遗传优化的 K 均值聚类算法①

胡　彧 (太原理工大学 测控技术研究所 山西 太原 030024)

毕晋芝 (太原理工大学 计算机与软件学院 山西 太原 030024)

摘　要： 在 K 均值聚类算法中，K 值需事先确定且在整个聚类过程中不能改变其大小，而按照经验 K 值划分所得的最终聚类结果一般并非最佳结果。通过求解所构造适应度函数的值，在变异操作中实现最佳聚类数 K 值的自动寻优，同时借助遗传操作完成聚类中心点的优化选取并利用遗传算法的全局寻优能力克服了 K 均值聚类算法的局部性。通过对 Iris 等数据集的实验分析，证明该算法具有良好的全局收敛性，且通过 K 值的自动调整，有效提高了聚类结果的划分。

关键词： K 均值算法；K 均值遗传算法；遗传算法；聚类算法；数据挖掘

Optimized K-Means Clustering Analysis Based on Genetic Algorithm

HU Yu[1], BI Jin-Zhi[2]

(1. The Institute of Measuring and Controlling Technology, Taiyuan University of Technology, Taiyuan 030024, China; 2. College of Computer and Software, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: For K-Means clustering algorithm, the k value must be determined in advance and can't be changed. However, the value is usually not the best if it is determined by experience. In this paper, fitness is taken into account to look for optimal number automatically in the mutation operations. Also, genetic operation is used to select the centers accordingly. In addition, the global optimization capability of genetic algorithm can overcome the locality of K-Means clustering algorithm. The experimental results show that this algorithm has better global searching capability and can efficiently improve the clustering result by adjusting the k value automatically.

Keywords: k-Means algorithm; the genetic k-Means algorithm; genetic algorithm; clustering algorithm; data mining

1　概述

聚类(cluster)做为数据挖掘技术的主要研究领域之一，近年来被广泛应用于各行各业。聚类分析方法做为一种无监督的学习方法，采用“物以类聚”的思想，将数据对象按某些属性分组成为多个类或簇，并且使得同类或簇中数据对象相似度尽可能大，而不同类或簇之间的差异尽可能大。K 均值聚类算法是聚类分析中一种基本的划分方法，因其思想可靠，算法简洁，而且能有效的应用于大数据集而被广泛使用。但是传统的 K 均值聚类算法往往受初始中心点选取的影响并且常常终止于局部最优。针对上述缺点，将遗传算法引入到 K 均值聚类算法中，通过遗传算法的一系列遗传操作实现对 K 均值聚类算法的改进。

目前基于遗传算法的 K 均值聚类算法主要是针对聚类中心点进行优化选取，或使 K 值能向最佳聚类数学习的问题，如：Murhty[1]，Sanghamitra Bandyopadhyay[2]，通过改进染色体编码与适应度函数,从而有效优化了 K 个中心点的选取。而傅景广[3]通过遗

① 基金项目:山西省自然科学基金(2009011019-2)
收稿时间:2009-10-10;收到修改稿时间:2009-11-14

传操作优化聚类中心点选取的同时采用特征向量来判断聚类划分的质量，使能得到聚类划分效果好的聚类中心点。刘婷[4]借助于类别数的上界的实验结果，通过染色体编码对应其类别数的方式来得到最佳聚类数，但是当数据量很大时，这种编码方式也将变的非常复杂。在本文中采用染色体基因值直接对应于 K 个聚类中心点的编码方式，通过求解所构造适应度函数值，在变异操作中完成 K 值的自动学习，同时也借助于遗传算法全局优化能力克服了 K 均值聚类算法的局部性，并且通过一系列遗传操作来优化聚类中心点的选取。

2 K均值聚类算法基本思想

K 均值聚类算法是一种基于划分方法的经典聚类算法之一，该算法的核心思想如下：首先从所给 n 个数据对象中随机选取 k 个对象作为初始聚类中心点，然后对于所剩下的其它对象，则根据它们与所选 k 个中心点的相似度(距离)分别分配给与其最相似的聚类，然后在重新计算所获聚类的聚类中心(该聚类中所有对象的均值)，不断重复这一过程直到标准测度函数开始收敛为止，其基本算法流程如下：

1) 从n个数据对象中任意选择k个对象作为初始聚类中心。

2) 根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离，并根据最小距离对相应对象进行划分。

3) 重新计算每个(有变化)聚类的均值(中心对象)。

循环上述流程 2 到 3，直到每个聚类不再发生变化或者标准测度函数开始收敛为止。

3 遗传优化的K均值聚类算法

3.1 遗传优化的 K 均值聚类算法

针对 K 均值聚类算法最佳聚类数难以确定，并且容易陷入局部最优的缺点。本文将具有自适应全局优化搜索能力的遗传算法引入到 K 均值聚类算法中，通过计算所构造的适应度函数值来完成一系列遗传操作，在优化 K 个聚类中心点选取的同时，主要借助变异操作实现对于 K 均值聚类算法中 K 值的自动学习。

3.2 遗传优化 K 均值聚类算法流程

1) 初始化，设置相关参数，如：初始聚类数 k，种群大小 m，交叉概率 p_c，变异概率 p_m，最大迭代次数 t 。

2) 随机产生初始种群。

3) 以种群个体为中心，采用 K 均值聚类算法对数据进行分类。

4) 计算种群个体的适应度值。

5) 执行选择、交叉、变异操作，产生新一代群体。

6) 重复执行(3)–(5)，直到达到最大迭代次数。

7) 计算种群个体的适应度值，以适应度值最大的个体进行分类输出。

3.2.1 适应度函数的构造

适应度函数做为衡量个体性能好坏的重要指标，在遗传进化过程中是对个体进行优胜劣汰的主要依据，在本文中适应度函数的选取不仅关系到下一代种群的优良性及数量，而且直接影响最佳 K 值的学习。在此将适应度函数定义如下：

$$f = \frac{D_{min}}{C(x)} \tag{1}$$

其中 D_{min} 为最小类间距，而 C(x)为平均类内距，其定义分别如下：

$$D_{min} = \min_{i,j=1}^{k} \left\| c_i - c_j \right\|^2 \tag{2}$$

$$C(x) = \frac{1}{k} \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} \left\| x_j - c_i \right\|^2 / n_i \right) \tag{3}$$

此适应度函数主要体现类间距应尽可能松散，而类内距应尽可能紧凑，即 D_{min} 应尽可能大，C(x)应尽可能小。在整个进化过程中，种群个体的长度即为 K 值的大小，如果 K 值小于最佳聚类数，随着 K 值的增加，D_{min} 与 C(x)都在减小，但聚类划分未达最佳，C(x)的变化明显大于 D_{min}，即适应度函数值增大。否则当 K 值大于最佳聚类数，由于聚类划分基本完成，此时 C(x)变化不明显，而因在基本聚类的基础上进行二次聚类，类间距已很小，所以 D_{min} 也会很小，整个适应度函数值变小。因此，此适应度函数在初始中心被优化的情况下，可以启发 K 值自动向最佳聚类数

学习。

3.2.2 最佳 K 值的调整

在本文中由遗传操作中的变异操作来控制 K 值的大小。变异操作的本质是挖掘群体中个体的多样性，同时提高算法的局部随机搜索能力，防止出现未成熟收敛[5]。在此通过对个体适应度函数的求解，决定聚类数 K 值的变化方向。由于最初所给的聚类数 K 值并非是最佳聚类数，因此将最初所给种群中具有最大适应度值的个体做为最佳聚类数的榜样个体，其它个体的长度(即 K 值)向榜样个体的长度靠扰。在此过程中，主要完成以下两方面工作：

1) 初始种群个体长度相同，在初次变异时，假设 K 值呈增长趋势，在设置初始聚类数 K 值时可以相对小些。然后再次变异时，如果变异个体长度小于榜样个体长度，说明最佳聚类数向增大方向发展，可在个体中增加一点，否则就减少一点。

2) 对于增加点，本文选取数据集中离目前个体中最大聚类数中心点距离最远的点加入个体中。而减少的点，应删除个体中与目前聚类数目最小的聚类中心点最近的点。

3.3 算法的实现

3.3.1 初始种群的确定

对于 K 均值聚类算法，K 值往往是事先凭借经验指定的，但是在某些应用方面根据经验所选取的 K 值一般并不是最佳聚类数，在本文中，事先指定一个 K 值，并且随机生成 K 个初始中心点做为初始个体，即如果给定的初始聚类数为 K，而数据维数为 N，则染色体的长度为 K*N。

3.3.2 染色体的编码

目前对于 K 均值遗传算法中对于染色体的编码主要有两种方式：1、将随机生成的 K 个初始聚类中心点做为染色体的基因值。2、将样本所属的聚类类别号做为染色体的基因值。在本文中采用第一种方法，即染色体的基因对应于 K 个初始聚类中心点，这样初始染色体的长度是固定的，但是随着 K 值的变化其长度又是可变的。由于聚类样本通常是多维的而且数据量也很大，在本文中针对染色体的编码选择浮点编码方式，如初始 K 值为 2，数据维数为 3 维。初始的 2 个中心点为(1，2，3)，(4，5，6)，则染色体编码为(1，2，3，4，5，6)，这样编码可以缩短染色体的长度，提高算法的效率。

3.3.3 选择操作

选择操作是根据适应度从群体中选择优良的个体，而将劣质个体进行淘汰，选择操作反映了“优胜劣汰”的进化准则，在本文中选择操作使用比例选择方法中最常用的方法：轮盘选择方法，其主要思想是：个体被选中的概率直接取决于该个体相对应的适应度值的大小。

3.3.4 交叉操作

交叉操作是指对一组将要进行交叉的染色体，按某种方式互相交换部分基因，从而形成新个体。在遗传算法中交叉操作是产生新个体的主要方法，并直接影响算法的全局搜索能力。本文中采用算术交叉(Arithmetical Crossover)，即：

$$x_i^{1'} = \alpha_i x_i^1 + (1-\alpha_i) x_i^2 \tag{4}$$

$$x_i^{2'} = (1-\alpha_i) x_i^1 + \alpha_i x_i^2 \tag{5}$$

其中 x_i^1、x_i^2 为父个体，$x_i^{1'}$、$x_i^{2'}$ 为产生的新个体，α_i 为[0，1]之间的任一随机数。

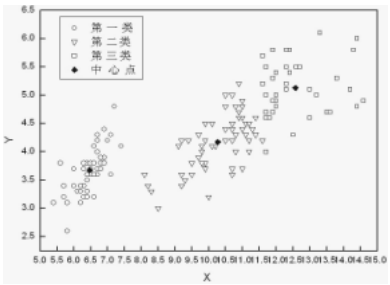
3.3.5 变异概率

对于变异概率的选取，由于刚开始 K 值选择具有不确定性，同时存在较大变数，因此开始时将变异概率选取大些，但是随着 K 值不断趋向于最佳 K 值，所以在后期应将变异概率减小。

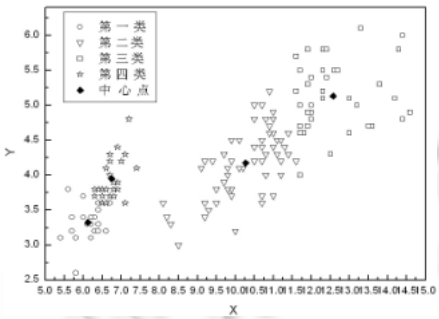
4 实验结果与分析

本实验操作系统采用的是 Windows 操作系统，实验环境主要应用 MyEclipse6.6 开发平台，采用 Mysql 做为后台数据库。分别将 iris 数据集与 glass 数据集[6]应用于原始的 K 均值聚类算法与本文所提出遗传优化的K均值聚类算法。其中iris数据集包含150个数据，标准聚类数为 3，每类包括 50 个数据，每个数据包含 4 个属性；而 glass 数据集包含 214 个数据，标准聚类数为 6 类，每个数据包含 9 个属性。在 K 均值聚类算法中，iris 数据集初始聚类数 K 为：3、4，对于 glass 数据集初始聚类数 K 为：5、6。而对于 K 均值遗传算法中参数设置如下：初始种群大小为 20，交叉概率 p_c 为 0.6，变异概率 p_{m1} 为 0.03，p_{m2} 为 0.001，iris 数据集初始聚类数 K 值为：2、3、4，

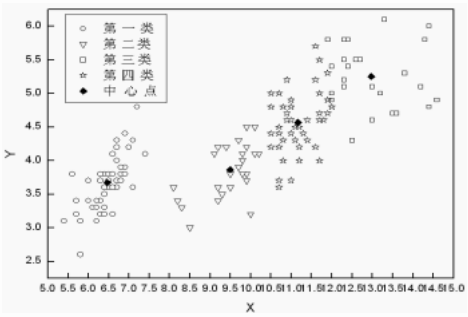
glass 数据集初始聚类数 K 为：4、5、6。



(a) k=3 时 iris 数据集分类情况



(b) k=4 时 iris 数据集分类情况



(c) k=4 时 iris 数据集分类情况

图 1 iris 数据分类情况

图 1 所示为 iris 数据集的分类情况，由图可以看出当 K 值为 4 时，K 均值聚类算法强行将本应属于同一类的数据分离为另一类，而且有可能将非常紧凑的一类分裂为两类(如图 1 中(b)、(c)所示)。

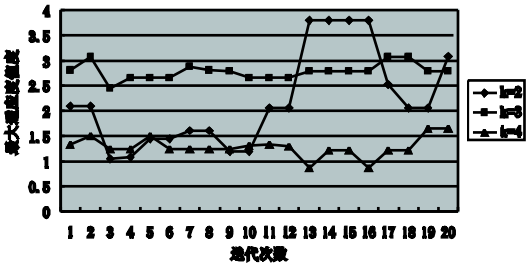


图 2 iris 数据集最大适值变化情况(迭代二十次)

图 2 为改进的 K 均值遗传算法应用于 iris 数据集迭代二十次，其最大适值的变化情况，可以看出根据最大适值的变化，K 值可以识别出最佳聚类数，并且向其靠近。

表 1 iris 数据集、glass 数据集聚类数与最小类内距比较

数据集	聚类数	初始聚类数	C(X) K-均值	最终聚类数	C(X) 新方法
iris	3	2	0.79419	3	1.08843
		3	0.634515	3	0.58023
		4	0.541498	5	0.52665
glass	6	5	1.371561	7	0.7998
		6	1.424232	6	0.693039
		7	1.210885	6	0.760484

从表 1 可以看出，改进的算法可以根据适值的变化识别最佳聚类数，并且自动向其学习。随着迭代次数的增加，平均类内距也变小，从而使所得聚类更紧凑。

5 结语

虽然改进的 K 均值遗传算法克服了 K 均值聚类算法的局部性，有效改进了聚类中心点的选取，并且使 K 值可以向最佳聚类数学习，但是也应该看到适应度函数的构造与变异概率的改变将会直接影响 K 值的学习过程，这也是某些最终 K 值会偏大的原因，而且对于随机选取的初始中心不同，也会直接影响其后的一系列操作。

参考文献

1 Murthy CA, Chowdhury N. In search of optimal clusters using genetic algorithms. Pattern Recognition Letter, 1996,17(8):825－832.

2 Sanghamitra Bandyopadhyay, Ujjwal Maulik. An evolutionary technique based on K-Means algorithm for optimal clustering . Information Sciences, 2002, 146(4):221－237.

3 傅景广,许刚,王裕国.基于遗传算法的聚类分析.计算机工程, 2004,30(4):122－124.

4 刘婷,郭海湘,诸克军,高思维.一种改进的遗传 k-means 聚类算法.数学的实践与认识, 2007,37(8):104－111.

5 潘伟,刁华宗.一种改进的实数自适应遗传算法.控制与决策, 2006,21(7):792－793.

6 赖玉霞,刘建平,杨国兴.基于遗传算法的 K 均值聚类分析.计算机工程, 2008,34(10):200－202.