空间聚类技术研究综述

柳 盛[1], 吉根林[2]

(1. 南京师范大学 虚拟地理环境教育部重点实验室, 江苏 南京 210046;
2. 南京师范大学 计算机科学与技术学院, 江苏 南京 210046)

[摘要] 空间数据挖掘是一种获取空间数据所蕴含知识的方法和技术. 空间聚类是空间数据挖掘的重要研究内容, 有着广泛的应用领域. 介绍了空间聚类算法的分类和性能要求、空间聚类过程和方法. 空间聚类算法主要有基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法以及其它形式的空间聚类算法.

[关键词] 空间数据挖掘, 空间聚类, 聚类分析
[中图分类号] TP311 [文献标识码] A [文章编号] 1672-1292(2010)02-0057-06

A Review of Researches on Spatial Clustering

Liu Sheng[1], Ji Genlin[2]

(1. Ministry of Education Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing 210046, China;
2. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China)

Abstract Spatial data mining is a kind of methods and techniques of obtaining the knowledge inherent in spatial data. Spatial clustering which has a wide area of applications takes up an important part in spatial data mining. This article introduces classification and performance requirements of spatial clustering algorithms, the process and methods of spatial clustering. In general, the major spatial clustering methods can be classified into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods and others.

Key words spatial data mining; spatial clustering; cluster analysis

据统计, 有 80% 以上的数据与地理位置相关. 事实上, 大量的空间数据是从遥感、医疗影像、地理信息系统(Geographic Information System, GIS)、计算机辅助设计(CAD)、物流系统等多种应用中收集而来, 其数据量之大、类型之多、结构之复杂远超过了人脑的分析能力[1]. 由此造成了空间数据虽多, 但知识贫乏的局面. 从这些空间数据中发现领域知识的迫切需求产生一个多学科、多领域综合交叉的新兴研究领域——空间数据挖掘[2]. 空间数据挖掘(Spatial Data Mining)是指从空间数据库中提取隐含的、用户感兴趣的空间和非空间模式、普遍特征、规则和知识的过程[3,4].

空间聚类(Spatial Clustering)是空间数据挖掘的重要组成部分, 是聚类研究在空间数据分析中的应用. 空间聚类应用广泛, 如地理信息系统、生态环境、军事、市场分析等领域. 通过空间聚类可以从空间数据集中发现隐含的信息或知识, 包括空间实体聚集趋势, 分布规律和发展变化趋势等. 如 Andrew 等利用空间聚类分析发现入侵物种分布的变化[5], Wan 等利用空间聚类获得客户群落的划分[6].

空间聚类分析研究引起了国内外学者的高度重视, 相关科研项目得到国家立项, 也提出了许多空间聚类算法. 本文在文献的基础上, 对空间聚类技术进行了综述, 介绍空间聚类算法的分类和性能要求、空间聚类过程和方法.

收稿日期: 2010-04-15.
基金项目: 国家自然科学基金(40871176).
通讯联系人: 吉根林, 博士, 教授, 博士生导师, 研究方向: 数据挖掘技术及其应用. E-mail: glj@njnu.edu.cn

1 空间聚类算法的分类与性能要求

要进行空间聚类,首先要分析空间数据的特性.由于空间数据具有空间实体的位置、大小、形状、方位及几何拓扑关系等信息,使得空间数据的存储结构和表现形式比传统事务型数据更为复杂.空间数据具有如下特性[4]:

(1) 空间属性间的非线性关系.由于空间数据中蕴含着复杂的拓扑关系,因此,空间属性间呈现出一种非线性关系.

(2) 空间数据的尺度特征.空间数据的尺度特征是指在不同的层次上,空间数据所表现出来的特征和规律都不尽相同.

(3) 空间信息的模糊性.空间信息的模糊性是指各种类型的空间信息中,包含大量的模糊信息,如空间位置、空间关系的模糊性,这种特性最终会导致空间聚类结果的不确定性.

(4) 空间数据的高维度.空间数据的高维度性是指空间数据的属性(包括空间属性和非空间属性)个数迅速增加.

目前空间数据聚类的研究主要是依据空间数据的特点对典型的聚类算法进行改进,从而使之适用于空间对象的特性,如 GDBSCAN算法[7]、CLATIN算法[8]、DDSC算法[9]等,或者针对某类空间数据类型进行空间聚类算法研究,如随着 GML(GeographyMarkup Language)数据在 GIS领域的广泛应用,提出了基于拓扑关系的 GML空间聚类算法[10-13].

空间聚类可以根据空间数据源类型、空间分析类型或系统结构类型进行分类.

(1) 按空间数据源类型,空间聚类可分为针对关系空间数据库的空间聚类、针对矢量数据的空间聚类、针对栅格数据(影像、DEM等)的空间聚类、针对 GML数据的空间聚类、针对视频数据等流格式数据的空间聚类等.

(2) 按空间分析类型,空间聚类可分为空间分布聚类、空间方位聚类、空间拓扑聚类、空间趋势聚类等.

(3) 按系统结构分类,空间聚类可分为基于单机的空间聚类、基于 C/S或 B/S结构的空间聚类、基于嵌入式的空间聚类、基于分布式数据库的空间聚类、基于网格的空间聚类等.

空间聚类算法在性能上应具备下列要求[14]:

(1) 可伸缩性.不受限于数据集的大小,聚类算法均能获得良好的结果.

(2) 能处理不同数据类型.

(3) 能发现任意形状的聚类.

(4) 不依赖领域知识等先验知识.

(5) 能处理噪声.聚类算法应该对孤立点、空缺、未知数据或者错误的数据不敏感,即使这些数据存在,也能够得到较好的结果.

(6) 对输入记录的顺序不敏感.

(7) 能处理高维数据.

(8) 能处理基于约束的聚类.实际应用的数据可能包含很多约束,要求算法能在实际应用中各种约束情况下得到较好的结果.

(9) 可解释性和可用性.要求算法得到的聚类结果是可以解释并具有实际意义的.

2 空间聚类过程与方法

空间聚类分析的任务是把空间数据对象分成多个有意义的簇,即根据相似性对数据对象进行分组,使每一个簇中的数据是相似的,而不同簇中的数据尽可能不同,即簇内相识,簇间不同.

2.1 空间聚类的过程

空间聚类的过程和通常的数据挖掘过程类似,如图 1所示.数据变换是利用相关变换或者降维技术对原始数据提取特征集.因为聚类是搜索簇的无监督学习过程,所以需要在聚类操作前确定相似性度量法

则.聚类结果的应用是将聚类结果作为其他挖掘算法的输入,从而得到更深层次的知识.

2.2　空间聚类方法

目前,国内外已有不少学者对空间聚类问题进行了较为深入的研究,提出了多种算法.根据空间聚类采用的不同思想,空间聚类算法主要可归纳为以下几种:基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法以及其他形式的空间聚类算法.

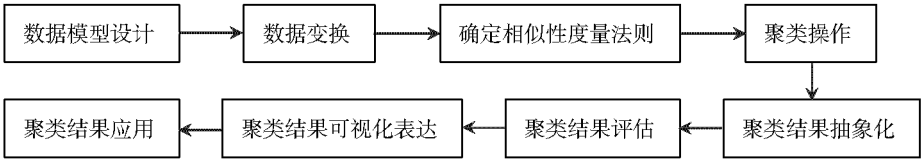

图1　空间聚类一般流程

Fig.1　General process of spatial clustering

2.2.1　基于划分的方法

给定一个包含 n 个对象或数据的集合,将数据集划分为 k 个子集,其中每个子集均代表一个聚类 ($k \leqslant n$),划分方法首先创建一个初始划分,然后利用循环再定位技术,即通过移动不同划分中的对象来改变划分内容,典型的划分方法包括 K-means[15]、K-medoids[16]、CLARA[16]和 CLARANS[17]算法等.

K-means算法是首先从 n 个数据对象随机地选择 k 个对象,每个对象初始地代表了一个簇中心,对剩余的每个对象,根据其与各个簇中心的距离,将它赋给最近的簇,然后重新计算每个簇的平均值.这个过程不断重复,直到准则函数收敛.K-medoids算法选用聚类中位置最中心的对象作为参照点.PAM 算法[14]是在初始选择 k 个聚类中心对象之后,不断循环对每两个对象(非中心对象和中心对象)进行分析,以选择出更好的聚类中心代表对象.CLARA算法是一种将 PAM 和采样过程结合起来的方法,提高了效率,其主要思想是不考虑整个数据集合,只考虑实际数据的一部分.CLARANS算法改进了 CLARA 算法的聚类质量,也拓展了数据处理量的伸缩范围,与 CLARA 算法的本质区别在于 CLARA 在搜索的开始是抽取节点的样本,而 CLARANS在搜索的每一步抽取邻居的样本.

2.2.2　基于层次的方法

层次聚类方法是通过将数据组织为若干组并形成一个相应的树来进行聚类的,可分为自顶向下的分裂算法和自底向上的凝聚算法两种.分裂聚类算法,首先将所有对象置于一个簇中,然后逐渐细分为越来越小的簇,直到每个对象自成一簇,或达到了某个终止条件.而凝聚聚类算法则相反,首先将每个对象作为一个簇,然后将相互邻近的簇合并为一个大簇,直到所有的对象都在一个簇中,或达到了某个终止条件.

AGNES和 DIANA[16]算法是早期的层次聚类方法,前者是一种凝聚聚类方法,后者是一种分裂聚类方法,两者都用各簇间距离度量来合并或分裂簇,在选择合并或分裂点时有一定困难,并且进行合并或分解后不能被撤销,聚类间对象也不能交换,因此会产生错误的簇从而降低聚类质量,且这种方法没有良好的可伸缩性.国内外学者在 AGNES和 DIANA 算法基础上提出了一些新的层次聚类算法,如 BIRCH[18]、CURE[19]、ROCK[21]和 CHAMELEON[21]算法.BIRCH 算法是一种综合的层次聚类方法.BIRCH 算法包括两个阶段,第一个阶段扫描数据库,动态建立一个初始存放于内存的 CF(Clustering Feature)树,CF树可以被看成是对数据的压缩;第二个阶段,采用某个聚类算法对 CF 树的叶节点进行聚类.CURE 算法选择基于质心和基于代表对象方法之间的中间策略,不用单个质心或对象来代表一个簇,而选择数据空间中固定数目具有代表性的点.ROCK算法是利用聚类间的连接进行聚类合并.CHAMELEON 算法是一种探索层次聚类中动态模型的聚类算法,首先利用一个图划分算法将数据对象聚合成许多相对较小的子聚类,然后再利用聚合层次聚类方法,通过不断合并这些子聚类来发现真正的聚类.

2.2.3　基于密度的方法

绝大多数基于划分方法的空间聚类算法都是基于对象之间的距离进行聚类,这类方法只能发现球状的类.基于密度的聚类方法与之不同,其主要思想是只要邻近区域的密度(对象或数据点的数目)超过某个阈值,就继续聚类,这样可以过滤“噪声”数据,发现任意形状的类,代表性算法有 DBSCAN[22]、OP-

TICS[23]和 DENCLUE[24]算法.

DBSCAN算法可以有效地发现具有任意形状的类,并正确地处理噪声数据.对于一个类中的每个对象,在其给定半径的领域中包含的对象不能少于某一给定的最小数目,不进行任何的预处理而直接对整个数据集进行聚类操作.其当数据量非常大时,必须有大内存量支持.该算法对参数 Eps和 M inpts非常敏感,且这两个参数很难确定. OPTICS算法是一种基于类排序方法.该算法并不明确产生一个聚类,而是为自动交互的聚类分析计算出一个增强聚类顺序. DENCLUE算法是一个基于一组密度分布函数的聚类算法.主要思想为:每个数据点的影响可以用一个数学函数来形式化地模拟,它描述了一个数据点在领域内的影响,被称为影响函数;数据空间的整体密度可以被模型化为所有数据点的影响函数的总和;聚类可以通过确定密度吸引点来得到,这里的密度吸引点是全局密度函数的局部最大.

2.2.4 基于网格的方法

基于网格的空间聚类方法采用了一个多分辨率的网格数据结构.该类算法首先将数据空间划分为有限个单元的网格结构,所有的处理都以单个的单元为对象.这样处理的一个突出的优点就是处理速度快,通常与目标数据库中记录的个数无关,只与把数据空间分成多少个单元有关.代表算法有 STING[25]、W aveCluster[26]和 CLIQUE[27]算法.

STING算法是基于网格的多分辨率方法.该种方法效率高,网格结构有利于并行处理和增量更新,但其降低了聚类的质量和精确性. W aveCluster算法也是一个多分辨率的聚类方法.它首先通过在数据空间上强加一个多维网格结构来汇总数据,然后采用一种小波变换来变换原特征空间,在变换后的空间中找到密集区域. CLIQUE算法综合了基于密度和基于网格的聚类方法,自动地发现最高维的子空间,对元组的输入顺序不敏感,不需要假设任何规范的数据分布,它随输入数据的大小线性扩展,当数据维数增加时具有良好的可伸缩性,但聚类结果的精确性一般较低.

2.2.5 基于模型的方法

基于模型的空间聚类方法包括基于统计的空间聚类方法和基于神经网络的空间聚类方法等.如 EM[28]、COBWEB[29]、SOM[30]算法等,是给每一个聚类假定一个模型,然后去寻找能够很好地满足这个模型的数据集.

2.2.6 其他形式的空间聚类算法

除了上述 5种空间聚类算法外,国内外学者根据空间聚类的要求,提出了多种结合其它思想的空间聚类方法,如带约束的空间聚类算法.

带约束的空间聚类算法是为了解决空间聚类中所面临的空间障碍问题而产生的,如城市中的河流、湖泊、道路等障碍,如果在实际分析中不考虑这些障碍(约束),获得的聚类结果必然与实际情况有较大的误差.比较典型的带约束的空间聚类算法有 COD_CLARANS[31], AutoClust+[32], DBCLuC[33]和 DBRS+[34]. COD_CLARANS算法在障碍物约束的条件下,计算任意两样本点的最近距离,将采样技术与 PAM相结合,通过迭代的方法来完成在障碍物约束下的聚类问题,它能够快速处理大量的障碍物,但需要先验知识且不适合大量的空间数据. AutoClust+算法不需要用户提供参数值,但用 Delaunay图处理约束代价高且缺乏灵活性. DBCLuC算法采用障碍线的方法来保证可视空间的不变,降低了障碍对象的处理时间,但该算法对参数的设置非常敏感且依赖经验. DBRS+算法提出了"Chop and Conquer"的方法来处理障碍对象,而随机样例的使用使算法对数据集的随机取例的顺序在一定程度上影响了算法的结果.

3 结语

国内外学者的研究使得空间聚类技术得到蓬勃的发展,但在空间聚类的功能与应用上仍存在不足.目前大部分的空间聚类技术主要针对关系空间数据库,而表达和承载空间数据的方式很多,如通过 Web、GML、图形影像等,因此面向各种不同类型数据的空间聚类研究非常必要.同时,空间数据一般同时包含空间属性(位置、拓扑、方位等)和非空间属性,在空间聚类过程中,如何考虑空间对象的空间属性和非空间属性,使之满足空间数据分析的需求,这也是值得研究的问题.此外,空间聚类技术要与各相关领域应用紧密结合,使之具有更广阔的应用前景.

[参考文献] (References)

[1] Vladimir E C, Lee I. Clustering with obstacles for geographical data mining[J]. ISPRS Journal of Hotogrammetry & Remote Sensing, 2004, 59: 21-34.

[2] Shashi S, Yan H. Discovering spatial colocation patterns: a summary of results[C] // Proc of the Seventh International Symposium on Spatial and Temporal Databases. London: Springer-Verlag, 2001: 236-256.

[3] Shashi S, Sanjay C. 空间数据库[M]. 谢昆青,马修军,杨冬青,译. 北京:机械工业出版社, 2004: 214-215.
Shashi S, Sanjay C. Spatial Databases: A Tour[M]. Xie Kunqing, Ma Xiujun, Yang Dongqing, Translated. Beijing: China Machine Press, 2004: 214-215. (in Chinese)

[4] Han J, Lee J, Kamber M. Geographic DataMining and Knowledge Discovery[M]. 2nd ed. Boca Raton: Taylor and Francis, 2009: 149-152.

[5] Andrew A, Thomas C, Kornissg. Ecological invasion: spatial clustering and the critical radius[J]. Evolutionary Ecology Research, 2007, 9: 375-394.

[6] Wan L, Li Y, Liu W, et al. Application and study of spatial cluster and customer partitioning[C] // Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. Guangzhou: IEEE, 2005: 1 701-1 706.

[7] Sander J, Ester M, Kriegel H P, et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications[J]. Data Mining and Knowledge Discovery[M]. Netherlands: Kluwer Academic Press, 1998: 169-194.

[8] Zhang Q, Couloigner I. A new and efficient *k*-medoid algorithm for spatial clustering[C] // Proceedings of the 2005 ICCSA. Singapore: Springer-Verlag, 2005: 181-189.

[9] Borah B, Bhattacharyya D K. DDSC: a density differentiated spatia clustering technique[J]. Journal of Computers, 2008, 2: 72-79.

[10] Ji G, Miao J, Bao P A. Spatial clustering algorithm based on spatial topological relations for GML data[C] // Proceedings of Intl Conf on Artificial Intelligence and Computational Intelligence. Shanghai: IEEE Computer Society, 2009: 298-301.

[11] Ji G, Miao J, Yang M. A novel spatial clustering algorithm based on spatial adjacent relations for GML data[C] // Proceedings of Intl Workshop on Education Technology and Computer Science. Wuhan: IEEE Computer Society, 2009: 278-281.

[12] Ji G, Zhang L A. Spatial polygon objects clustering algorithm based on topological relations for GML data[C] // Proceedings of Intl Conference on Information Engineering and Computer Science. Wuhan: IEEE Press, 2009: 363-366.

[13] Yang N, Ji G A. Spatial lines clustering algorithm based on adjacent relations for GML data[C] // Proceedings of Int Conf on Information Engineering and Computer Science. Wuhan: IEEE Press, 2009: 3 593-3 596.

[14] Han J, Kamber M. 数字挖掘概念与技术[M]. 范明, 孟晓峰,译. 北京:机械工业出版社, 2002: 224-225.
Han J, Kamber M. Data Minging Concepts and Techniques[M]. Fan Ming, Meng Xiaofeng, Translated. Beijing: China Machine Press, 2002: 224-225. (in Chinese)

[15] Lloyd S P. Least squares quantization in PCM[J]. IEEE Trans Information Theory, 1982, 28: 128-137.

[16] Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley & Sons, 1990.

[17] Ng A R, Han J. Efficient and effective clustering method for spatial data mining[C] //Proceedings of the 1994 Intl Conf Very Large Databases. San Francisco: Morgan Kaufmann, 1994: 144-155.

[18] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C] // Proceedings of the 1996 Intl Conf Management of Data. New York: ACM, 1996: 103-114.

[19] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases[C] // Proceedings of the 1998 ACM-SIGMOD Int Conf Management of Data. New York: ACM, 1998: 73-84.

[20] Guha S, Rastogi R, Shim K. ROCK: a robust clustering algorithm for categorical attributes[C] // Proceedings of the 1999 Int Conf Data Engineering. Washington DC: IEEE Computer Society, 1999: 512-521.

[21] Karypis G, Han E H, Kumar V. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling[J]. Computer, 1999, 32: 68-75.

[22] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases[C] // Proceedings of the 1996 Intl Conf Knowledge Discovery and Data Mining. Amsterdam: Elsevier Science, 1996: 226-231.

[23] Ankerst M, Breunig M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[C] // Proceedings of the 1999 Intl Conf Management of Data. New York: ACM, 1999: 49-60.

[24] Hinneburg A, Keim D A. An efficient approach to clustering in large multimedia databases with noise[C] // Proceedings of the 1998 Intl Conf Knowledge Discovery and Data Mining. San Francisco: Morgan Kaufmann, 1998: 58-65.

[25] Wang W, Yang J, Muntz R. STING: a statistical information grid approach to spatial data mining[C]// Proceedings of the 1997 Intl Conf Very Large Data Bases. San Francisco: Morgan Kaufmann, 1997: 186-195.

[26] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: a multiresolution clustering approach for very large spatial databases [C]// Proceedings of the 1998 Intl Conf Very Large Data Bases. San Francisco: Morgan Kaufmann, 1998: 428-439.

[27] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications[C]// Proceedings of the 1998 ACM SIGMOD. New York: ACM, 1998: 94-105.

[28] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. J Royal Statistical Society, 1977, 39: 1-38.

[29] Gennari J, Langley P, Fisher D. Models of incremental concept formation[J]. Artificial Intelligence, 1989, 40: 11-61.

[30] Kohonen T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 1982, 43: 59-69.

[31] Tung A K H, Hou J, Han J. Spatial clustering in the presence of obstacles[C]//Proceedings of the 2001 ICDE. Washington DC: IEEE Computer Society, 2001: 359-367.

[32] Estivill-Castro V, Lee I. Autoclust+: automatic clustering of point-data sets in the presence of obstacles[C]//Proceedings of the 2000 TSDM. London: Springer-Verlag, 2000: 133-146.

[33] Zaiane O R, Lee C H. Clustering spatial data in the presence of obstacles: A density-based approach[C]//Proceedings of the 2002 IDEAS. Washington DC: IEEE Computer Society, 2002: 214-223.

[34] Wang X, Rostoker C, Hamilton H J. Density-based spatial clustering in the presence of obstacles and facilitators[C]//Proceedings of the 2004 PKDD. New York: Springer-Verlag, 2004: 446-458.

[责任编辑: 严海琳]