一种优化初始中心的 K-means 聚类算法

邓 海,覃 华,孙 欣
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:针对传统 K-means 聚类算法对初始聚类中心的敏感性和随机性,造成容易陷入局部最优解和聚类结果波动性大的问题,结合密度法和最大化最小距离的思想,提出基于最近高密度点间的垂直中心点优化初始聚类中心的 K-means 聚类算法。该算法选取相互间距离最大的 K 对高密度点,并以这 K 对高密度点的均值作为聚类的初始中心,再进行 K-means 聚类。实验结果表明,该算法有效排除样本中含有的孤立点,并且聚类过程收敛速度快,聚类结果有更好的准确性和稳定性。

关键词:K-means 聚类;聚类中心;高密度点;垂直中心点
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2013)11-0042-04
doi:10.3969/j.issn.1673-629X.2013.11.011

A K-means Clustering Algorithm of Meliorated Initial Center

DENG Hai, QIN Hua, SUN Xin
(College of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract: The traditional K-means clustering algorithm has the sensitivity and randomness for initial clustering center. So it easily falls into local optimal solution and has unstable results. To solve the problem, proposed a K-means algorithm of meliorated initial clustering center based on vertical center point of the closest high density points. This algorithm selects K pairs of high density points that have the maximal distance between each other, and then uses the average values of K pairs of high density points as the initial clustering centers to implement the traditional K-means. The experimental results show that this algorithm is effective to eliminate isolated points and has better accuracy and stability.

Key words: K-means clustering; clustering center; high density points; vertical center

0 引 言

数据挖掘就是从大量的、不完全的、有噪声的、模糊的数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。聚类分析则是数据挖掘领域中重要的研究课题,无需任何先验知识,只需按照某种相似程度的度量,把数据集划分为若干组,使得同一组内数据对象具有较高的相似度,而不同组中的数据对象则相似度低[1]。

目前,研究者已经提出了许多聚类算法。常用的有基于划分的 K-means 算法[2]、基于密度的 DBSCAN 算法[3]、基于图论的 CHAMELEON 算法[4]、基于网格的 STING 方法[5]等等。K-means 聚类算法简单、收敛速度快,能有效处理中小数据集,但是该算法对初始的聚类中心点有依赖性和敏感性,从而导致聚类结果具有波动性。因此文中提出了一种基于最近高密度点间的中心点优化初始聚类中心的 K-means 聚类算法。在 UCI 数据集和文本数据集进行聚类实验,实验结果表明,该算法有效排除样本中含有的孤立点和噪声点,并且聚类过程收敛速度快,聚类结果相对传统 K-means 聚类有更好的准确性和稳定性。

1 K-means 聚类算法

1967 年,MacQueen 提出了 K-means 聚类算法[1]。该算法的核心思想是首先从输入的 n 个待聚类数据集对象中随机选取 K 个对象作为初始的聚类中心;然后,将其余对象根据其与聚类中心的距离分配到最近的簇中;再从新形成的簇中按照均值求出新的聚类中心;上述过程不断地迭代,直到目标函数收敛或者新形

收稿日期:2013-01-16 **修回日期**:2013-04-23 **网络出版时间**:2013-07-24
基金项目:教育部人文社会科学研究规划基金项目(11YJAZH080)
作者简介:邓 海(1986-),男,硕士研究生,研究方向为数据挖掘及电子商务技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130724.0945.012.html

成的聚类中心与前一次聚类中心接近,以达到聚类终止,输出各类对象。

设 x 表示簇 C_i 中的数据对象,c_i 表示簇 C_i 的中心点(均值),通常所用的目标函数为:

$$J_e = \sum_{i=1}^{k} \sum_{i \in C_i} \| x_i - c_i \|^2 \quad (1)$$

其中,J_e 是输入数据对象与它所在的簇的中心点的平方误差总和。

K-means 算法的具体实现步骤如下:

Input:包含 n 个对象的数据集 D 以及聚类数目 K。

Output:满足目标函数收敛的 K 个簇。

Step1:从数据集 D 中随机选取 K 个数据点作为初始聚类中心;

Step2:计算各对象与每个簇中心的距离,将每个对象赋给距离最近的簇;

Step3:重新计算簇 C_i 中的中心点;

Step4:不断重复执行 Step2 和 Step3,直到公式(1)目标函数收敛或者中心点不再发生变化。

虽然 K-means 聚类算法的思路简单,易于实现,并且当结果簇是密集的而且簇之间的区分明显时,它的效果较好,但是该经典的聚类算法也存在不足之处。K-means 算法要事先给出簇类个数,对初始聚类中心敏感,受噪声和孤立点的影响。另外,该算法采用迭代更新的方法,所以当初始中心选择在局部最小附近时,容易陷入局部最优值。

针对初始聚类中心的选择,国外有研究者提出了一些选取中心的方法[6-7],国内的张玉芳等人[8]提出基于取样的划分思想选取不同的聚类中心;孙可等人[9]提出了基于密度和最近邻相似度的初始质心选择算法等等。这些研究通过不同的方面在对 K-means 聚类初始中心选择上进行改进。文中结合密度法和最大化最小距离的思想,首先选取相互间距离最大的 K 对高密度点,并以这 K 对高密度点的均值作为聚类的初始中心,以便在一定范围扩大中心点的搜索范围,达到有效选取中心点的目的。

2　优化初始中心的 K-means 算法

2.1　基于密度思想的相关概念

一般在一个数据空间中,高密度点数据对象区域被低密度的对象区域所分割着,通常这些低密度区域的点为噪声点或孤立点。当进行聚类分析时,如果初始聚类中心选择到这些点,就会造成聚类结果不理想。为了排除噪声点的干扰,需要把初始聚类中心的选取集中到高密度的区域。

定义 1　高密度点:以数据集中的对象 x_i 为中心,在邻域半径 δ 内包含的样本个数不少于常数 Minpts 的点。

定义 1 中,邻域半径 δ 的取值综合输入数据集,将其取为 n 个样本间距离均方根的一半,采用以下公式:

$$\delta = \frac{1}{2}\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\| x_i - x_j \|^2} \quad (2)$$

通过引入高密度点,就可以有效排除数据样本中存在的噪声对象和孤立点对象,为 K-means 算法中选取初始中心点提供了一种方法。

2.2　最大化最小距离的思想

取尽可能离得远的数据对象作为初始聚类中心,从而避免初值选取时可能出现的聚类中心过于邻近的情况,努力得到数据集一个比较好的初始划分。因此在欧氏距离作为相似性度量的 K-means 算法中,取相互间距离最远的 K 个数据样本点比随机选取最初的 K 个样本点更具有代表性。 最大化最小距离的算法描述如下:

a)设有 n 个对象,$D_n = \{x_1, x_2, \cdots, x_n\}$,从中任取一个对象,如 x_1,作为第一个簇中心,则有 $Z_1 = x_1$;

b)在集合 D_n 中找出距离 Z_1 最远的对象作为第二个簇中心 Z_2;

c)找出集合中剩余的每个对象到已有簇中心最近的那个距离,即:对 D_n 中剩余的每一个对象 x_i,都分别计算到两个簇中心 Z_1 和 Z_2 的距离,记为 d_{i1} 和 d_{i2},并令其中较小值为 $\min(d_{i1}, d_{i2})$;

d) 计算 $\min(d_{i1}, d_{i2})$ 的最大值, 记为 $\max(\min(d_{i1}, d_{i2}))$,对应的那个对象记为 x_j;

e) 若 $\max(\min(d_{i1}, d_{i2})) > m \times | Z_2 - Z_1 |$,则取 x_j 作为第三个簇中心,其中 m 为算法中的检验参数,一般情况下取($\frac{1}{2} \leqslant m < 1$);

f) 再比较剩余的其他点,用同样的计算方法找到 $\max(\min(d_{i1}, d_{i2}, d_{i3}))$ 的对象;

g) 如果满足 $\max(\min(d_{i1}, d_{i2}, d_{i3})) > m \times [\text{average}(|Z_2 - Z_1|, |Z_3 - Z_2|)]$,则将该数据对象作为新的簇中心,重复第(f)步,直到找不到新的符合条件的簇中心,算法终止。

该方法依赖于初始点的选取以及检验参数 m,在没有样本分布的先验知识的前提下,算法只有通过多次试探的方式去得到最优的参数值,容易降低算法的收敛速度。由此在 K-means 聚类算法选取初始聚类中心时可以结合密度法以及最大化最小距离思想,可以把初始点的选取集中到高密度点中,排除密度区域比较小的点。为了在一定范围内扩大初始中心点存在的区域,现在考虑相邻最近高密度点的垂直中心点作为初始聚类中心,这样有利于高密度区域能够相互融合,提高聚类的效果。选取中心点的模型图如图 1 所

示。

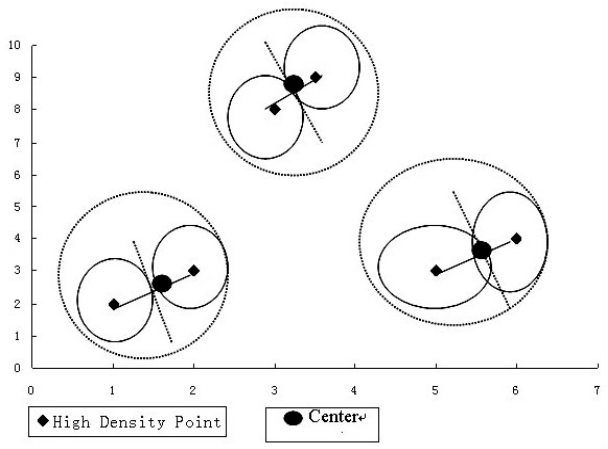

图 1 选取中心点的模型图

图中内圆为高密度点区域,外圆为最近相邻的高密度点区域融合的虚拟区域,实心黑点即为通过求解垂直中心点的初始中心。

2.3 优化初始聚类中心的 K-means 算法

通过前面的理论分析及选取中心点的模型图,得到基于最近高密度点间的中心点优化初始聚类中心的 K-means 聚类算法,该算法的描述如下:

Input:包含 n 个对象的数据集 D 以及聚类数目 K。

Output:满足目标函数收敛的 K 个簇。

Step1:for 1 to n, 计算每个样本点 x_i 与其余样本点之间的距离 $\mathrm{dis}(x_i,x_j)$,根据数据集求解邻域半径 δ ,统计到 x_i 距离 $< \delta$ 的样本个数,若个数 > min pts,则把该点定义为高密度点,得到高密度点集合 HP;

Step2:以集合 HP 中密度最大点 z_1 及其最近相邻的高密度点 z_1' 为一对,加入到高密度点对集合 C 中;

Step3:按最大化最小距离思想计算 z_2 及其最近相邻的高密度点 z_2' ,并加入到 C 中,直到 C 中高密度点对数目为 K;

Step4:计算高密度点对集合 C 中 K 个中心 $\bar{z}=(z_i,z'_i)/2$;

Step5:从 Step4 中 K 个中心点出发,执行 K-means 算法,得到聚类结果。

3 实验数据准备及实验

3.1 实验数据准备

为了验证改进算法的有效性,文中选择不同类别的数据集作为测试数据集。

首先是用专门用于数据挖掘算法的 UCI 数据集 Iris,它是 150 个关于三种花的生物统计数据,并且知道该数据集的实际聚类中心,有利于与文中算法得到的聚类中心做比较。

其次是中文文本数据集,主要来自中文文本分类语料库 TanCorp-12[10],预处理采用中科院分词工具 ICTCLAS 进行分词,去除停用词,并去掉数字和标点符号。文本表示采用向量空间模型(VSM)[11],并利用 TF-IDF 公式进行词项权重计算,最终形成文本数据集,TF-IDF 公式如下:

$$W(t,d)=\frac{\mathrm{tf}(t,d)\log\left(\frac{n}{n_t}+0.01\right)}{\sqrt{\sum_t(\mathrm{tf}(t,d))^2\times\log^2\left(\frac{n}{n_t}+0.01\right)}} \tag{3}$$

式中,$W(t,d)$ 为表示词 t 在文本 d 中的权重;$tf(t,d)$ 为词 t 在文本 d 中出现的词频;n 则表示全部的文本总数;n_t 表示包含词 t 的文本数;分母为归一化因子。

3.2 实 验

(1)Iris 数据集包含 4 个属性,150 个数据对象,可分为三类。对 Iris 数据进行传统 K-means 算法聚类和文中算法聚类结果如表 1 所示。

表 1 K-means 算法和改进算法聚类结果

聚类算法	Iris 数据集		
	初始中心	准确率/%	迭代次数
K-means 算法	(12,11,66)	51.33	3
	(1,10,25)	57.33	6
	(133,116,88)	89.33	6
	(17,61,126)	88.67	9
	(76,60,9)	88	10
	---	---	---
平均值	/	74.93	6.8
文中优化算法	[(127,124)/2, (14,39)/2, 126,130)/2]	90	4

另外文献[12]给出了 Iris 数据集的实际中心,结合该实验最终的聚类中心,得到以下误差平方和对比表,如表 2 所示。

实验分析:由表 1 中可以看出传统 K-means 算法随着不同的初始中心出现波动性比较大,而经过优化初始中心的 K-means 聚类算法能获得更高的准确率,并且在迭代次数上面也得到改进。而表 2 则显示了优化的初始中心算法得到的最终聚类中心更接近实际数据的原始中心,并且能够保持比较高的准确率。

(2)文中实验从中文文本分类语料库中随机选取人才、体育和娱乐三类数据各 50 篇文档,利用 VSM 构造文本数据集。文本对象具有维数高且稀疏的特点,容易受噪声点的影响,直接使用传统 K-means 聚类效果不佳。文献[13]提出了使用主成分分析(PCA)的降维法对高维数据降维后再用 K-means 聚类, 聚类效

果有所改善。

由此在利用文中算法进行文本聚类实验时,可以对文本数据进行 PCA 降维的预处理,最终实验结果如表 3 所示。

表 2　聚类中心误差对比

实验中各聚类中心	Iris 数据集原始中心	误差平方和
K-means 算法中心	5.00 3.42 1.46 0.24	10.785 1
	5.93 2.77 4.26 1.32	9.312 3
	6.58 2.97 5.55 2.02	0.155 3
		0.147 6
		0.147 6
平均值	----	4.109 6
文中优化算法中心	5.90 2.74 4.39 1.43	
	5.01 3.42 1.46 0.24	0.155 3
	6.85 3.07 5.74 2.07	

表 3　K-means 算法和改进算法的聚类结果

聚类算法	文本数据集		
	初始中心	准确率/%	迭代次数
K-means 算法	(53,66,145)	51.33	3
	(76,22,81)	39.33	1
	(104,106,61)	50.67	4
	(37,28,32)	75.33	11
	(82,23,62)	42.67	4
	----	----	----
平均值	/	51.86	6.8
文中优化算法	[(128,139)/2, (14,46)/2, (80,70)/2]	74	3

实验分析:从表 3 可以看出,文本聚类容易受孤立点影响,导致准确率波动很大;而文中算法能够排除孤立点干扰,从而得到稳定的准确率,并且平均迭代次数少于传统 K-means 算法。

4　结束语

K-means 算法是一种普遍使用的聚类算法,但容易受初始聚类中心的影响。

文中结合密度法和最大化最小距离的思想,提出基于最近高密度点间的垂直中心点来优化初始聚类中心的 K-means 聚类算法。在 UCI 数据集和文本数据集进行聚类实验,实验结果表明新算法排除了噪声点的干扰和避免陷入局部极值的情况,相对于传统的 K-means 算法能够获得更好的聚类效果。

参考文献:

[1] Han Jiawei, Kamber M. Data mining concepts and techniques [M]. 2nd ed. Beijing: China Machine Press, 2006.

[2] MacQueen J B. Some methods for clustering and analysis of multivariate observations[C]//Proc of 5th Berkeley Symp on Math Statist Prob. Berkeley: University of California Press, 1967:281-297.

[3] Ester M, Kriegel H P, Sander J, et al. A density based algorithm for discovering clusters in large spatial databases with noise [C]//Proc of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996:226-231.

[4] Karypis G, Han E H, Kumar V. CHANELEON: a hierarchical clustering algorithm using dynamic modeling[J]. IEEE Computer, 1999, 2(8):68-75.

[5] Wang W, Yang J, Muntz R. STING: a statistical information grid approach to spatial data mining[C]//Proceedings of the International Conference on Very Large Data Bases. [s. l.]: [s. n.], 1997:186-195.

[6] Khan S, Ahmad A. Cluster centre initialization algorithm for K-means clustering[J]. Pattern Recognition Lett, 2004, 25: 1293-1302.

[7] Redmond S J, Heneghan C. A method for initializing the K-means clustering algorithm using kd-trees[J]. Pattern Recognition Lett, 2007, 28:965-973.

[8] 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.

[9] 孙　可,刘　杰,王学颖. K-均值聚类算法初始质心选择的改进[J]. 沈阳师范大学学报(自然科学版),2009,27(4):448-450.

[10] Tan Songbo, Wang Yuefen. Chinese text classification Corpus-TanCorpV1.0 [DB/OL]. 2012. http://www.searchforum.org.cn/tansongbo/corpus.htm.

[11] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(5):613-620.

[12] Duda R O, Hart P E. Pattern Classification and Scene Analysis [M]. [s. l.]: Wiley and Son, 1973.

[13] Ding Chris, He Xiaofeng. K-means clustering via principal component analysis[C]//Proceedings of the 21st International Conference on Machine Learning. New York: ACM, 2004:225-232.

一种优化初始中心的K-means聚类算法

作者： 邓海， 覃华， 孙欣， DENG Hai， QIN Hua， SUN Xin
作者单位： 广西大学 计算机与电子信息学院,广西 南宁,530004
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013(11)