基于密度的空间聚类算法研究

聂跃光,陈立潮,陈 湖
(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:基于密度的聚类算法作为数据挖掘方法中的一种主要方法,不仅可以从数据集中发现任意形状的簇,而且可以观察到一个并发的、完整的聚类结构,以及具有对噪声数据不敏感的特点。针对目前常用的几种基于密度的聚类算法及改进算法进行讨论,分析了这些密度聚类算法各自的优缺点,并且以地理信息系统为应用研究背景,提出了基于密度的聚类算法与GIS相结合,通过对多维数据属性特征的提取,扩展到多维数据的处理,在三维空间地形数据中的分析中取得了高效的聚类结果。

关键词:空间数据挖掘;聚类;密度聚类;GIS
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2008)08-0091-04

Research of Spatial Clustering Algorithms Based on Density

NIE Yue-guang, CHEN Li-chao, CHEN Hu
(Institute of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: Spatial clustering algorithm based on density can find the cluster of random shape, and some subsequent and integrate clustering structure, and lack of sensitivity to noise. Expound several density based spatial clustering and some upswing method, and analyse the advantage and disadvantage of them. Then on background of GIS propose a method of combined density based spatial clustering algorithm and GIS, which extract the attribute of the multidimensional data can expand to handle the multidimensional data, and get the good effect from the analysis of the three dimension space.

Key words: spatial data mining; clustering; density clustering; GIS

0 引 言

随着数据挖掘研究领域技术的发展,特别是高维及空间数据挖掘的兴起,作为数据挖掘主要方法之一的聚类算法,也越来越受到人们的关注。数据挖掘(Data Mining),也称知识发现[1],是从数据库中便捷地抽取出以前未知的、隐含的、有用的信息,所挖掘出来的知识可应用于信息管理、决策支持、过程控制和其它许多应用。所谓聚类,就是把大量的 d 维数据样本(n 个)聚集成 k 个类(k,n),使同一类中样本的相似性最大,而不同类中样本的相似性最小。

基于密度的聚类算法将簇看作是数据空间中被较低密度的区域分割开的高密度对象区域,因此可以发现任意形状的簇,并能识别噪声数据。根据实现方法,该类算法分为基于局部连通性(local connectivity)和基于密度函数两种。前者将局部范围内密度相对高的区域连通起来,形成一个簇,代表算法有DBSCAN算法、OPTICS算法、CLIQUE算法等;后者用密度函数来模拟数据集的密度分布,代表算法有DENCLUE算法等。

收稿日期:2007-11-26
基金项目:山西省自然科学基金(200501044)
作者简介:聂跃光(1982-),男,山西忻州人,硕士研究生,研究方向为人工智能;陈立潮,教授,博士,研究方向为模式识别、人工智能、数据挖掘。

1 几种用于空间数据挖掘的密度聚类算法

1.1 DBSCAN算法

DBSCAN(Density Based Spatial Clustering of Applications with Noise)算法[2],是一种比较有代表性的基于密度的空间聚类算法,它将簇定义为密度相连的点的最大集合,能够把具有足够高密度的区域划分为簇,并可在有"噪声"的空间数据库中发现任意形状的聚类。

DBSCAN的算法思想是:从数据集 D 中的任意一个点 p 开始,查找 D 中所有关于Eps(最小半径)和MinPts(密度阈值)的从 p 密度可达的点。若 p 是核心

点,则其邻域内的所有点和 p 同属于一个簇,这些点将作为下一轮的考察对象(即种子点),并通过不断查找从种子点密度可达的点来扩展它们所在的簇,直至找到一个完整的簇;若 p 不是核心点,即没有对象从 p 密度可达,则 p 被暂时地标注为噪声。然后,算法对 D 中的下一个对象重复上述过程 …… 当所有种子点都被考察过,一个簇就扩展完成了。此时,若 D 中还有未处理的点,算法则进行另一个簇的扩展;否则,D 中不属于任何簇的点即为噪声。

DBSCAN 的算法描述如下:

1) 输入:包含 n 个对象的数据库,半径 Eps,最少数目 MinPts。

2) 输出:所有生成的簇,达到密度要求。

(1) 从数据集中任意选取一个点 p,并对其进行区域查询;

(2) 如果 p 是核心点,则寻找所有从 p 密度可达的点,最终形成一个包含 p 的簇;

(3) 否则,p 被暂时标注为噪声点;

(4) 访问数据集中的下一个点,重复上述过程,直到数据集中所有的点都被处理。

算法通过检查数据集 D 中每个点的 Eps 邻域来判断它是否是核心点进而决定如何扩展簇,因此算法对数据集中的每个点进行区域查询时,都要扫描整个数据集,因此时间复杂度是 $O(n^2)$。

DBSCAN 算法作为一种有代表性的基于密度的聚类算法,从提出至今,已经成功应用于城市规划、城市建设、选址等多个研究领域,并在发展过程中产生了几种有效的改进算法。具体主要有以下几种。

周水庚等人为了克服 DBSCAN 算法在处理大规模数据时的内存和 I/O 瓶颈,提出了 SDBSCAN (Sampling - based DBSCAN)算法[3]。算法基于数据取样技术来扩展 DBSCAN 算法,使之能够有效地对大规模数据集进行聚类分析,并且较大幅度地提高整个聚类过程的效率。何中胜等人为了解决不同密度簇聚类的问题,提出了一种基于数据分区的并行密度聚类算法 PDBSCAN (Partitioning - based DBSCAN)[4],算法基于数据分区并行技术来扩展 DBSCAN 算法,它根据数据库在某一维的分布特征,将整个数据库数据空间划分为若干个交集为空的局部区域,对每一个局部区域用 DBSCAN 进行局部聚类,最后将各局部聚类合并,从而完成整个数据库的聚类分析。不仅可以缓解用全局 Eps 聚类的质量恶化的问题,还提高了聚类的速度。周水庚等人提出了一种进行快速密度聚类的 FDBSCAN(Fast DBSCAN)[5]算法。选择核心点 p 邻域中的部分代表点用于簇的扩张,这样只有这些代表点需要进行区域查询操作,因此大大减少区域查询的次数,降低 I/O 开销,算法的速度也大大加快了。

1.2 OPTICS 算法

现实中的高维数据集经常分布不均匀,因而使用全局密度参数不能刻画其内在的聚类结构,为解决这个难题,Mihael 等人提出了 OPTICS(Ordering Points to Identify the Clustering Structure)聚类算法[6]。该算法扩展了 DBSCAN 算法,它并不产生数据集的聚类,而是生成代表基于密度的聚类结构的一个参数化的数据库的排序。这种排序包含的信息与对应于一个宽范围的参数设置的基于密度的聚类是相同的,并且是自动聚类分析与交互聚类分析的基础。聚类的结果可以用图或者其他可视化技术来表示。

对于一个固定的 MinPts 值,较高密度(Eps 取值较低)的聚类结果被完全包含在较低密度(Eps 取值较高)时所获得的密度连接中。因此,需要对 DBSCAN 算法进行扩展,使其能同时处理若干距离参数,并能同时生成不同密度的簇。为了产生一致的结果,进行聚类扩展时,对象应当以特定的顺序来处理。这个顺序根据最小的 Eps 值密度可达的对象来选择,以保证较高密度的簇能被首先完成。基于这个思路,每个对象需要存储两个值:核心距离(core - distance)和可达距离(reachability - distance)。

定义 1:对象 p 的核心距离是使得 p 成为核心对象的最小 Eps 值。若 p 无法成为核心对象,则 p 的核心距离没有定义。

定义 2:对象 q 关于对象 p 的可达距离是 p 的核心距离和 p 与 q 的欧几里得距离之间的较大值。若 p 不是核心对象,则 p 和 q 之间的可达距离没有定义。

OPTICS 算法生成了数据对象集合 D 的一个排序,并且为每一个对象存贮了核心距离和可达距离。该算法的思路是首先检查数据对象集合 D 中任一个对象的 ε - 邻域。设定其可达距离为“未定义”,并确定其核心距离,然后将对象及其核心距离和可达距离写入文件。如果 p 是核心对象,则将对象 p 的 ε - 邻域内的对象 $N\varepsilon(p)$ 插入到一个种子队列中,包含在种子队列中的对象按照其直接密度可达的最近的核心对象的可达距离排序。种子队列中具有最小可达距离的对象被首先挑选出来,确定该对象的 ε - 邻域和核心距离,然后将其该对象及其核心距离和可达距离写入文件中,如果当前对象是核心对象,则更多的用于扩展的后选对象被插入到种子队列中。这个处理一直重复到再没有一个新的对象被加入到当前的种子队列中。

由于 OPTICS 算法与 DBSCAN 在结构上等价,因此 OPTICS 算法具有和 DBSCAN 相同的时间复杂度,

即当采用空间索引时,复杂度为 $O(n\log n)$。

1.3 CLIQUE算法

CLIQUE(Clustering In Quest)[7]是基于密度和网格的聚类算法,它能自动识别高维数据的子空间中的簇,即子空间中相连接的高密度单元的并集。

算法只对原始空间的子空间进行搜索,而不在新的维度(如由原始的维度线性组合成新的维度)上搜索。聚类形成的簇可以用最小数量的、面积最大的、并可能互相重叠的矩形覆盖,所以可以用这些矩形的并集来描述簇。一个簇的最小描述是具有最大区域的矩形的非冗余覆盖。

算法主要分为四步:

(1)划分数据空间,并统计每个网格单元中包含的点数。

(2)使用类似于关联规则挖掘中的 Apriori 算法识别密度单元:首先确定 k 维的密度单元,……,在确定了 $k-1$ 维密度单元的基础上,生成候选的 k 维密度单元,直到不再有候选单元产生时结束。

(3) 识别聚类:采用在图中发现连通图的方法,确定数据集的划分,每个划分由相连的高密度单元组成,同时对应着一个簇。

(4) 生成簇的最小描述(minimal description):使用贪心算法生成聚类的最大矩形区域覆盖,然后去掉多余的矩形,从而生成簇的最小覆盖。

CLIQUE 能够自动识别出具有最高维数的子空间中的簇;对记录输入的次序不敏感;对数据的分布没有任何数学形式的假设;算法运行时间和输入数据的大小成线性关系,并且当维数增加时,具有良好的可伸缩性。但是算法在划分网格时没有或者很少考虑数据的分布,而且用网格内的统计信息来代替该网格内的所有点,从而导致了聚类结果的精确性下降。

1.4 DENCLUE算法

DENCLUE(DENsity - based CLUstEring)[8]是一个基于一组密度分布函数的聚类算法。算法主要思想是:每个数据点的影响可以用一个数学函数来形式化地模拟。数学函数描述了一个数据点在其邻域内的影响,因此被称为影响函数(influence function)。每个点的影响函数是其他所有数据点在该点的影响函数之和,则数据空间的整体密度可以被模型化为所有数据点的影响函数的总和,即全局密度函数。这样,簇可以通过确定密度吸引点(density attractor)来获得,这里的密度吸引点是全局密度函数的局部最大。

影响函数可以是一个任意的函数,例如抛物线函数、方波函数或高斯函数。可以定义密度函数的梯度和密度吸引点,如果存在一组点 x_0, x_1, x_k, $x_i = x$, $x_k = x*, 0 < i < k$, x_{i-1} 的梯度在 x_i 的方向上,则点 x 是被一个密度吸引点 $x*$ 密度吸引的。对于一个连续和可微的影响函数,可以用一个梯度指导的爬山算法来计算一组数据点的密度吸引点。

基于上述概念,算法形式化地定义中心定义的簇(center - defined cluster)和任意形状的簇(arbitrary - shape cluster)。一个中心定义的簇是一个被密度吸引点 $x*$ 密度吸引的子集 C,即点在 x_0 的密度函数不小于一个阈值;否则它被认为是孤立点一个任意形状的簇是子集 C 的集合,每一个都是密度吸引的,有不小于阈值的密度函数值,并从每个区域到另一个都存在一条路径 P,该路径上每个点的密度函数值都不小于 ξ。

该算法有坚实的数学基础,可以概括其他聚类方法:可以过滤大量"噪声";处理高维数据时效率很高,并且可以用数学方式简洁地描述任意形状的簇:它使用网格单元,并且只保存那些包含数据点的网格单元信息,同时以一个基于树的存取结构来管理这些单元,因此比一般聚类算法的速度要快。

2 密度聚类算法在GIS中的应用

地理信息系统(Geography Information System, GIS)[9]是集计算机科学、地理地质学、测绘科学、环境科学、空间科学、信息科学和管理科学等为一体的多学科结合的边缘科学,现有的地理信息系统(GIS)一般具有强大的空间数据管理、制图、查询和空间分析的功能,但缺乏对知识的发现、表达方法和机制。

空间数据挖掘技术是提高 GIS 智能化的重要手段。面对空间数据库的聚类算法所需要的参数能自动确定或用户容易确定,因此在 GIS 系统中加入空间数据挖掘技术,可以自动或半自动地从空间数据中发现一些特定的知识和普遍知识,直接提供给决策者使用,或指导技术人员的后续处理。

总之,研究基于密度聚类的空间数据挖掘技术,一方面可使 GIS 查询和分析技术提高到发现知识的新阶段,另一方面从中发现的知识可构成知识库用于建立智能化的 GIS 系统,为决策者提供有价值的知识,带来不可估的效益,因此基于聚类的空间数据挖掘方法与应用研究具有重要的理论意义和实用价值。

3 各种密度聚类算法的比较

基于上述分析,得到各密度聚类算法的比较结果,结论如表1所示。可以看出,每种算法都有其各自的特点和适用领域,在数据挖掘过程中,用户根据需要选

择合适的聚类算法。

表 1 密度聚类算法比较结果表

	算法效率	适合的数据类型	发现的聚类类型	对脏数据或异常数据的敏感性	对数据输入顺序的敏感性
DBSCAN	一般	数值型	任意	敏感	敏感
SDBSCAN	较高	数值型	任意	敏感	敏感
PDBSCAN	高	数值型	任意	敏感	敏感
FDBSCAN	较高	数值型	任意	敏感	敏感
OPTICS	一般	数值型	- - -	一般	不敏感
CLIQUE	较低	任意型	凸形或球形	一般	不敏感
DENCLUE	较高	数值型	任意	不敏感	不敏感

4 结束语

基于密度的聚类算法由于其可发现任意形状的簇，并且对噪声数据不敏感，常用于遥感图像的分析及选址、选线等工程中。在 GIS 中，基于密度的聚类可以从大量的空间数据中，通过数据点属性间的关联进行聚类，从而得到具有特别形状的聚类。由于高维空间技术的发展，如何从高维空间中发现大量的有用信息是今后研究的重点，基于密度的聚类算法由于其本身具有的特点，通过对高维数据属性特征的提取，可扩展到对高维数据的处理。

参考文献：

[1] Han Jiawei, Kambr M. Data Mining : Concepts and Techniques [M]. [s. l.]: Morgan Kaufmann Publishers, 2001.

[2] Ester M, Kriegel H P, Sander J, et al. A Density - Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//Proc. of 2nd Int. Conf, on Knowledge Discovery and Data Mining (KDD - 96). Portland, Oregon: [s. n.], 1996.

[3] 周水庚，范 晔，周傲英. 基于数据取样的 DBSCAN 算法[J]. 小型微型计算机系统，2000，21(12)：1270 - 1274.

[4] 何中胜，刘宗田，庄燕滨. 基于数据分区的并行 DBSCAN 算法[J]. 小型微型计算机系统，2006，27(1)：114 - 116.

[5] 周水庚，周傲英，曹 晶，等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展，2000，37(11)：1287 - 1292.

[6] Ankerst M, Breuning M M, Kriegel H P, et al. OPTICS: Ordering Points To Identify the Clustering Structure[C]//ACMSIGMOD Int. Conf. on Management of Data Philadephia PA: [s. n.], 1999.

[7] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]//ACM SIGMOD Int. Conf. USA: [s. n.], 1998: 94 - 105.

[8] Hinneburg A, Keim D C. An Efficient Approach to Clustering in Large Multimedia Databases with Noise [C]//ACM SIGKDD Int. Conf. USA: [s. n.], 1998: 58 - 65.

[9] 吴信才，白玉琪，郭玲玲. 地理信息系统——原理、方法和应用[M]. 北京：科学出版社，2001.

(上接第 90 页)

实验结果表明，改进后的算法较有效地抑制了噪声的放大，保护了图像的边缘信息，且改善了退化图像的视觉效果。

4 结束语

将零相位 RIF 和小波去噪技术引入到 NAS - RIF 图像盲复原算法中，由于小波变换具有对称性、正交性、光滑性和紧支性等特性和零相位 RIF 的保护图像的线条和边缘特性，改进后的算法不仅有较好的抗噪性能，而且保持退化图像边缘特征上也表现出一定的优势。但就此复原算法仅对背景均匀的退化图像有较好的复原性能，要将算法运用于非均匀背景的退化图像且达到一定的复原性能，仍需进一步研究。

参考文献：

[1] Banham M R, Katsaggelos A K. Digital image restoration[J]. IEEE signal processing magazine, 1997(3): 24 - 41.

[2] 张 航，罗大庸. 图像盲复原算法研究现状及其展望[J]. 中国图像图形学报，2004，9(10)：1145 - 1152.

[3] Kundur D, Hatzinakos D. Blind image deconvolution[J]. IEEE Signal Processing Mag, 1996, 13(3): 43 - 64.

[4] Kundur D. Blind deconvolution of still images using recursive inverse filtering[D]. Toronto: Toronto Univ. , 1995.

[5] Kundur D, Hatzinakos D. A novel blind deconvolution scheme for image restoration using recursive filtering[J]. IEEE Trans on Signal Processing, 1998, 26(2): 375 - 390.

[6] Chin Ann Ong, Chambers J A. An enhanced NAS - RIF algorithm for blind image deconvolution[J]. IEEE Trans on Image Processing, 1998, 8(7): 988 - 992.

[7] 谢杰成，张大力，徐文立. 小波图像去噪综述[J]. 中国图像图形学报，2002，7(3)：209 - 217.

[8] 柳 微，马争鸣. 基于边缘检测的图像小波阈值去噪方法[J]. 中国图像图形学报，2002，7(8)：788 - 793.

[9] Donoho D L, Johnstone I M. Ideal spatial adaptation via wavelet shrinkage[J]. Biometrika, 1994, 81: 425 - 455.