文章编号:1672 - 058X(2009)02 - 0144 - 04

一种改进的 K - 均值聚类算法*

但汉辉[1], 张玉芳[1], 张世勇[2]
(1. 重庆大学 计算机学院,重庆 400044;2. 重庆工商大学 计算机科学与信息工程学院,重庆 400067)

摘　要:为了改进 K - means 聚类算法的不足,把混合粒子群优化算法引入到 K - means 聚类算法中,重新选取编码方式并构造适应度函数,在此基础上提出了一种改进的 K - means 聚类算法;通过两个经典数据集的测试,实验结果表明:改进的算法比 K - means 算法具有更好的全局寻优能力、更快的收敛速度,且其解的精度更高对初始聚类中心的敏感度降低。

关键词:混合粒子群优化算法;K - 均值;聚类算法
中图分类号:TP 301.6　　**文献标识码**:A

K - means 算法是由 Macqueen 提出[1]的解决聚类分析问题的一种经典算法,它是基于聚类中心的划分方法,当结果簇是密集的且簇与簇之间的区别明显时,它的效果较好,对于较大的数据集该算法也具有较高的效率和相对的可伸缩性,目前该算法已经被广泛地应用于统计学、机器学习、空间数据库、生物学和市场研究等领域的实际应用和研究中。但在应用中越来越多的研究者发现该算法存在一些缺陷,主要有:不适用于有分类属性的数据,并且聚类结果随初始聚类中心的不同而可能不一样,有时甚至出现无解的情况;不适用于结果簇差别很大的数据集;必须事先给出簇的数目,增加了用户的负担;后期收敛速度较慢;通常以局部最优结束等。这些缺点极大地影响了 K - means 算法在各个具体领域中的进一步应用。近年来许多文献提出了针对 K - means 算法的改进方法,归纳起来主要有:结合遗传算法的改进,结合免疫算法的改进,结合群智能的改进,结合模拟退火的改进。如基于进化编程的聚类算法[2]、混合遗传聚类算法[3]、基于免疫规划的 K - means 聚类算法[4]和人工免疫 C - 均值混合聚类算法[5],这些算法采用不同方式对 K - means 算法进行改进,在一定程度上使 K - means 算法的性能更好。在此采用引入其他全局优化算法的思想来改进 K - means 算法,具体做法是把混合粒子群优化算法这种全局搜索能力强且具有并行思想的算法引入到 K - means 算法中,来改进 K - means 算法容易陷入局部最优、后期收敛速度慢和对初始聚类中心敏感的缺陷。通过经典数据集的对比测试,实验结果表明这种改进是有效的,改进后的算法明显优于 K - means 算法。

1　K - means 算法及其缺陷

K - means 算法思想比较简单,首先随机选择 k 个初始聚类中心,其中 k 是一个已知参数,即聚类(簇)个数,然后把每个点指派到最近的聚类中心,指派到一个聚类中心的所有点的集合构成一个簇,接下来根据指派到簇的点,更新每个簇的聚类中心,重复指派和更新过程,直到簇不发生变化或者聚类中心不发生变化,这时聚类准则函数已经收敛,算法结束。

算法 1:K - means 算法;输入:聚类数据集,聚类个数 k;输出:聚类中心,最小的误差平方和(sum of squared error,SSE)值 $\mathrm{SSE}=\sum_{i=1}^{k}\sum_{x\in c_i}D(c_i,x)^2$,簇(类)划分标识。

收稿日期:2008 - 12 - 05;**修回日期**:2009 - 01 - 11。
* **基金项目**:重庆市科委自然科学基金计划资助项目(*CSTC*. 2007*BB*2372)。
作者简介:但汉辉(1979 -),女,重庆市合川人,硕士研究生,从事信息网络、信息系统和数据挖掘研究。

步骤：

（1）给定含n个数据元素的数据集，随机选取k个元素或者通过随机指派n个数据元素到k个类中来获得初始聚类中心$c_j(I)$，$I=1,j=1,2,3,\cdots,k$。

（2）计算每个数据元素到每一个聚类中心的距离$D(x_i,c_j(I))$，其中$I=1,2,\cdots,n;j=1,2,\cdots,k$。如果满足$D(x_i,c_j(I))=\min\{D(x_i,c_m(I))\}\ m=1,2,\cdots,k$，则$x_i\in C_j$，$C_j$表示第$j$个类（簇）。

（3）根据新的簇划分情况，计算k个新的聚类中心$c_j(I+1)=\frac{1}{n_j}\sum_{x\in C_j}x$，$n_j$是簇$C_j$中的元素个数，$x$是划分到$C_j$中的输入数据集中的数据元素。

（4）判断$c_j(I+1)$是否等于$c_j(I)$或者误差平方和SSE是否变化，$j=1,2,3,\cdots,k$，如果不等或者SSE变化，则$I=I+1$，返回到（2）；如果相等或者SSE不变化，则聚类完成，结束算法。

从K－means算法的模型看，它实质上是求一个多峰（多极值点）函数的极小值问题，采用梯度下降法来获得目标函数SSE的极小点（在步骤（2）中）是导致K－means算法具有对初始聚类中心敏感以及陷入局部最优解难以获得全局最优解的最根本原因。为了克服此类问题，在此把具有很强全局搜索能力和更快收敛速度的混合粒子群优化算法（HPSO）[6]引入到K－means聚类算法中，从文献[6]中可以知道HPSO不但具有很快的收敛速度、很强的全局搜索能力，而且所获得的解的精度高，因此用它来改进K－means算法，可以帮助K－means算法摆脱局部极值的吸引，快速地找到全局最优，减小初始聚类中心对结果的影响。

2　改进的K－means聚类算法

为了把混合粒子群优化算法（HPSO）引入到K－means算法中，必须找到粒子的编码方式，通过仔细对比这两个算法的运行过程，可以发现对于混合粒子群优化算法（HPSO），为了找到最优解，每个粒子在搜索空间的不同区域搜索；对于K－means算法，为了找到最优解，k个聚类中心在搜索空间的不同区域搜索。粒子和聚类中心在各自的算法中扮演着相同的角色。因此可以在它们之间建立一个关系。在混合粒子群优化算法中每一个粒子就是一个候选解，对于K－means算法而言可以把确定好值的k个聚类中心整体看成是一个候选解，即一个粒子，也就是说此时的一个粒子就代表一种聚类方式。

这样设定粒子编码方式有一个明显的好处，因为对于n个粒子的种群，在算法开始的时候就要随机初始化n个粒子，对于K－means算法而言这就是随机给出了n种初始聚类中心，在算法以后的执行步骤中，就是从n种聚类中选择一种最优的，这样就可以明显地改善K－means算法对初始聚类中心敏感的缺陷。

对于数据对象的维度为d，聚类个数为k，样本数为n，样本为y的K－means算法其对应的粒子的编码的数学表达式如下：

$$x_i=(c_{i1},c_{i2},\cdots,c_{ik})',i=1,2,\cdots,n \tag{1}$$

$$c_{ij}=(c_{ij}^{(1)},c_{ij}^{(2)},\cdots,c_{ij}^{(d)}),j=1,2,\cdots,k \tag{2}$$

其中，c_{ij}表示第i种聚类方案的第j个类的聚类中心。

相应地误差平方和函数：

$$\mathrm{SSE}_i=\sum_{j=1}^{k}\sum_{y\in c_{ij}}D(c_{ij},y)^2 \tag{3}$$

那么原始的适应度函数，即目标函数是：

$$\mathrm{fitness}(x_i)=\mathrm{SSE}_i \tag{4}$$

$$p(x_i^1)=\Delta t_i \tag{5}$$

$$p(x_i^{k+1})=\begin{cases}\beta_1 p(x_i^k) & t_i\neq 0\\ \frac{1}{\beta_2}p(x_i^k) & t_i=0\end{cases} \tag{6}$$

其中罚式（5）和式（6）从文献[6]中引入，在引入罚函数$p(x_i)$的基础上重新构造的适应度函数为：

$$\mathrm{fitness}(x_i)=\mathrm{SSE}_i(1+p(x_i)) \tag{7}$$

算法2：改进的K－means算法；输入：待聚类数据集和聚类数目；输出：全局最优解（聚类中心，最小的误差平方和，簇（类）划分标识）。

步骤:

(1) 初始化粒子群;设置种群规模(粒子个数)n、粒子维数d、聚类数目k,搜索空间的大小、位置及速度的边界$x\max$、$x\min$、$v\max$、$v\min$和各个参数的值,把所有样本随机指派到k个簇中,并计算k个簇的聚类中心,从而得到一个粒子,重复执行n次,得到n个粒子。随机初始化各粒子的初始速度、初始化个体极值、全局极值和禁忌表[6]。

(2) 访问禁忌表,计算每个粒子对应罚函数$p(x_i)$的值,根据修改后的适应度函数计算每个粒子的适应度值。

(3) 对每个粒子,比较当前适应度值和它经历过的最好位置的适应度值。若更好,则更新;对每个粒子,比较自己当前最好适应度值和群体所经历的最好位置的适应度值。若更好,则更新。

(4) 根据粒子的速度公式[6]:$v_i^{k+1} = \omega v_i^k + c_1 r_1 (p_{ki}^{\text{best}} - x_i^k) + c_2 r_2 (g_k^{\text{best}} - x_i^k)$,和位置公式[6]:$x_i^{k+1} = x_i^k + v_i^{k+1}$,调整粒子的速度和位置。

(5) 样本的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:首先根据粒子的当前位置(k个新的聚类中心),按照最近邻法则,确定每一个样本应该在哪一个簇中;其次样本全部归类后,重新计算聚类中心(粒子的位置),重新计算适应度值,更新粒子的适应度值。

(6) 判断是否达到最大迭代次数或者满足最小误差,如果终止条件满足则输出全局极值以及获得全局极值的位置并结束算法,否则重复(2)。

3 改进的 K - means 算法有效性验证

为了验证改进的 K - means 算法的有效性,分别在 iris 数据集和 kdd cup 98 data 数据集上对算法进行测试。Iris 数据集是一个植物数据集,它的内容是关于鸢尾花的萼片的长度、宽度;花瓣的长度和宽度。该数据集共有 150 行记录,每行记录有 4 个属性,按照记录的自然属性可以分为:多刚毛的、杂色的和纯色的 3 个类别,每个类中各有 50 个记录。Kdd cup 98 data 数据集共有 95 413 条记录,每条记录有 56 个属性。

测试时根据文献[7]把c_1,c_2两个参数都设为 1.494 45,$\beta_1 = 3$,$\beta_2 = 2$、$\sigma = 500$、$a = 200$和$\Delta = 2.220\,4e-016$,根据文献[6]进行设置,禁忌表的每个元素都初始化为 0,每一次随机选择数据集中的样本作为初始聚类中心,并以这些聚类中心组成粒子,随机初始化粒子的速度,每个函数实验 40 次把所得结果求均值,对于 iris 数据集聚类数目为 3。两种算法在 Iris 数据集上的测试结果如表 1 所示。

表 1 改进 K - means 算法和 K - means 算法在 iris 数据集上的性能比较

改进 K - means 算法			K - means 算法		
收敛代数	最优解	收敛次数	收敛代数	最优解	收敛次数
10	78.854 1	30	14	86.753 0	11

从表 1 可以看出,对于 Iris 数据集而言,改进的 K - means 比 K - means 具有更强全局搜索能力,获得的解的精度更高,对初始聚类中心的敏感度有所降低。由于 Iris 数据集的第 4 个属性(花瓣的宽度)与前 3 个属性相关,这里用前 3 个属性来表示两种算法的某一次(随机选取一次)聚类结果,如图 1 所示。在 Iris 数据集上两种算法的 SEE 曲线图,如图 2 所示。两种算法在 Kdd cup 98 data 数据集上的测试结果如表 2 所示。

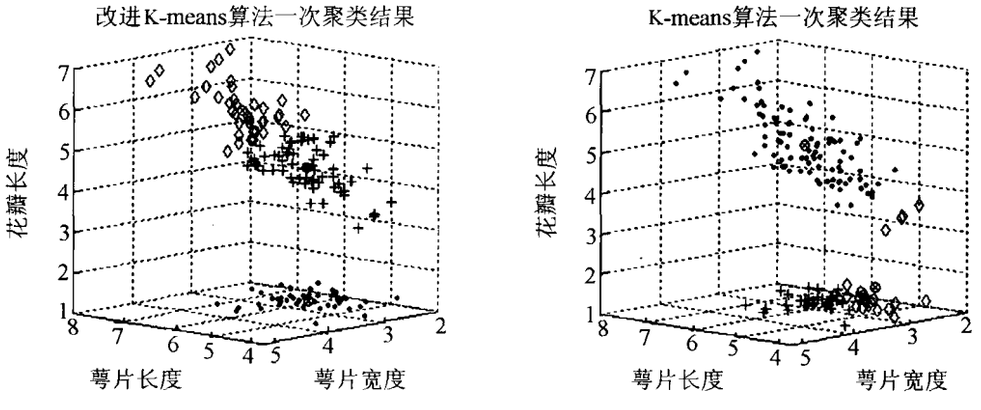

图 1 改进 K - means 算法与 K - means 算法的一次聚类

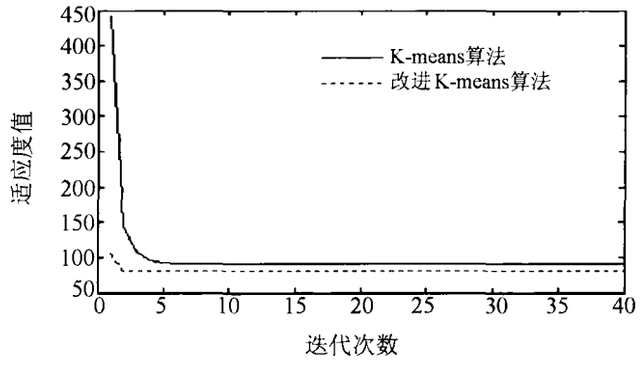

图 2　改进 K - means 算法与 K - means 算法在 iris 上的收敛曲线

表 2　改进 K - means 算法和 K - means 算法的在 Kdd cup 98 data 数据集上的性能比较

	改进 K - means 算法			K - means 算法		
	收敛代数	最优解	收敛次数	收敛代数	最优解	收敛次数
$K=3$	15	7.892e+05	25	23	9.876e+05	8
$K=10$	20	7.563e+05	20	30	9.438e+05	12
$K=50$	26	6.971e+05	24	36	8.022e+05	10

从表 1、表 2、图 1 和图 2 可以看出,改进 K - means 算法不论在收敛速度、算法的精度、克服陷入局部极值的能力都比 K - means 算法强,对初始聚类中心的敏感程度有了很大程度的降低,此外对于大型数据集改进 K - means 算法表现出更好的寻优能力。由于在此提出的算法是一种改进的 K - means 算法,因此,该算法跟 K - means 算法一样,不适用于有分类属性的数据和结果簇差别很大的数据集。

参考文献:

[1] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]. In: Proceedings of the 5th Berkeley Symposium on Mathematics Statistic Problem, 1967. 281 - 297

[2] SARKAR M, YEGNANARAYANA B, KHEMANI D. A clustering algorithm using an evolutionary programming - based approach [J]. Pattern Recognition Letters, 1997,18(10): 975 - 986

[3] KRISHNA K, MURTY M. *Genetic K - means algorithm* [J]. IEEE *Trans on System, Man and Cybernetics*: Part B, 1999, 29(3):433 - 439

[4] 行小帅,潘进,焦李成. 基于免疫规划的 K - means 聚类算法 [J]. 计算机学报,2003,26(5): 605 - 610

[5] 张雷,李人厚. 人工免疫 C - 均值聚类算法 [J]. 西安交通大学学报, 2005, 39(8):836 - 839

[6] 张世勇. 一种新的混合粒子群优化算法 [J]. 重庆工商大学学报:自然科学版, 2007,24(3):241 - 245

[7] CLERC M. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization [C]. In: Proceedings of the IEEE Congress on Evolutionary Computation, 1999. 1951 - 1957

[8] 代伟,刘敏,余永武. 基于 Aol Hoc 网络的混合入侵检测算法 [J]. 重庆工学院学报, 2008(3):82 - 85

An improved K - means cluster algorithm

DAN Han - hui[1], ZHANG Yu - fang[1], ZHANG Shi - yong[2]

(1. College of Computer, Chongqing University, Chongqing 400044, China;
2. College of Computer and Information Engineering, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: This paper incorporates hybrid particle swarm optimization algorithm into the K - means to overcome the local search of K - means algorithm, and adds the penalty function to reconstruct the fitness function, and proposes an improved K - means Cluster Algorithm, the computational experimental results on two benchmark dataset have shown that the improved K - means has better globe search capability, faster convergence velocity and is to attain higher precision value than K - means algorithm.

Key words: hybrid particle swarm optimization algorithm; K - means; cluster algorithm

责任编辑:代晓红