基于粒子群优化的模糊C-均值聚类算法研究

王纵虎 刘志镜 陈东辉
(西安电子科技大学计算机学院 西安 710071)

摘 要 针对用模糊C-均值聚类算法选择初始聚类中心敏感及模糊加权指数 m 对模糊C-均值聚类算法的聚类性能影响较大等问题,利用粒子群优化算法的全局寻优能力强及收敛速度较快的特点,结合模糊C-均值算法提出一种新的模糊聚类算法;采用了一种简单有效的粒子编码方法,将初始聚类中心和模糊加权指数 m 同时进行粒子群优化搜索,在得到最优适应度的同时,m 也收敛到一个稳定的最优解,从而有效地解决了上述问题。算法在人工合成数据集和多个UCI数据集上都取得了较好的效果。

关键词 聚类,模糊C-均值聚类,粒子群优化,粒子编码,初始聚类中心
中图法分类号 TP391 **文献标识码** A

Research of PSO-based Fuzzy C-means Clustering Algorithm

WANG Zong-hu LIU Zhi-jing CHEN Dong-hui
(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)

Abstract Fuzzy C-mean algorithm is sensitive to initial centroid and the choice of fuzzy weighting exponent m plays an import role in clustering result. PSO has the advantage of global optimization and good convergence speed. This paper proposed a new method by combining PSO and fuzzy C-Means to solve those problem. By a simple and effective particle encoding method, the best initial centroid and fuzzy weighting exponent were both searched in the process of PSO. Experiments on synthetic data sets and several UCI data sets achieve good results.

Keywords Clustering, FCM, PSO, Particle coding, Initial centroid

1 引言

聚类是数据挖掘、模式识别等研究的重要内容之一[1]。聚类是在事先不规定分组规则的情况下,将数据按照其自身特征划分为不同的群组,它是一种无监督的学习方法。聚类结果要求在不同群组间数据相似性越小越好,相同群组内数据相似性越大越好。

随着模糊集理论的提出,出现了用模糊集的方法来处理聚类问题的模糊聚类分析。在模糊聚类中,样本对象以一定的概率即隶属度隶属于某个类,与大多数硬划分聚类算法相比,这样的划分方式更能客观地反映现实世界。模糊C-均值(FCM)聚类算法最早由Dunn提出,并由Bezdek加以推广,它是基于目标函数的聚类算法中理论最为完善、应用最为广泛的一种方法。但是FCM算法是一种局部搜索算法,对初始聚类中心较敏感,容易陷入局部最优解。FCM算法需事先设置加权指数 m 的值,对于加权指数 m 对FCM算法性能的影响,Bezdek认为参数 m 控制着模糊类间的分享程度,并给出了一个经验范围[1.1,5],但没给出严格的证明[2]。如何选择合适的模糊加权指数 m 也是FCM算法要解决的关键问题之一。文献[3]基于模糊决策提出了一种 m 的优选方法,该方法同时考虑了样本结构的紧凑性与分离性,具有隶属度简单、不含参数等优点,但其得出的 m 值不稳定,不同类间相差太大;文献[4]通过引入一个新的隶属度约束函数,解决了IFP-FCM算法中 m 的一般化问题,但该方法只是解决了模糊加权指数的通用取值范围,并指出在图像分割领域 m 无论如何取值对聚类结果都无影响,而在其他应用领域 m 值的影响有待证明。文献[5]提出了一种基于模糊相关度的方法来研究 m 的最佳取值,在计算过程中,基于模糊相关度的聚类有效性变化反馈到 m 的变化中,通过多次递归迭代使 m 最终收敛到一个稳定值,但实验中需根据人为经验设置比例参数 r 的值。

粒子群优化算法是由美国的Kennedy和Eerhart受鸟群觅食行为的启发于1995年提出的。该算法是一种基于群体智能的具有全局寻优能力的优化工具,通过群体中粒子间的合作与竞争产生的群体智能指导优化搜索,每代种群中的解具有"自我"学习提高和向"他人"学习的双重优点,能够得到较优解,但是其局部搜索能力较弱。

针对模糊C-均值聚类算法存在的问题,有学者将群体智能优化算法引入聚类[6,7],以期达到全局寻优、快速收敛的目的。文献[6]中提出的基于粒子群优化算法的模糊C-均值聚

到稿日期:2011-10-20 返修日期:2012-02-17 本文受国家科技支撑计划项目(2007BAH08802),陕西省13115科技创新工程重大专项(2007ZDKG-57)资助。

王纵虎(1984-),男,博士生,主要研究方向为数据挖掘,E-mail:zonghuwang@gmail.com;**刘志镜**(1957-),男,教授.博士生导师,主要研究方向为数据挖掘和视觉计算;**陈东辉**(1984-),男,博士生,主要研究方向为数据挖掘。

类算法用PSO算法代替了FCM算法的基于梯度下降的迭代过程。文献[7]中提出的基于PSO的模糊聚类算法将PSO的聚类算法求得的解作为FCM的初始值，然后用FCM算法进一步求全局最优解。本文将模糊C-均值算法和粒子群优化算法相结合，采用一种简单有效的编码方式，同时将初始聚类中心和加权指数 m 进行编码，然后进行粒子群优化搜索，最后得到适应度最优的聚类结果，同时可以得到其对应的最优初始聚类中心和 m 值。已有的粒子群优化聚类算法都直接对 n 个聚类中心的 d 维样本向量进行编码，当样本维数较大时，计算量较大，且在进化过程中，d 维样本的值在一定范围内变化时可能会取到不符合样本实际情况的数值。由于各维的取值范围可能变化较大，造成粒子群搜索空间过大，在有限次迭代搜索时会影响算法的全局收敛性。在多个人工数据集与UCI数据集上的实验表明，本文提出的方法与传统的FCM算法和粒子群聚类算法的聚类结果相比，聚类结果的准确率和稳定性都有较大的提高。

2 模糊聚类算法

2.1 模糊C-均值聚类算法(FCM)

FCM把 n 个样本 $x_i(i=1,2,\cdots,n)$ 分为 c 个类簇，并计算出每类的聚类中心，使得目标函数达到最小。FCM用模糊划分使每个给定数据点用值在0,1间的隶属度来确定其属于各个类簇的程度。与引入模糊划分相适应，隶属矩阵 U 允许取值为0,1间的元素。令 u_{ij} 表示第 j 个样本点属于第 i 类的隶属度，则 u_{ij} 满足如下条件：

$$\sum_{i=1}^{c} u_{ij}=1,\ \forall j=1,\cdots,n \tag{1}$$

FCM的目标函数的一般化形式为：

$$J(U,c_1,\cdots,c_c)=\sum_{i=1}^{c} J_i=\sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

式中，u_{ij} 介于0,1之间，c_i 为模糊组 i 的聚类中心，$d_{ij}=\|c_i-x_j\|$ 为第 i 个聚类中心与第 j 个数据点间的欧氏距离，且 $m\in[1,\infty)$ 为加权指数。构造如下新的目标函数，可求得使式(2)达到最小值的必要条件：

$$\begin{aligned}\bar{J}(U,c_1,\cdots,c_c,\lambda_1,\cdots,\lambda_n)&=J(U,c_1,\cdots,c_c)+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)\\&=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^m d_{ij}^2+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)\end{aligned} \tag{3}$$

式中，$\lambda_j(j=1,\cdots,n)$ 是式(1)的 n 个约束式的拉格朗日乘子。对所有输入参量求导，使式(2)达到最小的必要条件为：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^m x_j}{\sum_{j=1}^{n}u_{ij}^m} \tag{4}$$

和

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{5}$$

由上述两个必要条件知，模糊C均值聚类算法是一个简单的迭代过程。FCM用下列步骤确定聚类中心 c_i 和隶属矩阵 U：

步骤1　用值在[0,1]间的随机数初始化隶属矩阵 U，使其满足式(1)中的约束条件。

步骤2　用式(5)计算 c 个聚类中心 c_i，$i=1,\cdots,c$。

步骤3　根据式(2)计算目标函数。如果聚类迭代次数大于预定的最大迭代次数或它相对上次目标函数值的改变量小于某个阈值，则算法停止。

步骤4　用式(5)计算新的 U 矩阵。返回步骤2。

上述算法也可以先初始化聚类中心，然后执行迭代过程。由于不能确保FCM收敛于一个最优解，算法的性能依赖于初始聚类中心，因此，我们要么用另外的快速算法确定初始聚类中心，要么每次用不同的初始聚类中心启动该算法，多次运行FCM。

基于目标函数的模糊C-均值聚类中，加权指数 m 对聚类过程和聚类结果有重要影响。随着 m 的增加，目标函数值单调下降，而且较大的 m 还有抑制噪声的功能，因此 m 取值越大越好；但同时，加权指数 m 控制着FCM聚类结果的模糊性，m 越大聚类结果越模糊，从这个角度又希望 m 的取值不要太大。

3 改进的粒子群优化模糊C-均值聚类算法

3.1 数据预处理

在对数据进行聚类前，本文采用式(3)对样本各维属性的值进行规范化，将其映射到[0,1]范围内；然后通过欧式距离计算样本之间的相异度，从而降低样本各维属性数值取值范围不一致对聚类结果的影响。

$$f'=\frac{f-\min_f}{\max_f-\min_f} \tag{6}$$

式中，$\min_f$ 为数据集中样本属性 f 的最小值，$\max_f$ 为数据集中样本属性 f 的最大值。

3.2 模糊粒子群聚类算法

在PSO算法中，粒子通过不断调整自己的位置来搜索新解，每个粒子都能记住自己搜索到的最优解 P_{id}，以及整个粒子群的最优解 P_{gd}。每个粒子都有一个速度，记作 V_{id}：

$$v_{id}(t+1)=\omega v_{id}(t)+c_1 rand()(p_{id}(t)-x_{id}(t))+c_2 rand()(p_{gd}(t)-x_{id}(t)) \tag{7}$$

式中，v_{id} 表示第 i 个粒子在第 d 维上的速度。ω 为用来控制粒子飞行速度的惯性权重，ω 越大，则粒子的飞行速度越大，粒子将以较大的步长进行全局探测；ω 越小，则粒子的速度步长越小，粒子将趋向于进行精细的局部搜索。c_1 和 c_2 为调节 P_{id} 和 P_{gd} 相对重要性的参数。$rand()$ 为随机数生成函数。则粒子移动的下一个位置为：

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1) \tag{8}$$

目前大多数粒子群聚类算法[6-9]都采用基于聚类中心的编码方式，也就是每个粒子的位置都是由 n 个聚类中心组成(n 为数据集的类别数)。由于样本向量维数为 d，因此粒子的位置是 $n\times d$ 维变量，则粒子编码结构如下所示：

$$P=(c_{11},c_{12},\cdots,c_{1d},\cdots,c_{i1},c_{i2},\cdots,c_{id},\cdots,c_{n1},c_{n2},\cdots,c_{nd})$$

本文提出了一种新的简单有效的编码方式，其先将样本按顺序编号为1到 k(k 为样本总数)，直接根据编号选择 n 个样本作为粒子的聚类中心，然后再增加一位存储加权指数 m 的值。在聚类过程中，算法根据粒子群的适应度值直接在样本数据集中随机搜索能较好代表聚类中心的样本作为粒子的聚类中心及合适的 m 值。本文提出的简化的粒子编码结构一共有 $(n+1)$ 位，粒子结构如下所示：

$$P=(c_1,c_2,\cdots,c_i,\cdots,c_n,c_{n+1})$$

根据本文提出的采用简化的粒子编码的粒子群聚类算法

的流程如下：

Begin

初始化：设定样本集合的编号为 1，…，k，则前 n 个粒子的位置 Z_i 的搜索空间为[1，k]。根据文献[2]中确定的 m 取值区间，本文取 m 的初始值范围为[1.1，5]，设置粒子编码加权指数位的搜索空间为[1.1，5]。

粒子的速度前 n 位 V_i 的搜索空间为$[-v_{max}, +v_{max}]$，设定 $v_{max}= p \cdot k, p\in[1/k, 1]$，最后一位加权指数 m 的速度为[0，0.2]。

输入样本数据集 X，聚类数目 n，设置粒子群体大小 M，初始化种群 P(0)。

For t=1 to 最大迭代次数 do

计算粒子群体 P(t)中各粒子个体的 FCM 聚类结果的适应度；

if 粒子适应度值 $<P_{id}$的适应度值

更新 P_{id}

end

if 粒子适应度值 $< P_{gd}$的适应度值

更新 P_{id}

end

根据式(7)、式(8)分别更新粒子速度和粒子位置；

End for

输出整个搜索空间找到的最小适应度值所对应的类簇划分。

End

本文采用第 2 节中的 FCM 算法对数据集进行划分。为了提高算法的运算速度，当聚类迭代次数大于预定的最大迭代次数时返回聚类结果，在实验中最大迭代次数设置为 10 次。由于第 2 节中的目标函数随着 m 的增大而减小，因此采用本文粒子编码方式后不能直接将第 2 节中的目标函数作为适应度值。本文采用的适应度计算公式如下：

$$f(X)=\arg\min\sum_{r=1}^{m}\sum_{d_i\in C_r}(d_i-\mu_r) \tag{9}$$

适应度函数通过最小化所有类簇中样本与类簇中心的相异度之和来寻找最优聚类。

通过直接用样本编码进行编码，直接将其代表的样本作为粒子比现有用样本各维对粒子进行编码随机性更强，搜索范围更大。在实验中我们发现，用数据集的样本代表聚类中心进行编码仍能取得较好的聚类效果，在搜索取得最优适应度的聚类结果的同时获得了最合适的加权指数 m 值。

4 实验结果

本文实验系统采用 C# 语言实现，在 CPU 为 3.2GHz、2G 内存环境下进行测试。

本文采用实验室开发的聚类数据生成器生成的两个人工数据集和来自权威的 UCI 机器学习数据库的 4 个真实数据集进行了实验。后者分别为 Iris、Wine、Zoo、Ecoli，它们经常被用来检验聚类算法的有效性。Iris 数据集样本的 4 个特征分别表示 Iris 数据的花瓣长度、花瓣宽度、萼片长度和萼片宽度。Wine 数据集由 3 类共 178 个样本组成，每个样本有 13 个属性，3 个类不存在交迭，但划分不太明显。Zoo 数据集共有 101 条记录，分为 7 类，由 16 个属性来描述样本，其中 15 个为布尔属性值，1 个为分类属性值属性(腿的数量){0，2，4，6，8}。Ecoli 数据来源于日本大阪大学细胞与分子生物学研究所，包含 8 个属性，共 336 条记录，数据记录用来预测生物蛋白在原核细胞中的位置。数据集的详细信息如表 1 所列。

表 1 数据集信息统计

数据集	样本数	类别数	属性数	类别分布
Dataset1	270	5	2	71,40,57,64,38
Dataset2	212	3	2	51,93,68
Iris	150	3	4	50,50,50
Wine	178	3	13	59,71,48
Zoo	101	7	18	41,13,20,10,8,4,5
Ecoli	336	8	8	143,77,2,2,35,20,5,52

两个人工数据集的数据分布情况如图 1、图 2 所示。

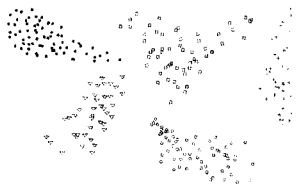

图 1 Dataset1 数据集数据分布图

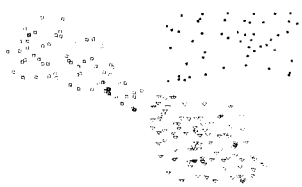

图 2 Dataset2 数据集数据分布图

粒子群优化过程的参数设置如下：最大迭代次数为 50，粒子群种群数目为 5，学习因子 $c_1=c_2=2$，p 取 $2/k$(k 为样本总数)，惯性权重 ω 在迭代过程中按式(10)线性下降，使粒子群在前期具有较高的全局搜索能力，在后续迭代过程降低移动速度，使其能够在最优值的邻域内逐步逼近最优解，避免较大的移动速度在优化解邻域产生震荡。

$$\omega=0.9-\frac{\text{当前迭代次数}}{\text{最大迭代次数}}\times 0.5 \tag{10}$$

将本文聚类算法运行 20 次，取聚类结果的平均值，将其与 K-Means 算法、FCM(模糊聚类算法，其中设置 $m=2$)、文献[8]改进的粒子群 K-Means 聚类算法及文献[10]中近邻传播算法(AP)中的聚类结果的精确度值进行比较，如表 2 所列。

表 2 聚类结果(精确度)

数据集	K-Means	FCM	AP	文献[7]	本文方法
Dataset1	0.79	0.93	0.952	0.862	0.959
Dataset2	0.83	0.94	0.959	0.948	0.989
Iris	0.78	0.88	0.88	0.884	0.894
Wine	0.68	0.74	0.76	0.72	0.964
Zoo	0.75	0.82	0.78	0.85	0.885
Ecoli	0.71	0.72	0.68	0.76	0.812

从表 2 中可见，在所有数据集上运行 5 种聚类方法的结果中，本文方法是最好的。与传统的 FCM 聚类算法相比，本文算法聚类结果精确度有明显提高。

本文算法在各个数据集上运行 20 次获得的聚类结果的精确度、召回率、FMeasure、熵值的平均值如表 3 所列。

表 3 数据集聚类结果

数据集	精确度	召回率	FMeasure	熵值
Dataset1	0.959	0.967	0.962	0.091
Dataset2	0.989	0.984	0.986	0.064
Iris	0.895	0.893	0.893	0.265
Wine	0.964	0.972	0.967	0.120
Zoo	0.885	0.786	0.799	0.090
Ecoli	0.812	0.644	0.652	0.212

表 4 给出了本文算法得到的最优适应度值及其对应的最优加权系数值。通过采用本文提出的新的粒子编码方式，在得到最优适应度值的同时，也得到了其对应的最优加权系数值 m。聚类结果中的 m 值都在[1.1，5]之间，与文献[2]的结论一致。

表 4　数据集聚类结果

数据集	最优适应度值	最优加权系数 m
Dataset1	34.54	3.73
Dataset2	34.485	2.23
Iris	29.305	1.52
Wine	88.697	2.56
Zoo	86.474	4.64
Ecoli	69.501	3.82

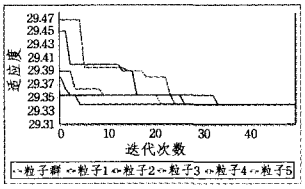

图 3　Iris 数据集粒子群聚类适应度收敛曲线图

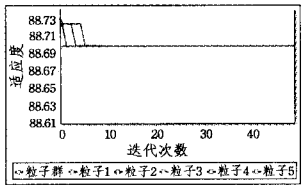

图 4　Wine 数据集粒子群聚类适应度收敛曲线图

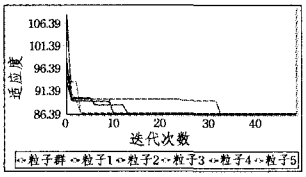

图 5　Zoo 数据集粒子群聚类适应度收敛曲线图

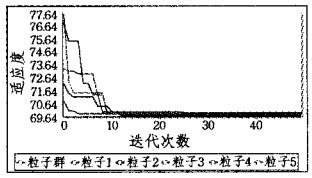

图 6　Ecoli 数据集粒子群聚类适应度收敛曲线图

图 3—图 6 为实验系统采用的 4 个 UCI 数据集的粒子群聚类的适应度收敛曲线，图中上方为种群中 5 个粒子的最优适应度变化曲线，最下方为粒子群全局最优适应度变化曲线。由于每代粒子之间社会信息的共享和各个粒子自我经验的提高，使得每代种群中的解具有自我学习提高和向他人学习的双重优点，所以其具有较快的收敛速度。实验中，4 个数据集在 50 次迭代内适应度函数值都能到达收敛状态。同时由于 FCM 具有较强的局部搜索能力，使得引入 FCM 后的粒子群算法的收敛速度和聚类准确率都大大提高。

结束语　本文将模糊 C-均值算法与粒子群优化有机结合，采用一种简化的粒子编码，同时对初始聚类中心和加权指数 m 进行搜索，有效地解决了 FCM 算法对初始聚类中心敏感和如何选择合适的加权指数 m 值的问题，取得了较好的实验结果。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宁. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

[2] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. New York：Plenum，1981

[3] 宫改云，高新波，伍忠东. FCM 聚类算法中的模糊加权指数 m 的优选方法[J]. 模糊系统与数学，2005，19(1)：143-147

[4] 朱林，王士同，邓赵红. 改进模糊划分的 FCM 聚类算法的一般化研究[J]. 计算机研究与发展，2009，46(5)：814-822

[5] 肖满生，阳娣兰，张居武，等. 基于模糊相关度的模糊 C 均值聚类加权指数研究[J]. 计算机应用，2010，30(12)：3388-3390

[6] 张利彪，周春光，马铭，等. 基于粒子群优化算法的模糊 C-均值聚类[J]. 吉林大学学报：理学版，2006，44(2)：217-222

[7] 许磊，张凤鸣. 基于 PSO 的模糊聚类算法[J]. 计算机工程与设计，2006，27(21)：4128-4129

[8] Esmin A A A，Pereira D L，de Araujo F P A. Study of different approach to clustering data by using the particle swarm optimization algorithm [C]//IEEE World Congress on Computational Intelligence. 2008：1817-1822

[9] van der Merwe D W，Engelbrecht A P. Data clustering using particle swarmoptimization [C]//Proceedings of Evolutionary Computation. 2003，1：215-220

[10] Frey B J，Dueck D. Clustering by Passing Messages between Data Points [J]. Science，2007，315(5814)：972-976

(上接第 165 页)

[15] von Luxburg U，Bousquet O，Belkin M. Limits of spectral clustering[J]. Advances in Neural Information Processing Systems，2005，17：857-864

[16] Dhillon I. Co-clustering documents and words using Bipartite Spectral Graph Partitioning[C]//KDD 2001 San Francisco. Callfornia，USA，2001：269-274

[17] Ding H Q. Unsupervised feature selection via two-way orderign in gene expression analysis[J]. Bioinformatics，2003，19(10)：1259-1260

[18] 李明华，刘全，等. 数据挖掘中聚类算法的新发展[J]. 计算机应用研究，2008，25(1)：13-17

[19] Von Luxburg. A tutorial on spectral clustering [J]. Statistics and Computing，2007，17(4)：395-416

[20] Fiedler M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory[J]. Czech，Math，Journal，1975，25：619-633

[21] Pothen A，Simon H，Liou K. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM J，Matrix Anal，1990，11(3)：430-452

[22] Mohar B. The Laplacian spectrum of graphs[J]. Graph Theory，Combinatorics and Applications，J. Wiley，New York，1991，2：871-898

[23] Hagen L，Kahng A. Fast spectral methods for ratio cut partitioning and clustering[C]//Proceedings of the ICCAD. 1992：10-12

[24] 王小平，曹立明. 遗传算法——理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002

[25] Berry M J A，Linoff G. Data Mining Techniques：For Marketing，Sales，and Customer Support[M]. John Wiley & Sons，Berlin，1997

[26] Pakhira M K，Bandyopadhyay S. A study of some fuzzy cluster validity indices，genetic clustering and application to pixel classification[J]. Fuzzy Sets and Systems，2005，155：191-214

[27] Oh S，Ahn CW，Jeon M. A New Evolutionary Approach to Cluster validation Index[J]. Journal of Computational and Theoretical Nanoscience，2010，7(5)：806-812

[28] Huang Jing-wei，Kang Li-shan，Chen Yu-ping. A new graph drawing algorithm for undirected graphs[J]. Journal of software，2000，11(1)：138-142

[29] 郑欣，林学. 学习非唯一的最佳聚类数[J]. 清华大学学报：自然科学版，2006，46(7)：1282-1285

[30] Li Xiao-hong，Luo Min. GAKC：a new GA_based k clustering algorithm[C]//the second international symposium on Information Science and Engineering(ISISE 2009). IEEE Computer Society，Dec 2009：334-338