基于密度的 K-means 聚类中心选取的优化算法*

周炜奔，石跃祥
（湘潭大学 信息工程学院 计算机科学与工程系，湖南 湘潭 411105）

摘　要：针对传统的 K-means 算法对于初始聚类中心点和聚类数的敏感问题，提出了一种优化初始聚类中心选取的算法。该算法针对数据对象的分布密度以及计算最近两点的垂直中点方法来确定 k 个初始聚类中心，再结合均衡化函数对聚类个数进行优化，以获得最优聚类。采用标准的 UCI 数据集进行实验对比，发现改进后的算法相比传统的算法有较高的准确率和稳定性。

关键词：K-均值；数据挖掘；聚类中心；垂直中点；密度
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)05-1726-03
doi:10.3969/j.issn.1001-3695.2012.05.033

Optimization algorithm of K-means clustering center of selection based on density

ZHOU Wei-ben, SHI Yue-xiang
(*Dept. of Computer Science & Engineering, College of Information Engineering, Xiangtan University, Xiangtan Hunan* 411105, *China*)

Abstract: Aiming at the problem of traditional K-means algorithm which is sensitive to initial clustering center and the number of cluster, this paper proposed a kind of optimization algorithm of initial clustering center selection. The algorithm was according to the distribution density of data and calculated the two vertical halfway points recently to determine the initial clustering center, then combined the equalization function to optimize the cluster number and got the optimal cluster. Used the standard UCI data sets as the contrast experiment objects, and found that the improved algorithm has the high accuracy and relative stability compared with traditional algorithm.

Key words: K-means; data mining; clustering center; vertical halfway point; density

数年来，聚类(clustering)已经成为数据挖掘领域中的一个研究热题[1]。它可以作为一个专门工具来发现数据分布信息，也可以作为数据挖掘算法的一个预处理步骤。到目前为止，已经提出了许多经典的聚类算法，如 K-means[2]、DBSCAN[3]、CURE[4]等，但总的来说，都存在着一定的问题。针对大规模、高维数据的聚类分析仍然是一个研究的热门课题。

在聚类算法中，K-means 算法的研究和应用相当广泛。为了提高 K-means 算法的性能，很多学者从不同方面对该算法进行了改进，尤其是针对 K-means 初始聚类中心点的选取问题。如王玲等人[5]提出一种密度敏感的相似度度量的方法；马帅等人[6]提出了一种基于参考点和密度的快速聚类算法，基本满足了聚类适应数据集分布的特征；袁方等人[7]提出了基于样本距离相似度和通过合适权值来初始化聚类的方法，对特定数据集采取合适权值进行聚类具有良好的效果。

K-means 算法是聚类分析中一种基本的划分方法，理论上可靠、算法简单、收敛速度快，能有效处理中小数据集而被广泛使用，但其结果对初始聚类中心敏感。因此本文提出了一种根据数据密度分布特征，引入高密度数据中最近两点的垂直中点作为初始聚类中心，从而很有代表性地反映了数据的分布特征，最后采用均衡化函数自动产生最优聚类个数作为最终结果。

1　K-means 算法

1967 年，MacQueen 提出了 K-means 算法。该算法的核心思想是在 n 个数据点中找出 k 个聚类中心 $c_1, c_2, \cdots, c_k$，使得每个数据 x_i 与其最近的聚类中心的 c_v 的平方距离和被最小化（该平方距离和被称为偏差 Δ），也即 Δ 收敛。算法具体描述如下：

输入：簇的数目 k 和包含 n 个记录的数据集。

输出：k 个簇，且 Δ 收敛或者最小。

a)初始化。随机指定 k 个聚类中心($c_1, c_2, \ldots, c_k$)。

b)分配 x_i。对每个样本 x_i 找到离它最近的聚类中心 c_v，将其分配到 c_v 所标明的类。

c)修正 c_v。通过不断计算簇的平均值，找更加合适的聚类中心 c_v。

d)计算偏差：

$$\Delta = \Sigma_{i=1}^{n} [\min_{r-1,\cdots,K} d(x_i, c_r)^2]$$

其中，$d(x_i, c_r)$ 的定义见 2.1 节。

e)Δ 收敛。如果收敛，则返回 $c_1, c_2, \cdots, c_k$ 并终止本算法；否则，返回步骤 b)。

收稿日期：2011-09-21；**修回日期**：2011-10-31　**基金项目**：湖南省教育厅创新平台开放基金资助项目(11K069)；湖南省自然科学基金资助项目(07JJ6115)；智能制造湖南省高校重点实验室资助项目(2009IM06)

作者简介：周炜奔(1986-)，男，硕士研究生，主要研究方向为数据仓库、数据挖掘(abenhao@126.com)；石跃祥(1964-)，男，教授，硕导，主要研究方向为图像处理、模式识别.

2　K-means 初始聚类中心改进算法

传统的 K-means 算法对初始聚类中心的选取比较敏感,因此选取的随机性就直接影响聚类结果的准确性和稳定性。本文针对这一缺点,提出了基于数据对象分布密度以及选取高密度数据样本中最近两点构成线段的中点作为初始聚类中心,即改进 K-means 算法当中的步骤 a)。

2.1　基本定义

定义 1　空间两点之间的距离公式:

$$d(i,j) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \quad (1)$$

定义 2　样本点之间的平均距离公式:

$$\text{avgDist} = \frac{1}{c_n^2} \times \sum d(i,j) \quad (2)$$

其中:c_n^2 是从 n 个点中取两个点的组合数,因此这里计算的是样本中所有点与点之间的距离。

定义 3　初始聚类中心

假设高密度数据样本中空间最近两点的坐标分别为(x_1, y_1, z_1)和(x_2, y_2, z_2),则聚类中心的坐标为

$$x = \frac{x_1 + x_2}{2}, y = \frac{y_1 + y_2}{2}, z = \frac{z_1 + z_2}{2} \quad (3)$$

定义 4　数据样本对象分布密度

已知样本数据总体 $D = \{x_1, x_2, \cdots, x_n\}$,其中样本 x_i 的密度可记为 $\text{dens}(x_i)$,式中 q 为密度参数$(q > 1)$,p_{ij}为连接数据点 x_i 和 x_j 之间的所有路径,l 为连接数据点 x_i 和 x_j 路径中数据的个数。

$$\text{dens}(x_i) = \sum_{j=1}^{n} \frac{1}{\min\limits_{p_{ij}} \sum\limits_{k=1}^{l-1} q^{d(x_k, x_{k+1})} - 1} \quad (4)$$

2.2　基于密度、中心点的初始化中心算法

本文算法的基本思想是首先计算样本数据集 D 中每个样本的密度,从而得到一个以密度为标准的样本集合 D'。然后在 D'集合基础上进行初始聚类中心的选取和簇的划分。每划分出一个簇,就从 D'集合中删除该簇所包括的数据点。改进算法的详细描述如下:

a)根据定义 4 计算数据集 D 中每个数据对象的密度,并进行密度点的排序,从而得到密度点的集合 D'。

b)选取当前密度最大的一个数据样本点并且通过定义 1 计算出与其最接近的样本点。

c)通过定义 3 计算这两点的线段中点,作为初始聚类中心 c_v,其中 $v \in (1, 2, \cdots, k)$。

d)以 c_v 为簇中心,以定义 2 中的 avgDist 为半径画取圆域$(r = \text{avgDist})$,则此圆域为一个划分,并且从集合 D'中删除该划分的数据点。

e)重复 b)~e),直到选取 k 个初始中心点。

f)划分 k 个簇以后,集合 D'中还有些没有划分给任何簇的数据点,则根据定义 1 计算这些点与各簇中心的距离,从而标明其属于哪个簇。

g)输出 k 个初始中心点,算法结束。

K-means 算法中,随机选取的聚类中心会产生不同的聚类结果且准确率不稳定。本文算法的目的是找到与数据在空间分布上相一致的初始聚类中心。对数据进行初始划分,以保证类内极大相似、类间相似度最小的原则。用实例来阐述算法,数据如图 1 所示。

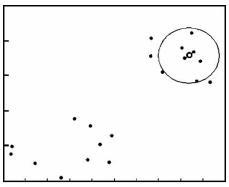

图1　某数据样本

根据本文提出的算法,首先计算每个数据点的密度,进行排序,构造新的集合 D',然后取得初始聚类中心,如该实例图中圆圈中的圆心(用空心点表示)表示当前密度最高点与其最近点构成的线段中点,这就是当前划分的一个初始聚类中心,并且图中圆的半径 $r = \text{avgDist}$。最后从集合 D'中删除圆圈内的点,然后重复以上步骤,直到选取 k 个初始聚类中心。

3　仿真实验分析

3.1　聚类均衡化评价函数

在本文算法的基础上,结合文献[8]中提到的均衡化评价函数作为聚类评价准则,运用评价并且优化聚类个数 k 的思想,对本文算法初始化的聚类进行评价和优化。

文献[9]已经证明,整数 k 的取值范围为 $1 \leqslant k \leqslant \sqrt{n}$。文献[8]算法从 $1 \sim \sqrt{n}$($\sqrt{n}$取整数)循环执行 K-means 聚类算法,并且记录均衡化评价函数及其对应的 k 值。最后选择函数最小值所对应的 k 值,即为最优聚类数目。

在这里只要将本文获取初始聚类中心的算法放在文献[8]中的步骤 1.1 即可通过调用均衡化评价函数的优化算法,获得最优聚类个数 k,同时也对奇异点和噪声点起到一定的抑制和消除作用。

3.2　实验结果

为了验证本文算法的有效性,选用实验平台 Windows XP, MATLAB 7.1 语言编程环境,Intel Pentium 3.00 GHz CPU, 2.0 GB内存,采用国际上专门用来测试机器学习、数据挖掘算法的 UCI 数据集中使用频繁的 Iris、Wine、Balance-scale 这三组数据集作为测试数据,分别测试五次,并且与传统的 K-means 算法进行效果比较,如表 1 所示。

表 1　与传统算法实验结果对比

算法	次数	数据集					
		Wine		Iris		Balance-scale	
		初始中心	准确率	初始中心	准确率	初始中心	准确率
传统算法	1	97,150,243	0.6372	49,56,19	0.6733	77,74,164	0.6458
	2	123,187,423	0.6214	51,121,65	0.6667	75,36,43	0.4826
	3	165,86,129	0.6152	52,129,96	0.5833	78,124,49	0.6134
	4	104,255,485	0.6565	54,118,89	0.5733	77,68,140	0.5633
	5	158,2352,287	0.5335	50,76,108	0.6333	70,72,17	0.5263
	—	—	—	—	—	—	—
传统算法平均		—	0.61276	—	0.62598	—	0.56628
本文算法		—	0.71353	—	0.86274	—	0.72858

注:选取多次实验中的五组,—代表省略余下实验组。

在实验过程中发现,两种算法的聚类个数划分相同,但传统的 K-means 算法,每一次测试其结果互不相同,准确率偏低,而采用本文基于密度改进的算法后,聚类中心是不会随测试次数而改变的,并且准确率也相对较高并且稳定,可见该算法可以取得良好的效果。

另外从实验结果来看,本算法在数据集 Wine 和 Balance-

scale 的测试效果优于文献[10]。

由于本文主要针对初始聚类中心进行优化选取,于是和文献[11]在同一测试数据集 Iris 上进行了聚类中心误差对比,对比结果如表 2 所示。

表 2 聚类中心误差对比

对比来源	初始化聚类中心	Iris 数据集原始中心	标准误差和
文献[11]	5.0154,3.3352,1.6158,0.3314		
改进	5.8891,2.7507,4.3582,1.4079	—	0.2157
算法	6.6568,2.9891,5.5481,1.9734		
本文	5.0082,3.1875,1.5362,0.2371	5.00,3.42,1.46,0.24	
改进	5.7886,2.7829,4.2287,1.3652	5.93,2.77,4.26,1.32	0.1776
算法	6.6059,2.9107,5.5562,1.9785	6.58,2.97,5.55,2.02	

由表 2 可知,本文提出的算法在选取聚类中心上比文献[11]更优。

3.3 实验结果分析

通过上面的实验结果可知,本文算法能够很好地反映数据的分布情况。采取基于密度的方法计算各数据点的密度值,这能良好地反映数据的大体分布状况;而本文选取基于密度的中点作为初始聚类的中心,非常具有代表性,不仅降低了复杂的聚类中心替换工作,也使得选取的初始聚类中心和实际聚类中心误差很小,可见该改进算法取得了良好的效果。

4 结束语

K-means 算法是一种被广泛应用的算法,但是它的聚类结果会随预先给定的聚类数目以及初始聚类中心的不同而波动,准确率相对偏低。本文算法在很大程度上消除了 K-means 算法对初始聚类中心的敏感性和随机选择性,克服了实际应用中 k 值的难易确定性;采用基于密度的方法,客观地反映了数据的分布状况,在此基础上选取中点作为初始聚类中心,并且通过均衡化评价函数对聚类个数进行优化,实验也表明该算法具有良好的综合性能。但还需要作进一步的研究:如利用该算法对高维数据集进行分析。同时,在优化 K-means 算法时,降低时间复杂度以及对海量数据进行分析将是今后研究的内容。

参考文献:

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[2] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. New York:Wiley,1990.

[3] ESTER M, KRIEGEL H P, SANDER J,*et al*. A density based algorithm for disco-vering clusters in large spatial databases with no ise [C]//Proc of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press,1996:226-231.

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

[5] 王玲,薄列峰,焦李成. 密度敏感的普聚类[J]. 电子学报,2007,35(8):1577-158.

[6] 马帅,王腾蛟,唐世渭,等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报,2003,14(6):1089-1095.

[7] 袁方,孟增辉,于戈. 对 K-means 聚类算法的改进[J]. 计算机工程与应用,2004,40(36):177-178,232.

[8] 汪中,刘贵全,陈恩红. 一种优化初始中心点的 K-means 算法[J]. 模式识别与人工智能,2009,22(2):299-304.

[9] 杨善林,李永森,胡笑旋,等. K-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践,2006,26(2):97-101.

[10] 姚跃华,史秀岭. 一种优化初始中心的 K-means 粗糙聚类算法[J]. 计算机工程与应用,2010,46(34):126-128.

[11] 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.