值. 如果结点 v_p 与结点 v_q 在属性 a_i 上具有相同的属性值, 则: $R(v_p, v_q)=1$, 否则为 0.

Δw_i^t 通过同一个簇中 $R(v_p, v_q)$ 的数量来进行计算. $R(v_p, v_q)$ 的数量越大, Δw_i^t 的值则越大. Δw_i^t 的定义如下:

$$\Delta w_i^t = \frac{\sum_{j=1}^k \sum_{v_p, v_q \in V_j} R(v_p, v_q)}{\frac{1}{m} \sum_{p=1}^m \sum_{j=1}^k \sum_{v_p, v_q \in V_j} R(v_p, v_q)} \quad (7)$$

则更新后的权重值 w_i^{t+1} 定义为

$$w_i^{t+1} = \frac{1}{2}(w_i^t + \Delta w_i^t) = \frac{1}{2} \left(w_i^t + \frac{m \sum_{j=1}^k \sum_{v_p, v_q \in V_j} R(v_p, v_q)}{\sum_{p=1}^m \sum_{j=1}^k \sum_{v_p, v_q \in V_j} R(v_p, v_q)} \right) \quad (8)$$

属性权重是增大、减小的或者不变依赖于 Δw_i^t 的值.

自动更新权重值的目标是在不断更新权重值的同时目标函数也在不断增大. 如果在同一个簇中的大量结点在属性 a_i 上具有相同的属性值, 权重将会增大, 相反, 如果同一个簇中的结点在属性 a_i 上拥有不同的属性值, 则权重将会减小. 由于 w_i^{t+1} 的增大, 在属性 a_i 上具有相同的属性值的不同结点间的随机移动距离将增大, 这些结点划分到同一簇中的可能性将增大, 同样目标函数也会增大.

4 算法例证

如图 1 所示, 图中存在 14 个结点分别为: $r_1 \dots r_{14}$, 其中结点 r_1, r_2, r_3, r_4, r_5 具有特有属性 a_1 , 结点 $r_7, r_8, r_9, r_{10}, r_{11}, r_{12}, r_{13}, r_{14}$ 具有特有属性 a_2 , 而结点 r_6 即具有属性 a_1 同时也具有属性 a_2 . 如果将特有属性 a_1 与 a_2 分别看作结点 v_{15}, v_{16} 加入图中. 各个原始结点具有不同的属性, 在各个结点与其相对应的新添加结点间添加一条边, 所得如图 2 所示.

初始化权重值 $w_0=w_1=1.0$, 计算该图的相似矩阵 P_A 为

$$P_A = \begin{matrix} & \begin{matrix} r_1 & r_2 & \dots & r_{13} & r_{14} & v_{15} & v_{16} \end{matrix} \\ \begin{matrix} r_1 \\ r_2 \\ \dots \\ r_{13} \\ r_{14} \\ v_{15} \\ v_{16} \end{matrix} & \left\{ \begin{array}{ccccccc} 0 & 1/3 & \dots & 0 & 0 & 1/3 & 0 \\ 1/4 & 0 & \dots & 0 & 0 & 1/4 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & \dots & 1/3 & 0 & 0 & 1/3 \\ 1/6 & 1/6 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 1/9 & 1/9 & 0 & 0 \end{array} \right\} \end{matrix}$$

设定 $l=2, c=0.2$, 在相似矩阵的基础上计算图 G 的邻接随机移动距离 R_A^l . 其中 $R_A^l(r_7, r_{11})=0.128$, 表示结点 r_7 以概率 0.128 经过属性结点 v_{16}

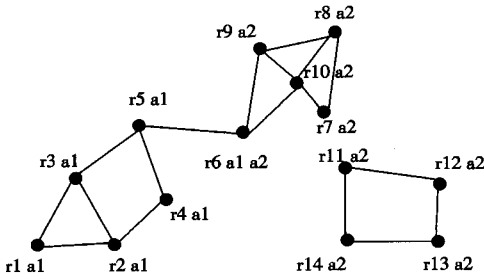


图1 原始图

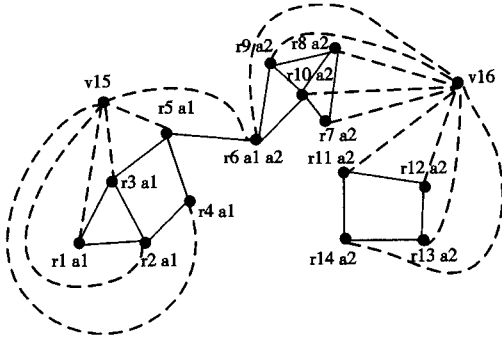


图2 属性扩展图

可到达结点 r_{11} , 其跳转的步数为 2. 如果仅仅按结构图, 结点 r_1 与结点 r_5 之间是不邻接的, 因为它们之间没有可以到达的路径, 而通过新添加结点 v_{16} , 两结点之间是相邻近的.

给定聚类数 $k=2$, 分别用不同的方法及文中所提的聚类方法对图 1 中的数据进行聚类, 得到的结果如图 3~图 5.

图 3 所示的是基于图拓扑结构聚类所得的结果, 具有不同属性的结点划分到了同一簇中.

图 4 所示的是基于属性相似性聚类所得的结果, 该聚类方法将所有具有相同属性的结点划分为同一簇, 但聚类结果破坏了原来图的拓扑结构, 使一些结点孤立了出来.

图 5 是本文所提出的基于结构及属性相似性进行聚类所得的结果. 该结果同时权衡了结点结构及

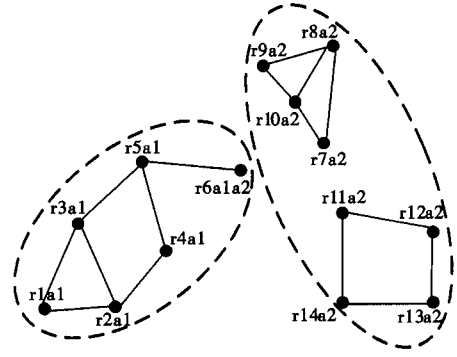


图4 基于属性的聚类结果

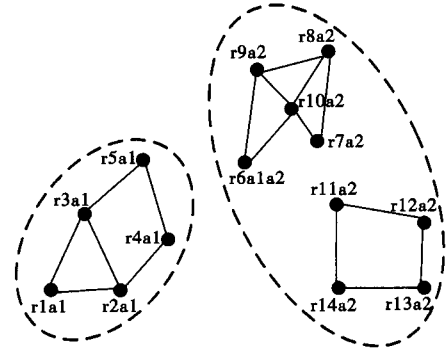


图5 基于结构及属性的聚类结果

属性的相似性, 在同一个簇中所有结点同时具有相同的属性, 并且图的拓扑结构也得到了很好的体现. 这正是我们想要的结果.

5 结束语

本文提出了一种新的权衡结点结构及属性相似性的方法, 将结点结构及属性两个看似独立甚至矛盾的性质结合起来, 并运用随机移动距离对结点间的相似性进行计算, 在此基础上提出了一种新的聚类算法, 该算法的目标在于最大化簇中结点的相似性. 通过相关实验充分说明了该聚类算法的有效性.

如何减小本文中所提算法的复杂度及在聚类过程中自动形成 k 个簇而不是事先确定等, 将是以后研究的重点.

参考文献:

- [1] Jeh G, Widom J. SimRank: a measure of structural-context similarity[C]// Proc 2002 ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, Edmonton, Canada: ACM, 2002:538-543.
- [2] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2000:888-905.

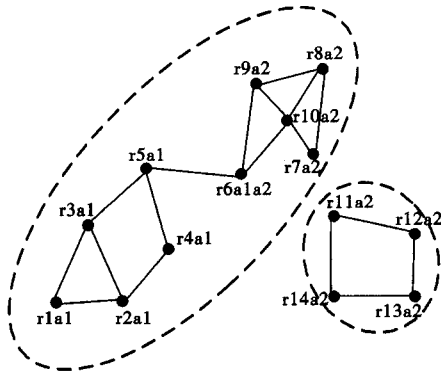


图3 基于结构的聚类结果

点增多,路由修复难度加大,平均跳数自然增加.对于大于0.2时,路由跳数减少的情况,分析认为当失效节点增多时,一次可能会绕过多个失效节点,路由跳数才未明显增加.图5显示了不同失效率下MA迁移能耗情况.虽经历修复过程,但仿真结果表明能耗并未明显增加,失效率0.2时仅增加18%.当前结果表明算法具有路由修复和低能耗的优点.

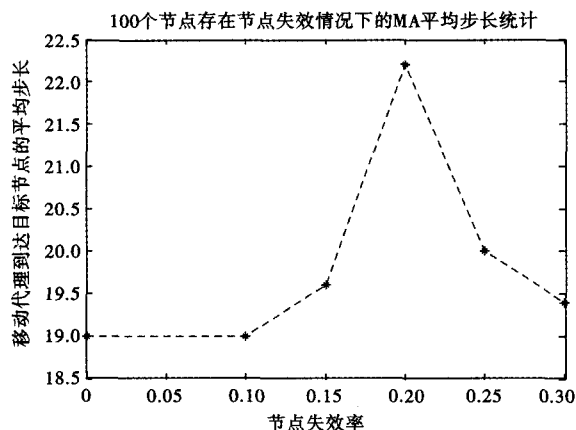


图4 不同失效率下的平均路由跳数

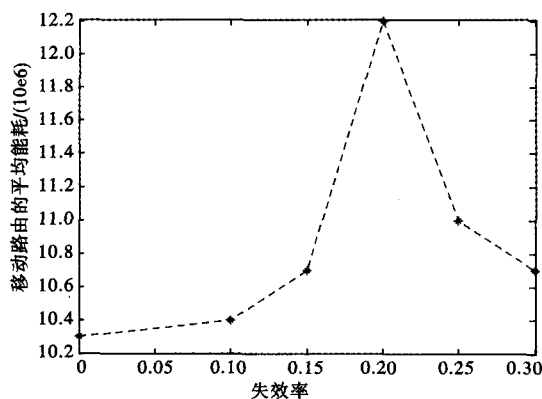


图5 不同失效率下MA路由平均路由能耗

6 结束语

针对基于移动代理的无线传感器中移动代理的

迁移路由问题,本文提出一种基于改进的自适应遗传算法的路由获取算法,并通过网格化和增加路由修复策略来提高算法适应网络拓扑变化的能力.但算法还需进一步优化处理,以获得更好的性能.

参考文献:

- [1] Hai-rong Q, Fei-yi W. Optimal itinerary analysis for mobile agents in ad hoc wireless sensor networks [C] // Proc 13th International Conference on Wireless Communications, Wireless' 2001. [s. l.]: IEEE, 2001: 147-153.
- [2] 李志宇,史浩山. 一种基于网格和移动代理的无线传感器网络数据融合算法[J]. 传感技术学报, 2008, 7(21):1232-1236.
- [3] Nen-Chung W, Yung-Fa H, Joug-shin C, et al. Energy-aware data aggregation for grid-based wireless sensor networks with a mobile sink [J]. Wireless pers. Commun. 2007, 43(2007): 1539-1551.
- [4] 刘向东,宋欣,王翠荣. 基于动态人工鱼群优化的 WSN 分簇算法[J]. 微电子学与计算机, 2011, 28(8):43-46.
- [5] Al-Karaki J N, Kamal A E. Efficient virtual-backbone routing in mobile ad hoc networks [J]. Computer networks, 2008, 52(2008):327-350.
- [6] 周艳,李海成. 基于 RSSI 无线传感器网络空间定位算法[J]. 通信学报, 2009, 30(6):75-79.
- [7] 王森,白文乐. 基于无线传感器网络的 TDOA-CTLS 定位算法研究[J]. 微电子学与计算机, 2010, 27(4): 149-152.
- [8] 王小平,曹立明. 遗传算法-理论、应用与软件实现 [M]. 西安:西安交通大学出版社,2002.

作者简介:

刘欢笑 女,(1983—),硕士.研究方向为软件工程、计算智能(移动代理方向).

杨晓飞 男,(1983—),博士.研究方向为无线传感器网络、电路设计.

(上接第113页)

- [3] Pons P, Latapy M. Computing communities in large networks using random walks[J]. Journal of Graph Algorithms and Applications, 10(2):191-218, 2006.
- [4] 史变霞,张明新. 一种改进的层次聚类算法[J]. 微电子学与计算机, 2010, 27(12): 87-89.
- [5] 王慧,申石磊. 一种改进的特征加权 K-means 聚类算法[J]. 微电子学与计算机, 2010, 27(7): 161-163.
- [6] 崔军,郭春艳,贾宗维. 一种连通非加权图的快速聚类

方法[J]. 计算机工程与应用, 2008, 44(7):179-181.

作者简介:

任巍英 女,(1986—),硕士研究生.研究方向为图聚类算法.

高媛 女,(1972—),副教授,硕士生导师.研究方向为人工智能与知识发现、软件设计方法及应用、网络技术应用.