第34卷 第20期 计 算 机 工 程 2008年10月
Vol.34 No.20 Computer Engineering October 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)20—0200—03 文献标识码：A 中图分类号：TP301

基于遗传算法的K均值聚类分析

赖玉霞[1]，刘建平[1]，杨国兴[2]
(1. 浙江理工大学信息电子学院，杭州 310018；2. 浙江天健会计师事务所，杭州 310012)

摘 要：传统K均值算法对初始聚类中心敏感，聚类结果随不同的初始输入而波动，容易陷入局部最优值。针对上述问题，该文提出一种基于遗传算法的K均值聚类算法，将K均值算法的局部寻优能力与遗传算法的全局寻优能力相结合，在自适应交叉概率和变异概率的遗传算法中引入K均值操作，以克服传统K均值算法的局部性和对初始中心的敏感性，实验证明，该算法有较好的全局收敛性，聚类效果更好。

关键词：K均值算法；聚类中心；遗传算法

K-Means Clustering Analysis Based on Genetic Algorithm

LAI Yu-xia[1], LIU Jian-ping[1], YANG Guo-xing[2]
(1. College of Information and Electronics, Zhejiang Science and Technology University, Hangzhou 310018;
2. Zhejiang PanChina Accounting Firm, Hangzhou 310012)

【Abstract】Traditional K-Means algorithm is sensitive to the initial centers and easy to get stuck at locally optimal value. To solve such problems, this paper presents an improved K-Means algorithm based on genetic algorithm. It combines the locally searching capability of the K-Means with the global optimization capability of genetic algorithm, and introduces the K-Means operation into the genetic algorithm of adaptive crossover probability and adaptive mutation probability, which overcomes the sensitivity to the initial start centers and locality of K-Means. Experimental results demonstrate that the algorithm has greater global searching capability and can get better clustering.

【Key words】K-Means algorithm; clustering center; genetic algorithm

1 概述

聚类分析方法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，使类间相似性尽量小，类内相似性尽量大[1]，实现对数据的分类。聚类分析作为数据挖掘系统中的一个模块，既可以作为一个单独的工具用以发现数据库中数据分布的深层信息，也可以作为其他数据挖掘算法的一个预处理步骤，因此，它是数据挖掘领域的一个重要研究课题，目前已被广泛应用于许多领域，如模式识别、数据分析、图像处理、客户关系管理。K均值算法是聚类分析中一种基本的划分方法[2]，因其理论可靠、算法简单、收敛速度快、能有效处理大数据集而被广泛使用，但传统的K均值算法对初始聚类中心敏感，其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，所以，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[3]。生物的进化过程主要是通过染色体之间的交叉和变异来完成的，与此相对应，遗传算法中最优解的搜索过程也模仿了生物的进化过程，使用遗传操作数作用于群体进行遗传操作，从而得到新一代群体，其本质是一种求解问题的高效并行全局搜索算法。它能在搜索过程中自动获取和积累有关搜索空间的知识，并自适应地控制搜索过程，从而得到最优解或准最优解。算法以适应度函数为依据，通过对群体个体施加遗传操作实现群体内个体结构重组的迭代处理。在这一过程中，群体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，本文给出了一种基于遗传算法的K均值聚类算法来克服K均值算法的局部性。

2 K均值算法的基本思想

K均值算法是一种使用最广泛的聚类算法。算法以 K 为参数，把 n 个对象分为 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，每个对象初始地代表了一个簇的平均值或中心，对剩余的每个对象根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平均值，不断重复该过程，直到准则函数收敛。准则函数如下：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-\overline{x_i}\right|^2 \tag{1}$$

其中，$\overline{x_i}$ 为簇 C_i 的平均值。

K均值算法的描述如下：

(1)任意选择 K 个记录作为初始的聚类中心。

(2)计算每个记录与 K 个聚类中心的距离，并将距离最近的聚类作为该点所属的类。

(3)计算每个聚集的质心(聚集点的均值)以及每个对象与这些中心对象的距离，并根据最小距离重新对相应的对象进行划分。重复该步骤，直到式(1)不再明显地发生变化。

3 基于遗传算法的K均值聚类算法

本文将遗传算法应用到聚类分析中，把遗传算法的全局优化能力与聚类分析的局部优化能力相结合来克服聚类算法的局部性，在种群进化过程中，引入K均值操作，同时，为

作者简介：赖玉霞(1979－)，女，讲师、硕士研究生，主研方向：计算机网络，信息通信系统；刘建平，副教授；杨国兴，中级会计师
收稿日期：2007-12-10 **E-mail：**lenalin2000@yahoo.com.cn

了避免早熟现象，在种群中采用自适应方法动态调节交叉概率和变异概率，使其能够随适应度自动改变。算法具体步骤如下。

3.1 染色体编码

染色体编码有很多种，在聚类分析中较常用的是基于聚类中心的浮点数编码和基于聚类划分的整数编码。由于聚类算法具有多维性、数量大等特点，聚类问题的样本数目一般远大于其聚类数目，因此采用基于聚类中心的浮点数编码，将各个类别的中心编码为染色体。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(1, 2), (5, 4), (8, 7)，则染色体编码为(1, 2, 5, 4, 8, 7)。这种基于聚类中心的编码方式缩短了染色体的长度，提高了遗传算法的速度，对于求解大量数据的复杂聚类问题效果较好。

3.2 初始群体的产生

为了获得全局最优解，初始群体完全随机生成。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个体的染色体编码串，共生成 m 个初始个体，由此产生第一代种群。

3.3 适应度函数的选取

适应度通常用来度量群体中各个体在优化计算中可能达到或接近于最优解的优良程度。本文采用式(1)构造适应度函数，由于式(1)的值越小说明聚类结果越好，越大说明聚类结果越差，因此选择如下的适应度函数：

$$f = b/(1+E) \tag{2}$$

其中，b 为常数，可以根据具体问题作调整。

3.4 遗传算子

3.4.1 选择算子

采用适应度比例法与最优保存策略相结合的混合选择算子。首先在每一代开始时，将群体中的最优个体记录下来，然后根据各个体的适应度计算个体被选中的概率，用轮盘赌方法进行个体的选择，最后在每次遗传操作后形成新群体时用当前所记录的最优个体替换新群体中的最差个体，以防止遗传操作破坏当前群体中适应度最好的个体。

3.4.2 交叉操作

交叉操作是指对 2 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 2 个新的个体，提高遗传算法的搜索能力。由于本文染色体采用浮点数编码，因此采用适合浮点数编码的算术交叉算子，即

$$X_A^{t+1} = aX_A^t + (1-a)X_B^t, \quad X_B^{t+1} = (1-a)X_A^t + aX_B^t$$

其中，a 是一个(0, 1)范围内的随机数。

3.4.3 变异操作

变异是一种局部随机搜索，与选择、交叉重组算子相结合可以保证遗传算法的有效性，使其具有局部随机搜索能力，同时保持种群的多样性，防止非成熟收敛[4]。本文采用均匀变异算子，其具体操作过程是：对于每个变异点，从对应基因位的取值范围内取一随机数代替原有基因值。即

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

其中，r 为(0, 1)范围内的随机数；$U_{\max}$, $U_{\min}$ 分别是该基因位的数值上下限。

3.4.4 交叉率和变异率的自适应调整

标准的遗传算法已经被证明无法收敛到问题的全局最优解[5]，尤其是在种群分布不均匀时易出现未成熟收敛，即“早熟现象”，在进化中后期由于个体竞争减弱而引起的随机搜索趋势还会导致算法收敛速度缓慢，其原因是进化算子在整个进化过程中都采用了固定的概率值。为了避免以上问题，本文采用了自适应遗传算子。自适应遗传参数的选择如下：

$$p_c = \begin{cases} p_{c1} & f' \le f_{\text{avg}} \\ p_{c1} - \dfrac{(p_{c1} - p_{c2})(f' - f_{\text{avg}})}{f_{\max} - f_{\text{avg}}} & f' > f_{\text{avg}} \end{cases}$$

$$P_m = \begin{cases} p_{m1} & f \le f_{\text{avg}} \\ p_{m1} - \dfrac{(p_{m1} - p_{m2})(f_{\max} - f)}{f_{\max} - f_{\text{avg}}} & f > f_{\text{avg}} \end{cases}$$

其中，f_{avg} 表示每代群体的平均适应度值；$f_{\max}$ 表示群体中的最大适应度值；f' 表示要交叉的 2 个个体中较大的适应度值；f 表示群体中要变异个体的适应度值。对于适应度大的个体，赋予其相应的交叉和变异概率，而对于适应度小的个体，其交叉概率和变异概率较大，自适应的交叉和变异概率能够提供相对某个解最佳的 p_c 和 p_m，使自适应遗传算法在保持群体多样性的同时，保证算法收敛。

3.5 K 均值操作

先以变异后产生的新群体的编码值为中心，把每个数据点分配到最近的类，形成新的聚类划分。然后按照新的聚类划分，计算新的聚类中心，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 均值操作后，遗传算法的收敛速度可以大大提高。

3.6 循环终止条件

循环代数开始为 0，每循环一次，代数加 1，若当前循环代数小于预先规定的最大循环代数，则继续循环；否则结束循环。

3.7 算法的设计

(1)设置遗传参数：聚类个数 c，种群大小 m，交叉概率 p_c，变异概率 p_m，最大迭代代数 T，适应度倍数参数 b。

(2)随机生成初始群体。

(3)计算群体各个体的适应度。

(4)进行选择、交叉、变异、K 均值操作，产生新一代群体。

(5)重复第(3)、第(4)步，直到达到最大迭代代数 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 均值聚类。

(7)输出聚类结果。

4 实验结果与分析

为了检验算法的有效性，对原始算法和改进算法进行了对比实验。在 Matlab 环境下分别编写 K 均值算法和本文算法，导入数据进行实验。实验数据有 2 组，一组模拟数据 Data1 包括 12 个样本，分为 4 类，每个样本有 2 个属性，分别是(0, 1), (1, 0), (1, 1), (2, 2), (2,3), (3, 2), (5, 6), (6, 5), (6, 6), (7, 7), (7, 8), (8, 7)，数据分布图如图 1 所示。

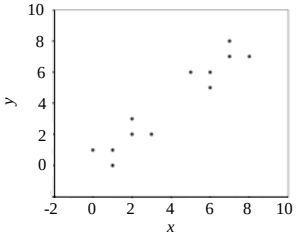

图 1 Data1 的数据分布

本文算法得到的聚类效果如图 2 所示，样本 1~3 为一簇，样本 4~6 为一簇，样本 7~9 为一簇，样本 10~12 为一簇，与

K 均值算法达到全局最优解的效果一样，与实际计算得出的聚类中心也一致。

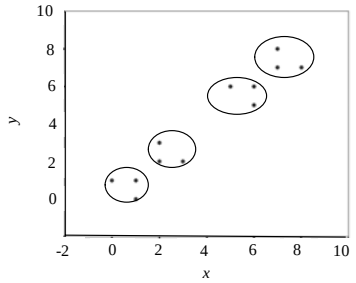

图 2　本文算法达到的聚类效果

第 2 组实验数据来自 http://ftp.ics.uci.edu/pub/machine-learning-databases/，数据集分别是 iris，glass 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；glass 数据集包含 214 个数据，分为 6 类，每个数据包含 9 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=100，交叉概率 p_{c1}=0.9, p_{c2}=0.6，变异概率 p_{m1}=0.1, p_{m2}=0.001, b=1 000，所有算法运行 20 次，运行情况如表 1 所示。根据表 1 的实验结果，K 均值算法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 glass 这种较高维度的数据集，没有一次达到全局最优解。而改进的算法对每组数据集的 20 次实验均能收敛到最优解，聚类效果较好。除数据集 iris 外，K 均值算法每组数据收敛到最优解的平均迭代次数都比本文算法多，所以，本文算法的收敛速度也比较快。

表 1　K 均值算法和本文算法的比较

聚类算法	数据集	E_{max}	E_{min}	E_{avg}	达到最优解次数	达到最优解的平均迭代次数	E 的最优解
K 均值	Data1	13.333 34	5.333 33	9.681 32	5	2.0	5.333 33
	iris	78.945 0	78.940 52	78.942 31	12	4.2	78.940 52
	glass	576.767 51	336.290 88	395.045 98	0	-	336.063 50
	wine	2.6336 57e6	2.370 79e6	2.394 692e6	18	6.2	2.370 79e6
本文	Data1	5.333 33	5.333 33	5.333 33	20	1.7	5.333 33
	iris	78.940 52	78.940 52	78.940 52	20	8.3	78.940 52
	glass	336.063 50	336.063 50	336.063 50	20	22.0	336.063 50
	wine	2.370 79e6	2.370 79e6	2.370 79e6	20	3.4	2.370 79e6

5　结束语

本文对 K 均值算法获得最优解的问题进行了研究，发现随机初始化会对该算法性能产生影响，不同的初始化中心会产生不稳定的聚类结果。本文提出的基于遗传算法的 K 均值聚类算法克服了上述缺陷。大量测试证明其不仅能够得到全局最优解，也能很好地解决 K 均值聚类方法对初始聚类中心敏感的问题，为聚类分析提供了一个新的思路。

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] 王　敞, 陈增强, 袁著祉. 基于遗传算法的 K 均值聚类分析[J]. 计算机科学, 2003, 30(2): 163-164

[3] 张云涛, 龚　玲. 数据挖掘原理与技术[M]. 北京: 电子工业出版社, 2004.

[4] 潘　伟, 刁华宗. 一种改进的实数自适应遗传算法[J]. 控制与决策, 2006, 21(7): 792-793.

[5] Rudolph G. Convergence Properties of Canonical Genetic Algorithms[J]. IEEE Trans. on Neural Networks, 1994, 5(1): 96-101.

(上接第 199 页)

表 2　在 breast-cancer 资料集上的实验对比结果

	F-measure	F_u-measure	双线性网格法
σ	0.014	0.030	0.148 7
C^+	149.413	27.725	-
C^-	199.517	89.808	-
C	-	-	0.088 4
少数类错分数量	7	1	8
多数类错分数量	13	15	12
错分率/(%)	2.93	2.34	2.93
F_u-measure 值	0.962 256 33	0.975 809 76	0.960 898 5

表 3　在 ionosphere 资料集上的实验对比结果

	F-measure	F_u-measure	双线性网格法
σ	0.534	0.396	0.354
C^+	296.647	4.951	-
C^-	71.719	299.300	-
C	-	-	4.00
少数类错分数量	4	3	8
多数类错分数量	11	8	7
错分率/(%)	4.27	2.44	4.27
F_u-measure值	0.939 191 34	0.962 441 31	0.939 191 34

相对F-measure，使用同样学习参数优化方法，但分类结果评价函数使用F_u-measure的模型，其小比例样本分类能力比F-measure有较大提高，小比例样本分类正确率较高，同时整体错分率也高于其他 2 种模型。使用F_u-measure对 3 种分类方法的实验结果进行了评价，F_u-measure模型对非平衡数据的分类性能要高于其他 2 种模型。相比双线性网格法，在速度上有很大优势。

5　结束语

本文针对实际分类数据类别不均匀的情况，提出了一种提高分类性能的分类改进算法。在基于高斯核函数的SVM分类方法中对多数类和少数类使用不同的惩罚参数C^+, C^-以获得高敏感度的超平面，并且提出利用遗传算法对SVM的学习参数进行优化调整。实验结果表明，利用本文的方法对非平衡数据进行分类的SVM的学习参数进行优化，对于非平衡数据的分类有较好的效果，使SVM对非平衡数据的分类性能有所提高，对少数类样本预测的准确性较高，取得了较好的改进效果。

参考文献

[1] Vapnik V N. The Nature of Statical Learning Theory[M]. New York, USA: Springer-Verlag, 1995.

[2] 张　琦, 吴　斌, 王　柏. 非平衡数据训练方法概述[J]. 计算机科学, 2005, 32(10): 181-186.

[3] Musicant D, Kumar V, Ozgur A. Optimizing F-measure with Support Vector Machines[C]//Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference. Florida, USA: AAAI Press, 2003: 356-360.

[4] Morik K, Brockhausen P, Joachims T. Combining Statistical Learning with a Knowledge-based Approach——A Case Study in Intensive Care Monitoring[C]//Proceedings of the International Conference on Machine Learning. San Diego, CA, USA: [s. n.], 1999.