基于障碍约束的空间聚类算法综述①

余冬梅

(陕西理工学院 数学与计算机科学学院, 汉中 723000)

摘 要: 传统的空间聚类算法解决的是未带障碍约束的空间数据聚类问题, 而现实的地理空间中经常会存在河流、山脉等阻碍物, 因此, 传统空间聚类算法不适用于带障碍数据约束的现实空间. 在解析了带障碍空间聚类相关概念和定义的前提下, 对带障碍约束条件的空间聚类算法进行梳理, 给出了这类算法的研究历史和沿袭关系, 并把这类算法按七个维度分为四大类, 分析了每类的技术优缺点, 最后给出了带障碍约束的空间聚类算法的未来研究趋向.

关键词: 空间聚类; 障碍约束; 分类; 障碍距离; 聚类算法

Survey of Spatial Clustering Algorithm with Obstacle Constrains

YU Dong-Mei

(School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723000, China)

Abstract: Classical algorithms of spatial clustering are performed in optimal data space without any obstacle. But many obstacle constrains exist in the real-world, such as rivers, mountains, etc. They may affect results of clustering substantially. In this paper, the knowledge of spatial clustering algorithm with obstacle constrains is illustrated in brief. And then, research history and inheritance relation of the algorithms is given. These algorithms are divided into four categories from seven respects. At last, technical feature of every category and trend of spatial clustering algorithm in the presence of obstacles are analyzed.

Key words: spatial clustering; obstacle constrain; category; obstacle distance; clustering algorithm

1 引言

基于空间数据的聚类是在地理空间数据集中, 按照对象间的距离、连接性或者相对密度等指标, 把一些具有一定相似性的数据对象归为一簇, 差异度大的数据对象分派到不同的簇中的过程. 真实地理环境中往往存在着江河、山脉、固定私有区等障碍物, 在空间聚类的分析应用中必须考虑到这些障碍物, 如快递公司投递站的选址问题、城市道路规划等问题, 而传统空间聚类算法没有考虑真实的地理环境特点, 未将障碍物考虑在算法中并加以处理. 但近几年随着空间信息技术的发展, 为了使空间聚类结果更加合理和实用, 有障碍要求的空间聚类算法逐渐成了研究热点, 也为空间数据挖掘、模式识别等领域的发展做出了不小的贡献. 为此, 本文将障碍约束下的空间聚类算法进行了整理分析, 以期为研究者的后继研究提供参考.

传统的空间聚类算法常常被分为基于划分的、基于密度的、基于层次的、基于模型的以及基于网格的等五种方法类型. 虽然这些方法在处理空间数据聚类分析时的侧重点、效果和特点不同, 但均已被实践检验是有效的无障碍约束的空间聚类算法, 也为空间信息研究与应用领域提供了很大的帮助. 基于障碍约束的空间聚类算法大多都是在传统聚类算法的基础上, 增加了障碍距离计算方法, 这些方法改变了传统空间聚类算法中数据之间的距离计算方法, 并且产生了围绕降低计算障碍距离代价为主要目标的改进方法. 经过查阅大量有关考虑障碍约束的空间聚类方法的资料,

① 基金项目:陕西省教育厅科学研究计划(自然科学专项) (14JK1132);陕西省科学技术研究发展计划(2014KJXX-75);汉中市科技发展专项(2013hzzx-38)
收稿时间:2014-04-20;收到修改稿时间:2014-05-12

本文将这些方法大致分为 4 类: 基于划分的方法、基于密度的方法、基于图论的方法、以及混合聚类方法, 将按两个指标对考虑障碍约束的空间聚类算法进行分析, 一是以时间为线索分析它们的发展研究轨迹, 二是从这些算法的特色上进行分类对比分析, 指出它们的共同特点和各自的优缺点, 并给出以后的研究方向.

2 相关概念

2.1 带障碍约束的空间聚类

带障碍约束的空间聚类是指在具有障碍物的空间数据集中, 将具有相似特性(如空间位置相邻)的数据归为一类(簇), 使得同一类(簇)中的数据相似性最大, 不同类(簇)之间的数据相异性最大. 其中在数据相似性的计算判断中, 须考虑障碍物具有分隔其他数据直接连接的特性. 在具有障碍的空间中, 传统的空间聚类算法的聚类效果见图 1, 而考虑障碍约束的空间聚类算法的效果见图 2.

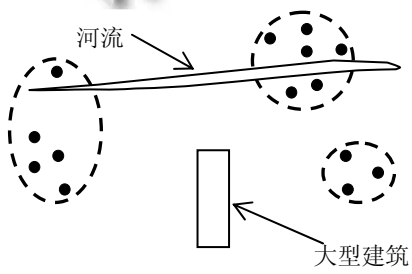

图 1　不考虑障碍约束的空间聚类

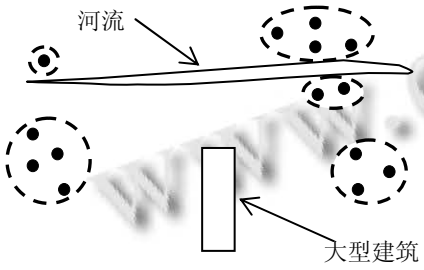

图 2　考虑障碍约束的空间聚类

2.2 带障碍距离

空间数据集中两点间最短障碍距离是指两点间不与任何障碍物相交的最短路径长度, 若使用 Dist(a,b) 表示障碍空间中任意 a,b 两点的带障碍距离, 则:

$$\mathrm{Dist}(a,b)=\begin{cases}a,b\text{间直线段长度}, & a,b\text{间直线段与障碍物不交}\\ a\text{绕过障碍到}b\text{的最短路径}, & a,b\text{间直线段与障碍物相交}\end{cases}$$

如图 3 所示, Dist(a,b)= Dist(a,p) + Dist(p,q)+ Dist(q,b), 其中 p 和 q 分别是障碍物的相对 a 和 b 的可视顶点.

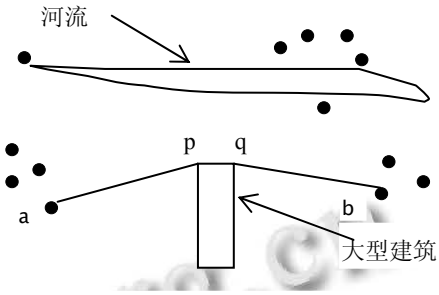

图 3　障碍距离示意图

在带障碍约束空间聚类算法中需要定义自己的带障碍距离计算方法, 不再是单纯的欧式距离, 致使这些算法的时间复杂度往往比传统空间聚类算法更高, 因此, 带障碍距离的计算是解决带障碍约束空间聚类分析问题的最关键技术之一.

3 带障碍约束的空间聚类算法的研究和发展

最早的带障碍约束的空间聚类算法是 2001 年 Anthony K.H.Tung 等提出的 COD-CLARANS 算法[1], 之后不断研究出了很多新算法. 分析这些算法, 绝大多数都是在传统空间聚类算法的基础上, 增加了障碍距离、障碍模型等概念而产生的, 且主要针对的是二维空间维度下点状数据的聚类, 如文献[1]中的 COD-CLARANS 算法是在传统的基于划分的聚类算法 CLARANS[2]的基础上首次引入障碍距离而产生的[3]; 文献[4]的 AUTOCLUST+算法是基于图论的 Delaunay 三角网而提出的; 文献[5]的 DBCLuC 算法是在传统的基于密度的聚类算法 DBSCAN[6]的基础上引入障碍模型而提出的, 即将障碍物模型化成一组多边形; 基于智能优化的障碍约束聚类算法[7-9]是在聚类的过程中加入智能优化算法中的优化模型而形成的. 下面从四种分类角度描述带障碍约束的空间聚类算法的研究发展情况, 如图 4 所示.

4 带障碍约束的空间聚类算法分类

研究中发现, 现有的带障碍约束的空间聚类算法以点目标数据的聚类为主, 未见有针对线目标和面目标数据的障碍约束聚类. 这与传统聚类算法中, 针对线目标和面目标空间聚类算法本身就少有一定关系.

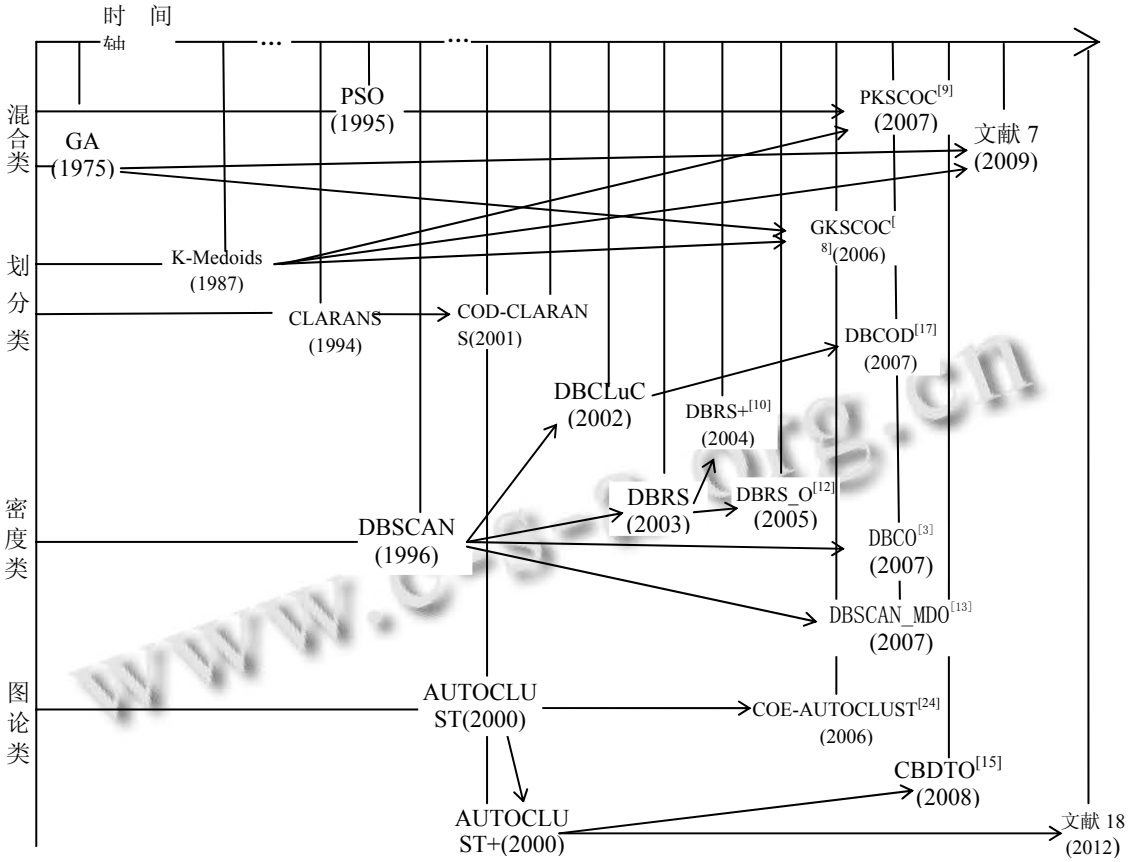

图 4 带障碍约束的空间聚类算法演变发展沿革

4.1 算法的分类方法

本文把考虑障碍约束的空间聚类算法分为 5 类: 划分类、密度类、图论类和混合类. 由于基于障碍约束的空间聚类算法间的典型区别在于, 聚类中数据对象间相似性的计算方法和时间效率等, 因此, 下面从多个维度出发分析这些算法: 依据的传统聚类算法的情况、有无预处理、是否需要输入参数、是否考虑障碍物上有连接点、代表性算法、优点和缺点, 如表 1 所示.

表 1 障碍空间聚类算法分类表

分类	依据的典型传统算法	预处理	需要输入参数值	考虑障碍物上有连接点	代表性算法	典型优点	典型缺点
划分类	CLARANS	有	是	否	COD-CLARANS[1]	在空间聚类中第一个引入障碍距离等概念，能处理大数量数据	预处理的开销非常大、检测球状或者近似球状的簇
	K-means	无	是	否	3S-UK-means[26]	将不确定性数据引入到障碍空间聚类算法中	带障碍距离计算过程复杂，不适合大数量数据空间
密度类	DBSCAN	有	是	有	DBCLuC[5] DBCOD[17]	能处理任意形状的簇，引入障碍域的可视性和可视空间概念	对噪声和数据输入顺序敏感，不能处理障碍物上有连接点的情况
	DBRS	无	是	有	DBRS+[10]	对障碍物处理细致，能处理连接型障碍	时间开销大

混合类	图论类	Delaunay 三角网 (Voronoi 图)	有	否	否	AUTOCLUST+[4] CBDTO[15] 文献 18	能处理任意形状的簇，支持多层关联分析	构造 Delaunay 三角网时处理约束代价大，缺少灵活性
			有	否	有	FOA[11]	能解决障碍物上有连接点的情况	构造 Voronoi 图计算障碍距离耗时大
	基于密度与网格	DBSCAN+网格划分	有	是	否	DCellO[21]	能够进行任意形状的带障碍的聚类,并且不需要指定聚类簇的数目。适合处理多维数据	网格步长大小的设定影响障碍区域的计算; 步长固定后在数据对象的分布密度不同时易造成聚类结果不合理。
	基于划分与智能优化	遗传算法	有	是	否	文献 7 GKSCOC[8]	对中小规模数据集的效率高	处理代价随数据量增长太快且计算速度慢
		群智能算法	有	是	否	PKSCOC[9]	收敛速度较快	容易陷入局部最优解

从以上的分类分析可以得出两个方面的结论. 一是障碍空间聚类算法的优劣很大程度上取决于它所依据的传统空间聚类算法, 也就是说它们继承了传统聚类算法的优点的同时也继承了一定的缺点. 如基于划分方法的典型算法 COD-CLARANS 和基于密度方法的典型算法 DBCLuC, 都由于聚类时需要用户输入相应的参数或阈值, 也正是这些参数或阈值的不同, 其聚类结果均有不同, 要么影响其聚类个数, 要么影响聚类形状. 另一方面, 这些算法中基于密度方法的障碍空间聚类算法的数目明显多于其他方法, 这主要由于传统的基于密度的空间聚类方法在时间复杂度(NlogN)上占有一定的优势.

4.2 算法特点分析

在进行带障碍距离计算的过程中, 这些障碍约束空间聚类算法所依据的空间数据距离计算方法也有很多, 主要有 Euclidean 距离、Voronoi 距离、基于优化算法的距离、Manhattan 距离、基于网格的距离等, 其中 Euclidean 距离的代表是 COD-CLARANS 算法[1]、Voronoi 距离的代表是 AUTOCLUST+算法[4]和 FOA 算法[11]、基于智能优化算法的距离有 GKSCOC 算法[8]和 PKSCOC 算法[9]、Manhattan 距离的代表是 DBSCAN-MDO 算法[13]、基于网格的距离的代表有 DcellO 算法[21]等.

对于智能优化混合类的带障碍约束的空间聚类算法是本文新提出的一种混合聚类方法, 它们的主要原理是, 利用智能优化算法计算空间中数据点间的障碍距离, 借助传统空间聚类算法在障碍距离下进行聚类计算. 这类算法的突出特点是通过智能优化算法自身的收敛特性加快空间障碍距离计算的速度, 已达到降低算法的时间复杂度的目的.

对于密度与网格混合类的带障碍约束的空间聚类算法是本文提出的又一种混合聚类方法, 如 DcellO 算法, 其主要原理是, 将数据对象映射到网格中, 随机选择一个包含数据对象的网格作为起始, 以该网格为核心不断扩展计算它到其他有数据对象的网格的带障碍最短距离, 并将满足传统密度聚类算法中密度半径和半径内数据量阈值要求的对象聚为一类, 之后继续选择其他包含数据对象的网格为新的核心重复上面的步骤, 完成下一轮聚类过程, 直至所有具有数据对象的网格被聚到某一簇为止. 该类算法的共同特点是只对具有数据对象的网格进行带障碍网格距离的计算, 这样可以大大减少计算的网格数量, 以提高计算效率.

纵观所有基于带障碍约束的空间聚类算法, 自从这类算法诞生以来, 绝大多数都是围绕点状数据和非点状障碍物组成的数据空间进行的研究, 而且在这些点状数据中只考虑空间属性, 未考虑非空间属性, 而现实的地理数据空间在应用中往往需要非点状数据的非空间属性的参与, 如在有河流经过的城市中, 依据空间区域中人口数量的分布情况规划银行 ATM 机的选址问题. 因此, 带障碍约束的空间聚类算法的研究与实际应用之间还存在一定的距离, 也为未来的研究指出了方向.

5 结语

由于拥有障碍物是现实空间的真实特点, 因此障碍约束下的空间聚类算法的研究在与空间数据有关的应用领域具有很强的现实意义, 也成为空间聚类算法研究的热点分支之一. 本文在阅读大量的考虑障碍约

束的空间聚类算法的基础上, 对现有的算法进行了分类研究, 总结了它们的共性、优势和不足. 这类算法的未来研究应朝着不仅考虑空间属性, 而且要考虑非空间属性, 以及非点状数据的障碍约束空间聚类方向.

参考文献

1 Tung AKH, Hou J, Han J. Spatial clustering in the presence of obstacles. Proc. of Int. Conf. on Data Engineering (ICDE 01). Heidelberg, Germany. 2001. 359–367.

2 Ng R, Han J. Efficient and effective clustering method for spatial data mining. Proc. of International Conference on Very Large Data Bases(VLDB'94). Santiago, Chile. 1994. 144–155.

3 卢炎生,娄强.障碍空间里基于密度的快速聚类算法.小型微型计算机系统,2007,28(11):1976–1980.

4 Estivill-CastroV, Lee IJ. Autoclust+: Automatic clustering of point-data sets in the presence of obstacles. Proc. of the International Workshop on Temporal Spatial and Spatial-Temporal Data Mining. Lyon, France. 2000.133–146.

5 Zaiane OR, Lee CH. Clustering spatial data when facing physical constraints. Proc. of the IEEE International Conference on Data Mining. Maebashi City, Japan. 2002. 737–740.

6 Ester M, Kriegel HP, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. of International Conference on Knowledge Discovery and Data Mining. 1996. 226–231.

7 王媛妮,边馥苓.基于演化算法的带故障约束空间聚类分析.计算机科学,2009,36(12):197–198.

8 Zhang XP, Wang JY, Wu F, et al. A novel spatial clustering with obstacles constraints based on genetic algorithms and K-medoids. Proc. of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA2006). Jinan, China. 2006, 1. 605–610.

9 Zhang XP, Wang JY, Fan ZS, Li B. Spatial clustering with obstacles constraints using ant colony and particle swarm optimization. Lecture Notes in Computer Science, 2007, 4819: 344–356.

10 Wang X, Rostoker C, Hamilton HJ. Density-based spatial clustering in the presence of obstacles and facilitators. ftp://cs.uregina.ca/Research/Techreports/2004-08.pdf. 2004.

11 Wang ZC, Xue LX, Li YS, Wang LL, Zhang XW. Voronoi diagram and spatial clustering in the presence of obstacles. Proc. of International Conference on Space Information Technology. Wuhan, China. 2005.

12 Wang X, Hamilton HJ. Clustering spatial data in the presence of obstacles. Proc. of International FLAIRS Conference. Miami Beach, FL. 2005. 177–198.

13 Park SH, Lee JH, Kim DH. Spatial clustering based on moving distance in the presence of obstacles. Lecture Notes in Computer Science, 2007, 4443: 1024–1027.

14 王莹.基于粒子群优化的带障碍约束 DBSCAN 算法研究[学位论文].哈尔滨:哈尔滨工程大学,2011.

15 李静.基于 Delaunay 三角网的有障碍物聚类算法研究[学位论文].太原:太原科技大学,2008.

16 Estivill-Castro V, Lee IJ. Autoclust: Automatic clustering via boundary extraction for mining massive point-data sets. Proc. of the 5th International Conference on Geocomputation. New South Wales, Australia. 2000. 23–25.

17 杨杨,孙志伟,赵政.一种处理障碍约束的基于密度的空间聚类算法.计算机应用,2007,27(7):1688–1691.

18 石岩,刘启亮,邓敏,王佳璆.一种顾及障碍约束的空间聚类方法.武汉大学学报(信息科学版),2012,37(1):96–100.

19 王立新,韩亚洪.涉及障碍物的聚类方法研究.计算机应用,2003,23(12):73–75.

20 郭薇,郭菁,胡志勇.空间数据库索引技术[学位论文].上海:上海交通大学出版社,2006.

21 陈克平,周丽华,王丽珍,等.Dcello-网格弥散聚类算法.计算机研究与发展,2004,41(增刊):205–212.

22 周丽华,王丽珍,陈克平.带障碍的空间分级聚类算法.计算机科学,2006,33(5):182–185.

23 孙宇清,赵锐,姚青,史斌,刘佳.一种基于网格的障碍约束下的空间聚类算法.山东大学学报(工学版),2006,36(3):86–90.

24 曾绍勤,李光强,廖志强.空间聚类算法的分类.测绘科学,2012,37(5):103–106.

25 张勇.COE-AUTOCLUST:对障碍空间上的实体进行自动聚类[学位论文].昆明:云南大学,2006.

26 曹科研,王国仁,韩东红,等.障碍空间中不确定数据聚类算法.计算机科学与探索,2012,6(12):1087–1097.