DOI:10.13657/j.cnki.gxkxyxb.2013.04.012

广西科学院学报
Journal of Guangxi Academy of Sciences

2013,29(4):262～264,268
Vol. 29, No. 4 November 2013

基于 GA 优化的 RBF 网络算法*
GA-Based RBF Network Optimization

杨 洁
YANG Jie

(柳州职业技术学院信息工程系,广西柳州 545006)
(Department of Computer Science, Liuzhou Vocational Technological College, Liuzhou, Guangxi, 545006, China)

摘要:针对 RBF 神经网络易于陷入局部最大值的缺点,把遗传算法引入 RBF 神经网络中,利用遗传算法具有全局搜索的优点,对 RBF 神经网络的权值进行优化,并把优化后的神经网络模型用于 DNA 序列的分类。仿真实验表明,采用遗传优化的 RBF 神经网络比传统 RBF 神经网络分类有更高的分类效率和正确率。

关键词:RBF 神经网络 DNA 序列分类 特征提取 遗传优化
中图法分类号:TP183 **文献标识码**:A **文章编号**:1002-7378(2013)04-0262-03

Abstract: Because RBF neural network is easy to fall into the defects of local maxima, the genetic algorithm is introduced into the RBF neural network. The advantage of genetic algorithm on global search can optimize the RBF neural network weights and the optimized neural network model is further used to classify DNA sequences. Compared with traditional RBF neural network, the genetic optimized RBF neural network shows higher classification efficiency and accuracy.

Key words: RBF neural network, DNA sequence classification, feature extraction, genetic optimization

DNA 序列隐藏了丰富的遗传信息,如何找出这些序列的编码方式是当前生物信息学最重要的课题之一。文献[1～4]分别从 RBF 神经网络、模糊聚类分析方法、支持向量机的角度对 DNA 序列的分类进行了讨论。RBF 网络是一种前馈网络[5],一般分为输入层、隐含层和输出层,每一层都由若干个神经元组成。RBF 神经网络算法以梯度下降法为基础进行网络参数寻优,存在的主要问题是训练时间长,易于陷入局部极小。相邻层之间通过权值实现联结,遗传算法(Genetic Algorithm,简称 GA)是一种全局优化算法,它能有效避免搜索过程中易于陷入局部最优解的问题[6]。把遗传算法引入 RBF 神经网络,利用该算法对神经网络的权值进行优化,可以大大提高网络训练的精度,同时又能提高网络训练的速度,避免落入局部极小值。

本文通过对 DNA 序列分类技术进行研究,采用遗传算法对 RBF 神经网络的权值进行优化,并把优化后的神经网络应用于 DNA 序列分类。仿真实验结果表明,改进的算法比传统 RBF 神经网络分类具有更高的分类效率和正确率,为解决分类、预测及模式识别等问题提供了新的途径。

1 遗传算法

遗传算法主要通过交叉、变异、选择等算子来实现,交叉或变异算子生成下一代染色体,称其为后代。根据适应度的大小,从上一代和后代中选择一定数量的个体作为下一代群体,再继续进化,这样经过若干代后,算法收敛于最好的染色体,它很可能就是问题的最优解或次优解。使用实数进行编码,首先将染色体表示为向量形式:

$$X=\{X(0),X(1),\cdots,X(N-1)\}。\qquad(1)$$

收稿日期:2013-02-15
修回日期:2013-05-10
作者简介:杨 洁(1977-),女,助教,主要从事生物信息学、数据挖掘研究。
*广西教育厅科研项目(200911LX486)资助。

其中，$X(i) \in [0,1], i=0,1,\cdots,N-1, X(k) \in [0,1], k=0,1,\cdots,(N-1)/2$，可以通过以下映射关系而获得

$$X(k)=[h(k)+1]/2。\tag{2}$$

其中，$h(k) \in [-1,1], k=0,1,\cdots,(N-1)/2$。

适应度函数决定了 GA 算法能否迅速收敛以及找到最优解，可见，适应度函数的选取至关重要。采用的适应度函数为

$$f=1/E^2。\tag{3}$$

采用的交叉算子基本思路：首先以概率 p_c 对 2 个父辈个体进行随机分割，重新组合后获得 2 个新的个体；然后依据分割点的数量，每个父辈个体随机选择 m 个没有重复的交叉点，产生 2 个新的子个体，从而实现了交叉操作。

变异算子是根据概率 p_m，把个体染色体上的某一个位置上的基因进行排列，发生突变。设父辈个体的向量表示为 $x=(x_1,x_2,\cdots,x_k)$。其中，分量 x_i 以 p_m 概率被选择作为变异。

2　基于 GA 优化的 RBF 网络算法

采用遗传算法学习 RBF 网络的步骤如下：

步骤 1　初始化控制参数：种群规模 $N=30$，交叉概率 p_c，突变概率 p_m，进化代数 $k=0$。采用实数方式对基因进行编码，随机产生初始种群。

步骤 2　计算每个个体评价函数，同时将其进行排序。可按下式概率值选择网络个体：

$$p_s=f_i/\sum_{i=1}^{N} f_i,$$

其中，f_i 为个体 i 的适配值，可用误差平方和 E 来衡量，即

$$f(i)=1/E(i),$$

$$E(i)=\sum_{p}\sum_{k}(V_k-T_k)^2,$$

其中：$i=1,2,\cdots,N$ 为染色体数；$k=1,2,\cdots,4$ 为输出层节点数；$p=1,2,\cdots,5$ 为学习样本数。

步骤 3　以概率 p_c 对个体 G_i 和 G_{i+1} 交叉操作产生新个体 G'_i 和 G'_{i+1}，未进行交叉操作的个体进行直接复制。

步骤 4　利用概率 p_m 突变产生 G_i 的新个体 G'_j。

步骤 5　将新个体插入到种群 p 中，并计算新个体的评价函数。

步骤 6　若找到了满意的个体，则结束，否则转步骤 3。

3　应用实例

本实验所用 PC 机的硬件配置为 Pentium(R) 4 CPU 2.4 GHz、1 G 内存。实验环境为 Matlab 7。对于任意一个 DNA 序列，提取的特征应该满足以下 2 个条件[7]：①所取特征必须可以标志 A 类和 B 类，即这些特征可以很好地区分已经标示分类的 DNA 序列。②所取特征必须具有一定的实际意义。本文采用序列中的碱基 A、G、C、T 的含量百分比作为该序列的特征。不同段的 DNA 中，每个碱基出现的概率并不相同。从生物理论中可知，编码蛋白质的 DNA 中 G、C 含量偏高，而非编码蛋白质的 DNA 中 A、T 含量偏高。可见，A、T、C、G 的频率有很多生物信息。

3.1　人工 DNA 序列分类

选用 40 组人工序列进行实验[8]，其中，1～20 组为已知类别的序列，21～40 组为未知类别的序列。首先，对 40 组 DNA 人工序列进行特征提取，得到 4 种碱基的丰度、状态转移特征和错一位置的特征三字符串；然后选择一些能明显区分 2 类 DNA 的特征：单碱基选择 T 和 G，状态转移特征选择 GC、GG、TA、TT，错一位置的特征三字符串选择 ATT、AGG、CGG、TAA、TAT、TTA、TTT、GGA、GGC。设 A 类序列的期望输出值为 0，B 类序列为 1，则 RBF 网络的目标向量可表示为 T=[0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1]。

这样，通过观察 RBF 神经网络的输出值，可以直观地判断未知序列的类型。本实验将提取的 3 组特征向量放入上述 3 个神经网络分别进行分类。若网络输出结果小于 0.5，则判为 0；否则判为 1。图 1 所示为人工 DNA 判别分类点阵图，在 0 线以上的属于 A 类，以下的则属于 B 类。

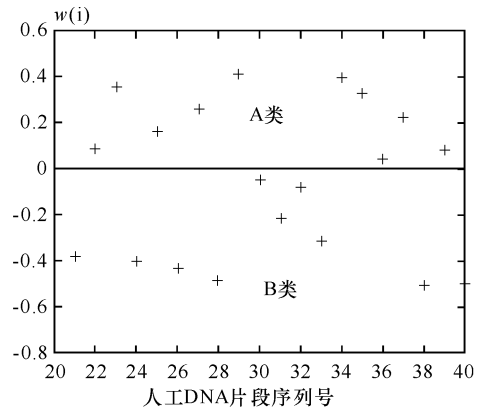

图 1　人工 DNA 判别分类点阵

表 1 给出了网络训练后的结果，表 2 给出了网

络分类后的结果。由表 2 可知，属于 A 类的序列号有 22、23、25、27、29、30、31、32、34、35、36、37、39；属于 B 类的序列号有 21、24、26、28、33、38、40。

表 1 优化后的 RBF 网络对自然序列的样本输出数据

序列号	期望输出	单碱基训练结果	双碱基训练结果	三碱基训练结果	优化后的网络
1	0	0.0085	0.0006	0.0148	0
2	0	0.0085	0.0006	−0.0054	0
3	0	0.0085	0.0006	0.0215	0
4	0	0.0085	0.0006	0.0090	0
5	0	0.0085	0.0006	−0.0043	0
6	0	0.0085	0.0006	−0.0115	0
7	0	0.0085	0.0006	0.0079	0
8	0	0.0085	0.0006	−0.0035	0
9	0	0.0085	0.0006	0.0120	0
10	0	0.0085	0.0006	−0.0046	0
11	1	0.9892	1.0086	1.0102	1
12	1	0.9892	0.9642	1.0033	1
13	1	0.9892	1.0026	0.9937	1
14	1	0.9892	1.0081	1.0067	1
15	1	0.9892	1.0084	1.0086	1
16	1	0.9892	0.9999	0.9946	1
17	1	0.9892	0.9827	0.9819	1
18	1	0.9892	1.0085	1.0121	1
19	1	0.9892	1.0077	0.9978	1
20	1	0.9892	1.0062	0.9971	1

3.2 自然 DNA 序列分类

本实验从选择的 DNA 序列上截取 3 000 个碱基，平均分成 25 组样本，每组样本 120 个碱基，前 10 组样本作为训练样本，后 15 组作为测试样本。单碱基选择 A、C、T 和 G，状态转移特征选择 AA、AT、CC、CG、GC、GG、TA、TT，错一位置的特征三字符串选择 AAA、AAC、AAT、AAG、ATA、ATT、CCC、CCT、CCG、CTT、CTG、CGG、TAA、TAT、TTA、TTC、TTT、GCC、GCG、GGC、GGG，建立 3 个 RBF 神经网络模型，表 1 和表 2 分别给出了经过 GA 优化的 RBF 神经网络对自然序列训练和分类的结果。可以看出，前 15 组 DNA 序列属于第一类，后 15 组 DNA 序列属于第二类，分类结果完全正确。

表 2 优化后的 RBF 网络对自然序列的实际输出数据

序列号	单碱基网络分类结果	双碱基网络分类结果	三碱基网络分类结果	优化后的网络
21	0.0085	0.0006	−0.0049	0
22	0.0085	0.0006	0.0410	0
23	0.0085	0.0006	0.0015	0
24	0.0085	0.0006	−0.0115	0
25	0.0085	0.0005	−0.0112	0
26	0.0085	−0.0036	0.1045	0
27	0.0085	0.0005	0.0391	0
28	0.0085	0.0006	0.0193	0
29	0.0085	0.0006	0.0728	0
30	0.0085	0.0006	−0.0021	0
31	0.0085	0.0006	0.0232	0
32	0.0085	0.0005	0.0120	0
33	0.0085	0.0006	−0.0031	0
34	0.0085	0.0003	0.0094	0
35	0.0085	0.0006	−0.0060	0
36	0.9892	1.0044	0.9809	1
37	0.9892	1.0083	1.0098	1
38	0.9892	1.0019	0.9946	1
39	0.9892	0.9917	0.9970	1
40	0.9892	1.0077	1.0079	1
41	0.9892	1.0065	1.0072	1
42	0.9892	1.0085	0.9947	1
43	0.9803	1.0057	0.9805	1
44	0.3409	0.9939	0.9935	1
45	0.0674	0.9716	0.9643	1
46	0.0088	0.9581	0.9732	1
47	0.9892	1.0023	1.0017	1
48	0.9892	1.0087	1.0144	1
49	0.9892	1.0087	1.0094	1
50	0.9892	1.0079	1.0060	1

4 结束语

通过对 DNA 序列分类技术进行了深入研究，针对 RBF 神经网络易于陷入局部最大值的缺点，采用遗传算法对 RBF 神经网络的权值进行优化，并把它应用于 DNA 序列分类中。该方法利用优化后的 RBF 神经网络对提取的每种特征分别进行训练分类。仿真实验结果表明，改进的算法比传统 RBF 神经网络分类有更高的分类效率和正确率，为解决分类、预测及模式识别等问题提供了新的途径。

参考文献：

[1] 李银山，杨春燕，张伟. DNA 序列分类的神经网络方法[J]. 计算机仿真，2003，20(2)：65-68.

[2] 冼广铭，曾碧卿，冼广淋. 最小二乘小波支持向量机的 DNA 序列分类方法[J]. 计算机工程与应用，2009，45(12)：222-225.

（下转第 268 页）

表 3 空间固定效应参数估计结果

地区	η_i	地区	η_i	地区	η_i	地区	η_i
北京	1.251	上海	1.252	湖北	0.955	云南	−0.756
天津	−0.170	江苏	3.366	湖南	1.101	西藏	−5.629
河北	1.589	浙江	2.273	广东	3.325	陕西	−0.132
山西	−0.005	安徽	0.277	广西	0.050	甘肃	−1.927
内蒙古	0.623	福建	0.661	海南	−3.892	青海	−4.192
辽宁	1.251	江西	−0.410	重庆	−0.434	宁夏	−3.980
吉林	−0.296	山东	3.279	四川	1.104	新疆	−1.142
黑龙江	0.269	河南	2.062	贵州	−1.596		

表 4 时间固定效应参数估计结果

年份	δ_t	年份	δ_t	年份	δ_t	年份	δ_t
2001	−2.571	2002	−1.845	2003	−1.144	2004	0.167
2005	0.255	2006	0.242	2007	0.836	2008	1.537
2009	0.624	2010	1.901				

3 结论

本文分别选用无固定效应模型、空间固定效应模型、时间固定效应模型和时空混合固定效应模型，分析我国 31 个省市 2001～2010 年的经济增长问题，不仅得出了回归系数的估计，还进一步做了空间和时间固定效应参数的估计，同时综合考虑了空间相关性和时间相关性，得到我国经济发展差异的影响因素。发现，我国区域经济增长呈现出明显的区域特征和阶段性特征。表明经济增长不仅依赖于地区自身条件，还与相邻地区的经济增长和经济周期息息相关。期初人均 GDP、年末人口总数和政府消费占 GDP 比重对我国经济增长的影响较为显著。

此外，由于空间权重矩阵的选取对研究结果有很大的影响，采用经济权重矩阵或更精确的权重矩阵是否可以得出更准确的结果，还有待于进一步研究。

参考文献：

[1] Anselin L. Spatial economics: methods and models [M]. Dordrecht :Kluwer Academic,1988.

[2] Elhorst J P. Specification and estimation of spatial panel data models[J]. International Regional Science Review ,2003,26(3):244-268.

[3] 胡洪胜. Two-way 空间自回归随机效应面板数据模型的检验[J]. 中央民族大学学报：自然科学报，2012，21(1):84-87.

[4] 季民河，武占云，姜磊. 空间面板数据模型设定问题分析[J]. 统计与信息论坛，2011，26(6):3-8.

[5] 梁亚民，臧海明. 区域差异视角下江苏省经济增长的实证分析——基于面板数据模型的研究[J]. 统计教育，2009，114(3):38-40.

[6] 苏良军，王芸. 中国经济增长空间相关性研究——基于“长三角”与“珠三角”的实证[J]. 数量经济技术研究，2007(12):26-37.

[7] 张志强. 空间面板数据参数估计的小样本特性探究[J]. 数量经济技术经济研究，2012(9):122-140.

[8] 中华人民共和国统计局. 国家数据[EB/OL]. [2013-06-10]. http://data. stats. gov. cn/workspace/index？m=hgnd.

（责任编辑：尹 闯）

（上接第 264 页）

[3] 马燕，范植华. 基于神经网络的基因分类器[J]. 计算机工程与设计，2005，26(2):308-311.

[4] 蔡春，苗立峰，邓乃扬. DNA 序列特征提取方法研究[J]. 北京联合大学学报：自然科学版，2008，22(4):70-72.

[5] 孙健，申瑞民，韩鹏. 一种新颖的径向基函数(RBF)网络学习算法[J]. 计算机学报，2003，6(11):45-47.

[6] Shu W，He B. A quantum genetic simulated annealing algorithm fortask scheduling[J]. ACM Computing Surveys，2006，33(1):115-127.

[7] 龚道雄，阮晓钢. 一种基于遗传算法的 DNA 多序列比对方法[J]. 北京工业大学学报，2003，3(1):19-22.

[8] Zhang Y，Waterman M S. An eulerian path approach to global multiple alignment for DNA sequences[J]. Journal of Computational Biology，2003，10(6):803-819.

（责任编辑：尹 闯）