

基于遗传算法的基因微阵列数据聚类

许珊娜, 祁云嵩

(江苏科技大学 计算机科学与工程学院, 江苏 镇江 212003)

摘 要: 微阵列基因数据用以挖掘特定的生物信息, 聚类分析对于研究基因功能和基因调控机制有重要意义. 结合改进的遗传算法对基因微阵列数据进行聚类分析, 并且通过实验与 K 均值聚类进行比较. 仿真实验表明, 该算法可以有效改进基因微阵列数据的聚类准确率.

关键词: 微阵列数据; 聚类; 遗传算法

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2012)04-0123-03

Clustering Microarray Gene Data Based on Genetic Algorithm

XU Shan-na, QI Yun-song

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: The microarray gene dataset is used to mine the specific biological information, and it is extremely important for the study of gene function and gene regulation mechanisms based on the clustering analysis. In the article, we propose a clustering analysis combined an improved genetic algorithm(GA), and compared with the K-means clustering. Simulation experiment shows that the algorithm can improve the clustering accuracy of the microarray data.

Key words: microarray gene dataset; clustering; Genetic Algorithm(GA)

1 引言

DNA 微阵列(DNA Microarray)数据, 是在一片芯片上同时测量不同样本的成千上万个基因在不同状态和不同组织中的 mRNA 表达水平而得到的一组数据. 它可以用来检测基因表达水平及识别基因序列、检测表达状况, 发现新基因、进行 DNA 序列分析. 微阵列基因表达数据的分析已经被成功应用于许多领域, 如临床诊断、指导用药、药物筛选等. 另外, 在预防医学、环境保护、军事、农业等领域也都有应用^[1-4].

随着基因表达数据的不断增长, K-均值聚类算法已经得到了广泛的应用. 它是基因表达谱数据分析最常用的多变量技术, 可以发现基因之间共表达或共调控关系、预测基因功能、提供疾病机理等. 它是基于选定的相似性测度对向量进行聚类. 常用的相似性测度包括欧氏距离、明考斯基距离、皮尔逊相

关距离等^[2-5]. K-均值聚类算法简单、收敛速度快, 局部搜索能力强, 不过易陷入局部最优.

由于遗传算法具有全局优化, 搜索空间大, 快速收敛等优点, 本文利用遗传算法, 采用浮点数编码和神经元激活函数, 并结合改进的 K-均值算法提出了一种改进的遗传聚类算法. 提高了聚类效果.

2 基于遗传算法的聚类分析

2.1 遗传 K-均值聚类算法

2.1.1 K-均值聚类算法

K-means 聚类算法是一种算法简单的常用划分方法. 该算法的核心思想如下: 首先从所给 n 个数据对象中随机选取 k 个对象, 每个代表初始聚类中心, 然后对于所剩下的其他对象, 则根据他们与所选 k 个中心点的相似度分别分配给与其最相似的聚类, 然后再重新计算所获聚类的聚类中心, 不断重复这一过程直到标准测度函数开始收敛为止^[3-5]. 其基本

的算法流程如下:

(1) 任意选 K 个初始聚类中心, $Z_2(1), \dots, Z_k(1)$, $k < N$. 括号内的序号代表寻找聚类中心的迭代运算的次序号, 聚类中心的向量可任意设定.

(2) 按照最小距离原则将其余样本分配到 k 个聚类中心中的某一个.

如果 $\min\{\|X - Z_i(k)\|, i=1, 2, \dots, k\} = \|X - Z_j(k)\| = D_j(k)$, 则 $X \in S_j(k)$. S_j 表示第 j 个聚类, 其聚类中兴路为 Z_j .

(3) 计算各个聚类中心的新的向量值:

$$Z_j(k+1) = \frac{1}{N_j} \sum_{X \in S_j(k)} X, \quad j=1, 2, \dots, k \quad (1)$$

(4) 如果 $Z_j(k+1) \neq Z_j(k)$, $j=1, 2, \dots, k$, 则回到(2), 将模式样本逐个重新分类, 重复迭代计算. 否则算法收敛, 计算完毕^[5].

2.1.2 遗传算法

生物个体的各种生命表征是由许多基因共同决定的. 同一种群的不同生物个体通常拥有不同的基因, 因此对外界环境的适应能力也是不同的. 在自然选择的作用下, 一部分环境适应能力较差的个体会死亡被淘汰, 而环境适应能力较强的则更多的存活下来并繁衍后代, 适应环境的基因会有较大的概率流传到下一代. 基因的传递过程中, 一部分基因会变异从而产生新的基因, 这种变异是随机的^[6-7].

遗传算法就是仿照上述理论, 每一个编码代表一个个体, 建立一个包含潜在的解的群体作为种群, 在环境作用下通过选择、交叉和变异代代繁衍, 由于子代的环境适应力一般优于附带, 因此算法最终能够得到问题的较优解.

2.2 改进的算法

2.2.1 编码方案

进行遗传算法前, 必须对问题的解空间进行编码, 以使它能够被遗传算法的算子操作. 最常用的编码方法是二进制编码, 但是由于基因微整列数据的维数较大, 所以不适合用二进制编码. 本文将采用的编码方式是浮点数编码^[7]. 所谓浮点数编码方法, 指个体的每个基因值用某一范围内的一个浮点数来表示, 个体的编码长度等于其决策变量的个数. 此编码方法适合应用于高精度的遗传算法, 搜索空间较大.

2.2.2 选择算法

本文算法中采用的选择策略是锦标赛选择策略. 锦标赛方法选择策略是每次从种群中取出一定数量个体, 然后选择其中最好的一个进入子代种群. 重复该操作, 直到新的种群规模达到原来的种群规

模^[8]. 具体步骤如下:

(1) 确定每次选择的个体数量 t .

(2) 从种群中随机选择个体构成组, 根据每个个体的适应度值, 选择其中适应度值最好的个体进入子代种群.

(3) 重复步骤 n 次, 得到的个体构成新一代种群.

2.2.3 交叉概率和变异概率

在神经网络中, 构造神经元激活函数 $y=f(x) = \frac{1}{1+e^{-\lambda x}}$. 可以利用此函数设计交叉概率 P_c 和变异概率 P_m .

$$P_c =$$

$$\begin{cases} \frac{P_{cmax} - P_{cmin}}{1 + \exp\left(-\lambda \left(\frac{2(f' - f_{cavg})}{f_{cmax} - f_{cavg}}\right)\right)} + P_{cmin}, & f' \geq f_{cavg} \\ P_{cmax}, & f < f_{cavg} \end{cases} \quad (2)$$

$$P_m =$$

$$\begin{cases} \frac{P_{mmax} - P_{mmin}}{1 + \exp\left(-\lambda \left(\frac{2(f' - f_{mavg})}{f_{mmax} - f_{mavg}}\right)\right)} + P_{mmin}, & f' \geq f_{mavg} \\ P_{mmax}, & f < f_{mavg} \end{cases} \quad (3)$$

式中, 参数 λ 称为 Sigmoidal 函数的增益, 其值决定了函数非饱和段的斜率, 其越大, 曲线越陡. P_c 表示交叉概率, P_{cmax} 表示最大交叉概率, P_{cmin} 表示最小交叉概率, P_m 表示变异概率, P_{mmax} 表示最大变异概率, P_{mmin} 表示最小变异概率, f_{cavg} 表示种群平均适应度值, f_{mmax} 表示种群最大适应度值. 交叉概率一般 $P_c=0.6 \sim 1.0$; 变异概率一般 $P_m=0.005 \sim 0.05$. 在本文中取 $P_{cmax}=1.0$, $P_{cmin}=0.6$, $P_{mmax}=0.05$, $\lambda=1$, $P_{mmin}=0.005$ ^[7].

2.2.4 适应度

适应度函数是整个遗传算法中极为关键的一部分. 好的适应度函数能够指导我们从非最优的个体进化到最优个体. 适应度函数是用来评价聚类结果的优劣, 好的聚类结果使得类间距离较大, 类内距离较小. 此算法的目标函数 J 及考虑类间距离、类内距离.

定义:

$$J = \frac{\sum_{i=1}^k \sum_{x_i \in C_i} \|x_i - C_i\|}{1 + \sum_{m \neq n} \|C_m - C_n\|},$$

$(m = 1, 2, \dots, k - 1; n = i + 1, \dots, k)$ (4)

式中, $\sum_{i=1}^k \sum_{x_l \in C_i} \|x_l - C_i\|$ 是类内距离, $\sum_{m \neq n} \|C_m - C_n\|$ 是类间距,类间距较大,类内距较小的聚类划分较优,从而目标函数 J 的值较小,因此,定义适应度函数为

$$\text{fitness} = \frac{1}{1 + J}$$
 (5)

2.2.5 算法流程

基于遗传算法聚类框架,得到一种新的基于遗传算法的聚类方法,算法具体描述如下:

- (1) 输入 k ,最大迭代数,利用 person 系数初始聚类,根据 2.2.1 节中给出的编码方案和 person 系数构成初始群体 $t=0$;
- (2) 根据给出的适应度评价方法,调用 K 均值算法,评价各个初始个体的适应度;
- (3) 选择个体,进行交叉、变异等操作,生成新一代个体群;
- (4) 计算新的聚类中心和类内距离平和;
- (5) 评价新一代个体群的适应度;
- (6) 判断是否达到最大迭代次数,若达到输出聚类结果,否则 $t=t+1$,执行第(3)步骤.

3 仿真实验与结果分析

仿真环境:软件:操作系统 Windows XP,编译软件 Matlab7.1;

硬件:Pentium(R) Dual-Core CPU 2.61 GHz,内存 1.99 GB.

本实验采用的数据集是白血病(LEUKEMIA72)数据集,白血病数据是由 Golub 等人在 1999 年给出的,它包括 72 个样本,每个样本含有 7 129 条基因,其中 47 个样本为急性淋巴白血病(ALL),25 个为急性骨髓白血病(AML)^[9].

首先利用相关系数 $\rho < 0.97$ 去除部分基因,减小基因数量,减少之后的基因数为 393.之后利用 2.2.5 节中的算法对 38 个训练样本的数据集(27 个 ALL 和 11 个 AML)进行基因聚类,再采用 34 个测试样本的数据集(20 个 ALL 和 14 个 AML)进行样本聚类.在上述仿真环境下,进行仿真实验,具体参数如下:

最大迭代次数 $T_{\max} = 100$; $\lambda = 1$, $P_{\text{cmax}} = 1.0$, $P_{\text{cmin}} = 0.6$, $P_{\text{mmax}} = 0.05$, $P_{\text{mmin}} = 0.005$.

针对白血病数据集,对不同的聚类数进行聚类,得到当聚类数为 13 时准确率最高,如表 1 所示.

表 1 不同聚类数聚类结果

聚类数	7	8	9	10	11	12	13	14	15
总样本数	34	34	34	34	34	34	34	34	34
正确划分	18	17	18	21	25	21	22	21	21
准确率/%	52.94	50.0	52.94	61.76	73.53	61.76	64.71	61.76	61.76

为了验证文中算法的有效性,文中做了一组对比实验,采用传统的 K 均值聚类算法在同样的实验条件下对基因聚类.从表 2 中,可以得到以下结论:本文算法的聚类数更小,算法得到的准确率更大(0.735 3).聚类的特征基因(距聚类中心最近的)如表 3 所示.

表 2 本文算法和 K 均值算法结果比较

聚类算法	聚类数	分类正确数	正确率
本文算法	11	25	73.53
K-means 算法	23	19	55.88

表 3 聚类数为 11 时的各类代表基因
(距聚类中心最近的基因)

类号	基因	类号	基因	类号	基因
1	X63527_at	5	X55954_at	9	X71428_at
2	U12465_at	6	L38941_at	10	U14971_at
3	X00351_3_at	7	M21142_cds2_s_at	11	U14970_at
4	X15940_at	8	X00351_5_at		

本文的算法在聚类的准确率上有所提高,显示了改进后的算法具有较好的聚类效果,验证了文中进行的算法改进是有效的.

4 结束语

本文对遗传算法的 K 均值聚类方法,进行了改进避免了 K 聚类算法容易陷入局部最优解的缺陷,并且对遗传算法的算子进行了改进,优化了算法的性能.实验表明,改进后的算法优于 K 均值聚类算法,提高了聚类的正确率.

在基因表达谱数据的采集过程中,某些基因参数取值不同,但表征的是同一种疾病,这些基因就构成冗余基因,在实际的基因表达谱数据的分析过程中,大量冗余基因的存在增加了诊断疾病的困难.冗余基因不仅影响聚类的收敛时间,且会影响聚类结果,因此冗余基因处理是下一步研究重点.

5 结束语

插值预测的计算复杂度相当高,是 AVS 解码器的瓶颈所在,占用整个解码器大部分的处理周期。本文提出的高速自适应的硬件结构,重点在并行自适应流水线结构的设计,以减少处理周期;根据亮度插值样本位置的对称性,提出一种转置滤波器组的结构,减少了 26 个 8 比特的缓存空间和 37 个 4 抽头滤波器;提取色度插值中复用结构,降低了硬件资源的使用。在 SMIC 0.18 μm 工艺库下综合,最高工作频率为 200 MHz,占用逻辑门数约为 82 k。一个宏块的亮度插值的计算过程平均占用 43 个周期,预测一个参考帧为 2 的宏块最多只需要 512 个时钟周期。仿真和综合结果表明该硬件结构在控制硬件资源使用的前提下,极大的提高了处理速度,能够满足 1080p@30fps 的 AVS 视频实时解码的需求。

参考文献:

- [1] 李霞,黄玄,周莉,等. AVS 帧内预测算法分析及 VLSI 实现[J]. 微电子学与计算机,2009,26(1):29-33.
- [2] 中国国家标准管理委员会. GB/T 20090.2-2006. 信

息技术 先进音视频编码 第 2 部分:视频[S]. 北京:中国标准出版社,2006.

- [3] 黄玄,陈杰,周莉,等. AVS 高清视频环路滤波结构与电路实现[J]. 微电子学与计算机,2010,27(3):11-15.
- [4] Ke Xu, Chiu-Sing Choy. A power-efficient and self-adaptive prediction engine for H 264/AVC Decoding [J]. Very Large Scale Intergraton (VLSI) Systems, 2008,16(3):302-313.
- [5] 王方晴,王祖强. AVS 分数像素插值算法的 VLSI 实现 [J]. 电子技术应用,2006,(4):44-47.
- [6] Zhang Duo-Li, Cheng Xian-Wen, Du Gao-Ming et al. A novel VLSI architecture of lum interpolator of H 264 decoder[C]//IEEE International Conference on Acoustics spech and signal proceesing. San Francisco, CA, USA: IEEE,2009:175-178.
- [7] 胡倩,虞露. AVS 插值算法的一种高效的硬件结构设计 [J]. 电路与系统学报,2008,13(3):148-152.

作者简介:

黄有源 男,(1986—),硕士研究生. 研究方向为视频图像处理的数字集成电路设计.

何明华 男,(1971—),博士,教授. 研究方向为超大规模集成电路设计及系统级芯片设计.

(上接第 125 页)

参考文献:

- [1] Edmundo BonillaHuerta, Be' a trice Duval, Jin-Kao-Hao. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data[J]. Neurocomputing, 2010(73):2375-2376.
- [2] Zhaohui S Qin. Clustering microarray gene expression data using weighted Chinese restaurant process [J]. Bioinformatics, 2006, 22(16):1988-1999.
- [3] Nabil Belacel, Qian Wang, Miroslava Cuperlovic-culf. Clustering Methods for Microarray Gene Expression Data [J]. OMICS,2006, 10(4): 507-508.
- [4] Michael Laszlo, Sumitra Mukherjee. A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 2006, 28(4):633-634.
- [5] 边肇祺,张学工. 模式识别[M]. 2 版. 北京:清华大学

出版社,2000.

- [6] 王婧. 遗传算法及其在聚类分析中的应用[D]. 湖北:华中师范大学,2009.
- [7] 杨淑莹. 模式识别与智能计算——Matlab 技术实现 [M]. 北京:电子工业出版社,2008.
- [8] Potts C J, Terri D. The development and evaluation of an improved genetic algorithm based on migration and artificial selection [J]. IEEE Transactions on Systems, Man, and Cybematics,1994,24(1):73-86.
- [9] Golub T R, Slonim D K, Tamayo C P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999,286(15):531-537.

作者简介:

许珊珊 女,(1985—),硕士. 研究方向为生物信息学.