遗传神经网络在邮件过滤器中的应用

王 波[1]，黄迪明[2]

(1. 中国科学院成都计算机应用研究所 成都 610041; 2. 电子科技大学计算机科学与工程学院 成都 610054)

【摘要】针对目前反垃圾邮件技术的缺点，提出一种基于遗传优化神经网络的垃圾邮件过滤器模型，利用遗传算法全局搜索能力优化神经网络连接权值，克服神经网络局部极小值点问题，提高神经网络的学习速度和识别能力。通过对遗传算法和人工神经网络算法的实现，证明它们在垃圾邮件过滤器中有很好的应用效果。

关 键 词 垃圾邮件; 遗传算法; 人工神经网络算法; 局部极小点

中图分类号 TP393.08 文献标识码 A

ANNA Optimized by GA and Its Application in E-mail Filter

WANG Bo[1]，HUANG Di-ming[2]

(1. Chengdu Institute of Computer Applications of Chinese Academy of Science Chengdu 610041;
2. School of Computer Science and Engineering, UEST of China Chengdu 610054)

Abstract This paper presents a model of spam mail filter based on artificial neural network which is optimized by genetic algorithms. By using genetic algorithms, which is good at wide searching ability in solution space on optimizing connection weight matrix of artificial neural network , artificial neural network can get over the inherent limitation of local minimal and improve its learning speed and recognition ability. The application of the model in spam filter has been successfully proved by its implementing.

Key words spam; genetic algorithm; artificial neural network algorithm; local minimal

电子邮件业务会产生大量的垃圾邮件(Spam)。为了提高电子邮件的服务质量，必须遏制垃圾邮件的产生和泛滥，多种反垃圾邮件基本技术应运而生。实践表明，采用基于邮件外部简单特征的过滤技术容易被垃圾邮件生产者欺骗，不能有效控制垃圾邮件，只有针对邮件内容整体特征进行邮件过滤，才可能有效地控制垃圾邮件。目前，基于内容统计特征的垃圾邮件过滤器主要采用的技术有Naive Bayes概率模型、人工神经网络算法(Artificial Neural Network Algorithms，ANNA)、粗糙集理论等。本文针对人工神经网络算法存在的不足，提出了一种模拟人脑智能，基于遗传优化神经网络的垃圾邮件过滤模型。

1 人工神经网络算法的局部极小值点问题

人工神经网络算法应用于邮件过滤，是通过它对模式的学习/识别能力来实现对垃圾邮件特征模式识别的一项新技术。人工神经网络算法在对输入样本学习时，初始连接权值是一个随机值，人工神经网络算法以该点为起点，搜索该局部的连接权值并使输出均方误差最小。在多层非线性网络中，作为局部搜索算法

收稿日期：2004 – 03 – 23
基金项目：四川省科技厅重点科技攻关资助项(03GG-006-021)
作者简介：王 波(1978 –)，男，硕士，主要从事网络信息技术方面的研究； 黄迪明(1944 –)，男，教授，主要从事网络信息技术方面的研究.

的人工神经网络算法，均方误差可能有多个局部极小值点，正常情况下，初始连接权值只有落在全局最小值点附近，均方误差才能经过人工神经网络算法收敛至最小值。然而，由于初始连接权值的随机性，均方误差可能落在全局最小值点附近，也有可能落在局部极小值点附近，前者人工神经网络算法可以正常收敛，后者只能收敛于一个局部最小值点。人工神经网络算法初始权值的非确定性带来了搜索的局限性，为了人工神经网络算法能正常收敛到全局最小值点，需要人工神经网络算法尽可能逃离局部极小值点，进入全局最小值点域内。一种可行的办法就是在人工神经网络算法中引入相应的全局优化技术，利用该技术，全局搜索并确定全局最小值点的范围，然后再由人工神经网络算法局部搜索到最小值点。

2 遗传算法与人工神经网络算法的结合

遗传算法(Genetic Algorithms，GA)与人工神经网络算法的结合是学习和进化之间交互作用的一种模式。图1表示人工神经网络算法的遗传设计过程[1]。人工神经网络算法建立学习模型，而遗传算法构成进化模型。遗传算法的进化建立在人工神经网络算法的基础之上，是人工神经网络算法学习能力的进化模拟，遗传算法的个体与具有学习能力的人工神经网络算法通过基因型和表现型来表达。遗传算法与人工神经网络算法的结合，不仅能发挥人工神经网络算法泛化的映射能力，而且使人工神经网络算法具有很快的收敛性和较强的学习能力，两者在反垃圾邮件过滤器中的应用，提高了过滤器的学习速度和识别能力。

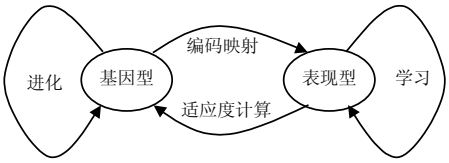

图1 GA +ANN

3 垃圾邮件过滤模型

3.1 GA+ANNA模型

GA+ANNA模型原理如图2所示。图中，文本表示采用向量空间模型，文本特征向量$\boldsymbol{X}=(x_1, x_2, \cdots, x_n)$，$W^i$ 和 B^i 是第i层神经元所对应的连接权值和偏置量。

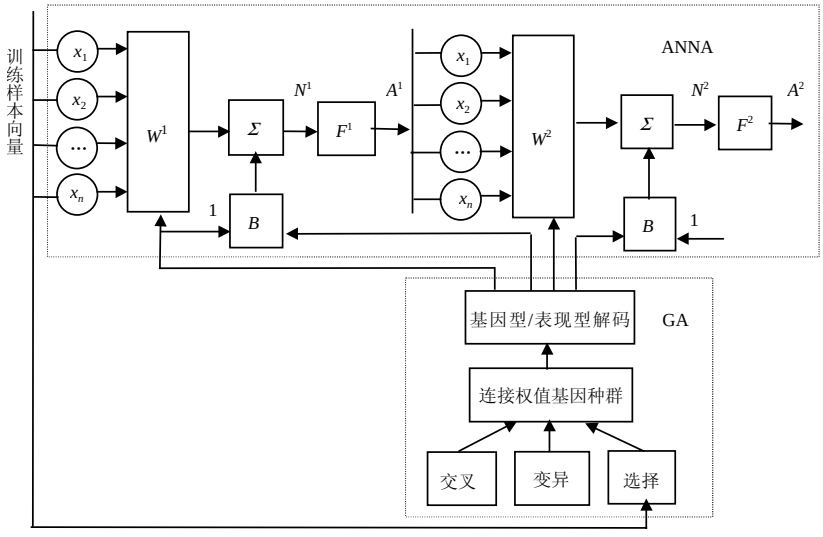

图2 GA+ANNA模型

3.2 GA+ANNA工作机制

1) 连接权值空间全局搜索阶段，利用遗传算法的随机性模拟人脑想象求解过程全局搜索人工神经网络算法权值空间，在全局解空间中找到最小值点的局部范围，从而提高人工神经网络算法收敛于最小值点的概率和收敛速度。

2) 连接权值空间局部搜索阶段，即识别/学习阶段，经过遗传算法搜索后的初始连接权值可能是最小值

点，也可能在最小值点附近，遗传算法搜索由于其交叉/变异算子带来的“扰动”，很难把连接权值收敛于最小值点，所以它的局部搜索能力有限，而且遗传算法本身不具备记忆/识别模式的能力，只能通过人工神经网络算法进行局部连接权值搜索和记忆，完成遗传算法不能完成的工作。

3) 识别阶段，经过遗传算法的全局搜索和人工神经网络算法的局部搜索，连接权值收敛于一个可以容忍的识别误差范围后，人工神经网络算法就可以用于对邮件文本特征向量的计算分类。

3.3 GA+ANNA过滤器实现

3.3.1 ANNA参数说明

人工神经网络算法使用对数-*S*型函数，有：

$$\begin{cases} F^{(i)}=\dfrac{1}{1+\mathrm{e}^{n}} \\ N^1=W^1X^1+B^1 \\ A^1=F^1(N^1) \\ N^2=W^2A^1+B^2 \\ A^2=F^2(W^2F^1(W^1X^1)) \end{cases} \tag{1}$$

式中　F^1，F^2 是传输函数。人工神经网络算法的学习算法采用标准三层神经网络反向传播(Back Propagation, BP)的监督学习算法，即：

1) 前向传播　$A^0=X$；$A^{m+1}=F^{m+1}(W^{m+1}A^m+B^{m+1})$，$m=0, 1, \cdots, M-1$；$A=A^M$；

2) 前后传播　$S^M=-2F^M(N^M)(Y-A)$；$S^m=F^m(N^m)(W^{m+1})^TS^{m+1}$，$m=M-1, \cdots, 2, 1$；

3) 更新权值和偏置　$W^m(k+1)=W^m(k)-\alpha S^m(A^{m-1})^T$；$B^m(k+1)=B^m(k)-\alpha S^m$ [2]。

为了验证该模型效果和特征向量输入方便，令输入层I=10；输出层O =2；隐含层H=6；最大迭代次数T=1 000；均方误差E=0.01；学习速度a=0.5

3.3.2 GA参数说明

遗传算法采用基本遗传算法，令种群β=15，选择淘汰率γ=0.4，变异率δ=0.01，适应度函数为：

$$f=(n\ \text{SAMPLE}-e)/\text{SAMPLE} \tag{2}$$

式中　迭代误差 $e=\sum_{i=0}^{sample}E_i^2$，$n$为人工神经网络算法输入层向量维数，SAMPLE为样本个数。

基因编码采用浮点数编码。神经网络是有向图，可用矩阵来确定一个神经网络。如果神经网络有n个神经元，可用一个$n\times n$矩阵表示神经网络的互连结构，“$w_{(i)(j)}$”代表该神经网络互连结构的神经元结点间的连接权值，“×”表示无连接。例如在某神经网络结构中，结点序号从1～18排列，1~10为输入层结点，11~16为隐层结点，17，18为输出层结点，根据神经网络节点间的限制，可列出该神经网络的连接矩阵，如表1所示。

表1　基因编码

节点	1	…	10	11	…	16	17	18
1	×	×	×	×	×	×	×	×
…	×	×	×	×	×	×	×	×
10	×	×	×	×	×	×	×	×
11	$w_{(1)(11)}$	$w_{(\cdots)(11)}$	$w_{(10)(11)}$	×	×	×	×	×
…	$w_{(1)(\cdots)}$	$w_{(\cdots)(\cdots)}$	$w_{(10)(\cdots)}$	×	×	×	×	×
16	$w_{(1)(16)}$	$w_{(\cdots)(16)}$	$w_{(10)(16)}$	×	×	×	×	×
17	×	×	×	$w_{(11)(17)}$	$w_{(\cdots)(17)}$	$w_{(16)(17)}$	×	×
18	×	×	×	$w_{(11)(18)}$	$w_{(\cdots)(18)}$	$w_{(16)(18)}$	×	×

该表中，将神经网络对应的连接权值编码按照从左到右、从上到下的顺序连接起来，可组成染色体编码。为了基因型/表现型的解码方便，把偏置量B^1和B^2嵌入连接权值中，得一染色体编码：

$$\langle\ w_{(1)(11)}\ w_{(\cdots)(11)}\ w_{(10)(11)}\,|\,w_{(1)(\cdots)}\ w_{(\cdots)(\cdots)}\ w_{(10)(\cdots)}\,|\,w_{(1)(16)}\ w_{(\cdots)(16)}\ w_{(10)(16)}\,|\,b_1^1\ b_2^1\cdots b_6^1\,|\,w_{(11)(17)}\ w_{(\cdots)(17)}$$

$$w_{(16)(17)} \mid w_{(11)(18)}\ w_{(\cdots)(18)}\ w_{(16)(18)} \mid b_1^2\ b_1^2\ \rangle$$

该染色体编码构成一个神经网络，对应于种群矩阵的一个行向量，染色体编码(一维向量)与表现型(二维连接权值矩阵)之间通过一个算法实现编码映射。

3.4 GA+ANNA学习/识别测试

为不失一般性，表2只提供两个10维模拟输入样本和神经网络的期望输出结果。

表2 训练样本

训练样本特征向量										期望输出	
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y_1	y_2
1	2	3	4	5	6	7	8	9	10	1	0
10	9	8	7	6	5	4	3	2	1	0	1

图3是在表2输入样本下没有用遗传算法全局搜索连接权值时，某次人工神经网络算法的迭代次数和它的均方误差。可以看出，误差收敛于0.771。图4中的实线是遗传算法运行结果的最大适应度曲线，虚线是该运行结果的平均适应度曲线，经过60代遗传后，平均适应度已经很接近最大适应度，算法可以就此终止。图5是经过某次遗传算法全局搜索后，人工神经网络算法在局部搜索迭代的一次误差，该误差可收敛到0.006附近。由于遗传算法已经把初始连接权值收敛到误差范围内，人工神经网络算法只需学习该模式即可。对比图3和图5，遗传算法下的人工神经网络算法误差能收敛于0.006，而非遗传算法下的人工神经网络算法误差只能收敛到0.771，说明图3中人工神经网络算法的误差可收敛到局部极小值点。图5说明遗传算法的引入不仅克服了人工神经网络算法的局部极小值点问题，同时也提高了人工神经网络算法的收敛速度，经过一次迭代学习便可使输出误差收敛到设定范围之内。

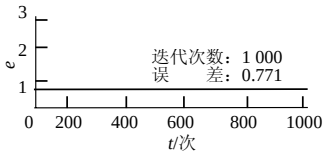

图3 ANN运行结果

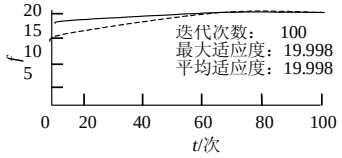

图4 GA运行结果

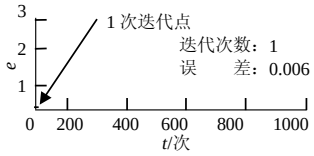

图5 GA+ANN运行结果

表3是两个不同邮件特征向量和它对应的实际输出值。由表3可以看出，对样本模式的“扰动”输入同样能得到和样本模式标准输出接近的实际输出。第一个输入样本被识别为y_1类，第二个输入样本在一定的阀值下被识别为y_2类，实践表明，基于遗传算法的人工神经网络算法垃圾邮件过滤模型是可行和有效的。

表3 识别结果

邮件文本特征向量										实际输出	
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	a_1	a_2
1	0	3	0	5	0	7	0	9	0	1.000	0.000
10	0	8	0	6	0	4	0	2	0	0.000	0.998

4 结 束 语

本文提出了基于遗传优化神经网络的智能邮件过滤器。人工神经网络算法中引入遗传算法，利用两者的互补优势，提高了人工神经网络算法收敛于全局最小值点的可能性和速度，能够满足基于遗传优化神经网络的智能邮件过滤器的应用要求，实现对垃圾邮件的有效过滤。

参 考 文 献

[1] 王小平, 曹立明. 遗传算法–理论、应用与软件实现[M]. 西安: 西安交通大学出版社, 2002. 18-50

[2] Martin T H, Howard B D, Mark H. Neural network design[M]. 北京: 机械工业出版社, 2002. 197-221

编　辑　熊思亮