

沙尘暴综合预报系统中场特征提取方法的研究

路志英, 杨乐, 赵智超, 杨玉峰

(天津大学电气及自动化学院, 天津 300072)

摘要:沙尘暴天气频繁发生, 严重干扰和影响着人们的正常生活, 对社会经济和生态环境均造成一定程度的危害。根据沙尘暴天气的成因及特点, 在分析现有特征提取方法弊端的前提下, 提出了 ANN 沙尘暴综合预报系统场特征提取方法。首先进行距平运算数据预处理, 获取沙尘暴成因物理场样本; 然后利用论文方法进行场特征提取。在此基础上, 借助综合预报思想构建并实现了遗传 ANN 沙尘暴预测系统。仿真结果表明, 论文所研究的场特征提取方法在提高沙尘暴预报的准确率方面做出了较大的贡献。

关键词:特征提取; 人工神经网络; 沙尘暴; 综合预测

中图分类号:TP391.9 **文献标识码:**B

A Field Feature Extraction Method of Sand - dust Storm Ensemble Forecast System Based on ANN

LU Zhi - ying, YANG Le, ZHAO Zhi - chao, YANG Yu - feng

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

ABSTRACT: The sand - dust storm happened more and more frequently in recent years. It not only seriously affects the normal life of people, but also do certain harm to the social economy and the environment. According to the origin and characteristic of sand - dust storm weather, a method of field feature extraction of sand - dust storm ensemble forecast system on GA - neural network is proposed in this paper, which is based on the analysis of the shortcomings of existing feature extraction methods. Firstly, the origin physics field sample of the sand - dust storm can be acquired after the preprocessing of characteristic data. Then the field feature extraction is performed by the presented method. On this basis, the dust storm forecast system on GA - neural network is constructed and realized with the aid of the ensemble forecast theory. The emulational results show that the presented method obtains a better achievement in enhancing the veracity of the sand - dust storm forecast.

KEYWORDS: Feature extraction; Artificial neural network; Sand - dust storm; Ensemble forecast

1 引言

沙尘暴是沙暴和尘暴两者兼有的总称, 是指强风把地面大量沙尘卷入空中, 使空气特别混浊, 水平能见度低于 1km 的天气现象^[1]。近年来, 沙尘暴在我国频频发生, 沙暴过处, 农田受损、沃土流失, 加剧土地沙化, 造成大气环境严重污染, 侵蚀自然生态环境, 影响交通和供电线路的正常运作, 给人民生命财产造成严重损失。因此对沙尘暴的检测与预报日渐成为重要研究课题。

作为气象预报基础的历史观测数据, 无论来自遥感图像还是来自地面站点, 都具有空间上的场分布性。鉴于天气系

统的互相关性, 对某一地点的天气进行预报应综合考虑本地的历史数据及临近区域的影响力。因此无论是某地点预报还是某区域的预报, 所用方法都要面对区域场才能实现。正是由于数据源上的这一特点, 增加了研究处理方法的难度。

所有原本接纳向量输入的分类与回归方法(处理器), 当面对场数据输入时, 需要在场数据源与处理器入口间插入若干预处理环节完成场数据映射为向量数据的转换, 这在大部分情况下往往需要做出取舍以满足表达数据量的限定。然而信息选取的标准因应用的目标不同而异, 例如: 某种方法能够成功地从场源中提取出影响降雨的相关信息, 但它不一定能在提取的过程中有效地保留与沙尘暴有关的信息, 从而适用于沙尘暴预报。因此, 对于沙尘暴预报系统, 有必要探索新的与之更相配的物理场及其特征提取的方法。

本文针对沙尘暴数据的高维度的特点,对原有物理场特征提取方法进行进一步的改进,将每个物理场作为独立的预报信息源进行分级特征提取,从而达到简化向量规模的目的,克服了原有基于聚类方法^[2]的数据冗余使计算负担沉重的问题。通过基于遗传神经网络的沙尘暴综合预报系统的预报结果证实本文所提出的特征提取方法的有效性,提高了综合预报的准确性。

2 基于遗传神经网络的沙尘暴综合预报系统

沙尘暴历史数据集具有场分布广,维度高,数据量大的特点。文献[2]针对沙尘暴数据的特点,提出了一种基于聚类的场特征提取方法。但是该方法得出的40个特征个数太多且存在重复的表征,从而阻碍了区分尺度的进一步细化。本文针对上述方法的缺陷,在沿用分场聚类思路的基础上,通过计算沙尘暴各场的典型模式,并求取各场数据与典型模式距离,从而获取了数据压缩的表示,降低了数据维度,提高了预报系统数据输入的精确度。

遗传神经网络的沙尘暴综合预报系统框架如图1所示。该系统共有24个因子作为输入(4×5个物理场特征,4个单场预报输出结果),通过遗传算法对LM神经网络的优化,最终实现综合预报。

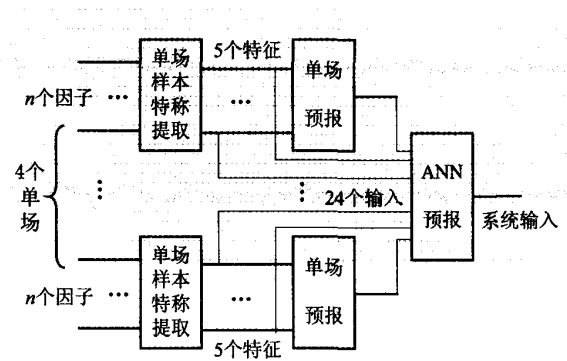


图1 遗传神经网络的沙尘暴综合预报系统框架

3 沙尘暴物理场的特征提取方法

3.1 沙尘暴物理场数据预处理

目前气象部门对于沙尘暴发生的因果关系还没找到明确的因果关系表达式,正处于探索的过程中。根据文献[3]对沙尘暴的成因分析,我们选取了体现主要因素的4个物理场格点场,即高度值场(H5)、高度场形(H5_dm)、位温场形(SE_dm)和风形场(UV_dm),它们反映了各个源场的分布信息。

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。前三个物理场的数据维数均为19×9=171,风形场的数据维数为19×9×2=342,因此,每一个样本数据的维数高达855。由于沙尘暴数据源的多因子、高维、样本数据量大、建模样本量不均衡以及发生沙尘暴的小概率等特点,所以,必须通过样本的特征提取来降低样本维数,否则,无法建立基于遗传ANN的综合预报系统,继而实现较高预报准确率的目的。选定数据如表1所示。

以500hpa高度值场(H5)作为值参照信息保留,对三个源场(H5、SE、UV)分别做距平运算,即可得到H5_dm(高度场形)、SE_dm(位温场形)、UV_dm(风形场)三个物理场。

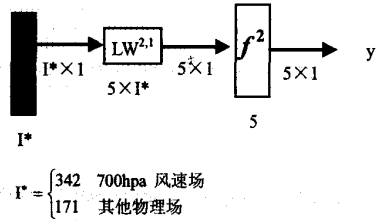


图2 SOM 聚类网络

3.2 物理场的特征提取

将上述4个物理场看作4个独立的信息源,分别进行场的特征提取。具体方法:

时间 (年份)	样本数(单场)			地点		源场(测量场)		
	沙尘暴	非沙尘暴	发生概率	东经	北纬			
1981 ~ 1997				70° ~ 115°	35° ~ 55°	500hpa	850hpa	700hpa
2月11日 ~ 6月10日	575	1469	28.1%	(均分18格)	(均分8格)	高度场(H5)	位温场(SE)	风速场(UV)

* 风速为投影到东西和南北两个垂直方向上的值;

1) 选取物理场中历史上的沙尘暴日样本数据,构造新的样本集(每个场样本数575),它们分别是:H5_sa、H5_dm_sa、UV_dm_sa、SE_dm_sa;

2) 利用自组织映射神经网络^[4]分别对H5_sa、H5_dm_sa、UV_dm_sa、SE_dm_sa进行聚类,其网络结构如图2所示。由图1可见,每个样本集可聚5类;

3) 对每一类,求其类中心。具体计算:

$$M_k^{(L)} = \frac{1}{n} \sum_{i=1}^n s_k^{(L)}(i)$$

其中: $s_k^{(L)}(i)$ 为第L个物理场第k类第i个样本; $L=1, 2, 3, 4; k=1, 2, 3, 4, 5; n$ 为第L个物理场的样本数。

4) 计算历史样本与类中心的距离,以获取单场特征。

设 $a_{ij}^{(L)}$ 为第 L 个物理场第 i 个样本第 j 个点数据,则第 L 个物理场第 k 个特征为:

$$f_k^{(L)} = \sqrt{\sum_{i,j} (a_{ij} - M_k^{(L)})^2}$$

其中: $L = 1, 2, 3, 4, 5$; $i = 1, 2, \dots, 2044$; $j = \begin{cases} 342 & 700hpa \text{ 风速场} \\ 171 & \text{其他物理场} \end{cases}$ 。

根据此方法分别对高度值场、高度形场、位温形场、风形场进行了特征提取,其结果为每个物理场 5 个特征,因此,对于遗传神经网络沙尘暴综合预报系统的输入特征来说共计 20 个。

4 遗传神经网络预报模型的实现与实例运行

4.1 分级与综合回归

综合预报系统中,当输入向量包含多个特征时,需将这些特征分级进行局部回归,然后将局部结果也作为总回归的输入,这会提升总回归的效果^[5]。所以,首先对不同物理场的特征进行分级,即将不同物理场的特征分别作为一个反向传播神经网络分类器(BP 分类器)的输入,进行训练,得到最佳拟合的神经元输入权值;然后将输出结果提供给分级回归的中间层,通过分级回归得到单场拟合局部值;最后再将局部值与单场特征一起作为综合回归总输入,这样每个场都将对总输入有六个值的贡献。所以,沙尘暴样本数据的维数从 855 降至 24。维数的大幅降低,使综合回归预报模型的建立成为可能。

4.2 系统实现及参数设置

基于遗传 ANN 沙尘暴综合预报系统通过 C++ 语言和 Matlab 语言实现的。由于 C++ 语言具有面向对象的特性,因此程序段可以得到充分的重用。系统中所涉及的 SOM 和 BP 神经网络是从同一神经网络基类继承而来的。

为与文献[1]中的场特征提取方法进行比较,本文使用了与之相同的数据集,将 81 ~ 95 年的样本作为训练数据,96 ~ 97 年的样本作为测试数据,样本信息如表 2 所示。

表 2 样本信息

	样本数	沙尘暴天气	非沙尘暴天气
训练数据(81 - 95 年)	1803	550	1253
测试数据(96 - 97 年)	239	25	214

用于单场预报的 BP 神经网络采用四层结构^[6],包括 2 个隐层,1 个输入层,1 个输出层;对于综合预报神经网络,通过对结构和方法的优化,得到 LM 算法的四层网络结构。系统结构如图 3 所示。

上述单场和综合 LM 神经网络中各神经元连接权值及结构通过遗传算法获取,神经元的激活函数 f 为 Sigmoid 函数,可以实现特征的非线性回归。

4.3 测试结果

将 96 ~ 97 年的样本输入基于遗传 ANN 沙尘暴综合预

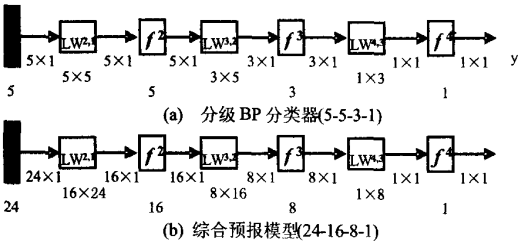


图 3 所用神经网络结构示意图

(结构参数根据多次实验,选择拟合效果最佳者)

报系统进行测试,得到的预报结果可能出现四种情况:正确报有,正确报无,空报(系统输出为有,但实测结果为零),漏报(系统输出为零,但实测结果为零)。为了说明预报模型的性能,采用气象上常用的成功界限指数 CSI 可作为衡量指标。CSI 的计算方法为:

$$CSI = \frac{c_f}{c_f + w_f} \times 100\%$$

其中: c_f 为正确报有的样本数, w_f 为漏报与空报的样本数之和。将测试结果带入上式计算,CSI 值为 41.3%,而文献[1]所用聚类方法得到的最高 CSI 值为 29.4%。

4.4 测试结果分析

由上述结果可以看出,本文的遗传神经网络综合预报方法的效果较以前有明显提高。

遗传 ANN 综合预报方法能取得上述性能提升的原因主要有如下几个方面:

1) 原方法^[2]在分场聚类后进行了场间的类组合,以便沟通各场的信息,但这样做可能造成冗余场特征的出现。由于本文构建的综合预报可以在场间作出权衡取舍,聚类后的场间组合过程没有必要存在,因此本文提出的特征提取新方法省去类组合,从而消除了冗余特征。

2) 原方法从场间组合出的 24 个沙尘暴类中选择了包含样本多的十个类,这势必造成其它类的舍弃及其信息的丢失。本文提出的新方法没有过高的类数目,没有场间组合,因此没有人为地舍弃任何信息。

3) 由于没有场间组合,场内的聚类数可以进一步增加,使沙尘暴的类型可以进一步细分,从而提高预报的准确性。

5 结论

本文提出了遗传 ANN 沙尘暴综合预报系统中的场特征提取方法,并利用这一方法针对 96 ~ 97 年的样本进行了综合预报,有效地提高了沙尘暴预报的准确性,表明了方法的有效性。尽管如此,由于沙尘暴预报问题的特点及其数据的复杂性,还需对其作更深入的研究。例如,单场特征的提取方法可以采用时间序列预测方法得出法特征序列预测值,从而实现完全统计意义上的沙尘暴预报。

参考文献:

- [1] 王式功, 董光荣. 沙层暴研究的进展[J]. 中国沙漠, 2000, 20(4): 349 - 356.
- [2] 王萍, 等. 一种大规模样本数据的特征提取方法[J]. 天津轻工业学院学报, 2003, 18: 48 - 52.
- [3] 路志英, 等. 基于神经网络的多模型综合预报方法[J]. 计算机应用, 2004, 24: 50 - 51, 88.
- [4] M T Hagan, H B Demuth, M H Beale, Neural network design(electrical engineering)[D]. Brooks Cole, 1995.
- [5] 路志英, 等. 基于种群多样度的变参数遗传算法的研究[J]. 计算机仿真, 2006, 23(1): 96 - 99.
- [6] H Goldstein. Multilevel Statistical Models[D]. Edward

Arnold, London. 1995.

[作者简介]



路志英(1964 -), 女(汉族), 天津人, 硕士, 副教授, 主要研究方向: 进化算法、人工神经网络。

杨乐(1982 -), 男(汉族), 河北人, 硕士生, 主要研究方向: 自然语言处理, 数据挖掘算法。

赵智超(1980 -), 男(汉族), 河北人, 博士生, 主要研究方向: 数据挖掘算法。

杨玉峰(1981 -), 男(汉族), 河北人, 硕士生, 主要研究方向: 数据挖掘算法, 图像处理。

(上接第76页)

变化的需要, 可以调整线性调频范围 B 。根据式(6)和式(22), 定义随机脉位线性调频脉冲表达式为

$$s(t) = s_g(t) \cdot s_e(t) \quad (23)$$

其部分时域图如图4所示。

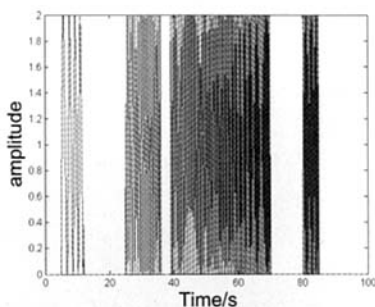


图4 随机脉位线性调频脉冲信号

根据 Woodward 的定义^[4,5], 随机脉位信号的平均模糊函数的表达式为

$$X(\tau, \xi) = K \cdot E |x(\tau, \xi)|^2 \quad (24)$$

式中, $x(\tau, \xi) = \int_{-\infty}^{+\infty} u(t)u^*(t+\tau)e^{j2\pi\xi t} dt$; $E(\cdot)$ 表示取均值。

将式(6)代入式(24), 计算机仿真的随机脉位线性调频脉冲信号的模糊函数以及等高线图如图5所示。

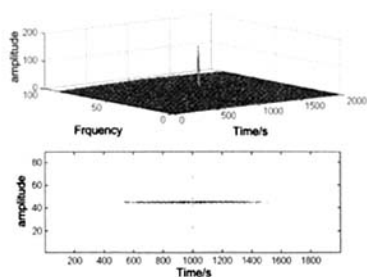


图5 随机脉位线性调频脉冲信号的模糊函数以及等高线图

由图5可见, 由噪声幅度过门限产生的随机脉位、随机脉宽线性调频脉冲信号的模糊函数, 其特性极其类似于噪声的模糊函数, 说明它具有很好的类噪声性。

5 结论

研究了采用由视频噪声过门限形成的随机脉位脉冲的平均脉冲密度、脉冲平均宽度、平均脉冲重复周期、脉冲相关性, 以及随机脉位线性调频脉冲信号的模糊函数等, 由分析可知, 随机脉位线性调频脉冲信号既具有雷达信号的特点, 可以作为雷达信号使用, 又具有宽带噪声的遮盖特性, 可以作为电子攻击信号使用, 因此该信号具有作为未来一体化系统使用的共用信号的潜力。

参考文献:

- [1] 徐崔春. 雷达与干扰机一体化中信号共享的概念与波形设计[D]. 成都: 电子科技大学, 2002.
- [2] 宋耀良. 随机多元码脉位调制和脉间二元序列相位调制的混合波形设计[D]. 电子学报, 1994, 22(3): 23 - 27.
- [3] 刘隆和, 任献彬. 随机脉位调制雷达的抗干扰性能分析[D]. 现代雷达, 1996, 16(3): 97 - 103.
- [4] Shen Ying, Shang Weihua, Liu Guosui. Ambiguity function of chaotic phase modulated radar signals Signal Processing Proceedings[C]. 1998. ICSP 98, 1998 Fourth International Conference on, 1998, 2: 1574 - 1577.
- [5] G Wetzker, F Jondral. Ambiguity function of periodic pulse shaped sequences[J]. Electronics Letters, 1999, Vol. 35(8): 611 - 613.

[作者简介]



万明杰(1966 -), 男(汉族), 河南郑州人, 防空兵指挥学院副教授, 空军工程大学导弹学院博士后, 主要从事电子对抗、防空作战建模与仿真研究。

林志远(1969 -), 男(汉族), 河南新蔡人, 电子科技大学博士后, 主要从事雷达信号处理、电子对抗技术研究。

李涛(1978 -), 男(汉族), 河南新野人, 防空兵指挥学院讲师, 主要从事防空导弹工程技术研究。