

## 基于数据分布特性的聚类中心初始化方法

禹贵辉, 潘志斌, 乔瑞萍, 邹 彬

(西安交通大学 电子与信息工程学院, 西安 710049)

**摘 要:** 文中提出了一种新的基于数据局部和全局分布特性的 K-Means 初始化方法. 算法通过对数据空间进行网格化后统计每个网格中数据点数目, 选取具有数目局部最大值的网格, 再利用距离优化方法全局的估算出 K 个初始聚类中心. 在人工和真实数据集上, 进行了与传统的聚类中心初始化算法的比较. 实验结果表明, 该算法利用局部最大值网格和距离优化的方法估算的聚类中心能够在保持及改善聚类效果的同时, 明显减少迭代次数, 提高收敛速度.

**关键词:** 初始聚类中心; K-均值算法; 网格化; 局部最大值; 距离优化

**中图分类号:** TP391.4

**文献标识码:** A

**文章编号:** 1000-7180(2011)11-0152-05

## New Method for the Initialization of Clusters Based on Sata Distribution

YU Gui-hui, PAN Zhi-bin, QIAO Rui-ping, ZOU Bin

(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract:** A new initializing algorithm based on data distribution is proposed for K-Means in this paper. First we partition data space into grid and find the local-maximum cell which counts more data points than its neighborhood cells. Then we use distance optimization method to choose the seed clusters from local-maximum cells globally. Benchmark experiments evaluate the proposed method and five other typical initialization methods on both synthetic and real-life data sets, and the results demonstrated that our proposed algorithm gives faster convergence speed without descending in clustering performance.

**Key words:** initialization; K-Means; grid; local maximum; distance optimization

### 1 引言

K-Means 作为一种典型的聚类方法在模式识别、数据挖掘等领域被广泛使用, 其聚类结果对聚类中心的初始化选取比较敏感. 如果初始聚类中心选取不合适, 有可能使算法陷入局部最优, 并且影响收敛的速度.

为了确定算法的初始聚类中心, 已经提出五种初始化算法. 其中, 基于统计概率假设的随机选取方法(R-SEL)<sup>[1]</sup>被最早提出; R-MEAN<sup>[2]</sup>通过在数

据中心处产生随机微小扰动生成 K 个初值点, 能够避免随机选取方法出现某些类中数据点包含的数目为零的情况; SCS<sup>[3]</sup>和 KKZ<sup>[4]</sup>是两种典型的距离优化的算法; Kaufman 提出一种典型的密度估计 KR 方法<sup>[5]</sup>, 它通过计算两两距离估计数据分布的密度, 然后从具有较高局部密度的区域中的数据里选取初始聚类中心.

文中提出一种新的基于网格的初始化方法, 其出发点在于考虑到数据中各个类中心附近的数据分布较密集, 距类中心越远数据分布的密度越小, 那么

收稿日期: 2010-12-30; 修回日期: 2011-01-22

基金项目: 国家自然科学基金项目(60672054); 陕西省科学技术攻关项目(2008K04-01); 高校博士点基金项目(20100201110030)

对于特定聚类中心,聚类中心附近单位面积分布的数据点的数目一定大于边缘区域.

2 K-Means 算法

K-Means 是一种被广泛使用的基于划分的聚类方法. 给定数据集  $X = \{x_i \mid x_i \in R\}$ , 其中  $i = 1, 2, \dots, N$ , 将数据划分为  $K$  类:  $\{C_1, C_2, \dots, C_K\}$ , 每类样本的数目为  $N_k, k = 1, 2, \dots, K$  均值为  $\{m_1, m_2, \dots, m_K\}$  则

$$m_k = \frac{1}{N_k} \sum_{x_i \in C_k} x_i \tag{1}$$

K-均值聚类试图找出使平方误差函数值最小的划分,其目标函数为

$$J = \sum_{k=1}^K \sum_{i=1}^{N_k} \|x_i - m_k\| \tag{2}$$

算法具体流程<sup>[6]</sup>如下:

- (1) 初始化算法选取  $K$  个初始聚类中心点;
- (2) 循环下述流程(3)到(4),直到达到收敛条件;
- (3) 计算每个数据点与这些中心的距离,根据最小距离对数据点进行划分;
- (4) 根据划分结果更新聚类中心;

针对聚类中心初始化方法计算复杂度和不同的初始中心点可能得到不同的聚类结果两个问题,文中提出了一种新的基于网格的初始化方法,性能测试实验表明,该算法得到的结果要优于其他典型的初始化算法<sup>[1-5]</sup>.

3 文中算法

将数据的每一维等分为  $\eta$  个区间,数据空间划分为互不交叠的网格单元,数据集集中的每个点属于一个特定的网格,那么对于某个网格可以求出落入该网格中的数据点的数目. 对于  $D$  维数据集,一个网格单元共与  $3^D - 1$  个网格单元相邻,如果该网格单元中包含的数据点数目大于其相邻单元,则称该单元为局部最大网格单元. 一般情况下各个类中心附近的数据分布较密集,离类中心越远,数据分布的密度越小,类边缘的密度分布是从中心向四周逐步递减的,同时高密度的数据对象区域被低密度的对象区域分开,因此类中心所在的网格为局部最大网格. 而传统的基于密度估计算法需要计算任意数据点间距离,计算复杂度较高,采用网格化算法其复杂度仅与  $n$  成线性增加关系<sup>[7]</sup>.

因此,一般来说,在网格划分合适的情况下,对于特定的一个类,类中心网格包含数据的数目大于其相邻网格中的数据数目<sup>[7]</sup>,类中心所在的网格构成一个局部最大值,而边缘处的网格因与中心网格相邻而无法构成一个局部最大值. 图 1 给出了一个简单实例,(a)中包含 3 类服从混合高斯分布的二维数据集,类中包含数据的数目分别为 50、500、1000,(b)(c)中给出了将数据每维等分为 5 个区间,然后统计每个网格中数据点总数的结果,(b)中深色网格与(c)中阴影区域相对应,包含的数据数目大于各自邻域,定义为局部最大值网格,选取这些网格中

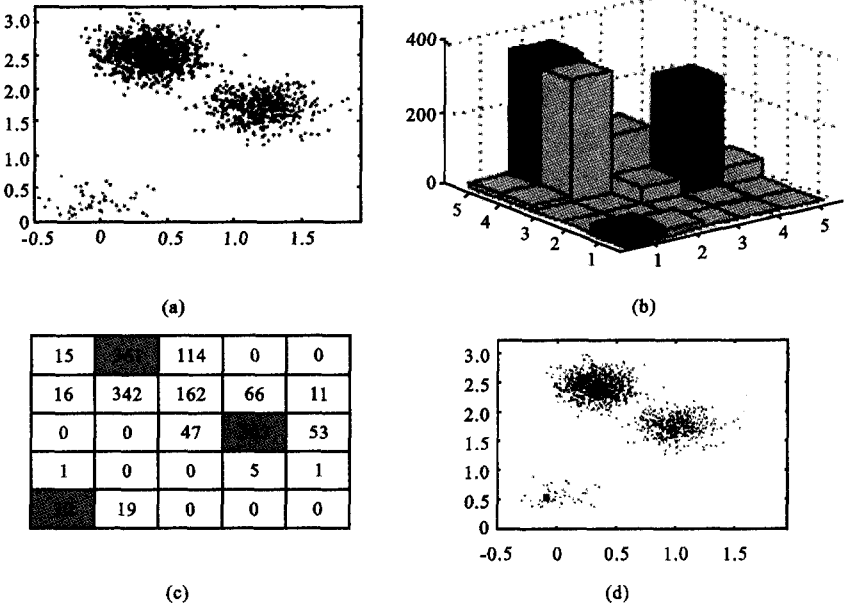


图 1 利用网格化局部最大值选取初始中心示意图

所有数据的均值为初始中心点,结果如(d)所示。

考虑到噪声的影响,可能得到的局部最大值网格数目会大于  $K$ ,文中提出了一种结合网格中心点数目与网格间距离的新的距离优化算法<sup>[8]</sup>,来从局部最大值网格中估算出  $K$  个初始中心点。如图 2 所示,纵坐标表示网格中数据点数目,粗实线  $C$  表示一个已经选取为中心点的局部最大值网格,细实线  $G_1, G_2$  为备选的网格,传统的距离优化算法仅根据  $G_1, G_2$  距  $C$  的距离  $\text{dis}$  而选取  $G_2$  作为下一个初始中心,如果存在因噪声而产生的局部最大值,这种单纯考虑距离的方法并不合适。实际上选取  $G_1$  作为下一个初始中心应该比  $G_2$  更为合理,文中采取  $\alpha = \frac{\Delta \text{dis}}{\Delta N}$  作为选取标准,即优先选取距离越远,数目越接近甚至超过已选网格的备选网格作为下一个初始中心。

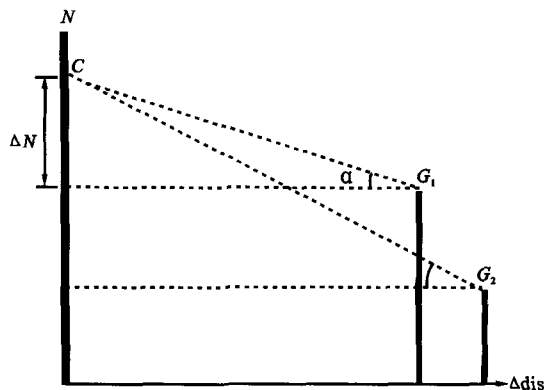


图 2 结合网格数据数目和距离的优化方法

基于上述的考虑,认为用局部最大值的网格来估算初始中心点是比较合理的,同时考虑到噪声的影响,再利用结合网格数据数目和距离的优化方法估算得到初始  $K$  个初始聚类中心,文中提出的初始化算法如下:

输入:包含  $n$  个数据点的  $D$  维数据集  $X$ ,聚类数目  $K$ ,以及每个维度上所划分的区间个数  $\eta$ 。

输出:数据集的  $K$  个类。

(1) 将数据的每一维等分为  $\eta$  个区间,数据空间划分为互不交叠的网格单元,统计落入网格中的数据点的数目;

(2) 遍历所有网格,寻找具有局部最大值的网格  $G_i (i = 1, 2, \dots, M)$  共  $M (M \geq K)$  个,若  $M < K$ ,则增加  $\eta$ ,然后返回(1);

(3) 计算每个具有局部最大值的网格中数据点的均值

$$Y = \{y_i \mid y_i = \frac{1}{N_{G_i}} \sum_{x_j \in C_i} x_j, i = 1, 2, \dots, M\} \quad (3)$$

其中,  $N_{G_i}$  表示网格包含的数据点数目;

(4) 从  $Y$  中选取  $K$  个点作为初始中心点,其过程为:

① 将包含数据点最多的网格作为第一个类中心  $C_1 = Y_{j_1}$ ;

② 对于  $K = 2, \dots, K$ ,  $C_k$  通过如下方式得到:对于每个未选取中心  $y_j$ ,计算其与所有已经存在的聚类中心的

$$\alpha_j = \max_{i=1, \dots, k-1} \left\{ \frac{N_{G_i} - N_j}{\|y_j - C_i\|} \right\} \quad (4)$$

其中,  $N_j$  表示未选中心  $y_j$  所在网格中包含的数据点数目,  $N_{G_i}$  为已确定中心  $y_{j_i}$  所在网格中包含的数据点数目,那么

$$C_k = y_{j_k} = \arg \min \{\alpha_j\} \quad (5)$$

(5) 以  $\{C_1, C_2, \dots, C_K\}$  作为初始聚类中心,应用  $K$ -均值聚类,得到聚类结果。

## 4 实验

为验证文中方法的有效性,采用人工数据集和真实数据集与典型的五种传统初始化算法在如下 3 个方面进行性能比较<sup>[9]</sup>。五种方法为 R-MEANS、R-SEL、SCS、KKZ 和 KR 方法,包括迭代次数  $I$ ,紧凑度  $\text{Cmp}$ ,分离度  $\text{Sep}$  等方面。如表 1 所示。

实验结果表明,文中算法的迭代次数都小于其他 5 种典型的算法,可以使  $K$  均值算法尽快到达收敛,其中最为明显的是 S13 数据集,迭代次数只有其他 5 种算法中最好结果的 53%;同时在聚类结果性能方面,得到紧凑度均小于其他 5 种算法。因此,相对于 5 种典型的初始化算法,在人工数据集上,文中算法能够在改善聚类性能的同时,明显减少迭代次数,使聚类过程尽快到达收敛。

在真实数据集中,采用 UCI 数据库<sup>[10]</sup> 中的数据集: Iris 数据集和 Letter Recognition 数据集。UCI 数据库是一个专门用于测试机器学习、数据挖掘算法的数据库。表 2 给出了这些数据集的具体描述,实验结果如表 3 所示。

从实验的结果来看,对于 Iris 数据集,文中算法能够最快达到收敛,对于 Letter Recognition 数据集,文中算法在迭代次数和分离度上优于其他 5 种算法。相对真实数据算法,人工数据集的性能改善更为明显,这是由于人工数据集的聚类中心数目与数据维数的相对值大于真实数据,而且人工数据集的分布更加符合离类中心越近密度越高的假设,这样选取的局部最大值更能反映聚类中心的位置。综上

所述,文中算法在人工数据集和真实数据集上都能 次数.  
 够在保持及改善聚类效果的同时,明显减小迭代

表1 文中算法与5种典型算法在3组人工数据集上的实验结果

|                                                            | I                 | Cmp                 | Sep                 |
|------------------------------------------------------------|-------------------|---------------------|---------------------|
| Data Set: S01( $v=0.05$ , $r=0.0$ ), $\sigma=0.5$ , $K=10$ |                   |                     |                     |
| R-MEANS                                                    | $7.67 \pm 2.85$   | $0.1411 \pm 0.0087$ | $0.6306 \pm 0.0287$ |
| R-SEL                                                      | $6.29 \pm 2.79$   | $0.1403 \pm 0.0130$ | $0.5728 \pm 0.0351$ |
| SCS                                                        | $4.25 \pm 2.02$   | $0.1265 \pm 0.0072$ | $0.5553 \pm 0.0160$ |
| KKZ                                                        | $3.75 \pm 0.85$   | $0.1330 \pm 0.0049$ | $0.5399 \pm 0.0159$ |
| KR                                                         | $6.80 \pm 2.33$   | $0.1225 \pm 0.0059$ | $0.5661 \pm 0.0053$ |
| Proposed                                                   | $1.65 \pm 0.49$   | $0.1198 \pm 0.0003$ | $0.5677 \pm 0.0002$ |
| Data Set: S13( $v=0.10$ , $r=0.2$ ), $\sigma=0.5$ , $K=10$ |                   |                     |                     |
| R-MEANS                                                    | $16.17 \pm 5.04$  | $0.2695 \pm 0.0068$ | $0.5828 \pm 0.0120$ |
| R-SEL                                                      | $17.34 \pm 7.44$  | $0.2703 \pm 0.0072$ | $0.5766 \pm 0.0126$ |
| SCS                                                        | $12.36 \pm 4.73$  | $0.2711 \pm 0.0058$ | $0.5648 \pm 0.0085$ |
| KKZ                                                        | $11.35 \pm 3.73$  | $0.2683 \pm 0.0054$ | $0.5680 \pm 0.0098$ |
| KR                                                         | $8.30 \pm 1.78$   | $0.2637 \pm 0.0043$ | $0.5790 \pm 0.0036$ |
| Proposed                                                   | $4.45 \pm 1.23$   | $0.2596 \pm 0.0021$ | $0.5811 \pm 0.0016$ |
| Data Set: S25( $v=0.15$ , $r=0.4$ ), $\sigma=0.5$ , $K=10$ |                   |                     |                     |
| R-MEANS                                                    | $22.58 \pm 7.97$  | $0.3079 \pm 0.0040$ | $0.5755 \pm 0.0074$ |
| R-SEL                                                      | $22.66 \pm 10.39$ | $0.3082 \pm 0.0040$ | $0.5735 \pm 0.0074$ |
| SCS                                                        | $18.55 \pm 8.12$  | $0.3072 \pm 0.0031$ | $0.5678 \pm 0.0050$ |
| KKZ                                                        | $21.45 \pm 9.13$  | $0.3063 \pm 0.0026$ | $0.5672 \pm 0.0031$ |
| KR                                                         | $17.10 \pm 5.67$  | $0.3092 \pm 0.0016$ | $0.5637 \pm 0.0018$ |
| Proposed                                                   | $14.55 \pm 2.82$  | $0.3051 \pm 0.0009$ | $0.5713 \pm 0.0016$ |

表2 实验采用的2组真实数据的具体描述

| 数据集                | 维数 Dim | 类的个数 K | 数据点总数 N |
|--------------------|--------|--------|---------|
| Iris               | 4      | 3      | 150     |
| Letter Recognition | 16     | 26     | 20 000  |

表3 文中算法与5种典型算法在2组真实数据集上的实验结果

|                                      | I                 | Cmp                 | Sep                 |
|--------------------------------------|-------------------|---------------------|---------------------|
| Data Set: Iris $\sigma=1$ , $K=3$    |                   |                     |                     |
| R-MEANS                              | $7.15 \pm 3.63$   | $0.3415 \pm 0.0226$ | $0.3114 \pm 0.0391$ |
| R-SEL                                | $5.55 \pm 2.96$   | $0.3409 \pm 0.0082$ | $0.3405 \pm 0.0369$ |
| SCS                                  | $5.60 \pm 2.68$   | $0.3361 \pm 0.0004$ | $0.3192 \pm 0.0009$ |
| KKZ                                  | $4.00 \pm 0.00$   | $0.3358 \pm 0.0000$ | $0.3187 \pm 0.0000$ |
| KR                                   | $9.00 \pm 0.00$   | $0.3366 \pm 0.0000$ | $0.3206 \pm 0.0000$ |
| Proposed                             | $3.00 \pm 0.00$   | $0.3366 \pm 0.0000$ | $0.3206 \pm 0.0000$ |
| Data Set: Ltrrec $\sigma=5$ , $K=26$ |                   |                     |                     |
| R-MEANS                              | $31.65 \pm 10.09$ | $0.6028 \pm 0.0038$ | $0.1404 \pm 0.0080$ |
| R-SEL                                | $30.65 \pm 7.15$  | $0.6045 \pm 0.0045$ | $0.1410 \pm 0.0061$ |
| SCS                                  | $32.30 \pm 8.18$  | $0.6033 \pm 0.0053$ | $0.1317 \pm 0.0053$ |
| KKZ                                  | $27.40 \pm 4.86$  | $0.6119 \pm 0.0053$ | $0.1377 \pm 0.0044$ |
| KR                                   | $28.80 \pm 10.42$ | $0.6012 \pm 0.0047$ | $0.1325 \pm 0.0047$ |
| Proposed                             | $21.00 \pm 0.00$  | $0.6042 \pm 0.0000$ | $0.1258 \pm 0.0000$ |

## 5 结束语

文中提出了一种新的基于网格的初始化 K-Means 算法. 该算法首先统计网格中数据出现的数目, 寻找局部最大值, 再利用基于网格的距离优化的方法估计聚类中心. 实验结果表明, 相对于传统初始化算法, 文中算法得到的初始化聚类中心能够在保持及改善聚类效果的同时, 明显减少迭代次数, 提高收敛速度.

### 参考文献:

- [1] Forgy E. Cluster analysis of multivariate data: efficiency vs interpretability of classifications [J]. *Biometrics*, 1965, 21(3): 768-769.
- [2] He J, Tan A H, Tan C L. ART-C: A neural architecture for self-organization under constraints [C]// *Proceedings of International Joint Conference on Neural Networks*. Honolulu, HI: IEEE, 2002: 2550 - 2555.
- [3] Tou J T, Gonzalez R C. *Pattern recognition principles* [M]. [s. l.]: Addison Wesley Pub Co, 1974.
- [4] Katsavounidis I, Kuo C, Zhang Z. A new initialization technique for generalized Lloyd iteration [J]. *IEEE Signal Processing Letters*, 1994, 1(10): 144 - 146.
- [5] Kaufman L, Rousseeuw P J. *Finding groups in data: an introduction to cluster analysis* [M]. [s. l.]: John Wiley & Sons, 1990.

- [6] 边肇祺, 张学工. *模式识别* [M]. 北京: 清华大学出版社, 2000: 235-237.
- [7] 牛琨, 张舒博, 陈俊亮. 融合网格密度的聚类中心初始化方案 [J]. *北京邮电大学学报*, 2007, 30(2): 6-10.
- [8] 苏锦旗, 薛惠锋, 詹海亮. 基于划分的 K-均值初始聚类中心优化算法 [J]. *微电子学与计算机*, 2009 (26): 8-11.
- [9] He Ji, Lan M, Tan C L. Initialization of cluster refinement algorithms: a review and comparative study [C]// *Proceedings of International Joint Conference on Neural Networks*, Budapest: IEEE, 2004: 297-302.
- [10] Asuncion A, Newman D. UCI machine learning repository [EB/OL]. [2010-12-30]. [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, School of Information and Computer Science, 2007.

### 作者简介:

禹贵辉 男, (1985-), 硕士研究生. 研究方向为图像识别.  
 潘志斌 男, (1963-), 教授. 研究方向为图像编码、图像识别.  
 乔瑞萍 男, (1966-), 副教授. 研究方向为图像编码算法及其 DSP 实现.  
 邹彬 男, (1986-), 硕士研究生. 研究方向为基于内容的图像检索

(上接第 151 页)

- [5] 田祖伟, 李勇帆. 基于汇编代码的指令调度器的设计与实现 [J]. *计算机科学*, 2009, 36(3): 45-47.

### 作者简介:

董锐 女, (1986-), 硕士研究生. 研究方向为编译器设

计和优化.

王志君 女, (1981-), 助理研究员. 研究方向为 DSP/CPU 处理器设计和专用集成电路.  
 梁利平 男, (1969-), 研究员. 研究方向为 DSP/CPU 处理器设计和 VLSI 信号处理.