一种改进的 K-means 聚类算法*

周爱武,崔丹丹,肖 云
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:K-means 算法是最常用的一种基于划分的聚类算法,但该算法需要事先指定 K 值、随机选择初始聚类中心等的缺陷,从而影响了 K-means 聚类结果的稳定性。针对 K-means 算法中的初始聚类中心是随机选择这一缺点进行改进,利用提出的新算法确定初始聚类中心,然后进行聚类,得出最终的聚类结果。实验证明,该改进算法比随机选择初始聚类中心的算法性能得到了提高,并且具有更高的准确性及稳定性。

关键词:欧氏距离;K-means;优化初始聚类中心
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1674-7720(2011)21-0017-03

An improved K-means clustering algorithm

Zhou Aiwu, Cui Dandan, Xiao Yun
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: K-means algorithm is one of the most commonly used clustering algorithm. But in actual application, there are some defects, for example, the value of K need to be specified ahead, and initial clustering center is a random choice and so on. This influences the performance of the K-menas algorithm. Aiming at the defect that the initial algorithm center of K-means is a random choice, this essay gives an improvement algorithm. Using this improved algorithm to comfirm clustering center to do clustering. After analysis, this improved algorithm makes the performance and accuracy better than the algorithm that random selection of initial clustering center.

Key words: Euclidean distance; K-means; optimization initial clustering center

聚类分析[1](clustering)是数据挖掘研究的重要领域,借助聚类分析将大量的数据对象聚成不同的类簇,使不同簇之间的相似度低,簇内的相似度高,它是一种无监督的学习算法。为了实现对数据对象的聚类,人们提出了不同的聚类算法。聚类算法主要分成基于划分、基于密度、基于分层、基于网格和基于模型的五大类[2]。K-means(均值)聚类算法是典型的基于划分的聚类算法,同时也是应用最广泛的一种聚类算法。K-means 聚类算法[3]主要针对处理大数据集,不但处理快速简单,而且算法具有高效性以及可伸缩性。但是 K-means 聚类算法存在 K 值需要事先指定、随机选择初始聚类中心等的局限性。人们针对 K-means 聚类算法的这些局限性提出了不同的改进算法。刘涛等人[4]提出了基于半监督学习的 K-means 聚类算法的研究,用粒子群算法以及迭代搜索的思想找到优质的聚类中心进行聚类;李飞等人[5]提出了基于遗传算法的全局搜索能力来解决初始聚类中心选择的敏感性问题。

K-means 聚类算法由于初始聚类中心是随机选择的,容易造成算法会陷入局部最优解甚至是无解的情况,而聚类结果的好坏直接取决于初始聚类中心的选择。因此初始聚类中心的选择十分重要。本文主要针对随机选择初始聚类中心这一缺点,提出了一种新的改进的 K-means 聚类算法。

1 传统的 K-means 聚类算法

K-means 聚类算法是解决聚类问题的一种经典算法,该算法具有简单、快速并且能够有效处理大数据集的特点。K-means 聚类算法首先从 n 个数据对象中任意选取 k 个对象作为初始聚类中心;而对于所剩下的其他对象,则根据它们与这些聚类中心的相似度(距离),分

* 基金项目:安徽省教育厅自然科学基金(KJ2009A57)

别将它们分配给与其最相似的类簇；然后计算该类簇中所有对象的均值；不断重复这一过程直到标准准则函数开始收敛为止。具体步骤如下[6]：

输入：k，data[n]；输出：k 个簇的集合，满足聚类准则函数收敛。

(1)任意选择 k 个对象作为初始中心点，例如 c[0]=data[0]，…c[$k-1$]=data[$k-1$]；

(2)根据簇中对象的均值，将每个对象指派给最相似的簇；

(3)更新簇均值，即计算每个簇中对象的均值；

(4)重复步骤(2)和步骤(3)，直到准则函数不再发生变化。

其准则函数定义如下：$E=\sum_{i=1}^{k}\sum_{p\in c_i}^{\infty}|p-\bar{x}_i|^2$，一般采用均方差作为标准准则函数。其中，$E$ 是指数据集中的对象与该对象所在簇的中心的平方误差的综合，E 越大说明对象与聚类中心的距离越大，簇内的相似性越低，反之则说明相似性越高；p 是簇内的一个对象；C_i 表示第 i 个簇；$\bar{x}_i$ 是簇 C_i 的中心，k 是簇的个数。

2 改进 K-means 算法描述

2.1 相关定义

定义 1 数据对象 $x_i(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$、$x_j(x_{j1},x_{j2},\cdots,x_{jp})^{\mathrm{T}}$ 之间的距离用欧氏距离 $d(x_i,x_j)$ 表示如下：

$$d(x_i,x_j)=\|x_j-x_i\|=\sqrt{\sum_{k=1}^{p}|x_{ik}-x_{jk}|^2} \tag{1}$$

定义 2 二维数据样本点中心 center(x_i,x_j)：

$$\mathrm{center}(x_i,x_j)=\left(\frac{x_{i1}+x_{j1}}{2},\frac{x_{i2}+x_{j2}}{2}\right) \tag{2}$$

2.2 改进算法的思想及基本步骤

影响 K-means 聚类算法性能的主要原因有：样本集中孤立点以及随机选择初始聚类中心而造成聚类结果的不稳定以及不准确。针对 K-means 的这种不足，本文提出了一种新的思想：首先将样本点中影响聚类结果的孤立点去除，然后利用坐标平移的思想来确定初始聚类中心，利用 K-means 算法进行聚类，最终得到可以满足平方误差准则函数收敛的聚类结果。

算法具体步骤：

首先排除样本点中的孤立点：

(1)输入样本点，利用 unique 函数排除样本点中重复的数据；

(2) 计算每个样本点与其余样本点之间的距离存入矩阵 *cid* 中；

(3)指定孤立点的个数 acnodenum，执行孤立点查找程序，即计算每个点与其余点的距离之和，找出距离最大的前 acnodenum 个点，即为孤立点；排除孤立点，将孤立点存入集合 acnode 中，并将这些点从原始数据集中删除得到新的数据集 datanew，即为本文算法第一次去除孤立点之后的样本点集合。在第一次去除了孤立点之后，可以得到新的样本点集合 datanew。

其次对 datanew 样本进行处理，从中找出 k 个初始聚类中心：

(4)求出样本点集合 datanew 中的两两之间的距离存入矩阵 $\boldsymbol{D}$ 中；

(5) 从矩阵 $\boldsymbol{D}$ 中找出距离最大的两个点 A 和 B，其最大距离记为 max$in D$，根据式(2)计算其中心 center 和半径(r=max$in D$/2)；

(6)第二次去除孤立点：求 datanew 中的每个样本点与 center 的距离，将大于 r 的样本点加入到集合 acnode 中并将其从 datanew 中去除得到第二次去除孤立点之后的样本点 datanewsec；

(7)利用坐标平移的思想求解初始聚类中心：

①将步骤(5)中求出的 A、B 中的任一点加入初始聚类中心集合 *nc* 中作为第一个初始聚类中心；

②循环 $k-1$ 次实现以 center 为参照点，将 A 坐标顺时针移动圆心角等于 $2\times pi/k$ 的度数；

③最终得到包含 A 在内的 k 个点，将这个 k 个点作为初始的聚类中心存入矩阵 *nc* 中；

(8)利用步骤(7)中求得的初始的聚类中心 *nc*，用 K-means 算法进行聚类得出满足聚类准则函数收敛的聚类结果。

(9) 计算 acnode 中的每个点与每个初始聚类中心的距离，将 acnode 中的点加入到距离初始聚类中心最近的簇中。

3 实验结果及分析

3.1 实验数据及实验环境

为了便于对比分析与计算，本实验采用的是二维数据，并且数据类型是数值型的。实验采用了两组测试数据：一组是随机数据，一组是 UCI 数据库中的标准数据集 Iris 数据集。实验工具采用 MATLAB 环境编程实现。

3.2 实验方案

3.2.1 采用随机数据

采用传统的随机选择初始聚类中心的 K-means 算法将本文的改进算法对随机产生的 80 个样本进行聚类，聚类的簇数设为 k=4，比较其聚类结果图。

传统 K-means 算法随机选取 4 组初始聚类中心对同一样本集进行聚类，其聚类结果图如图 1 所示。

第 1 组：(0.660 2，0.207 1)、(0.342 0，0.607 2)、(0.289 7，0.629 9)、(0.341 2，0.370 5)。

第 2 组：(0.767 6，0.274 6)、(0.261 0，0.193 1)、(0.719 7，0.827 6)、(0.315 8，0.620 6)。

第 3 组：(0.580 8，0.104 6)、(0.815 8，0.400 6)、(0.211 4，0.445 7)、(0.623 2，0.807 5)。

第 4 组：(0.568 1，0.846 9)、(0.781 2，0.575 2)、(0.211 4，0.445 7)、(0.628 6，0.122 5)。

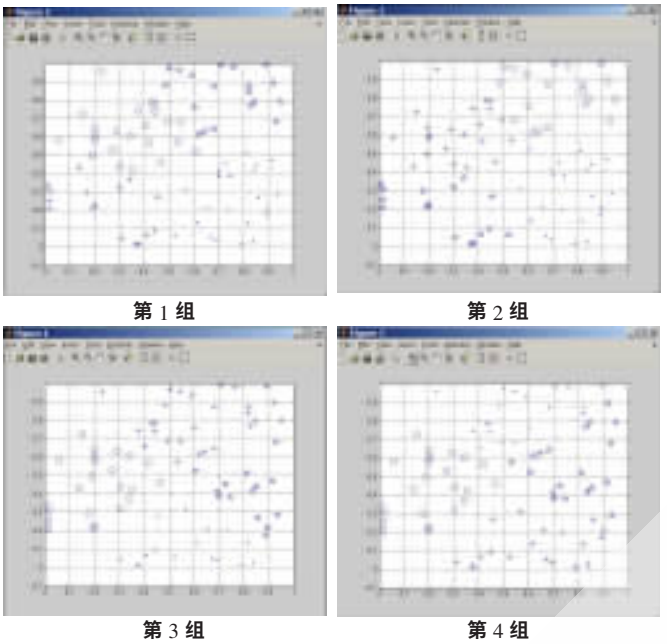

图 1 针对随机数据的选取四种初始聚类中心的结果图

采用改进算法选出的初始聚类中心为(0.231 1，0.956 8)、(0.999 6，0.795 7)、(0.838 5，0.027 2)和(0.070 0，0.188 3)，其聚类结果如图 2 所示。

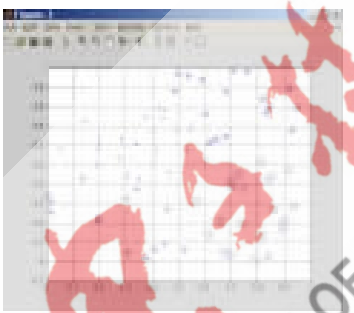

图 2 针对随机数据的改进算法聚类结果

由图 1、图 2 可以看出，利用本文改进算法选出的初始聚类中心进行聚类，其聚类结果比较接近数据分布。

3.2.2 采用 Iris 数据集

Iris 数据集是 UCI 数据库中的一个标准数据集，包含有 4 个属性，150 个数据对象，可分为 3 类。选用 Iris 数据集中间二维的数据进行聚类，分别用原算法和改进算法进行实验。对实验结果从运行时间以及准确度上进行分析，实验结果汇总以及分析如表 1 所示。

表 1 K-means 算法与本文改进算法准确率与运行时间比较

序号	数据个数	正确数据个数 传统算法	正确数据个数 改进算法	初始聚类中心 传统算法	初始聚类中心 改进算法	准确率/% 传统算法	准确率/% 改进算法	运行时间/s 传统算法	运行时间/s 改进算法
1	30	21	27	(3.6，1.4)，(3.9，1.7)(3.4，1.4)	(3.0，6.6)，(5.445，2.711 6)，(0.855 0，2.538 4)	70	90	0.281	0.172
2	60	50	55	(3.4，1.4)，(3.5，1.5)(3.4，1.6)	(4.0，1.2)，(0.481 8，4.868 8)(5.418 2，6.082 12)	83.3	91.67	0.344	0.321
3	90	80	80	(2.9，4.5)，(2.6，3.5)(3.3，6.0)	(3.6，1.0)，(0.295 2，4.992 0)(5.404 8，5.858 0)	88.89	88.89	0.484	0.435
4	120	100	114	(3.8，6.7)，(2.6，6.9)(2.2，5.0)	(3.2，1.2)，(1.508 1，4.563 4)(5.041 9，4.336 6)	83.3	95	0.640	0.589
5	150	125	137	(2.6，6.9)，(2.2，5.0)(3.2，5.7)	(3.6，1.0)，(1.264 8，4.351 8)(5.335 2，4.698 2)	83.3	91.33	0.781	0.692
平均	–	–	–	–	–	81.76	91.34	0.506	0.442

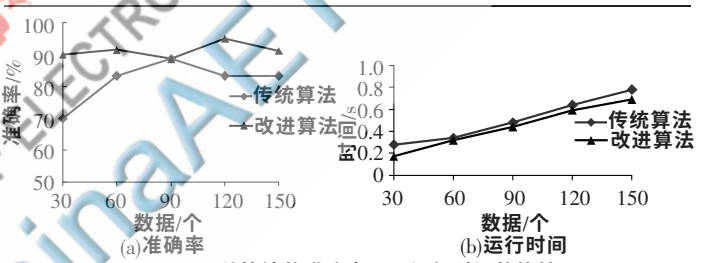

图 3 两种算法的准确率以及运行时间的比较

从表 1 可以看出，改进算法的运行时间比传统 K-means 算法的运行时间要小，尤其当数据集比较大时，其运行时间小得多。从图 3 中可以看出，采用改进算法其准确度明显提高。

本文提出的改进算法虽然在查找孤立点以及计算样本点之间的距离方面，会增加时间消耗，但是改进算法准确度较高，聚类效果较好。实验证明该算法是切实可行的，与传统的 K-means 算法相比较，有较好的聚类结果。

参考文献

[1] Han Jiawei, KAMBER M. Data mining concepts and techniques, second edition[M]. Elsevier(Singapore)Pte Ltd, 2006:251-263.

[2] 张建辉.K-means 聚类算法的研究与应用 [D]. 武汉：武汉理工大学，2007：10-14.

[3] 冯超.K-means 聚类算法的研究[D].大连：大连理工大学，2007：15-19.

[4] 刘涛，尹红健.基于半监督学习的 K-均值聚类算法的研究[J].计算机应用研究，2010，27(3)：913-917.

[5] 李飞，薛彬，黄亚楼.初始中心优化的 K-Means 聚类算法[J].计算机科学，2002，29(7)：94-96.

[6] Shi Na, Liu Xumin, Guan Yong. Research on k-means clustering algorithm [C]. Third International Symposium on Intelligent Information Technology and Security Informatics, 2010:63-67.

(收稿日期：2011-07-15)

作者简介：

周爱武，女，1965 年生，副教授，主要研究方向：数据库与 Web 技术，数据仓库与数据挖掘，信息系统安全。

崔丹丹，女，1986 年生，硕士，主要研究方向：数据库与 Web 技术，数据挖掘。

肖云，女，1985 年生，硕士，主要研究方向：数据库与 Web 技术，数据挖掘。