一种基于加权复杂网络特征的K-means聚类算法

赵 鹏[1,2],耿焕同[2],蔡庆生[2],王清毅[2]
(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.中国科学技术大学 计算机系,安徽 合肥 230027)

摘 要:在分析了传统的基于划分的K-means聚类算法的优越性和存在不足的基础上,根据近两年复杂网络研究中部分新的理论成果,提出了复杂网络加权度、加权聚集度与加权聚集系数的定义,并将数据聚类转换为复杂网络上的节点聚类,提出基于加权复杂网络特征的K-means聚类算法(简称WCNFC算法)。实验结果表明,该算法根据节点加权复杂网络特征值,能够较好地找到聚类中心,有效地避免了对初始化选值敏感性的问题,从而使得聚类质量大大提高。

关键词:聚类;复杂网络;聚集度;聚集系数
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)09-0035-03

A Novel K-means Clustering Algorithm Based on Weighted Complex Networks Feature

ZHAO Peng[1,2],GENG Huan-tong[2],CAI Qing-sheng[2],WANG Qing-yi[2]
(1.Ministry of Education Key Lab.of Intelligent Computing & Signal Processing,Anhui Univ., Hefei 230039,China;
2.Dept. of Computer Sci. and Tech.,University of Science and Technology of China, Hefei 230027,China)

Abstract:After analyzing the advantages and disadvantages of the traditional partitioned K-means clustering algorithm and based on the new theory results achieved in the field of complex networks, the definitions of weighted degree, weighted clustering degree, and weighted clustering coefficient of complex networks and a novel K-means clustering algorithm based on the weighted complex networks feature were proposed. The clustering of datum was transformed into clustering of nodes in complex networks. The experimental results show that this algorithm can find clustering centers better based on the weighted complex networks feature of nodes and it is robust to initialization, so the quality of clustering is improved greatly.

Key words:clustering;complex networks;clustering degree;clustering coefficient

0 引 言

聚类是指将事物按有关特性的相似程度进行分组的过程。聚类在数据挖掘、图像分割、模式识别、空间遥感技术、特征提取和信号压缩等诸多领域中有着广泛的应用。K-means聚类算法[1~3]是由MacQueen提出的基于划分的聚类算法,该算法是目前应用最为广泛的聚类算法之一。它具有算法简单且收敛速度快的特点。但是该算法的性能依赖于聚类中心的初始位置,即对于随机的初始值选取可能导致不同的聚类结果,甚至存在无解的情况,而且算法对孤立点和噪声数据很敏感。

复杂网络自20世纪末逐渐兴起以来,正迅速地在深度和广度上与其它学科进行交叉[4~8]。一方面,从现实世界存在的网络中不断地发现新结构与新现象,大量重要的应用问题涌现出来;另一方面,以研究复杂网络一般规律为目标的理论研究工作也迅速发展,不断提出新的理论模型和新的分析方法。文献[4]对复杂网络中的几个经典模型及其聚集性质作了理论的研究与细致的刻画,并取得了一些新的理论成果。然而对于这些新的理论成果的应用却未见于文献。

笔者在对复杂网络的重要特征研究的基础上,提出了加权复杂网络特征的定义,并将加权复杂网络特征的度和聚集性质应用于基于划分的聚类分析中初始值的选取上,提出了基于加权复杂网络特征的K-means聚类算法。实验结果表明,该算法根据节点加权复杂网络特征值,能够较好地找到聚类中心,有效地

收稿日期:2006-11-14
基金项目:国家自然科学基金项目(70171052);安徽省高校青年教师基金项目(2006jq1040)
作者简介:赵 鹏(1976-),女,安徽安庆人,博士研究生,讲师,研究方向为人工智能;蔡庆生,教授,博士生导师,研究方向为人工智能、机器学习、复杂系统。

避免了对初始化选值敏感性的问题，从而使得聚类质量大大提高。

1 相关理论

1.1 复杂网络及其重要特征

复杂网络无所不在，包括：互联网、生物的神经网络、人类的食品供应网络、细胞新陈代谢系统、飞机航班的时刻表、公司之间合作的关系、人们参加相同社会活动的网络，甚至语言中词与词之间的语法联系等等。

复杂网络是一个颇难界定的概念，研究者们还没有对其给予严格的定义。目前，学界公认的复杂网络的重要特征有度分布、平均最短路径、聚集度与聚集系数。文中提出的聚类算法利用了复杂网络中节点的度与聚集系数这两个重要特征。

节点的度是指该节点与其它节点相关联的边数。节点的聚集系数是指与该节点相连的近邻节点之间互连的比例。其定义如下[5]：

定义 1：设 $V=\{v_1,v_2,\cdots,v_N\}$ 为一节点集合，令无序偶对 (v_i,v_j) 表示节点 $v_i\in V$ 与 $v_j\in V$ 之间的边，设 $G(V,E)$ 是以 V 为节点集合，以 $E\subset\{(v_i,v_j):v_i,v_j\in V\}$ 为边集合的图，节点 v_i 的度 D_i 为：

$$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}| \tag{1}$$

节点 v_i 的聚集度 K_i 为：

$$K_i=|\{(v_j,v_k):(v_i,v_j)\in E,(v_i,v_k)\in E,v_j,v_k\in V\}| \tag{2}$$

节点 v_i 的聚集度 C_i 为：

$$C_i=K_i/\binom{D_i}{2}=\frac{2K_i}{D_i(D_i-1)} \tag{3}$$

聚集度和聚集系数反映了网络的模块性质。同一模块内部节点互连度高，聚集性强，而处于模块之间的节点聚集系数较弱。节点的聚集度和聚集系数体现了在此节点局部范围内的互相连接密度，而对很多网络，例如蛋白质相互作用网络，其节点的连接代表节点属性上的某种相似性，因此可以利用聚集度和聚集系数为特征，对网络的节点进行聚类。

1.2 加权复杂网络及其重要特征

目前对复杂网络的研究主要针对非加权复杂网络，即只考虑所有边的权值都一样的情况。而现实的网络中，边的权值往往是不一样的，并且会影响整个网络的性能。加权复杂网络能够比较完整地表达复杂网络的结构。对照第 1.1 节中节点度、聚集度和聚集系数的定义，提出加权复杂网络中节点加权度、加权聚集度和加权聚集系数的定义。

定义 2：设 $V=\{v_1,v_2,\cdots,v_N\}$ 为一节点集合，令无序偶对 (v_i,v_j) 表示节点 $v_i\in V$ 与 $v_j\in V$ 之间的边，w_{ij} 为边 (v_i,v_j) 的权值。设 $WG(V,E,W)$ 是以 V 为节点集合，以 $E\subset\{(v_i,v_j):v_i,v_j\in V\}$ 为边集合，以 $W=\{w_{ij}:(v_i,v_j)\in E\}$ 为权值集合的图，节点 v_i 的加权度 WD_i 为：

$$WD_i=\sum_{(v_i,v_j)\in E}w_{ij} \tag{4}$$

节点 v_i 的加权聚集度 WK_i 为：

$$WK_i=\sum_{(v_j,v_k)\in R}w_{jk} \tag{5}$$

其中 $R=\{(v_j,v_k):(v_i,v_j)\in E,(v_i,v_k)\in E,v_i,v_j,v_k\in V\}$

节点 v_i 的聚集系数 WC_i 为：

$$WC_i=WK_i/\binom{D_i}{2}=\frac{2WK_i}{D_i(D_i-1)} \tag{6}$$

节点的加权度反映了该节点与其它节点的连接强度。节点的加权聚集系数则体现了此节点局部范围内的相互连接密度和强度。

2 基于加权复杂网络特征的 K－means 聚类算法

为了克服传统的基于划分的 K－means 聚类算法对初始选值敏感性的缺点，提出基于加权复杂网络特征的 K－means 聚类算法（简称 WCNFC 聚类算法）。WCNFC 聚类算法将聚类问题转换为复杂网络上节点聚类的问题。复杂网络中作为聚类中心的节点不仅具有与同类其它节点较强的连接强度，而且与之相连的节点之间也具有较大的相互连接密度和强度，即具有较强的局部聚集性。因此，WCNFC 算法选取加权聚集系数和加权度高的节点作为聚类的初始中心节点，然后采用基于划分的方法对网络上节点进行聚类。WCNFC 聚类算法主要步骤如下：

STEP1：根据原始数据集计算相似度，得到相似度矩阵。

STEP2：根据相似度矩阵建立加权复杂网络。

首先以数据为节点，相似度作为数据之间连边的权值，表示数据之间的连接强度，然后将权值小于阈值 β 的弱连接边删去。这样在保证网络性能的前提下，大大简化了计算量。

STEP3：计算各个节点的加权复杂网络综合特征值。

首先计算各个节点的加权度 WD_i，加权聚集系数 WC_i，然后计算得到该节点的加权网络综合特征值（公式 7）WCF_i：

$$WCF_i=\alpha WC_i+(1-\alpha)WD_i/N \tag{7}$$

其中 α 为可调节的参数，$0<\alpha<1$，N 为网络中

节点个数。

STEP4:对各节点的加权复杂网络综合特征值进行排序,形成由大到小的队列 queue。

STEP5:从队列 queue 中依次选取 k 个加权网络综合特征值,且与已被选作初始聚类中心的节点之间没有连边的节点作为初始聚类中心。

STEP6:以所选的 k 个节点作为初始聚类中心,采用 K－means 算法,根据相似度矩阵对数据集进行划分迭代,形成聚类。

3　实验结果与分析

为了测试 WCNFC 聚类算法的有效性,文中分别用 K－means 算法和 WCNFC 算法作了比较实验。

实验使用的数据来源于"中文自然语言处理开放平台"网站提供的中文文档分类测试集,该测试集共分为 10 个类别:环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治。从每个类别中随机取出 20,40,60,80,100 篇文档,来组成 5 个文本测试集 D_1,D_2,D_3,D_4,D_5,其中 $|D_1|=200$、$|D_2|=400$、$|D_3|=600$、$|D_4|=800$、$|D_5|=1000$。

实验采用向量空间法将每个文本表示成关键词向量,即每个文档 $d_i=(t_{1i},t_{2i},\cdots,t_{mi})$,其中 t_{ji} 表示关键词 j 在文档 i 中的权重,这里采用 TFIDF 值[9] 作为权重。

实验相似度函数采用夹角余弦式(8):

$$\mathrm{Sim}(d_i,d_j)=\frac{\sum_{k=1}^{m}t_{ki}t_{kj}}{\sqrt{\sum_{k=1}^{m}t_{ki}^2\sum_{k=1}^{m}t_{kj}^2}} \tag{8}$$

其中 $d_i\neq d_j$。

实验采用纯净度(Purity)作为聚类结果的一个质量指标,其定义如下:

定义 3:对于数据集 D,令 P 表示 D 上的一个聚类结果(即一个划分),$p\in P$,表示该划分中的一个簇;令 L 表示 D 上的一个手工分类的结果(也是一个划分),$l\in L$,表示该手工分类的一个类别。则有:

对 $\forall p\in P$,p 相对于 $l\in L$ 的查准率定义为:

$$\mathrm{Precision}(p,l)=\frac{|p\cap l|}{|p|}$$

Purity 定义为:

$$\mathrm{Purity}(P,L)=\sum_{p\in P}\frac{|p|}{|D|}\max_{l\in L}\mathrm{Precision}(p,l)$$

由定义 3 可以看出,Purity 是各个簇的查准率的加权平均值,而各簇的查准率取决于它相对于各手工分类类别的最大查准率,因此,该指标能比较全面地反映聚类的效果。

实验分别设定聚类簇数目 $k=12,15$,迭代的最大次数设为 10,其中每种情况下,K－means 算法分别重复执行 20 次,以获得平均性能指标,WCNFC 算法中设定参数 $\alpha=0.5$,$\beta=0.2$。由实验结果(见表 1、表 2)可以看出,WCNFC 聚类算法所获得的聚类纯净度要高于 K－means 聚类算法,平均聚类纯净度高出近 10%,并且 WCNFC 算法的性能比较稳定,与 K－means 算法相比,纯净度上下波动较小。实验结果显示:节点的加权复杂网络特征值能够更加完整、准确地表达节点在聚类中的中心地位。WCNFC 算法根据节点的加权复杂网络特征值,较好地选取了聚类中心初始值,从而使得聚类质量大大提高。

表 1　聚类结果纯净度对比表($k=12$)

算法	1000	800	600	400	200
K－means	56.74%	56.23%	58.76%	51.34%	52.11%
WCNFC	65.38%	65.12%	65.43%	64.16%	64.71%

表 2　聚类结果纯净度对比表($k=15$)

算法	1000	800	600	400	200
K－means	61.87%	61.23%	61.45%	58.69%	59.02%
WCNFC	71.03%	71.89%	71.24%	70.13%	70.65%

4　总　结

根据近两年复杂网络研究中一部分新的理论成果,提出了复杂网络加权度、加权聚集度与加权聚集系数的定义,并根据节点的加权度与加权聚集系数提出了基于加权复杂网络特征的 K－means 聚类算法。实验结果表明,根据节点加权复杂网络特征值,能够较好地找到聚类中心,从而使得聚类质量大大提高。下一步的工作是针对大规模的数据集,如何降低加权复杂网络特征值的计算复杂度,设计出高效的聚类算法,投入到实际工程应用中。

参考文献:

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000.

[2] Ordonez C, Omiecinski E. Efficient disk－based K－means clustering for relational databases[J]. IEEE Trans. Knowledge and Data Engineering, 2004, 16(8): 909－921.

[3] Ordonez C. Clustering binery data streams with K－means [C]//ACM DKMD Workshop. San Diego, California: [s. n.], 2003.

[4] YAO Xin. Studies on Complex networks and its Clustering Degree[D]. Beijing: Tsinghua University, 2005.

[5] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167－256.

(下转第 40 页)

的方差小，因此，AR 模型确实能够有效地预测主机负载。

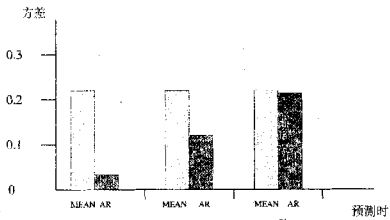

图 2 AR 模型和 MEAN 模型

此外，预测器的初始化时间和预测时间也是评价预测模型的一个重要因素[9,10]。对在所用的几种模型的初始化时间和预测时间进行比较后，结果如图 3 所示。AR 模型能够在 0.1 秒的时间内初始化完毕，并且在能够接受的足够短的时间内完成预测任务。

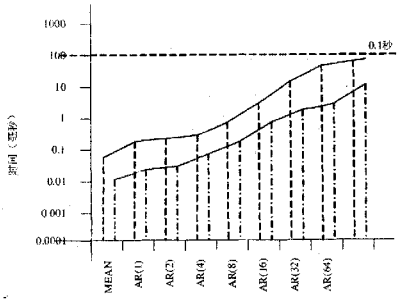

图 3 预测器初始化时间和预测时间比较图

4 结 语

在多种行业中得到广泛应用的线性时间序列[11]同样也可以被运用到分布式实时系统中，根据线性时间序列的统计功能，利用过去时刻的负载在较准确的范围内计算将来可能发生的负载，使预测分布式实时系统的在未来一段时间内的状况成为可能[12]。文中利用一种简单的线性时间模型——AR 模型在一个较小误差的范围内完成预测任务，还证实了 AR 模型能够在较短的时间内完成模型的初始化和预测任务，具有高效性和稳定性的特点。

如果能够较准确地预测主机负载，分布式实时系统的调度器就能够为任务选择主机，为预测任务在某台主机上的运行时间提供了前提。利用主机负载预测任务在某台固定主机上的运行时间，根据预测结果调度任务，最终不仅能够满足尽可能多的任务的最终时限，并且能够很大程度上提高系统的性能。

参考文献：

[1] 吴怀宇. 时间序列分析与综合[M]. 武汉：武汉大学出版社，2004.

[2] Liu J W S. 实时系统[M]. 姬孟洛，李 军，王 馨，译. 北京：高等教育出版社，2003.

[3] 姚天任，孙 洪. 现代数字信号处理[M]. 武汉：华中理工大学出版社，1999.

[4] Dinda P A. A Prediction - based Real - time Scheduling Advisor[C]// Proceedings of the 16th International Parallel and Distributed Processing Symposium (IPDPS 2002). Washington D. C. USA：IEEE Computer Society，2002.

[5] Dinda P A，O'Hallaron D R. An Evaluation of Linear Models for Host Load Prediction[C]// Proceedings of the 8th IEEE Symposium on High - Performance Distributed Computing (HPDC - 8). Redondo Beach，CA：[s. n.]，1999.

[6] 屈志坚，陈剑云. 分布式远动调度系统中实时数据库的研究与实现[J]. 微计算机信息，2004(8)：115 - 116.

[7] 徐浩峰，应宏伟，朱向容. 时间序列分析方法预报基坑支撑轴力[J]. 水利学报，2004(1)：105 - 109.

[8] Dinda P A. Online Prediction of the Running Time of Tasks [J]. Cluster Computing，2002，5(3)：225 - 236.

[9] Ghyselsy E，Valkanovz R. Linear Time Series Processes with Mixed Data Sampling and MIDAS Regression Models[EB/OL]. 2006 - 07. http：//ssrn. com/abstract = 920610.

[10] Dinda P A，Lowekamp B，Kallivokas L F，et al. The Case For Prediction - based Best - effort Real - time Systems[C]// Proceedings of the 7th International Workshop on Parallel and Distributed Real - Time Systems (WPDRTS 1999). San Juan：[s. n.]，1999：309 - 318.

[11] 王双强，陈 强，李 江. 基于 AR 模型的车辆声车型自动分类技术[EB/OL]. 2005 - 10. 中国科技论文在线. http：//www. paper. edu. cn/paper. php? serial_number = 200511 - 308.

[12] 刘江辉，徐海水. 基于 RT - CORBA 的任务运行时间预测研究[D]. 广州：广东工业大学，2005.

（上接第 37 页）

[6] Albert R，Barabosi A L. Statistical mechanics of complex networks[J]. Review of Modern Physics，2002，74：47 - 97.

[7] Strogatz S H. Exploring complex networks[J]. Nature，2001，410：268 - 276.

[8] Watts D J，Strogatz S H. Collective dynamics of 'small - word' networks[J]. Nature，1998，393：440 - 442.

[9] Lee D L，Chuang H，Seamons K. Document ranking and the vector - space model[J]. IEEE Software，1997，14(2)：67 - 75.