第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·软件技术与数据库· 文章编号：1000—3428(2007)03—0065—02 文献标识码：A 中图分类号：TP39

初始聚类中心优化的 k-means 算法

袁 方，周志勇，宋 鑫
（河北大学数学与计算机学院，保定 071002）

摘 要：传统的 k-means 算法对初始聚类中心敏感，聚类结果随不同的初始输入而波动。为消除这种敏感性，提出一种优化初始聚类中心的方法，此方法计算每个数据对象所在区域的密度，选择相互距离最远的 k 个处于高密度区域的点作为初始聚类中心。实验表明改进后的 k-means 算法能产生质量较高的聚类结果，并且消除了对初始输入的敏感性。

关键词：数据挖掘；聚类；k-means 算法；聚类中心

K-means Clustering Algorithm with Meliorated Initial Center

YUAN Fang, ZHOU Zhiyong, SONG Xin
（College of Mathematics and Computer, Hebei University, Baoding 071002）

【Abstract】 The traditional k-means algorithm has sensitivity to the initial start center. To solve this problem, a new method is proposed to find the initial start center. First it computes the density of the area where the data object belongs to; then finds k data objects all of which are belong to high density area and the most far away to each other, using these k data objects as the initial start centers. Experiments on the standard database UCI show that the proposed method can produce a high purity clustering result and eliminate the sensitivity to the initial start centers.

【Key words】 Data mining; Clustering; K-means algorithm; Clustering center

聚类分析是数据挖掘领域的一个重要分支。聚类就是一个将数据集划分为若干组或类的过程，通过聚类使得同一组内的数据对象具有较高的相似度，而不同组中的数据对象则是不相似的[1]。

为了实现对数据对象的聚类，人们提出了很多种不同的算法。常用的有 k-means 算法[2]、STING 算法[3]、CLIQUE 算法[4]和 CURE 算法[5]。文献[6]分析了各种聚类算法的性能和适用范围。k-means 是一种基于划分的聚类算法，目的是通过在完备数据空间的不完全搜索，使得目标函数取得最大值。由于局部极值点的存在以及启发算法的贪心性，传统的 k-means 算法对初始聚类中心敏感，从不同的初始聚类中心出发，得到的聚类结果也不一样，并且一般不会得到全局最优解。在实际应用中，由于初始输入不同而造成结果的波动是不能接受的。因此怎样找到一组初始中心点，从而获得一个较好的聚类效果并消除聚类结果的波动性对 k-means 算法具有重要意义。

本文提出了一种寻找初始聚类中心的方法，使得初始聚类中心的分布尽可能体现数据的实际分布。实验表明了这种方法的可行性和有效性。

1 优化初始聚类中心算法的基本思想

设待聚类的数据集：

$\mathbf{X}=\{\mathbf{x}_i \mid \mathbf{x}_i \in R^p, i=1,2,\ldots,n\}$

k 个聚类中心分别为 $\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_k$。

用 $w_j(j=1,2,\ldots,k)$ 表示聚类的 k 个类别。有如下定义：

定义 1 两个数据对象间的欧氏距离为

$d(\mathbf{x}_i,\mathbf{x}_j)=\sqrt{(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)}$

定义 2 属于同一类别的数据对象的算术平均为

$\mathbf{z}_j=\frac{1}{N_j}\sum_{x\in w_j}\mathbf{x}$

定义 3 目标函数：

$J=\sum_{i=1}^{k}\sum_{j=1}^{n_j}d(\mathbf{x}_j,\mathbf{z}_i)$

k-means 算法[1]描述如下：

输入：聚类个数 k 以及包含 n 个数据对象的数据集；

输出：满足目标函数值最小的 k 个聚类。

算法流程：

(1)从 n 个数据对象中任意选择 k 个对象作为初始聚类中心；

(2)循环下述流程(3)到(4)，直到目标函数 J 取值不再变化；

(3)根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离，并且根据最小距离重新对相应对象进行划分；

(4)重新计算每个聚类的均值（中心对象）。

传统的 k-means 算法对初始聚类中心敏感，不同的初始中心往往对应着不同的聚类结果。本文的主要目的就是找到一组能反映数据分布特征的数据对象作为初始聚类中心，也就是改进上述算法中的第(1)步。

在用欧氏距离作为相似性度量的 k-means 算法中，相互距离最远的 k 个数据对象比随机取的 k 个数据对象更具有代表性。不过在实际的数据集中往往有噪声数据存在，如果只是单纯地取相互距离最远的 k 个点来代表 k 个不同的类别，有时会取到噪声点，从而影响聚类效果。一般在一个数据空间中，高密度的数据对象区域被低密度的对象区域所分割，通常认为处于低密度区域的点为噪声点[1]。为了避免取到噪声点，取相互距离最远的 k 个处于高密度区域的点作为初始

基金项目：河北省科技厅攻关计划基金资助项目(05213573)；河北省教育厅科研计划基金资助项目(2004406)

作者简介：袁 方(1965－)，男，教授，主研方向：数据挖掘，信息检索；周志勇，硕士生；宋 鑫，学士、助教

收稿日期：2006-04-27 **E-mail：**yuanfang@mail.hbu.edu.cn

聚类中心。

为了计算数据对象 x_i 所处区域的密度，定义一个密度参数：以 x_i 为中心，包含常数 *Minpts* 个数据对象的半径称之为对象 x_i 的密度参数，用ε表示。ε越大，说明数据对象所处区域的数据密度越低。反之，ε 越小，说明数据对象所处区域的数据密度越高。通过计算每个数据对象的密度参数，就可以发现处于高密度区域的点，从而得到一个高密度点集合 D。

表 1 传统 k-means 算法与改进 k-means 算法实验结果

算法	数据集									
	Iris		Balance-scale		New-thyroid		Haberman		Wine	
	初始中心	准确率	初始中心	准确率	初始中心	准确率	初始中心	准确率	初始中心	准确率
随	18,9,101	82.00%	466,561,599	49.92%	185,206,59	62.79%	198,291	51.96%	47,105,3	56.74%
机	35,6,21	88.67%	384,49,465	46.88%	110,25,122	78.14%	74,120	51.96%	25,117,141	70.22%
选	12,11,66	53.33%	616,496,589	50.88%	84,73,183	84.19%	19,182	50.00%	104,55,19	57.30%
择	143,36,85	57.33%	362,18,139	46.88%	204,202,91	65.12%	240,222	51.96%	35,119,34	57.30%
初	11,56,53	84.67%	22,115,401	50.72%	198,194,171	62.79%	161,216	50.98%	15,89,131	70.22%
始	16,2,29	57.33%	247,121,529	44.96%	94,200,46	86.05%	158,202	51.96%	133,35,101	70.22%
聚	47,5,42	52.00%	472,111,1	48.48%	1,102,202	69.77%	78,246	51.63%	116,153,86	57.30%
类	14,10,127	82.00%	316,128,345	50.40%	21,207,152	62.79%	108,145	75.82%	113,18,96	57.87%
中	12,49,134	76.77%	281,510,362	47.04%	55,116,207	62.79%	18,176	50.00%	26,112,169	70.22%
心	13,10,96	66.67%	501,127,10	46.72%	160,6,39	84.19%	162,190	52.29%	31,154,59	57.30%
平均		70.08%		48.29%		71.86%		53.86%		62.47%
改进	17,61,126	88.67%	1,475,601	51.20%	26,202,154	84.19%	146,63	75.82%	61,19,3	57.30%

在 D 中取处于最高密度区域的数据对象作为第 1 个聚类中心 $\mathbf{z}_1$；取距离 $\mathbf{z}_1$ 最远的一个高密度点作第 2 个聚类中心 $\mathbf{z}_2$；计算 D 中各数据对象 $\mathbf{x}_i$ 到 $\mathbf{z}_1, \mathbf{z}_2$ 的距离 $d(\mathbf{x}_i, \mathbf{z}_1)$，$d(\mathbf{x}_i, \mathbf{z}_2)$，$\mathbf{z}_3$ 为满足

$$\max(\min(d(\mathbf{x}_i, \mathbf{z}_1), d(\mathbf{x}_i, \mathbf{z}_2)) \; i = 1, 2, \ldots, n$$

的数据对象 $\mathbf{x}_i$；$\mathbf{z}_m$ 为满足

$$\max(\min(d(\mathbf{x}_i, \mathbf{z}_1), d(\mathbf{x}_i, \mathbf{z}_2), \ldots, d(\mathbf{x}_i, \mathbf{z}_{m-1})) \; i = 1, 2, \ldots, n$$

的数据对象 $\mathbf{x}_i$，$\mathbf{x}_i \in D$。依此得到 k 个初始聚类中心。

2 优化初始聚类中心的 k-means 算法

优化初始聚类中心的 k-means 算法描述如下：

输入：聚类个数 k 以及包含 n 个数据对象的数据集；

输出：满足目标函数值最小的 k 个聚类。

(1)计算任意两个数据对象间的距离 $d(\mathbf{x}_i, \mathbf{x}_j)$；

(2)计算每个数据对象的密度参数，把处于低密度区域的点删除，得到处于高密度区域的数据对象的集合 D；

(3)把处于最高密度区域的数据对象作为第 1 个中心 $\mathbf{z}_1$；

(4)把 $\mathbf{z}_1$ 距离最远的数据对象作为第 2 个初始中心 $\mathbf{z}_2$，$\mathbf{z}_2 \in D$；

(5)令 $\mathbf{z}_3$ 为满足 $\max(\min(d(\mathbf{x}_i, \mathbf{z}_1), d(\mathbf{x}_i, \mathbf{z}_2)), \; i = 1, 2, \ldots, n$ 的数据对象 $\mathbf{x}_i$，$\mathbf{z}_3 \in D$；

(6) 令 $\mathbf{z}_4$ 为满足 $\max(\min(d(\mathbf{x}_i, \mathbf{z}_1), d(\mathbf{x}_i, \mathbf{z}_2), d(\mathbf{x}_i, \mathbf{z}_3)))$，$i = 1, 2, \ldots, n$ 的 $\mathbf{x}_i$，$\mathbf{z}_4 \in D$；

…

(7)令 $\mathbf{z}_k$ 为满足 $\max(\min(d(\mathbf{x}_i, \mathbf{z}_j)))$，$i = 1, 2, \ldots, n$，$j = 1, 2, \ldots, k-1$ 的 $\mathbf{x}_i$，$\mathbf{z}_k \in D$；

(8)从这 k 个聚类中心出发，应用 k-means 聚类算法，得到聚类结果。

3 实验结果与分析

3.1 实验描述

为验证上述算法，选用 UCI 数据库上的 Iris、Balance scale、New-thyroid、Haberman、Wine 5 组数据作为测试数据。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，库中的数据都有确定的分类，因此可以用准确率直观地表示聚类的质量。

3.2 实验结果

用随机选择初始聚类中心的传统 k-means 算法和本文提出的优化初始聚类中心的 k-means 算法，得到表 1 中的结果。

3.3 实验分析

在 Iris、New-thyroid 和 Haberman 3 个数据集中，随机初始聚类中心的 k-means 算法，随初始聚类中心的不同，最高准确率和最低准确率之间的差值都在 20 个百分点以上，即使差别最小的数据集 Balance scale 也将近 6 个百分点，结果随初始输入的不同波动性比较大。应用改进后的算法，能得到一个稳定的结果，所选的 5 组数据中，4 组数据的准确率较平均准确率有明显提高，数据集 Wine 有所降低。

Wine 数据集有 178 个数据，分成 3 类，每个数据项有 14 个属性，各个属性的取值范围差距较大。应用决策树进行分析发现：Alcohol(属性 1)，Flavanoids(属性 8)，Color_intensity(属性 11)这 3 个属性对分类最为重要，它们的取值范围分别为：(0.34~5.08)，(1.28~13)，(11.03~14.83)，而对分类贡献不大的属性 Proline（属性 13）的取值范围是(278~1680)，在计算距离时，属性 Proline 起到了绝对的作用，从而导致用相互距离较远的点并不能很好地代表数据的实际分布情况。

4 结论

k-means 聚类算法是一种广泛应用的聚类算法，但是聚类结果随不同的聚类中心波动的特性影响了这种算法的应用范围。本文提出了一种选择 k-means 初始聚类中心的算法，实验表明，用这种方法得到的初始聚类中心用于 k-means 算法，能消除算法对初始聚类中心的敏感性，并能得到较好的聚类结果。

参考文献

1 朱　明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社, 2002: 138-139.

2 MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

3 Wang Wei, Yang Jiong, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining[C]//Proc. of the 23rd International Conference on Very Large Data Bases, 1997.

4 Agrawal R, Gehrke J, Gunopulcs D. Automatic Subspace Clustering of High Dimensional Data for Data Mining Application[C]//Proc. of ACM SIGMOD Intconfon Management on Data, Seattle, WA, 1998: 94-205.

5 Guha S, Rastogi R, Shim K. Cure: An Efficient Clustering Algorithm for Large Database[C]//Proc. of ACM-SIGMOND Int. Conf. Management on Data, Seattle, Washington, 1998: 73-84.

6 汤效琴, 戴汝源. 数据挖掘中聚类分析的技术方法[J]. 微计算机信息, 2003, 19(1).