

一种基于高斯分布的自适应 DBSCAN 算法

陈 刚¹, 刘秉权², 吴 岩³

(1 广东科技学院 计算机系, 广东 东莞 523083; 2 哈尔滨工业大学 计算机科学与技术学院,
黑龙江 哈尔滨 150001; 3 哈尔滨工业大学 软件学院, 黑龙江 哈尔滨 150001)

摘 要: 针对传统的 DBSCAN 算法只能依靠经验来设置阈值($\minPts, Eps$)和无法对多密度数据集进行有效聚类的不足, 提出了一种可适用于所有密度分布特征的数据集的基于高斯分布的自适应 DBSCAN 算法. 该算法根据数据集的特点寻找最大的聚类效果指数(CEI)来确定 $\minPts$, 通过 Dist_k 图的层次数确定 Eps 个数并通过高斯分布中的参数估计来确定每个密度层次 Eps 大小, 最后用所求得阈值进行聚类. 将该算法与传统的 DBSCAN 算法分别应用于单密度数据集和多密度数据集, 结果显示该算法更有效.

关键词: 数据集; 高斯分布; 密度层次; 聚类效果; 噪音率

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2013)03-0027-04

A Adaptive DBSCAN Algorithm Based on Gauss Distribution

CHEN Gang, LIU Bing-quan, WU Yan

(1 Department of Computer, Guandong Univesity of Science & Technology, Dongguan 523083, China;

2 School of Computer Scicence and Technology, Harbin Institute of Technology, Harbin 150001, China;

3 School of Sofeware, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Traditional DBSCAN algorithm usually set the threshold ($\minPts, Eps$) depending on empirical value, and merely fit for single-density data sets, in allusion to the shortage, the article puts forward a new adaptive DBSCAN algorithm which based on Gaussian distribution and can be used to multi-density data sets. The new algorithm can generate proper threshold according to the data set characteristics. Firstly, to get the $\minPts$ according to the max CEI value. Then, to confirm the number of Eps according to the curve density level, so that to get Eps data by the Gaussian distribution law. At last, to make clustering for the data set with the $\minPts$ and Eps value. Also, to apply the new algorithm and the traditional DBSCAN algorithm to single-density and multi-density data sets, the results show that the new algorithm is more efficient.

Key words: data sets; gaussian distribution; density level; clustering effect; noise ratio

1 引言

基于密度的聚类算法(Density-Based Spatial-Clustering of Application with Noise, DBSCAN)是数据挖掘聚类算法中的一个重要分支. DBSCAN 就能够发现任意形状的簇, 但前提是要提供合适的参数, 并且仅仅用于均匀密度的数据集, 对于密度变化较大的数据集时, 例如当此算法用于气象云图^[1], 地

震带等密度变化较大的簇时^[2], 聚类效果不明显.

针对 DBSCAN 算法的现存缺点, 本文提出了一种基于高斯分布的自适应 DBSCAN 算法(Adaptive DBSCAN Algorithm Based on Gauss Distribution, AGD-DBSCAN). 该算法能根据聚类数据的自身特点构建函数, 通过对函数进行有效的分析, 求出不同密度的簇的阈值参数, 然后根据不同的阈值生成合适的簇, 实验结果显示算法在处理变密度聚类时

非常有效.

2 DBSCAN 及已有的改进工作

2.1 DBSCAN 相关概念

DBSCAN 算法是基于中心的密度聚类方法, DBSCAN 算法的相关概念如下:

定义 1 Eps 邻域: 给定某对象 p , p 的邻域 $N_{Eps}(p)$ 定义为以 p 为核心, 以 Eps 为半径的 d 维超球体的区域, 公式表示为:

$$N_{Eps}(p) = \{q \in D \mid \text{dist}(p, q) \leq Eps\} \quad (1)$$

其中 $D \subseteq R^d$, d 为空间 R 的维度. $\text{dist}(p, q)$ 表示对象 p 和 q 之间的直线距离.

定义 2 核心点、边界点、噪音点: 对于数据对象 p , 且 $p \in D$. 如果以 p 为中心, 以 Eps 为半径, 若 $N_{Eps}(p)$ 内的点数超过给定 minPts , 则称 p 为核心点, 若 p 不是核心点, 但在某核心点区域内, 则称 p 为边界点, 其余为噪音点或离群点.

定义 3 直接密度可达: 给定一个对象 $D \in R^d$, 若 p 在 q 的邻域内, 若 q 为核心点, 则称对象 p 出发到对象 q 是直接密度可达.

定义 4 密度可达: 在给定数据集 D 中, 若存在一个对象链 $p_1, p_2, \dots, p_n$, 其中, $p_1 = q, p_n = p$, 对于 $p_i \in D (1 \leq i \leq n)$, 若在 (minPts, Eps) 条件下, p_{i+1} 从 p_i 直接密度可达, 则称对象 p 从对象 q 密度可达.

定义 5 密度相连: 若存在一个对象 $o \in D$, 使得对象 p 和 q 是从 o 在 (minPts, Eps) 条件下密度可达, 则称对象 p 到 q 密度相连. 且密度相连是对称的.

定义 6 簇: 从任何一个核心点开始, 从该对象点密度可达的所有对象点构成一个簇.

2.2 DBSCAN 算法现有改进工作

在 DBSCAN 算法中, 设置阈值参数的方法是: 通常设定 $\text{minPts}=4$, 通过观察来判断设置 Eps . 对于数据集 D 中的对象 p , 计算出所有与 p 最近的第 k 个对象的距离 $\text{dist}_k(p)$. p 遍历 D 中所有对象, 并计算对象之间的距离, 得到集合 $\text{Dist}_k = \{\text{dist}_k(p)\}$, 将 Dist_k 绘制成排序图, 若数据集的簇密度分布较均匀, 则 Dist_k 图大致如图 1 所示. 从图中可以看出, 所有曲线在某点突然变陡, 传统算法认为那个突然变陡点的对应的距离为所求阈值 Eps . 因此该算法仅仅只能对密度均匀的数据集有效.

目前很多学者已针对不足提出了大量改进建

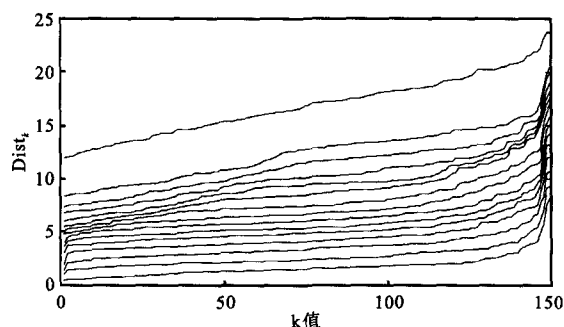


图 1 单密度 Dist_k 图

议, 文献[3-7]针对问题的某些方面提出了各种改进方法, 但目前改进工作仍然无法真正实现数据集自适应确定阈值(minPts 和 Eps).

3 AGD-DBSCAN 算法思想及实现

AGD-DBSCAN 算法中密度分层思想: 用数据集所有点到第 k 个最近点的距离建立距离分布矩阵 $\text{DIST}_{n \times n}$, 绘制 Dist_k 图, 选取其中一条曲线, 如图 2 所示, 距离相同或者相近的点在 Dist_k 图中表现为一条近似平缓曲线, 如线段 a, c, e, 可认为 a 线段上所有点密度为同一层次密度, c 和 e 同之. 再看图 2 中的线段 b, d, f, 曲线不平缓, 对应各线段的点为噪音点或离群点.

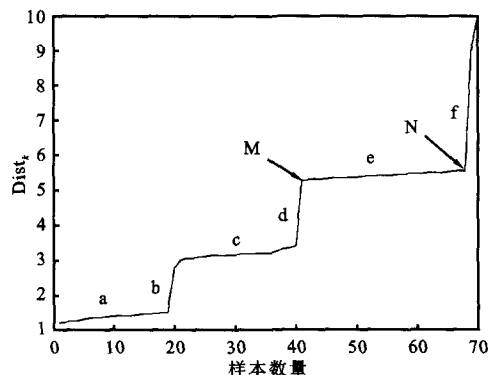


图 2 多密度 Dist_k 图

定义 7 $\text{DIST}_{n \times n}$: 计算所有点到其第 k 个最近点的距离, 并建立距离分布矩阵 $\text{DIST}_{n \times n}$:

$$\text{DIST}_{n \times n} = \{\text{dist}(i, j) \mid 1 \leq i \leq n, 1 \leq j \leq n\} \quad (2)$$

其中 n 表示数据集对象数目, $\text{DIST}_{n \times n}$ 为实对称矩阵, 将每列进行升序排列, 排序后进行转置, 得到矩阵 KNN, KNN 第 k 列表示每个点到其第 k 个最近点距离的集合.

定义 8 增长距离与曲线增长距离之和: 在 Dist_k 图中, 点的距离为它们的 Dist_k 值, 若相邻两点坐标为 (i, y) 和 $(i+1, b)$, 则 $(i+1, b)$ 点的增长距离

和曲线增长距离之和分别为:

$$S = b - y$$

$$MS_j = \sum_{i=2}^n S_i$$

其中 S_i 代表每条曲线第 i 个点的增长距离, MS_j 代表第 j 条曲线的增长距离之和. n 代表数据集点的数量.

定义 9 跳跃间断点:在 Dist_k 图中,若在曲线 a 上,某第 q 点相对第 $q-1$ 个点,其 Dist_k 值急速增大,且在第 q 点之后的点的 Dist_k 值仍然急速增大或者到急速增大一段时间后基本不变,则此点为此密度层次的跳跃间断点. Safter_{km} 表示第 k 密度层次跳跃间断点和之后的所有的点到 m 点(下一个密度层次开始的第 1 个点,若 k 为最后一个密度层次,则 m 为数据集最后一点)之间的增长距离之和.

$$\text{Safter}_{km} = \sum_{i=k}^m S_i \quad (3)$$

定义 10 聚类效果指数(Clustering Effect Index, CEI):表示每条曲线所有的 Safter_k 与 MS_j 的比值,如公式(4)所示:

$$\text{CEI} = \frac{\text{Safter}_{km} + \text{Safer}_{(k-1)n} + \dots + \text{Safer}_{1q}}{MS_j} \quad (4)$$

设数据集共有 k 密度层次,其中 Safter_{km} 表示第 k 密度层次跳跃间断点($m1$)之后的到 m 点(结束点)之间所有点的增长距离之和; $\text{Safer}_{(k-1)n}$ 表示 $k-1$ 密度层次跳跃间断点($n1$)之后的到 n 点(k 密度层第一个点)之间所有点的增长距离之和;同理 Safer_{1q} 表示达到第 1 密度层次跳跃间断点($q1$)之后到 q 点(第 2 密度层第一个点)之间所有点的增长距离之和.

定义 11 噪音率 P :基于定义 10,噪音率可表示为:

$$P = \frac{(m-m1) + (n-n1) + \dots + (q-q1)}{n} \times 100\% \quad (5)$$

AGD-DBSCAN 算法的聚类思想:通过 Dist_k 图根据公式(4)生成所有 k 值的 CEI 曲线,找出最大 CEI 值对应的 k ,令此 k 为 minPts ,根据公式(5)计算此时的噪音率 P ,接着根据 Dist_k 曲线是否分层的特征确定此数据集密度层次,然后用 $1-P$ 作为置信度,依据高斯分布的参数估计确定公式(7)中的 y ,并用公式(7)计算 Eps 值,最后根据求得的阈值参数进行 DBSCAN 聚类.

AGD-DBSCAN 算法分为 2 大步骤.具体如

下:

步骤 1:确定 minPts 和 Eps 值.

① minPts 的确定

计算所有点到第 k 个最近点的距离,用其距离建立 KNN 矩阵,用 MATLAB 将 KNN 矩阵所有列进行拟合,得到如图 1 所示的 Dist_k 图.从所有的曲线中找出曲线聚类效果指数 CEI 最大的曲线,该曲线对应的 k 即本文所求的 minPts ,用此 minPts 进行聚类,可合理地将核心点与噪音点区分;当 CEI 值达到最大,即为 1 时,此时聚类效果最好.图 2,即 a, c, e 均接近水平的,也就是说相同密度层次的点到第 k 个最近点的距离基本相同.

② Eps 的确定

当确定 minPts 后.若对应的数据集的 Dist_k 图与图 1 相似,则称该数据集为单密度数据集,仅需求一个 Eps ;若对应的 Dist_k 图如图 2 所示,则称该数据集为多密度数据集,需分别求多个 Eps .但是,不管是单密度还是多密度数据集,都要计算每个密度层次上 Dist_k 值的平均值 $\bar{x} = \frac{1}{k}(x_i + \dots + x_j)$,其中 $k = j - i + 1$, i 表示该层第 1 个点, j 表示该层最后一点,分别为图 2 曲线的 M 点和 N 点.还要计算每个密度层次的均方差 σ_n :

$$\sigma_n = \sqrt{\frac{1}{k} \sum_{m=1}^k (x_m - \bar{x}_n)^2} \quad (6)$$

其中 n 表示密度的层数, k 表示某个层次密度的点数量, m 表示在此密度层中点序号.根据高斯分布性质,可以确定:

$$Eps_n = \bar{x}_n + y \sigma_n \quad (7)$$

其中 y 是一个与 P 有关的变量,确定 Eps 值的条件是,要保证所有的核心点都要在 Eps 内,若 Eps 选择过大,会导致噪音点错误地标记为核心点;若 Eps 选择过小,会导致核心点错误地标记为噪音点或同一簇误分裂为多簇,所以需要选择合适的 Eps .

本文从统计学角度分析,由于数据集所有点均为独立同分布事件,且有相同的均值和方差,所以这些点的概率分布符合高斯分布,即在这个层次的点的概率分布符合以 $\bar{x}_n$ 为均值以 σ_n 为标准差的正态分布,公式如下:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(x-\bar{x}_n)^2}{2\sigma_n^2}} \quad (8)$$

均数 $\bar{x}_n$ 决定正态曲线的中心位置;标准差 σ_n 决定正态曲线的陡峭或扁平程度. σ_n 越小,曲线越陡峭; σ_n 越大,曲线越扁平.即 σ_n 越小,说明样本越集

中,越靠近均值点,越大说明样本越分散,偏离均值点越大.根据高斯分布参数估计可知,点到第 k 个点的距离的值处于区间 $(\bar{x}_n - y\sigma_n, \bar{x}_n + y\sigma_n)$ 内的概率为 $1 - P$,如公式(9)

$$p\{|x_j - \bar{x}_i| < y\sigma_i\} = 1 - P \quad (9)$$

若以 $Eps_n = \bar{x}_n + y\sigma_n$ 为半径, $\minPts = k$ 进行聚类,则保证所有的点都要在 Eps 内的概率为 $1 - P$,即保证数据集中有 $1 - P$ 的点在 Eps_n 内,理论上可保证所有的核心点均在 Eps_n 内.

步骤 2:进行 DBSCAN 聚类.

用自适应 DBSCAN 算法进行聚类时,若仅有一个 Eps 就直接聚类;若有多个 Eps 时,则应先聚较高密度簇,再聚次低密度簇,直到将所有的簇聚类完成,即按从小到大的顺序选择 Eps_i ,依次使用阈值 $(\minPts, Eps_i)$ 来调用 DBSCAN 算法,将成功聚类结果标记为 C_{i-t} ,可称为密度层次为 D_i 的第 t 个簇.并且被聚类过的点不会重复参与聚类.当所有阈值 $(\minPts, Eps_i)$ 均被使用之后,没有被聚类的点,标注为离群点或者噪音点.

4 实验

4.1 实验过程

本实验采用 MATLAB7.0 仿真软件.为了说明本算法的有效性,本实验用两个数据集进行说明,实验所用的数据集 DS1、DS2 是具有不同密度特点的数据集.需要指出的是,AGD-DBSCAN 算法对数据集的数据分布特点和数据集密度是否均匀没有要求.

DS1 是 150 个点单密度数据集,DS2 为 205 个点多密度数据集.由于已有证明 DBSCAN 算法适用

于任何形状的聚类,所以本文在形状方面不再分析.通过 AGD-DBSCAN 算法得到 DS1、DS2 的 $\minPts$ 分别为 4 和 2,因为此时的 CEI 值最大,分别为 0.7451 和 1;DS1 与 DS2 对应的 P 分别为 5.33% 和 2.44%;DS1 的 Eps 为 1.885,DS2 的 Eps 分别为 0.1 和 0.5.

4.2 实验评价

在 DBSCAN 算法中,采用 R* 树伸展算法进行搜索,时间复杂度为 $O(n \log n)$,搜索时间主要消耗在对象间的距离计算和第 k 个最近点的距离计算,在 AGD-DBSCAN 算法中,由于 DIST 及 KNN 矩阵的中已经包含了这些数据,因此在聚类过程中不必重复计算.相对 DBSCAN 算法,AGD-DBSCAN 算法时间主要是用于确定 $\minPts$ 和 Eps 值.

为了证明 ADG-DBSCAN 算法的准确性与优越性,将数据集 DS1 和 DS2 同时都采用 DBSCAN 和 AGD-DBSCAN 算法进行聚类.DBSCAN 算法是采用传统的方法确定阈值,而 AGD-DBSCAN 算法采用本文所提到的步骤 1 进行确定阈值,所有的实验都是在 AMD Athlon64X2 双核+1GM DDR+windows XP 平台上,应用 MATLAB 7.0 计算并测得.

表 1 中分别给出 ADG-DBSCAN 与传统 DBSCAN 算法的相关指标.从表可知,在时间性能方面 AGD-DBSCAN 比 DBSCAN 要慢,原因是需要时间用于确定 $\minPts$ 和 Eps 上,但是这方面的时间差距不存在数量级别;表中还显示,直接给定阈值的 DBSCAN 聚类的准确度不如 AGD-DBSCAN 高.因此,这种自适应的计算阈值的 ADG-DBSCAN 算法是可行的.

表 1 算法运算时间与聚类准确度对比

数据集	样本数量	密度层次	聚类算法	$\minPts$	Eps	运行时间	准确率
DS1	150	1	AGD-DBSCAN	4	1.885	639	88.7%
			DBSCAN	4	2.016	424	60.8%
DS2	205	2	AGD-DBSCAN	3	0.1, 0.5	1527	92.07%
			DBSCAN	4	0.323	895	46.58%

5 结束语

本文提出了基于高斯分布的自适应 DBSCAN 算法 ADG-DBSCAN,该算法可应用于密度特征的数据集.实验结果表明,在时间性能相差不大的情况下,ADG-DBSCAN 算法比传统的 DBSCAN 算法对所有密度特征的数据集聚类效果更优.今后,将此算法应用于单发地震带,连发地震带划分或强雷暴

区,雷暴区的划分,进一步探讨此算法的应用效果和应用前景.

参考文献:

- [1] 侯荣涛,朱斌.基于 DBSCAN 聚类算法的闪电路径预报模型[J].计算机应用,2012,32(3):847-851.
- [2] 赵康,陆介平.一种基于密度的文本聚类挖掘算法[J].计算机应用研究,2009,26(1):124-126.

(下转第 34 页)

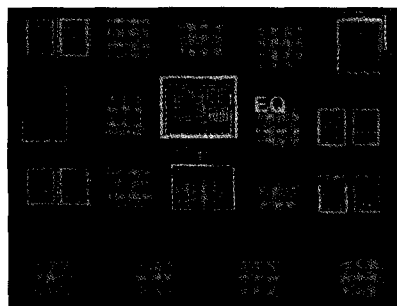


图 7 芯片版图

6 结束语

文中使用 $0.13\mu\text{m}$ CMOS 实现了一款可以补偿高频信号传输线损耗的新型均衡器结构. 仿真结果验证得到, 该电路可以接收速率高达 5Gbit/s 数据信号, 并对传输线损耗造成的高频衰减提供了有效的补偿. 在设计中, 使用了对称负载和有源电感等技术, 避免了使用片上电感. 对电路的噪声抑制和均衡效果起到了明显的促进. 电路仿真使用 Hspice, 功耗仿真结果满足设计需求.

参考文献:

- [1] Yoshiharu Kudoh, Muneo Fukaishi, Masayuki Mizuno. A $0.13\mu\text{m}$ CMOS 5-Gb/s 10-m 28AWG cable transceiver with no - feedback - loop continuous - time post - equalizer[J]. IEEE Journal of Solid - State Circuit, 2003, 38(5): 741-746.
- [2] Jin Liu, Xiaofeng Lin. Equalization in high - speed

communication systems[J]. IEEE Circuits And Systems Magazine, 2004, 4(2): 4-17.

- [3] Jri Lee. A 20-Gb/s a adaptive equalizer in $0.13\text{-}\mu\text{m}$ CMOS technology[J]. IEEE Journal of Solid - State Circuit, 2006, 41(9): 2058-2066.
- [4] Yasumoto Tomita, Masaya Kibune, Junji Ogawa, et al. A 10-Gb/s receiver with series equalizer and on-chip ISI monitor in $0.11\text{-}\mu\text{m}$ CMOS[J]. IEEE Journal of Solid - State Circuit, 2005, 40(4): 986-993.
- [5] Jong - Sang Choi, Moon - Sang Hwang, Deog - Kyoon Jeong. A $0.18\text{-}\mu\text{m}$ CMOS 3.5-Gb/s continuous-time adaptive cable equalizer using enhanced low - frequency gain control method[J]. IEEE Journal of Solid - State Circuit, 2004, 39(3): 419-425.
- [6] John G Maneatis, Mark A Horowitz. Precise delay generation using coupled oscillators[J]. IEEE Journal of Solid - State Circuit, 1993, 28(12): 1273-1282.
- [7] Zhenyu Zhao, Jianjun Wang, Shaoqing Li, et al. A 2.5-Gb/s $0.13\mu\text{m}$ CMOS current mode logic transceiver with preemphasis and equalization[C]// ASIC, 2007. ASICON '07. 7th International Conference. China: Guilin, 2007.

作者简介:

周振宇 男, (1987-), 硕士. 研究方向为模拟数字混合信号集成电路.

王忆文 男, (1966-), 教授. 研究方向为大规模集成电路与系统.

(上接第 30 页)

- [3] FengP J, Ge L D. Adaptive DBSCAN - based algorithm for constellation reconstruction and modulation identification[C]// Proceedings of Radio Science Conference 2004. Beijing: Pub House of Electronics Industry, 2004: 177-180.
- [4] Yue S H, Li P, Guo J D, et al. A statistical information-based clustering approach in distance space[J]. Journal of Zhejiang University Science, 2005, 6A(1): 71-78.
- [5] 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.
- [6] 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报, 2009, 26(4):

530-537.

- [7] 谭颖, 胡瑞飞, 殷国富. 多密度阈值的 DBSCAN 改进算法[J]. 计算机应用, 2008, 28(3): 745-748.

作者简介:

陈刚 (1977-), 硕士, 讲师. 研究方向为数据挖掘、人工智能.

刘秉权 (1970-), 博士, 副教授. 研究方向为自然语言处理、人工智能.

吴岩女, (1965-), 博士, 副教授. 研究方向为自然语言处理、人工智能.