

DOI: 10.3901/JME.2013.13.145

基于均值漂移的 R*-树结点分裂优化算法*

孙殿柱 宋 洋 刘华东 李延瑞
(山东理工大学机械工程学院 淄博 255091)

摘要: R*-树可有效提高散乱点云、网格曲面等数据的处理效率。为降低 R*-树结点的重叠度,提高其空间利用率,将结点分裂作为模式聚类问题,采用高斯核均值漂移对结点进行模式聚类,将收敛后的模式点数量作为最佳分裂数,并以模式点为初始值结合 k -均值实现 R*-树的结点自适应分裂。试验证明,该算法可实现各类复杂几何对象的 R*-树结点分裂问题,降低 R*-树结点分裂的参数依赖性,并能有效避免 k -均值的局部收敛问题,提高 R*-树空间数据查询效率。

关键词: R*-树结点分裂 均值漂移 最优带宽 k -均值聚类

中图分类号: TP391

The Node Splitting Optimization Algorithm of R*-tree Based on Mean Shift

SUN Dianshu SONG Yang LIU Huadong LI Yanrui
(School of Mechanical Engineering, Shandong University of Technology, Zibo 255091)

Abstract: The R*-tree can improve the processing efficiency of unorganized point cloud and surface meshes. In order to reduce the overlap degree of R*-tree nodes and increase the space utilization rate, the node splitting of R*-tree is regarded as a pattern clustering problem, pattern clustering the nodes of R*-tree using Gauss mean shift algorithm, the count of mode points is considered as the best splitting number, then splitting the nodes of R*-tree with k -means algorithm whose initial values are the mode points. Experiments show that the newly proposed algorithm has good performance to solve the node splitting problems for any complex geometric object, reduce the parameter dependence, avoid the local convergence problem of k -mean effectively, and improve the R*-tree spatial query efficiency.

Key words: Node splitting of R*-tree Mean shift Optimal bandwidth k -means clustering

0 前言

R*-树是一种允许结点相互重叠的高度平衡树,具有优良的空间数据动态索引性能^[1],其原始定义适合处理二维数据,将其扩展为三维索引结构并应用于逆向工程领域,可有效提高散乱点云、三角网格、分片曲面等数据的处理效率^[2-3]。

R*-树的构建过程主要由结点插入与结点分裂两个步骤组成,其中结点分裂本质上是一个典型的聚类问题,可以基于现有成熟的聚类算法实现^[4-5]。文献[5]采用 k -均值算法,将结点分裂由传统的两路分裂改进为由聚类技术支持的多路分裂,有效降低了 R*-树结点间的重叠度,但在选择子树、结点分裂等过程中使用了结点最小外接矩形(Minimum

bounding rectangle, MBR)增量、重叠区域增量等评优指标,使得索引只对同维数据有效,若高维空间中存在低维数据(如多点共线、多点共面等),会导致 MBR 各维度尺寸严重不均,使 R*-树结点分裂失效^[6-7]。文献[8-9]将几何对象统一表示为四维点对象,以结点包围盒外接球之间的重合度作为结点间的相似度,并结合 k -均值算法实现 R*-树结点分裂;该算法可处理各种几何对象的结点分裂问题,但需要由用户输入分簇的簇数,而输入簇数的不同可能导致差别很大的聚类结果,并且由用户设置簇数的聚类分簇通常不能反映真实的结点分布,而且 k -均值算法对初始值敏感,初始值选取不当可能导致局部收敛,不能获得全局最优解。

鉴于 k -均值结点分裂算法存在参数依赖和对初始值敏感的问题,本文基于高斯核均值漂移(Mean shift, MS)算法对结点进行模式聚类,并以渐近积分平方差为评价规则自动计算全局最优带宽,对得到

* 国家自然科学基金(51075247)和山东省自然科学基金(ZR2010EM008)资助项目。20120802 收到初稿,20130418 收到修改稿

的模式点迭代使用高斯核均值漂移直到模式点收敛为止, 此时模式点的数量即为最佳分裂数, 以模式点为初始值结合 k -均值算法实现最佳分裂数下的结点分裂, 进而实现 R^* -树的结点自适应分裂。该算法可降低 R^* -树结点分裂的参数依赖性以及有效避免 k -均值的局部收敛问题, 能准确反映结点的空间分布, 降低结点间的重叠度, 提高 R^* -树的空间查询能力。

1 基于高斯核函数的均值漂移算法

设 d 维欧氏空间 R^d 中 n 个采样点集 $S = \{x_i | 1 \leq i \leq n\}$, 利用高斯核函数 $K_N(x) = (2\pi)^{-\frac{d}{2}} \times \exp(-\|x\|^2)$ 以及正定的 $d \times d$ 的带宽矩阵 H_i 对点集 S 进行核密度估计, 公式为

$$\hat{f}(x) = (2\pi)^{-\frac{d}{2}} \sum_{i=1}^n w_i |H_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|x - x_i\|_{H_i}^2\right) \quad (1)$$

式中, $w_i \geq 0$ 表示采样点的权重, 满足 $\sum w_i = 1$ 。

设带宽固定且各向一致, 均值漂移矢量 $M(x)$ 可表示为

$$M(x) = m(x) - x = H \frac{\nabla \hat{f}(x)}{\hat{f}(x)} \quad (2)$$

$$m(x) = \frac{\sum_{i=1}^n w_i \exp\left(-\frac{1}{2}\left\|\frac{x - x_i}{h}\right\|^2\right) x_i}{\sum_{i=1}^n w_i \exp\left(-\frac{1}{2}\left\|\frac{x - x_i}{h}\right\|^2\right)} \quad (3)$$

式中, $\nabla \hat{f}(x)$ 表示密度梯度估计, x_i 和 x 分别表示采样点和核中心, h 为带宽^[10]。

由式(2)可知均值漂移矢量 $M(x)$ 总是指向密度大的方向, 因此均值漂移算法最终收敛至密度极大值点。均值漂移算法根据式(3)反复迭代搜索特征空间中样本密集的区域, 即样本点沿着密度增加的方向“漂移”到局部极值点, 其过程如图 1 所示。

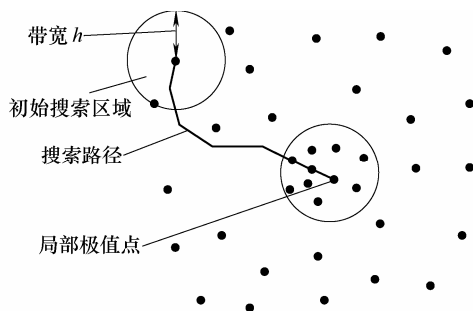


图 1 均值漂移示意图

带宽 h 是均值漂移算法的一个重要参数, 决定了参与算法迭代计算的数据量, 且会影响算法的收

敛速度与准确性, 故带宽选择的优劣直接影响结果的正确性。

评价带宽的优劣可采用密度函数估计与真实密度函数之间的误差作为评价准则, 合适的带宽可使密度函数估计接近真实密度函数, 因此把误差最小时对应的带宽作为最优带宽。而误差的度量可基于渐近积分均方差实现^[11]。即将密度函数估计与真实密度函数的渐近积分均方差定义为

$$A(\hat{p}, p) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 R(p'') \quad (4)$$

式中, $R(g) = \int_R g(x)^2 dx$, $\mu_2(g) = \int_R x^2 g(x) dx$, p'' 为真实密度函数的二阶导数, 由式(4)知 $A(\hat{p}, p) \geq 0$ 。最优带宽即是使式(4)取得最小值时对应的带宽, 因此式(4)对带宽 h 进行微分并令微分所得算式等于零, 进而解得

$$h = \left[\frac{R(K)}{\mu_2(K)^2 R(p'') N} \right]^{\frac{1}{5}} \quad (5)$$

式(5)虽给出最优带宽 h 的求解方法, 但 $R(p'')$ 依赖于真实密度函数的二阶导数, 不能直接进行带宽的计算, 故需要对密度函数的导数进行估计。密度函数的 r 阶导数估计公式为

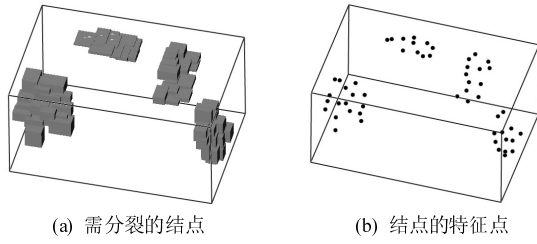
$$\hat{p}^{(r)}(x) = \frac{(-1)^r}{\sqrt{2\pi n} h^{r+1}} \sum_{i=1}^n H_r\left(\frac{x - x_i}{h}\right) \exp\left(-\frac{(x - x_i)^2}{2h^2}\right) \quad (6)$$

式中, $H_r(u)$ 是 r 阶的埃尔米特多项式。将式(6)代入式(5)即可计算出全局最优带宽。

2 基于均值漂移求解结点最优分裂数

R^* -树结点分裂的最优目标是使分裂后的结点能反映其真实空间分布, 结点的空间分布可采用其密度函数表示。

以图 2 中的结点为例分析结点最优分裂数与结点特征点密度函数的关系, 结点的特征点可采用其 MBR 的中心表示, 如图 2 b 所示。显而易见特征点的密度函数能表示结点的空间分布, 密度函数的极值点处的特征点密集, 即结点的内聚性强, 因此以极值点处的结点为聚类中心进行聚类可获取内聚较好的聚类方案。而均值漂移算法的本质是搜索密度的局部极值点, 故可采用均值漂移算法求解密度函数的极值点; 为避免存在相近的局部极值点, 可对极值点迭代使用均值漂移算法进行处理, 直到极值点收敛为止。

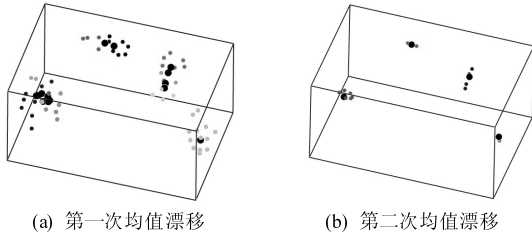


(a) 需分裂的结点 (b) 结点的特征点
图 2 R*-树的一个结点与其特征点

依据上述分析, 基于均值漂移的结点最优分裂数的求解算法如下。

- (1) 求解结点的特征点集 P , 即结点 MBR 中心点的集合。
- (2) 利用式(5)求解特征点集 P 的最优带宽 h 。
- (3) 基于均值漂移算法求解特征点集 P 的模式点(密度的局部极值)集 C 。
- (4) 以带宽 h 采用均值漂移算法对模式点集 C 进行模式划分, 得到模式点集 C' , 如若 C' 与 C 相同则跳转步骤(5), 否则令 $P=C'$, 跳转步骤(3)。
- (5) 如果 C' 中模式点的数量大于 1 则直接返回该数量值, 否则返回数值 2。

采用上述算法求解图 2 所示结点的最佳分裂数, 图 3a 为第一次均值漂移结果, 所得模式点集包含 20 个模式点, 其中存在较多距离相近的模式点, 对其再次进行均值漂移, 结果如图 3b 所示, 此时模式点集包含 4 个模式点。当对模式点集进行第三次均值漂移时, 模式点集无变化, 这意味着算法至此收敛, 结点最优分裂数为 4。



(a) 第一次均值漂移 (b) 第二次均值漂移
图 3 基于均值漂移算法的最佳分裂数求解过程

3 基于 k-均值的结点分裂

对数据点进行聚类需先定义其相似度, 一般采用数据点的欧氏距离作为相似度的定义, 但 R*-树的结点常用 MBR 表示, 直接使用 MBR 的中心距离无法反映结点的空间位置与分布, 故两结点相似度的定义如下

$$\xi_{(i,j)} = \frac{r_i + r_j}{d} \quad (7)$$

式中, r_i 、 r_j 分别为结点 N_i 、 N_j 外接球半径, d 为结点中心距。

采用上述求解的结点最优分裂数和最佳初始

点作为 k -均值算法的参数, 对 R*-树的结点进行聚类分簇, 实现 R*-树结点的最优分裂, 具体算法如下。

- (1) 基于均值漂移求解结点的最优分裂数 k 和最佳初始聚类中心 $O\{o_i | i=0,1,\dots,k-1\}$ 。
- (2) 分别计算其他结点与各聚类中心的相似度, 并将其归并到与各聚类中心相似度最大的分簇中。
- (3) 对各个分簇中的结点, 计算任一结点与该分簇中的其他结点的相似度之和, 取其最大的作为新的聚类中心。
- (4) 如果第 i 次聚类中心 O_i 与第 $i-1$ 次聚类中心 O_{i-1} 相同, 则跳转步骤(5), 否则返回步骤(2)继续执行。
- (5) 根据聚类结果中的结点划分实现 R*-树的结点分裂。

采用上述结点分裂算法对图 2a 所示的结点进行分裂, 其效果如图 4 所示。

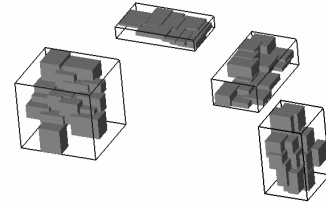


图 4 结点的分裂效果图

4 时间复杂度分析

均值漂移算法是一种高效的统计迭代算法, 它完全依靠特征空间中的样本点进行分析, 对任意形状分布都有效, 不需要任何先验知识, 原理简单、迭代效率高、收敛速度快。设结点的数量为 n , 自动计算带宽的时间复杂度为 $O(n)$, 均值漂移算法的时间复杂度为 $O(n^3)$, 设求解最佳分裂数时迭代次数为 m , 则其时间复杂度为 $O(mn^3)$, 基于 k -均值的结点分裂的时间复杂度为 $O(nkl)$, k 为簇数, l 为算法迭代次数, 故本文算法的时间复杂度为 $O(n) + O(mn^3) + O(nkl)$, 其中 m 一般取 $2 \sim 5$, $kl \ll n$, 故本文算法时间复杂度的数量级为 $O(n^3)$ 。虽然本文算法的时间复杂度较高, 但 R*-树结点包含子结点数目的上限 M 一般取 $40 \sim 50$, 且 $n = M + 1$, 即基数 n 的取值范围一般较小, 因此结点分裂时间对 R*-树的构建时间影响不显著。

5 应用实例

分别基于本文算法和文献[8]算法对图 5 所示的四个模型建立 R*-树索引, 构建时间如表 1 所示;

基于该索引对所有数据点进行 k -近邻查询, 查询时间如表 2 所示; 各层结点平均重叠区域体积分别如表 3 和表 4 所示, 其中获得以上数据所用计算机 CPU 主频为 2.5 GHz, 内存为 1 GB。由表 1、2 可知, 虽然基于本文算法构建 R*-树的效率有所降低, 但其 k -近邻查询效率得到明显提高, 为 12%~20%; 由表 3、4 可知, 采用本文算法构建的 R*-树结点重叠区域降低了 10%~20%, 提高了 R*-树的查询效率。



图 5 网格模型

表 1 本文算法及文献[8]算法构建 R*-树时间 ms

算法	图 5a 模型	图 5b 模型	图 5c 模型	图 5d 模型
本文算法	5 011	4 013	5 512	7 976
文献[8]算法	1 478	1 323	1 972	2 892

表 2 本文算法及文献[8]算法 k -近邻查询时间 s

模型	数据规模	k	本文算法	文献[8]算法
图 5a 模型	89 308	15	0.137	0.136
		30	0.170	0.189
图 5b 模型	56 361	15	0.05 89	0.069 3
		30	0.07 19	0.089 4
图 5c 模型	97 874	15	0.109	0.137
		30	0.180	0.195
图 5d 模型	109 310	15	0.190	0.217
		30	0.257	0.301

表 3 本文算法建树重叠区域体积 mm^3

结点层数	图 5a 模型	图 5b 模型	图 5c 模型	图 5d 模型
1	0	0	0	0
2	66 693.418	3 171 614.210	34 907.175	190 745.789
3	87 514.630	797 152.549	93 045.743	54 019.741
4	22 888.580	257 946.473	27 970.460	7 879.140
5	4 786.178	—	40 174.176	4 917.187

表 4 文献[8]算法建树重叠区域体积 mm^3

结点层数	图 5a 模型	图 5b 模型	图 5c 模型	图 5d 模型
1	0	0	0	0
2	94 587.833	4 301 114.199	35 474.618	391 372.279
3	90 3951.023	1 252 112.752	94 407.748	76 821.840
4	34 642.517	291 644.180	35 147.032	8 012.688
5	6 777.533	—	50 475.974	5 471.538

采用本文算法和文献[8]算法对图 5 a 模型构建的 R*-树根结点相同, 其他各层结点的效果如图 6 和图 7 所示, 图 8 为叶结点局部放大图, 文献[8]在构建 R*-树索引时出现轴向包围盒体积过大的奇异结点, 这些结点会影响 R*-树的结点的聚合度, 降低其索引性能, 而本文算法构建的 R*-树结点重叠度与体积较小。

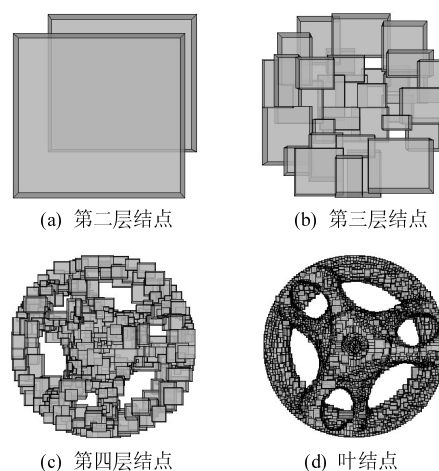


图 6 本文算法的 R*-树各层结点效果图

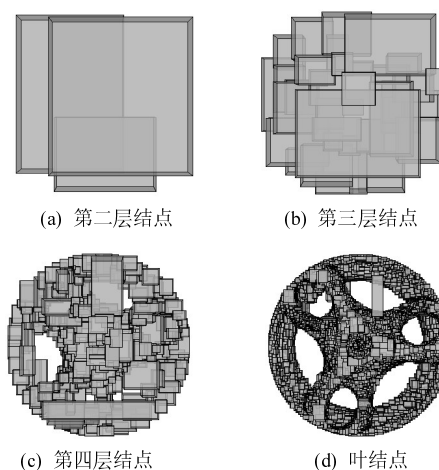


图 7 文献[8]算法的 R*-树各层结点效果图

6 结论

(1) 提出了多重自适应带宽的均值漂移算法, 基于该算法可根据 R*-树结点 MBR 的空间分布自动计算结点的最优分裂数, 从而降低了 R*-树结点

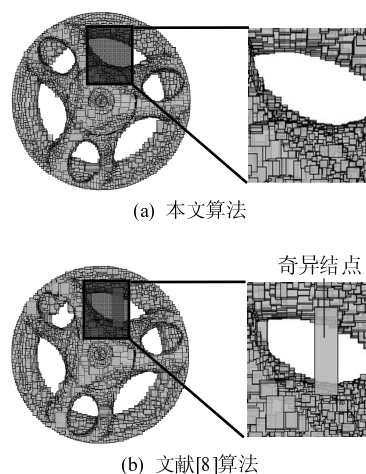


图 8 不同算法的聚类效果

分裂对聚类参数的依赖, 增强了 R*-树对散乱数据的适应性。

(2) 将多重自适应带宽的均值漂移算法的模式点集收敛结果作为 R*-树结点 k -均值聚类的初始值, 解决了 k -均值聚类算法容易陷入局部最优解的问题, 保证了结点分裂结果的最优性, 从而显著降低了结点重叠度。

(3) 基于本文算法构建 R*-树索引, 虽然构建效率有所降低, 但是索引的整体性能得到了显著提高, 鉴于实际应用中索引构建周期远小于索引应用周期, 因此以较小的索引构建效率的损失换取索引性能的提升是非常有意义的。

参 考 文 献

- [1] BECKMANN N, KRIEGEL H P, SCHNEIDER R, et al. The R*-tree: An efficient and robust access method for points and rectangles [J]. ACM IGMOD, 1990, 19(2): 322-331.
- [2] 孙殿柱, 范志先, 李延瑞, 等. 散乱数据点云型面特征分析算法的研究与应用[J]. 机械工程学报, 2007, 43(6): 133-136.
SUN Dianzhu, FAN Zhixian, LI Yanrui, et al. Research and application of surface feature analysis for scatter data points[J]. Chinese Journal of Mechanical Engineering, 2007, 43(6): 133-136.
- [3] 孙殿柱, 康新才, 李延瑞, 等. 三角 Bézier 曲面数控精加工刀轨快速生成算法[J]. 机械工程学报, 2011, 47(15): 187-192.
SUN Dianzhu, KANG Xincan, LI Yanrui, et al. Efficient algorithm for finishing NC tool path generation based on triangular Bézier surface[J]. Journal of Mechanical Engineering, 2011, 47(15): 187-192.
- [4] 张明波, 陆锋, 申排伟, 等. R 树家族的演变和发展[J]. 计算机学报, 2005, 28(3): 289-300.

- ZHANG Mingbo, LU Feng, SHEN Paiwei, et al. The evolution and progress of R-tree family[J]. Chinese Journal of Computers, 2005, 28(3): 289-300.
- [5] BRAKASTOULAS S, PFOSE D, THEODORIDIS Y. Revisiting R-tree construction principles[C]// Proceedings 6th ADBIS, London, Lecture Notes In Computer Science, 2002: 149-162.
- [6] ZHU Qing, GONG Jun, ZHANG Yeting. An efficient 3D R-tree spatial index method for virtual geographic environments[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2007, 62(3): 217-224.
- [7] 黄继先, 鲍光淑, 夏斌. 基于混合聚类算法的动态 R-树[J]. 中南大学学报, 2006, 37(2): 366-370.
HUANG Jixian, BAO Guangshu, XIA Bin. A dynamic R-tree index based on hybrid clustering algorithm[J]. Journal of Central South University, 2006, 37(2): 366-370.
- [8] 孙殿柱, 田中朝, 李延瑞, 等. 基于四维聚类的 R*-树结点分裂算法[J]. 机械工程学报, 2009, 45(10): 180-184.
SUN Dianzhu, TIAN Zhongchao, LI Yanrui, et al. Node splitting algorithm of R*-tree based on four-dimensional clustering[J]. Journal of Mechanical Engineering, 2009, 45(10): 180-184.
- [9] 孙殿柱, 李延瑞, 朱昌志, 等. 几何对象统一表示的 R*-tree 结点分裂算法[J]. 华中科技大学学报, 2010, 38(2): 55-58.
SUN Dianzhu, LI Yanrui, ZHU Changzhi, et al. Node splitting algorithm for R*-tree based on united expression of all geometry objects[J]. Journal of Huazhong University of Science and Technology, 2010, 38(2): 55-58.
- [10] 周芳芳, 樊晓平, 叶榛. 均值漂移算法的研究与应用[J]. 控制与决策, 2007, 22(8): 841-847.
ZHOU Fangfang, FAN Xiaoping, YE Zhen. Mean shift research and applications[J]. Control and Decision, 2007, 22(8): 841-847.
- [11] VIKAS R, RANMANI D. Fast optimal bandwidth selection for kernel density estimation[C]// Proceedings of the Sixth SIAM International Conference on Data Mining. 2006: 524-528.

作者简介: 孙殿柱, 男, 1956 年出生, 博士, 教授。主要研究方向为 CAD/CAM 一体化, 逆向工程。

E-mail: dianzhus@sdu.edu.cn

宋洋, 男, 1987 年出生。主要研究方向为 CAD/CAM 一体化。

E-mail: sy198704@163.com

刘华东, 男, 1985 年出生。主要研究方向为 CAD/CAM 一体化。

E-mail: liuhuadong0610@163.com

李延瑞, 男, 1979 年出生。主要研究方向为 CAD/CAM 一体化。

E-mail: lyanyr@gmail.com