

# K-Means 聚类算法的研究

周爱武, 于亚飞

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

**摘要:** K-Means 算法是一种经典的聚类算法, 有很多优点, 也存在许多不足。比如初始聚类数  $K$  要事先指定, 初始聚类中心选择存在随机性, 算法容易生成局部最优解, 受孤立点的影响很大等。文中主要针对 K-Means 算法初始聚类中心的选择以及孤立点问题加以改进, 首先计算所有数据对象之间的距离, 根据距离和的思想排除孤立点的影响, 然后提出了一种新的初始聚类中心选择方法, 并通过实验比较了改进算法与原算法的优劣。实验表明, 改进算法受孤立点的影响明显降低, 而且聚类结果更接近实际数据分布。

**关键词:** K-Means 算法; 初始聚类中心; 孤立点

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1673-629X(2011)02-0062-04

## The Research about Clustering Algorithm of K-Means

ZHOU Ai-wu, YU Ya-fei

(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

**Abstract:** The algorithm of K-means is one kind of classical clustering algorithm, including both many points and also shortages. For example must choose the initial clustering number. The choose of initial clustering centre has randomness. The algorithm receives locally optimal solution easily, the effect of isolated point is serious. Mainly improved the choice of initial clustering centre and the problem of isolated point. First of all, the algorithm calculated distance between all data and eliminated the effect of isolated point. Then proposed one new method for choosing the initial clustering centre and compared the algorithm having improved and the original algorithm using the experiment. The experiments indicate that the effect of isolated point for algorithm having improved reduces obviously, the results of clustering approach the actual distribution of the data.

**Key words:** K-Means; initial clustering centre; isolated point

聚类分析是数据挖掘领域中重要的研究课题, 用于发现大规模数据集中未知的对象类。它广泛地应用于文本搜索、模式识别、人工智能、图像分析<sup>[1]</sup>等领域。目前已经存在许多的聚类算法, 比如基于划分的 K-Means<sup>[2]</sup>算法, 基于层次的 CURE<sup>[3]</sup>算法, 基于密度的 DBSCAN<sup>[4]</sup>算法, 基于网格的 STING<sup>[5]</sup>方法, 基于模型的 COBWEB<sup>[6]</sup>方法等等。K-Means 算法是一种以平均值作为聚类中心的分割聚类方法, 简单而且快速, 但是本身存在着许多问题。文中主要针对 K-Means 算法的不足之处加以改进。

## 1 K-Means 算法

### 1.1 K-Means 算法

K-Means 算法首先需要选取初始聚类中心, 然后

对所有数据点进行分类, 最后计算每个聚类的平均值调整聚类中心, 不断的迭代循环。最终使类内对象相似性最大, 类间对象相似性最小<sup>[7]</sup>。具体的流程如下:

输入: 聚类的数目  $K$  和包含  $N$  个对象的数据库。

输出:  $K$  个聚类簇, 使平方误差准则最小。

方法:

(1) 从所有数据样本中随机选择  $K$  个对象, 作为初始聚类中心。

(2) 根据距离中心最近的原则, 计算其他数据对象到各聚类中心的距离, 将其分配到各个相应的类中。

(3) 对每一个类, 计算其所有对象的平均值, 作为新的聚类中心。

(4) 根据距离中心最近的原则, 重新进行数据对象的分配。

(5) 返回第(3)步循环执行, 当目标函数不再变化时算法结束。

### 1.2 K-Means 算法的优缺点

当聚类是密集的, 而聚类之间区别明显时, K-Means 算法的效果较好。另外, 对处理大数据集, 该算

收稿日期: 2010-06-01; 修回日期: 2010-09-17

基金项目: 安徽省教育科研重点项目 (KJ2009A57)

作者简介: 周爱武 (1965-), 女, 副教授, 研究方向为数据库与 web 技术、数据仓库与数据挖掘、信息系统安全。

法是高效率的,因为它的复杂度是  $O(nkt)$ , 其中,  $n$  是所有对象的数目,  $k$  是聚类的数目,  $t$  是迭代的次数。通常  $k \gg n$  且  $t \gg n$ 。但是, K-Means 算法只有在聚类的平均值被定义的情况下才能使用。如果处理符号属性的数据并不适用。K-Means 算法对初始聚类中心和样本的输入顺序敏感,对于不同的初始聚类中心和样本输入顺序,聚类结果会有很大差别。由于采用迭代更新的方法,所以当初始聚类中心落在局部值最小附近时,算法容易生成局部最优解。另外,算法的效果受孤立点的影响很大。

## 2 改进的 K-Means 算法

针对 K-Means 算法存在的不足,参考文献[8]采用聚类均值点与聚类种子相分离的思想,减小了孤立点的影响。参考文献[9]通过区域划分方法估算出  $K$  个中心点作为初始聚类中心,减少了迭代次数,提高了算法质量。参考文献[10]采取对数据进行预处理的方式选取初始中心,提高了算法效率。参考文献[11]中从数据预处理与初始聚类中心选择方面加以改进,降低了孤立点的影响,提高了聚类质量。

文中主要对初始聚类中心的选择和孤立点的问题加以改进。K-Means 算法对于初始聚类中心敏感,目前的主要方法是把全部样本直观的分为  $K$  类,计算各类的平均值,作为初始聚类中心,或者进行多次的初始聚类中心的选择,然后聚类,通过比较找出最好的一组聚类结果。

聚类的数据不可避免地会出现孤立点,即少量数据点远离数据密集区的情况,由于随机的选取初始聚类中心,可能会将孤立点选为初始聚类中心,这样会使聚类结果产生很大的偏差。另外,在进行聚类计算时,是将聚类均值点(类中所有数据的平均值)作为新的聚类中心进行新一轮的聚类计算,在这种情况下,新的聚类中心可能偏离真正的数据密集区,从而导致聚类结果出现偏差。由此可见,孤立点对于 K-Means 算法有很大的影响。所以改进算法首先运行孤立点查找算法,排除孤立点,然后进行聚类。孤立点在聚类算法之后单独聚类。

本文是在欧式距离的框架内,主要针对二维空间的数据样本进行分析,首先根据参考文献[12]中距离和的思想,排除孤立点的影响,通过计算数据集中各个对象之间的距离,筛选掉与其他对象的距离之和最大的点,如果需要,进一步筛选掉距离之和次大的点,根据精确度的要求,筛选掉  $M$  个数据对象,使孤立点不参与初始聚类中心的计算,从而避免聚类结果出现大的偏差。容易知道,为检测基于距离和的孤立点,算法将需要  $N$  平方次的数据对象间的距离计算,当  $N$  很大

时,计算量将非常大。为此,可以基于随机抽样的近似计算。从数据集中均匀地抽出一部分对象,假定抽出的对象可以有效地代表原数据集,则在计算每个数据对象与其它对象的距离和时,就不需计算对象与原始数据集中所有对象的距离,而是仅仅计算对象与抽出的对象的距离,一般抽取的对象数很少,所以算法的复杂度将极大地降低。

定义二维空间任意两点之间的距离  $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ ,  $(x_1, y_1), (x_2, y_2)$  是两个二维数据点,  $x$  为横坐标,  $y$  为纵坐标。初始聚类中心的选择具体有以下几个步骤:

(1) 通过运行孤立点查找程序计算  $N$  个对象两两之间的距离,输出距离矩阵  $cid$ ,  $cid$  为  $n * n$  的矩阵。

(2) 在 matlab 中执行  $a = \text{sum}(cid, 2)$  求出矩阵每行元素的和,即每个对象与其他所有对象的距离之和。

(3) 执行  $[q, l] = \max(a)$  求出距离之和最大的点所在的位置  $l$ , 从而找到第一个孤立点。

(4) 删除 1 个孤立点之后,返回第一步,在剩余  $N - 1$  个数据的输出矩阵中即可找到距离之和次大的点,以此类推,排除  $M$  个数据对象,满足精度要求。

(5) 删除  $M$  个孤立点之后,剩余的  $N - M$  个数据点再次运行孤立点查找程序,输出距离矩阵  $cid$ 。

(6) 在 matlab 中执行  $[t, c] = \max(\max(cid))$  求出距离最大的值以及列位置,从而找到距离最大的两个点之一,然后运行  $[t, a] = \max(cid)$  求出矩阵各列元素的最大值以及它们的行下标。找出刚才第  $c$  列数据的行坐标,确定另一个距离最大的点。作为开始的两个聚类中心。

(7) 然后将两点连成直线,以直线中心为圆心,以直线为直径画圆。基本可以把所有样本包括在内。然后在直线垂直方向画出另一条直径,即把圆四等分。距离直径端点最近的数据对象作为另一个初始聚类中心,如果需要划分更多的聚类,就需要对称画出另外的直径,把圆八等分,选取其中的直径端点作为参考点,然后计算距离端点最近的样本作为聚类中心。根据预先指定的  $K$  值,选取  $K$  个数据样本作为初始聚类中心。

对于第 7 步,端点坐标可以计算求得,对于数据比较大的情况,查找距离直径端点最近的数据会增加很大的计算量,所以对于大规模的数据,可以比较得到端点附近的数据点作为初始聚类中心,从而有效降低算法的时间复杂度。

通过输入初始聚类中心和  $K$  值,运行 K-Means 算法,得到剩余数据的聚类图,聚类算法会得到最后的聚类中心,对于开始排除的孤立点,根据距离中心最近原则,计算孤立点与最后聚类中心的距离,决定孤立点属于哪一类。

### 3 实验分析

实验利用 MATLAB 环境, 为了便于分析与观察, 实验采用二维数据, 且数据类型为实型。数据属性分别对应平面直角坐标系的横轴和纵轴。分别利用原 K-Means 算法和改进后的 K-Means 算法进行聚类。

#### 3.1 随机数据

对随机生成的 80 个数据样本进行聚类。用  $x = \text{rand}(80, 2)$  随机生成 80 个数据样本, 数据坐标范围在 0 和 1 之间。预先指定  $K = 4$ , 即把样本分为四类。聚类结果图中实心点表示第一类, 加号表示第二类, 圆形表示第三类, 五角星表示第四类。聚类首先得到 K-Means 算法迭代 100 次的精确结果, 然后比较分析相同结果的最少迭代次数。

利用改进的 K-Means 算法进行聚类, 运行孤立点查找程序, 排除距离之和最大的 6 个孤立点 (0.8936, 0.9883), (0.0153, 0.2091), (0.0185, 0.2523), (0.2311, 0.9568), (0.0099, 0.3340), (0.0196, 0.3050)。剩余的 74 个点再次运行孤立点查找算法输出距离矩阵。执行  $[t, c] = \max(\max(\text{cid}))$  确定 1 个初始聚类中心为 (0.9218, 0.8939), 运行  $[t, a] = \max(\text{cid})$  确定另一个初始聚类中心 (0.3093, 0.0439)。然后以这两点为直径画圆, 找到其他两个初始聚类中心为 (0.4289, 0.7889), (0.8913, 0.1730)。

输入初始聚类中心和  $K$  值, 运行 K-Means 算法之后, 改进算法得到最后的聚类中心 (0.7197, 0.8267), (0.3427, 0.1658), (0.3422, 0.6044), (0.7752, 0.2628)。根据孤立点与聚类中心的距离, 结果表明, 孤立点 (0.8936, 0.9883) 属于实心点区域, 第一类。孤立点 (0.2311, 0.9568) 属于圆形区域, 第三类。孤立点 (0.0153, 0.2091), (0.0185, 0.2523), (0.0099, 0.3340), (0.0196, 0.3050) 属于加号区域, 第二类。

原算法不排除孤立点, 随机选取初始聚类中心 (0.7382, 0.1991), (0.3420, 0.6072), (0.6038, 0.7604), (0.5936, 0.9334)。聚类结果如图 1 所示, 改进算法聚类结果如图 2 所示。聚类表见表 1。

表 1 随机数据的聚类表

算法	原算法 (图 1)	改进算法 (图 2)
迭代 100 次聚类结果	实心点 27 个	实心点 18 个
	加号 20 个	加号 12 个
	圆形 19 个	圆形 23 个
	五角星 14 个	五角星 21 个
CPU 用时 (秒)	至少迭代 9 次	至少迭代 5 次
	0.0470	0.0310

从聚类图和聚类表可以看出, 原 K-Means 算法初始聚类中心随机产生, 所以聚类很容易产生偏差。迭代次数较多, 与实际数据分布差别比较大。改进算法由于开始排除了孤立点的影响, 选择了较好的初始聚类中心, 所以聚类结果更加准确。

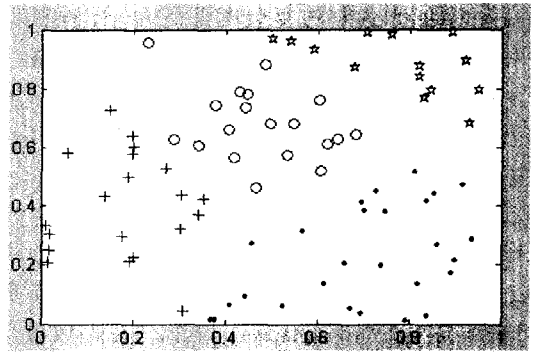


图 1 原算法 80 个数据聚类图

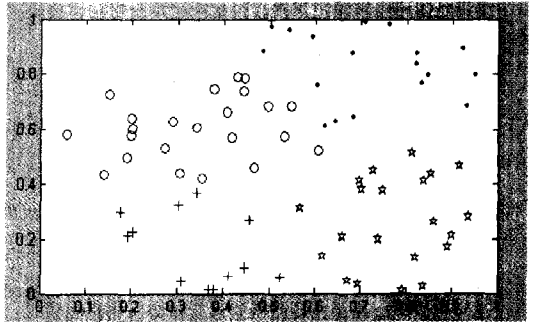


图 2 改进算法排除 6 个孤立点后聚类图

#### 3.2 标准数据

实验采用 UCI 数据库里的 Iris 数据集, Iris 数据集包含有 4 个属性, 150 个数据对象, 可分为 3 类。选用 Iris 数据集中 2 维的数据进行聚类, 排除重复数据后, 还有 121 个数据。分别用原算法和改进算法进行聚类。实验结果图中, 加号表示第一类, 圆形表示第二类, 五角星表示第三类。

因为 Iris 数据集比较标准, 所以改进算法只需排除 7 个孤立点。执行孤立点查找程序, 找到距离之和最大点 (2.6, 6.9), 循环查找其余 6 个孤立点 (3.8, 6.7), (3.6, 1.0), (4.0, 1.2), (4.4, 1.5), (4.2, 1.4), (2.3, 1.3) 达到精确度要求。

排除孤立点后, 对剩余的 114 个数据运行孤立点查找程序, 找到距离最大的两个点作为初始聚类中心 (3.0, 1.1), (2.8, 6.7)。以这两点画圆和另一条垂直的直径, 找到第三个聚类中心 (2.0, 3.5), 满足  $k = 3$  的要求。执行 K-Means 算法进行聚类。

改进算法聚类之后, 得到最后的聚类中心 (3.4161, 1.5000), (3.0515, 5.6727), (2.7500, 4.3240)。根据距离中心最近原则, 决定孤立点的位置, 孤立点 (2.6, 6.9), (3.8, 6.7) 属于圆形区域, 第二类。孤立点 (3.6, 1.0), (4.0, 1.2), (4.4, 1.5), (4.2, 1.4), (2.3, 1.3) 属于加号区域, 第一类。

原算法随机选取初始聚类中心 (4.2000, 1.4000), (3.0000, 5.1000), (3.9000, 1.7000) 聚类结果如图 3 所示。改进算法聚类结果如图 4 所示。

表 2 标准数据的聚类表

算法	原算法(图 3)	改进算法(图 4)
迭代 100 次聚	加号 13	加号 31
类结果	圆形 84	圆形 33
	五角星 24	五角星 50
CPU 用时(秒)	至少迭代 3 次 0.0310	迭代 1 次 0.0160

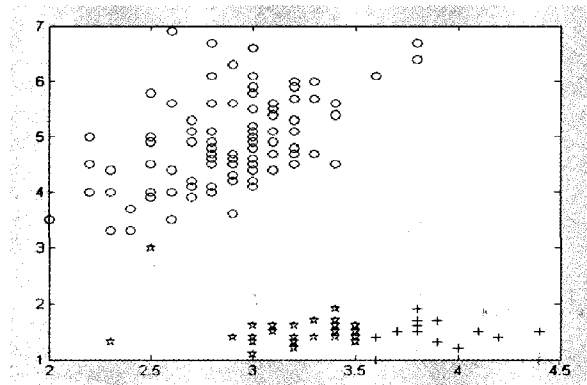


图 3 原算法 121 个数据聚类图

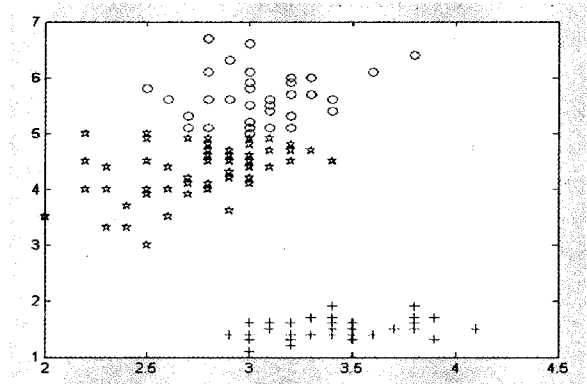


图 4 改进算法排除 7 个孤立点后聚类图

从聚类图和聚类表可以看出,对于标准数据来说,由于分布比较均匀,改进算法迭代次数明显减少,执行时间明显减少。

4 结束语

K-Means 算法作为一种常用的聚类算法,对球状分布的数据具有很好的效果,但是算法对初始聚类中心敏感,容易受到孤立点的影响。文中在聚类之前排除了孤立点的影响,提出了一种新的选取初始聚类中

心的方法。

实验结果表明,改进算法更接近实际数据分布。虽然需要查找少量孤立点,会增加时间消耗,但是改进算法准确度较高,聚类效果较好。

参考文献:

[1] 周卫星,廖 欢. 基于 K 均值聚类和概率松弛法的图像区域分割[J]. 计算机技术与发展,2010,20(2):68-70.

[2] Mac Q J. Some methods for classification and analysis of multivariate observations [C]// In: Proc. 5th Berkeley Symposium in Mathematics. Berkeley, USA; Univ of California,1967.

[3] GUHA S, RASTOGI R, SHIM K. CURE: An efficient clustering algorithm for large databases [C]// Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1998: 73-84.

[4] Ester, Martin, Hans Peter Kriegel, et al. A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon: [s. n. ], 1996.

[5] Wang W, Yang J, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining [C]// Proc. of 1997 Intl. Conf. on Very Large Databases. Athens, Greece: [s. n. ], 1997:186-195.

[6] Kohonen T. Self-organized Formation of Topologically Correct Feature Maps [J]. Biological Cybernetics, 1982, 43 (1): 59-69.

[7] 朱 明. 数据挖掘 [M]. 合肥: 中国科学技术大学出版社, 2002.

[8] 李业丽,秦 臻. 一种改进的 k-means 算法 [J]. 北京印刷学院学报, 2007, 15 (2): 63-65.

[9] 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 (1): 8-11.

[10] 步媛媛,关忠仁. 基于 K-means 聚类算法的研究 [J]. 西南民族大学学报: 自然科学版, 2009, 35 (1): 198-200.

[11] 连凤娜,吴锦林,唐 琦. 一种改进的 K-means 聚类算法 [J]. 电脑与信息技术, 2008, 16 (1): 38-40.

[12] 陆声链,林士敏. 基于距离的孤立点检测研究 [J]. 计算机工程与应用, 2004 (33): 73-75.

(上接第 61 页)

[10] 张 程,陈自郁,古 平,等. 基于 DOM 树结构的 Blog 网页自动识别 [J]. 计算机应用研究, 2008 (5): 1489-1491.

[11] 王洪伟,吴家春,蒋 馥. 基于描述逻辑的本体模型研究

[J]. 系统工程, 2003, 21 (3): 101-107.

[12] 郑冬冬,赵朋朋. Deep Web 爬虫研究与设计 [J]. 清华大学学报 (自然科学版), 2005, 45 (S1): 1896-1902.