ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 4,Jul. 2015,pp. 888－895
DOI:10.16337/j.1004-9037.2015.04.022

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于密度的快速聚类方法

张　晓[1]　张媛媛[2]　高　阳[2]　周新民[3]

(1.伊犁师范学院电子与信息工程学院,伊宁,835000;2.南京大学计算机软件新技术国家重点实验室,南京,210023;3.江苏省公安厅物证鉴定中心,南京,210046)

摘　要:具有噪声的基于密度的聚类方法(Density-based spatial clustering of applications with noise,DBSCAN)在数据规模上的扩展性较差。本文在其基础上提出一种改进算法——具有噪声的基于密度的快速聚类方法(Fast-density-based spatial clustering of applications with noise,F-DBSCAN),对核心对象邻域中的对象只作标记,不再进行扩展检查,通过判断核心对象邻域中是否存在已标记对象来实现簇合并,对边界对象判断其邻域中是否存在核心对象来确认是否为噪声。此方法避免了原始算法中对重叠区域的重复操作,在不需创建空间索引的前提下,其时间复杂度为 $O(n\log n)$。通过实验数据集和真实数据集,验证其聚类效果及算法效率。实验表明 F-DBSCAN 算法不仅保证了有良好的聚类效果及算法效率,并且在数据规模上具有良好的扩展性。

关键词:聚类;密度;F-DBSCAN;算法效率
中图分类号:TP301　　**文献标志码**:A

Fast Density-Based Clustering Approach

Zhang Xiao[1], Zhang Yuanyuan[2], Gao Yang[2], Zhou Xinmin[3]

(1. School of Electronics and Information Engineering, Yili Teachers College, Yining, 835000, China; 2. State Key Laboratory for Software Technology, Nanjing University, Nanjing, 210023, China; 3. Public Security Material Evidence Identification Center of Jiangsu Province, Nanjing, 210046, China)

Abstract: Density-based spatial clustering of applications with noise (DBSCAN) has poor scalability on the data size, especially when the amount of data increases. Here an improved adaptive fast-density-based spatial clustering of applications with noise (F-DBSCAN) algorithm is proposed, with no longer checks of the objects inside the neighborhood of core objects, but just the mark of them. Merging clusters is performed by determining whether there exist the marked objects in the neighborhood of core objects. Noisy objects are recognized by checking whether the neighborhood of border ones contains a core ones. The proposed algorithm can avoid the repeated checking of overlapping area of the original DBSCAN without building the spatial index, thus improving its efficiency substantially with time complexity approaching $O(n\log n)$. The clustering quality of F-DBSCAN is validated on both artificial and real datasets, and its efficiency is also validated on two real datasets from different industries. The empirical results suggest that F-DBSCAN can achieve good clustering quality as well as better efficiency and scalability.

Key words: clustering; density; F-DBSCAN; algorithm efficiency

---

基金项目:江苏省社会发展(BE2010638)资助项目。
收稿日期:2014-01-10;**修订日期**:2014-10-10

引 言

聚类是数据挖掘领域中的重要技术之一。聚类是将数据集中的数据按照某种相似性准则划分成若干簇,使得同一簇中的数据具有较高的相似性,不同簇中的数据尽可能不同。基本的聚类算法可分为5类:基于划分的、基于层次的、基于模型的、基于网格和基于密度的聚类算法。其中,基于密度的聚类算法无需预先指定划分的簇数,能够在含有噪声的数据中发现任意形状的簇。DBSCAN(Density-based spatial clustering of applications with noise)算法是一种经典的基于密度的聚类算法,该算法可以过滤低密度区域,发现稠密样本点,有效处理噪声数据,但在数据规模上的扩展性较差,其时间复杂度为$O(n^2)$。对于那些在速度上要求较高的聚类问题,其在时间上的消耗则是一个瓶颈,如大数据聚类问题、在线聚类问题等。此外,该算法使用两个全局参数 Eps 和 MinPts,对于密度不均匀的数据集,聚类效果较差。DBSCAN 算法从一个对象出发,逐步扩展去找与这个对象所有密度可达对象,从而形成一个簇。由于某些核心对象的邻域有重叠,所以存在重复考查现象。本文提出了基于 DBSCAN 算法的改进算法 F-DBSCAN(Fast-density-based spatial clustering of applications with noise),通过合并重叠区域来实现簇增长,在不改变数据结构、无需创建空间索引的前提下使算法效率得到大幅度提升。

根据数据集的密度分布情况,基于 DBSCAN 算法的研究可分为单密度数据和多密度数据的聚类问题。对单密度数据[1-9]聚类的研究主要集中于聚类效果和聚类效率两方面。对多密度数据[10-13]的聚类一般通过先采用某种方法将数据集根据密度进行划分,然后再利用 DBSCAN 算法或改进算法对各个密度区域进行聚类。近年来有不少学者致力于 DBSCAN 算法效率的的改进。文献[6,8]针对核心对象邻域存在重叠现象,提出了寻找种子对象的新方法。文献[2]还进行了基于并查集等树型结构的 DBSCAN 算法的设计。这些工作都具有一定的参考价值,但只有当使用空间索引 R^*-tree 或 Kd-tree 时,其效率才能达到 $O(n\log n)$,否则为 $O(n^2)$。本文在用于异常检测的早期版本 IDBSCAN[9](Improved density-based spatial clustering of applications with noise)算法的基础上,对其存在的簇合并的问题进行改进,提出 F-DBSCAN 算法。

1 DBSCAN 算法与 F-DBSCAN 算法

1.1 基本概念[14]

(1)Eps-邻域:指给定对象的半径 Eps 以内的区域。

(2)核心对象:如果一个对象的 Eps-邻域包含不少于 MinPts 个对象,则该对象被称为核心对象。

(3)边界对象:如果一个对象的 Eps-邻域包含的对象个数少于 MinPts,则该对象被称为边界对象。

(4)直接密度可达:给定一个数据集 D,如图 1(a)所示,p,q 为数据集 D 中的对象,如果 p 在 q 的 Eps 邻域内,且 q 是一个核心对象,则称 p 是从 q 直接密度可达的。

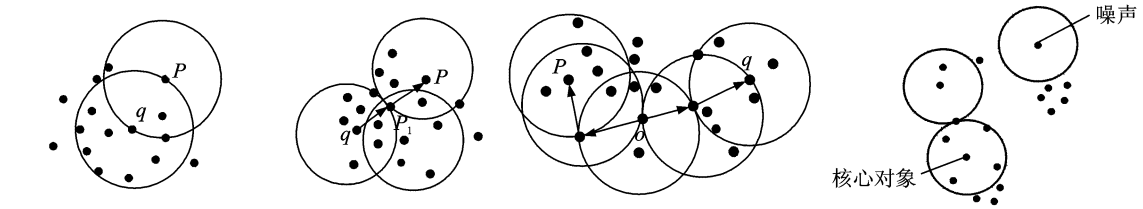
(5)密度可达:如图 1(b)所示,如果存在一个对象序列 $p_1,p_2,\cdots,p_n,p_1=q,p_n=p$,对 $p_i\in D,1\leqslant i\leqslant n,p_{i+1}$ 是从 p_i 关于 Eps 和 MinPts 直接密度可达的,那么 p 是从 q 关于 Eps 和 MinPts 密度可达的。

(6)密度相连:如图 1(c)所示,o 是数据集合 D 中一个对象,如果对象 p 和 q 都是从 o 关于 Eps 和 MinPts 密度可达的,那么对象 p 和 q 是关于 Eps 和 MinPts 密度相连的。

(7)噪声:如图 1(d)所示,不包含在任何簇中的对象被认为是“噪声”。

1.2 DBSCAN 算法[13]

DBSCAN 算法通过检查数据集中每个对象的 Eps 邻域来寻找聚类。如果一个对象 p 的 Eps 邻域包含多于 MinPts 个对象,则创建一个以 p 为核心对象的新簇,然后反复寻找从这些核心对象直接密度可达的对象,这个过程可能涉及一些密度可达簇的合并,当没有新的对象可以被添加到任何簇时,该过



(a) 直接密度可达 (a) Directly density-reachable (b) 密度可达 (b) Density-reachable (c) 密度相连 (c) Density-connected (d) 噪声 (d) Noise

图 1 基本概念

Fig. 1 Basic concepts

程结束。伪代码如下：

DBSCAN 算法

```
REPEAT
    任取数据库中一个未处理过的对象
    IF 抽取的对象为核心对象
        找出从该对象所有密度可达对象，聚为簇
    ELSE
        结束本次循环，抽取一下对象考查
UNTIL 所有对象均被考查过
```

1.3 F-DBSCAN 算法

1.3.1 算法基本思想

DBSCAN 算法的聚类策略是对核心对象的邻域中包含的所有对象进行考查，找出包含的核心对象，以这些为种子继续扩展。算法的聚类过程存在以下问题：(1)如果两个核心对象的邻域有重叠，则这两个邻域的对象就应属于同一个簇，应将其合并。而没有必要重复对其考查。(2)核心对象邻域中的数据最终必将归为某个簇，而边界对象则有可能是噪声，也有可能是属于某个簇。基于此，应进一步考查边界对象的归属问题。基于以上分析，本文的 F-DBSCAN 算法主要对 DBSCAN 算法做如下改进。

(1)对于核心对象，其邻域不再做进一步考查，而是将其归为某个簇。该簇有可能是核心对象所在簇，也有可能是与其他簇合并过的簇。(2)对于边界对象，进一步考查其邻域中是否存在核心对象。如果存在核心对象，则该边界对象归为该核心对象所在簇。反之，则该边界对象为噪声。伪代码如下。

```
REPEAT
  任取数据库中一个未处理过的对象
  IF 抽取的对象为核心对象
    寻找其 Eps 邻域
    IF 该邻域中已有对象归为某个簇 m
        该核心对象及其邻域中所有对象
        都归为簇 m
    ELSE
        创建新簇 n
  ELSE
    IF 被抽取对象邻域中存在核心对象
```

被抽取边界对象归为该核心对象所属簇
或创建新簇
ELSE
被抽取边界对象为噪声数据
UNTIL 所有对象均被考查过

图 2 实例说明具体聚类过程。

(a)对图 2 数据集假设 Eps=1,MinPts=4,若先抽到 a 点,a 是边界对象,则考查其邻域是否存在核心对象。其邻域中的 c 为核心对象,则 a 归为 c 的簇 1,a,b,c,d,e 均为簇 1 中对象,且不再从它们出发循环考查。h,j 也如上处理。

(b)抽到 g 点为核心对象,此时邻域中 e 点已标记为 1 簇,则将 g 及 h,j 合并到 1 簇,而并非产生新簇。即当两个簇有重叠时应合并。

(c)对 f 和 i 点,因其邻域中无核心对象,因此为噪声数据。

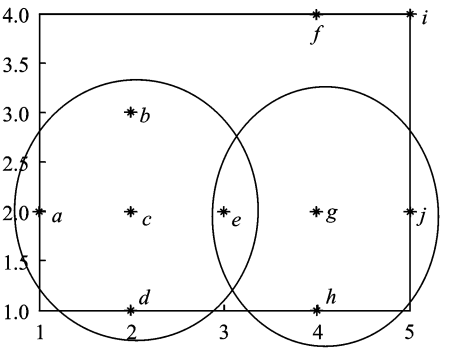


图 2 实例数据分布图

Fig. 2 Distribution of instance data

1.3.2 算法效率分析

从表 1 中看出,本文的 F-DBSCAN 算法时间开销接近 $O(n\log n)$。相比文献[6,8]提出的基于空间索引 kd-tree 或 R*-tree 的聚类算法,本文算法的优势在于无需创建空间索引,即可达到 $O(n\log n)$的效率。

表 1 式$(N * \log_4 N) * d$ 中,d 为常量,主要完成将曲线缩放至 F-DBSCAN 曲线处,实现两曲线增长趋势的对比。表 1 运行时间比较效果图见图 10。

表 1 各算法运行时间

Table 1 Running time of each algorithm s

样本数量	DBSCAN	K-means	F-DBSCAN	$(N * \log_4 N) * d$
5 000	4.288 80	2.233 802	0.260 932	0.260 932
10 000	12.546 0	6.594 758	0.744 271	0.564 335
15 000	26.973 9	13.338 90	1.449 244	0.883 767
20 000	57.758 0	26.489 16	2.13 7203	1.213 610
25 000	91.157 9	36.875 84	2.683 459	1.551 194
30 000	129.739 8	53.190 94	3.602 572	1.894 946
35 000	172.517 3	71.945 75	4.278 538	2.243 824
39 000	221.264 4	90.70 292	5.244 346	2.526 125

2 实验结果与分析

2.1 聚类效果

对聚类效果实验,本文从实验数据集和真实数据集两方面进行验证。

2.1.1 实验数据集聚类效果

为了直观显示,本文实验数据集选择二维数据。两数据集标识效果如图 3 和图 4。

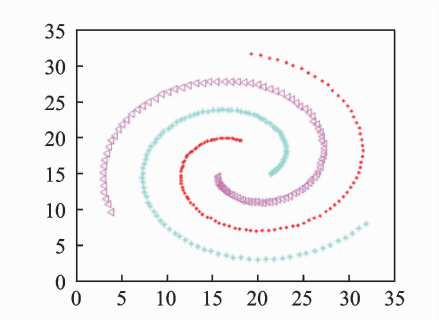


图 3 数据集 1 标识图

Fig. 3 Dataset1 identification

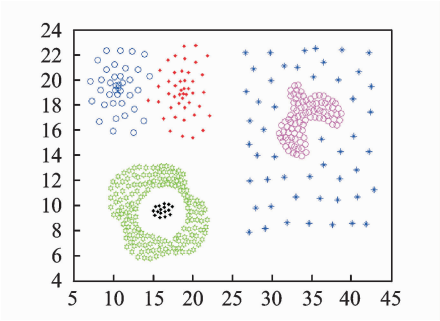


图 4 数据集 2 标识图

Fig. 4 Dataset2 identification

数据集 1(Spiral)共 312 个数据[15],当 $K=3$ 时 K-means 算法聚类效果如图 5(a)所示;当 Eps=1.843 0,MinPts=4 时,DBSCAN 和 F-DBSCAN 算法聚类效果如图 5(b)和(c)所示。

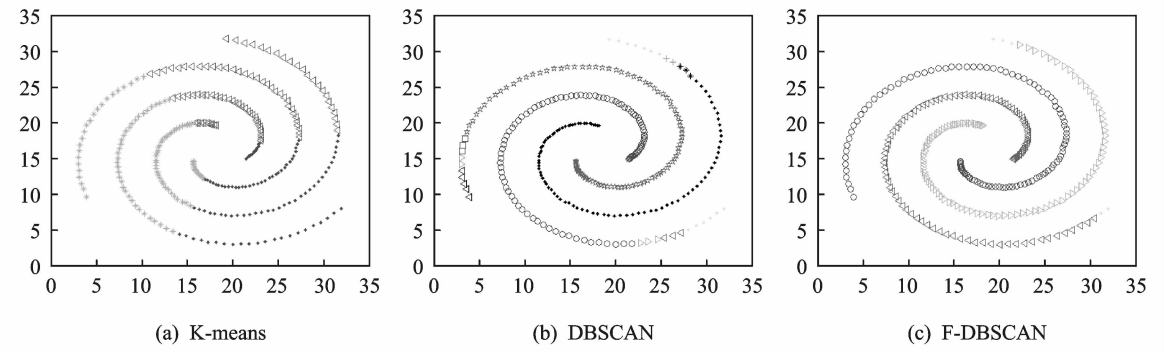


图 5 数据集 1 聚类结果

Fig. 5 The clustering results of Dataset1

数据集 2(Compound)共 399 个数据[16],当 $K=6$ 时 K-means 算法聚类效果如图 6(a)所示;Eps=1.400 0,MinPts=4 时,DBSCAN 和 F-DBSCAN 算法聚类效果如图 6(b)和(c)所示。由表 2 可见,针对两数据集,F-DBSCAN 较 DBSCAN 算法的聚类效果较有提高,而较 K-means 算法则有显著提高。

2.1.2 真实数据集

该数据为某通讯行业数据 5 000 份,为未标识数据,其分布如图 7 所示,图 8 给出 DBSCAN 算法和本文算法聚类结果,其中 Eps=0.004 8。

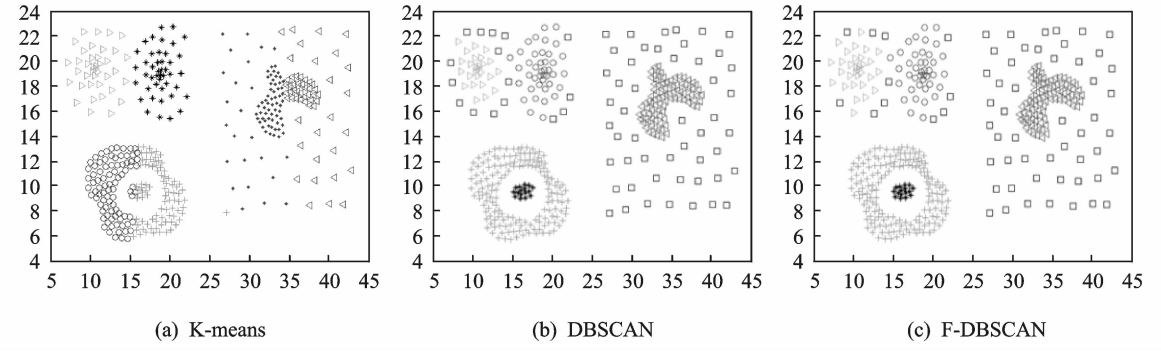


图 6 数据集 2 聚类结果

Fig. 6 The clustering results of Dataset2

表 2 各算法聚类效果比较

Table 2 The comparison of clustering results for each algorithm

数据集	样本数	算法	误分样本个数	误分率
Spiral	312	K-means	211	68%
		DBSCAN	34	11%
		F-DBSCAN	5	1%
Compound	399	K-means	228	57%
		DBSCAN	16	4%
		F-DBSCAN	12	3%

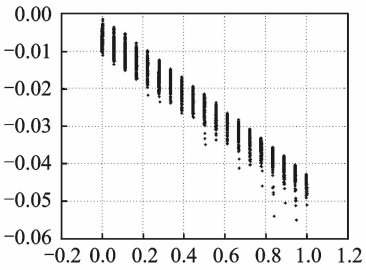


图 7 通讯数据二维分布图

Fig. 7 Two-dimensional distribution of communication data

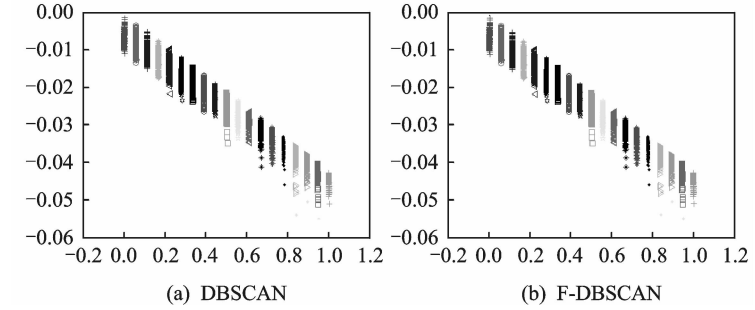


图 8 真空数据集的聚类结果

Fig. 8 Clustering results of real dataset

该数据集两算法作用结果相同,都能达到较好的聚类效果。从以上 3 个数据集聚类结果可以看出,本文算法在聚类效果上略优于 DBSCAN 算法,相对 K-means 算法则有较大提高。

2.2 聚类效率

本文采用两组来自不不同行业的真实数据进行验证,数据集 1 为教育统计数据,是长春某师范院校的评教数据,共 15 482 份。数据集 2 为通讯行业数据,是江苏某通讯公司客户数据,共 39 000 份。图 9 和图 10 分别给出以上 3 个算法的针对两数据集实验结果。

测试环境为 CPU Core i3,内存 2 GB,OS win7;实验平台为 Matlab 2010a。从图 9,10 可以看出,随

着样本数的不断增加,本文算法的运行时间基本没有大幅度的改变,说明本文算法在数据规模上具有良好的扩展性。在图 9,10 中给出了 DBSCAN 算法和 K-means 算法运行时间曲线,它们的时间复杂度分别为 $O(n^2)$和 $O(N)$,本文算法时间曲线与$(N * \log_4 N) * d$ 非常接近(实验中 MinPts=4),进而说明本文算法的时间复杂度接近 $O(n\log n)$。

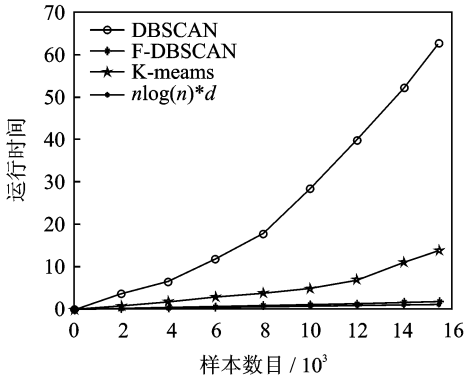


图 9 数据集 1 实验结果

Fig. 9 The experimental results of Dataset1

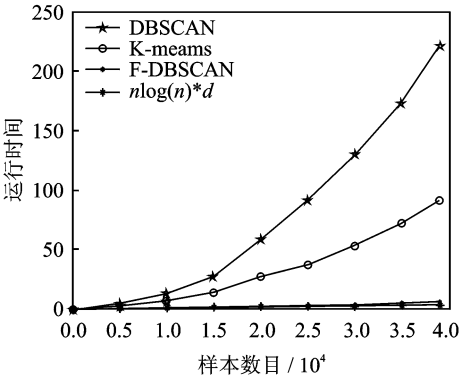


图 10 数据集 2 实验结果

Fig. 10 The experimental results of Dataset2

3 结束语

本文以 DBSCAN 算法为基础提出了一种基于密度的快速聚类算法 F-DBSCAN,分别采用实验数据和真实数据对 F-DBSCAN 算法的聚类效果和性能进行测试。实验证明,相比 DBSCAN 算法,F-DBSCAN 在聚类效果略有改善的前提下,能够显著提高聚类速度。今后的工作重点将集中以下两个方面:尝试将算法应用于在线聚类问题;将数据分区和并行技术与本文算法相结合用于大数据的聚类分析。

参考文献:

[1] Khani F, Hosseini M J, Abin A A, et al. An algorithm for discovering clusters of different densities or shapes in noisy data sets[C]// Processing of the 28th Annual ACM Symposium on Applied Computing. [S. l.]: ACM, 2013:144-149.

[2] Patwary M, Palsetia D, Agrawal A, et al. A new scalable parallel DBSCAN algorithm using the disjoint-set data structure [C]// High Performance Computing, Networking, Storage and Analysis. Salt Lake City, Utah, USA: IEEE, 2012:10-16.

[3] Thanh T N, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters[J]. Chemometrics and Intelligent Laboratory Systems, 2013,120:92-96.

[4] Singh S, Awekar A. Incremental shared nearest neighbor density-based clustering[C]// Proceeding of the 22nd ACM International Conference on Information & Knowledge Management. [S. l.]: ACM, 2013:1533-1536.

[5] 潘玲玲,张育平,徐涛. 核 DBSCAN 算法在民航客户细分中的应用[J]. 计算机工程,2012,38(10):70-73.
Pan Lingling, Zhang Yuping, Xu Tao. Application of kernel DBSCAN algorithm in civil aviation customer segmentation[J]. Computer Engineering, 2012,38(10):70-73.

[6] 许虎寅,王治和. 一种改进的基于密度的聚类算法[J]. 微电子学与计算机,2012,29(2):44-47.
Xu Huyin, Wang Zhihe. An improved clustering algorithm based on density[J]. Journal of Microelectronics and Computer, 2012,29(2):44-47.

[7] 许宏玮. 基于密度与路径的谱聚类算法研究[D]. 广州:广东工业大学,2013.
Xu Hongwei. Spectral clustering algorithm based on density and path research[D]. Guangzhou: Guangdong University of Technology, 2013.

[8] 周水庚,周傲英,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.
Zhou Shuigeng, Zhou Aoying, et al. A fast density-based clustering algorithm[J]. Computer Research and Development, 2000,37(11):1287-1292.

[9] 张晓. 教师评价中基于聚类算法的异常点分析的研究[D]. 长春:东北师范大学,2009.

Zhang Xiao. The research of the clustering algorithm abnormal point analysison the teachers evaluation[D]. Changchun: Northest Normal University, 2009.

[10] 赵文,夏桂书. 一种改进的 DBSCAN 算法[J]. 四川大学学报,2013,36(2):312-316.
Zhao Wen, Xia Guishu. An improved DBSCAN algorithm[J]. Journal of Sichuan University, 2013,36(2):312-316.

[11] 陈刚,刘秉权,吴岩. 一种基于高斯分布的自适应算法[J]. 微电子学与计算机,2013,30(3):27-30.
Chen Gang, Liu Bingquan, Wu Yan. A adaptive DBSCAN algorithm based on Gauss distribution[J]. Journal of Microelectronics and Computer, 2013,30(3):27-30.

[12] 夏鲁宁,荆继武. SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院研究生院学报,2009,26(4):530-537.
Xia Luning, Jing Jiwu. SA-DBSCAN: An adaptive algorithm based on density clustering[J]. Journal of the Graduate School of the Chinese Academy of Sciences, 2009,26(4):530-537.

[13] 钱美旋,叶东毅. 利用一维投影分析的无参数多密度聚类算法[J]. 小型微型计算机系统,2013,34(8):1866-1871.
Qian Meixuan, Ye Dongyi. Parameter free multi-density clustering using one-dimensional projection analysis[J]. Journal of Chinese Computer Systems, 2013,34(8):1866-1871.

[14] Han J, Kanber M. Data mining: Concepts and techniques[M]. 北京:机械工业出版社,2000.

[15] Chang H, Yeung D Y. Robust path-based spectral clustering[J]. Pattern Recognition, 2008,41(1):191-203.

[16] Zahn C T. Graph-theoretical methods for detecting and describing gestalt clusters[J]. Computers, IEEE Transactions on, 1971,100(1):68-86.

作者简介:

张晓(1967-),女,高级实验师,研究方向:数据挖掘、计算机教育,E-mail:Zhx0125@163.com。

张媛媛(1991-),女,硕士研究生,研究方向:机器视觉,机器学习等,E-mail:zhangyuanyuan2013nju@gmail.com。

高阳(1972-),男,教授,博士生导师,研究方向:强化学习、智能 agent、智能应用等,E-mail:gaoy@nju.edu.cn。

周新民(1961-),男,教授级高级工程师,研究方向:图像处理与模式识别,E-mail:jszhouxinmin@sina.com。