基于遗传算法的 K 调和均值聚类算法

李家成,苏一丹,覃 华,吴 丹
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:K 调和均值算法(KHM)用数据点与所有聚类中心的距离的调和平均值替代了数据点与聚类中心的最小距离,是一种对初始值不敏感、收敛速度快的有效聚类算法,但它容易陷入局部最小值。而遗传算法具有良好的全局优化能力。文中结合了 KHM 和遗传算法各自的优点,采用 KHM 计算每一代种群的聚类中心,并构造适应度函数,通过遗传算法进行一系列择优操作,成功地解决了 KHM 容易陷入局部最小值的问题。实验结果表明,所提出的算法不仅优化了聚类中心,而且还改善了聚类质量。

关键词:遗传算法;K 调和均值;聚类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2013)09-0055-04
doi:10.3969/j.issn.1673-629X.2013.09.014

K-Harmonic Means Clustering Algorithm Based on Genetic Algorithm

LI Jia-cheng,SU Yi-dan,QIN Hua,WU Dan
(College of Computer Science and Electronics Information,Guangxi University,Nanning 530004,China)

Abstract:In K-harmonic means clustering was an effective algorithm which was not sensitive to the initial value and converged quickly, it used harmonic means distance from the data point to all clustering centers to replace the minimum distance between the data point and all clustering centers. But it also easily converged to the local minimum,and genetic algorithm had a good global optimal capacity. Combined the advantages of KHM and genetic algorithm,used the KHM to calculate the clustering center of every population,and structure fitness function,through the genetic algorithm conduct a series of preferential operation,successfully solved the problem of KHM easily converged to the local minimum. The experiment showed the algorithm not only optimized the cluster centers,but also improved the cluster quality.

Key words:genetic algorithm;K-harmonic mean;clustering

0 引 言

聚类是数据挖掘和非监督机器学习的一项重要技术,它根据数据的内在性质将数据划分为若干类,使同一类中的对象尽可能相似,不同类间对象的差异性尽可能大[1]。根据聚类采用的方法可分为:层次聚类、基于网络的聚类、划分聚类、基于密度的聚类以及基于模型的聚类[2]。KHM 是一种类似于 K-means 的聚类算法,属于划分聚类,该算法用数据点与所有聚类中心的距离的调和平均值代替了 K-means 算法中数据点与聚类中心的最小距离[3],成功地解决了 K-means 算法对初值敏感的问题。

KHM 实现简单,收敛速度快,但容易陷入局部最优。因此,很多学者将各种优化方法引入 KHM 聚类算法中。沈明明等人提出了融合 K-调和均值的混沌粒子群聚类算法[3]。毛力等人提出了融合 K-调和均值和模拟退火粒子群的混合聚类算法[4]。赵恒等人提出了一种基于调和均值的模糊聚类算法[5]。刘国丽提出了基于模拟退火的 K 调和均值聚类算法[6]。

遗传算法(GA)是由美国 Holland 教授在 1975 年提出的,该算法是一种自适应全局搜索算法,具有较强的鲁棒性和全局寻优能力[7-8]。很多学者利用遗传算法进行聚类[9-10],也有学者把遗传算法和 K-means 算法相结合[11-14]。目前,还没有人把遗传算法应用在 K 调和均值聚类算法中。鉴于此,文中提出了基于遗传算法的 K 调和均值(GAKHM)聚类算法。

收稿日期:2012-12-06 **修回日期**:2013-03-08 **网络出版时间**:2013-05-09
基金项目:教育部人文社会科学研究项目(11YJAZH080)
作者简介:李家成(1985-),男,硕士研究生,研究方向为智能系统与智能 CAD;苏一丹,博士,教授,研究方向为电子商务、智能系统与智能 CAD;覃 华,博士,副教授,研究方向为电子商务、数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130509.1059.043.html

1 K-调和均值(KHM)算法

KHM 聚类算法是基于中心迭代过程的算法。假设 $Z = \{Z_1, Z_2, \cdots, Z_n\}$ 是一组数据元组,其中 $Z_i = (Z_{i,1}, Z_{i,2}, \cdots, Z_{i,m})$ 表示具有 m 个属性的数据对象。聚类个数为 k,聚类中心集合为 $C = \{C_1, C_2, \cdots, C_k\}$。

(1)KHM 算法中用调和平均函数取代了 KM 算法中的最小距离函数,调和平均公式为:

$$\frac{k}{\sum_{c \in C} \frac{1}{d^2(z,c)}} \tag{1}$$

其中,$z \in Z$ 代表数据集中的对象;$c \in C$ 代表聚类中心;$d^2(z,c)$ 是距离测度。

(2)KHM 算法目标函数为:

$$E_{\mathrm{KHM}} = \sum_{z \in Z} \frac{k}{\sum_{c \in C} \frac{1}{d^2(z,c)}} \tag{2}$$

即为所有数据点到每个聚类中心的调和平均值的和。

(3)中心迭代公式为:

$$c_k = \frac{\sum_{z \in Z} \frac{1}{(\sum_{y \in C} \frac{d^2(z,c_k)}{d^2(z,y)})^2} z}{\sum_{z \in Z} \frac{1}{(\sum_{y \in C} \frac{d^2(z,c_k)}{d^2(z,y)})^2}} \tag{3}$$

2 基于遗传算法的 K 调和均值聚类算法

由于 KHM 聚类算法容易陷入局部最优值,因此,文中将具有自适应全局优化搜索能力的遗传算法引入到 KHM 聚类算法中,通过计算适应度函数值来进行一系列遗传操作,对 K 个聚类中心点进行优化,利用变异操作来完成对 KHM 聚类算法中 K 值的自动学习。

2.1 初始种群确定

在 KHM 聚类算法中,K 值一般是事先通过经验设置的,但是,通过经验所获取的 K 值,在某些应用中,往往并不是最好的聚类数。在文章中,事先设置一个 K 值,然后随机生成 K 个初始中心值作为初始个体。即,如果把初始聚类数设置为 K,把数据维数设置为 M,那么染色体的长度等于 $K * M$。

2.2 构造适应度函数

适应度函数是对遗传进化过程中的个体进行优胜劣汰的主要标准,也是评价个体性能好坏的主要根据。在文章中,对适应度函数的选取,直接影响到最佳 K 值的学习和下一代种群的优良性及数量。在此,适应度函数定义如下:

$$f = \frac{D_{\min}}{E_{\mathrm{KHM}}} \tag{4}$$

其中,$D_{\min}$ 是最小类间距离;E_{KHM} 为 KHM 算法目标函数。

2.3 染色体编码

在聚类分析中,常用的染色体编码有:基于聚类划分的整数编码、基于聚类中心的浮点数编码。通常情况下,聚类具有数据量大、多维性等特征,使得聚类样本数量远远大于其聚类数目。因此,文中采用基于聚类中心的浮点数编码,对每个聚类中心进行编码。假设把初始 K 值设为 3,数据维数设为 3,把 3 个聚类中心分别初始化为(1,2,3),(4,5,6),(7,8,9),则染色体的编码为(1,2,3,4,5,6,7,8,9)。这种编码,不仅可以把染色体的长度缩短,而且还提高了算法的效率。

2.4 选择操作

选择操作反映了自然界“优胜劣汰”的原理,根据适应度值的大小,从种群中选择优秀的个体,淘汰较差的个体。在文章中,使用“轮盘选择方法”完成选择操作,其主要思想为:个体适应度值的大小决定该个体被选中的概率,适应度值越大的个体被选中的概率就越大;否则,被选中的概率就越小。

2.5 交叉操作

在遗传算法中,主要采用交叉操作产生新个体。交叉操作是按照某种方式,对一组将要进行交叉的染色体相互交换其部分基因,从而形成新个体。交叉操作直接影响算法的全局搜索能力。在文章中,采用算术交叉(Arithmetical Crossover)进行交叉操作,即:

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

$$z_i^{2'} = \alpha_i z_i^2 + (1 - \alpha_i) z_i^1 \tag{6}$$

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

2.6 变异操作

变异是一种局部随机搜索,它根据自然界生物变异的原理,改变染色体的某些基因位,从而保持种群的多样性,防止“早熟”现象的产生。在文章中,采用“均匀变异算子”对染色体进行变异操作,操作过程为:对每个变异点,随机生成一个基因取值范围内的数代替原来的基因值。即:

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

其中,r 为(0,1) 范围内的随机数;$U_{\max}$ 表示该基因位最大值;$U_{\min}$ 表示该基因位最小值。

2.7 算法流程

算法流程见图 1。

3 实 验

3.1 聚类结果评价方法

F-measure 是由信息检索领域中的查准率(Precision)和查全率(Recall)进行综合得到的,文中采用 F-

measure 对聚类结果进行评价。对已知分类 i 和任意聚类 j,其查准率和查全率定义为:

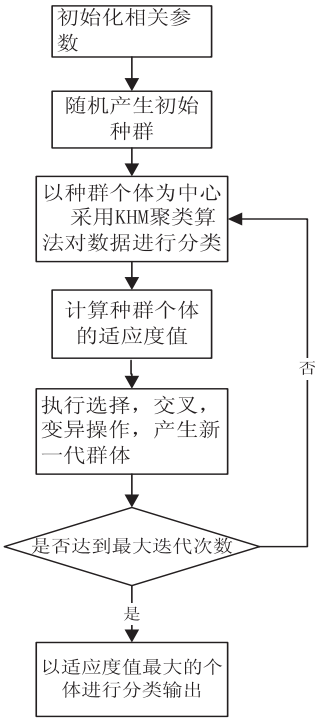

图 1　算法流程

$$P(i,j)=\frac{M_{ij}}{M_i} \tag{7}$$

$$R(i,j)=\frac{M_{ij}}{M_j} \tag{8}$$

公式中的 M_i 表示分类 i 中的对象数目,M_j 表示聚类 j 中的对象数目,M_{ij} 表示在聚类 j 中属于已知分类 i 的对象数目。分类 i 的 F - measure 可定义为:

$$F(i)=\frac{2PR}{P+R} \tag{9}$$

公式中的 $P=P(i,j)$,$R=R(i,j)$。把 $F(i)$ 看作系统对分类 i 的评分,把具有最高 $F(i)$ 值的聚类当作相应的分类 i。总的 F - measure 是每个分类 i 的加权平均值,即:

$$\text{F - measure}=\frac{\sum_{i\in N}|i|\times F(i)}{\sum_{i\in N}|i|} \tag{10}$$

其中,N 表示所有的已知分类,$|i|$ 表示分类 i 中的对象数目。F - measure 越大,说明聚类效果越好。

3.2　实验结果及评价

该实验在 Win7 系统下,采用 matlab,C++编程语言完成。为了测试 GAKHM 聚类算法的性能,实验采用 Iris,Glass,Wine 数据集作为测试样本。其中 Iris 数据集共有 150 个实例,每个实例有 4 个属性。Iris 数据集的实际聚类中心位置分别为:(5.00　3.42　1.46　0.24),(5.93　2.77　4.26　1.32),(6.58　2.97　5.55　2.02)。Iris 数据集分为 3 类,每类各有 50 个实例,其中一类与其他两类有较好的分离,而另外两类之间存在交迭。

在相同的实验环境下,分别对 KHM,遗传 K 均值(GAKM),GAKHM 聚类算法进行 20 次实验,并用 F-measure 评价函数对三种算法所得的聚类结果进行评价。实验中设定种群规模 $p_{size}=50$,交叉概率 $P_c=0.75$,变异概率 $P_m=0.15$,最大迭代次数 MaxGen = 100。表 1 和表 2 是三种算法进行 20 次实验的聚类结果。

表 1　三种算法的平均聚类中心与实际值的差异(Iris)

聚类算法	平均聚类中心				与实际中心的误差平方和
KHM	(5.003 5	3.403 0	1.484 9	0.251 5)	0.075 0
	(5.889 1	2.761 2	4.364 1	1.397 3)	
	(6.775 1	3.052 4	5.646 9	2.053 6)	
GAKM	(5.006 7	3.407 1	1.482 5	0.253 8)	0.895 3
	(6.044 1	2.809 6	4.623 4	1.562 7)	
	(7.182 8	3.128 6	6.087 8	2.125 0)	
GAKHM	(5.001 7	3.404 0	1.474 8	0.246 6)	0.059 4
	(5.753 4	2.683 7	4.216 7	1.239 2)	
	(6.656 8	3.036 2	5.518 0	2.043 6)	

表 2　三种算法的平均 F-measure

数据集	KHM	GAKM	GAKHM
Iris	0.894 4	0.810 3	0.917 9
Glass	0.807 1	0.677 1	0.849 4
Wine	0.813 2	0.728 4	0.853 9

由表 1 得出,在三种算法中,GAKHM 算法的平均聚类中心与实际中心的误差平方和最小,说明 GAKHM 算法所求出的聚类中心比其他两种算法更接近于实际聚类中心。由表 2 可知,在不同的数据集下,GAKHM 的平均 F-measure 值最大,表明 GAKHM 的聚类精确度比 KHM,GAKM 的高。

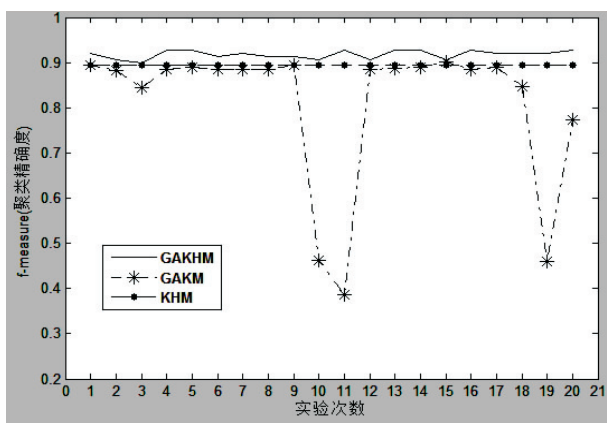

图 2　三种算法 20 次实验的 F-measure(Iris)

图 2,图 3 和图 4 分别是三种算法在数据集(Iris,Glass,Wine)下的 F-measure 曲线。由图可知,除了图

3 的一次实验外,在每次实验中,GAKHM 的 F-measure 都明显高于 GAKM 和 KHM 的 F-measure,表明 GAKHM 有较好的聚类效果。因为 GAKHM 和 KHM 均不受初值的影响,所以图中 GAKHM 和 KHM 的 F-measure 曲线波动很小;而 GAKM 受初值的影响,所以它的 F-measure 曲线波动较大。说明 GAKHM 和 KHM 的聚类结果比 GAKM 的聚类结果稳定。综上可得,GAKHM 在聚类中心优化,聚类精确度,算法稳定性上都比其他两种算法占优势。

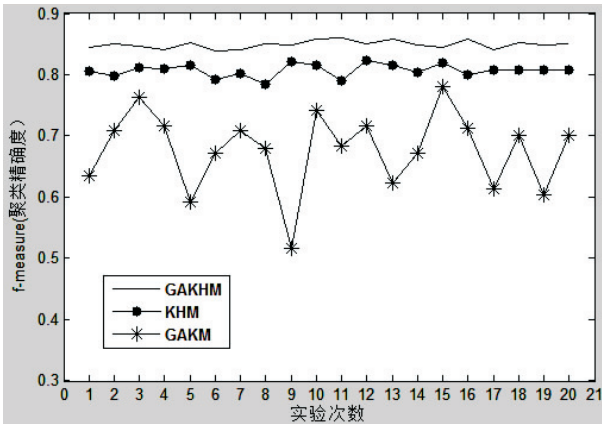

图 3 三种算法 20 次实验的 F-measure(Glass)

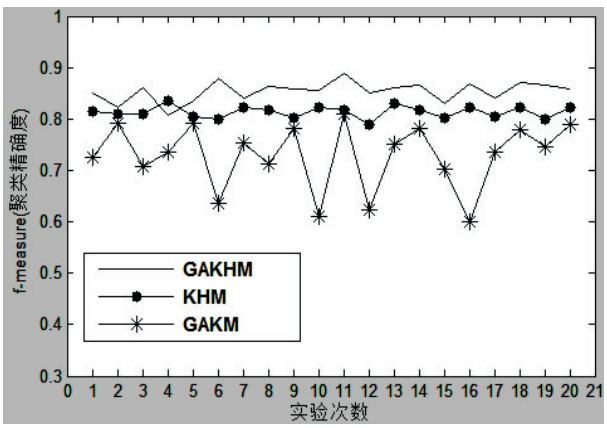

图 4 三种算法 20 次实验的 F-measure(Wine)

4 结束语

文中根据 K 调和均值和遗传算法各自的优缺点,提出了一种新的算法:基于遗传算法的 K 调和均值聚类算法(GAKHM)。接着通过实验证明了 GAKHM 优于 KHM 和 GAKM,最终得出 GAKHM 是一种聚类精确度高,稳定,能够优化聚类中心的聚类算法。虽然 GAKHM 是一种较好的聚类算法,但也存在缺点,比如要事先确定 K 值,并且在做实验的时候,GAKHM 算法的运行速度明显慢于 KHM 和 GAKM,说明 GAKHM 有较大的时间复杂度。接下来将对 GAKHM 的这些缺陷进行改进。

参考文献:

[1] 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,2006.

[2] 陆林花. 一种新的基于遗传算法的动态聚类算法[J]. 计算机仿真,2009,26(7):122-125.

[3] 沈明明,毛 力. 融合 K-调和均值的混沌粒子群聚类算法[J]. 计算机工程与应用,2011,47(27):144-146.

[4] 毛 力,刘兴阳,沈明明. 融合 K-调和均值和模拟退火粒子群的混合聚类算法[J]. 计算机与应用化学,2011,28(2):177-180.

[5] 赵 恒,杨万海. 一种基于调和均值的模糊聚类算法[J]. 电路与系统学报,2004,9(5):114-117.

[6] 刘国丽,甄晓敏. 基于模拟退火的 K 调和均值聚类算法[J]. 计算机系统应用,2011,20(7):90-93.

[7] 徐家宁,张立文,徐素莉,等. 改进遗传算法的 k 均值聚类算法研究[J]. 微计算机应用,2010,31(4):11-15.

[8] Fogel D B. An introduction to simulated evolutionary optimization[J]. IEEE Trans. on Neural Network,1994,5(1):3-14.

[9] Bhuyan J N,Raghavan V V,Elayavalli V K. Genetic algorithm for clustering with an ordered representation[C]//Proc. of 4th Int. Conf. on Genetic Algorithms. San Mateo: Morgan Kaufman,1991:408-420.

[10] Guo Haixiang,Zhu Kejun,Gao Siwei,et al. Animproved genetic k-means algorithm for optimal clustering[C]//Proc. of Sixth IEEE International Conference. Leipzig: IEEE Press,2006.

[11] Jones D R,Beltramo M A. Solving partitioning problems with genetic algorithms[C]//Proc. of 4th Int. Conf. on Genetic Algorithms. San Mateo:Morgan Kaufman,1991:442-457.

[12] 王 颖,刘建平. 基于改进遗传算法的 K-means 聚类分析[J]. 工业控制计算机,2011,24(8):78-79.

[13] 赖玉霞,刘建平,杨国兴. 基于遗传算法的 K 均值聚类分析[J]. 计算机工程,2008,34(20):200-202.

[14] 王 康,颜雪松,金 建,等. 一种改进的遗传 k 均值聚类算法[J]. 计算机与数字工程,2010,38(1):18-20.

(上接第 54 页)

on MIP table in IPv4/v6 mixed networks[C]//Proc. of International Conference on Computer Science and Network Technology. [s. l.]:[s. n.],2011:1026-1030.

[8] Jabid T,Kabir M H,Chae O. Facial expression recognition using Local Directional Pattern[C]//Proc. of ICIP. [s. l.]:[s. n.],2010:1605-1608.

[9] 柏 勇,何 春,王 赏. 基于 IP 组播的 MPLS 组播架构[J]. 通信技术,2010,43(10):62-64.

[10] Ko J,Park S,Lee E. An extended PIM-SM for efficient data transmission in IPTV services[C]//Proc. of IC-NIDC. [s. l.]:[s. n.],2010:115-119.

基于遗传算法的K调和均值聚类算法

作者： 李家成， 苏一丹， 覃华， 吴丹， LI Jia-cheng， SU Yi-dan， QIN Hua， WU Dan
作者单位： 广西大学 计算机与电子信息学院,广西 南宁,530004

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(9)