K-Means 算法的研究与改进

周爱武,陈宝楼,王 琰
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:K-Means 算法是一种基于划分方法的经典聚类算法,已经在很多领域得到广泛的应用。虽然该算法有很多优点,但其也存在自身的局限性,比如需要用户输入聚类簇个数,初始聚类中心是随机性选择的,算法容易陷入局部最优解,对孤立点比较敏感等。文中首先应用统计学中的标准分数对样本进行孤立点分析,然后提出一种新的初始聚类中心确定策略。对改进的算法和原算法分别做实验进行比较,实验结果表明,改进的算法在准确率、收敛速度和稳定性方面都有很大的提高。

关键词:K-Means 算法;孤立点;初始聚类中心
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)10-0101-04

Research and Improvement of K-Means Algorithm

ZHOU Ai-wu, CHEN Bao-lou, WANG Yan
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: K-Means algorithm is a classic clustering algorithm based on the classification method has been widely applied in many fields. Although the algorithm has many advantages, there are also their own limitations, such as user input the number of clusters, initial cluster centers is random selection, the algorithm is easy to fall into local optimal solution is more sensitive to outlier and so on. It firstly analyses sample outlier by statistics standard scores, and then puts forward a new strategy to determine the initial clustering centers. Improved algorithm and the original algorithm were doing experiments to compare, the experimental results show that the improved algorithm's accuracy rate, convergence speed and stability are improved greatly.

Key words: K-Means; outlier; initial clustering centers

0 引 言

随着数据挖掘的应用不断发展,聚类分析已经广泛地应用到了很多领域,包括市场研究、数据分析、模式识别、图像处理、人工智能和 Web 文档分类等领域[1,2]。K-Means 算法是最为经典的基于划分的聚类方法,是十大经典数据挖掘算法之一[3]。

K-Means 算法的基本思想就是把用户提供的 n 个样本点划分到 k 个簇中,其中每个样本点属于距离簇中心最近的那个簇。结果使得每个簇中的样本点有较高的相似性,不同簇中的样本点有较高的相异性。1967 年 MacQ J 最早在文献[4]中提出 K-Means 算法的基本思想,并且在文献中给出了经典的基于划分的算法 MacQueen。但是由于 MacQueen 算法的代价较大且聚类结果较差,人们后来对初始聚类中心和循环迭代进行优化。在初始聚类中心选择方面,主要有以下三种经典的改进算法。文献[5]提出的一种简单聚类探索方法,文献[6]给出的一种二元分裂方法,文献[7]给出的 K 值评估方法。

传统 K-Means 算法随机选取初始中心,算法容易陷入局部最优。如何选取合理的一组初始聚类中心,在降低聚类结果波动性的同时又能得到较高的聚类准确率具有重要的现实意义。针对随机选取初始聚类中心的缺陷,提出了一种带孤立点检测根据距离和密度动态选取初始聚类中心的方法,得到的聚类中心能够很好地代表 K 个聚类簇。实验结果表明改进的算法聚类结果较原算法在准确率、收敛速度和稳定性方面都有明显的提高。

1 K-Means 算法

1.1 K-Means 算法基本介绍

K-Means 聚类算法是一种基于划分的聚类方法,即在欧几里得空间里把 n 个数据对象组织分为 K 个划分($k \leqslant n$),每个划分分别代表一个簇。首先,由用户输入所要聚类的数目 K,并通过某种初始中心策略初

收稿日期:2012-02-17;**修回日期**:2012-05-28
基金项目:安徽省教育科研重点计划项目(KJ2009A57)
作者简介:周爱武(1965-),女,副教授,研究方向为数据库与 web 技术、数据仓库与数据挖掘、信息系统安全;陈宝楼(1987-),男,硕士研究生,研究方向为数据库与 web 技术、数据挖掘。

始选择 K 个对象作为聚类中心,对剩余的每个对象,根据其与各中心的距离将它置于最近的类中。然后,重新计算每个类中数据对象的平均值形成新的聚类中心[8]。这个过程重复迭代进行,直到聚类结果收敛为止。K - Means 聚类算法的步骤可描述如下[2]:

输入:K:簇的数目,D:包含 n 个对象的数据集。

输出:K 个簇的集合。

算法:

(1) 从 D 中任意选择 K 个对象作为初始簇中心;

(2) repeat;

(3) 根据簇中对象的均值,将每个对象(再)指派到最相似的簇;

(4) 更新簇均值,即计算每个簇中对象的均值;

(5) until 目标函数不再发生变化。

步骤(5)返回第(3)步循环执行,当目标函数不再变化时算法结束。

1.2 K-Means 算法的优缺点

K-Means 算法是解决聚类问题的一种基于划分的经典算法,由于算法过程的简单快捷,所以在数据挖掘中应用比较广泛。当聚类的数据是密集的(凸型的),并且簇与簇之间的数据差异较大,K-Means 算法的聚类效果较好。当处理大量数据集时,该算法是高效并且是相对可伸缩的。

然而,K-Means 算法也有自身的局限性。该算法在只有簇均值可定义的情况下才可以运用,例如在涉及分类属性的数据集时就不可用。用户必须实现给出待聚类的具体簇数目也是该算法的一个缺点,并且这个过程要靠用户的经验。K-Means 算法不易发现非凸状的簇,或者每个簇规模差异很大的簇。由于算法的初始聚类中心是随机的,所以算法不但不稳定而且容易陷入局部最优解。此外,它对噪声和离群点数据是敏感的[4]。

2 改进的 K-Means 算法

文中主要针对 K-Means 算法在孤立点分析和初始聚类中心的选择方面加以改进。

2.1 孤立点分析

在数据挖掘中经常存在一些数据对象,它们与数据的一般行为或模型不协调。这样的数据对象称为孤立点,它们与数据集中的其它样本点有着很大的差异。孤立点在聚类挖掘中的主要体现是,孤立点会远离数据密集的区域。如果聚类算法的初始聚类中心的选取过程是针对整个样本点的话,那么选出的初始聚类中心可能会是孤立点或者严重偏离实际中心的点。此外,聚类过程在进行新一轮聚类迭代时,计算均值如果把孤立点也算在内的话,则新的初始聚类中心也会存在误差。综上,孤立点的存在会影响聚类的整个过程,并且还会造成错误积累。所以在改进的算法中首先去除孤立点,然后再进行聚类过程。孤立点在常规聚类过程完成后,再单独把它们放到聚类较近的簇中。

标准分数(standard score)是以标准差为单位来衡量某一分数与平均数之间的离差情况,是反映个体在团体中相对位置的最好统计量[9]。在统计中,变量值与其平均数的离差除以标准差的值,称为标准分数,也称为标准化值或 Z 分数。计算过程是对变量数值进行标准化处理的过程,它并没有改变该组数据的分布情况,而只是将数据变为平均数为 0,标准差为 1 的数据。在统计中,当一组数据对称分布时,经验法则表明:约有 68% 数据的 Z 分数的绝对值小于等于 1,约有 95% 数据的 Z 分数的绝对值小于等于 2,约有 99% 数据的 Z 分数的绝对值小于等于 3,如果数据 Z 分数的绝对值大于 3 就可以认为是离散点。

文中采用"Z 分数的绝对值大于 2 的数据作为孤立点"的方法来对数据进行预处理,处理过程如下:

设 point[i][j] 表示第 i 个点的第 j 维的值,则 i 和 j 点之间的欧式距离可表示为 $\mathrm{dist}[i][j] = \sqrt{\sum_{k=1}^{d}(\mathrm{point}[i][k] - \mathrm{point}[j][k])^2}$,$i$ 点到其他所有点距离之和 $\mathrm{Dist}[i] = \sum_{j=1}^{N}\mathrm{dist}[i][j]$,$d$ 为样本点的维数。

定义 1:样本点 i 的标准分数 $z[i] = \frac{1}{\sigma}(\mathrm{Dist}[i] - \mu)$,其中:

$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Dist}[i], \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{Dist}[i] - \bar{x})^2}$$

传统去除孤立点的方法主要是通过计算欧式距离,去除离其它点距离之和最远的 λ % 个点。这里的阈值 λ %,是通过人为凭经验指定,这样的实验结果受人为因素干扰。而标准分数是统计学中描述数据分布的重要方法,它能够克服人为指定阈值的不足,公平、公正、客观、科学地对待群体中的每一个成员。

2.2 初始聚类中心的确定

原 K-Means 算法的初始聚类中心确定策略是随机地选择 K 个数据对象,每个数据对象各自代表一个簇的初始聚类中心。这样的 K-Means 算法对于初始聚类中心就很敏感,对于不同的初始聚类中心,聚类结果会有很大的差别。由于 K-Means 聚类算法的聚类过程是采用迭代更新的方法,所以当初始聚类中心落在局部值最小区间附近时,整个聚类算法很容易生成局部最优解[10]。因此,要想获得较好的聚类结果,可以从初始中心的选择上进行改进,降低随机性,以达到

优化的目的。

一般说来确定初始聚类中心问题没有一个简单、普遍适用的解决办法。很多算法采用的方法是:(1)随机方式确定;(2)通过设计的算法确定。前者可能选取“孤立点”、类边缘点或者一个类中两个以上的点作为初始聚类中心,这样聚类效果肯定不理想。后者因为设计出的算法带有主观性,同时给计算带来了负担,因此,设计一个合理的初始聚类中心确定算法对于实际问题很有意义。

本算法的思想是每次把相对集中的数据先划分出来,这样可以保证每类划分的样本有着较高的相似性。文中改进后的初始聚类中心确定算法描述如下:

输入:聚类个数 K 及包含 n 个样本点的数据集。

输出:K 个初始聚类中心。

算法:

For $i = 1:K - 1$

(1) 找出到其他点距离之和最远点,记为 O_{i1};

(2) 找出距离 O_{i1} 点最远点 O_{i2};

(3) 把距 O_{i2} 点距离小于等于第 N/K 个小元素的点(这些 N/K 点较其它点来说离 O_{i2} 点距离较近)归为类 i;

(4) 从样本集合中删除已归类的点,求出类 i 的聚类中心;

End

把集合中剩下的点归为类 K,同时也求出类 K 的聚类中心。其中,K 是指簇的个数。

本算法的每个步骤只是简单的欧式距离计算,没有涉及其它的运算,例如文献[11]中的密度函数,在运算量上较其它初始中心确定策略相对较少。

3　实验分析

为了检验改进算法的有效性,本实验采用 UCI 数据库的 Iris 和 Libras Movement 数据集作为测试对象,比较的性能指标主要为准确率和收敛速度(每次测试的循环次数)两个指标。算法采用 C 语言编写,在 Pentinum(R)4 3.0GHz,1GB 内存,Dev-C++ 4.9.9.2 环境下运行。

首先用 Iris 数据集对改进的算法和原算法分别进行随机 10 次测试,在准确率方面的比较结果如图 1 所示。

对改进的算法和原算法分别进行随机 10 次测试,在循环次数方面的结果如图 2 所示。

图 1 和图 2 呈现了实验的运行结果:原算法的准确率在 79% ~89% 之间波动、循环次数在 3 ~13 次之间波动,而改进算法的准确率是 92%、循环次数为 3 次。

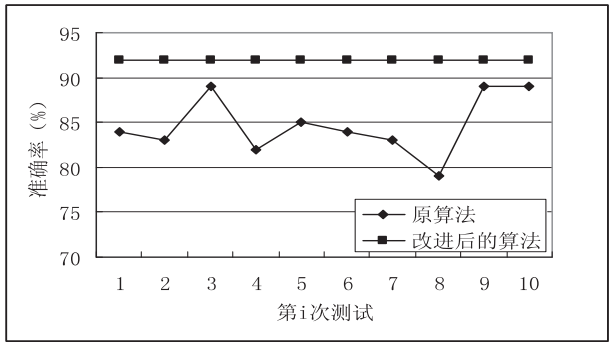

图 1　原算法和改进的算法随机运行 10 次的准确率比较图

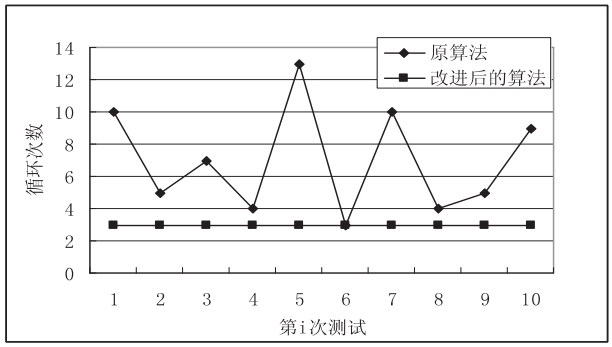

图 2　原算法和改进的算法随机运行 10 次的循环次数比较图

原算法之所以既不稳定准确率又不高,是因为受孤立点和初始聚类中心的影响。孤立点的存在不但会影响初始聚类中心的确定,而且还会影响后面的聚类过程。改进的算法在聚类过程进行之前先去除孤立点,这样可以避免孤立点对整个聚类过程的影响,通过文中使用的方法去除的孤立点有 5 个,分别是(7.6,3.0,6.6,2.1)、(7.7,3.8,6.7,2.2)、(7.7,2.6,6.9,2.3)、(7.7,2.8,6.7,2.0)和(7.9,3.8,6.4,2.0)。去除孤立点后的数据集分布相对均匀,聚类效果有所改善。但是孤立点也是实际应用中的一部分,所以在聚类完成后再把孤立点加入到数据集中再重新聚类一次。文献[11]的实验结果是 88.67%,文献[12]的实验结果是 88%,改进后的算法准确率是 92%、循环次数为 3 次,改进后的算法准确率是 92%、循环次数为 3 次,该实验结果相对原算法的实验结果不但准确率提高而且相对比较稳定。

原算法的时间复杂性为 $O(nkt)$。文中改进算法的第一步去除孤立点的时间复杂性为 $O(n*n)$,第二步初始中心确定的时间复杂性为 $O(nk)$,第三步算法循环运行的时间复杂性为 $O(nkt)$,其中 n 表示样本点数目,k 表示簇的个数,t 表示循环的次数。如果用 Iris 数据集对原算法和改进后的算法进行性能测试,那么测试结果基本相同。Iris 数据集样本数相对较少,聚类过程迭代的次数也不多,所以无法体现出改进后的算法在性能上的优越性。所以为了验证算法的效率,

选用更大的数据集 Libras Movement 进行测试,该数据集共有 360 个样本点,每个样本点是 90 维浮点型数据,样本点共分为 15 类。为了让实验结果更加明显,给出程序运行 100 次的结果:原算法花费 228031 毫秒时间,改进的算法花费 206200 毫秒时间。聚类过程之前的去除孤立点和初始聚类中心确定操作使得初始聚类中心更加符合实际的情况,这样才会使得聚类过程的迭代次数减少,在总的运行时间上才会少于原算法的运行时间。该实验结果表明,改进的算法在孤立点和初始聚类中心的处理时间花费是值得的,尤其对于大数据集来说。

由此可以得出,改进的算法在准确率、稳定性和收敛速度方面都有很大的提高。

4 结束语

K-Means 算法作为经典的聚类算法,对数据集是紧凑的并且簇与簇之间明显分离的情况聚类效果较好,但是它受孤立点和初始中心影响较大。文中首先运用统计方法对孤立点进行检测,然后提出一种新的基于分布的初始聚类中心确定策略。实验结果表明,改进的算法在准确率、稳定性和收敛速度方面都有很大的提高。

参考文献:

[1] 周卫星,廖 欢. 基于 K 均值聚类和概率松弛法的图像区域分割[J]. 计算机技术与发展,2010,20(2):68-70.

[2] Han Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. Beijing: China Machine Press, 2007.

[3] Wu Xindong, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowl. Info. Syst., 2008(14): 1-37.

[4] MacQ J. Some methods for classification and analysis of multivariate observations[C]//Proc of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: [s. n.], 1967: 281-297.

[5] Tou J. Pattern Recognition Principles[M]. USA: Addison Wesley, 1974.

[6] Linde Y, Buzo A, Gary R. An Algorithm for Vector Quantizer Design[J]. IEEE Trans on Communication, 1980, 28(1): 84-95.

[7] Chomicki J, Godfrey P, Gryz J, et al. Skyline with Presorting Theory and Optimization[C]//Proc of the International Conference on Intelligent Information Systems. Wroclaw, Poland: [s. n.], 2005: 216-225.

[8] 黄震华,向 阳,张 波,等. 一种进行 K-Means 聚类的有效方法[J]. 模式识别与人工智能,2010,23(4):517-521.

[9] 贾俊平. 统计学[M]. 北京:中国人民大学出版社,2006.

[10] 朱颢东,钟 勇,赵向辉. 一种优化初始中心点的 K-Means 文本聚类算法[J]. 郑州大学学报(理学版),2009,41(2):30-30.

[11] 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.

[12] 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):66-66.

(上接第 100 页)

4 结束语

上述试验结果显示,文中提出的适应度函数能够使遗传算法更高效的收敛,比单纯的“分支函数叠加法”[12]有更好的效率。遗传算法是一个随机算法,初始种群的适应度值将会影响算法的收敛性和解的适应性。在初始化种群的时候,文中给出了两种算法来生成初始群体,提高种群的适应度。鉴于文中存在的一些不足,如参数的类型、范围等问题,还有待进一步思考,对于参数范围如何自适应处理,多路径覆盖如何高效产生等相关工作,简化了服务组合的工作量,尚需要进一步完善。这是下一步研究的重点。

参考文献:

[1] Korel B. Automated Software Test Data Generation[J]. IEEE Trans. on Software Eng., 1990, 16(8): 870-879.

[2] 荚 伟. 基于遗传算法的软件结构测试数据生成技术研究[J]. 北京航空航天大学学报,1997,23(1):36-40.

[3] 荚 伟. 遗传算法在软件测试数据生成中的应用[J]. 北京航空航天大学学报,1998,24(4):434-436.

[4] 潘祖烈. 基于遗传算法的黑箱测试用例自动生成模型[J]. 计算机工程,2008,34(9):205-210.

[5] 赵 明. 基于遗传算法的测试用例生成工具研究[J]. 计算机工程,2005,31(13):151-153.

[6] 伦立军. 基于遗传算法的测试数据生成研究[J]. 计算机工程,2005,31(23):82-84.

[7] 金 虎. 基于面向路径的遗传算法的测试用例自动生成[J]. 计算机工程,2007,33(3):21-23.

[8] 高海昌. 改进的遗传算法在测试数据自动生成中的应用[J]. 系统工程与电子技术,2006,28(7):1077-1081.

[9] 王元珍. 产生多条路径上测试用例的改进遗传算法[J]. 计算机工程,2006,32(13):196-205.

[10] Chaiyarataiia N. Recent Developments in Evolutionary and Genetic Algorithms: Theory and Applications[C]// Second International Conference on (Conf. Publ. No. 446) Genetic Algorithms in Engineering Systems: Innovations and Applications. [s. l.]: [s. n.], 1997: 270-277.

[11] Zeng L Z, Boualem B, Ngu A H H, et al. QoS-aware middleware for web services composition[J]. Software Engineering, 2004, 30(5): 311-327.

[12] 陈 亮,孙 敏. 基于免疫遗传算法的 Web 服务组合方法[J]. 计算机工程,2010,36(10):226-230.

K-Means算法的研究与改进

作者： 周爱武， 陈宝楼， 王琰
作者单位： 安徽大学 计算机科学与技术学院,安徽 合肥 230039
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(10)