

基于概念格的空间聚类方法^①

殷俊华¹, 李光强², 陈 翼², 邓 敏²

¹(中南大学 湘雅二医院, 长沙 410011)

²(中南大学 地球科学与信息物理学院, 长沙 410083)

摘 要: 空间聚类一直是空间数据挖掘研究的热点之一。现有的聚类方法大都局限于根据空间位置来进行空间聚类的, 忽略了空间对象的专题属性, 从而导致空间聚类结果有时完全不符合人的空间认知, 缺乏合理的解释。为此, 综合考虑空间对象的位置和专题属性, 提出了一种基于概念格的空间聚类(Concept Lattices Based Spatial Cluster, CLBSC)方法。该方法通过构建多维专题属性的概念格, 简化了空间聚类计算。最后, 通过两组实验对 CLBSC 算法进行了验证分析, 研究结果表明: 所提出的 CLBSC 算法是一种具有高可靠性和抗噪性的空间聚类算法。

关键词: 空间聚类; 概念格; 空间位置属性; 专题属性; CLBSC

A Spatial Clustering Method Based on Concept Lattices

YIN Jun-Hua¹, LI Guang-Qiang², CHEN Yi², DENG Min²

¹(The second hospital, Central South University, Changsha 410011, China)

²(School of Geosciences and info-physics, Central South University, Changsha 410083, China)

Abstract: Spatial clustering is a hot issue in the field of spatial data mining. For a spatial object, the spatial location and the thematic attributes of spatial data are the inherent characteristics. However, the existing approaches mostly regard only the distance of spatial location as the similarity metric of spatial clustering, ignoring the thematic attributes of spatial objects. The results of these spatial clustering methods are not reasonable. Thus, a new spatial clustering method, named Concept Lattices Based Spatial Cluster (CLBSC for short) is proposed in this paper. The method considers both the spatial distance and attribute distance, and it simplifies the computation via building multi-dimensional attribute lattices. Furthermore, many concepts about CLBSC are expounded and its algorithm is narrated in detail. Finally, two experiments demonstrate that CLBSC algorithm is able to find more outlier and improve the reliability of spatial clustering using the Same Lattices Number.

Keywords: spatial clustering, concept lattices, spatial location; thematic attributes; CLBSC

空间数据挖掘是从大数据量的地理空间数据库中提取隐藏的、有用的信息和知识的过程, 是空间信息领域研究的重要方向之一^[1,2]。空间数据挖掘主要包括空间数据分类、空间关联规划挖掘、空间聚类、空间趋势探索和异常探测等内容。其中, 空间聚类通过将空间对象集合划分为若干个空间簇, 使得簇内空间目标间差异较小, 簇间差异较大, 从而获取空间对象的某种分布特征, 是空间数据挖掘领域里非常重要的挖

掘方法之一。传统的空间聚类方法大致可以分为: 基于划分的聚类方法^[3-6]、基于层次的聚类方法^[7-10]、基于密度的聚类方法^[11-13]、基于网格的聚类方法^[14-16]、基于模型的聚类方法^[17,18]等。在这些空间聚类方法中, 大都仅依据对象之间的空间距离进行聚类, 忽略了空间对象的专题属性特征。由于任何空间对象都具有空间属性和非空间属性(亦称专题属性)的双重特征, 仅利用空间对象间的距离来进行空间聚类, 很难发现

① 基金项目: 国家高技术研究发展计划(863)(2009AA12Z206); 湖南省自然科学基金(09JJ6061)

收稿时间: 2010-11-10; 收到修改稿时间: 2010-12-13

隐藏在空间数据中的专题属性分布特征,亦会降低空间聚类结果的可靠性,甚至可能得出与实际情况相悖的结论。因此,需要研究同时顾及空间特征和专题属性特征的空间聚类方法。

目前,同时考虑空间位置和专题属性的空间聚类研究成果相对较少。李新运^[19]等提出了坐标与属性一体化的空间聚类方法,该方法将空间实体的空间位置和专题属性特征都视为空间对象的属性维,进而使用属性距离函数计算相似度,然后结合 K-means 算法进行聚类。该方法虽同时考虑了空间位置和专题属性,但是将二者同等处理,弱化了空间实体的空间特征,聚类结果不理想。李光强^[20]等提出了基于双重距离的空间聚类方法,该方法通过研究空间距离和专题属性距离的双重可达或相连性,进而将两种距离均可达或相连的空间对象归为同一空间簇。该方法能够发现任意形状的空间簇,且效率较高,但是该算法没有充分考虑空间对象专题属性的自然分布特征,且不能处理非数值型属性。此外, Lin^[21]等提出的 DC (Dual Clustering) 算法以及 Zhou^[22]等提出的 DCAD (Dual Clustering Algorithm for Distributed Spatial Databases) 方法是从空间位置和专题属性两方面分别进行聚类,进而综合生成最终聚类结果。该方法计算复杂,需要输入的参数较多,同时也不易扩展和推广。

本文在充分考虑空间对象的专题属性分布的自然特性的基础上,论述了空间对象的专题属性概念格的定义,进而提出了一种基于概念格的空间聚类方法 (Concept Lattices based Spatial Clustering, 简称 CLBSC)。该方法同时考虑了空间对象的空间位置和专题属性,依据多维专题属性构建的概念格,将每个空间对象分配到相应的概念格中,从而可以简化计算专题属性偏离差异的过程,同时亦可以将非数值型专题属性引入概念格,可以提高空间聚类的可靠性和效率。

1 基于距离的属性概念格构建

1.1 概念格

概念格 (Concept Lattices, CL) 在数据挖掘、软件工程、信息检索和聚类分析等方面,有着重要的应用,最早由 Wille 于 1982 年提出。概念格提供了一种支持数据分析的有效工具,是从数据集 (又称形式背景) 中生成概念格的过程,实质上也是一种概念聚类

过程。概念格的每个节点是一个形式概念,由两部分组成:外延,即概念所覆盖的实例;内涵,即概念的描述,是该概念覆盖实例的共同特征^[23-25]。

假设给定形式背景为三元组 (O, I, R) , 其中 O 是实例集, I 是数据的属性集, R 是 $O \times I$ 的二元关系,则存在唯一的一个偏序集合与之对应,并且这个偏序集合产生一种格结构,称为概念格。概念格中的每个节点是一个序偶,称为概念,记为 (x, y) , 其中 $x \in O$, 称为概念的外延, $y \in I$ 称为概念的内涵。

在定义概念时,通常使用不同的尺度对 I 进行不同级别的划分,形成一个概念分层,即概念分层定义一个映射序列,将低层概念映射到更高层概念。例如,在空间数据库中,行政区划包括中国、湖南省、长沙市、雨花区等,则有雨花区 $\subset$ 长沙市 $\subset$ 湖南省 $\subset$ 中国,这些映射形成行政区划维的概念分层。

1.2 基于距离的专题属性概念格划分

对于数值型属性的划分方法,通常使用等深、等宽两种方法,前者是将样本平均分配到每个概念格中,使每个概念格中的样本数相等,后者是使用固定宽度划分属性,不保证每个概念格中的样本数相等^[2]。由于等深、等宽的划分方法都不能体现数据分布的自然特征,所以本文研究一种基于自然距离的概念格划分方法。

若给定样本集合 S , 其中专题属性项集 $I = \{i_1, i_2, \dots, i_k\}$, 若每个属性项均为数值型,则分别对 S 的不同属性项进行升序排序,得到不同的集合,分别记为 $\angle S_1, \angle S_2, \dots, \angle S_k$; 然后计算每个序列中的前后相连数据的差值,组成差序列,分别记为 $\text{Diff}_1, \text{Diff}_2, \dots, \text{Diff}_k$; 升序排序的差序列分别记为 $\angle \text{Diff}_1, \angle \text{Diff}_2, \dots, \angle \text{Diff}_k$ 。用 $\max_p^q$ 表示第 q 个属性项的升序差序列中第 p 个差值。

基于自然距离的概念格构造方法使用不同的 $\max_p^q$ 值构造不同层次的概念格,例如湖南省气象站点专题属性包括日均温度、天气状况两维属性 (表 1-a), 各气象站点的日均温度、天气状况两维属性 (表 1-a), 各气象站点的日均温度差序列如表 1-b 所示。若将各气象站点的日均温度划分为两个概念,则日均温度的概念包括 $\text{Temp}_1 (0, 23.25]$ 和 $\text{Temp}_2 [24.5, +\infty)$, 而非数值型属性天气状况可以泛化为三个概念,包括晴天、阴天、雨天,进而气象站点日均温度的两个概

念与天气状况的三个概念形成 6 个概念格（图 1），简 记为 cl^1 、 cl^2 、 cl^3 、 cl^4 、 cl^5 和 cl^6 。

表 1 湖南省气象站点属性值及差分序列

(a)			(b)		(c)	
站点号	日均 温度	天气 状况	日均 温度	日均温度差序列	天气 状况	泛化后 天气状况
1	22.75	中雨	20.25		多云	阴天
2	21.25	小雨-中雨	21.25	1	小雨-中雨	雨天
3	22	小雨-中雨	22	0.75	小雨-中雨	雨天
4	24.75	阴天	22.75	0.75	中雨	雨天
5	20.25	多云	22.75	0	中雨	雨天
6	23.25	多云	23.25	0.5	多云	阴天
7	24.5	阴天	24.5	1.25	阴天	阴天
8	22.75	中雨	24.75	0.25	阴天	阴天
9	25	晴天	25	0.25	晴天	晴天
10	25.5	晴天	25.5	0.5	晴天	晴天

Temp1, 晴天	Temp1, 阴天	Temp1, 雨天
Temp2, 晴天	Temp2, 阴天	Temp2, 雨天

图 1 概念格

对于样本集 S ，如果要对每个属性项划分为 p 个概念，则需要从每个属性有序差序列中取出前 $p-1$ 个差值，并以每个差值的被减数和减数为界，对样本集进行划分，然后将每个属性项划分后的概念集进行组合，得到相应的概念格。因此，基于自然距离构建的概念格也就是依据样本属性的最大差距划分的结果，体现了数据的自然分布特性。

当使用不同的划分宽度定义概念时，也就形成了多尺度（多种距离）的概念树。本文不研究概念树的生成和不同层次概念之间的关系，仅利用构建的单层次概念格进行空间聚类。

2 基于概念格的空间聚类原理

若有空间对象集 $F=\{f_1, f_2, \dots\}$ ，专题属性项集 $I=\{i_1, i_2, \dots\}$ ，空间数据库 $SD=F \times I$ ，那么空间对象 f_p 的专题属性记为 ap 。空间对象 f_p 的专题属性所属的概念格编号记为 $f_p.LatticeID$ 。则有以下定义：

定义 1. 同格对象是指在 f_p 的空间邻近域中，所有专题属性概念格编号与 f_p 相同的对象。

用 $SN(f_p)$ 表示 f_p 的空间邻近域，且 $f_p \in SN(f_p)$ 。若存在空间数据对象 $f_q \in SN(f_p)$ ，且 $f_p.LatticeID =$

$f_q.LatticeID$ ，则 f_q 是 f_p 的一个同格对象，记为 $f_p \otimes f_q$ ，否则记为 $\neg f_p \otimes f_q$ 。

空间对象 f_p 的所有同格对象组成 f_p 的同格对象集合，记为 Γ_p ，表达为：

$$\Gamma_p = \{ \forall f_q | f_p \otimes f_q \wedge f_q \in SN(f_p) \}$$
 (1)

定义 2. 同格数是指空间对象 f_p 的同格对象集的数目，记为 τp ，表达为：

$$\tau_p = \|\Gamma_p\|$$
 (2)

定义 3. 若有 $f_x \otimes f_y$ 和 $f_y \otimes f_z$ 成立，且 $\neg f_x \otimes f_z$ ，则称 f_x 和 f_z 格连接，记为 $f_x \sim f_z$ 。

定义 4. 若 $\tau p \geq \varepsilon$ ，则称 f_p 为核，记为 $f_p.isCore = true$ 。其中 ε 是同格数最小阈值。

定义 5. 空间簇是所有同格或格连接的核及其同格对象组成的集合，记为 sc 。

下面举例说明 CLBSC 聚类过程，空间示例数据如图 2 所示，图中共有八个空间对象，分别为 1~8，其中每个空间对象的下标代表该空间对象的专题概念格编号。

若取同格数阈值 $\varepsilon = 4$ ，空间对象 1~4 互为空间近邻，空间对象 5~8 也互为空间近邻，则基于概念格的聚类结果为：空间对象 1~4 构成一个空间簇，空间对象 5~8 虽然在空间上邻近，但是其中空间对象 7 的概念格编号与 5、6、8 不同，因此不能构成一个空间簇。

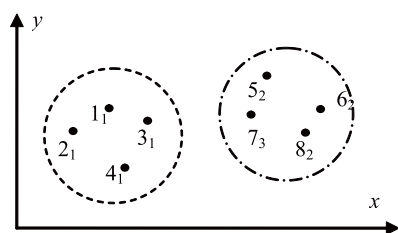


图2 空间数据示例

空间邻近域的定义有多种方法,通常使用固定距离和固定近邻数目的方法,前者将固定距离范围内的对象组成目标对象的空间邻近域,后者将所有与 fp 最近的 k 个空间对象组成 fp 的空间邻近域,亦称为 k -近邻法(或 k -NN),其中 k 称为 k 空间邻近对象数目阈值。

3 实验算例

为了验证 CLBSC 聚类算法的实用性和可靠性,本节共设计了两组实验。实验一使用 2008 年湖南省 96 个气象观测站点的日均观测数据进行空间聚类,并通过试凑法取得最优的聚类效果。实验二则将 CLBSC 聚类算法与基于双重距离的空间聚类算法^[20]进行比较,分析 CLBSC 算法的优点。

3.1 实验一

本组实验选用 2008 年湖南省 96 个气象观测站的日均观测数据进行实验,验证 CLBSC 聚类算法的实用性。湖南省气象监测数据表的数据项包括站号、站名、日期、日均温度、天气状况,部分数据如表 2 所示。

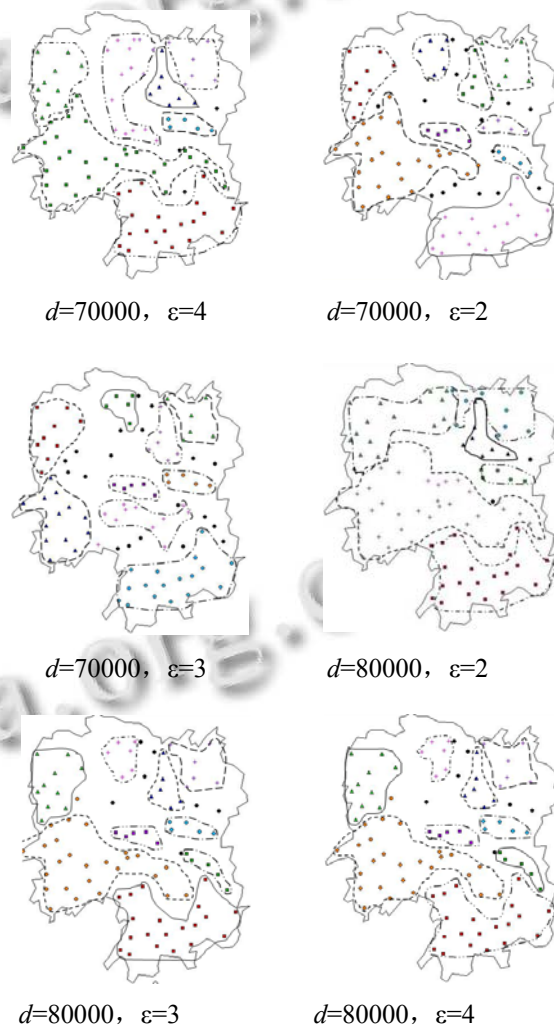
表2 气象监测数据

站号	站名	日期	日均温度	天气状况
1	龙山	2008-7-28	28.25	晴天
2	桑植	2008-7-28	29.5	晴天
3	永顺	2008-7-28	28.5	阴天
4	张家界	2008-7-28	30.5	晴天
5	保靖	2008-7-28	28.25	晴天
6	花垣	2008-7-28	28.25	多云
...	...	...	...	...

本实验步骤如下:①在气象数据库中,随机抽取一定比例的气象站点作为训练样本,将各气象站点的日均气温按升序排列,并计算升序排列后的差值,同时将各站点的天气状况属性进行概念泛化,然后根据

小节 2.2 构建相应的概念格;②依据气象站点的专题属性,为每个气象站点分配概念格编号;③根据试凑法确定相应的空间邻近域距离 d 和同格数阈值 ε ;④判断每个气象站点是否为核;⑤递归检索所有同格或格连接的核及其同格对象,并聚为空间簇。

在聚类过程中,为了选取合适的空间邻近域距离 d 和同格数阈值 ε ,本实验使用试凑法,经过多次试验来确定较为理想的参数,从而获得较可靠的聚类结果。聚类结果如图 3 所示。其中黑色圆点代表空间孤立点,其余点按其形状构成不同的空间簇。

图3 CLBSC 空间聚类结果图。 d 空间邻近域阈值, ε 为同格数阈值

同时参数取值(空间邻近域阈值和同格数阈值)与聚类结果统计(聚类数目与孤立点数目)的关系如下表 3 所示。

表 3 参数取值与聚类结果统计的关系表

参数取值			聚类结果统计	
编号	空间邻近域距离 阈值(m)	同格数 阈值	聚类 数目	孤立点 数目
1	70000	4	9	19
2	70000	3	9	15
3	70000	2	7	3
4	80000	4	9	14
5	80000	3	9	7
6	80000	2	7	2

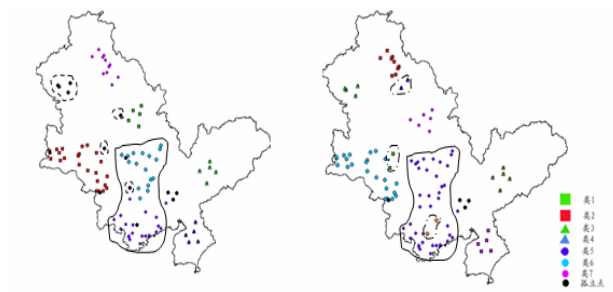
分析图 3 和表 3 可以发现, (1)CLBSC 算法聚类结果中同一簇中的气象站点不仅在空间上相对邻近, 而且其专题属性也较接近, 即其概念格编号相同; (2)孤立点主要是由于其空间邻近域内找不到足够多的气象站或者其气象站点的概念格编号与作为核的气象站点的概念格编号不同所致; (3)当空间邻近距离一定时, 同格数阈值越大, 孤立点数目亦越大, 当同格数阈值一定时, 空间邻近距离越大, 孤立点数目越小。总之, CLBSC 聚类结果同时考虑了空间实体的空间特征和专题属性特征, 使每个空间簇具有明确的地学含义, 同时该算法能够发现任意形状的簇, 能够有效区分孤立点, CLBSC 聚类结果受空间邻近距离阈值与同格数阈值的选取影响较大。

3.2 实验二

本组实验使用华南某市的土壤 Cu 浓度和 Zn 浓度的监测数据对 CLBSC 算法与 DDBSC 算法进行比较分析。图 4(a)是 CLBSC 算法的聚类结果, 图 4(b)是 DDBSC 算法的聚类结果。

分析图 4 可以发现, (1)图 4(b)的 DDBSC 算法未能将该市中部两个空间簇有效的区分开来, 而图 4(a)的 CLBSC 算法则进行了合理的区分, 如图 4 的实线框所示; (2)在典型空间簇区分清楚的前提下, 可以发现 CLBSC 算法能发现更多的孤立点信息, 如图 4(a)虚线框所示。因此, 可以说明 CLBSC 算法能获得更好的聚类效果, 同时能更好的发现孤立点信息; (3)在 DDBSC 算法的聚类结果中, 有可能出现空间对象个数很少的空间簇, 这必然会影响到聚类结果的可靠性和抗噪性, 因此 DDBSC 算法在运行结束时, 必须给定簇内要素数目下限, 当簇内要素数目小于给定的下限时, 则该簇不可靠^[20], 如图 4(b)虚线框所示。而 CLBSC 算法通过引入基于自然距离的概念格来划分专题属性并引入

同格数阈值来控制每个空间簇要素个数的下限, 从而有效的克服了这一问题, 提高了空间聚类的可靠性和抗噪性。



(a)CLBSC: $d=12000$, $\epsilon=4$ (b) DDBSC: $d=12000$, $\epsilon_{attr}=35$

图 4 CLBSC 和 DDBSC 算法空间聚类结果比较图. d 为空间邻近域阈值, ϵ 为同格数阈值

分析图 4 可以发现, (1)图 4(b)的 DDBSC 算法未能将该市中部两个空间簇有效的区分开来, 而图 4(a)的 CLBSC 算法则进行了合理的区分, 如图 4 的实线框所示; (2)在典型空间簇区分清楚的前提下, 可以发现 CLBSC 算法能发现更多的孤立点信息, 如图 4(a)虚线框所示。因此, 可以说明 CLBSC 算法能获得更好的聚类效果, 同时能更好的发现孤立点信息; (3)在 DDBSC 算法的聚类结果中, 有可能出现空间对象个数很少的空间簇, 这必然会影响到聚类结果的可靠性和抗噪性, 因此 DDBSC 算法在运行结束时, 必须给定簇内要素数目下限, 当簇内要素数目小于给定的下限时, 则该簇不可靠^[20], 如图 4(b)虚线框所示。而 CLBSC 算法通过引入基于自然距离的概念格来划分专题属性并引入同格数阈值来控制每个空间簇要素个数的下限, 从而有效的克服了这一问题, 提高了空间聚类的可靠性和抗噪性。

4 结论

本文提出基于概念格的空间聚类方法, 首先依据概念格原理构建空间对象的专题属性概念格, 再利用空间邻近距离阈值构建空间邻近域, 进而计算每个空间对象的同格数, 最后将同格或格相连的核及其同格对象组成一个空间簇。该方法考虑了空间对象的空间特征和专题属性特征, 使得空间邻近且专题属性邻近的对象聚成一类, 因此每个空间簇具有较为明确的地

学含义。通过实验分析,表明①CLBSC算法能获得较好的聚类效果,同时亦能更好的发现孤立点;②通过基于自然距离的概念格来划分专题属性,并引入同格数阈值可以有效的提高空间聚类的可靠性和抗噪性。但是,本文研究的方法受到空间邻近距离阈值和同格数阈值的影响,这两个参数的确定具有一定的主观性和不确定性,该参数的确定方法及聚类结果的评价是本文后续研究的主要内容。

参考文献

- 1 王劲峰.空间分析.北京:科学出版社,2006.
- 2 Han JW, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001. 223–259.
- 3 Macqueen J. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symposium in Mathematics, University of California, Berkeley, USA, 1967. 281–297.
- 4 Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990.
- 5 Ng R, Han J. CLARANS: A method for clustering objects for spatial data mining. IEEE Trans. Knowledge & Data Engineering, 2002,14(5):1003–1016.
- 6 Hopeener F, Klawonn F, Kruse R. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, 1999.
- 7 Guha S, Rastogi R, Skim K. CURE: An efficient clustering algorithm for large databases. Proc. 1998 ACM—SIGMOD Int. Conf. Management of Data (SIGMOD'98). 1998. 73–84.
- 8 Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. Proc. 1996 ACM — SIGMOD Int'l Conf. Management of Data (SIGMOD'96), 1996. 103–114.
- 9 Sudipto G, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes Proc. 1999 Int'l Conf. Data Engineering (ICDE'99), 1999. 512–521.
- 10 Karypis G, Han EH, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. Computer, 32, 1999. 68–75.
- 11 Ester M, Krigel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases. Proc. 1996 Int'l Conf. Knowledge Discovery and Data Mining (KDD'96), 1996. 226–231.
- 12 Ankerst M, Breunig M, Kriegel HP, et al. OPTICS: Ordering points to identify the clustering structure. Proc. 1999 ACM—SIGMOD Int'l Conf. on Management of Data (SIGMOD'99), 1999. 49–60.
- 13 Hinneburg A, Keim D A. An efficient approach to clustering in large multimedia databases with noise. Proc. 1998 Int'l Conf. Knowledge Discovery and Data Mining (KDD'98), 1998.58–65.
- 14 Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining. Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97), Athens, Greece, 1997.186–195.
- 15 Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A multi resolution clustering approach for very large spatial databases. Proc. 1998 Int'l Conf. Very Large Data Bases (VLDB'98), 1998. 428–439.
- 16 Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications. Proc. 1998 ACM—SIGMOD Int.'l Conf. Management of Data (SIGMOD'98), 1998. 94–105.
- 17 Fisher D. Improving inference through conceptual clustering. Proc. AAAI Conf.Seattle, WA, 1987. 461–465.
- 18 Cheeseman P, Stutz J. Bayesian classification (Autoclass): Theory and results. Advances in Knowledge Discovery and Data Mining, MA: AAAI/MIT Press, 1996.153–180.
- 19 李运新,郑新奇,闫弘文.坐标与属性一体化的空间聚类方法研究.地理与地理信息科学,2004,20(2):38–40.
- 20 李光强,邓敏,程涛,等.一种基于双重距离的空间聚类方法.测绘学报,2008,37(4):482–488.
- 21 Lin CR, Liu KH, Chen MS. Dual clustering: integrating data clustering over optimization and constraint domains. Proc. of the IEEE Transactions on Knowledge and Data Engineering, 2005.628–637.
- 22 Zhou JG, Guan JH, Li PX. DCAD: A dual clustering algorithm for distributed spatial databases. Geo-spatial Information Science, 2007,10(2):137–144.
- 23 Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Proceeding of the Ordered Sets' 82, Boston: Reidel, 1982.445–470.
- 24 胡可云,陆玉昌,石纯一.概念格的应用及进展.清华大学学报(自然科学版),2000,40(9):77–81.
- 25 Eklund P. Concept Lattices. Berlin: Springer, 2004.