数据集的分形维数及其算法实现

吴新玲[1], 周国强[2]
(1 广东技术师范学院 计算机科学学院, 广东 广州 510262;
2 广东技术师范学院 教育技术与传播学院, 广东 广州, 510262)

摘 要: 基于数据库和数据挖掘等应用领域的需求, 介绍了计盒维数和广义分形维数的定义, 重点介绍了基于多层网格结构的分形维数的线性计算算法, 并对算法的数据结构提出了改进方案. 新的数据结构保持了算法的线性, 降低了算法的空间复杂性, 有效提高了算法的效率.

关键词: 分形几何; 分形维数; 盒维数; 算法; 数据结构
中图分类号: TP391 文献标识码: A 文章编号: 1000-7180(2011)06-0012-04

Fractal Dimension of Data Set and Its Algorithm Realization

WU Xin-ling[1], ZHOU Guo-qiang[2]
(1 Institute of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510262, China;
2 Institute of Educational Technology and Communication,
Guangdong Polytechnic Normal University, Guangzhou 510262, China)

Abstract: In this paper, the definition of the box-counting dimension and generalized fractal dimension are introduced based on the requirements of database and data mining application. The linear calculation algorithm of fractal dimension based on the multi-level grid structure is introduced mainly. An improved method about the data structure of the linear algorithm is proposed. The new data structure maintain the linear nature of the algorithm and reduce the space complexity of the algorithm, effectively improve the algorithm efficiency.

Key words: fractal geometry; fractal dimension; box dimension; algorithm; data structure

1 引言

自然界存在大量不够光滑和不够规则的事物, 如高山、树木、海岸线等等. 传统的经典数学主要研究规则、光滑的对象, 而将那些不够光滑和不够规则的集和函数, 当着是“病态”的, 认为不值得进行研究[1]. 1967 年, 美籍数学家曼德布罗特(B. B. Mandelbrot) 在《科学》杂志上发表的著名论文: 《英国的海岸线有多长》, 告诉人们对不光滑、不规则对象是值得也是能够进行详细的数学描述的. 并且, 在 1975 年正式创立了分形几何学, 为研究这类不规则集提供了总的理论框架.

分形集合最基本的特性就是自相似性, 即任意小比例的细节都具有某种程度的自相似. 这种自相似性可以是几何意义上的严格自相似, 也可以是统计意义上的近似. 因此, 对形态、功能、信息、时间、空间等任一方面具有自相似特性的对象, 我们都能应用分形几何学的基本思想进行分析、描述. 作为一种新的概念和方法分形正在许多领域开展应用研究[2-3].

经典几何中, 维数是描述规整几何图形几何特征的基本概念. 如, 点是 0 维, 光滑的曲线是 1 维, 曲面是 2 维, 空间图形是 3 维等等, 这里维数都是整数. 整型维数已不够度量分形集合, 如典型的三分康托尔分形集, 其集合中有无穷个数据点, 但其长度为 0, 按传统维数的观点, 其维数既不能是 0, 也不能是 1, 那么只能是介于 0 和 1 之间的一个分数. 因此, 分形几何定义了多种形式的维数, 以描述分形集合不

收稿日期: 2010-08-24; 修回日期: 2010-11-01

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同的比例特征和自相似性. 分形维数是分形几何的中心概念.

近年来, 分形的概念也被广泛应用于数据库和数据挖掘等计算机应用领域. 尤其是将分维数作为衡量数据集的本质维, 用于对数据集进行属性选择[4], 空间数据查询与连接[5], 聚类与分类[6] 等分析. 这些应用分析中采用的分形维数不尽相同, 但无论采用何种形式的分形维数, 面向大数据集的分维数计算都是一项耗费空间和时间的任务. 因此如何提高分形维数计算的效率成为这类应用任务的一个关键问题. 本文面向数据库和数据挖掘等应用领域, 介绍了数据集的常用分形维数及其计算算法, 并对算法的数据结构进行了讨论, 给出了降低算法空间复杂性和时间复杂性的实现方法以及算法的应用实例.

2　数据集的分形维数定义

分形几何已产生了多种形式的分形维数定义, 这些定义不尽等价, 当然也只能用于适合它们的应用领域内. 然而大部分维数的定义都基于"用尺度 δ 进行度量"的思想[1], 即忽略尺寸小于 δ 时的不规则性, 观察 $\delta \to 0$ 时的测量值. 例如, 假设 F 是 R^n 上的任意非空的有界子集, $M_\delta(F)$ 是 δ 下 F 的测量值, 且 $M_\delta(F)$ 服从幂定理, 即当 $\delta \to 0$ 时, 有常数 c, s, 使得

$$M_\delta(F) \sim c\delta^{-s} \tag{1}$$

则可说 F 具有"分配维数" s. 由上式两边取对数可得到

$$s = \lim_{\delta \to 0} \frac{\ln M_\delta(F)}{-\ln \delta} \tag{2}$$

而式(2) 中的 s 可以通过 δ 在一个有限的范围内, $\ln M_\delta(F)$ 和 $\ln\delta$ 的双对数图的斜率来估算, 因此便于计算和实践, 这也正是基于此思想定义的"计盒维数"被广泛应用的原因之一.

计盒维数也称盒维数是目前应用最普遍的分维数之一, 它是将式(2) 中的测量值定义为直径最大为 δ 可以覆盖 F 的集的最少个数. 因此, 定义 F 的下计盒维数和上计盒维数分别为

$$\underline{D_b}(F) = \varliminf_{\delta \to 0} \frac{\ln M_\delta(F)}{-\ln \delta} \tag{3}$$

$$\overline{D_b}(F) = \varlimsup_{\delta \to 0} \frac{\ln M_\delta(F)}{-\ln \delta} \tag{4}$$

如果 $\underline{D_b}(F)$ 与 $\overline{D_b}(F)$ 相等, 则为 F 的计盒维数或盒维数, 记为

$$D_b(F) = \lim_{\delta \to 0} \frac{\ln M_\delta(F)}{-\ln \delta} \tag{5}$$

如果 F 是一个有限数据点集合, 且 δ 在 (δ_1, δ_2) 范围内, 数据集具有统计自相似特性, 那么 $D_b(F)$ 可以由下式获取.

$$D_b(F) = -\frac{\partial \ln(M_\delta(F))}{\partial \ln \delta} \tag{6}$$

式(6) 在 $\delta_1 < \delta < \delta_2$ 时是一个常量, 该常量即 F 的盒维数, 且等于 $\ln M_\delta(F)$ 与 $\ln\delta$ 双对数图中对应 $\delta_1 < \delta < \delta_2$ 直线段的斜率.

根据文献[1] 中的等价定义, 式(5) 中的 $M_\delta(F)$ 又可以是以下五个数中的任意一个:

(1) 覆盖 F 的半径为 δ 的最少闭球数;

(2) 覆盖 F 的边长为 δ 的最少的立方体数;

(3) 与 F 相交的 δ 网立方体的个数;

(4) 覆盖 F 的直径最大为 δ 的集的最少个数;

(5) 球心在 F 上, 半径为 δ 的相互不交的球的最多个数.

关于分形维数的其他定义可详见文献[1, 7], 下面主要介绍基于盒维数, 针对有限数据点集合定义的广义分形维数[5], 在数据库和数据挖掘等应用中常据此计算相关数据集的分维.

广义分形维数(Generalized fractal dimension): 对于 R^n 上的有限数据点集合 F, 如果 F 在 $\delta_1 < \delta < \delta_2$ 范围内具有自相似性, 则 F 的广义分形维数 $D_q(F)$ 定义为

$$D_q(F) = \frac{1}{q-1} \frac{\partial \ln \sum_i p_i^q}{\partial \ln \delta} \quad \delta \in (\delta_1, \delta_2) \wedge q \neq 1 \tag{7}$$

式中, p_i 是边长为 δ 的与 F 相交的第 i 个网立方体中包含 F 中数据点的频数, 在 $\delta_1 < \delta < \delta_2$ 内式(7) 是一个常量.

显然, 式(7) 定义了一簇分形维数, 且 $D_0(F)$ 就是 F 的盒维数, 而当 $q = 2$ 时, 称之为"相关分维(correlation fractal dimension)", 在数据分析中, 常采用此分维数定义分析数据点属性之间的关联性[4-6].

3　分形维数的计算算法

实践中计算式(7) $D_2(F)$ 最常用的方法就是采用"计盒"方式[4-6, 8]. 即将数据点集合 F 所在的 E 维空间划分为边长为 δ 的 E 维网格单元, 计算网格中每个单元包含的数据点数的平方之和 $S(\delta)$, 即

$$S(\delta) = \sum_i C_{\delta,i}^2 \tag{8}$$

式中, $C_{\delta,i}$ 是第 i 个网格单元包含的数据点数. 由于

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

假设数据集是自相似集, 那么必然存在区间(δ_1, δ_2), 当δ在此区间内取值, 形成的$\ln(S(\delta))$与$\ln(\delta)$的双对数图将对应直线, 其常量斜率即为F的相关分维$D_2(F)$.

采用"计盒"方法计算的算法复杂性就在于需要计算不同δ值的$S(\delta)$. 目前, 最有效的计算算法是文献[4]中提出的 LiBOC(Line Box Occupancy Counter) 算法. 该算法通过使用多层网格结构避免了对数据集的重复扫描, 使算法在时间复杂性上与数据集规模N呈$O(N)$的线性关系.

所谓多层网格结构是指将δ按 1/2 的比例逐层缩减, 形成多层网格结构, 即下层网格单元的边长是上层网格单元的一半. LiBOC 算法采用树形存储结构存储、计算多层网格单元的数据点数, 如图 1 所示. 图 1 对应的是一个二维空间 5 个数据点数据集的 3 层网格树形存储结构. 不妨假设数据集已经规范化, 即包含数据集的网格边长为 1. 由图 1 不难知道: 任意E维、N个数据点的数据集合F的多层网格树形存储结构的第r层对应边长$\delta = 1/2^r$的网格层, 且上一层的网格单元在下一层将分解为2^E个网格单元, 当然每个网格单元只有当其包含的数据点数不为 0 时才有必要往下分解; 树中每个结点是一个有2^E个元素的数组; 树的深度K对应$\ln(S(\delta))$与$\ln(\delta)$双对数图的数据点数.

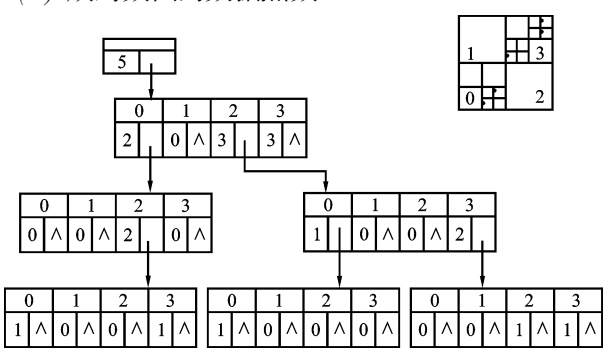


图 1 多层网格树形存储结构示意图

LiBOC 算法如下:

输入: N 行E 维的规范化数据集F

输出: 分形维数 D

(1) 构造树 T 的第一层结点, 并初始化结点中每个元素的计数 $C_i = 0$;

(2) 对 F 中每个数据点d, 作

r 从 1 到 K

计算第 r 层 d 所在网格单元序号 i,

$C_i = C_i + 1$, 并进入下层结点, 若下层结点不存在则创建;

(3) 扫描树 T, 对 $\delta = 1/2, 1/2^2, \cdots, 1/2^K$, 分别计算 $S(\delta)$;

(4) 生成与 $\ln\delta$、$\ln S(\delta)$ 对数图;

(5) $D =$ 图中直线部分斜率.

显然 LiBOC 算法的时间复杂性为$O(N * E * K)$, 而在空间复杂性上则会随着数据维数的增加树中每个结点空间呈2^E指数增长, 且随着层次的增加结点中会出现越来越多的计数为 0 的空单元, 因此图 1 所示的树形结构中的结点数组并不适合维数较高的数据集, 文献[4]在实际计算时, 对于维数大于 3 的数据集, 结点是采用链表存储结构.

下面我们给出多层网格的另一种链式存储结构, 该结构如图 2 所示. 图 2 中的示例数据点集与图 1 相同, 其中顺序表L的第r个元素指向第r层网格单元链表$(r = 1, 2, \cdots, K)$, 每层网格单元统一编号, 序号从 0 至$r * 2^E$, 链表结点的数据域包含单元序号和该单元包含的数据点数, 且链表中只包含数据点数不为 0 的网格单元结点. 为了查找和插入方便, 每层链表都含头结点, 且结点按网格单元序号顺序链接.

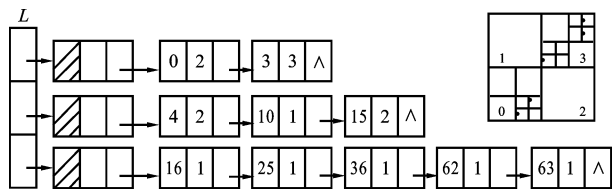


图 2 多层网格链式存储结构示意图

采用图 2 所示的数据结构, 算法思想与 LiBOC 一样, 因此不会改变时间复杂性, 但与图 1 所示的树形结构相比, 减少了大量的空单元, 结点也不必增加额外的指针信息, 降低了空间复杂性, 且结构简单易实现.

4 分形维数计算实例

基于 LiBOC 算法思想和图 2 所示数据结构对两个数据集分别计算了其分形维数, 这两个数据集一个是典型的 Sierpinski 三角形(如图 3 所示)分形数据集, 计算时取了集合中 2 000 个数据点, 记为 Sierpinski2K. 另一个数据集包含a、b、c三个属性 5 000个数据点, 其中a、b属性值随机生成, $c = a^2 + b^2$, 记为 Random5K. 两个数据集的分形维数理论值分别为

$D_2(\text{Sierpinski2K}) = D_0 = \ln(3)/\ln(2) = 1.585$,

$D_2(\text{Random5K}) = 2$

计算结果呈现在图 4、表 1 中. 图 4 中的(a)、(b)分别对应两个数据集的双对数图, 表 1 列出了两

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个数据集分形维数的计算值与理论值, 对比可知计算精度超过 10^{-2}.

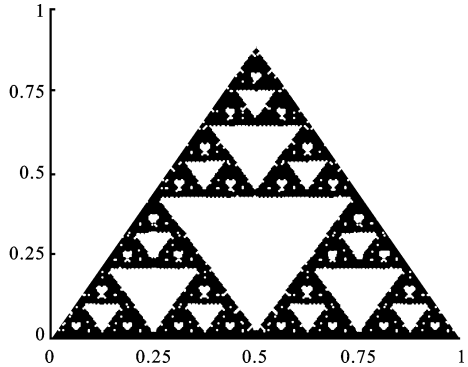


图 3　Sierpinski 三角形

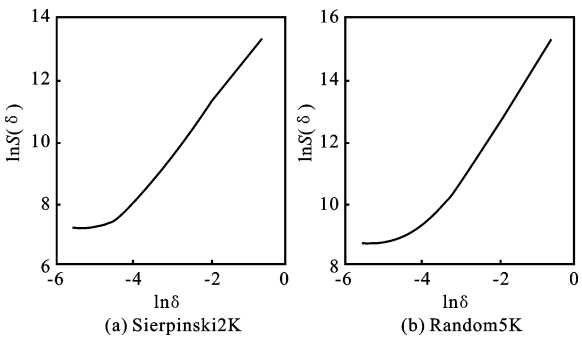


图 4　Sierpinski2K 与 Random5K 数据集双对数图

表 1　数据集分维计算值与理论值

数据集	D_2 计算值	D_2 理论值
Sierpinski2K	1. 583	1. 585
Random5K	1. 988	2

5　结束语

分形维数是分形几何在各个领域应用的主要测评工具, 采用“计盒”方式计算分形维数是常用的方法, 但“计盒”所需空间必然会随着数据点的维数增加而大量增加, 因此降低算法的空间复杂性很有必要. 本文提出的数据结构虽然简单, 但能有效降低“计盒”的空间复杂性. 另外, 实践数据还表明: 利用双对数图估算数据集的分形维数, 多层网格单元的边长缩减幅度越小, 计算值越接近理论值, 即网格单元从上往下逐层细化得越缓慢, 得到的分形维数越准确.

参考文献:

[1] Kenneth F. 分形几何数学基础及其应用[M]. 北京: 人民邮电出版社, 2007.

[2] 刘琼, 李世银, 王秀娟. 自相似网络流量建模及分析[J]. 微电子学与计算机, 2009, 25(11): 131- 134.

[3] 娄莉, 刘天时. 基于小波与分形相结合的图像压缩优化算法[J]. 微电子学与计算机, 2010, 27(6): 145- 148.

[4] Traina C, Traina A, Wu L, et al. Fast feature selection using fractal dimension[C]// Anais do XV Simpósio Brasileiro de Banco de Dados(SBBD'00), João Pessoa, Brasil: Giteseer, 2000: 1- 15.

[5] Belussi A, Faloutsos C. Estimating the selectivity of spatial queries using the Correlation Fractal dimension[C]// Proceedings of the 21st international conference on very large data bases(VLDB'95), San Francisco, CA, USA: Morgan Kaufndnn, 1995.

[6] Sousa E P M, Traina C, Traina A J M, et al. A fast and effective method to find correlations among attributes in databases[J]. Data Mining and Knowledge Discovering, 2007(14): 367- 407.

[7] Peitgen H O, Jürgens H, Saupe D. Chaos and fractals: New Frontiers of Sicence[M]. New York: Springer- Verlag, 2004.

[8] Schroeder M. Fractal, chaos, power laws: minutes from an infinite paradise[M]. New York: W. H. Freeman and Company, 1991.

作者简介:

吴新玲　女, (1963-), 教授. 研究方向为数据挖掘、软件开发方法.

周国强　男, (1962-), 高级实验师. 研究方向为软件设计与开发、多媒体技术.

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net