一种优化初始聚类中心的 K-means 聚类算法*

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

摘 要: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,利用提出的新算法选出初始聚类中心,并进行聚类。这种算法比随机选择初始聚类中心的算法性能有所提高,具有更高的准确性。

关键词:欧氏距离;K-means;优化初始中心
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1674-7720(2011)13-0001-03

An optimization initial clustering center of K-means clustering algorithm

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

Abstract: This article is an improvement aiming at the defect that the initial algorithm center of K-means is a random choice. Using this improved algorithm to select new clustering center to do clustering. After analysis, this new algorithm improves performance and accuracy better than the algorithm that random selection of initial clustering center.

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

数据挖掘技术研究不断深入与发展,作为数据挖掘技术中的聚类分析,也越来越被人们关注与研究。聚类分析是数据挖掘中一个非常活跃的研究领域,并且具有广泛的应用。聚类就是将数据集划分成若干簇或者类的一个过程[1]。经过聚类之后,使得同一簇中的数据对象相似度最大,而不同簇之间的相似度最小。

聚类是一种无监督的学习算法,即把数据对象聚成不同的类簇,从而使不同类之间的数据相似度低,而同一个类中的相似度高,并且将要划分的类是之前不知道的,其形成由数据驱动。聚类算法[1]分成基于划分的、密度的、分层的、网格的、模型的。其中基于划分的聚类算法中的 K-均值算法(K-means 算法)是最常用的一种聚类算法,同时也是应用最广泛的一种算法。K-means 聚类算法主要针对处理大数据集时[2],处理快速简单,并且算法具有高效性和可伸缩性。但是 K-means 算法也有一定的局限性[3],如 K 值必须事先给定,只能处理数值型数据,初始聚类的中心是随机选择的,而其聚类结果的好坏直接取决于初始聚类中心的选择。并且由于初始聚类中心随机选择,容易造成算法陷入局部最优解。因此初始聚类中心的选择十分重要。

本文针对随机选择初始聚类中心的缺点,提出了一种新的改进的 K-means 聚类算法。该算法产生的初始聚类中心不是随机的,能够很好地体现数据的分布情况,使得初始中心尽可能地趋向于比较密集的范围内,从而进行更好的聚类,最终消除了传统 K-means 算法中由于初始聚类中心选择是随机的而产生的缺点。最后实验证明了这种算法的有效性与可行性。

1 传统 K-means 算法

1.1 传统 K-means 算法的思想

传统的 K-means 算法的工作流程[1,4]:首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心;而对于所剩下其他对象,则根据它们与这些聚类中心的相似度(距离),分别将它们分配给与其最相似的(聚类中心所代表的)聚类;然后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值)。不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。其准则函数定义如下:$E=\sum_{i=1}^{K}\sum_{p\in C_i}|P-\overline{x_i}|^2$。其中,$E$ 为数据集中的对象与该对象所在簇中心的平方误

* 基金项目:安徽省教育厅重点项目(KJ2009A57)

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，$\overline{x_i}$ 是簇 C_i 的中心，k 是簇的个数。

传统的 K-means 算法具体描述如下[5]：

输入：k，data[n]；

输出：K 个簇的集合。

(1)任意选择 k 个对象作为初始中心点，例如 $c[0]$=data[0]，$\cdots c[k-1]$=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)、(3)，直到不再发生变化。

1.2 传统 K-means 算法的局限性

传统的 K-means 算法中对于 K 个中心点的选取是随机的[3]，而初始点选取的不同会导致不同的聚类结果。为了减少这种随机选取初始聚类中心而导致的聚类结果的不稳定性，本文提出了一种关于初始聚类中心选取的方法，用来改变这种不稳定性。

2 优化初始聚类中心的改进 K-means 算法

2.1 基本定义

设需要聚类的数据集：$\boldsymbol{X}=\{x_i|x_i\in R\ P,i=1,2,\cdots,n\}$，$k$ 个聚类中心分别用 $z_1,z_2,z_3,\cdots z_k$ 表示。有如下定义：

定义 1　2 个 p 维向量 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jp})^{\mathrm{T}}$ 数据对象间的距离用欧氏距离[6]表示：

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

定义 2　二维数据样本点中心 center(x_i,x_j)[6]：

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

定义 3　样本点之间的平均距离 Meandist：

$$\mathrm{Meandist}=\frac{\sum d(x_i,x_j)}{c_n^2}$$

即所有样本点的两两之间的距离之和除以样本点 n 的组合数。

2.2 改进算法流程

本算法的改进建立在没有离群点的数据集上，针对没有离群点的数据进行分析。

输入：样本点，初值 k。

输出：k 个簇的聚类结果，使平方误差准则最小。

步骤：

(1)求出两两样本点之间的距离存入矩阵 $\boldsymbol{D}$ 中。

(2)初始化集合 A 以及中心点集合 Center，最小距离的样本点放入集合 A 中，并求出其中心最为第一个初始的聚类中心 z_1。

(3)求出次小距离的样本点的中心，然后求出此中心与 z_1 之间的距离，与 Meandist 进行判断。如果小于 Meandist，则将此样本点加入 A 中，再求第三距离小的样本点，重复步骤(3)；如果大于 Meandist，则求出此中心存入 Center。

(4)Until 集合 Center 中的个数等于 k，初始聚类中心全部找到。

(5)用找到的初始聚类中心进行 K-means 聚类。

算法举例：

如图 1 所示，假设有 20 个点数据集，并且已经将孤立点排除，需要将其聚成 k=3 类。首先计算两两之间的距离，利用定义 2 求出 Meandist，并找出最小的距离(如图中的 x1、x2)；然后求出其中心，用红色表示；找出距离次小的距离(如图中的 x3、x4)，计算出 x3、x4 的中心，并加一步判断。如果这个中心与前面求出的一个聚类中心之间的距离小于 Meandist，那么就排除这个聚类中心，接着执行找第三小的距离，并求其中心，直到找到 K 个初始聚类中心为止；反之，则求下一个初始聚类中心，直到找到 k 个初始聚类中心为止。

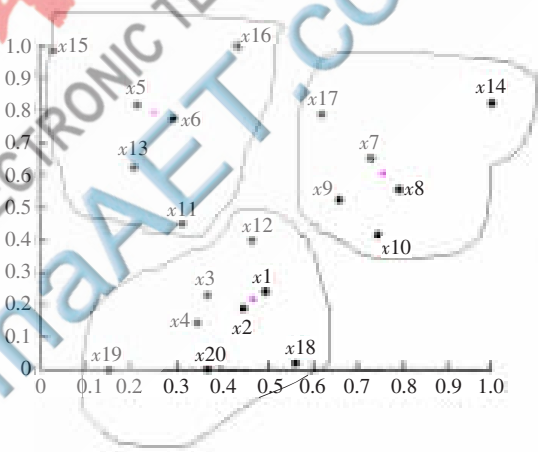

图 1　改进算法聚类举例

3 实验分析

为了便于分析与计算，本文采用的是二维数据，并且数据类型是实型的，实验环境为 MATLAB。为了进行对比，分别采用了传统的 K-means 算法与本文改进的 K-means 算法进行比较。

本文实验采用了两组实验进行验证，一组是随机数据，一组是标准数据库集。

(1)采用随机数据

本文用随机产生的 80 个样本分别采用传统的 K-means 算法进行聚类与本文的改进算法进行聚类，比较其聚类结果图。

传统算法采用随机选取初始聚类中心有 (0.950 1，0.794 8)、(0.231 1，0.956 8)、(0.606 8，0.522 6)，其聚类结果如图 2 所示。

采用改进算法的初始聚类中心有 (0.339 9、0.028 4)，(0.200 7，0.591 4)、(0.724 8，0.381 9)，其聚类结果如图 3 所示。

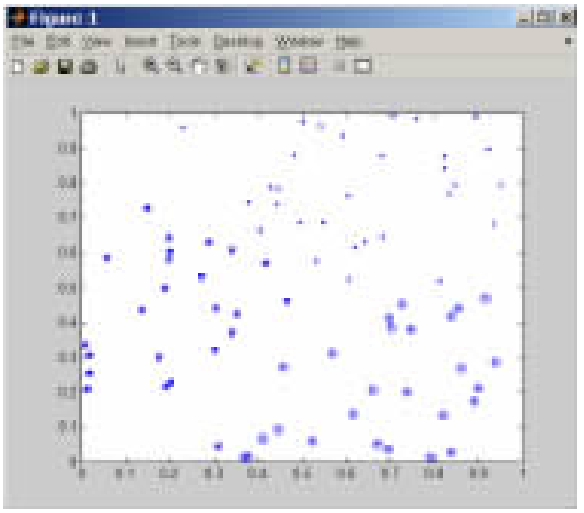

图 2 针对随机数据的传统的 K-means 聚类结果

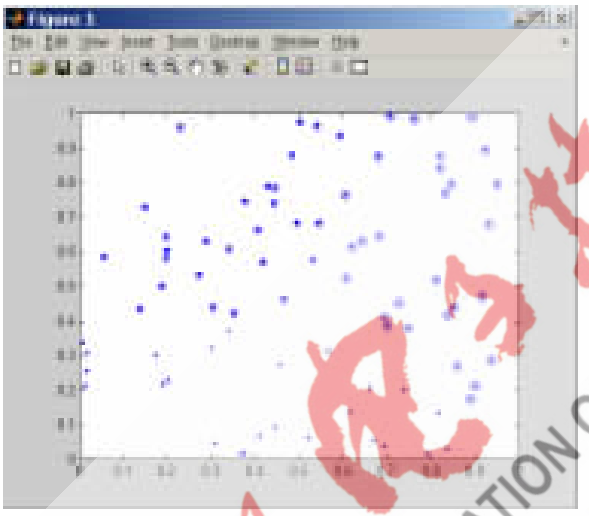

图 3 针对随机数据的改进算法聚类结果

(2)采用标准数据集:Iris 数据集

本文采用了 Iris 数据集,它是 UCI 数据库中的一个标准数据集。Iris 数据集包含有 4 个属性,150 个数据对象,可分为三类。选用 Iris 数据集前二维的数据进行聚类。分别用传统算法和改进算法进行聚类,其中分别用实心点、圈实心点以及五角星表示这三类。

传统算法采用随机选取初始聚类中心有(0.950 1,0.582 8),(0.231 1,0.423 5)、(0.606 8,0.515 5),其聚类结果如图 4 所示。

采用改进算法的初始聚类中心有(0.009 9,0.015 0)、(0.294 2,0.639 2)、(0.651 2,0.190 5),其聚类结果如图 5 所示。

对比这两幅图的聚类结果可以看出,采用改进算法产生聚类结果比较稳定准确。

运用 K-means 算法和本文改进算法针对随机数据和 Iris 数据分别实验得出的时间如表 1 所示。

K-means 算法是应用最为广泛的一种基于划分的算法,但是由于其初始中心的选择是随机的,从而影响了聚类结果,使得聚类结果不稳定。本文主要是针对传统 K-means 算法的这一缺点,提出了一种新的改进算法,即基于平均距离的思想,进行初始聚类中心的选择。实验证明,该算法是切实可行的,与传统的 K-means 算法比较,有较好的聚类结果以及较短的运行时间。但本文算法是基于先将噪声点排除掉之后应用此改进算法进行聚类、且是在点的分布比较均匀的前提下应用,才有良好的效果。如果对于具有噪声点的数据集有一定的局限性、而且是比较密集的点的情况下,这将在以后的学习研究中进行探讨。

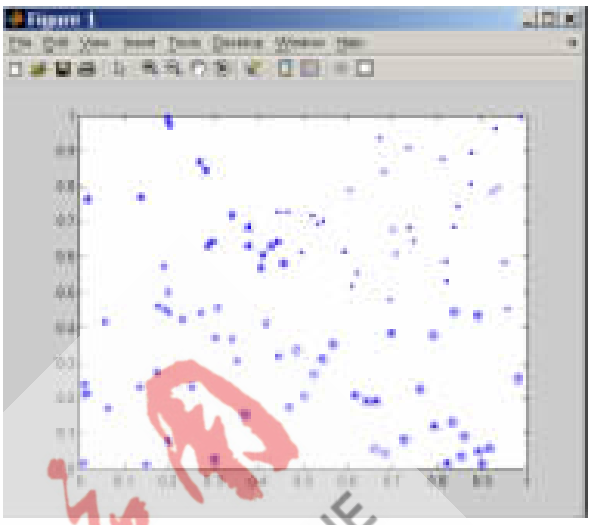

图 4 Iris 数据集传统 K-means 算法聚类结果

图 5 Iris 数据集改进算法聚类结果

表 1 两种算法不同数据集的执行时间比较

数据集	算法	执行时间/s
随机数据	K-means	1.484
	改进算法	0.516
Iris	K-means	7.423
	改进算法	2.168

参考文献

[1] HAN J, KAMBER M. 数据挖掘概念与技术[M].范明,孟小峰,等译.北京:机械工业出版社,2006.

[2] 孟海东,张玉英,宋飞燕.一种基于加权欧氏距离聚类方法的研究[J].计算机应用,2006,26(22):152-153.

[3] 包颖.基于划分的聚类算法研究与应用[D].大连:大连理工大学,2008:18-20.

[4] 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.

[5] 张玉芳,毛嘉莉,熊忠阳.一种改进的 K-means 算法[J].计算机应用,2003,23(8):31-33.

[6] 袁方,周志勇,宋鑫.初始聚类中心优化的 k-means 算法[J].计算机工程,2007,33(3):65-66.

(收稿日期:2011-03-13)

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,硕士生,主要研究方向:数据库与 Web 技术、数据挖掘。

潘勇,男,1985 年生,硕士生,主要研究方向:数据库与 Web 技术、数据挖掘。