基于不确定性的空间聚类*)

何彬彬[1,2] 方 涛[2] 郭达志[1]

(中国矿业大学环境与测绘学院 徐州221008)[1] (上海交通大学图像处理与模式识别研究所 上海200030)[2]

摘 要 空间数据挖掘是指从空间数据库(空间数据仓库)中提取用户感兴趣的空间的和非空间的模式和规则的过程,空间聚类是空间数据挖掘最基本的任务。本文首先分析了空间数据挖掘过程中的不确定性,并以空间聚类为例具体分析空间聚类过程中的数据预处理方法。在此基础上,以EM算法为基础,顾及空间聚类的随机性和模糊性以及基于Delaunay图的空间对象的空间邻近关系,构建了基于不确定性的空间聚类算法。

关键词 不确定性,空间数据挖掘,空间聚类,EM算法,Delaunay图

Uncertainty-Based Clustering Method for Spatial Data Mining

HE Bin-Bin[1,2] FANG Tao[2] GUO Da-Zhi[1]

(School of Environment & Spatial Informatics, China University of Mining and Technology, Xuzhou 221008)[1]

(Institute of Image Processing & Pattern Recognition, Shanghai Jiao-Tong University, Shanghai 200030)[2]

Abstract Spatial data mining, is to extract the hidden, implicit, valid, novel and interesting spatial or non-spatial patterns and rules from practical spatial databases (repositories). Spatial clustering is the most fundmental task of spatial data mining. In this paper, uncertainties in spatial data mining are analyzed firstly. Then, the data preprocessing method of spatial clustering is introduced. Finally, based on EM algorithm, a uncertainty-based algorithm of spatial clustering is constructed considering the randomness and fuzziness of spatial clustering and the spatial neighborhood relationship among spatial objects based on Delaunay diagram.

Keywords Uncertainty, Spatial data mining, Spatial clustering, EM algorithm, Delaunay diagram

1 引言

空间数据挖掘(Spatial Data Mining),或称从空间数据库(空间数据仓库)中发现知识(Knowledge Discovery from Spatial Databases/ Repositories),是指从空间数据库(空间数据仓库)中提取隐含的、用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其他一些隐含在空间数据库中的普遍的数据特征的过程[1]。该过程可分为数据准备与选择、数据预处理、数据挖掘、模式评价与知识表示四个阶段(图1)。空间数据挖掘按功能又可分为空间聚类、空间分类、空间关联规则挖掘等,其中空间聚类是其最基本的功能。空间聚类是指根据空间对象特征的聚散程度将它们划分为不同簇(类别)的空间概括和综合,使在同一簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大,它是一个非监督的过程。

聚类方法主要有六类:划分方法(如K-means、K-medoid)、层次方法(如CURE、BIRCH)、基于密度的方法(如DBSCAN)、基于模型的方法、基于网格的方法和基于图的方法,每一种方法都有其优缺点。通常空间数据聚类就是利用这些典型聚类算法进行,并未考虑空间数据的空间信息,其结果经常在地理位置上很混乱。鉴于此,空间数据的空间信息必须考虑,一些基于空间约束的方法已在不同文献中出现,归纳起来主要有以下三种方法:第一种方法,将空间数据的空间坐标作为空间数据的一个额外变量,并加以不同的权重[2];第二种方法,定义空间数据之间的一个邻近关系图,作为传统聚类算法的一个约束[3];第三种方法,建立一个将地理空间距离和非空间变量结合的相异度矩阵,然后在此基础上运行传统的聚类算法[4]。这三种方法都有其明显的缺点,前两种方法只适合地理位置很靠近的空间数据,如果空间对象之间的地理位置相距较远,很可能会将很相似的空间对象划分为不同的类别。第三种方法则没有统计证据。此外这些方法还有一个共同的缺点,就是将空间对象集合划分为互不相交的子集,其中每个空间对象最多只属于一个聚类。但是当空间对象与两个聚类之间的距离相等,或者空间对象与两个聚类区域同时相交时,强制聚类之间不能相交的约束就不符合实际。

另一个重要的事实是空间数据自身具有不确定性,空间数据挖掘过程中也会带来一系列的不确定性,而且这些不确定性在空间数据挖掘过程中会不断传播和积累。而传统的空间数据挖掘并未较这些特性考虑进去,并且一般认为挖掘出来的知识都是有用的和确定的。目前国内外对于空间聚类的研究主要集中在聚类方法的可收缩性,方法对聚类复杂形状和类型的数据的有效性,高维聚类分析技术,以及针对大型数据库中混合数值和分类数据的聚类方法[5]。而对于空间聚类过程中的不确定性研究很少。本文首先分析空间数据挖掘过程中的不确定性。接下来,具体讨论怎样处理空间聚类过程中数据预处理阶段的不确定性,最后基于EM算法、顾及空间聚类的随机性和模糊性以及基于Delaunay图的空间对象的空间邻近关系,构建基于不确定性的稳健空间聚类算法。

*)本文得到国家自然基金项目(编号:60275021)资助。何彬彬 博士研究生,研究方向:空间数据挖掘与知识发现,遥感与GIS应用。方 涛 副教授,研究方向:智能信息获取与处理。郭达志 教授,博士生导师,研究方向:遥感和GIS理论及其应用。

2 空间数据挖掘过程中不确定性分析

不确定性在现实世界是客观存在的，确定性是相对的，不确定性是绝对的。空间数据和空间数据挖掘过程中同样存在不确定性，而且这些不确定性在空间数据挖掘过程中会不断传播与积累(图1)。空间数据的不确定性在GIS(Geographic Information System)理论中已进行了大量研究，限于篇幅，这里不进行阐述。以下简要分析空间数据挖掘过程中不确定性。

存在不完整、含噪声的和不一致的空间数据是大型的、现实世界空间数据库或空间数据仓库的共同特点。从形式上看，空间数据库中数据的不确定性包括位置不确定性、属性不确定性、时域不确定性、逻辑不一致性和数据不完整性。这些不确定性大都是在数据采集、解译和数据入库等过程中由于仪器设备精度限制和人为因素造成的，很多是不可避免的，即使通过后期数据处理也只能减少空间数据的部分不确定性。

空间数据选择阶段的不确定性主要是指用户根据挖掘任务的要求，主观选择目标数据过程中带来的不确定性，包括那些数据应该被选择、多少的数据量才足够等。

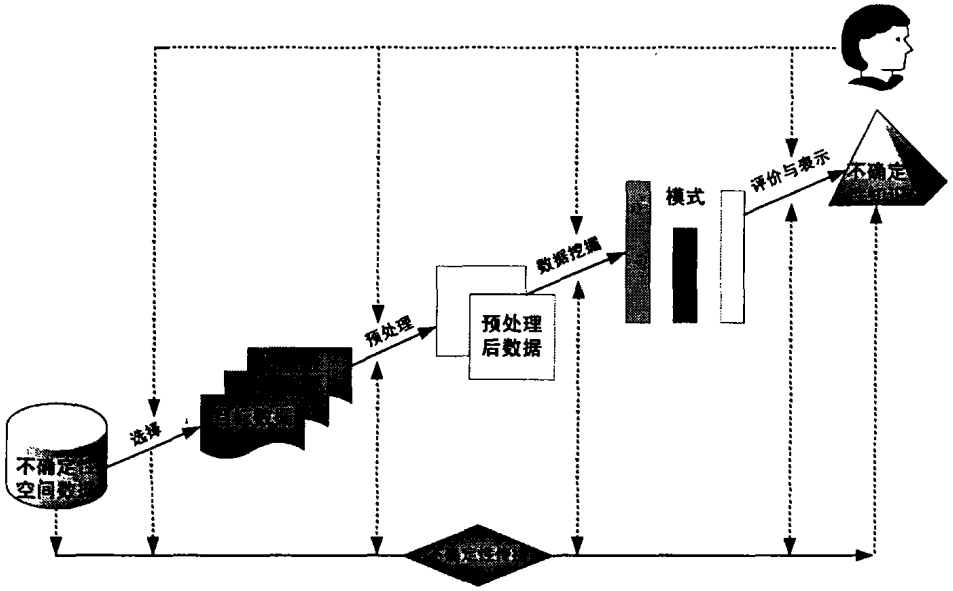

图1 空间数据挖掘过程中的不确定性及其传播

空间数据预处理主要包括数据清理、数据变换和数据规约。数据清理主要是试图填充空缺的值、识别孤立点、消除噪声和纠正数据中的不确定性；数据变换是将数据转换成适合于挖掘的形式，主要包括平滑(去掉数据中的噪声)、聚集(对数据进行进一步的聚集)、数据概化(用高层次的概念替换低层次"原始"数据)、规范化(将属性数据按比例缩放)和属性构造(构造新的属性并添加到属性集中，以帮助挖掘过程)。这一阶段，一方面处理空间数据中较明显的不确定性；另一方面，处理方法本身也会带来不确定性。

空间数据挖掘本身带来的不确定性主要是指挖掘算法的局限性而造成挖掘结果与真实情况的不完全一致，这也是造成数据挖掘不确定性的最重要原因之一。

知识表示中的不确定性主要是指知识中隐含不确定性，包括随机性、模糊性和不完全性。由于目前对人类知识的结构及机制还没有完全搞清楚，因此关于知识表示的理论及规范尚未建立起来。同一知识可以用多种方法表示。有些知识用这种方法比较好，而有些知识可能采用另一种表示方法比较合适。空间数据挖掘所获得的知识，大都是经过归纳和抽象的定性知识，或是定性和定量相结合的知识。对这些知识的最好表示方法是自然语言，至少在知识表示方法中含有语言值，即用语言值表达其中的定性概念。

3 空间聚类的数据预性处理

空间聚类过程中的不确定性来源与前述空间数据挖掘过程中的不确定性来源类似。结合空间聚类的具体特点采用如下不确定性处理方法，重点针对空间数据预处理和空间聚类算法改进。

假设要聚类的空间数据集合包含 n 个数据对象，空间聚类算法采用对象-对象数据结构(相异度矩阵)存储 n 个对象两两之间的近似性，表现形式是 $n\times n$ 维的矩阵。

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

$d(i,j)$是对象 i 和对象 j 之间相异性的量化表示，$d(i,j)=d(j,i)$，而且 $d(i,i)=0$。空间数据的变量类型主要包括区间标度变量、二元变量、标称型变量、序数型变量和比例标度型变量。限于篇幅，下面重点介绍两类最常见变量(区间标度变量和混合类型的变量)的数据预处理方法，其它类型变量的预处理方法可参见文[5]。

3.1 区间标度变量

空间聚类过程中，区间标度变量是一个粗略线性标度的连续度量。选用的度量单位将直接影响聚类分析的结果。为了避免对各度量单位选择的依赖，对数据进行标准化。给定一个变量 f 的度量值，变换和处理方法如下：(1)计算平均的绝对偏差 S_f：$S_f=\frac{1}{n}(|x_{1f}-m_f|+|x_{2f}-m_f|+\cdots+|x_{nf}-m_f|)$，其中 $x_{1f},\cdots,x_{nf}$是 f 的 n 个度量值，m_f 是 f 的平均值，即：$m_f=\frac{1}{n}(x_{1f}+x_{2f}+\cdots+x_{nf})$，这个平均的绝对偏差 S_f 比标准偏差 σ_f 对于孤立点具有更好的鲁棒性；(2)计算标准化的度量值 Z_{if}：$Z_{if}=\frac{x_{if}-m_f}{S_f}$；(3)计算对象间的相异度：对象间的相异度是基于对象间距离来计算，常用的距离度量方法有：欧几里得

距离、曼哈坦距离和明考斯基距离，其定义如下：

欧几里得距离：$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$；

曼哈坦距离：$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}|$；

明考斯基距离：$d(i,j)=(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\cdots+|x_{ip}-x_{jp}|^q)^{1/q}$；以上公式的 $i=(x_{i1},x_{i2},\cdots,x_{ip})$ 和 $j=(x_{j1},x_{j2},\cdots,x_{jp})$ 是标准化后的两个 p 维的数据对象。如果对每个变量根据其重要性赋予一个权重，加权的欧几里得距离可以计算如下：

$$d(i,j)=\sqrt{w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+\cdots+w_p|x_{ip}-x_{jp}|^2}$$

类似地，加权也可以用于曼哈坦距离和明考斯基距离。

3.2 混合类型的变量

许多真实的空间数据库中，对象是被混合类型的变量描述的。对于这种类型的变量，是将不同类型的变量组合在单个相异度矩阵中，把所有有意义的变量转换到共同的值域区间 $[0.0,1.0]$ 上。

假设空间数据集包含 p 个不同类型的变量，则对象 i 和 j 之间的相异度 $d(i,j)$ 定义为：$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$，其中，如果 x_{if} 或 x_{jf} 缺失，或者 $x_{if}=x_{jf}=0$，且变量 f 是不对称的二元变量，则指示项 $\delta_{ij}^{(f)}=0$；否则，$\delta_{ij}^{(f)}=1$。变量 f 对 i 和 j 之间相异度的计算方式与其具体类型有关：如果 f 是二元变量或标称变量，则当 $x_{if}=x_{jf}$，$d_{ij}^{(f)}=0$，否则 $d_{ij}^{(f)}=1$；如果 f 是区间标度变量，$d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{\max_h x_{hf}-\min_h x_{hf}}$，其中 h 遍取变量 f 的所有非空缺对象；如果 f 是序数型或比例型标度变量，计算 r_{if} 和 $Z_{if}=\frac{r_{if}-1}{M_f-1}$，并将 Z_{if} 作为区间标度变量值对待。

4 基于不确定性的空间聚类算法

根据以上分析，以 EM 算法为基础，顾及空间聚类的随机性和模糊性以及基于 Delaunay 图的空间对象的空间邻近关系，构建基于不确定性的稳健空间聚类算法。以下为该算法的主要方法和步骤：

4.1 EM 聚类算法

期望最大化（Expectation Maximization，EM）算法是解决数据残缺问题的一种出色的聚类算法。具体来讲，令 $D=\{x(1),\cdots,x(n)\}$ 为 n 个观察到的数据向量。设 $H=\{z(1),\cdots,Z(n)\}$ 表示隐藏变量 Z 的 n 个值，与观察到的数据点 D 一一对应；也就是说 $Z(i)$ 与数据点 $x(i)$ 相联系。假定 D 是离散的，这样便可以把未知的 $z(i)$ 值想像为数据的不可见分类（聚类）标签。

我们可以把观察到的数据的对数似然写作：$l(\theta)=\log p(D|\theta)=\log\sum_H p(D,H|\theta)$，其中右侧的项表明观察到的似然可以观察到数据和隐藏数据的似然对隐藏值的求和，这里假定了一个以未知参数 θ 为参量的概率模型 $p(D,H|\theta)$。

设 $Q(H)$ 为残缺数据 H 的任意概率分布。可以用下式表示似然：

$$l(\theta)=\log\sum_H p(D,H|\theta)=\log\sum_H Q(H)\frac{p(D,H|\theta)}{Q(H)}\geqslant\sum_H Q(H)\log\frac{p(D,H|\theta)}{Q(H)}=\sum_H Q(H)\log p(D,H|\theta)+\sum_H Q(H)\log\frac{1}{Q(H)}=F(Q,\theta)$$

函数 $F(Q,\theta)$ 是要最大化的函数（似然 $l(\theta)$）的下限。EM 算法在以下二者间交替：固定参数 θ，使 F 相对于分布 Q 最大化；固定 $Q=p(H)$，使 F 相对于参数 θ 最大化。具体地说：E 步骤：$Q^{k+1}=\arg\max_Q F(Q^k,\theta^k)$；M 步骤：$\theta^{k+1}=\arg\max_\theta F(Q^{k+1},\theta^k)$。

E 步骤中当 $Q^{k+1}=p(H|D,\theta^k)$ 时似然达到最大值，对于很多模型可以有相当直接的方法明确地计算出 $p(H|D,\theta^k)$。此外，对于这个 Q 值不等式变成了一个等式 $l(\theta^k)=F(Q,\theta^k)$。

在 M 步骤中，最大化问题简化为最大化 F 中的第一项（因为第二项不依赖于 θ），因此可以看到：$\theta^{k+1}=\arg\max_\theta\sum_H p(H|D,\theta^k)\log_p(D,H|\theta^k)$，这个表达式也经常可以得到闭合形式的解。显然根据定义 E 和 M 步骤在每一步中不会降低 $l(\theta)$：在 M 步骤的开始根据定义我们有 $l(\theta)=F(Q^{k+1},\theta^k)$，而且以后的 M 步骤调整 θ 来使 F 最大化。

要确定一个精确的算法我们需要取一个初始起点（如从一个初始的随机选取的 Q 或 θ 的值开始）和一种探测收敛的方法（如 Q,θ，或 $l(\theta)$）中的任一个在一次迭代后和上一次迭代后没有明显变化。

4.2 EM 算法的混合估计

对于正态混合模型：$f(x_i|\theta)=\sum_{k=1}^{K}\pi_k f_k(x_i|\mu_k,\sum_k)$，其中，$\mu_k$ 是第 k 个分量的均值，$\sum_k$ 是第 k 个分量的标准差，π_k 是数据点属于分量 k 的验前概率（$\sum_k\pi_k=1$）。因此，参数向量为 $\theta=\{p_1,\cdots,p_k,\mu_1,\cdots,\mu_k,\sum_1,\cdots,\sum_k\}$。假定如果知道 θ 的值，那么一个测量向量为 x 的对象来自第 k 个分量的概率为：$\hat{p}(k|x)=\frac{\pi_k f_k(x_i|\mu_k,\sum_k)}{f(x)}$，这就是基本的 E 步骤。据此，可以根据以下各式估计 π_k,μ_k 和 $\sum_k$：$\hat{\pi}_k=\frac{1}{n}\sum_{i=1}^{n}p(k|\hat{x}(i))$；$\hat{\mu}_k=\frac{1}{n\pi_k}\sum_{i=1}^{n}p(k|\hat{x}(i))x(i)$；$\hat{\sum}_k=\frac{1}{n\pi k}\sum_{i=1}^{n}\hat{p}(k|x(i))(x(i)-\hat{\mu}_k)^2$，其中的汇总求和是针对数据集中的 n 个数据点的。这三个等式就是 M 步骤。

4.3 EM 算法的模糊聚类

模糊聚类把对象集合划分为若干模糊子集，允许对象对不同的类别有不同的隶属度，模糊聚类反映出聚类过程中的不确定性，有利于根据用户进行决策。如果聚类是多级的，模糊化的聚类结果蕴涵了更丰富的信息，有利于下一层聚类。

正态混合模型的 EM 聚类算法可以用下面最优化函数—Hathaway 标准表示[6]：

$$D(c,\theta)\triangleq\sum_{k=1}^{K}\sum_{i=1}^{n}c_{ik}\log(\pi_k f_k(x_i|\mu_k,\sum_k))-\sum_{k=1}^{K}\sum_{i=1}^{n}c_{ik}\log(c_{ik})\qquad(1)$$

其中 $c=(c_{ik})$，$i=1,n$ $k=1,K$ 定义了一个模糊分类，c_{ik} 表示 x_i 属于类别 k 的隶属度

$(0\leqslant c_{i,k}\leqslant 1,\sum_{k=1}^{k}c_{ik}=1,\sum_{i=1}^{n}c_{ik}>0,1\leqslant i\leqslant n,1\leqslant k\leqslant K)$。

根据上式，从一个给定的初始参数 θ^0 开始，函数 $D(c,\theta)$ 的最优化按如下两个步骤进行：

（下转封四）

(上接第 198 页)

E 步骤：为了最大化标准，更新分类矩阵：$c^{m+1} = \arg\max_c D(c,\theta^m)$，考虑约束 $\sum_{k=1}^{K} c_{ik} = 1, \forall i, D(c,\theta^m)$ 的拉格朗日式为：$D(c) = D(c,\theta^m) + \sum_{i}^{n} \lambda_i ((\sum_{k=1}^{K} c_{ik}) - 1)$，其中 λ_i 是相应约束的拉格朗日系数。

则产生最优化的必要条件是：

$$\begin{cases} \dfrac{\partial D}{\partial c_{ik}} = \log(\pi_k^m f_k(x_i|\mu_k^m, \sum_k^m)) - 1 - \log c_{ik} + \lambda_i = 0 \\ \sum_{k=1}^{K} c_{ik} = 1 \end{cases}$$

则 $c_{ik}^{m+1} = \dfrac{\pi_k^m f_k(x_i|\mu_k^m, \sum_k^m)}{f(x_i|\theta^m)}$；M 步骤：根据 $\theta^{m+1} = \arg\max_\theta D(c^{m+1},\theta)$，参数被重新估计，结合前面公式：$D(c^{m+1},\theta) = Q(\theta|\theta^m) - \sum_{k=1}^{K}\sum_{i=1}^{n} c_{ik}^{m+1} \log(c_{ik}^{m+1})$

4.4 基于 Delaunay 图的空间邻近关系

根据空间数据的空间分布，建立离散空间数据的 Delaunary 图(图 2)，并计算空间对象间的空间邻近关系矩阵 V：

$$v_{ij} = \begin{cases} \alpha > 0 & \text{如果 } x_i \text{ 和 } y_i \text{ 邻近} \\ 0 & \text{如果 } x_i \text{ 和 } y_i \text{ 不邻近} \end{cases}$$

4.5 Hathaway 标准的改进

在上述已建算法基础上，考虑空间数据的空间信息，改进函数(1)。首先，建立正则化条件：

$$G(c) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ik} \cdot c_{jk} \cdot v_{ij}$$

然后，建立新的标准函数为：

$$U(c,\theta) = D(c,\theta) + \beta \cdot G(c) (\beta \geqslant 0) \quad (2)$$

则最优化的条件如下：

$$\begin{cases} \dfrac{\partial U}{\partial c_{ik}} = \log(\pi_k^m f_k(x_i|\mu_k^m, \sum_k^m)) + 1 - \\ \quad \log c_{ik} + \lambda_i + \beta \sum_{j=1}^{n} c_{jk} v_{ij} \\ \sum_{k=1}^{k} c_{ik} = 1 \end{cases}$$

最后得到如下方程：

$$c_{ik}^{m+1} = \frac{\pi_k^m f_k(x_i|\mu_k^m, \sum_k^m) \cdot \exp\{\beta \sum_{j=1}^{n} c_{jk}^{m+1} v_{ij}\}}{\sum_{l=1}^{k} \pi_l^m f_l(x_i|\mu_l^m, \sum_l^m) \cdot \exp\{\beta \sum_{j=1}^{n} c_{jl}^{m+1} v_{ij}\}}$$

M 步骤：$\theta^{m+1} = \arg\max_\theta U(c^{m+1},\theta) = \arg\max_\theta D(c^{m+1},\theta)$。

参 考 文 献

1 Miller, Harvey J, Han Jiawei. Geograp- hic data mining and knowledge discove- ry[M]. London: Taylor & Francis, 2001

2 Jain A K, Farrokhnia F. Unsupervis- ed texture segmentation using gabor fol- ters[J]. Pattern Recognition, 1991, 24 (12): 1167～1186

3 Legendre P. Constrained clustering[A]. In: Developments in Numerical Ecology NATO ASI Series G: Ecological Sciences[C]. Springer-Verlag. Heidelberg, 1991, 14: 289～307

4 Oliver M A, Webster R . A Geostati- stical basis for spatial weighting in multi- variate classification[J]. Mathematical Geology, 189, 21: 15～35

5 Han Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. San Francisco California. Morgan Kaufmann Publishers, 2001

6 Hathaway R J. Another interpretation of the EM algorithm for mixture distributi- ons[J]. Journal of statistics & Probability Letters, 1986, 4: 53～56

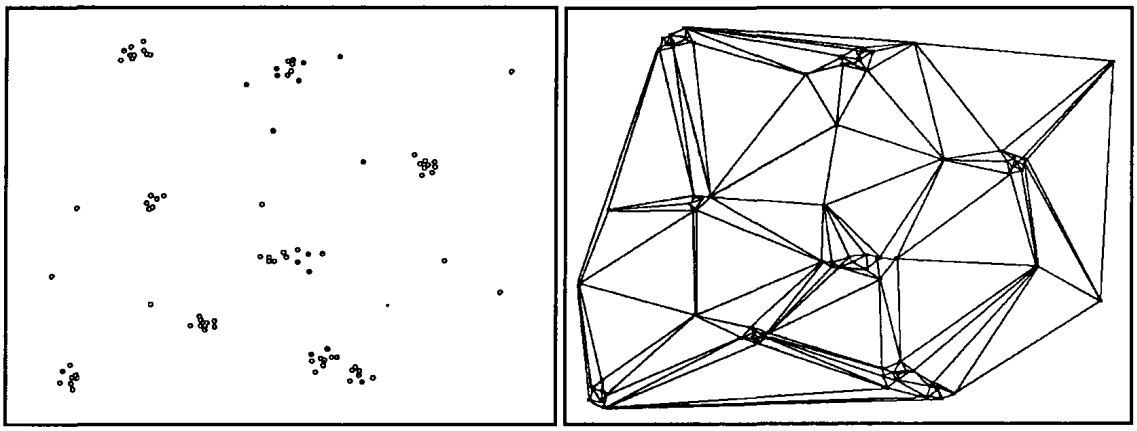

图 2 空间数据集和它的 Delaunay 图

计算机科学

(1974 年 1 月创刊)
第 31 卷第 11 期 (月刊)
2004 年 11 月 25 日出版

ISSN 1002—137X
CN50—1075/TP

定价：20.00 元 国外定价：5 美元
邮发代号：78—68
发行范围：国内外公开

主管单位：国 家 科 学 技 术 部
主办单位：国家科技部西南信息中心
编辑出版：《计 算 机 科 学》杂 志 社
重庆市渝中区胜利路 132 号 邮政编码：400013
电话：(023) 63500828 E-mail：jsjkx@swic.ac.cn
社 长：牟炳林
主 编：彭 丹
副 主 编：朱宗元
印 刷 者：重 庆 科 情 印 务 有 限 公 司
总发行处：重 庆 市 邮 政 局
订 购 处：全 国 各 地 邮 政 局
国外总发行：中国国际图书贸易总公司 (北京399 信箱)
国外代号：6210—MO