DOI:10.11721/cqnuj20140213

基于邻域影响的改进粒子群算法的聚类算法*

温凤文，王洪春
(重庆师范大学 数学学院，重庆 401331)

摘要:K-均值算法是一种传统的聚类分析方法,具有思想与算法简单的特点,因此成为聚类分析的常用方法之一。但 K-均值算法的分类结果过分依赖于初始聚类中心的选择,对于某些初始值,该算法有可能收敛于一般次优解,在分析 K-均值算法和粒子群算法的基础上,提出了一种基于邻域影响的改进的粒子群算法的聚类算法,通过对粒子群算法的改进来优化与 K-均值结合的聚类算法。该算法将局部搜索能力强的 K-均值算法和全局搜索能力强的粒子群算法结合,提高了 K-均值算法的局部搜索能力、加快收敛速度,有效阻止了早熟现象的发生,达到那些离群的孤立点。实验表明该聚类算法有更好的收敛效果,一方面聚类所用的时间更短,另一方面聚类的准确率更高。

关键词:聚类分析;K-均值算法;粒子群算法
中图分类号:TP38　　**文献标志码**:A　　**文章编号**:1672-6693(2014)02-0059-04

聚类分析是数据挖掘领域中重要的技术之一[1-2],用于发现数据中未知的分类。聚类算法是机器学习、数据挖掘和模式识别等研究方向的重要研究内容之一,在识别数据对象的内在关系方面,具有及其重要的作用。聚类分析根据数据的内在性质将数据划分为若干个聚类集合类,使每一类中的元素尽可能具有相同的特性,不同聚合类之间的特性差别尽可能大。在已有的聚类算法中,K-均值聚类算法是经常用到的一种算法[3-11]。这种方法首先选定某种距离度量作为元素间的相似性度量,然后确定一个评价聚类划分结果质量的准则函数,在给出聚类中心点后,用迭代的方法找到准则函数极大值或者极小值的最好聚类结果[4,6]。

粒子群算法(Particle swarm optimization, PSO)[1]是一种有效的全局寻优算法,最早由美国的 Kennedy 和 Eberhart[8-9]于 1995 年提出。它是基于群体智能理论的优化算法,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。与传统的进化算法相比,粒子群算法保留了基于种群的全局搜索策略,但是题材用的速度-位移模型,操作简单。由于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,从而能在较少的迭代次数内找到最优解。

为此,本文研究了基于改进粒子群算法的 K-均值聚类算法,同时对算法进行了仿真实验分析,与文献[4]所介绍的聚类算法进行了比较。仿真实验结果表明,基于改进粒子群算法的 K-均值聚类算法较文献[4]中介绍的聚类算法收敛效果要好,收敛速度更快,而且可以找到更好的解。

1 K-均值聚类算法

$\mathbf{R}^n$ 空间中的聚类[4]问题可以这样来描述:对于给定的 $\mathbf{R}^n$ 中的 N 个点组成的模式样本集 $X=\{X_1,X_2,X_3,\cdots,X_N\}$,按照它们之间的相似性将其划分为 m 个(m 事先给定)集合 $C_1,C_2,\cdots,C_m$,满足:1) $C_i\neq\varnothing$, $i=1,2,\cdots,m$;2) $C_i\cap C_j=\varnothing, i,j=1,2,\cdots,m; i\neq j$;3) $\bigcup_{i=1}^{m}C_i=\{X_1,X_2,\cdots,X_N\}$。并且使得总的类间离散度和[11]

$$J_c=\sum_{k=1}^{m}\sum_{x_i\in C_k}d(X_i,Z_k) \tag{1}$$

达到最小,其中 $Z_k=[Z_1,Z_2,\cdots,Z_m]$为第 k 类的聚类中心,$d(X_i,Z_k)$为样本到对应聚类中心的距离。聚类准则函数 J_c 即为各类样本到对应聚类中心距离的总和。本文中 $d(X_i,Z_k)$为欧式空间的距离 $d(X_i,Z_k)=\|X_i-Z_k\|$。

K-均值聚类算法步骤如下:1) 任选 m 个初始聚类中心 $c_1,c_2,\cdots,c_m$;2) 将样本集 $X=\{X_i,i=1,2,\cdots,N\}$ 中各个样本按照最小距离原则分配给 m 个聚类中心的第 i 个聚类中心 c_i;3) 计算新聚类中心 $c_i'(i=1,2,\cdots,k,)$,即

* **收稿日期**:2013-01-01　　**网络出版时间**:2014-03-10　19:23
资助项目:重庆市自然科学基金计划项目(No. CSTC2011BB2116);重庆师范大学教改项目(No. 932126-0009-35)
作者简介:温凤文,女,研究方向为智能计算,E-mail: 765693826@qq.com;**通讯作者**:王洪春,E-mail: wanghongchun@swsc.com.cn.
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20140310.1923.013.html

$c_i'=\frac{1}{N_i}\sum_{X\in S_i}X$,其中 N_i 为第 i 个聚类域 S_i 包含的个数;4) 若 $c_i'\neq c_i(i=1,2,\cdots,k)$,转步骤 2),否则算法收敛,计算结束。

K-均值算法通过迭代的方法来得出聚类的划分结果,为了防止第 4)步骤中的终止条件不能满足而出现死循环,一般在算法执行的初始时给出一个固定的最大迭代次数避免这种情况发生。

2 标准粒子群优化算法

标准的粒子群算法[1,10]可描述为:设粒子群在一个 n 维空间中搜索,由 m 个粒子组成种群 $Z=\{Z_1,Z_2,\cdots,Z_m\}$,其中的每个粒子所处的位置 $Z_i=\{z_i^1,z_i^2,\cdots,z_i^n\}$ 都表示问题的一个解。在解空间中每个粒子受局部最优信息和全局最优信息的影响,以一定得速度在整个解空间飞行,飞行速度和位置由个体的飞行经验和群体的飞行经验动态调整,以便用于信息的交换。粒子通过不断调整自己的位置 Z_i 来搜索新解。每个粒子都能记住自己搜索到的最好解,记作 p_i^d(历史最优),以及整个粒子群经历过的最好的位置,即目前搜索到的最优解,记作 G^d(全局最优)。此外每个粒子都有一个速度,记作 $V_i=\{v_i^1,v_i^2,\cdots,v_i^n\}$,当 2 个最优解都找到后,每个粒子根据(2)式来更新自己的速度。

$$v_i^d(t+1)=wv_i^d(t)+\eta_1 rand()(p_i^d-z_i^d(t))+\eta_2 rand()(G^d-z_i^d(t)) \tag{2}$$

$$z_i^d(t+1)=z_i^d(t)+v_i^d(t+1) \tag{3}$$

$v_i^d(t+1)$表示第 i 个粒子在 $t+1$ 次迭代中第 d 维上的速度,w 为惯性权重系数,η_1,η_2 为加速常数(目前大多数文献都采用 $\eta_1=\eta_2=2$),$rand()$为 0~1 之间的随机数。此外为使粒子速度不致过大,可设置速度上限 $v_{\max}$,即当(2)式中 $v_i^d(t+1)>v_{\max}$时,$v_i^d(t+1)=v_{\max}$;$v_i^d(t+1)<-v_{\max}$时,$v_i^d(t+1)=-v_{\max}$。

从(2)、(3)式可以看出,粒子的移动方向由 3 部分决定:自己原有的速度 $v_i^d(t)$与自己最佳经历的距离 $p_i^d-z_i^d(t)$与群体最佳经历的距离 $G^d-z_i^d$,并分别由权重系数 w,η_1,η_2 决定其相对重要性。

3 改进粒子群算法的聚类分析

3.1 基于粒子群算法的 K-均值聚类算法

粒子群算法[11]采用实数编码,一个编码对应于一个可行解。在本文的粒子群聚类算法中,采用的是基于聚类中心的编码方式,也就是每个粒子的位置是由 m 个聚类中心组成,由于样本向量维数为 n,因此粒子的位置是 $m\times n$ 维变量,所以粒子的速度也应当是 $m\times n$ 维变量,另外每个粒子还有一个适应度 f_i。对于某个粒子,可以按照以下的方法来计算其适应度[11]:1) 按照最近邻法则,确定对应粒子的聚类划分;2) 根据聚类划分,按照(1)式计算类间离散度和 J_c;3) 个体的适应度本文采用的是 $f_i=h/J_c$,其中 J_i 是总类间离散度和,h 是常数,根据具体情况而定。这样个体的适应度与离散度和负相关,离散度和越小,个体适应度越大。

这样,粒子就可以采用以下的编码结构:$c_1^1,c_1^2,\cdots,c_1^n,\cdots,c_m^1,c_m^2,\cdots,c_m^n,v_1^1,v_1^2,\cdots,v_1^n,v_m^1,v_m^2,\cdots,v_m^d,f_i$。当聚类中心 Z_j 确定时,聚类的划分可以按照最近邻法来确定。即若 X_i,j 满足

$$\|X_i-Z_j\|=\min_{k=1,2,\cdots,m}\|X_i-Z_k\| \tag{4}$$

则 X_i 属于第 j 类。

3.2 改进粒子群算法的聚类分析的描述

3.2.1 基于邻域影响的粒子群算法介绍(AOI) 在自然界中,信息的传递受地理位置因素的影响,距离近的两个个体能够很快得到信息并受其影响很大,相反,地理位置远则受到社会影响就小。从(2)式可以看出,式子的右端可以分为 3 个部分:第 1 部分为原先的速度项,第 2 部分为该微粒历史最优位置对当前位置的影响(个体认知),第 3 部分微粒群体历史最好位置对当前位置的影响(社会影响)。因而 Kevin J Birkley 等[3]对标准 PSO(2)式等号右边第 3 部分做了修改,研究了基于 AOI 的 PSO

$$v_i^d(t+1)=v_i^d(t)+\eta_1\times rand()\times(p_i^d-x_i^d)+\eta_2\times f(|G^d-x_i^d|)\times rand()(G^d-x_i^d) \tag{5}$$

其中 $f(x)$是一个对角函数,有 $f(x)=\begin{cases}-\frac{h}{w_0}x, x\leqslant w_0\\ b, x>w_0\end{cases}$,$b,h$ 为给定的常数,w_0 是一个可调节的参数。在文献[3]中,w_0 表示粒子之间的距离。

因为 $f(x)$的光滑性,它不具有渐变性质,这里引入一个核函数 $K:\mathbf{R}^n\to[0,\infty)$,核函数具有以下性质:1) $K(-u)=K(u)$;2) 如果 $|u|<|v|$,那么 $K(u)>K(v)$,并且 $\lim\limits_{|u|\to\infty}K(u)=0$。

把性质(2)称为核函数的局部性质,核函数的局部特性在基于区域影响的 PSO 算法中起着重要作用。满足

上述性质的 K 有很多，在本文中选择 Gaussian 核函数 $K_\sigma(u)=e^{-|u|^2/2\sigma^2}$，其中 σ 是控制核函数的窗口大小的参数。与文献[3]相比，$K(u)$ 较 $f(x)$ 有更少的参数。在本文中将带有 Gaussian 核函数的 PSO 算法简记为 GPSO。

基于上述讨论得到新的基于区域影响的 PSO 算法的粒子群速度更新公式

$$v_i^d(t+1)=v_i^d(t)+\eta_1\times rand()\times(p_i^d-x_i^d)+\eta_2\times K(D)\times rand()(G^d-x_i^d) \tag{6}$$

其中 D 是当前粒子位置到历史最优位置的欧氏距离。

3.2.2 改进粒子群算法的聚类分析算法步骤　类似文献[4]，基于 GPSO 的 K-均值算法的具体流程如下。

步骤 1，种群的初始化。在初始化粒子时，先将每个样本随机指派为某一类，作为最初的聚类划分，并计算各类的聚类中心，作为初始粒子的位置，计算粒子的适应度，并初始化粒子的速度位置。

步骤 2，对每个粒子，比较它的适应度值和它经历过的最好位置（历史最优）$p_i=(p_i^1,p_i^2,\cdots,p_i^d)$，如果更好，更新 p_i；

步骤 3，对每个粒子，比较它的适应度值和群体所经历的最好位置 G^d 的适应度值，如果更好，更新 G^d；

步骤 4，根据(3)式和(6)式调整粒子的速度和位置。

步骤 5，新个体的 K-均值优化。对于新一代粒子，按照以下的 K-均值算法进行优化：1) 根据粒子的聚类中心编码，按照最近邻法则来确定对应该粒子的聚类划分；2) 按照聚类划分，计算新的聚类中心，更新粒子的适应度值，取代原来的编码值。由于 K-均值具有较强的局部搜索能力，因此引入 K-均值优化后的粒子群算法的收敛速度可以大大提高。

步骤 6，如果达到结束条件（最够好的位置或最大迭代次数），则结束，否则转步骤 2。

4 仿真实验与实验结果分析

下面通过一组 Fisher 的 Iris 植物样本数据的聚类分析来比较 K-均值算法、CPSO 方法、文献[4]介绍的聚类算法和本文提出的基于改进粒子群 K-均值聚类算法的性能。本实验软件环境为：操作系统 windows XP，编译软件 Matlab 7.10.0(R2010a)；硬件环境为 AMD Athlon(tm) 64，X2 Dual Core Processor 4400+ 2.31GHz，2.00 GB 的内存。

该数据集由属于 3 种植物的 150 个样本组成，每个样本均为 4 维向量，代表植物的 4 种特征数据。3 种方法都对实验的数据进行了归一化处理。每种算法进行 20 次，实验结果见表 1。为了使试验结果更具有说服力，实验运行时间和准确率都是取的 20 次实验的平均值。粒子种群的规模 $N=50$，迭代次数 M，加速常数 $\eta_1=\eta_2=2$，$v_{max}=1$，$v_{min}=0$，类别数 $k=3$，惯性权重系数 $w=v_{max}-t\times(v_{max}-v_{min})/M$（$w$ 也可以取常数，参考陈贵敏[10]等人对粒子群算法的惯性权值递减策略）。

表 1　文献[4]聚类的实验结果与本文聚类的实验结果的比较

序号	算法	运行时间/s	准确率/%	迭代次数
1	K-均值算法	0.525 784	71.4	100
2	CPSO	30.058 522 55	83.7	250
3	文献[4]方法	25.279 837	86.7	200
4	本文方法	13.389 847	91.9	100

在表 1 中，分别给出了 K-均值算法、CPSO 方法、文献[4]方法和本文方法运行的时间、准确率，通过比较 CPSO 方法、文献[4]方法和本文方法的最佳迭代次数，由迭代次数的减少可以看出本文方法的运行时间明显减少，准确率也比 CPSO 方法提高了近 10 个百分点。

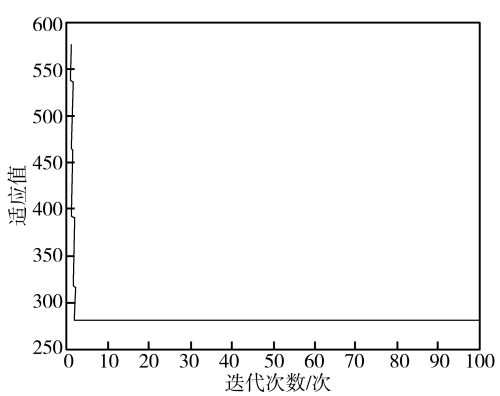

图 1　适应值曲线图

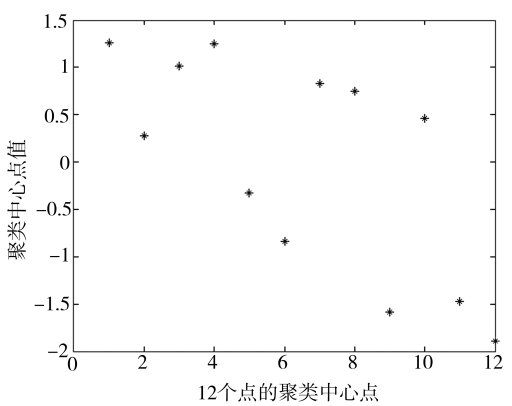

图 2　聚类中心点图

通过计算可以知道本文方法的最终平均适应值为 146.590 34，下面图 1 为适应值曲线图，从图中可以看出大约迭代

20 次该粒子就到到了自己的最优适应值,可以设置最大实验迭代次数为 100 次。图 2 为 3 个类 12 个点的最终聚类中心图,可以看出经过聚类后 Iris 植物样本聚类中心数据之间的关系比较明显,起到了分类的效果。

5 结论

在传统的 K-均值聚类方法的基础上,本文研究了改进的 CPSO 方法算法,在基本 PSO 算法的第三项引用 Gaussian 核函数,并和 Kevin J Binkley 算法比较。理论分析和数据试验结果都表明该方法克服了传统聚类算法存在的问题,其全局寻优能力优于 CPSO 方法,且有较快的收敛速度。

参考文献:

[1] 杨淑莹. 模式识别与智能计算[M]. 北京:电子工业大学出版社,2008.
Yang S Y. Pattern Recognition and Intelligent Computing [M]. Beijing: Beijing Electronics Industry University Press, 2008.

[2] 张军,詹志军. 计算智能[M]. 北京:清华大学出版社,2009.
Zhang J, Zhan Z J. Computational Intelligence[M]. Beijing: Tsinghua University Press, 2009.

[3] Binkley K J, Hagiwara M. Particle swarm optimization with area of influence: increasing the effectiveness of swarm[J]. Swarm Intelligence Symposium, 2005(1): 45-52.

[4] 陈小全,张继红. 基于改进粒子群算法的聚类算法[J]. 计算机研究与发展,2012,49(增刊):87-91.
Chen X Q, Zhang J H. Clustering algorithm based on improved particle swarm optimization[J]. Computer Research and Development. 2012, 49(Supplement): 87-91.

[5] 孙洋,罗可. 基于改进粒子群算法的聚类算法[J]. 计算机工程与应用,2009,45(33):132-134.
Sun Y, Luo K. Clustering algorithm based on improved particle swarm optimization[J]. Computer Engineering and Applications, 2009, 45(33): 132-134.

[6] 刘向东,沙秋夫,刘勇奎,等. 基于粒子群优化算法的聚类分析[J]. 计算机工程,2006,32(6):201-202.
Liu X D, Sha Q F, Liu Y K, et al. Cluster analysis based on particle swarm optimization algorithm[J]. Computer Engineering, 2006, 32(6): 201-202.

[7] 刘靖明,韩丽川,侯立文. 一种新的聚类算法-粒子群聚类算法[J]. 计算机工程与应用,2005,41(20):183-185.
Liu J M, Han L C, Hou L W. A new clustering algorithm - particle clustering algorithm[J]. Computer Engineering and Applications, 2005, 41(20): 183-185.

[8] Kennedy J, Eberhart R C. Particle swarm optimization[J]. Proc IEEE Int Conf Neural Networks, 1995(4): 1942-1948.

[9] Eberhart R, Kennedy J. A new optimizer using particle swarm theory[EB/OL]. (2004-03-09)[2013-01-01]. http://www.ppgia.pucpr.br/~alceu/mestrado/aula3/PSO_2.pdf.

[10] 陈贵敏,贾建援,韩琪. 粒子群优化算法的惯性权值递减策略研究[J]. 西安交通大学学报,2006,40(1):51-56.
Chen G M, Jia J Y, Han Q. Particle swarm optimization algorithm inertia weight decreasing Strategy[J]. Xi'an Jiaotong University Journal, 2006, 40(1): 51-56.

[11] 刘靖明,韩丽川,侯立文. 基于粒子群的 K 均值聚类算法[J]. 系统工程理论与实践,2005(6):55-58.
Liu J M, Han L C, Hou L W. K-means clustering algorithm based on PSO[J]. Systems Engineering Theory and Practice, 2005(6): 55-58.

Clustering Algorithm Based on Improved Particle Swarm Optimization

WEN Feng-wen, WANG Hong-chun

(School of Mathematics, Chongqing Normal University, Chongqing 401331, China)

Abstract: K-mean algorithm, a traditional clustering method with simple characteristic of thought and algorithm, has therefore become one of the methods commonly used in cluster analysis. But the K-means algorithm classification results depend on the initial cluster centers choice. For some initial value, the algorithm may converge to the general sub-optimal solution. This paper proposes a clustering algorithm based on the influence of neighborhood improvement particle swarm optimization (PSO). Through the improved PSO algorithm, we can optimize the combination of K-means clustering algorithm. Both the K-mean, which has strong capacity of local searching, and the PSO, which has power global search ability, are combined. It not only improves the K-mean's local searching capacity, accelerates the convergence rate, but also effectively prevents the premature convergence. The experiments show that this clustering algorithm has better convergence. On the one hand the use of clustering is shorter; on the other hand the accuracy is higher.

Key words: clustering analysis; K-mean; PSO

(责任编辑 黄 颖)