DOI:10.11721/cqnuj20160127

一种改进的 K-means 动态聚类算法*

张 阳[1], 何 丽[2], 朱颢东[3]
(1. 郑州轻工业学院 数学与信息科学学院, 郑州 450002;2. 重庆师范大学 计算机与信息科学学院, 重庆 401331;
3. 郑州轻工业学院 计算机与通信工程学院, 郑州 450002)

摘要:传统的 K-means 算法通过不断的重复计算来完成聚类,聚类中心点的不断变化产生的一些动态变化信息将对聚类产生一定的干扰,且当数据量过大时,算法的时间开销和系统的 I/O 开销将大大增加,这严重影响了算法的性能。为此,论文提出一种改进的 K-means 动态聚类算法,该算法充分考虑了 K-means 聚类过程中信息的动态变化,通过为算法的终止条件设定标准值,来减少算法迭代次数,减少学习时间;通过删除由信息动态变化而产生的冗余信息,来减少动态聚类过程中的干扰,使算法达到更准确更高效的聚类效果。实验结果表明,当数据量较大时,相比于传统的 K-means 算法,改进后的 K-means 算法在准确率和执行效率上都有较大的提升。

关键词:K-means;聚类分析;数据挖掘;动态聚类
中图分类号:TP301 **文献标志码**:A **文章编号**:1672-6693(2016)01-0097-05

在数据大爆炸,“知识太贫乏”的时代,数据挖掘应运而生,聚类分析作为数据挖掘的重要分支,在信息化时代起着举足轻重的作用。聚类分析的目标在于将数据集分成若干个簇,并保证同一簇内的数据点相似度尽可能大,簇与簇之间数据点的相似度尽可能小。聚类操作是对事先未知的数据对象进行类的划分,而类的形成是由数据驱动来完成。在数据挖掘领域中,聚类分析既可以作为数据挖掘过程中的一个环节,又可以作为获取数据分布情况的工具。聚类分析的应用前景较为广泛,比如:生物种群的划分、目标客户的定位、市场趋势分析、模式识别及图像处理等。

K-means 算法是聚类分析最常用的方法之一,最早由 MacQueen 提出[1],该算法的精妙之处在于简单、效率高且宜于处理大规模的数据,已经被应用到众多领域,包括:自然语言处理、天文、海洋、土壤等。自 K-means 算法提出以来,大量有关 K-means 算法的研究如雨后春笋般涌现,算法的弊端纷纷暴露出来,主要包括以下 4 点:第一,必须事先确定 K 值;第二,聚类结果会受到初始聚类中心影响;第三,处理分类属性数据较为困难且易产生局部最优解;第四,当数据量过大时,不仅使算法的时间开销非常大,且由聚类的动态变化导致的冗余信息也将对算法产生影响。针对以上 K-means 算法的不足,国内外学者提出众多的解决方法:有的提出基于密度的改进 K 均值算法,该算法针对由初始中心点的随机产生导致的聚类结果的不稳定提出了改进算法[2];有的提出基于密度和最邻近的 K-means 文本聚类算法[3];有的提出聚类模式下一种优化的 K-means 文本特征选择算法,该算法针对 K-means 算法对类中心点初始值机孤立点过于敏感的问题提出的一种改进算法[4];有的提出基于信息熵的精确属性赋权 K-means 聚类算法[5];还有提出一种基于余弦值和 K-means 的植物叶片识别方法[6]。

本文针对 K-means 算法的第四点不足,提出 K-means 动态聚类方法,该算法充分考虑了 K-means 聚类过程中信息的动态变化,通过为算法的终止条件设定标准值,来减少算法迭代次数,减少学习时间;通过删除由信息动态变化而产生的冗余信息,来减少动态聚类过程中的干扰,使算法达到更准确更高效的聚类效果。

1 相关知识

1.1 相关定义

假设 $A=\{a_i \mid a_i \in \mathbf{R}^m, i=1,2,\cdots,n\}$ 为给定的数据集,$T_i(i=1,2,\cdots,k)$ 代表 k 个类别,$c(T_1)$,$c(T_2)$,$\cdots$,

* **收稿日期**:2015-02-11 **修回日期**:2015-04-22 **网络出版时间**:2015-12-02 13:26
资助项目:河南省科技攻关项目(No. 122102210024;No. 102102210544);国家自然科学基金(No. 61201447)
作者简介:张阳,讲师,研究方向为数据库、信息安全和信息处理,E-mail:zhangyang@zzuli.edu.cn
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20151202.1326.012.html

$c(T_k)$分别是 k 个聚类中心。有如下定义：

定义 1 设向量 $a_i=(a_{i1},a_{i2},\cdots,a_{im})$和向量 $a_j=(a_{j1},a_{j2},\cdots,a_{jm})$分别代表两个数据对象，那么它们之间的欧式距离定义为：

$$d(a_i,a_j)=\sqrt{\sum_{d=1}^{m}(a_{id}-a_{jd})^2}。\tag{1}$$

定义 2 同一类别的数据对象的质心点定义为：

$$c(T_i)=\frac{1}{|T_i|}\sum_{a_j\in T_i}a_j,\tag{2}$$

其中$|T_i|$是 T_i 中数据对象的个数。

1.2 K-means 算法

K-means 算法是一种基于样本间相似性度量的间接聚类方法，也被称为 K-均值算法。算法的主要思想是通过迭代的过程把数据集划分为不同的类别，使得评价聚类性能的准则函数达到最优[7]。算法描述如下：

输入：簇的个数 K 与包含 n 个对象的数据集合。

输出：k 个簇，使得准则函数值满足条件。

步骤 1 为每个聚类确定一个初始聚类中心点；

步骤 2 将数据集中的数据按照欧氏距离原则分配到最邻近簇；

步骤 3 使用每个簇中的样本数据均值作为新的聚类中心；

步骤 4 重复步骤 Step2 与 Step3 直至算法收敛；

步骤 5 结束，得到 K 个结果簇。

2 K-means 动态聚类算法

2.1 K-means 算法存在的缺点

在样本数据聚类的过程中，不仅需要计算每个聚类对象与它们中心对象的距离，还需要重新计算中心对象发生变化的聚类的均值，且计算是在一次次迭代中重复完成，当数据样本较多时，过大的计算量会严重影响算法的性能[8]。其次，由于 K-means 聚类是个动态变化的过程，聚类的过程中将产生一些冗余信息，会对聚类产生一些不必要的干扰。

2.2 改进的 K-means 算法

针对 K-means 算法的以上缺陷，提出两点优化原则：1)减少聚类过程中的迭代次数；2)减少聚类过程中的数据量。K-means 动态聚类算法的基本思想：由于 K-means 算法是通过迭代的过程把数据集划分为不同的类别，现为中心点的该变量设定一个值 σ_1，σ_1 的初始值为 0，在迭代的过程中获得 σ_1 值，过程如下：计算用新中心点 O_i 代替原中心点 O_j 所造成的绝对误差，公式：

$$E=\sum_1^k\sum_{p\in C_j}|p-O_j|,\tag{3}$$

其中，p 为空间中的数据点，O_j 为簇 C_j 的中心点。

计算 O_i 造成的绝对误差与原中心点 O_j 造成的绝对误差的差值，计算公式为：

$$e=E_i-E_j。\tag{4}$$

当 $e<0$ 时，此时中心的的改变量就为设定的 σ_1。一旦获得 σ_1 值，该值就不再改变。

在迭代的过程中，当中心点的改变量小于 σ_1 时，将整个簇加入到已选数据集，并将其从样本集中删除，使得原始样本数据集中只保留未被正确识别的样本。由定义 2 求得中心点的改变量计算公式为：

$$\sigma_r=\frac{1}{|T_i|}\sum_{a_i\in T_{r,j}}a_i-\frac{1}{|T_{i-1}|}\sum_{a_j\in T_{r-1,j}}a_j,\tag{5}$$

其中，r 为算法迭代次数，$T_{r,j}$表示第 r 次迭代的第 j 个类别。当 $\sigma_r\leqslant\sigma_1$时，满足条件，然后对其他样本进行筛选，直到所有样本数据都被正确识别。

算法流程如图 1 所示。算法描述如下：

步骤 1 依据经验规律确定聚类数 K 值在 2 到$\sqrt{N}$之间，其中 N 为数据空间中的所有数据点的个数。通过

在[2，$\sqrt{N}$]区间逐个选取 K 值，并利用聚类有效性函数来评价聚类的效果。最终得到最优的 K 值。

步骤 2　使用 K-中心值法选出初始聚类中心。

所谓 K-中心值法就是为了避免孤立点的干扰而采用簇的平均值作为参照点，此种方法仍然是基于最小化所有对象与参照点之间的相异度之和的原则来执行的。

步骤 3　将样本集中的样本按照欧式距离原则分配到最邻近的簇中。

步骤 4　计算每个类的质心点。

步骤 5　判断聚类中心点的改变量是否满足设定的条件，如果满足，将其加入到已选特征集，同时将它从数据样本集中删除。

步骤 6　判断数据样本集是否为空，如果为空，结束算法。如果不为空，遍历中心点个数 N，当 $N<K$ 时，转向步骤 3，当 $N=K$ 值时，继续下一步。

步骤 7　更新中心点。计算每个聚类中心点的改变量大于设定值的簇的质心，并将其作为新的聚类中心，然后转向步骤 3。

步骤 8　结束，数据样本为空集，得到 K 个结果簇。

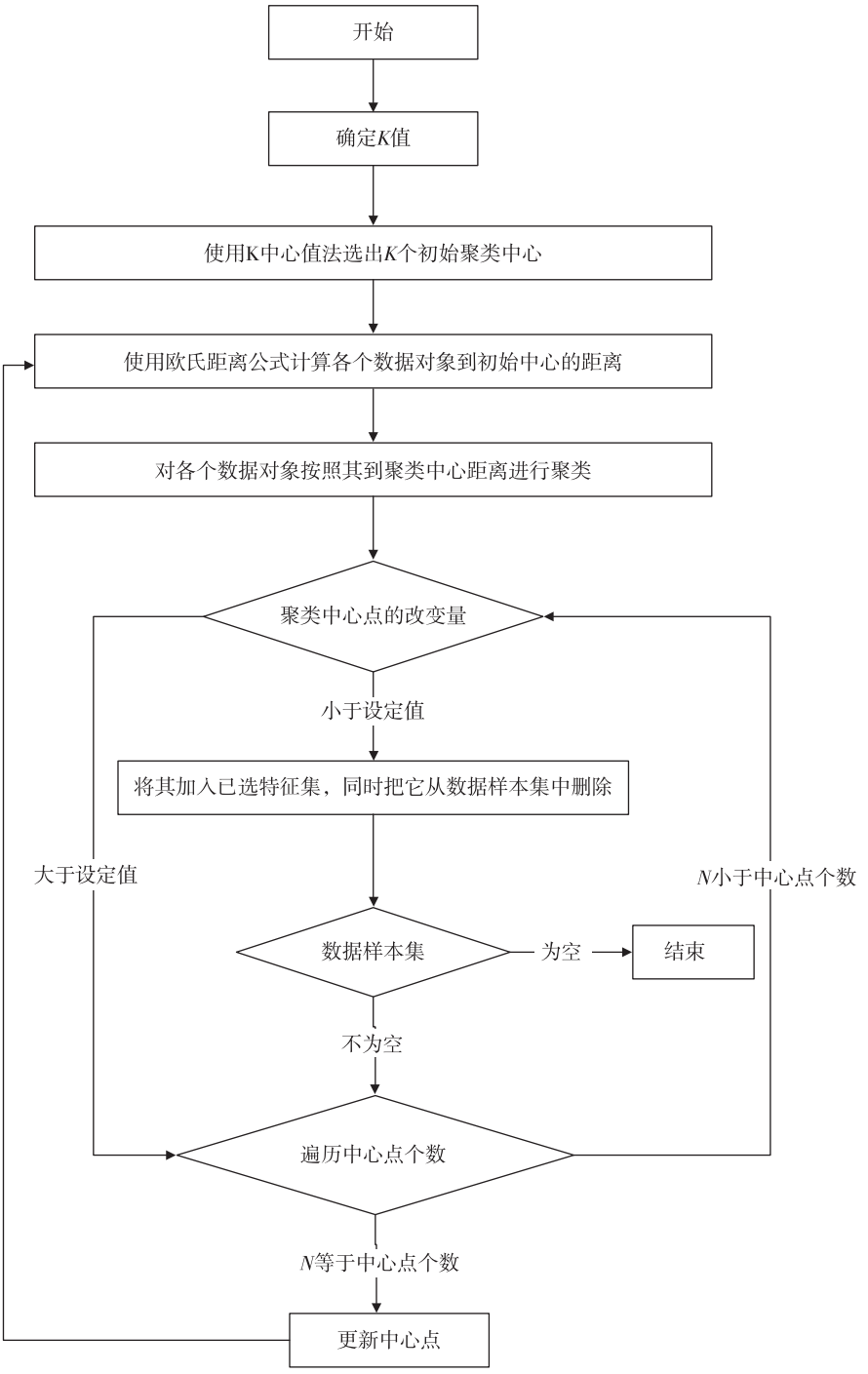

图 1　K-means **动态聚类流程图**

算法优点：该算法充分考虑了 K-means 聚类过程中信息的动态变化，通过为算法终止条件设定标准值，减少了算法迭代次数，减少了学习时间。通过删除由信息动态变化而产生的冗余信息，来减少动态聚类过程中的干扰，使算法达到更准确更高效的聚类效果。

3 实验研究

3.1 实验数据集

为了分析 K-means 动态聚类算法的聚类性能，模拟实验使用 5 种不同的公用数据集合。数据样本集合均来自 UCI 机器学习数据库，UCI 数据库是一个专门用于数据挖掘算法和测试机器学习的公用数据库。库中的数据均有确定的属性类别，因此，可以用准确率和时间效率来衡量聚类性能的优劣。为验证传统 K-means 算法和 K-means 动态聚类算法的准确率和时间效率，不对任何测试数据集的数据分布做任何人为处理。表 1 描述了 5 组数据的概要信息，如名称、样本数和类别数等。

从表1可以看出,5组数据分别由不同的数量样本数和类别数组成,数据集的多元性在一定程度上验证了它们在不同条件下的性能,保证了实验结果具有普遍性。

表1 实验数据集

序号	数据库	样本数/个	类别数/个
1	Lung-cancer	3 203	27
2	Promoter	10 608	16
3	Splice	31 909	32
4	Coil	58 223	49
5	Isolet	156 090	57

3.2 实验结果对比

为了避免K-means算法本身固有的缺陷对实验结果造成影响,现对实验数据做预处理。首先确定聚类数 K 值,依据经验规律 K 的取值在2到$\sqrt{N}$之间,其中 N 为数据空间中的所有数据点的个数。通过在$[2,\sqrt{N}]$区间逐个选取 K 值,并利用较为传统的有效性函数 W_n 指数来评价聚类效果,进而得出最优的 K 值。5组数据的 K 值分别为:23,18,42,39,73。然后使用K-中心值法选出初始聚类中心。再使用欧氏距离计算各个数据样本到中心点的距离并对各个数据对象按照其到聚类中心距离进行聚类。以上为数据的预处理操作部分;在算法迭代过程中所获得的 σ_1 值为 1.31×10^{-7}。接下来依据算法流程分别对传统的K-means算法和K-means动态聚类算法完成聚类的整个流程。分别对聚类算法精度和时间效率进行了对比,实验结果如表2和表3所示。

表2 传统K-means算法与K-means动态聚类算法精度比较

算法	数据库	聚类精度/%
传统K-means	Lung-cancer	85.53
	Promoter	83.24
	Splice	82.87
	Coil2000	75.09
	Isolet	73.32
K-means动态聚类	Lung-cancer	65.76
	Promoter	70.23
	Splice	80.83
	Coil2000	85.28
	Isolet	89.79

表3 传统K-means算法与K-means动态聚类算法执行时间的比较

算法	数据库	执行时间/ms
传统K-means	Lung-cancer	35
	Promoter	72
	Splice	117
	Coil2000	185
	Isolet	402
K-means动态聚类	Lung-cancer	16
	Promoter	16
	Splice	17
	Coil2000	35
	Isolet	56

由表2可见,K-means动态聚类算法对数据量较大的Coil2000和Isolet数据集取得了较高的聚类精度,对数据量较小的Lung-cancer和Promoter数据集,K-means动态聚类算法精度低于传统K-means算法的聚类精度,说明数据量越大,改进后的算法就越有优势,说明通过删除聚类过程中的冗余信息,逐步减少聚类的过程中的干扰确实可以提高聚类准确率。说明当数据量较小时,改进后的算法并不可取。由表3可见,传统的K-means算法较改进后的算法在执行时间上要长,分别是改进后算法的2.19倍、4.50倍、6.88倍、5.29倍和7.19倍。改进后的算法在执行效率上较传统的算法有较大的提升,数据表明,数据量越大,效率就越高。

4 结语

K-means算法是一种应用广泛的聚类算法,在众多领域都取得了不亚于传统聚类算法的聚类效果,虽然如此,该算法固有的缺陷依然对聚类的性能造成了一定的影响,如K-means算法需要事先确定 K 的个数;需要事先选取初始中心点;对孤立点数据很敏感等;目前已有不少学者针对这些问题提出来改进算法,然而这些方法目前只能适用于一些特定的领域,尚不具有通用性。那么是否可以借鉴前人的研究成果,将K-means算法应用于文本分类中的特征选择环节也是值得科学研究工作者深思的问题,也是作者下一步研究的方向。

参考文献:

[1] 袁利永,王基一.一种改进的半监督K-means聚类算法[J].计算机工程与科学,2011,33(6):138-143.
Yuan L Y, Wang J Y. An improved semi-supervised K-means clustering algorithm[J]. Computer Engineering & Science, 2011, 33(6): 138-143.

[2] 傅德胜,周辰.基于密度的改进K均值算法及实现[J].计

算机应用,2011,31(2):432-434.
Fu D S,Zhou C. Iproved K-means algorithm and its implementation based desnsity[J]. Journal of Computer Applications,2011,31(2):432-434.

[3] 张文明,吴江,袁小蛟. 基于密度和最近邻的 K-means 文本聚类算法[J]. 计算机应用,2010,30(7):1933-1935.
Zhang W M,Wu J,Yuan X J. K-means text clustering algorithm based on density and nearest neighbor[J]. Journal of Computer Applications,2010,30(7):1933-1935.

[4] 刘海峰,刘守生,张学仁. 聚类模式下一种优化的 K-means 文本特征选择[J]. 计算机科学,2011,38(1):195-197.
Liu H F,Liu S S,Zhang X R. Clustering-based improved K-means text feature selection[J]. Computer Science,2011,38(1):195-197.

[5] 原福永,张小彩,罗思标. 基于信息熵的精确属性赋权 K-means 聚类算法[J]. 计算机应用,2011,31(6):1675-1677.
Yuan F Y,Zhang X C,Luo S B. Accurate property weighted K-means clustering algorithm based on information entropy[J]. Journal of Computer Applications,2011,31(6):1675-1677.

[6] 谢娟英,高红超. 基于统计相关性与 K-means 的区分基因子集选择算法[J]. 软件学报,2014,25(9):2050-2075.
Xie J Y,Gao H C. Statistical correlation and K-means based distinguishable gene subset selection algorithms[J]. Journal of Software,2014,25(9):2050-2075.

[7] LEE S S,Lin J C. An accelerated K-means clustering algorithm selction and erasure rules[J]. Zhejiang University-SCIENCE C:Computers Electronics,2012,13(10):761-768.

[8] Orakoglu F E,Ekinci C E. Optimization of constitutive parameters of foundation soils K-means clustering analysis [J]. Sciences in Cold and Arid Regions,2013,5(5):0626-0636.

An Improved K-means Dynamic Clustering Algorithm

ZHANG Yang[1], HE Li[2], ZHU Haodong[3]

(1. College of Mathematics and Information Science, Zhengzhou University Of Light Industry, Zhengzhou 450002;
2. College of Computer and Information Science, Chongqing 401331;
3. School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: The traditional K-means algorithm clusters by repetitive computing. the changing cluster centers bring some of the dynamic change information, It will produce interference for clustering. And the large amounts of data will increase the algorithm's time overhead and system I/O overhead, even affect the performance of the algorithm,So,this paper proposed an improved K-means dynamical clustering algorithm. The proposed algorithm takes into account the dynamic information of K-means clustering process and reduces algorithm iterations and learning time by setting the standard value for termination condition of the algorithm, and reduces interference of dynamic clustering by removing redundant information from the changing information to make the algorithm to achieve more accurate and efficient clustering effect. Experimental results show, when the amount of data is large, the improved K-means algorithm is better than the traditional algorithms in accuracy and efficiency.

Key words: K-means; cluster analysis; data mining; dynamic clustering

(责任编辑 游中胜)