计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):118−123 [doi: 10.15888/j.cnki.csa.006077]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于类中心与边界自寻优的聚类算法①

张文军[1], 王建平[2], 范世平[2], 张柳霞[1]

[1](中华女子学院 计算机应用技术研究所, 北京 100101)
[2](天地科技建井研究院, 北京 100013)

摘　要: 随着互联网应用的普及和深入, 涌现了许多新的应用场景和数据类型, 导致许多经典的聚类算法不能有效地适应新的发展形势, 成为数据挖掘中的棘手问题和研究热点, 为此提出一种新颖的基于类中心与边界自寻优的数据聚类算法. 该算法引入数据点“距离半径”分布矩阵 R 及其“距离半径累计”分布矩阵 Σ*R* 概念表征数据聚合度, 并依据广度优先原则自寻优 *R* 与 Σ*R* 中皆为最小的数据点作为类中心; 同时, 提出“距离半径偏导”分布矩阵 *R*’, 描述簇类之间的松散度, 并采用广度优先原则自寻优矩阵 *R*’中的突变跃迁增长点, 作为簇类之间的分界. 通过经典的 Aggregation 聚类数据集的仿真实验测试, 表明该算法能够有效地对多种形状、大小和不同密度分布的数据集进行聚类分析, 能较好地识别出孤立点和噪声, 具有较高的鲁棒性和分析精度.

关键词: 聚类; 类中心自寻优; 类边界自寻优; 广度优先; 数据挖掘

引用格式: 张文军,王建平,范世平,张柳霞.基于类中心与边界自寻优的聚类算法.计算机系统应用,2017,26(11):118–123. http://www.c-s-a.org.cn/1003-3254/6077.html

Clustering Algorithm Based on Self-Optimizing Center and Boundary of Classes

ZHANG Wen-Jun[1], WANG Jian-Ping[2], FAN Shi-Ping[2], ZHANG Liu-Xia[1]

[1](Research Institute of Applied Computer Technology, China Women University, Beijing 100101, China)
[2](Mine Construction Institute of Tian-Di Science and Technology, Beijing 100013, China)

Abstract: With the deep development and popularization of Internet, new data types emerge in new application fields so that many classic clustering algorithms are no longer effectively adapted to new situations, so data mining is becoming thorny issues and research focus. Therefore the article proposes a novel clustering algorithm based on self-optimizing the centers and boundaries of classes. The algorithm contains the points’ distance-radius-distribution matrix-R and the cumulative radius-distribution matrix-Σ*R* characterizing the degree of data aggregation. The data points with the minimum *R* and Σ*R* as the class centers are searched under the breadth-first. The algorithm also includes the partial derivative matrix-*R*’ of the distance-radius distribution to describe the gradient change of the loose degree between different points. According to self-optimizing and breadth-first, the transition point of matrix-*R*’, which its partial derivative is the biggest one in adjacent points, is found as the class boundary, inside which all points belong to the class. After emulating and testing the algorithm by typical clustering data sets of Aggregation, the result shows that the algorithm can effectively cluster the data sets with different shapes, sizes and different densities, identify the isolated points and noises, and also have better robustness and accuracy.

Key words: clustering; self-optimizing class center; self-optimizing class border; breadth-first; data mining

随着云计算、物联网、移动互联以及社交媒体的发展和普及, 产生了许多新的数据类型和应用场景[1-3]. 聚类分析广泛应用于模式识别、信息检索、机器学习、图像处理、生物种群划分等研究领域, 已成为数

① 基金项目: 国家科技支撑计划资助项目 (2012BAB13B00); 中华女子学院科研基金重点资助项目 (KG2014-02002)
收稿时间: 2017-02-23; 修改时间: 2017-03-23; 采用时间: 2017-03-29

据挖掘领域重要的研究课题. 如何对大量涌现的新型数据进行快速地处理并挖掘出有用的信息、知识以及预测未来, 已成为极具挑战性的迫切任务.

聚类分析是数据挖掘的重要任务之一. 其目的是, 针对数据对象集, 依据数据之间联系的紧密度和相似度进行分类, 同一类中的数据对象之间具有较高的相似度, 不同类中的数据对象差异较大[4]. 针对不同的数据类型和应用场景的聚类分析, 已提出了许多经典的聚类算法, 例如, 基于划分的聚类[5,6]、基于层次的聚类[7] 、基于密度的聚类[8,9]、模糊聚类[10,11]、基于网格的聚类[12,13]等, 也有许多算法集成了两种或两种以上的聚类方法[14,15], 以获得更好的性能. 然而不幸的是, 这些经典的聚类算法已不能适应新型数据的聚类分析[16-18].

因此, 本文提出了一种新颖的基于类中心与边界自寻优的数据聚类分析算法 (SOCBC, Self-Optimizing Center and Boundary of Class). 该算法引入数据点“距离半径”分布矩阵 R 和“半径累计”分布矩阵 ΣR 概念表征数据聚合度, 并依据广度优先原则自寻优 R 与 ΣR 皆为最小的数据点作为类中心; 同时引入“距离半径偏导”分布矩阵 R', 描述簇群之间的松散度, 采用广度优先原则自寻优矩阵 R'中的突变跃迁点, 作为簇类之间的分界. 仿真结果表明, 该算法能够有效地对多种形状、大小和不同密度分布的数据进行聚类分析, 并能较好地识别出孤立点和噪声. 为数据的聚类分析, 提供了一种新的有效的聚类算法.

1 相关定义

定义 1. 设 m 维数据集 X, 其属性 $(a_1, a_2, \ldots, a_m)$, 则构成 m 维数据空间. 给定 m 维数据样本 n 个, 构成数据集 $X=\{x_1, x_2, \ldots, x_n\}$, 则每个数据样本对应一个 m 维数据空间中的数据点.

定义 2. 给 n 个 m 维数据点分别编号 (i=1, 2, 3, …, n), 计算各点之间的欧氏距离半径$\{R_{ij}(i, j)|i, j=1, 2, \ldots, n; i\neq j\}$.

$$R_{ij}=\sqrt{\sum_{k=1}^{m}\left(a_{ik}-a_{jk}\right)^2} \tag{1}$$

定义 3. 从第一个数据点开始, 将每一个点 i 与其他 n-1 个点之间的距离值按递增排序, 组成 n-1 维距离半径行向量 R_i 及其各距离值对应的点编号向量 P_i, 共计 n 个行向量$\{R_i|i=1, 2, 3, \ldots, n\}$及其对应的点编号向量$\{P_i|i=1, 2, 3, \ldots, n\}$, 分别构成 $n(n$-1) 距离半径矩阵 $R=\{R_{ij}\}n(n$-1) 及其对应的点编号矩阵 $P=\{P_{ij}\}n(n$-1); 也即, 矩阵第 i 行向量表示: 第 i 个数据点与其他各点的距离半径的递增排列及其对应点的编号.

$$R=\begin{pmatrix}R_1\\R_2\\\vdots\\R_n\end{pmatrix}=\begin{pmatrix}R_{1,1}&R_{1,2}&\cdots&R_{1,n-1}\\R_{2,1}&R_{2,2}&\cdots&R_{2,n-1}\\\vdots&\vdots&\vdots&\vdots\\R_{n,1}&R_{n,2}&\cdots&R_{n,n-1}\end{pmatrix} \tag{2}$$

$$P=\begin{pmatrix}P_1\\P_2\\\vdots\\P_n\end{pmatrix}=\begin{pmatrix}p_{1,1}&p_{1,2}&\cdots&p_{1,n-1}\\p_{2,1}&p_{2,2}&\cdots&p_{2,n-1}\\\vdots&\vdots&\vdots&\vdots\\p_{n,1}&p_{n,2}&\cdots&p_{n,n-1}\end{pmatrix} \tag{3}$$

定义 4. 针对距离半径矩阵 R 的各距离元素 R_{ij}(其中 $j\neq$1, n-1), 计算其一阶行偏导数, 如式 (4), 表征第 i 点的邻域中与距其第 j 近的点的距离的变化率, 并组成一阶行偏导矩阵, 如式 (5), 用于判定数据中心点的类边界. 也即, 当点 i 为某一聚类的中心点时, 在其对应的第 i 行向量中随着列数 j 增加 (与中心点的距离渐远) 会出现距离半径剧增的跳变点: 一阶行偏导数 R'出现剧增的跳变, 则该点为聚类边界点.

$$\frac{\partial R_{ij}}{\partial i}=\frac{R_{ij+1}-R_{ij-1}}{2} \tag{4}$$

$$R'=\frac{\partial R}{\partial I}=\left(\frac{\partial R_{ij}}{\partial i}\right)_{n(n-2)} \tag{5}$$

定义 5. 针对距离半径矩阵 R 的各距离元素 R_{ij}, 计算行向量元素的左累计, 如式 (6), 并构成距离半径累计矩阵 ΣR 或 S, 用于表征各数据点周围数据点分布的整体疏密度. 那么将类中心点与其周围的各点相比, 则依据广度优先, S 矩阵中其对应的行向量各元素数值 S_{ij} 为最小, 作为寻优、确定类中心点的判据之一. 为使形象与表述方便兼顾, 特此声明: S 与 ΣR 为同一概念, 不同形式.

$$S_{ij}=\sum_{k=1}^{j}R_{ik} \tag{6}$$

$$\sum R=S=\left(S_{ij}\right)_{n(n-1)} \tag{7}$$

定义 6. 在距离半径矩阵 R 的第 j 列 (j 可以取: 数据集中数据个数 n*10% 的整数值) 中选择距离值最小者所对应的行 i(也即第 i 个数据点), 定义为类中心点搜索的初始点.

2　基于类中心与边界自寻优算法

2.1　算法的理论分析

(1) 假设以数据点 P_1 为起始点, 则以它为中心的某个邻域内 (如图 1(a)) 包含数据点{1, 2, 3, 4, 5}. 在距离半径矩阵 R 中, 比较各数据点对应的行向量{R_1, R_2, R_3, R_4, R_5}各列的距离半径值{R_{1j}, R_{2j}, R_{3j}, R_{4j}, R_{5j}}, 会发现随着列数 j 的递增数据点 P_5 的 R_{5j} 数值为最小; 同理, 在距离半径累计矩阵 S 中的{S_{1j}, S_{2j}, S_{3j}, S_{4j}, S_{5j}}, 也会发现 S_{5j} 的数值也为最小, 因为 P_5 更靠近区域数据中心, 所以其邻域内包含同样数量的数据点时, 其距离半径值相对最小.

(2) 接下来再以 P_5 为中心的某个邻域内 (如图 1(b)), 去掉原先已经参与过比较的点{1, 2, 3, 4}后, 所包含的新数据点{5, 6, 7, 8, 9, @, a, b}. 在距离半径矩阵 R 中, 比较新数据点对应的行向量{R_5, R_6, R_7, R_8, R_9, $R_@$, R_a, R_b}各列的距离半径值{R_{5j}, R_{6j}, R_{7j}, R_{8j}, R_{9j}, $R_{@j}$, R_{aj}, R_{bj}}, 会发现随着列数 j 的递增数据点 $P_@$的 $R_{@j}$ 数值为最小; 同理, 也会比较发现 S 矩阵中的 $S_{@j}$ 数值也为最小, 则表明 $P_@$更靠近区域数据中心.

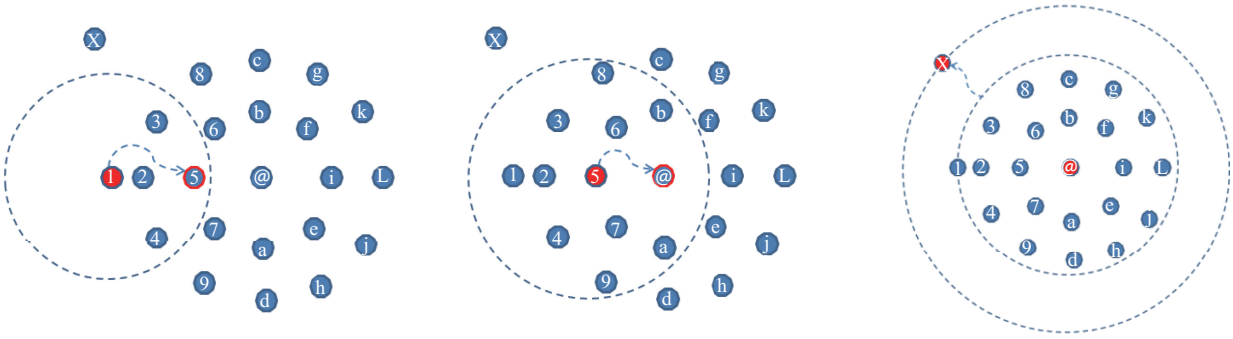
(3) 再以点 $P_@$为中心的某个邻域内 (如图 1(c)), 去掉之前已经参与过比较的点后, 所包含的新数据点为{@, c, d, e, f, g, h, i, j, k, L}. 在距离半径矩阵 R 中, 比较新数据点对应的行向量各列的距离半径值, 会发现随着列数 j 的递增数据点 $P_@$的 $R_{@j}$ 数值仍然为最小; 同理, 也会比较发现 S 矩阵中的 $S_{@j}$ 数值也为最小, 则表明点 $P_@$就是区域数据中心点, 也即该聚类的类中点.

(4) 在确定聚类的中点 $P_@$之后, 以其为中心逐步扩大其邻域半径 (如图 1(c)), 也即, 在如下各矩阵的第@行向量中逐渐递增列数 j, 分别寻找出现数值剧增的跳变点 **X**, 即为本聚类的边界点:

- 距离半径矩阵 R 中的{R(@, j)}
- 一阶距离半径偏导矩阵 R'的{R'(@, j)}

由于聚类的边界点的特征是点分布稀疏, 点之间相对距离较远, 因此, 当 j 点处于边界点时, 如果 j 继续增加, $P_@$点邻域将包含了其他聚类的数据点或孤立点 X, 则导致距离半径突增, 也会使 R(@, j) 和 R'(@, j) 激增. 那么, 以突变点 X 为邻域的边界点以内的各数据点属于本聚类, 而其余的数据点则属于其他的聚类或为孤点和噪声.

(5) 其余新聚类的搜索, 将在剩余的数据集之中按照上述方法继续搜寻, 直至只剩下孤立点和噪声点. 最后将包含共同数据点的类合并为同一个类.



(a) 从起始点开始寻优　　(b) 寻优中间过程　　(c) 类中心与边界的寻优

图 1　算法的理论分析示意图

上述算法充分体现了聚类分析的本质和内涵: 同一类中的数据对象之间具有较高的相似和聚合度, 不同类中的数据对象差异较大、相距较远.

2.2　算法过程描述

给定一数据集, 数据点离类中心越近, 其邻域聚合度越高, 反之越低, 当到达聚类边界时, 聚合度会发生剧烈的跳变和下降. SOCBC 算法的基本思想是: 基于聚合度寻优和广度优先策略逐步从聚合度较低的数据点向较高的点迁移直至到达类中心点, 再以其为中心向周围邻域逐渐扩展直至搜寻到聚合度剧减的跳变点作为类的边界, 则其边界内的各数据点属于此类.

首先, 计算各数据点之间的欧氏距离, 然后归类、排序、生成与各点对应的距离半径矩阵 R 与编号矩阵 P、半径累计矩阵 ΣR 或 S、以及偏导矩阵 R'; 再依据广度优先在矩阵 R 和 ΣR 中寻优到一聚合度最大的点, 也即其某一邻域内其距离半径及其累计均为最小, 作为搜索的起始点; 在以其为中心的动态邻域中搜索 R 和 ΣR 最小的点, 如此往复直至最小点不发生迁移,

即为类中心点; 再以其为中心逐渐扩展其邻域在偏导矩阵 R'中搜索跳变点作为类中心点边界, 其内各点构成一类; 如此往复在剩余的数据点集中继续寻找其他的类中心及其边界, 直至剩下不属于任何类的孤立点或噪声; 最后, 将具有共同数据点的类合并为同一个类.

算法. SOCBC

输入: 数据集 X

输出: 聚类、孤立点或噪声

步骤如下:

① 对数据集 X 中的 n 个数据点进行编号 (i=1, 2, …, n);

② 计算各数据点之间的欧氏距离、距离半径矩阵 R 及其对应的编号矩阵 P、一阶行偏导矩阵 R'、以及距离半径累计矩阵 ΣR;

③ 在 R 矩阵的第 j 列 (j 可以取 n*10% 的整数值) 中选择数值最小者所对应的行 i(也即第 i 个数据点); 同理, 在 ΣR 矩阵的第 j 列中也选择数值最小者对应的数据点 k. 如果 $i\neq k$, 则按照广度优先搜索的方式调整 j 的取值, 重复上述寻优直至 $i=k$, 则将此时的 i 点作为数据中心搜索的起始点;

④ 在起始点的动态邻域内比较 R 矩阵中各点对应的行向量$\{R_i\}$各列的距离半径值$\{R_{ij}\}$, 逐渐递增列数 j, 按照广度优先的寻优方式取其中最小值对应的点 i; 同理, 在同一邻域内各点对应于 ΣR 矩阵的第 j 列中也选择数值最小者对应的数据点 k. 如果 $i\neq k$, 则调整 j 的取值, 重复上述寻优直至 $i=k$, 将 i 点作为新的数据中心点, 如此反复递归直至数据中心点@不发生变化, 则此时的数据中心点为一聚类的中心点@;

⑤ 以@为中心点逐步扩大其邻域半径, 也即: 点@在各矩阵 R 和 R'中对应的第@行向量$\{R(@, j)\}$和$\{R'(@, j)\}$中, 逐渐递增列数 j, 分别搜寻和综合评判确定跳变点 X, 作为本聚类的边界点. 则, 以@为中心点, 以突变点 X 为邻域边界点以内的各数据点属于本聚类;

⑥ 在剩余的数据集之中重复 345 步骤继续搜寻、确定其余的新聚类中心点, 直至只剩下孤点和噪声点;

⑦ 最后, 将具有共同数据点的类合并为同一个类, 则聚类中心搜索和数据归类结束;

⑧ 输出聚类结果.

3 实验结果及分析

为更进一步地论证“类中心与边界自寻优聚类”SOCBC 算法的正确性和鲁棒性, 用一些经典的聚类数据集对其进行了聚类实验, 并对类中心点与周围其他数据点的距离半径 R、半径累计 ΣR 及其偏导 R'在一定的邻域内的分布曲线进行了对比分析, 如图 3-图 6. 为了能够直观形象地表示数据之间的聚类关系, 实验选用了经典的二维数据集 Aggregation(如图 2), 其包含 788 条记录, 共分为 7 类.

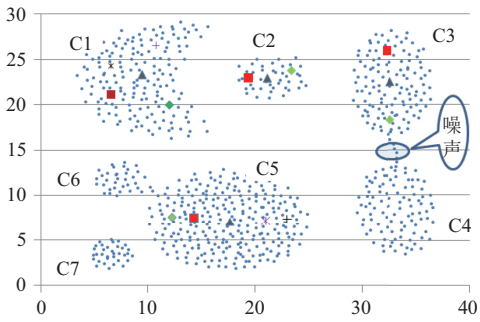
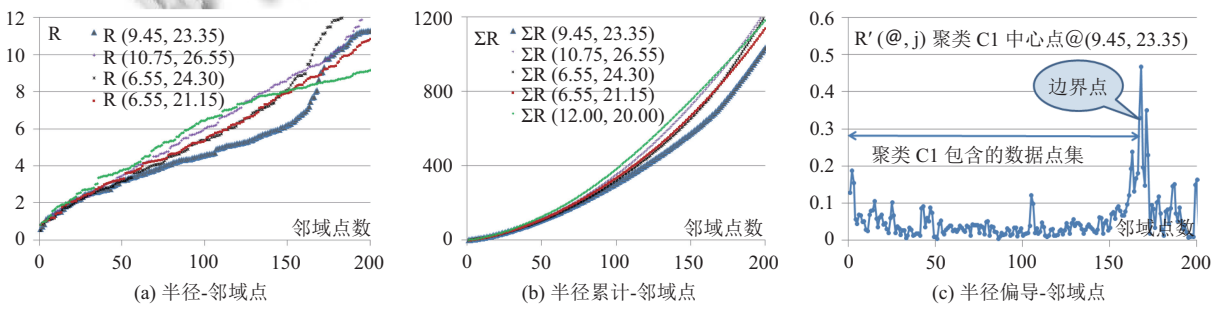
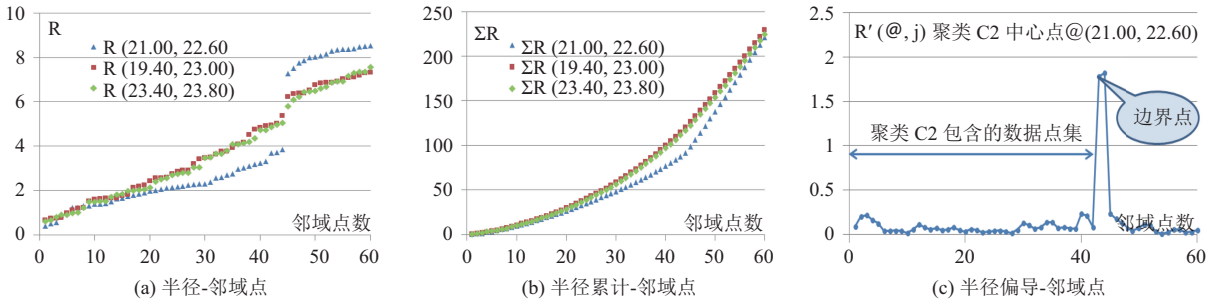


图 2 Aggregation 数据集与聚类



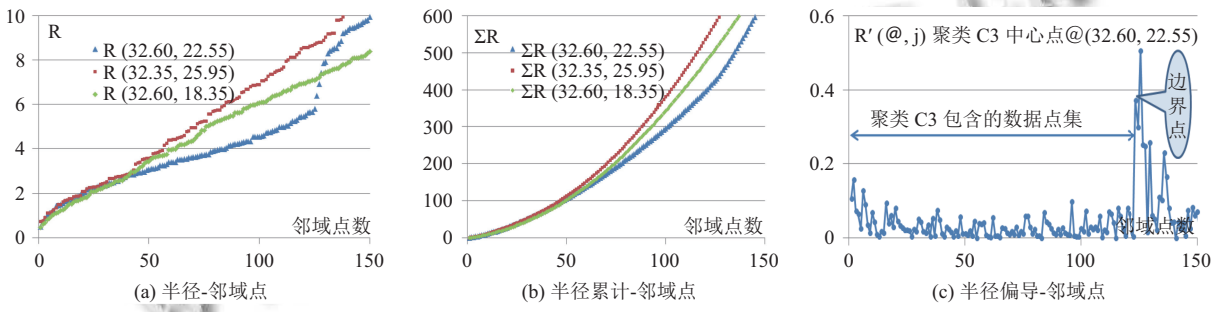
(a) 半径-邻域点 (b) 半径累计-邻域点 (c) 半径偏导-邻域点

图 3 聚类 C1 的中心点与周边点比较



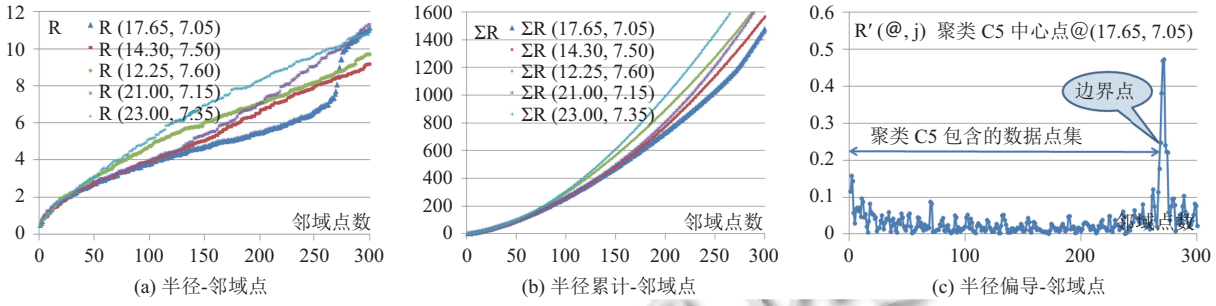
(a) 半径-邻域点 (b) 半径累计-邻域点 (c) 半径偏导-邻域点

图 4 聚类 C2 的中心点与周边点比较



(a) 半径-邻域点 (b) 半径累计-邻域点 (c) 半径偏导-邻域点

图 5 聚类 C3 的中心点与周边点比较



(a) 半径-邻域点 (b) 半径累计-邻域点 (c) 半径偏导-邻域点

图 6 聚类 C5 的中心点与周边点比较

算法对 Aggregation 数据集中的 C1 类数据群进行分析计算时, 表明:

(1) 图 3(a): 表示 5 个数据点在各自的邻域内, 包含相同的点数时其对应的邻域半径 R, 比较各点曲线可发现点@(9.45, 23.35) 对应的邻域半径为最小, 也即该点的邻域内各点最稠密、聚合度最高, 为数据聚类的中心点.

(2) 图 3(b): 表示 5 个数据点在各自的邻域内, 包含相同的点数时其邻域内各点距其的半径累计 ΣR, 比较各点曲线也可以发现点@(9.45, 23.35) 对应的邻域内各点距其的距离半径累计和为最小, 也反映了该点的邻域内点的聚合度最高, 也即为数据聚类的中心点.

(3) 图 3(c): 表示在中心点@(9.45, 23.35) 的邻域内各点的半径偏导 R', 反映邻域半径的增长率的分布曲线, 可发现其邻域内第 168 点处对应的“半径偏导”剧增至 0.33, 说明在该点处邻域半径发生剧增跳变, 也即此处的数据点聚集度剧减, 表明邻域已经处于本聚类的边界, 则本聚类所包含的数据点数为 168=168-该边界点+类中心点. 那么, 构成该聚类的数据集为: 该中心点在编号矩阵 P 中对应的行向量中的前 167 个数据点的编号, 再加上该聚类的中心点@本身.

同理, SOCBC 算法对其余的 C2、C3、C5 以及其他类进行自寻优聚类计算分析时, 也能准确地确定相应的类中心和边界及其各类所包含的数据点集, 如

表 1 所示. 另外, 值得一提的是, 该聚类算法将 C3 和 C4 类之间的几个数据点视为其边界之外的孤立点, 也即数据噪声.

表 1 聚类实验的结果分析

聚类	类中心点	边界点偏导	聚类点数
C1	(9.45, 23.35)	0.33	168
C2	(21.00, 22.60)	1.75	43
C3	(32.60, 22.55)	0.36	124
C5	(17.65, 7.05)	0.38	269

除此之外, 还用一些其他典型的高维聚类数据集测试了该算法, 也得到了满意的结果.

4 结语

理论研究、仿真和实践表明, 提出的距离半径矩阵 R、累计矩阵 ΣR 和偏导矩阵 R'概念及其基于类中心与边界自寻优的聚类算法, 能充分体现聚类的核心思想: 同一类中的数据对象之间具有较高的聚合度, 而非同类中的数据对象差异较大、相距较远. 该算法对多种类型的数据集的聚类具有较好的鲁棒性, 还能有效地识别孤立点和噪声. 但对环形、条形和缠绕等分布奇特的数据集的聚类, 还尚需进一步研究.

参考文献

1 张文军, 王建平, 范世平, 等. 深井冻结施工远程监测与故障诊断物联网的设计. 煤炭科学技术, 2015, 43(4): 82–87.

2 邹裕. 云计算平台的海量数据知识提取框架. 计算机系统应用, 2016, 25(11): 216–220. [doi: 10.15888/j.cnki.csa.005409]

3 张可文, 赵庆展, 周可法, 等. 基于移动互联网的兵团旅游信息系统. 计算机系统应用, 2016, 25(5): 65–70.

4 金建国. 聚类方法综述. 计算机科学, 2014, 41(11A): 288–293.

5 张云伟, 宋安军. 基于 K-Means 改进算法在微博话题发现中的应用研究. 计算机系统应用, 2016, 25(10): 308–311. [doi: 10.15888/j.cnki.csa.005461]

6 翟东海, 鱼江, 高飞, 等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究. 计算机应用研究, 2014, 31(3): 713–715, 719.

7 张永, 浮盼盼, 张玉婷. 基于分层聚类及重采样的大规模数据分类. 计算机应用, 2013, 33(10): 2801–2803.

8 杨善红, 梁金明, 李静雯. 基于网格密度影响因子的多密度聚类算法. 计算机应用研究, 2015, 32(3): 743–747.

9 苏辉, 葛洪伟, 张欢庆, 等. 密度敏感的数据竞争聚类算法. 计算机应用, 2015, 35(2): 444–447. [doi: 10.11772/j.issn.1001-9081.2015.02.0444]

10 肖春景, 乔永卫, 贺怀清, 等. 基于最佳聚类准则的多级模糊态势评估方法. 计算机应用研究, 2013, 30(4): 1011–1014.

11 林建辉, 严宣辉, 黄波. 基于 SVD 与模糊聚类的协同过滤推荐算法. 计算机系统应用, 2016, 25(11): 156–163. [doi: 10.15888/j.cnki.csa.005474]

12 赵慧, 刘希玉, 崔海青. 网格聚类算法. 计算机技术与发展, 2010, 20(9): 83–85, 89.

13 苏勇, 黄烨, 周冬. 基于网格结构的二次 CLARANS 聚类算法. 计算机应用与软件, 2013, 30(3): 287–290.

14 盛洪波, 汪西莉. 基于局部聚类的自适应线性近邻传递分类算法. 计算机应用, 2014, 34(1): 255–259. [doi: 10.11772/j.issn.1001-9081.2014.01.0255]

15 周炜奔, 石跃祥. 基于密度的 K-means 聚类中心选取的优化算法. 计算机应用研究, 2012, 29(5): 1726–1728.

16 楼巍. 面向大数据的高维数据挖掘技术研究[博士学位论文]. 上海: 上海大学, 2013: 82–89.

17 丁世飞, 贾洪杰, 史忠植. 基于自适应 Nyström 采样的大数据谱聚类算法. 软件学报, 2014, 25(9): 2037–2049. [doi: 10.13328/j.cnki.jos.004643]

18 Shirkhorshidi AS, Aghabozorgi S, Wah TY, *et al*. Big data clustering: A review. Proc. of the 14th International Conference on Computational Science and Its Applications. Guimarães, Portugal. 2014. 707–720.