基于格点场数据的沙尘暴双预报模型

王 萍[1]，刘 颖[1]，王汉芝[1]，刘环珠[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 100080）

摘 要：为提高沙尘暴预报的准确率，以描述大气环流形式的物理场格点数据作为建模样本，采用自组织神经网络对物理格点场数据样本进行聚类，构建出由大规模阵列式数据格式表示的建模样本的低维特征，再用模糊神经网络综合建模样本的一般性规律，用非典型样本进行二次建模以反映建模样本的特殊性，并设计隶属度调整方案对一般性和特殊性进行协调，由此形成兼顾建模样本一般性和特殊性的双预报模型. 测试结果表明，基于特征提取方案的双预报模型体系使沙尘暴预报准确率达到80.4%.

关键词：建模样本特征；沙尘暴；预报模型；神经网络
中图分类号：TP183；TP273.2　　**文献标志码**：A　　**文章编号**：0493-2137(2006)03-0329-05

Dual Forecasting Model of Sand-Dust Storm Based on the Lattice Position Field Data

WANG Ping[1], LIU Ying[1], WANG Han-zhi[1], LIU Huan-zhu[2]
(1. School of Electrical and Automation Engineering, Tianjin University, Tianjin 300072, China;
2. National Meteorological Center, Beijing 100080, China)

Abstract: To improve the forecasting precise ratio of the sand-dust storm, the physical lattice position field data, which is used to describe atmosphere circumfluence, constitutes modeling sample. The physical lattice position field data are clustered by self-organizing neural network. Based on these processed data, the low-dimensional features for modeling samples, which is depicted by large-scale point lattice data, are extracted successfully. A fuzzy neural network is trained to embody the generic rules hidden in modeling samples. Sequentially based on non-typical samples, further model is constructed to reflect their particularity. Then a membership degree adjusting formula is proposed to blend the universality with the particularity. Thus the dual forecasting model comes into being, which includes the generic rules and the particularity of modeling samples. The experimental results show that the forecasting precise ratio of the sand-dust storm predicted by the proposed dual forecasting model is 80.4%.

Keywords: feature for modeling sample; sand-dust storm; forecasting model; neural network

用于天气预报的格点场数据，以二维阵列的格式记载高空物理场的分布，它的每一数据点称为格点，代表确定的地理位置. 一般每纵向或横向相邻格点间取相等距离（如经度2.5°或纬度2.5°）[1]. 其中，物理场涉及高度场、风场（南北向和东西向）和位温场等.

经研究表明，沙尘暴天气的形成除了与地表存在裸露的沙尘源有关以外，还与高度场、风场和位温场等所形成的特定的大气环流形式和天气系统密切相关[2]. 沙尘暴和非沙尘暴是两种不同类型的天气，它们在特定的区域和季节，具有不同的相关物理场分布规律. 笔者试图利用模式识别建模技术，在近20年的大气物理格点场数据中寻找这种类间差异的规律，并通过训练（学习）将其归纳于预报模型之中. 所选资料包括发生在东经70°～115°、北纬35°～55°地域内的1981年开始的17年间（2月11日～6月10日）的572个沙尘暴日（其中包括242个强沙尘暴日）和1 469个

收稿日期：2004-12-30；**修回日期**：2005-04-08.
作者简介：王 萍（1955— ），女，教授，wangps@tju.edu.cn.

非沙尘暴日. 对每一份资料来讲,按照常规的数据密度(相隔)计算,由值相似高度场、形相似高度场、形相似风场以及形相似位温场组成的数据规模将达到855个,显然,直接使用具有如此数据规模的样本进行统计建模是不可能的,必须合理、客观地抽取反映沙尘暴和非沙尘暴的样本特征.

1 用统计分析方法形成建模样本特征

1.1 基本思想

据专家经验,沙尘暴日的高度场在数值分布和形态分布上以及风场和位温场在形态分布上有几种典型的类型. 假设这些类型分布在242个强沙尘暴样本之中,首先将该样本集做子类划分,并求出各子类物理均值场,设子类数为 n,则物理均值场有 $4n$ 个. 考虑到每4种物理场联合描述一个沙尘暴子类,于是以每4种物理均值场为度量基准,在所有的历史样本中筛选出沙尘暴子类样本和相应的非沙尘暴样本子集,用它们组织 $k-l$ 变换(n 个)[3],根据变换结果构造出 n 个综合因子,作为描述沙尘暴建模样本的 n 个特征.

1.2 沙尘暴样本子类划分

据专家经验,设值相似高度场、形相似高度场、形相似风场以及形相似位温场的类别数分别为2、3、2、2,它们将形成24种不同的组合,按照以上类别数的设定,用自组织神经网络对242个强沙尘暴日进行聚类[4—5],结果发现沙尘暴日基本集中在其中的10个组合之中,选出这10个组合,并称它们为沙尘暴子类.

1.3 样本特征的形成

上述10个子类在各个场的分布上,特别是在每4种物理场的关联关系上各不相同,属于沙尘暴的10个典型类型,利用主成分分析方法分别针对10个子类构造出10个综合特征,并应将其作为建模样本的特征进行描述.

(1)计算10个沙尘暴子类的 10×4 个均值场 $C(k,q)=[c_{ij}(k,q)]$,其中 $k=1,2,\cdots,10$, $q=1,2,3,4$, $c_{ij}(k,q)=\frac{1}{n_k}\sum x_{ij}(k,q)$, $x_{ij}(k,q)$ 表示位于第 k 个子类中第 q 个物理场格点 (i,j) 上的数据, n_k 表示第 k 个子类的样本数.

(2)计算 k 类样本与均值场的距离 $d(k,q)$ 和方差 $\sigma(k,q)$,记最大距离为 $d_{\max}(k,q)$,并由此设定距离阈值 $\mu(k,q)$.

(3)计算样本 x 的物理场与 k 类物理均值场的距离 $d_x(k,q)$. 对沙尘暴样本 x 或对非沙尘暴样本 x,若 $d_x(k,q)<\mu(k,q)$,则 x 被选中. 设共选中样本数 m_k.

(4)将每个选中样本的4个物理场距离 $d_x(k,1)$、$d_x(k,2)$、$d_x(k,3)$ 和 $d_x(k,4)$ 作为该样本特征,记作 $Z_1^{(i)},\cdots,Z_4^{(i)}$, $i=1,\cdots,m_k$,并用 m_k 个样本进行 k-l 变换[5],得到4个主成分 $y_1(k),\cdots,y_4(k)$.

经过对各 k 值下的第1主成分 $y_1(k)$ 的统计计算得知,它们所能代表的分类信息在33.5%~47.6%之间. 为兼顾其余主成分作用而又不至增加复杂性,将各子类的4个主成分按照如下方案进行综合,得

$$y(k)=\frac{\lambda_1(k)}{\lambda_{\text{sum}}(k)}y_1(k)+\frac{\lambda_2(k)}{\lambda_{\text{sum}}(k)}y_2(k)+\frac{\lambda_3(k)}{\lambda_{\text{sum}}(k)}y_3(k)+\frac{\lambda_4(k)}{\lambda_{\text{sum}}(k)}y_4(k) \tag{1}$$

式中 $\lambda(k)=\sigma[y(k)]$ 为用方差 σ 表示的主成分分量 $y(k)$ 所代表的分类信息,分类信息总量为

$$\lambda_{\text{sum}}(k)=\sum_{j=1}^{4}\lambda_j(k)=\sum_{j=1}^{4}\sigma[y_{ij}(k)] \tag{2}$$

则称 $y(k)$ 为第 k 个沙尘暴子类的总量特征.

1.4 样本特征的统计检验及分析

首先,为检验在第1主成分下和在总量特征下沙尘暴与非沙尘暴的类间显著性差异,组织了方差检验. 即在显著性水平 $\alpha=0.01$ 的前提下,分别算出10个第1主成分和10个总量特征的值,如表1所示. 不难看出,与第1主成分相比,总量特征仅在第5个分量上使分类信息模糊化,但在第1、第6和第8个分量上分类信息被突显出来,使显著性特征数从4个增加到了6个,沙尘暴和非沙尘暴的差异总体上比较显著.

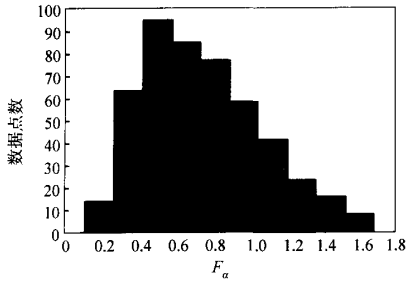
表1 第1主成分和总量特征的方差检验

Tab.1 Variance test of the first principal component and the synthesis features

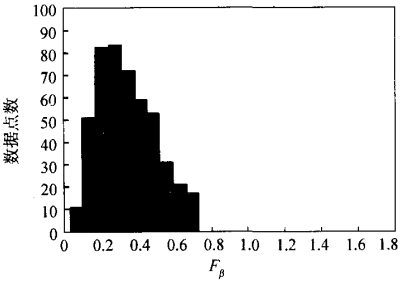
特征号 k	第1主成分 F_α	总量特征 F_β
1	4.9[1)]	14.3
2	0.7[1)]	3.3[1)]
3	17.4	10.5
4	1.2[1)]	0.2[1)]
5	48.0	2.4[1)]
6	6.4[1)]	69.8
7	21.6	14.5
8	0.2[1)]	22.1
9	73.9	175.2
10	4.8[1)]	0.0[1)]

注:1)不具有显著差异.

另外,总量特征下的样本分布更集中,即类内聚类性更强. 图1给出了它们的一个子类的示例.



(a) 第 1 主成分



(b) 总量特征

图 1　样本特征下沙尘暴样本的聚集性

FIg. 1　Compactness of sand-dust storm samples using different features

2　总量特征描述下的神经网络模型的建立

2.1　拓扑结构及模糊 BP 算法

为避免因过多隐层及过多神经元而引发过拟合现象. 经过多次试验比较,确定整个神经网络由输入层、两个隐层和输出层组成,总节点数为 10 × 25 × 15 × 1. 算法选用模糊权的误差反向传播算法,输入、输出、权系数和阈值均选用三角模糊量[6].

2.2　对训练样本的编辑

针对沙尘暴预报问题具有小概率和两类样本量不均衡的特点,设计了 3 种样本编辑方案,并相对同样的学习率和惯性系数进行训练,再用剩余样本做了检测. 表 2 展现了 3 种样本编辑[7—8]方案下的检测结果.

从表 2 中看出,若以 1981 ~ 1987 年的样本作为训练样本,循环次数少,且沙尘暴(正例) 和非沙尘暴(反例)的预报率均衡,预报结果相对较好. 1981 ~ 1985 年学习样本较少,代表性差,神经网会漏掉对测试样本中的一些模式的学习,因而使识别率降低. 而 1981 ~ 1989 年虽然囊括了较多的样本,但由于沙尘暴和非沙尘暴数量极不均衡造成沙尘暴的预报率较低. 从样本的代表性和样本数量的均衡性两个方面考虑,对训练样本的筛选,认为选择 1981 ~ 1987 年的样本为训练样本比较合适.

表 2　几种样本编辑方案下的预报结果比较

Fig. 2　Forecasting results under several schemes edting samples

训练样本	训练次数	测试样本	沙尘暴报对率/%	非沙尘暴报对率/%	总报对率/%
1981 ~ 1985 沙 257 个 非 344 个	900	1986 ~ 1997	67.3	69.0	68.6
		1988 ~ 1997	68.7	68.9	68.9
		1990 ~ 1997	69.1	69.6	69.5
1981 ~ 1987 沙 342 个 非 499 个	600	1988 ~ 1997	73.0	71.1	71.4
		1990 ~ 1997	73.0	71.2	71.4
1981 ~ 1989 沙 397 个 非 685 个	2 100	1990 ~ 1997	43.8	85.3	77.7

2.3　参数调整

在采用模糊权的神经网络中,其输出与权系数都是三角模糊量. 权值有 w_l 和 w_u 之分,相应的,学习率有 η_l 和 η_u 之分,惯性系数有 β_l 和 β_u 之分. 研究发现对误差反向传播算法中的参数 η 和 β 的合理调整,会提高网络模型的工作品质,使预报结果有所改善(见表 3). 由表 3 可知,将 1981 ~ 1987 年样本用于训练,将 1988 ~ 1997 年样本用于试报,在学习率 $\eta_u = 0.5$, $\eta_l = 0.3$、惯性系数 $\beta_u = 0.08$, $\beta_l = 0.04$ 下训练权值 w_l 和 w_u,得到的训练模型预报结果相对较好.

表 3　惯性系数与学习率的调整与试报结果

Tab. 3　Forecasting results under different inertial coefficients and learning rates

%

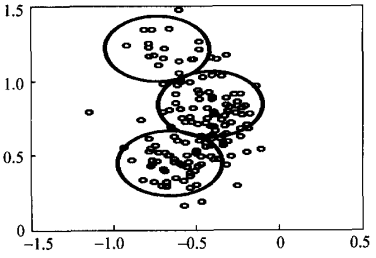
参数	$\eta_u = 0.5, \eta_l = 0.3$ $\beta_u = 0.08, \beta_l = 0.04$	$\eta_l = \eta_u = 0.5$ $\beta_l = \beta_u = 0.08$	$\eta_l = \eta_u = 0.3$ $\beta_l = \beta_u = 0.04$
沙尘暴报对	73.0	61.4	69.5
非沙尘暴报对	71.1	78.3	68.4

客观地讲,以上神经网络仍然对相当数量的样本不能很好地识别. 对判错样本做进一步的分析,发现它们大多表现为:有一个或几个物理场与本类物理场相偏离,属于非典型的沙尘暴或非沙尘暴样本,在这里统称为非典型建模样本. 很显然,以上网络不具备正确识别非典型样本的能力,为此应特别进行再次建模.

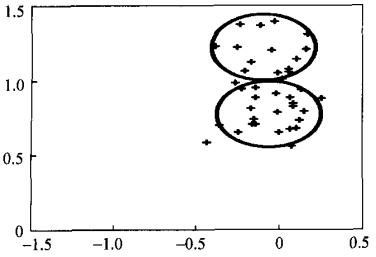
3 基于非典型样本的再次建模

3.1 非典型样本的界定和聚类

非典型样本，即以上方法训练得到的模糊神经网报错的样本. 将 1988 ~ 1997 年间前 5 年的报错样本选作参考样本，后 5 年的报错样本选作检测样本. 首先，对参考样本进行聚类分析，结果可见，非沙尘暴参考样本被聚为 3 类；沙尘暴参考样本聚为 2 类. 将这 5 个子类视为非典型样本群. 图 2 给出了降维处理后的非典型样本群的分布及其聚类示意.



(a) 非沙尘暴样本



(b) 沙尘暴样本

图 2 非典型样本的分布及聚类

Fig. 2 Distribution of non-typical samples and their clustering

3.2 样本的非典型性判定

样本聚类后，可算出各非典型样本子集的聚类中心，子集中的任一样本与其对应的非典型类中心的距离反映了该样本的非典型程度，同时也可通过距离计算来判别某一样本是否落入非典型样本区. 由于各非典型样本区最大距离 $d^i_{\max}$ 不同，故用相对距离 γ^i（$\gamma^i = d^i/d^i_{\max}$，$d^i$ 为样本与第 i 个非典型区中心的距离）作为判断样本是否落入非典型区的条件.

(1) 计算待调整样本离 5 个非典型区的相对距离 $\gamma^i, i = 1, \cdots, 5$，并取其中的最小值 $\gamma_{\min}$，相应的非典型区称为该样本最近非典型区.

(2) 根据 $\gamma_{\min}$ 的大小判断该样本是否落入最近非典型区.

3.3 隶属度的调整策略及实现

将经以上模糊神经网判定后的样本做非典型性考察，会出现 2 种可能：①判定为沙尘暴或非沙尘暴，又未落入非典型区；②落入非典型区.

为参照样本的非典型性对其类别归属进行调整，将样本隶属度调整公式设计为 2 项之和，即

$$\mu = a\mu_1 + b\mu_2 \tag{3}$$

式中：μ_1 为模糊神经网络给出的样本隶属度；μ_2 为反应样本的非典型程度；a 和 b 为侧重度系数.

3.3.1 调整侧重度系数

根据样本 x 是否落入非典型区，选用不同的侧重度系数. 当样本落入非典型区时，适当增加对样本非典型程度的侧重度，取 $a = 0.7, b = 0.3$. 反之，若未落入非典型区，适当减小对样本非典型程度的侧重度，取 $a = 0.9, b = 0.1$.

3.3.2 构造函数，使其随样本的非典型程度单调变化

当 x 落入沙尘暴的非典型区时，$\mu_2 \geqslant 0.5$ 且随 x 的非典型程度单调递增；当 x 落入非沙尘暴的非典型区时，$\mu_2 \leqslant 0.5$ 且随 x 的非典型程度单调递减.

设某非典型区的相对距离均值为

$$u = \frac{1}{n}\sum_{k=1}^{n}(d_k/d_{\max}) \tag{4}$$

且 x 落入该区，用 x 与 u 的绝对差描述其非典型程度（差值越小，非典型程度越强），将 μ_2 构造为

$$\mu_2 = \begin{cases} \dfrac{1}{2}(1 + \mathrm{e}^{-|\gamma_{\min} - u|}) & x \text{ 落入沙尘暴非典型区} \\ \dfrac{1}{2}\mathrm{e}^{-|\gamma_{\min} - u|} & x \text{ 落入非沙尘暴非典型区} \end{cases} \tag{5}$$

容易验证，式(5)符合对 μ_2 的要求. 称上述非典型样本群及隶属度调整方案为隶属模型，它与前述模糊神经网络模型合称为双预报模型.

3.4 对隶属模型的测试

对 1993 ~ 1997 年的样本运用模糊神经网络进行判定，再通过非典型样本群获知其非典型程度，利用式(3)进行隶属度调整，结果，进入非典型区和未进入非典型区的报对、报错样本数量均有所变化（见表 4）. 表 4 的数据显示，在双预报模型下，有 25.8%的模糊神经网的报错样本得以纠正，而报对样本的出错率只有 0.67%. 从而使沙尘暴预报的总报对率从单模型（仅神经网）时的 74.3%升至双模型时的 80.4%.

3.5 双模型预报的框架

本沙尘暴双预报模型系指用 7 年（1981 ~ 1987）的格点场数据训练而得的模糊神经网络和经该网络判断的 5 年（1988 ~ 1992）样本建立的非典型样本群及隶属度调整方案. 用它们形成的沙尘暴双预报模型框架如图 3 所示.

表 4　双预报模型与单预报模型报对率比较

Tab. 4　Comparision of the right ratio between the dual and the single forecasting model

落入/未落入非典型区	报错样本			报对样本		
	模糊神经网单模型报错数	双模型对单模型纠正数	纠正率/%	模糊神经网单模型报对数	双模型的再出错数	出错率/%
落入	69	27	39.13	169	1	0.59
未落入	86	13	15.12	279	2	0.72
合计	155	40	25.81	448	3	0.67

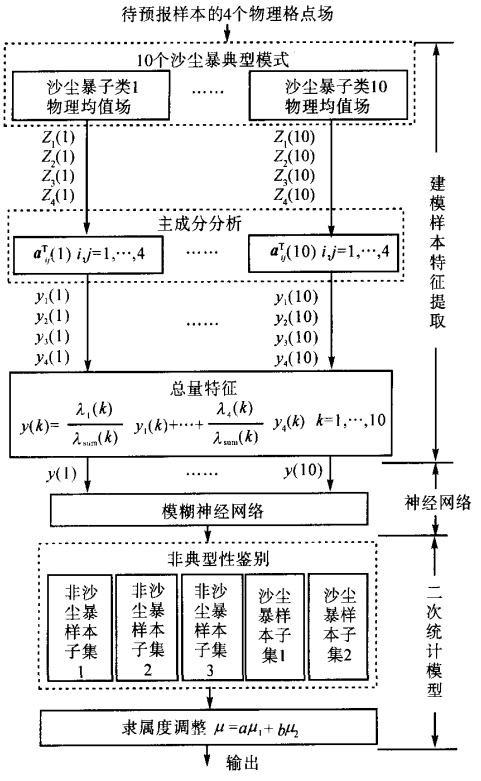


图 3　沙尘暴双预报模型框架

Fig. 3　Frame of dual forecasting model of sand-dust storm

4　结　语

用神经网络综合建模以反映样本中的一般性规律和再次建模以兼顾建模样本中的特殊性,由此形成的双预报模型可有效提高对沙尘暴的预报准确率.

另外,格点场数据格式与图像数据格式类似,均以二维数组的形式描述样本,在聚类分析、相似性计算及主成分分析的基础上构建总量特征的方法,将大规模阵列式数据样本转化为低维特征向量描述,是一种以建模为目的,并将大规模数据阵转换为特征描述的有效方法.

参考文献:

[1] 罗　坚,黄　峰. 基于统计模型的气象数据无损压缩新方法[J]. 地球科学进展,2003,18(4):637—642.
Luo Jian, Huang Feng. A new scatheless compression encoding scheme for meteorological grid data based on statistical model [J]. *Advance in Earth Sciences*, 2003, 18(4): 637—642(in Chinese).

[2] 王式功,董光荣. 沙尘暴研究的进展[J]. 中国沙漠,2000,20(4):349—356.
Wang Shigong, Dong Guangrong. Advances in studying sand-dust storms of china [J]. *Journal of Desert Research*, 2000, 20(4): 349—356(in Chinese).

[3] 蒋惠园,王餐香. 主成分分析法在综合评价中的应用[J]. 武汉理工大学学报:交通科学与工程版,2004,28(4):467—470.
Jiang Huiyuan, Wang Wanxiang. Synthetic appraisal for multi-objects decision-making [J]. *Journal of Wuhan University of Technology: Transportation Science and Engineering*, 2004, 28(4): 467—470(in Chinese).

[4] Wu Zheng, Yen G G. A SOM projection technique with the growing structure for visualizing high-dimensional data [J]. *International Journal of Neural Systems*, 2003, 13(5): 353—365.

[5] Tsai Chengfa, Wu Hanchang. A new data clustering approach for data mining in large databases[C]// *International Symposium Parallel Architectures, Algorithms and Network*. Manila, the Philippines, 2002: 278—283.

[6] Lei Hongli, Zhang Jianbang, Zhang Dianzhi. A new algorithm of fuzzy neural networks with multiple form fuzzy weights [C]// *The 4th World Congress on Intelligent Control and Automation, Intelligent Control and Automation*. Shanghai, China, 2002, 4: 3252—3255.

[7] 张秀玲. 神经网络自适应控制的研究进展及展望[J]. 工业仪表与自动化装置,2002(1):10—14.
Zhang Xiuling. The advancement of vista of a network adaptive control [J]. *Industrial Instrumentation and Automation*, 2002(1): 10—14(in Chinese).

[8] 乔志骏,刘其真,易维列,等. 一个基于模糊社经网络的数据逼近和泛化建模方法[J]. 模式识别与人工智能,2001,14(2):253—256.
Qiao Zhijun, Liu Qizhen, Yi Weilie, et al. An approach for modeling of approximation and generalization based on fuzzy neural network [J]. *Pattern Recognition and Artificial Intelligence*, 2001, 14(2): 253—256(in Chinese).