AGNES 算法在 K-means 算法中的应用 *

周爱武,潘 勇,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:提出一种新的选取初始聚类中心的算法,该算法结合了凝聚层次聚类算法 AGNES,利用该算法选出初始聚类中心,再应用到 K-means 算法中进行聚类。实验表明,改进的算法聚类效果更好,准确率得到了提高,迭代次数也明显减少,还能够发现异常点。

关键词:K-means 算法;AGNES 算法;初始聚类中心;密度;簇
中图分类号:TP301 文献标识码:A 文章编号:1674-7720(2011)23-0079-03

The application of AGNES in K-means algorithm

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

Abstract: This paper puts forward a new algorithm to choose the initial clustering center, which combines the AGNES algorithm. After applying this algorithm to choose the initial clustering center, we use the K-means algorithm to cluster. Experiments have shown that with the improved algorithm, it is easy to get clustering results characterized by better clustering effects, higher accuracy and a obvious reduction of iteration times.

Key words: K-means algorithm; AGNES algorithm; initial clustering center; density; cluster

K-means 算法是基于划分的经典聚类算法,该算法简单、快速,得到了广泛应用,但是该算法对孤立点是敏感的,其次该算法要求用户必须事先给出簇数 k 值,该算法聚类结果受初始聚类中心随机性选择影响比较大,聚类中心选择不当,容易导致聚类准确度下降,甚至无法得到聚类结果。针对以上缺点,国内外许多专家和学者提出很多对 K-means 算法的改进,比如参考文献[1]~参考文献[3]是基于密度的对 K-means 算法进行改进来找到密度比较高的代表点作为初始聚类中心,参考文献[4]是把遗传算法应用到 K-means 算法中来确定 k 个初始聚类中心,这种结合能够进行一种全局优化,可提高算法的准确率,将智能算法与 K-means 算法的结合也是近年来研究的热点。K-means 算法以及其改进的算法已经应用到了各个领域,取得了很好的效果,比如参考文献[3]是改进的 K-means 算法在客户划分中的应用,参考文献[5]是在图像标注和检索的应用,参考文献[6]是在存在异常孤立点大数据集中进行聚类发现簇的应用。由此可见,K-means 算法的效率高、时间复杂度低,应用前景宽广。

本文针对初始聚类中心随机选择这个缺陷,提出将凝聚层次聚类算法 AGNES 应用到改进的算法中,得到 k 个密度比较高的初始聚类中心,并将其应用到 K-means 中去进行聚类。实验表明,改进的算法运行效率高,并能获得比较高的准确率。

1 算法的基本思想

1.1 K-means 算法

算法接收输入量 k,然后将 n 个数据对象划分为 k 个聚类以便使得所获得的聚类满足:同一聚类中的对象相似度较高;而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“聚类中心对象”来进行计算的。

K-means 算法步骤描述如下:

输入:n 个数据对象,簇数 k

输出:k 个簇的集合。

(1) 从 n 个数据对象中任意选择 k 个数据对象作为初始聚类中心。

(2)根据簇中数据对象的平均值,将每个数据对象依

* 基金项目:安徽省教育厅重点项目(KJ2009A57)

次划分到最近距离聚类中心标志的簇中。

(3) 更新簇的平均值，即计算每个簇中对象的平均值，$\bar{x}_i=\sum_{x\in C_i}\frac{x}{|C_i|}$，其中$|C_i|$为 i 簇中数据对象的个数，x 为簇 i 中的数据对象，$\bar{x}_i$ 为簇 i 中数据对象的平均值。

(4)计算 $E=\sum_{i=1}^{k}\sum_{x\in C}|x-\bar{x}_i|^2$，其中 x 为 i 簇中的数据对象，i 为簇类别，簇数目为 k，E 为所有簇内数据对象到各自簇聚类中心的距离和。

(5)如果 E 值变化明显，将返回步骤(2)。

1.2 AGNES 算法

参考文献[7]中提到的 AGNES 算法是凝聚的层次聚类方法。AGNES 算法最初将每个对象作为一个簇，然后这些簇根据某些准则被一步步地合并。

AGNES 算法描述如下：

输入：包含 n 个数据对象的数据库，终止条件簇的数目 c；

输出：c 个簇，达到终止条件规定簇数目。

(1)将每个数据对象当成一个初始簇；

(2)REPEAT；

(3)根据两个簇中最近的数据点找到最近的两个簇；

(4)合并两个簇，生成新的簇的集合；

(5)UNTIL 达到定义的簇的数目；

2 改进 K-means 算法

K-means 算法初始聚类中心是随机选取的，选取不当的话，将会无法得到正确的聚类结果。而 AGNES 算法将最近的数据对象聚集在一起形成簇，根据 AGNES 算法的这种思想，提出一种新的算法来确定 k 个初始聚类中心。

2.1 相关概念

定义 1 欧式距离

数据对象 x_i，x_j 都为 n 维数据，其中 $x_i=(x_{i1},x_{i2},\cdots,x_{in})$，$x_j=(x_{j1},x_{j2},\cdots,x_{jn})$；两对象距离记为：$Distance(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2}$

定义 2 点的密度

数据 x_i 和距离 α，以 x_i 为圆心，α 为半径的区域内对象的个数（包括圆心自身）记为该对象的密度，记为 $Density(x_i)=(p$ 的个数$|Distance(x_i,p)\leq\alpha)$，其中 $Dis\tan ce(.)$ 为欧式距离，α 取所有数据对象间距离和的平均值。

定义 3 平均密度

n 个数据对象，平均密度记为：$Ave=\frac{\sum_{i=1}^{n}Density(x_i)}{n}$

2.2 初始聚类中心的确定算法

Initial Clustering Center 算法如下：

输入：n 个数据对象的数据集 DataSet，聚类簇数 k，参数 θ，ClusterSet、HighSet、CenterSet 3 个集合开始均为空。

输出：k 个初始聚类中心数据集 CenterSet。

(1)将每个对象当成一个初始簇；

(2)REPEAT；

(3) 根据两个簇中最近的的数据点找到最近的两个簇；

(4)合并两个簇，生成新的簇的集合；

(5)UNTIL 达到定义的簇的数目 $c=\theta\times k_{max}$，其中 $k_{max}=\sqrt{n}$，这是根据人们长期根据经验得出的结论，簇数 k 值不会超过 k_{max}，即 $2\leq k\leq k_{max}$，θ 是为了调节生成簇的数目 $\theta\times k_{max}$，防止步骤(6)~步骤(18)生成的簇数比指定的簇数 k 值小；

(6)计算数据对象的平均密度 Ave；

(7)对步骤(5)中生成的 c 个簇将其加入到 ClusterSet 集合中，分别计算每个簇中数据对象的个数，将个数大于 Ave 的簇加入到 HighSet 集合中；

(8)令 $i=1$；

(9) 如果 HighSet 集合不为空，执行步骤(10)~步骤(13)；否则执行步骤(15)~步骤(18)；

(10)REPEAT；

(11)取出 HighSet 集合中数据对象个数最多的簇 c_x；

(12)计算该簇 c_x 中数据对象的平均值做为第 i 个初始聚类中心，将其加入到 CenterSet 集合中去，并将簇 c_x 从 HighSet 和 ClusterSet 中删除，$i=i+1$；

(13)UNTIL 得到 k 个初始聚类中心；

(14)如果步骤(10)~步骤(13)生成的初始聚类中心个数小于 k，并且 HighSet 集合为空，即执行步骤(15)~步骤(18)；

(15)REPEAT；

(16)取出 ClusterSet 集合中数据个数最多的簇 c_x；

(17)计算簇 c_x 中数据对象的平均值做为第 i 个初始聚类中心，将其加入到 CenterSet 集合中去，并将簇 c_x 从 ClusterSet 中删除，$i=i+1$；

(18)UNTIL 得到 k 个初始聚类中心；

(19)算法结束。

2.3 改进的 K-means 算法

改进 K-means 算法如下：

(1)运行 Initial Clustering Center 算法，得到 k 个初始聚类中心集合 CenterSet；

(2)输入 k 值、k 个初始聚类中心集合 CenterSet 和数据对象集合 DataSet；

(3)运行 K-means 算法，输出 K 个聚类，算法结束。

3 改进后的实验及分析

本实验采用 C# 编写，参数 θ 取 1。实验第一组数据采用 UCI 数据库里的 Iris 数据集作为实验数据对象，该数据集共有 150 条记录，每个记录有 4 个属性，sepal

length, sepal width, petal length, petal width, 可以分为3类,所以簇数 k 取值为3,其中 $k_{max}=\sqrt{n}\approx12$。第二组数据是随机选取10个二维数据点 (1,1),(2,1),(1,2),(2,2),(4,3),(5,3),(4,4),(5,4),(2,6),(2,7),($k_{max}=\sqrt{n}\approx3$),实际上10个二维数据点分为3类是比较合适的,所以在实验中,把聚类簇数 k 值也取为3。该实验从初始聚类中心、迭代次数和准确率三方面来对K-means和改进算法进行比较,实验结果如表1所示(改进的K-means算法得到的初始聚类中心是通过算法计算平均值得到的,表1显示的是具体数据对象的维度值,传统K-means算法聚类中心是数据对象点标号,是通过随机选取得到的)。

表1 传统K-means算法和改进K-means算法的比较

算法	Iris数据集			随机数据集		
	初始聚类中心	迭代次数	准确率/%	初始聚类中心	迭代次数	准确率/%
传统K-means	(17,49,64)	8	48.67	(3,4,6)	4	60
	(30,80,135)	10	88.67	(1,6,7)	5	100
	(12,60,141)	14	85.45	(1,2,7)	3	40
	(1,20,51)	5	53.78			
改进K-means	(5.3,3.7,1.4,0.4)	6	89.33	(1.6,6.5)	2	100
	(5.5,2.8,3.5,1.2)			(1,1.5)		
	(7.2,3.1,6.1,1.9)			(4.6,3.2)		

从表1 Iris数据集可以看出传统K-means算法的迭代次数有时比改进K-means算法要小,但是其准确率却显得很低。这说明K-means算法受初始聚类中心的影响比较大,一旦选择不当,准确率将会很低,而优化的初始聚类中心是比较稳定的,符合数据对象的实际分布,同时可以尽快收敛最优解并且准确率高。从随机数据集整体上来看,改进的K-means算法的迭代次数减少,收敛速度比较快,准确率很高。所以改进的K-means算法收敛速度比较快,得到的初始聚类中心符合数据对象的实际分布,提高了聚类准确率,达到了更好的聚类效果,同时该算法在烟草配送点分配问题等领域都得到了很好的应用。

本文结合层次聚类算法AGNES算法,提出一种确定初始聚类中心的算法,使得到的初始聚类中心比较稳定,符合数据的实际分布,避免选到孤立点,大大提高了算法的准确率和效率。同时改进的算法一个优点是能进行异常点的发现,能够对网络异常进行检测。但是改进K-means算法的时间复杂度不如传统K-means算法,所以今后的研究方向主要是针对大数据集在时间复杂度的提高上作进一步的研究,并将改进的算法应用到其他实际领域中。

参考文献

[1] 赖玉霞, 刘建平.K-means算法的初始聚类中心的优化[J].计算机工程与应用,2008,44(10):147-149.

[2] 刘艳丽,刘希云.一种基于密度的K-均值算法[J].计算机工程与应用,2007,43(32):153-154.

[3] 向坚持,刘相滨,资武成.基于密度的K-Means算法及在客户细分中的应用研究[J].计算机工程与应用,2008,44(35):246-248.

[4] 孙秀娟,刘希玉.基于初始中心优化的遗传K-means聚类新算法[J].计算机工程与应用,2008,44(23):166-168.

[5] 潘崇,朱红斌.改进K-means算法在图像标注和检索中的应用[J].计算机工程与应用,2010,46(4):183-185.

[6] ESTER M, RIEGEL H P. A density based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Ortland, Oregon, 1996.

[7] 毛国君,段立娟,王实,等.数据挖掘原理与算法[M].北京:清华大学出版社,2000年.

(收稿日期:2011-06-09)

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。

潘勇,男,1985年生,硕士,主要研究方向:数据库与web技术,数据挖掘,计算机软件与理论。

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,计算机应用。