doi:10.6041/j.issn.1000-1298.2016.10.044

DBSCAN 算法优化及在村镇管理决策中的应用

彭　波　史春雷　高万林
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 作为空间数据挖掘技术中的一种,带有噪声的空间聚类应用算法(DBSCAN 算法)是基于密度的聚类算法,其可以从空间数据库中发现任意形状的聚类。本文研究了基于密度的空间聚类算法优化原理及实现过程,分析了原始 DBSCAN 算法存在的问题,通过避免公共领域对象的重复查询,减少对核心对象邻域查询的计算,优化后算法的时间效率提高了 33.73%。将优化后的 DBSCAN 算法应用于村镇网格化管理,可对网格化管理系统中的数据记录进行有效挖掘,为村镇管理工作提供信息和辅助决策。

关键词: 数据挖掘; 空间聚类; 村镇管理; DBSCAN 算法
中图分类号: TP311　　**文献标识码**: A　　**文章编号**: 1000-1298(2016)10-0346-05

Optimization and Application of DBSCAN Algorithm in Management of Villages and Towns

Peng Bo　Shi Chunlei　Gao Wanlin
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: As one of the spatial data mining technologies, DBSCAN algorithm is a density-based clustering algorithm. Since it can find clusters with any forms from the spatial database, DBSCAN algorithm becomes more and more popular. The optimization principle and realization process of density-based spatial clustering algorithm were studied in detail, and the existing problems of original DBSCAN algorithm were analyzed. By avoiding repeated searches of objects in the public domain, the computation of searches on the neighborhood of core object was reduced, and the time efficiency of the algorithm was improved. After analyzing the distribution of roadside stall business in rural areas, two key parameters, i.e., *Eps* and *MinPts*, of the algorithm and the searching zone of neighborhood of core object were determined. The experiment results showed that the time efficiency of optimized algorithm was improved by approximately 33.73%. Finally, the optimized algorithm was applied to the community grid management in rural areas. By data mining of the rural area grid management system, the most frequent regions were successfully identified for roadside stall business. Using this algorithm, the hot spots of problems in rural area management can be found out in time, which uncovered the common rules hidden behind the routine business. Hence, the corresponding management can be performed to a certain region, which can provide information and auxiliary decisions for rural area management.

Key words: data mining; spatial clustering; rural area management; DBSCAN algorithm

引言

空间数据挖掘是指从空间数据库中提取事先未知、隐含其中、最终可理解的空间或非空间的一般知识规则的过程。空间数据挖掘的空间知识主要包括空间分类、关联、聚类和特征等规则及例外[1]。

收稿日期: 2016-03-14　修回日期: 2016-04-10
基金项目: 国家星火计划项目(2015GA600002)
作者简介: 彭波(1960—),女,教授,博士生导师,主要从事媒体信息处理研究,E-mail: pengbo_cau@126.com

目前已有大量的聚类算法，总体上，可以分为以下几类[2-12]：①基于划分的聚类算法，对于一个 n 个元组或对象的数据库，将其构建为 k 个划分（$k \leqslant n$）。②基于层次的聚类算法，将数据对象集合组成一棵聚类的树。层次方法根据层次分解的形成，可以分为分裂的和凝聚的。③基于密度的聚类算法，在数据空间中将簇看作是被低密度区域分隔开的高密度对象区域。④基于网格的聚类算法，采用一个多分辨率的网格数据结构，将空间划分为有限个单元。⑤基于模型的聚类算法，通常基于一个假设：数据是根据潜在的概率分布生成的，然后对给定的数据和某些数学模型之间的适应性进行优化。

本文基于密度的空间聚类算法（Density-based spatial clustering of application with noise，DBSCAN）对大规模数据聚类时间开销大的问题，描述算法的原理。DBSCAN 算法对公共邻域对象的重复查询，会导致算法对核心对象邻域查询计算量的增加，使得算法的时间效率降低。本文对 DBSCAN 算法进行优化，从减少算法对公共邻域对象的查询次数来提高算法的时间效率。

1　DBSCAN 算法及村镇网格化管理

1.1　DBSCAN 算法思想

DBSCAN 算法是一种基于密度的空间数据聚类算法，由 MARTIN 等[13]提出。其基本思想是：对于空间数据库中的一个点，在给定半径的邻域内只要其包含的数据对象个数大于某个给定值，就继续聚类[14]。算法包含 2 个参数：*Eps* 和 *MinPts*。*Eps* 是给定的区域半径的大小，*MinPts* 是指定点的邻域内包含的数据对象的最小个数。

1.2　DBSCAN 算法聚类过程

（1）建立原始数据集 D，对数据集中所有数据对象进行初始化，遍历所有的点，若找到其中一个点 q_j 为核心点，将其存放于搜索集 S 中。

（2）遍历搜索集 S，把 S 中的所有点作为种子点进行检验，对于点 p_i，若 p_i 目前不属于任何一个聚类，搜索其邻域并判断是否是一个核心对象，若 p_i 是核心对象，说明其和点 q_j 同属于一个聚类，并将其邻域内数据对象纳入搜索集 S 中。若 p_i 是边界对象，则将其和点 q_j 划为同一个聚类。最后把点 p_i 从 S 中删除。

（3）依次遍历搜索集 S 直到为空。

（4）遍历步骤（1）数据集 D 中不属于任何聚类的数据对象，重复步骤（2）、步骤（3）。

至此，对数据集 D 完成了 DBSCAN 聚类操作，各个数据对象分属于不同的簇，数据集中没有包含在任何簇中的数据对象就构成了噪点，算法执行流程如图 1 所示。

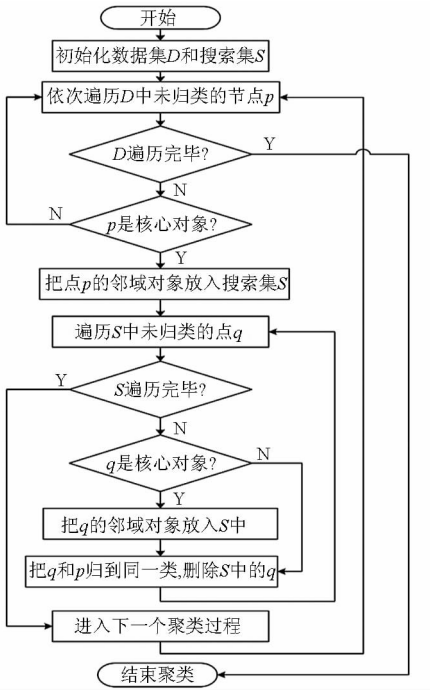

图 1　原始 DBSCAN 算法执行流程

Fig. 1　Implementation process of original DBSCAN algorithm

1.3　村镇网格化管理

村镇网格化管理是近几年我国在基层社会管理建设中摸索出的一种新型管理模式[15-21]，它主要是指按照一定标准把村镇社区划分为多个网格，以信息技术和网格间的协调机制为依托，实现网格间的资源共享，最终达到资源优化配置，提高村镇社区管理的效率，它可以解决社区信息化由于跨部门的协同工作和信息整合、共享带来的问题。村镇网格化管理信息系统中积累了海量数据记录，可为针对村镇社区管理业务的数据挖掘工作提供前提条件。

2　DBSCAN 算法优化及村镇占道经营事件的分布分析

2.1　避免公共邻域对象的重复查询

分析 DBSCAN 算法的原理可知，算法聚类过程中对于初始对象的选取是随机的，对于给定的原始数据集，可能存在多个核心对象，如图 2 所示。图 2 中，p、q、r 均是核心对象，由于三者邻域存在交集，在对数据集进行遍历的过程中造成对交集内的公共数据对象邻域查询的重复操作。这就使得整个聚类过程遍历数据量增加，时间开销变大，若避免这些公共数据对象的重复遍历，则可提高算法的时间效率。

2.2　村镇占道经营事件的分布规律

将 DBSCAN 算法应用在村镇管理中的目的是

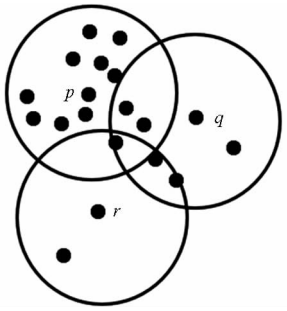

图 2 邻域相交的核心对象

Fig. 2 Core objects of neighborhood intersection

找出某类村镇管理事件的高发区域,高发区域即某区域在某段时间内事件发生的频度达到一定程度。结合实际中发生的村镇占道经营事件,本文规定半径 10 m 的范围内一年内发生 20 ~ 30 件占道经营事件视为一个占道经营的高发区域。DBSCAN 算法中需要人工输入 *Eps* 和 *MinPts*,一般由用户根据经验设定。根据村镇占道经营高发区域的定义,并能保障由算法得到的聚类结果的准确性,设定 *Eps* 为 10、*MinPts* 为 30。如图 3 所示,假设圆心 *O* 是一个核心对象,则图示圆形区域内至少分布有 30 件占道经营事件。现实中商贩选择摊点时一般沿街道直线分布,可视为 30 件事件在某一半径上沿直线均匀分布,故距圆心 *O* 2/3 半径范围内至少分布有 20 件占道经营事件。

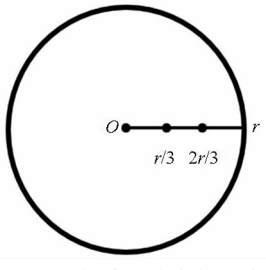

图 3 占道经营事件分布

Fig. 3 Roadside stall business distribution

2.3 DBSCAN 算法优化

对 DBSCAN 算法优化的基本思想是:结合村镇占道经营事件的分布规律及高发区域的定义,距核心对象 *O* 2/3 邻域半径内的所有对象也可形成一个高发区域,再对此范围内的数据点进行遍历会造成大量重复查询。如图 4 所示,当前数据对象 p 操作完毕后,选择下一个待处理数据对象前先排除当前数据对象 2/3 邻域半径内的点,选择邻域内其他区域的点,避免部分数据对象的重复查询,从而减少查询时间。2 个数据对象之间的距离采用欧几里得数学模型,即

$$d(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

其中 $d(x,y)$ 满足以下条件:$d(x,y) \geqslant 0$,当 $x = y$ 时,

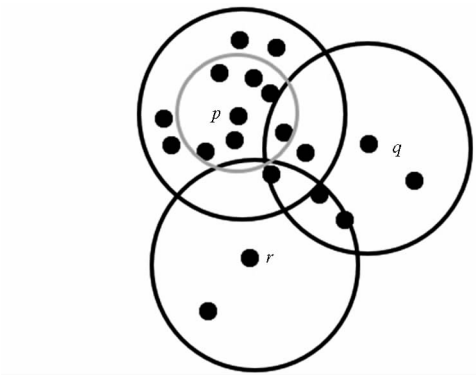

图 4 公共数据对象的限定查询

Fig. 4 Limited query for public data objects

$d(x,y) = 0$,$d(x,y) = d(y,x)$。

优化算法执行步骤如下:

(1)原始数据集的初始化

①在数据对象的数据结构定义中添加一个新的字段 *ID*,表示聚类分类结果,初始化为零。②定义一个临时搜索数据集合 *S*,用于存储检索结果。③初始化参数 *MinPts* 和 *Eps*。

(2)遍历原始数据集

①将数据集每一个点依次作为种子节点进行考察,令 $i = 1$,$j = 1$,*Cluster* = 1。对于当前对象 q_j,若 $q_j ID = 0$,则搜索它的邻域,若邻域内包含点数大于 *MinPts*,说明其为核心对象,把 q_j 及其邻域内所有对象的 *ID* 置为 *Cluster*,并将 q_j 邻域内距离其 2/3 邻域半径外的所有点存入 *S* 中。②对 *S* 进行遍历,把 *S* 中每一个点作为种子点依次进行考察,对于点 p_i,若 p_i 当前不属于任何类即 $p_i ID = 0$,搜索它的邻域。若 p_i 是核心对象,说明它是 q_j 的直接密度可达点,令 $p_i ID = Cluster$,同时将其邻域内所有对象的 *ID* 置为 *Cluster*,并将 p_i 邻域内距离其 2/3 邻域半径外的所有点存入 *S* 中。最后删除 *S* 中的 p_i。③令 $i = i + 1$,执行步骤②直到 *S* 为空。

(3)令 $j = j + 1$,*Cluster* = *Cluster* + 1,重复步骤(2),直到对数据集中所有对象都进行了遍历。

优化后的 DBSCAN 算法流程如图 5 所示。

2.4 实验结果与分析

采用 5 组仿真数据分别对原始 DBSCAN 聚类算法和优化后聚类算法进行了对比分析,实验结果如表 1 所示。

原始 DBSCAN 算法通过判断每个点 *Eps* 邻域内点的个数判断其是否是核心对象,进而扫描其所有的核心对象并存放进搜索集。算法的一大部分工作都在对核心对象邻域的扫描上,判断每个点是否是核心对象都要扫描整个数据集,因此算法时间开销较大。本文对算法进行优化后,在将某点邻域内对象存入搜索集时选择其2/3邻域半径外的核心对

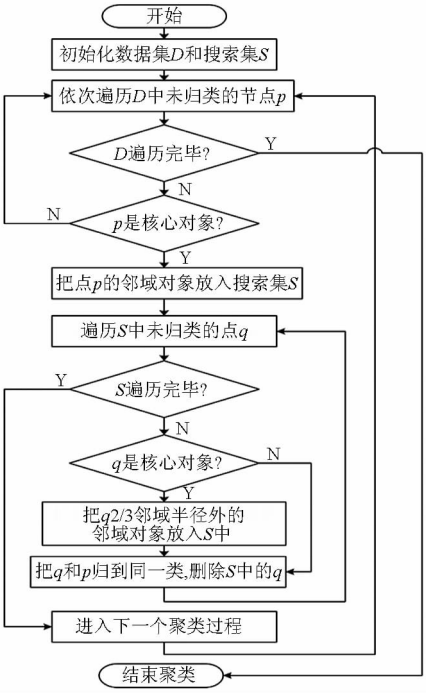

图 5　优化后的 DBSCAN 算法流程图

Fig. 5　Flow chart of optimized DBSCAN algorithm

表 1　改进前后的聚类算法耗时比较

Tab. 1　Time consuming comparison of clustering algorithms

数据集样本容量/k	2	6	10	15	20
原始算法耗时 t_1/ms	531	1 140	2 143	2 917	5 869
本文算法耗时 t_2/ms	265	765	1 686	2 241	3 452
减少率/%	50.09	32.89	21.33	23.17	41.18

象，从而使算法的扫描范围大大下降，算法时间复杂度的系数减小。

由表 1 的算法耗时可以计算得到减少的平均比率，可知优化后的 DBSCAN 聚类算法比原始 DBSCAN 聚类算法在时间性能上提高约 33.73%。改进前后算法的耗时随着数据样本容量的增多差距逐渐加大，主要原因在于随着数据集样本容量的增多，改进后的算法避免了大量公共邻域对象的重复查询。

3　优化 DBSCAN 算法的应用

3.1　实例研究

以某地区村镇社区网格化管理系统运行情况为例，一年内某区域在系统内产生的商贩占道经营类事件记录有 657 条。对这 657 条数据的地理坐标进行处理，二维空间位置分布如图 6 所示。

采用本文优化的 DBSCAN 算法对数据记录进行聚类分析，参数设置结合上文分析：$Eps = 10$，$MinPts = 30$。

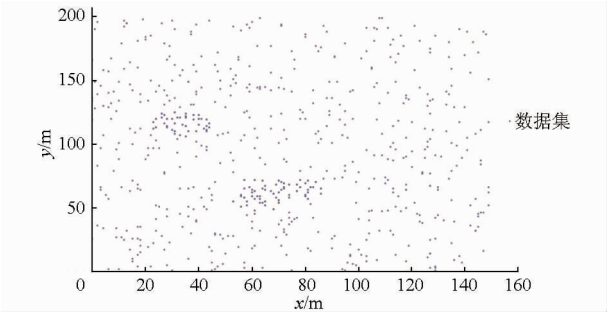

图 6　村镇社区占道经营类事件分布

Fig. 6　Two-dimensional distribution of roadside stall business event

3.2　结果分析

经过本文算法的聚类处理，该区域商贩占道经营类事件的聚类分析结果如图 7 所示，图中空心区域代表算法生成的第 1 个聚类，实心区域代表算法生成的第 2 个聚类。相比较原始 DBSCAN 算法，本文算法可以更快速地发现聚类，找出村镇占道经营的高发区域，这在处理大数据量的情况下具有明显优势。

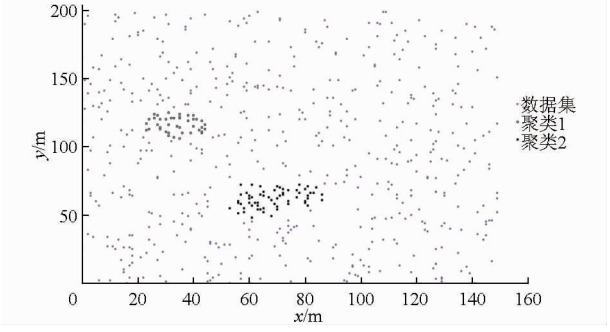

图 7　村镇社区占道经营类事件聚类分析结果

Fig. 7　Results of cluster analysis for roadside stall business event

从图 7 分析可知，商贩占道经营类事件有 2 个高发区域，第 1 个高发区域即聚类 1 所代表区域，坐标系范围是：23 m≤x≤44 m，106 m≤y≤124 m。第 2 个高发区域即聚类 2 所代表区域，坐标系范围是：53 m≤x≤86 m，49 m≤y≤71 m。

平日看似无序的村镇管理问题数据，通过空间聚类分析，得到了问题高发区域的的空间分布规律，政府管理人员可根据空间聚类分析的结果，具体分析形成问题高发区域的原因，对该区域有针对性的采取管理措施。

4　结束语

研究了基于密度的空间聚类算法优化原理及实现过程，分析了原始 DBSCAN 算法存在的问题，通过避免公共领域对象的重复查询，减少对核心对象邻域查询的计算，提高了算法的时间效率。经实验

结果分析,优化后的算法在时间效率上提高约33.73%。最后本文将其应用于村镇社区网格化管理,进行实例研究。应用该算法可以及时发现村镇管理中问题事件的高发区域,挖掘日常业务中隐藏的规律性知识,为村镇管理工作提供信息和辅助决策。

参 考 文 献

1 李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论与方法[J]. 武汉大学学报:信息科学版, 2002,27(3):224 - 229.
LI Deren, WANG Shuliang, LI Deyi, et al. Theories and technologies of spatial data mining and knowledge discovery[J]. Geomatics and Information Science of Wuhan University, 2002,27(3):224 - 229. (in Chinese)

2 WAZAVKAR S V, MAINJREKAR A A. Text clustering using HFREC - CA and rough k-means cluster algorithm[J]. Discovery, 2014,15(40):44 - 47.

3 ZHANG Chunfei, FANG Zhiyi. An improved k-means clustering algorithm[J]. Journal of Information & Computational Science, 2013,10(1):193 - 199.

4 TRIKHA P, VIJENDRA S. Fast density based clustering algorithm[J]. International Joural of Machine Learning and Computing, 2013,3(1):10 - 12.

5 ZHONG Luo, TANG Kunhao, LI Lin, et al. An improved clustering algofithm of tunnel monitoring data for cloud computing[J]. The Scientific World Journal, 2014(2014):2 - 5.

6 李晓黎,刘继敏,史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器[J]. 计算机学报,2001,24(1):62 - 68.
LI Xiaoli, LIU Jimin, SHI Zhongzhi. A Chinese web page classifier based on support vector machine and unsupervised clustering [J]. Chinese Journal of Computers, 2001,24(1):62 - 68. (in Chinese)

7 BLATT M, WISEMAN S, DOMANY E. Superparamagetic clustering of data[J]. Physical Review Letters, 1996,76(18):3251 - 3254.

8 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):49 - 58.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research[J]. Journal of Software, 2008,19(1):49 - 58. (in Chinese)

9 韩忠明, 陈妮, 乐嘉锦,等. 面向热点话题时间序列的有效聚类算法研究[J]. 计算机学报,2012,35(15):2337 - 2346.
HAN Zhongming, CHEN Ni, LE Jiajin, et al. An efficient and effective clustering algorithm for time series of hot topics[J]. Chinese Journal of Computers, 2012,35(15):2337 - 2346. (in Chinese)

10 张敏,于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004,15(6):858 - 862.
ZHANG Min, YU Jian. Fuzzy partitional clustering algorithms[J]. Journal of Software, 2004,15(6):858 - 862. (in Chinese)

11 黄韬,刘胜辉,谭艳娜. 基于 k-means 聚类算法的研究[J]. 计算机技术与发展,2011,21(7):54 - 57.
HUANG Tao, LIU Shenghui, TAN Yanna. Research of clustering algorithm based on k-means[J]. Computer Technology and Development, 2011,21(7):54 - 57. (in Chinese)

12 BASU B, SRINIVAS V V. Regional flood frequency analysis using kernel-based fuzzy clustering approach[J]. Water Resource Research, 2014,50(4):3295 - 3316.

13 MARTIN Ester, HANS-PETER Kriegel, JERG Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]// Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996:226 - 229.

14 DHAMI C, BNASAL M. An improvement of DBSCAN algorithm to analyze cluster for large datasets[C]// Proceedings of IEEE International Conference on MOOC Innovation and Technology in Education, 2013: 42 - 46.

15 JAEGER P T, SHNEIDERMAN B, FLEISCHMANN K R, et al. Community response grids: E-government, social networks, and effective emergency management[J]. Telecommunications Policy, 2007, 31(10):592 - 604.

16 王爽. 城市网格化管理模式研究[J]. 软件导刊, 2009,8(3):115 - 116.

17 谭丽,曹香淳. 网格化管理与支撑系统的实现及改造[J]. 通信管理与技术, 2012,6(3):12 - 13.

18 万学斌. 关于推进农村网格化管理的思考[J]. 政策, 2014,4(4):62 - 64.

19 李平,卢立. 宜昌市网格化管理建设中网格划分方法及其应用研究[J]. 城市勘测,2011,12(6):31 - 34.

20 林青. 拓展网格化管理与创新社区治理机制研究[J]. 南京理工大学学报,2014,27(5):35 - 41.
LIN Qing. A research on promotion of grid management and innovation mechanism to community governance[J]. Journal of Nanjing University of Science and Technology, 2014,27(5):35 - 41. (in Chinese)

21 杨堂堂. 从数字城市到智慧城市的建设思路与技术方法研究[J]. 地理信息世界, 2013,20(1):63 - 67.
YANG Tangtang. The solution and methods for the construction of smart city from the digital city[J]. Geomatics World, 2013, 20(1):63 - 67 (in Chinese)