一种基于遗传算法的 K-means 聚类算法

王 娟
(贵州民族学院 计算机与信息工程学院,贵州 贵阳 550025)

摘 要:传统 K-means 算法对初始聚类中心的选取和样本的输入顺序非常敏感,容易陷入局部最优。针对上述问题,提出了一种基于遗传算法的 K-means 聚类算法 GKA,将 K-means 算法的局部寻优能力与遗传算法的全局寻优能力相结合,通过多次选择、交叉、变异的遗传操作,最终得到最优的聚类数和初始质心集,克服了传统 K-means 算法的局部性和对初始聚类中心的敏感性。

关键词:遗传算法;K-means;聚类
中图分类号:TP18 **文献标识码**:A **文章编号**:1674-7720(2011)20-0071-03

A K-means clustering algorithm based on genetic algorithm

Wang Juan
(College of Computer and Information Engineering, Guizhou University for Nationalities, Guiyang 550025, China)

Abstract: Traditional K-means algorithm is sensitive to selecting initial clustering centers and input sequence, it is easy to get into the local best. In view of the above-mentioned problems, this paper proposes a K-means clustering algorithm(GKA) based on genetic algorithm. It combines local optimization of K-means algorithm with global optimization of genetic algorithm. By multiple selection, crossover and mutation, it can get optimal clustering number and initial centroid collection. So it overcomes the locality of traditional K-means algorithm and sensitivity of initial clustering centers.

Key words: genetic algorithm; K-means; clustering

聚类分析是一个无监督的学习过程,是指按照事物的某些属性将其聚集成类,使得簇间相似性尽量小,簇内相似性尽量大,实现对数据的分类[1]。聚类分析是数据挖掘技术的重要组成部分,它既可以作为独立的数据挖掘工具来获取数据库中数据的分布情况,也可以作为其他数据挖掘算法的预处理步骤。聚类分析已成为数据挖掘主要的研究领域,目前已被广泛应用于模式识别、图像处理、数据分析和客户关系管理等领域中。K-means 算法是聚类分析中一种基本的划分方法,因其算法简单、理论可靠、收敛速度快、能有效处理较大数据而被广泛应用,但传统的 K-means 算法对初始聚类中心敏感,容易受初始选定的聚类中心的影响而过早地收敛于局部最优解,因此亟需一种能克服上述缺点的全局优化算法。

遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化搜索算法。在进化过程中进行的遗传操作包括编码、选择、交叉、变异和适者生存选择。它以适应度函数为依据,通过对种群个体不断进行遗传操作实现种群个体一代代地优化并逐渐逼近最优解。鉴于遗传算法的全局优化性,本文针对应用最为广泛的 K-means 方法的缺点,提出了一种基于遗传算法的 K-means 聚类算法 GKA(Genetic K-means Algorithm),以克服传统 K-means 算法的局部性和对初始聚类中心的敏感性。

用遗传算法求解聚类问题,首先要解决三个问题:

(1)如何将聚类问题的解编码到个体中;

(2)如何构造适应度函数来度量每个个体对聚类问题的适应程度,即如果某个个体的编码代表良好的聚类结果,则其适应度就高;反之,其适应度就低。适应度函数类似于有机体进化过程中环境的作用,适应度高的个体在一代又一代的繁殖过程中产生出较多的后代,而适应度低的个体则逐渐消亡;

(3)如何选择各个遗传操作以及如何确定各控制参数的取值。

解决了这些问题就可以利用遗传算法来求解聚类问题,这也显示了遗传算法与求解问题无关的特性。

1 K-means 算法

K-means 聚类算法的目标是把包含 n 个对象的数据集 x 分为 k 个簇，使簇内具有较高的相似度，而簇间相似度较低。算法首先随机选择 k 个对象作为初始聚类中心，再计算剩余数据对象到各聚类中心的距离并将其赋给最近的簇，然后重新计算每个簇的平均值，不断重复此过程，直到准则函数收敛。

准则函数定义如下：

$$J=\sum_{i=1}^{k}\sum_{x_j\in c_i}||x_j-z_i||^2 \tag{1}$$

其中，k 为类别数，x_j 为样本对象，z_i 为类 c_i 的聚类中心。

K-means 聚类算法的描述如下：

(1)给定大小为 n 的样本数据集 x，类别数 k，从样本集中随机选择 k 个初始聚类中心 $z_j,j=1,2,\cdots,k$。

(2)计算剩余的每个数据对象与聚类中心的距离 $d(x_i,c_j),i=1,2,\cdots,n,j=1,2,\cdots,k$，如果满足 $d(x_i,c_m)=\min\{d(x_i,c_j),i=1,2,\cdots,n,j=1,2,\cdots,k\}$，则将 x_i 划分到类 c_m 中。

(3)根据划分后各集合中的点计算新的聚类中心，计算公式为：

$$c_j^*=\frac{1}{n_j}\sum_{x_m\in c_j}x_m \quad j=1,2,\cdots,k \tag{2}$$

其中，n_j 为类 c_j 中样本的个数。

(4)判断：如果 $c_j\neq c_j^*$，则用 c_j^* 取代 c_j，并返回(2)继续执行；否则表示数据划分不再变化，此时算法运行结束，当前中心点为最终的聚类划分结果。

2 基于遗传算法的 K-means 聚类算法(GKA)

GKA 的基本思想是：首先从要聚类的样本集选出初始种群，并对其执行遗传算法；对执行完遗传算法后产生的新种群执行 K-means 操作。如此反复循环，直到寻找出聚类问题的最优解。

2.1 染色体编码

遗传算法的编码方法分为三大类：二进制编码、符号编码和浮点数编码，其中二进制编码方法是遗传算法中最主要和常用的一种编码方法。由于聚类样本具有多维性、数据量大等特点，如果采用传统的二进制编码，染色体的长度会随着维数的增加或精度的提高而显著增加，从而使得搜索空间急剧增大，大大降低了计算效率，因此本文采用基于聚类中心的浮点数编码方法。

例如对于一个类别为 3 的聚类问题，假设数据集为2维，初始的 3 个聚类中心点为(10,20)、(30,40)和(50,60)，则染色体编码为(10,20,30,40,50,60)。这种基于聚类中心的编码方式意义明确、直观，缩短了染色体的长度，提高了运算效率，对于求解大量数据的复杂聚类问题效果较好。

2.2 初始化种群

初始群体完全随机生成。首先从样本空间中随机选出 k 个个体，每个个体表示一个初始聚类中心，然后根据所采用的编码方式将这组个体(聚类中心)编码成一条染色体。然后重复进行 m 次染色体初始化(m 为种群大小)，直到生成初始种群。

2.3 适应度函数的设计

适应度函数[2]是用来评价个体的适应度、区别群体中个体优劣的标准。个体的适应度越高，其存活的概率就越大。本文依据准则函数 J 构造适应度函数，由于 J 越小说明聚类划分的质量越好，J 越大说明聚类划分的质量越差，因此设计如下的适应度函数：

$$f=\frac{1}{1+J} \tag{3}$$

由式(3)可以看出，目标函数值越小的聚类中心，其适应度越大；目标函数值越大的聚类中心，其适应度越低。

2.4 遗传操作

2.4.1 选择操作

遗传算法使用选择操作来实现对群体中的个体进行优胜劣汰操作：适应度高的个体被遗传到下一代群体中的概率大；适应度低的个体被遗传到下一代群体中的概率小。本文采用锦标赛选择法，首先，随机地从种群中挑选一定数目的个体，然后从中选出适应度最大的个体作为父个体，重复迭代以上步骤直到父个体的总数达到种群规模。

2.4.2 交叉操作

交叉操作[2]是指对两个相互配对的染色体按某种方式相互交换部分基因，从而形成两个新的个体。由于本文采用的是浮点数编码方式，因此采用适合浮点数编码的算术交叉算子。

算术交叉是指由两个个体的线性组合而产生出两个新的个体。假设在两个个体 X_1 和 X_2 之间进行算术交叉，则交叉后产生的新个体为：

$$\begin{cases}X_1'=\alpha X_2+(1-\alpha)X_1\\X_2'=\alpha X_1+(1-\alpha)X_2\end{cases} \tag{4}$$

其中，α 是一个参数，可以是常数(此时为均匀算术交叉)，也可以是一个由进化代数所决定的变量(此时为非均匀算术交叉)。

2.4.3 变异操作

变异[2]是指将个体染色体编码串中的某些基因座上的基因值用该基因座的其他等位来替换，从而形成一个新的个体。变异的目的是改善遗传算法的局部搜索能力；维持群体的多样性，防止早熟收敛。本文采用均匀变异算子，其具体操作过程是：

(1)依次指定个体编码串中的每个基因座为变异点，并确定每个基因点的取值范围$[U_{\min},U_{\max}]$；

(2)对每一个变异点，以变异概率 P_m 从对应基因的取值范围内取一个随机数来代替原有值。其中变异点的新基因值为：

$$X_i' = U_{min} + r(U_{max} - U_{min})$$

其中，r 为(0，1)范围内符合均匀概率分布的一个随机数。

2.5 K-means 优化操作

由于 K-means 是一种局部搜索能力强的算法，本文算法在每一代执行完遗传操作后引入了 K-means 算法中的一个操作步骤 K-means 操作，对新生种群中的每个个体进行 K-means 优化，优化后的群体作为下一代种群进入演化。这样不仅可以提高混合算法的局部搜索能力，同时也有利于提高其收敛速度。具体的优化操作如下：先以变异后产生的新群体的编码值作为中心，把每个数据对象分配到最近的类，形成新的聚类划分；然后计算新的聚类中心，取代原来的编码值；经 K-means 优化操作后产生新一代种群开始下一轮遗传操作。

2.6 算法设计

基于遗传算法的 K-means 聚类算法(GKA)流程描述如下：

(1)设置遗传参数：聚类数 k，种群规模 m，最大迭代次数 T，交叉概率 P_c，变异概率 P_m；

(2)种群初始化：从样本中随机选取 k 个点作为聚类中心并进行编码，重复 m 次，产生初始种群；

(3)计算群体中各个体的适应度；

(4)通过选择、交叉、变异、K-means 操作，产生新一代群体；

(5)重复步骤(3)和步骤(4)，直到达到最大迭代次数 T；

(6)计算新一代群体的适应度，以最大适应度的最佳个体为中心进行 K-means 聚类；

(7)输出聚类结果。

3 实验

为了验证算法的有效性，本文对 K-means 算法和 GKA 算法进行了对比实验。在 Matlab 环境下分别编写 K-means 算法和 GKA 算法，导入数据进行实验。实验数据来自 KDD CUP[3]，数据集分别是 iris 和 wine。其中，iris 包含 150 个数据，分为 3 类，每类 50 个数据，每个数据包含 4 个属性；wine 数据集包含 178 个数据，分为 3 类，每个数据包含 13 个属性。本文算法的参数设置如下：种群大小 m=30，算法的最大迭代次数 T=50，交叉概率 P_c=0.9，变异概率 P_m=0.001。所有算法各运行 20 次，运行结果如表 1 所示。

表 1 K-means 和 GKA 算法比较

数据集	算法	平均聚类准确度/%	达到最优解的平均迭代次数
iris	K-means	96.7	5.9
	GKA	98	6.7
wine	K-means	84.3	9.5
	GKA	95.5	4.3

从表 1 可以看出，K-means 算法对初始聚类中心的选取敏感性很大，容易陷入局部最小值，并不是每次都能得到最优解，特别是对于 wine 这种较高维度的数据集，有时聚类准确度不够理想。除数据集 iris 外，K-means 算法每组数据收敛到最优解的平均迭代次数都比 GKA 算法多，所以 GKA 算法的收敛速度也较快。

本文针对应用最为广泛的 K-means 算法的缺点，提出了一种基于遗传算法的 K-means 聚类算法 GKA，将 K-means 算法的局部寻优与遗传算法的全局寻优相结合，通过多次选择、交叉、变异的遗传操作，最终得到最优的聚类数和初始质心集，克服了传统 K-means 算法的局部性和对初始聚类中心的敏感性。实验表明，GKA 算法在聚类准确度和收敛速度上均比 K-means 算法更优。

参考文献

[1] 韩家炜，堪博.数据挖掘：概念与技术[M].北京：机械工业出版社，2007.

[2] 吴多比.数据挖掘中基于遗传算法的聚类方法应用研究[D].重庆：重庆大学，2009.

[3] UCI Mechine Learning Repository[EB/OL].http://archive.ics.uci.edu/ml/datasets.html.

(收稿日期：2011-07-18)

作者简介：

王娟，女，1983 年生，讲师，硕士，主要研究方向：数据挖掘、网络安全。