

基于遗传神经网的沙尘暴预测模型的研究

路志英¹ 杨玉峰¹ 赵智超¹ 庞 勇¹ 刘还珠²

¹(天津大学电气与自动化工程学院,天津 300072)

²(国家气象中心,北京 100081)

E-mail: syge@tju.edu.cn

摘要 神经网络结构和学习算法在很大程度影响模型建立的时间和预测结果的准确性,论文针对现有BP神经网络的缺陷,提出了基于遗传算法的神经网络预测模型,并利用该模型研究了沙尘暴的预测问题。仿真预测研究表明,提出的方法无论是收敛速度方面还是预报准确率方面,都取得了比较满意的结果。

关键词 遗传算法 神经网络 沙尘暴预报

文章编号 1002-8331-(2005)33-0220-03 文献标识码 A 中图分类号 TP18

The Study of the Sand-bust Storm Forecasting Model Based on GA-neural Network

Lu Zhiying¹ Yang Yufeng¹ Zhao Zhichao¹ Pang Yong¹ Liu Huanzhu²

¹(School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072)

²(National Meteorological Center, Beijing 100081)

Abstract: The original structure of the network and the learning algorithm may lead to significant impact in the training duration and veracity of the forecasting result. To improve it, a novel design of artificial neural network based on genetic algorithm for sand-dust storm forecasting is proposed in this paper. Some practices prove that this model can perform better than the traditional neural network.

Keywords: genetic algorithm, artificial neural network, sand-dust storm forecast

1 引言

近年来,我国沙尘暴、扬沙和浮尘天气频繁发生,严重干扰和影响着人们的正常生活,对社会经济和生活环境均造成一定程度的危害,使人们愈来愈认识到沙尘暴是不可忽视的大气、生态以及生存环境的重要问题之一。所以,人们加强了对沙尘暴预报的研究,但由于沙尘暴预报的复杂性,一直不能取得比较满意的结果^[1],其主要原因是沙尘暴预报的数据量非常大。

BP网络是一种有导师的学习方法,算法的核心是通过一边向后传播误差,一边利用最小二乘法的最陡下降梯度法进行权值的调整,以实现或逼近所希望的输入输出映射关系。神经网络的结构和学习算法在很大程度上影响着预报的准确率和训练时间,这已成为共识。如果只采用神经网络来进行预报,则网络结构的设置和学习算法的选择将是一个较难的问题。而遗传算法(GA)是基于自然选择和遗传规律的并行全局搜索算法,具有较强的宏观搜索能力,具有寻优的全局性,克服了BP算法中容易出现的局部极小问题,对根据预报要求而进行的网络初始结构和学习算法的选择能起到一定的指导作用。

本文针对神经网络的初始网络结构设置和学习算法选择比较难的问题,采用遗传算法对神经网络结构和学习算法进行优化,并根据优化后的网络结构和学习算法,构建出了适合沙尘暴的自适应预报的模型,通过对数据的仿真测试,取得了比

较满意的结果。

2 遗传算法优化神经网络模型的建立

2.1 神经网络模型

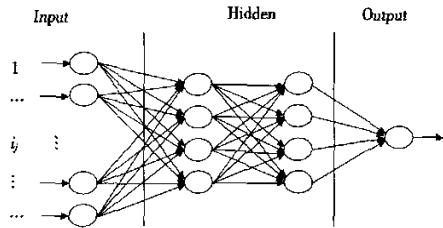


图1 多层前馈神经网络

图1为多层前馈神经网络模型,网络输入节点和输出节点视具体情况而定,而隐层数、各隐层节点数及所使用的学习算法均由遗传算法优化决定。隐层节点采用 *tansig* 函数作为激励函数,其表达式为:

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} \quad (1)$$

输出节点的激励函数采用线性函数 *purelin*,即:

作者简介:路志英(1964-),女,硕士,副教授,硕士生导师,主要从事人工智能、模式识别等方面的研究。杨玉峰(1981-),男,模式识别与智能系统专业硕士研究生(在读),研究方向为人工智能。赵智超(1980-),男,模式识别与智能系统专业硕士研究生(在读),研究方向为人工智能。庞勇(1979-),男,模式识别与智能系统专业硕士研究生,研究方向为人工智能。刘还珠(1944-),女,正研高工,国家气象中心,主要从事人数值预报方面的研究。

$$f(x)=x \quad (-1 \leq x \leq 1) \quad (2)$$

公式(1)和公式(2)中的 x 均为神经元输入值。每个神经元 j 的输出可通过下式计算:

$$Y_j = f(\sum_{i=1}^n w_{ij} x_i + b) \quad (3)$$

其中 f 是激励函数, x_i 是前一层神经元的输出, w_{ij} 为神经元 j 与前一层每个神经元的连接权值, b 是这个神经元的阈值。为了简化问题复杂性, 限制最大隐层数为 2。

2.2 遗传算子的编码方式

遗传编码采用二进制编码方式。编码串共 10 位, 由三部分组成, 即:



其中: 第一部分表示第一隐层节点数。

第二部分表示第二隐层节点数。

第三部分表示神经网络所采用的学习算法,

00 表示 Levenberg-Marquardt 算法(LM);

01 表示 Resilient Backpropagation 算法(RP);

10 表示 Scaled Conjugate Gradient 算法(SCG);

11 表示 Quasi-Newton 算法(QN)。

如果某一隐层节点数为 0, 则表示该隐层不存在, 但由于第一隐层必须含有节点, 所以节点数采用加 1 即范围为 1~16。例如编码串: 0111 1100 01, 表示第一隐层为 8 个节点, 第二隐层为 12 个节点, 采用 RP 算法进行学习。

2.3 适应值函数

适应值函数定义如下:

$$f = \frac{1}{e} (1 + coef_1 \times \frac{1}{N_1} + coef_2 \times \frac{1}{N_2}) \quad (4)$$

其中 N_1, N_2 分别是第一、二隐层节点数; $coef_1$ 和 $coef_2$ 是根据对网络运算速度的需要设定的系数, 系数值越大, 表示对网络速度的要求越高; e 是所有测试样本的平均泛化误差, 其定义如下:

$$e = \frac{1}{n} \sum_{i=1}^n e(i) \quad (5)$$

n 为测试样本个数, $e(i)$ 为第 i 个测试样本泛化误差, 其定义如下:

$$e(i) = \frac{|y_i - \bar{y}_i|}{\bar{y}_i} \quad (6)$$

y_i 为测试样本实际输出, \bar{y}_i 为目标输出。

对于每次进化之后形成的新个体, 首先使用学习算法进行训练, 然后计算其适应值。由于学习过程占用大量的时间, 为了减少进化时间, 将学习算法的最大训练周期定为 50, 这大大节省了进化所需要的时间。从大量的仿真实验结果来看, 所选择的 4 种算法在前 50 代时误差迅速下降, 而在此之后下降缓慢, 所以减少训练周期对适应值评价的准确性影响不大。当遗传算法对网络的结构和学习算法进化完成之后, 再将网络结构设置为进化后的节点数, 同时将所选择的学习参数调整为最大 500 个周期, 用相应的学习算法进行训练, 构造出所需要的神经网络模型。建立模型的具体过程如下:

Procedure GA-ANN

Begin

 Initialize $P(0)$;

$t=0$;

$N=50$; // 进化代数;

 Use the train method to train $P(t)$;

 Begin

 while($t \leq N$) do // 种群大小;

 Evaluate fitness of $P(t)$;

 End for

 for $i=1$ to M do

 Select operation to $P(t)$;

 End for

 for $i=1$ to $M/2$ do

 Crossover operation to $P(t)$;

 End for

 for $i=1$ to M do

 Mutation operation to $P(t)$;

 End for

 Use the train method to train $P(t)$;

 for $i=1$ to M do

$P(t+1)=P(t)$;

 End for

 for $i=1$ to M do

 Evaluate fitness of $P(t)$;

 End for

$t=t+1$;

End while

Choose three best $P(t)$;

Change the parameter of the learning algorithm and train the ANN

Choose the best structure

End

End

3 遗传神经网在沙尘暴预报中的应用

3.1 沙尘暴预报问题及样本数据的特点

沙尘暴预报问题具有小概率、多因子、高维、样本数据量大、建模样本量不均衡等特点, 并且气象部门对于各因子与沙尘暴发生的因果关系也在探索的过程中, 还没找到明确的因果关系表达式。

本文采用的数据源来自美国环境气象中心(NCEP)每天所提供的资料, 它为 2.5×2.5 (每格 2.5°) 格点场数据阵。通过对沙尘暴的成因的分析, 其主要因素体现在 4 个物理格点场上, 即高度值场(H5)、高度形场(H5_dm)、位温形场(SE_dm)和风形场(UV_dm), 前三个物理场的数据维数均为 $19 \times 9 = 171$, 风形场为 $19 \times 9 \times 2 = 342$, 因此每一个样本数据的维数高达 855, 数据量非常大。由于高维数据量太大, 很难用来预报, 因此实验室人员在数据处理方面进行了多方面的研究, 并应用数据挖掘进行了预处理, 经过处理后, 综合预报模型的总输入为 24 个。

3.2 试验结果与分析

为了全面分析所用神经网络综合预报性能, 实验选用了 1981 年~1997 年的数据样本, 其中 1981 年~1995 年样本为训练样本, 1996 年和 1997 年样本作为测试样本。样本的具体构

表 2 不同网络的预报结果

训练算法	网络结构	训练时间/s	正确报无	正确报有	空报	漏报	CSI	
传统网络	LM	24-20-10-1	845.2970	165	13	51	12	17.10%
遗传神经网络	RP	24-16-8-1	37.031	183	20	33	5	34.48%

* 赛扬 2.4GPC 机 ** 阈值为 0.6,0.4, 大于 0.6 认为预报正确, 小于 0.4 认为预报错误

成如表 1 所示。

表 1 试验样本

	样本数	沙尘暴天气	非沙尘暴天气
训练用数据(1981 年~1995 年)	1803	550	1253
测试用数据(1996 年~1997 年)	241	25	216

借助 matlab 平台, 实现系统的仿真, 可得到 24-16-8-1 的优化网络结构, 其学习算法为 RP。有关系统的性能如表 2 所示。

其中 CSI 为成功界限指数, 气象上常用 CSI 来衡量预报模型的性能。CSI 的定义如下:

$$CSI = \frac{c_f}{c_f + w_f} \times 100\% \quad (7)$$

c_f 为正确报出的沙尘暴日数, w_f 为漏报与空报数之和。将表 3 中的数据代入计算式(7)可得: 传统多层网络的 CSI 值为 17.10%, 遗传神经网络的 CSI 为 34.48%, 比传统多层网络提高了 2 倍。另外, 由表 3 还可以看出: 遗传神经网的训练时间也大大地缩短了。由此可以表明, 本文的遗传神经网络综合预报方法的效果较传统网络有明显提高。

图 2 为本试验的线性拟合。图中横坐标 0 表示没有沙尘暴

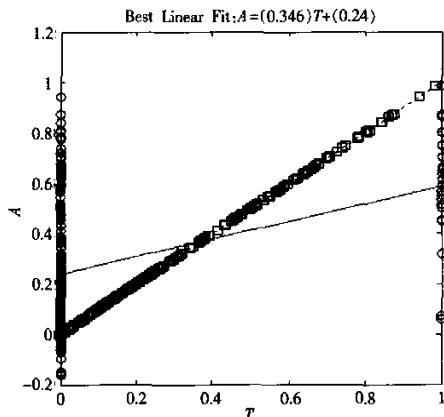


图 2 试验结果拟合曲线

(上接 155 页)

communications, Computers and signal Processing PACRIM, Pacific Rim Conference, 2003; 1: 405~408

10.V M Vokkarane, J P Jue. Prioritized burst segmentation and composite burst-assembly techniques for QoS support in optical burst-switched networks[J]. IEEE Journal on Selected Areas in Communications, 2003; 21(7): 1198~1208

11.王建新,曾庆济,张治中等.光突发交换网络中支持 QoS 的一种竞争处理机制[J].电子与信息学报,2004;26(6):1000~1004

12.K Dolzer, C Gauger, J Spath et al. Evaluation of reservation mechanisms in optical burst networks[J]. AEU Int. Journal of Electronic Communication, 2001; 55(21): 1~8

13.K Dolzer, C Gauger. On burst assembly in optical burst switching networks-a Performance evaluation of just-enough-time[C]. In: Proceeding Information Theory Conference 17, Salvador, Brazil, 2001; 4: 222~226

2005.33 计算机工程与应用

万方数据

发生的日子, 1 表示有沙尘暴发生的日子, 圆圈点代表预报出来的数据点。带有方块的线为理想情况下的预报结果, 即全部预报正确时值全为 0 或全为 1。实线则为根据预报结果所得的拟合线。

通过上述结果可以看出, 采用遗传算法进行神经网络结构的优化和学习算法的选择, 可以避免初始网络结构选择的盲目性, 同时也可以选取合适的学习函数, 从而在一定程度上提高了建模的快速性和预报的准确性。由此表明, 通过遗传算法进行网络优化对网络结构的设置和训练函数的选择起到了指导作用, 并产生了较好的效果。

4 结论

本文针对 BP 神经网络的缺陷, 即初始网络结构和学习算法的设置将很大程度影响模型建立的时间和预测结果的准确性, 并根据沙尘暴天气的特点, 提出了基于遗传算法的神经网络预测沙尘暴的模型。通过传统神经网络与遗传神经网络仿真结果的比较, 表明: 遗传算法优化后得到的模型无论在收敛速度方面还是在预报准确率方面, 都取得了比较满意的结果。当然由于沙尘暴预报问题的特点及其数据的复杂性, 对其研究还有待于进一步深入。若能对原始数据作更深入的分析, 挖掘出尽可能少而精的特征, 并利用遗传算法进一步设计相应的神经网络来提高预报的准确度将是下一步研究的方向。

(收稿日期: 2005 年 3 月)

参考文献

1. 正式功等. 沙尘暴研究的进展[J]. 中国沙漠, 2000-12: 349~356
2. Wenbo Zhao, De-Shuang Huang, Ge yunjian. The Structure Optimization of Radial Basis Probabilistic Neural Networks Based on Genetic Algorithms[C]. In: Neural Networks, IJCNN '02, 2002; 1086~1091
3. 周明, 孙树栋编著. 遗传算法原理及应用[M]. 国防科技大学出版社
4. 贺哲, 韩雪英, 孔海江等. 沙尘暴天气的成因及其天气形势分析[J]. 河南气象, 2000; (4): 3~4
5. 冯利华. 神经网络在天气预报中的应用[J]. 信息与控制, 2001-08: 365~367
14. M Neuts, Z Rosberg, H L Vu et al. Performance analysis of optical composite burst switching[J]. IEEE Communications Letters, 2002; 6(8): 346~348
15. M Zukerman, E W M Wong, Z Rosberg et al. On teletraffic applications to OBS[J]. IEEE Communication Letters, 2004; 8(2): 116~118
16. M Zukerman, E W M Wong, Z Rosberg et al. Teletraffic modeling of optical burst switching[C]. In: Transparent optical networks, 2003 Proceedings of 2003 5th International conference on, ICTON 2003; 1: 82~86
17. J Hui. Switching and Traffic Theory for Integrated Broadband Networks. Kluwer, 1990
18. H L Vu, M Zukerman. Blocking probability for priority classes in optical burst switching networks[J]. IEEE Communications Letters, 2002; 6(5): 214~216