天津大学

硕士学位论文

基于模糊神经网的沙尘暴预报模型

姓名：王汉芝

申请学位级别：硕士

专业：模式识别与智能系统

指导教师：王萍;林孔元

20031201

摘要

目前，人工智能和模式识别技术已经在各个具体领域得到广泛的应用。本文所做的研究工作，正是围绕着“样本特征提取和沙尘暴的预报”这一具体问题，利用人工智能和模式识别技术展开研究的。

首先，从沙尘暴和非沙尘暴样本在四个基本物理场上反映出的特点出发，选出用于建模的代表性样本，通过对主成分分析结果的研究，设计出特征综合方案，形成兼顾各个主成分的“总量特征”。利用这种特征综合方法，实现了对40维样本进行再次的特征提取，最终形成10维更为合理的建模样本。然后，展开基于模糊神经网络的沙尘暴建模方案研究，并从网络的拓扑、训练参数、样本集合等方面对模型进行优化，使对沙尘暴的预报达到了一定的预报效果。

本文对基于模糊神经网的预报结果展开进一步研究，指出影响预报准确率的主要因素是训练样本中含有为数较多的非典型性样本。于是，通过聚类建立非典型样本区，再构建基于非典型样本的统计模型，并设计出一种兼顾模糊神经网预报结果和样本非典型程度的沙尘暴隶属度调整方案，使建立在模糊神经网预报（1级）结果之上的统计模型再预报（2级），在基本不影响1级报对率的前提下，纠正了相当比例的报错样本，与文献[5]的基于40个特征的神经网络相比，本文提出的模糊神经网与统计模型的联合预报方案，使沙尘暴的报对率从60%提高到73.3%以上，*CSI*值也由25.9%提高到38.7%，预报效果得到明显改善。

最后对建立预报系统中遇到的问题作了总结，并提出了特征重构、集成神经网络、神经网络与专家系统结合的解决方案，同时就沙尘暴的动态预报问题提出了实现框架。

关键词：特征分析 模糊权 模糊神经网 统计建模

Abstract

At present, the technology of artificial intelligence and pattern recognizing has been widely applied in various fields. The implementation of extracting feature and forecasting sand-dust storm, which use the technology of artificial intelligence and pattern recognize is presented in this paper.

At first, starting from the character of sand-dust storm and non-sand-dust storm samples reflected on four physical fields, we choose representative samples used in modeling. By studying the result of principal component analysis (PCA), a scheme to integrate features is designed. "gross features" which give attention to all the principal components are formed hereby. Thus, 40 dimensions features of samples are extracted again by the method of integrating features, 10 dimensions features of modeling samples are formed at last, which are more reasonable than ever. Second, the sand-dust storm forecasting model based on fuzzy nerve net (FNN) is researched in succession. By means of model optimized in several aspects such as FNN topology, parameters and compilation of the samples, a more reasonable forecasting result is acquired by the FNN model.

It is pointed that the primary factor in influencing the right forecast ratio is lots of non-typical samples among training samples by researching the forecasting result based on FNN again. Then, by clustering, the region of non-typical samples is founded, so does the statistical modeling based on non-typical samples. A method to adjust the degree subjecting to sand-dust storm has been designed at the same time, which gives attention to the forecasting result of FNN and the degree of non-type. Then samples are forecasted again using statistical model (second stage) based on the FNN model (first stage), some samples which are forecasted wrong at first stage are corrected at second stage, and the statistical model has little effect on samples which are forecasted right at first stage. Compared with the forecasting result of NN based on 40 dimensions from literature [5], the right ratio of sand-dust storm is improved from 60% to 73.3%, *CSI* is improved from 25.9% to 38.7% by the model combined FNN with statistical model pointed in this paper, the result is better than ever.

At last, some problems encountered in constructing forecast system are summarized in the paper. Some resolutions to these problems such as reconstructing features, integration NN, combining expert system and NN are presented at the same time, as well as the realizing frame of dynamical forecasting system.

Key Words: Analyzing Feature Fuzzy Weight FNN Statistical Modeling

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 天津大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：王汉芝 签字日期：2003 年 12 月 20 日

学位论文版权使用授权书

本学位论文作者完全了解 天津大学 有关保留、使用学位论文的规定。特授权 天津大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：王汉芝 导师签名：王春 林

签字日期：2003 年 12 月 20 日 签字日期：03 年 12月 20 日

第一章 绪论

1.1 沙尘暴

1.1.1 沙尘暴的危害

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]。沙尘暴，特别是特强沙尘暴是一种危害极大的灾害性天气。当其形成后会以排山倒海之势滚滚向前移动，携带沙粒的强劲气流所经之处，通过沙埋、风蚀沙割、狂风袭击、降温霜冻和污染大气等作用方式，使大片农田或受沙埋、或遭风蚀刮走沃土，或者农作物受霜冻之害。它能加剧土地沙漠化，对大气环境造成严重污染，对生态环境造成巨大破坏，对交通和供电线路产生重要影响，给人民生命财产造成严重损失[2]。

沙尘暴多发生在北方地区的冬春季节，我国西北许多地区人口过快增长，资源过度开发利用，生态环境急剧恶化，土壤沙化、水土流失日益严重，局部地区已到了十分严重的程度，这就为沙尘暴的发生提供了丰富的物质资源。而频繁发生的沙尘暴对于下游的广大地区来说是一个严重的污染源，沙尘顺风飘溢，从北到沈阳、北京、天津，南到上海和武汉的很多大城市都深受其害。

1.1.2 沙尘暴的成因

专家研究指出，沙尘暴形成有 3 个基本条件，一是大风，这是形成沙尘暴的动力条件；二是地面上裸露的沙尘物质，它是沙尘暴的物质基础；三是不稳定的空气状态，这是重要的热力条件。即强风因子、沙源因子和热力不稳定因子是沙尘暴产生的宏观条件。在宏观条件满足的前提下，还需有利的环流形势和天气系统相配合。易产生沙尘暴的主要环流形势和天气系统有：经纬向环流调整、冷锋活动、低空东风急流、中尺度系统等。总之，强沙尘暴的发生、发展是在特定的地形条件、沙尘源条件和各种不同尺度天气条件下共同作用的结果。[2][3][4]

1.1.3 沙尘暴的预报

减轻沙尘暴的危害，首先对荒漠化土地进行治理，在退耕还林还草的同时，大幅度降低草场载畜量，并适当采取人工措施，恢复天然草原植被，就可以有

效地抑制主要沙尘来源，形成绿色生态屏障。而大面积恢复林草植被也需要较长时间。故还需建立和完善沙尘天气的动态监测、预警系统，做好防灾减灾的科学研究，以降低强沙尘天气造成的损失。

目前，针对中国沙尘天气的特点，主要应用气象卫星、高空探测、地面自动气象站等多种探测与监测手段，实现对沙尘暴天气发生、发展和传输过程的跟踪和定量监测。对沙尘暴的预报主要基于传统的天气学分析和个人对数值预报产品的理解并凭借预报员的经验来完成，现在急需发展沙尘暴天气预报的新方法，建立具有较高业务应用价值的沙尘暴监测、预警系统。并及时发布信息，以利于提前安排好生产、交通和群众生活，尽可能减少损失。

在对沙尘暴进行预报时，必须综合考虑形成沙尘暴的三个因子及其相互作用。沙源因子反映地理条件，相对来说比较固定，主要由冷空气路径来决定，只要冷空气经过的路径下面垫有丰富的沙源，就要看另外两个因子的表现如何。而对于冷空气路径、强风和空气稳定性目前气象部门都是可观测的，并且能达到一定的预测水平，这就使得沙尘暴的预报具有了可能。

1.2 预报模型建模基础[5]

沙尘暴样本的特征提取是建立沙尘暴预报模型的前提。文献[5]从数据源特点出发，根据专家经验依次通过聚类分析、建立典型模式类、计算中心场，再以样本与中心场的距离作为样本的特征。在每个样本的 855 个数据中提取到 40 个特征。建立和优化了基于人工神经网的沙尘暴预报模型。虽然预报效果不很理想，但 40 维的特征实现了建模的可能，神经网的预报模型则为进一步建模提供了借鉴。

1.2.1 沙尘暴样本

一、格点场数据源

美国环境气象中心每天所提供的资料（NCEP）是我国数值预报的重要数据源，内含的丰富气象信息有待进一步开发和利用。

鉴于我国沙尘暴发生的时间和地区特点[6][7]，将 NCEP 资料圈定在我国西北部从初春到初夏的范围之内，详见表 1-1。

按照 NCEP 资料每 2.5 度记一格的数据格式，可知表 1-1 给定的选定区域，东西向跨 45°分为 18 格、南北向跨 20°分为 8 格。东西向和南北向的交叉记为一个格点，这样，一个格点场的数据量就是 19*9=171 个，即每个格点场提供 19×9 的数据阵。考虑到沙尘暴形成的基本条件，选取 500hpa 的高度、700hpa 的东南

风和西北风、850hpa 的温度和比湿共五种格点场作原始数据源。然后，用 850hpa 的温度和比湿推导出 850hpa 的位温，再将东南风和西北风合并在一起，最终得到反映沙尘暴信息的三种物理场，即 500hpa 的高度(*Hgt*)场、700hpa 的两个风(*UV*)场和 850hpa 的位温(θ_{se})场。

表 1-1 NCEP 资料范围

	范围	说明	沙尘暴日频数
时间	1981~1997 年 2 月 11 日~6 月 10 日	17 年初春至初夏	$\frac{\text{沙尘暴日}}{\text{非沙尘暴日}} = \frac{575}{1469} = 39.14\%$
地域	东经 70°~115° 跨度 45°	俄罗斯、蒙古；新疆、内蒙古、甘肃、宁夏、青海、山西、陕西、河北、北京、天津、山东等省市	
	北纬 35°~55° 跨度 20°		

二、沙尘暴样本的数据规模

距平、值相似和形相似是气象上处理数据和度量格点场数据相似度的常用办法，其中，距平是求取观测值对平均值的差值，值相似是利用原始数值比较样本间的相似程度，形相似是利用距平值比较样本间的相似程度，它能够直观地反映格点场形状的相似程度。

根据专家经验，分别对高度场数据进行值相似和形相似度量、对两个风场和位温场进行形相似度量，于是每个样本的数据规模高达 171*2+342+171=855。若简单地将样本数据数量等同于样本维数，沙尘暴样本数据维数则高达 855。

三、建立沙尘暴子模式

首先，利用自组织特征映射网络聚类方法对 1981 年—1997 年 242 个强沙尘暴日 (出现沙尘暴的站点数目为 9 以上的沙尘暴日) 样本进行聚类。根据专家经验，从值相似角度，高度场可分为两类；从形相似角度，高度场可分为三类、风场分为两类、位温场分为两类。

这样，242 个样本被聚为 2*3*2*2=24 种类型，聚类结果列于表 1-2 中。表中类别号的四位数字中，第一位到第四位依次表示高度场值相似的类型、高度场形相似的类型、风场形相似的类型和位温场形相似的类型。例如 1211 表示其样本属于 *Hgt* 值相似的第 2 类，*Hgt* 形相似的第 3 类，*UV* 形相似的第 2 类和θ_{se}形相似的第 2 类。

沙尘暴有强沙尘暴和少站点沙尘暴（出现沙尘暴的站点数目为 10 以下）之分，在强沙尘暴中，又将站点数较多的沙尘暴界定为严重沙尘暴，据各子类中聚集的样本数和沙尘暴严重程度，大体有样本数为 0、样本数少、样本数多以及样本数据量相对较多且含有特别严重的沙尘暴类型几种情况。

1．样本数较多的子类客观上反映了三种物理场各自特定格局的组合与沙尘暴天气的联系，应该作为典型的沙尘暴模式，如子类 1 和子类 10。

2．而那些样本数相对较多且包含着历史上特别严重的沙尘暴天气的子类，其物理场组合格局不容忽略，也作为典型的沙尘暴模式，如类别 0010 样本数（9 个）虽少于类别 0211（12 个），但其中包括站点数为 63 的严重沙尘暴样本，故选择为子类 2。

由此从 24 个聚类结果中筛选出 10 个沙尘暴子类（模式）如表 1-2 所示。

表 1-2 聚类结果

类别	样本数	站点数	备注	类别	样本数	站点数	备注
0000	35	10~80	子类 1	1000	16	10~58	子类 6
0001	4	13~55		1001	8	10~27	
0010	9	11~63	子类 2	1010	20	10~57	子类 7
0011	0			1011	1	18	
0100	8	10~35		1100	0		
0101	17	11~67	子类 3	1101	11	10~24	子类 8
0110	6	10~33		1110	15	10~37	子类 9
0111	0			1111	9	10~15	
0200	0			1200	0		
0201	15	10~25	子类 4	1201	1	18	
0210	17	10~41	子类 5	1210	1	10	
0211	12	10~21		1211	37	10~49	子类 10

四、提取沙尘暴模式特征

借助沙尘暴的 10 个子模式，可以建立起这样的概念，一种子模式下三种物理场值的特征及形的特征能够表征一类沙尘暴天气，10 个模式下全部物理场值的特征和形的特征联合起来可以鉴别沙尘暴和非沙尘暴样本。于是，首先计算 10 个沙尘暴模式的 *Hgt* 值和形的中心场、*UV* 形的中心场以及 θ_{se} 形的中心场共

计 40 个($C_{ij}, i=1,2,\cdots,10, j=1,2,3,4$)。计算公式如下：

$$c_{ij}(k)=\frac{1}{n_i}\sum_{l=1}^{n_i}x_{ij}^{(l)}(k) \tag{1-1}$$

其中 $i=1,2,\cdots,10;\ j=1,2,3,4;\ k=\begin{cases}1,2,\cdots,342 & \text{风的中心场}\\ 1,2,\cdots 171 & \text{其他中心场}\end{cases}$

n_i— 类i中的样本个数，

$x_{ij}^{(l)}(k)$— 类i中第l个样本的第j个格点阵的第k个格点数据，

$c_{ij}(k)$— 中心场C_{ij}的第k个格点数据。

然后计算样本的 4 个值或形的格点阵与每个子模式（共 10 个）的 4 个中心场的相似度，并将每一个相似度记为该样本的一个特征取值。40 个特征则构成样本的特征向量 Z。

$$Z=(z_{11},z_{12},z_{13},z_{14},\cdots,z_{10,1},z_{10,2},z_{10,3},z_{10,4})$$

$$z_{ij}=\sqrt{\sum_{k=1}^{m}(x_j(k)-c_{ij}(k))^2} \tag{1-2}$$

其中 $i=1,2,\cdots,10;\ j=1,2,3,4;\ m=\begin{cases}342 & \text{风的中心场}\\ 171 & \text{其他中心场}\end{cases}$

$x_j(k)$—样本第j个格点阵的第k个格点数据

1.2.2 基于人工神经网的沙尘暴预报基础

一、沙尘暴建模

确定沙尘暴预报模型建模框架如图 1-1 所示。

建模过程一般先选定一种训练集合和测试集合，确定一种网络拓扑和训练参数作为初始模型，训练完成后，用测试样本集合测试模型预报效果，若结果不好，重新选定样本集合、网络拓扑、训练参数和特征提取方式，再训练、测试，反复试探，直至测试效果满意为止。

选用 BP 网络训练沙尘暴预报模型，以成功界限指数 *CSI* 作为衡量模型质量的指数。

$$CSI=\frac{c_f}{c_f+w_f}\times 100\% \tag{1-3}$$

其中，c_f为正确报出的沙尘暴日数，w_f是漏报与空报数之和。[8]

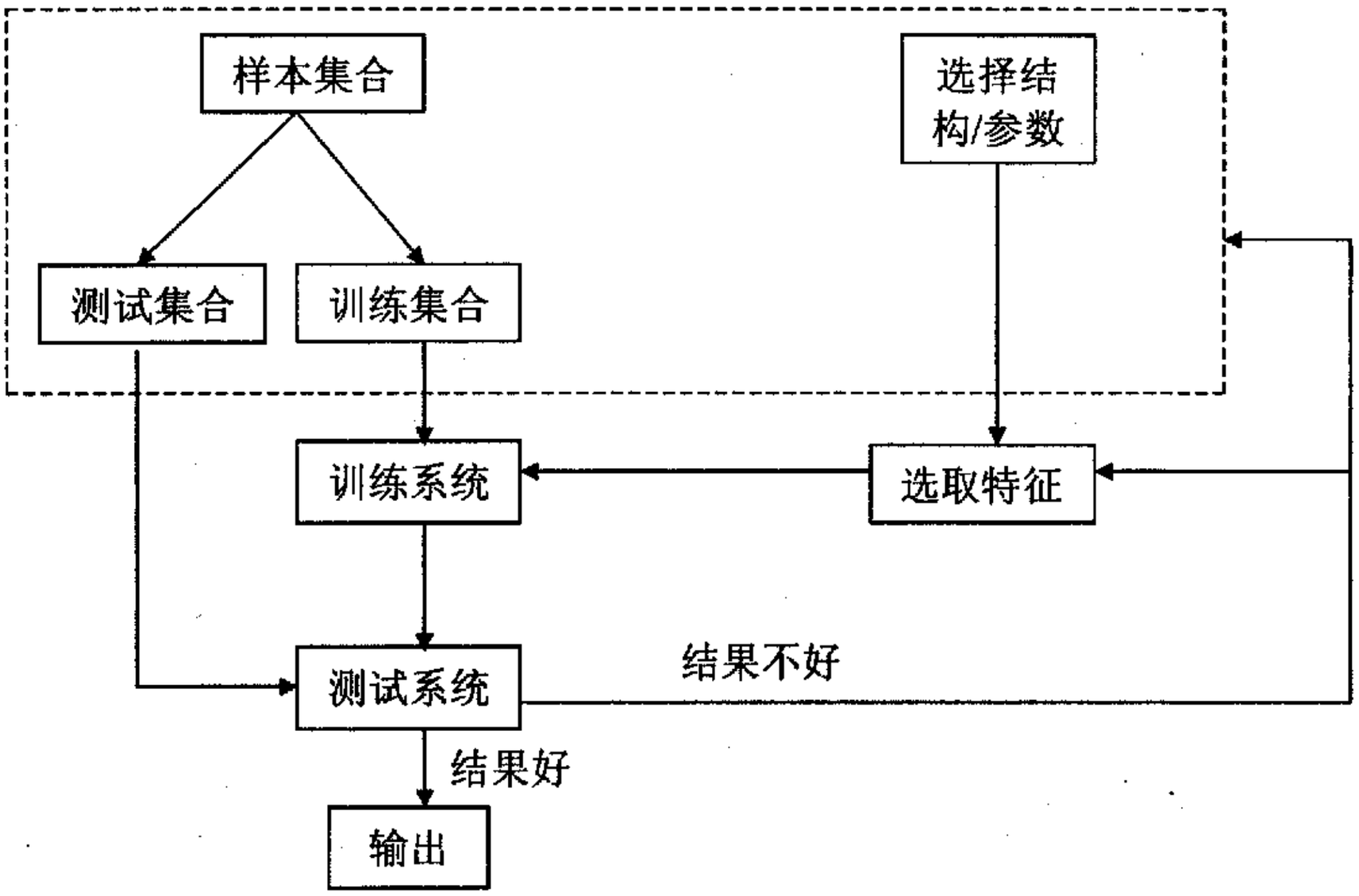


图 1-1 建模过程流程图

二、网络拓扑结构试探

表 1-3 为采用手工试探法试出的各种拓扑结构的试报结果。

其中，训练参数：学习率 0.4，惯性系数 0.02，迭代 1000 次。

训练集合：90~95 年样本集。测试集合：96~97 年样本集。

表 1-3 各种拓扑结构的试报 *CSI* 值

第一隐层节点数	第二隐层节点数	试报 *CSI*(%)
40	20	12.7
30	30	11.1
20	20	21.7
△ 20	10	21.6
20	0	15.1

在上述训练参数和样本集合条件下，40*20*10*1 这种拓扑结构，试报得到的 *CSI* 值较高，是一种较优的拓扑结构。

三、训练样本的编辑

神经网络中记忆的规律，是从已知样本和已知变量组成的数据集中学习训练得到的。很明显，只有当已知样本的数据可靠时，学习到的规律才可能是正确的；只有当选取的样本在表达的规律中具有代表性时，才能得到对预报有用的数学模型。在建模过程中，对训练集合的组织进行尝试，以下为其中两种情况的效果比较。

训练集合 1：90-95 年共 6 年的样本作为训练集合（样本数为 715 个，其中含沙尘暴日样本 153 个，占总数的 21.4%。）。

训练集合 2：81-95 年共 15 年的样本作为训练集合（样本数为 1788 个，其中含沙尘暴日样本有 548 个，占总数的 30.6%）。

测试集合：两种情况都选用 96-97 年共 2 年的样本作为测试集合（共有样本数 239 个，其中含沙尘暴日样本 25 个，占总数的 10.5%）。

训练参数：两网络均为学习率 0.4，惯性常数 0.02，迭代 1000 次。

表 1-4 各种训练样本试报结果

训练集合年限	样本个数	网络拓扑	拟和 *CSI*(%)	试报 *CSI*(%)
90-95 年	715	40*20*10*1	100	22
△81-95 年	1788	40*20*9*1	68.4	25.9

表 1-4 说明，虽然 90-95 年样本比较接近 96-97 的大气环境，但却未能包含所有沙尘暴类型，神经网对一些类型的沙尘暴缺乏学习机会，故不具有识别能力。适当增加训练集的样本数目，同时也增加了沙尘暴的类型，使模型预报水平提高。

除样本集合编辑、网络拓扑和训练参数的优化有助于预报水平的提高外，积极进行样本的特征提取也是提高沙尘暴预报率的一种方式，提取富含分类信息的样本特征，并建立样本的典型性分析，将更助于预报水平的提高。

1.3 本文的主要工作

本文所做工作是国家气象中心“沙尘天气中短期及短时预报系统”研制工作的一部分，任务是在岳斌所做的工作[5]的基础上，进一步提取合理的特征，并建立合适的沙尘暴预报模型，使沙尘暴预报率达到一定的预报精度。

一、岳斌工作的分析

1．从样本的 855 个数据中提取到 40 维的样本特征，从而使沙尘暴预报模型的建立有了可能，但缺乏对样本特征的分析，对样本特征所含的分类信息及相互间的独立性都属未知。

2．建立了基于人工神经网的沙尘暴预报模型，并从样本、拓扑结构和参数等方面优化预报模型，但对样本集合的编辑缺乏更多的尝试和必要的分析，仅从时间的接近和背离上编辑样本，却忽略了学习样本中正例与反例个数的差异。另外，对网络的拓扑结构优化仅着眼于隐层节点数的变化，却没有对网络进行结构性的调整。

3．仅注意到典型样本的贡献，却忽视了非典型样本的干扰，缺乏对样本个例的分析。

二、本文主要工作

1．分析样本 40 维特征的独立性和携带的分类信息，在特征检验的指导下，利用主成分分析法降维和消除相关的特点，重新提取特征，实现特征相互之间独立且富含分类信息。

2．分析利用神经网建模失利的原因，用模糊神经网建模，并优化沙尘暴预报模型，尝试多种模糊神经网结构和参数、合理编辑学习样本和试报样本，组织不同的预报方案，实现沙尘暴预报率的进一步提高。

3．分析非典型的样本特征，根据样本的分布，建立非典型样本区域，利用样本与非典型区域的接近程度合理调整隶属度，实现统计建模。

第二章　模式建模

模式是对所研究的物理对象或过程的定量或者结构的描述。具有某些共同特征的模式的集合称为模式类。

模式识别（Pattern recognizing）是用数学、物理和技术的方法实现对模式的自动处理、描述、分类和解释，是信息科学的一个分支。其目的在于用机器部分地实现人的这种智能活动。不仅有很大的实用价值，而且对于探索人类对外部环境感知能力的机理也有重要的意义。模式识别技术现已在天气预报、生产控制、质量检验、疾病诊断、遥感监测、文字识别、地震探测、指纹分析、细胞分类、遗传工程等方面得到广泛应用。模式识别的经典方法有统计模式识别和句法模式识别。统计模式识别以统计决策理论为基础，建立统计学识别模型。句法模式识别又叫结构模式识别，它立足于分析模式的结构，以形式语言理论的概念为基础。

模式建模：建模是一种对未知世界的逼近方法，当建立的模型用于分类问题时，则模型便是对类间差异规律的总结，亦称分类器。相应的，传统的模式建模法分为统计建模和结构建模法，智能模式建模方法包括人工神经网络法和模糊神经网络法。

2.1 传统模式建模方法

2.1.1 统计建模法

统计模式识别技术是使用统计信息和估计理论去获取从表达空间到解释空间的映射。统计建模法是采用统计数学的方法在大量的样本中寻找类别差异规律，确定分类函数，结合具体问题的性质提出判别规则，由分类函数和判别规则组成分类器，从而形成分类模型。统计决策理论是处理模式分类问题的基本理论之一，它对模式分析和分类器的设计有着实际的指导意义。[9][10]

统计方法主要包括：回归分析、判别分析、聚类分析、探索性分析等。其中，主要的判别分析法为：

1．贝叶斯（Bayes）准则

贝叶斯决策理论方法是统计模式识别中的一个基本方法。在连续情况下，假设要识别的 d 维特征向量为 $x=[x_1,x_2,\cdots,x_d]^T$，各类别状态用 ω_i 来表示，其中 $i=1,2,\cdots,c$ 。$P(\omega_i)$ 和 $p(x|\omega_i)$ 分别为对应于各个类别 ω_i 出现的先验概率和类条件概率密度函数，则后验概率可由贝叶斯公式(2-1)确定，常用的判别规

则有最小错误率判别规则、最小风险判别规则、最大似然比判别规则等。

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)} = \frac{p(x \mid \omega_j)P(\omega_j)}{\sum_{j=1}^{c} p(x \mid \omega_j)P(\omega_j)} \tag{2-1}$$

将式(2-1)和根据具体问题选用的判别规则相结合就可进行分类判决。

2．费歇（Fisher）准则

Fisher 提出的判别分析方法并不对样本 d 维空间 R 直接划分，而是把 d 维空间中的样本投影到较低维空间，再在低维空间进行分类。所以，Fisher 判别分析的中心思想是找出最有利于低维空间分类的投影方法。Fisher 准则要求投影的结果满足各类之间差异尽量大，而同类样本之间的差异尽量小。

2.1.2 结构建模法[9]

结构模型是从概念模型到定量模型的中介．其重点在于元素间关系的测辨和模型结构的确定[11]。结构建模法着眼于模式结构特征。对于一个复杂的模式可以把其按结构分解成若干较简单子模式的组合，而子模式再按其结构分解成若干基元。辨认模式中的每一基元，完成待识模式的分类，参见图 2-1。

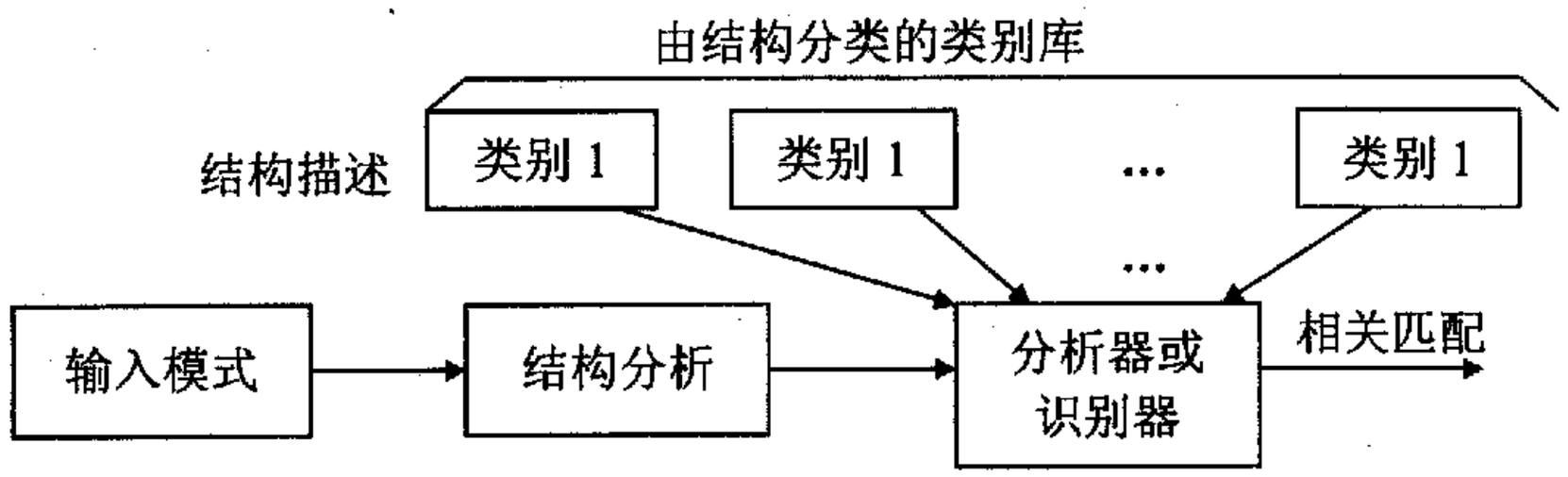


图 2-1 结构建模法

2.1.3 统计模式建模过程

模式识别系统都由两个过程组成，即设计和实现。设计是指用一定数量的样本（叫做训练集或学习集）确定出一套分类模型和判别规则(分类器)，使得按这套分类模型和判别规则对待识别样本进行分类所造成的错误识别率最小或引起的损失最小。实现是指用所设计的分类器对待识别的样本进行分类决策。

模式识别系统主要由 4 个部分组成：数据获取、预处理、特征提取和选择、分类决策，如图 2-2 所示。

1．数据获取。利用可以用于计算机运算的符号来表示所研究对象。

2．预处理。预处理的目的是去除噪声，增强有用的信息，并对相应的退化现象进行复原以及观测数据的无量纲归一化处理等。

3．特征提取和选择。对原始数据进行变换，得到并选择最能反映分类本质的特征以便有效的实现分类识别，这就是特征提取和选择的过程。

4．分类器设计及分类决策。分类决策就是在特征空间中用统计方法把被识别对象归入某一类别。

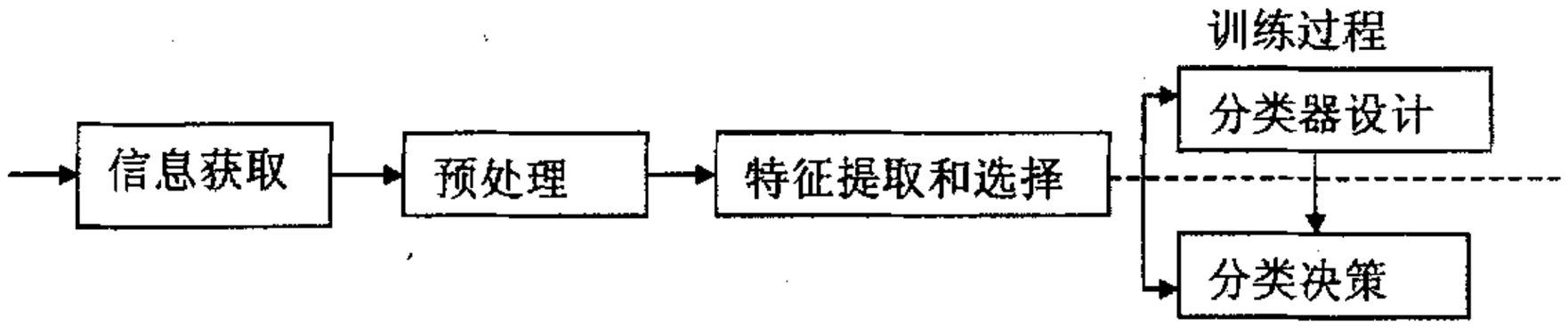


图 2-2 模式识别系统基本构成

2.1.4 多类问题策略

许多模式识别问题都是多分类的，这就要求解决多分类问题。因此，研究多类分类算法是很重要的。

1．“一对一”策略

对c类样本数据两两组合，构建$c(c-1)/2$个判别函数，把样本分为c个类别，每个判别函数只对其中的两个类别分类。

2．“一对多”策略

把c类问题化为$c-1$个两类问题，每个问题负责区分本类数据和非本类数据。

3．决策树

利用树分类器可以把一个复杂的多类别分类问题采用分级的形式转化为若干个简单的分类问题来解决。

2.2 智能模式建模方法[12][13][14]

近年来人们对人工神经网络法、模糊识别法应用于模式建模的研究日益增多，这两种方法更接近人的智能。

2.2.1 人工神经网络法[10][15][16][17]

人工神经网络已在各个领域得到广泛的应用，尤其是在智能系统中的非线性建模及其控制器的设计、模式分类与模式识别、联想记忆和优化计算等方面更是得到人们的极大关注。[18]

人工神经网络(ANN)由大量功能简单而具有自适应能力的信息处理单元—人工神经元按照大规模并行的方式，通过一定的拓扑结构连接而成。ANN 模型有多种形式，它取决于神经元特性函数、网络拓扑、学习算法这三大因素，具有联想记忆与回忆、分类、优化决策与计算等功能。

1．神经元特性函数的类型

常用的神经元特性函数有阶跃函数、准线性函数、Sigmoid 函数和双曲正切函数，其特性函数曲线分别如图 2-3(a)、(b)、(c)、(d)所示：

2．神经网络拓扑[19]

在神经网络应用中，各种模型层出不穷，但总的来说，大致可以归结为以下几类：

1)前馈式网络：该种网络结构是分层排列的，每一层的神经元输出只和下一层神经元相连。这种网络结构特别适用于 BP 算法，如今已得到了非常广泛的应用。

2)输出反馈的前馈式网络：该种网络结构与前馈式网络的不同之处在于这种网络存在着一个从输出层到输入层的反馈回路。该种结构适用于顺序型的模式识别问题。

3)前馈式内层互连网络：该种网络结构中，同一层之间存在着相互关联，神经元之间有相互制约的关系，但从层与层之间的关系来看还是前馈式的网络结构。许多自组织神经网络大多具有这种结构，如 ART 网络等。

4)反馈型全互连网络：在该种网络中，每个神经元的输出都和其它神经元相连，从而形成了动态的反馈关系，如 Hopfield 网络。

5)反馈型局部互连网络：该种网络中，每个神经元只和其周围若干层的神经元发生互连关系，形成局部反馈，从整体上看是一种网格状结构。该种网络特别适合于图像信息的加工和处理。

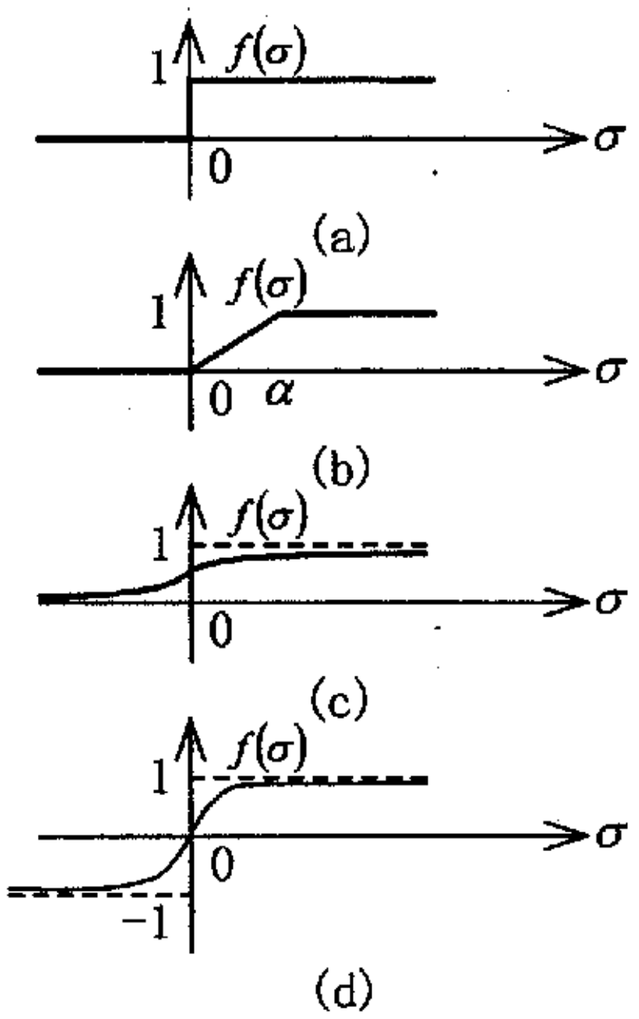


图 2-3 神经元特性函数
(a)阶跃函数(b)准线性函数
(c)Sigmoid 函数(d)双曲正切函数

3．神经网络的学习规则

根据学习样本有无期望的输出，可将学习方法分为有导师学习和无导师学

习两种。

1）有导师的学习网络：主要是根据网络的实际输出向量与期望输出向量之间的偏差来调整网络中的各权值。衡量偏差的准则不同，则学习的规则亦不同，人工神经网络有导师的学习规则主要有 *D.D.Hebb* 学习规则、纠错规则、感知器学习规则、δ 学习规则、广义 δ 学习规则、模拟退火算法和梯度下降算法。

2）无导师的学习网络：主要强调网络自身根据外界环境获取样本数据的信息特征来进行分类判别，无导师学习网络中比较典型的有 Kohonen 的自组织特征映射和 Grossberg 的自适应共振理论。两者均按照样本的相似程度重新组织样本空间，使得在输出空间里，相似的样本自动聚在一起。与有导师学习网络不同，无导师学习网络的学习准则不是基于误差代价函数，而是基于样本间的相似度或距离。无导师学习的学习规则主要有：用于 Hopfield 网络的 Hebbian 学习规则和常用于 Kohonen 自组织特征映射的竞争学习规则。

2.2.2 模糊神经网

一、模糊性及模糊神经网

在自然科学或社会科学研究中,存在着许多定义不很严格或者说具有模糊性的概念。这里所谓的模糊性，主要是指客观事物在中间过渡中差异的不分明性，如某一生态条件对某种害虫、某种作物的存活或适应性可以评价为“有利、比较有利、不那么有利、不利”；灾害性霜冻气候对农业产量的影响程度为“较重、严重、很严重”，等等。模糊集合论是处理分析这些“模糊”概念数据的主要数学工具。

模糊集合论的提出虽然较晚，但目前在各个领域的应用十分广泛。实践证明，模糊数学在农业中主要用于病虫测报、种植区划、品种选育等方面，在图像识别、天气预报、地质地震、交通运输、医疗诊断、信息控制、人工智能等诸多领域的应用也已初见成效。从该学科的发展趋势来看，它具有极其强大的生命力和渗透力。

虽然模糊逻辑和神经网络是两个截然不同的领域，它们的基础理论相差较远，但它们都是智能的仿真方法，站在实践和理论的角度完全可以将它们相结合。把模糊逻辑和神经网络相结合就产生了一种新的技术领域，这就是模糊神经网络。

二、 理论研究现状[20]

近年来，模糊神经网络的研究已取得了一些成果，主要体现在以下几个方

面：

1．模糊系统与神经网络系统作为一般自适应模型无关估计的研究。神经网络作为一般函数估计器，已广泛地适用于各种应用领域。

2．利用神经网络对模糊控制规则的获取、细化等方面的研究。

3．在神经网络学习算法中引入模糊控制技术的研究。

模糊逻辑和神经网络取得了一些成功的实际应用，如应用于窑炉、工业机器人控制等：但是，由于工业过程的复杂性，尤其在连续生产过程中干扰大，可变因素多，用模糊神经网络处理系统问题仍然有它的不足之处。主要表现在：

1．达不到真正的实时性要求，不能实现真正的实时自学习、自调整、自适应。在连续生产过程中，往往会存在一些干扰因素，或者生产条件出现变化，这时要求模糊神经网络能够识别这一变化，并通过自学习作出相应的处理。

2．抗干扰性能不强。在离线学习时，模糊神经网络具有较强的抗干扰性能。但是，在实时过程中，由于相关软件及硬件的限制、抗干扰的能力被削弱。

3．模糊神经网络的工程化应用还缺乏实用的开发平台。

2.2.3 神经网络与模糊神经网络模式建模过程

图 2-4 中的建模过程与统计法模式建模过程基本一致。但因人工神经网络作为模式分类器，有它自身的特点，两者又略有不同。“一般神经网络模式分类器兼有模式变换和模式特征抽取的作用，所以，一般的神经网络分类器不需要对输入的模式作明显的特征提取，网络的隐层本身就具有特征提取的功能”。[21]

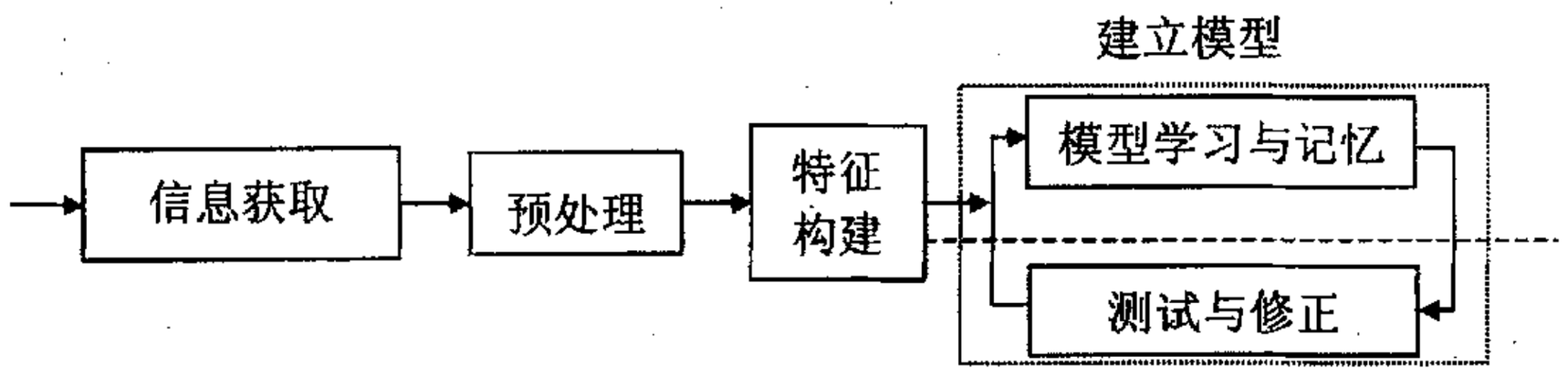


图 2-4 神经网络及模糊神经网络用于模式识别一般过程

神经网络具有强大的功能，从理论上讲，三层前馈神经网络就可以实现任意的非线性方程。但在建模初期，具有很多的不确定因素，一般都需要经过模型优化这一环节，通过反复试探寻找最适合的网络拓扑、训练参数、训练样本集合等，从而获取更优的模型。即使在其中一次建模中，也需要有学习和记忆的过程，就如同人脑对新的人或事物都有认识、了解的过程，通过长期的正确

样本的训练，才能得到正确的认识。在这点上，它与传统的建模方法是有区别的。

2.3 沙尘暴模式建模

2.3.1 沙尘暴预报问题的特点

沙尘暴是小概率事件，在本文研究的区域范围和时间段内，沙尘暴发生的气候概率约为 20%。

与沙尘暴相关的因子很多，地面情况、气压场、风场、湿度、温度等都与沙尘暴是否发生相关。由于所考虑的区域范围较广，又增加了区域内的这些因子的多样性，所以说沙尘暴属于高维问题。

沙尘暴样本的数据量十分庞大，每年的 2 月中旬至 6 月中旬时间段内沙尘暴日与非沙尘暴日数据共约一百多组，若以近二十年的相关数据作为沙尘暴模式建模的背景资料，就有近二千组数据。在这近二千组数据组成的数据集合中，正例（沙尘暴类样本）所占的比率仅约 20%，两类样本量很不均衡。

由上所述，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，还没有找到明确因果关系表达式。

2.3.2 沙尘暴模式建模与数据挖掘[22]

从信息来源的角度来看，建立沙尘暴预报模型的过程，是数据挖掘过程的主要部分。绝大部分有用信息隐藏在大量的沙尘暴相关数据资料中，依靠神经网络的数据挖掘能力，学习、总结其间的规律，并将这些规律分布式记忆在网络的权值中。

本文的沙尘暴预报模型的建立大致分为特征分析、基于模糊神经网的沙尘暴预报模型、统计建模三个阶段：

1．特征分析：虽然人工神经网络的隐层可理解为对输入变量做了变换，但这代替不了的特征(变量)的选择，特别是当样本量有限时提高网络性能的一个重要措施是压缩原始变量的维数[23]，从岳斌所做的神经网建模基础看，预报结果并不理想，特征是否富含分类信息是影响预报结果的重要方面。故对 40 维特征进行特征分析，检验沙尘暴 10 个模式间以及沙尘暴与非沙尘暴间的差异是否显著，以全面了解现有的 40 个特征。并依据对特征的统计分析结果寻找合理的特征提取方法，以便更好地组织沙尘暴预报模型的建立。

2．基于模糊神经网的沙尘暴预报模型：将模糊技术和神经网相结合，形成

既有神经网络的学习能力和非线性表达能力，又能表达近似于定性知识预报模型。通过对输入的模糊化，使原本分类结果与特征之间非常复杂的非线性关系有可能变为线性关系，这样不仅简化了分类器的设计更提高了分类器的性能。利用提取到的特征重新建立沙尘暴预报模型，并进行模型的优化，以期达到较理想的预报效果。

3．统计建模：由于地面情况、气压场、风场、湿度、温度等都是发生沙尘暴天气的相关因子，复杂的天气变化、特别是在沙尘与非沙尘天气的过渡期，将出现沙尘和非沙尘天气的某些物理场相类似的现象，这将导致模型对相当规模的非典型样本误报。既考虑沙尘与非沙尘天气的典型情况又兼顾到为数较多的特例，才能使模型更全面、客观地反映问题的实际。因此特别组织非典型样本进行统计建模，以求模型预报准确率的进一步提高。

另外，40 维特征的形成过程中，通过聚类挑选出沙尘暴的 10 个典型的模式类。即典型的沙尘暴模式具有 10 类，而非沙尘暴只有 1 类，由此，可以将沙尘暴和非沙尘暴的分类问题看为“一对十”的多类分类问题，新出现的样本，若属于沙尘暴的典型模式，则必为沙尘暴样本。而非沙尘暴样本则应背离沙尘暴的 10 个典型模式，这也是一种应该尝试的建模策略。

第三章 模糊神经网

模糊神经网络是一种新型的神经网络，它是在网络中引入模糊算法或模糊权系数的神经网络[24]。主要有三种结构[25]：

1.输人信号为普通变量，连接权为模糊变量；

2.输人信号为模糊变量，连接权为普通变量；

3.输人信号与连接权均为模糊变量。

3.1 模糊神经网建模

Zadeh 首先提出模糊建模的思想[26]，使模糊建模作为模糊系统研究中的一个关键问题而受到人们的关注。用于建模的模糊模型通常有如下 3 种类型：

1．Mamdani 模型[27]

规则：

R^l：if x_1 is G_1^l, x_2 is $G_2^l, \cdots, x_n$ is G_n^l then y is H^l, $l=1,2,\cdots,M$

其中 $G_i^l(i=1,2,\cdots,n)$和H^l均是模糊集合。

2．T-S 模型[28][29][30][37][38][39]

规则：

R^l：if x_1 is G_1^l, x_2 is $G_2^l, \cdots, x_n$ is G_n^l then y is $f^l(x_1, x_2, \cdots, x_n)$, $l=1,2,\cdots,M$

其中：$G_i^l(i=1,2,\cdots,n)$是模糊集合，$f^l(x_1, x_2, \cdots, x_n)$是输入 x 的函数，通常

$$f^l(x_1, x_2, \cdots, x_n) = a_0^l + a_1^l x_1 + \cdots + a_n^l x_n \text{。}$$

3．实常数后件模糊模型[31]

规则：

R^l：if x_1 is G_1^l, x_2 is $G_2^l, \cdots, x_n$ is G_n^l then y is θ^l, $l=1,2,\cdots,M$

其中：$G_i^l(i=1,2,\cdots,n)$是模糊集合，θ^l 是实常数。

3.2 隶属度函数

正确地确定隶属度函数，是运用模糊集合解决实际问题的基础，是能否用好模糊集合的关键。目前隶属度函数的确定方法大致有以下几种：

1．模糊统计方法：用对样本统计实验的方法确定隶属度函数。

2．例证法：从有限个元素的隶属度值来估计模糊子集隶属度函数。

3．专家经验法：根据专家的经验来确定隶属度函数。

4．机器学习法：通过神经网络的学习训练得到隶属度函数。

目前常用的隶属度函数有三角形、梯形、正态型、Γ型、和 Sigmiod 型五种。其函数曲线分别如图 3-1 所示：

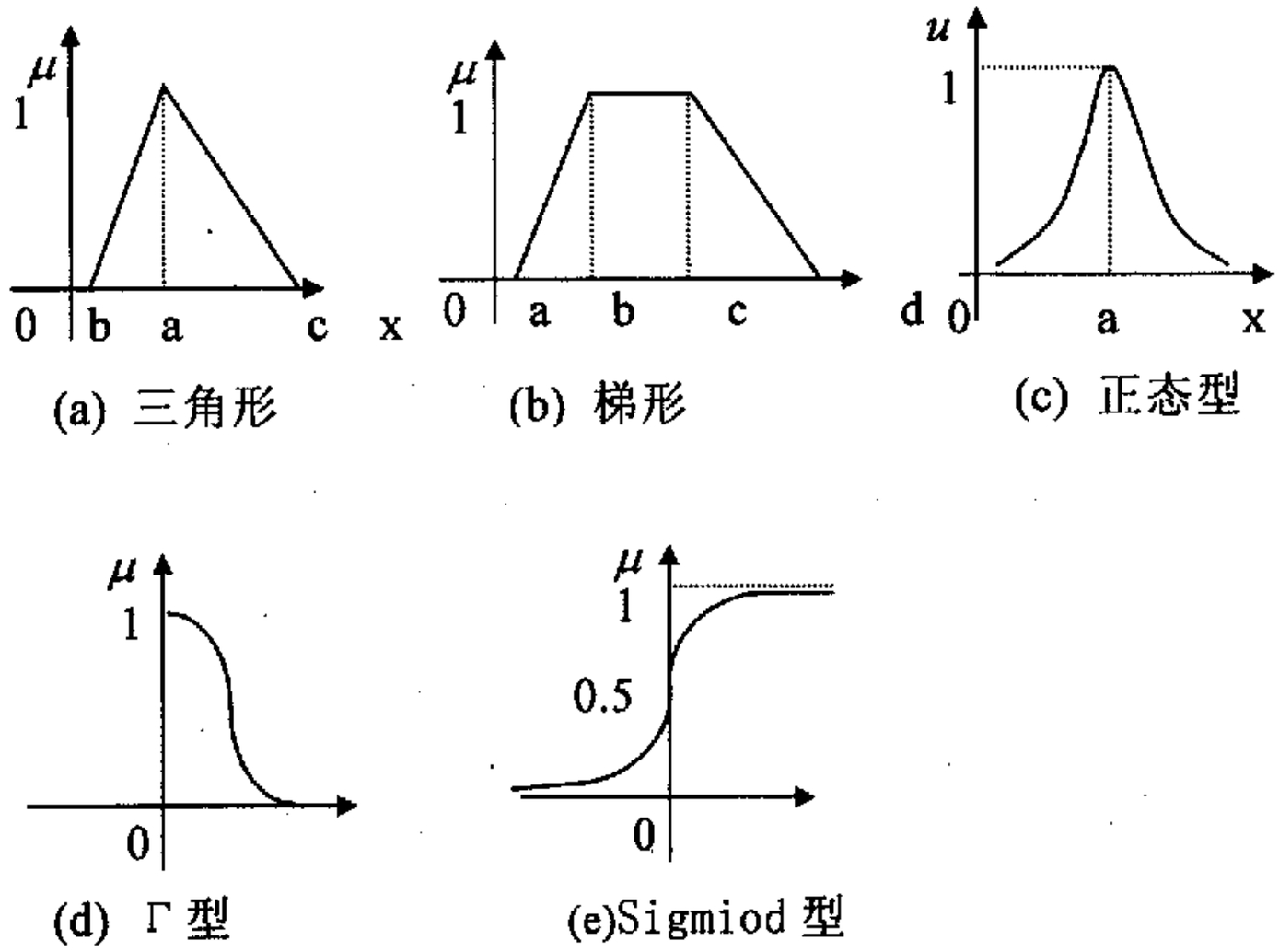


图 3-1 隶属度函数曲线

3.3 误差反向传播算法

3.3.1 误差反向传播算法简介[32]

误差反向传播算法又称 BP 学习算法，由正向传播和反向传播组成。正向传播是输入信号从输入层经隐层传到输出层．若输出层得到期望的输出，则学习算法结束；否则，误差转至反向传播。反向传播就是将误差信号按原连接通道反向计算，由梯度下降法调整各层神经元的权值和阈值、使误差信号减小。

由于梯度下降法的不足之处是算法收敛速度相对较慢，为此有下列改进的 BP 算法。

1．附加动量法

附加动量法使网络在修正其权值时，不仅考虑误差在梯度上的作用，而且考虑在误差曲面上变化趋势的影响，其作用如同一个低通滤波器，在没有附加动量的作用下，网络可能陷入局部极小值。该方法是在反向传播的基础

上，在每一个权值变化的基础上加上一项正比于前次权值变化量的值，并根据反向传播算法来产生新的权值变化。

2．自适应学习率

调节学习速率的准则是：当新误差超过旧误差一定的倍数时，学习速率将减少；否则其学习速率保持不变；当新误差小于旧误差时，学习速率将被增加。该方法可以保证网络总是以最大的可接受的学习速率进行训练。

3.3.2 算法步骤[33]

误差反向传播算法如下：

1．初始化权值及阈值为一个小的随机数；

2．施加输入矢量 $x_0, x_1, \cdots\cdots x_{n-1}$ 及期望输出 $d_0, d_1, \cdots\cdots d_{m-1}$；

3．从第一隐层开始，逐层计算输出矢量 $y_0, y_1, \cdots\cdots y_{m-1}$；

4．按式(3-1)修正权值：

$$\omega_{ij}(t+1) = \omega_{ij} + \eta\delta_i X_i' \tag{3-1}$$

其中：$\eta>0$ 为学习常数，ω_{ij} 是从节点 i 或输入端到节点 j 的权值，X_i' 是节点 i 的输出或一个输入，δ_i 是节点 j 的误差项：

$$\delta_i = y_i(1-y_i)(d_i - y_i) \qquad \text{（对输出层）} \tag{3-2}$$

$$\delta_i = x_i'(1-x_i')\sum\delta_k\omega_{ik} \qquad \text{（对隐层）} \tag{3-3}$$

重复 3~4，直到对所有样本权值不再变化或达到指定的迭代次数。

其中 2 和 3 是前向过程，4 是反向过程。

3.3.3 模糊权[34][35][36]的误差反向传播算法

一、模糊神经网络输出的计算

考虑一个四层的前向神经网络，用 I 表示输入层，H 表示隐层，O 表示输出层。假设其输入、输出、权系数和阈值都是三角模糊量，如对给定论域 U，输出 o 的 α 水平截集为 $o_\alpha = \{o \in \mathrm{U} \mid u_{\mathrm{A}}(o) \geq \alpha\}$，其中，A 为沙尘暴集合。其示意图如图 3-2 所示：

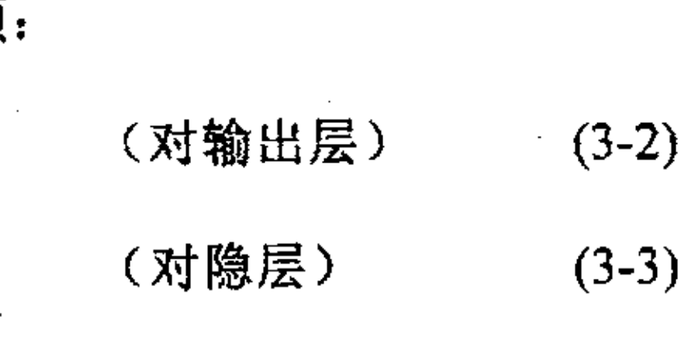


图 3-2 输出 o 的 α 截集图

则模糊神经网络输入输出关系可以写为

下面式子：

1．对于输入层：

$$I_i = x_i \tag{3-4}$$

2．对于第一隐层：

$$H_{1j}(\alpha)=\left[h_{1jl}(\alpha),h_{1ju}(\alpha)\right]=\left[f\left(U_{1jl}(\alpha)\right),f\left(U_{1ju}(\alpha)\right)\right] \tag{3-5}$$

其中：$U_{1jl}(\alpha)=\sum_{i=0}^{c} w_{ijl}(\alpha)\cdot x_i$

$$U_{1ju}(\alpha)=\sum_{i=0}^{c} w_{iju}(\alpha)\cdot x_i$$

c为输入层的节点数。

3．对于第二隐层：

$$H_{2j}(\alpha)=\left[h_{2jl}(\alpha),h_{2ju}(\alpha)\right]=\left[f\left(U_{2jl}(\alpha)\right),f\left(U_{2ju}(\alpha)\right)\right] \tag{3-6}$$

其中：$U_{2jl}(\alpha)=\overbrace{\sum_{i=0}^{m} w_{ijl}(\alpha)\cdot h_{1jl}(\alpha)}^{w_{ijl}(\alpha)\geq 0}+\overbrace{\sum_{i=0}^{m} w_{ijl}(\alpha)\cdot h_{1ju}(\alpha)}^{w_{ijl}(\alpha)<0}$

$$U_{2ju}(\alpha)=\overbrace{\sum_{i=0}^{m} w_{iju}(\alpha)\cdot h_{1ju}(\alpha)}^{w_{iju}(\alpha)\geq 0}+\overbrace{\sum_{i=0}^{m} w_{iju}(\alpha)\cdot h_{1jl}(\alpha)}^{w_{iju}(\alpha)<0}$$

4．对于输出层

$$o_j(\alpha)=\left[o_{jl}(\alpha),o_{ju}(\alpha)\right]=\left[f\left(U_{jl}(\alpha)\right),f\left(U_{ju}(\alpha)\right)\right] \tag{3-7}$$

其中：$U_{jl}(\alpha)=\overbrace{\sum_{i=0}^{n} w_{ijl}(\alpha)\cdot h_{2jl}(\alpha)}^{w_{ijl}(\alpha)\geq 0}+\overbrace{\sum_{i=0}^{n} w_{ijl}(\alpha)\cdot h_{2ju}(\alpha)}^{w_{ijl}(\alpha)<0}$

$$U_{ju}(\alpha)=\overbrace{\sum_{i=0}^{n} w_{iju}(\alpha)\cdot h_{2ju}(\alpha)}^{w_{iju}(\alpha)\geq 0}+\overbrace{\sum_{i=0}^{n} w_{iju}(\alpha)\cdot h_{2jl}(\alpha)}^{w_{iju}(\alpha)<0}$$

另外，对于各层

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

二、目标函数

对于网络的输出端 o 的 α 截集以及相应的目标输出 Y 的 α 截集，可以定义目标函数 e，对单输出神经网络，其目标函数定义为：

$$e=e_l(\alpha)+e_u(\alpha)=\frac{(Y_l-o_l)^2}{2}+\frac{(Y_u-o_u)^2}{2} \tag{3-8}$$

三、权值修正算法

根据梯度法，可以用目标函数对 w_{ij} 进行修正，过程如下：

$$\Delta w_{ijl}(\alpha)(t)=-\eta\frac{\partial e}{\partial w_{ijl}(\alpha)}+\beta*\Delta w_{ijl}(\alpha)(t-1) \tag{3-9}$$

$$\Delta w_{iju}(\alpha)(t)=-\eta\frac{\partial e}{\partial w_{iju}(\alpha)}+\beta*\Delta w_{iju}(\alpha)(t-1) \tag{3-10}$$

其中： η 是学习常数，β 是惯性常数。

1．隐层和输出层之间的权系数

$$\frac{\partial e}{\partial w_{iol}(\alpha)}=\frac{\partial e}{\partial o_l(\alpha)}*\frac{\partial o_l(\alpha)}{\partial u_l(\alpha)}*\frac{\partial u_l(\alpha)}{\partial w_{iol}(\alpha)}=d^3{}_l(\alpha)*\frac{\partial u_l(\alpha)}{\partial w_{iol}(\alpha)}$$

$$\frac{\partial e}{\partial w_{iou}(\alpha)}=\frac{\partial e}{\partial o_u(\alpha)}*\frac{\partial o_u(\alpha)}{\partial u_u(\alpha)}*\frac{\partial u_u(\alpha)}{\partial w_{iou}(\alpha)}=d^3{}_u(\alpha)*\frac{\partial u_u(\alpha)}{\partial w_{iou}(\alpha)}$$

其中： $d^3{}_l(\alpha)=o_l(\alpha)*(1-o_l(\alpha))*(Y_l(\alpha)-o_l(\alpha))$

$$d^3{}_u(\alpha)=o_u(\alpha)*(1-o_u(\alpha))*(Y_u(\alpha)-o_u(\alpha))$$

$$\frac{\partial u_l(\alpha)}{\partial w_{iol}(\alpha)}=\begin{cases}h_{2il}(\alpha) & if\ w_{iol}(\alpha)\geq 0\\ h_{2iu}(\alpha) & else\end{cases}$$

$$\frac{\partial u_u(\alpha)}{\partial w_{iou}(\alpha)}=\begin{cases}h_{2iu}(\alpha) & if\ w_{i0u}(\alpha)\geq 0\\ h_{2il}(\alpha) & else\end{cases}$$

2．隐层和隐层之间的权系数

$$\frac{\partial e}{\partial w_{ijl}(\alpha)}=d^2 1_{jl}(\alpha)*\frac{\partial u_{2jl}(\alpha)}{\partial w_{ijl}(\alpha)}+d^2 2_{jl}(\alpha)*\frac{\partial u_{2jl}(\alpha)}{\partial w_{ijl}(\alpha)}$$

$$\frac{\partial e}{\partial w_{iju}(\alpha)}=d^2 1_{ju}(\alpha)*\frac{\partial u_{2ju}(\alpha)}{\partial w_{iju}(\alpha)}+d^2 2_{ju}(\alpha)*\frac{\partial u_{2ju}(\alpha)}{\partial w_{iju}(\alpha)}$$

其中：

$$d^2 1_{jl}(\alpha)=\begin{cases}d^3{}_l(\alpha)*(1-h_{2jl}(\alpha))*h_{2jl}(\alpha)*w_{ijl}(\alpha) & if\ \ w_{ijl}(\alpha)\geq 0\\ 0.0 & else\end{cases}$$

$$d^2 2_{jl}(\alpha)=\begin{cases}d^3{}_u(\alpha)*(1-h_{2jl}(\alpha))*h_{2jl}(\alpha)*w_{iju}(\alpha) & if\ w_{ijl}(\alpha)<0\\ 0.0 & else\end{cases}$$

$$d^2 1_{ju}(\alpha)=\begin{cases}d^3{}_l(\alpha)*(1-h_{2ju}(\alpha))*h_{2ju}(\alpha)*w_{ijl}(\alpha) & if\ w_{iju}(\alpha)<0\\ 0.0 & else\end{cases}$$

$$d^2 2_{ju}(\alpha)=\begin{cases}d^3{}_u(\alpha)*(1-h_{2ju}(\alpha))*h_{2ju}(\alpha)*w_{iju}(\alpha) & if\ w_{iju}(\alpha)\geq 0\\ 0.0 & else\end{cases}$$

$$\frac{\partial u_{2jl}(\alpha)}{\partial w_{ijl}(\alpha)}=\begin{cases}h_{1il}(\alpha) & if\ w_{ijl}(\alpha)\geq 0\\ h_{1iu}(\alpha) & else\end{cases}$$

$$\frac{\partial u_{2ju}(\alpha)}{\partial w_{iju}(\alpha)}=\begin{cases}h_{1iu}(\alpha) & if\ w_{iju}(\alpha)\geq 0\\ h_{1il}(\alpha) & else\end{cases}$$

3．输入层和隐层之间的权系数，

$$\frac{\partial e}{\partial w_{ijl}(\alpha)}=\left(d^1 1_{jl}(\alpha)+d^1 2_{jl}(\alpha)+d^1 3_{jl}(\alpha)+d^1 4_{jl}(\alpha)\right)*\frac{\partial u_{1jl}(\alpha)}{\partial w_{ijl}(\alpha)}$$

$$\frac{\partial e}{\partial w_{iju}(\alpha)} = \left(d^1 1_{ju}(\alpha) + d^1 2_{ju}(\alpha) + d^1 3_{ju}(\alpha) + d^1 4_{ju}(\alpha)\right) * \frac{\partial u_{1ju}(\alpha)}{\partial w_{iju}(\alpha)}$$

其中： $d^1 1_{jl}(\alpha) = \left(1 - h_{1jl}(\alpha)\right) * h_{1jl}(\alpha) * \sum_{w_{jkl}(\alpha)\geq 0} d^2 1_{kl}(\alpha) * w_{jkl}(\alpha)$

$$d^1 2_{jl}(\alpha) = \left(1 - h_{1jl}(\alpha)\right) * h_{1jl}(\alpha) * \sum_{w_{jkl}(\alpha)\geq 0} d^2 2_{kl}(\alpha) * w_{jkl}(\alpha)$$

$$d^1 3_{jl}(\alpha) = \left(1 - h_{1jl}(\alpha)\right) * h_{1jl}(\alpha) * \sum_{w_{jku}(\alpha)< 0} d^2 1_{ku}(\alpha) * w_{jku}(\alpha)$$

$$d^1 4_{jl}(\alpha) = \left(1 - h_{1jl}(\alpha)\right) * h_{1jl}(\alpha) * \sum_{w_{jku}(\alpha)< 0} d^2 2_{ku}(\alpha) * w_{jku}(\alpha)$$

$$d^1 1_{ju}(\alpha) = \left(1 - h_{1ju}(\alpha)\right) * h_{1ju}(\alpha) \sum_{w_{jkl}(\alpha)< 0} d^2 1_{kl}(\alpha) * w_{jkl}(\alpha)$$

$$d^1 2_{ju}(\alpha) = \left(1 - h_{1ju}(\alpha)\right) * h_{1ju}(\alpha) * \sum_{w_{jkl}(\alpha)< 0} d^2 2_{kl}(\alpha) * w_{jkl}(\alpha)$$

$$d^1 3_{ju}(\alpha) = \left(1 - h_{1ju}(\alpha)\right) * h_{1ju}(\alpha) * \sum_{w_{jku}(\alpha)\geq 0} d^2 1_{ku}(\alpha) * w_{jku}(\alpha)$$

$$d^1 4_{ju}(\alpha) = \left(1 - h_{1ju}(\alpha)\right) * h_{1ju}(\alpha) * \sum_{w_{jku}(\alpha)\geq 0} d^2 2_{ku}(\alpha) * w_{jku}(\alpha)$$

$$\frac{\partial u_{1jl}(\alpha)}{\partial w_{ijl}(\alpha)} = x_i$$

$$\frac{\partial u_{2ju}(\alpha)}{\partial w_{iju}(\alpha)} = x_i$$

四、学习步骤

1．对模糊权系数和模糊阈值赋初值；

2．正向计算：利用式(3-4)、(3-5)、(3-6)、(3-7)计算各层输出值；

3．利用式(3-8)计算误差值，如果预定的结束条件满足，则结束，否则继续；

4．反向传播：利用式(3-9)、(3-10)修正权值；

5．转入 2。

第四章　特征分析

4.1 统计检验理论[40]

在气象分析中，特别关注两个或多个气象要素之间的关系，在天气预报过程中，往往通过前后期各种气象要素变化及有关理论和经验确定一些预报指标，来预报未来天气。这些预报指标及其统计关系是否反映了大气变化的规律、可靠性如何，通常需要统计检验。

统计检验，是根据实践经验和实际实行的可能性，对需要了解的总体的某些特征，如参数的取值或取值的范围、总体的分布类型、两个总体之间的相关性、两个总体分布之间的差异、时间序列的随机性、不同预报指标及预报方法的效果之间的差异等等，做出一定的假设，然后再根据实测数据，按一定的公式计算来判断此假设是否合理，从而决定是接受或舍弃假设。所以统计检验又称为假设检验、统计假设检验，也有的称为显著性检验。

4.1.1 均值检验

均值检验是用来检验两类样本的平均数差异是否显著，若已知二个正态变量母体的子样为：$x_1,x_2,\cdots,x_{n1}$ 和 $y_1,y_2,\cdots,y_{n1}$，他们的数学期望分别为 M_X,M_Y，其均值检验的步骤如下：

第一步：假设两类样本平均数相等，计算子样的平均数

$$\overline{X}=\sum_{i=1}^{n_1}x_i/n_1\,,\qquad \overline{Y}=\sum_{i=1}^{n_2}y_i/n_2$$

第二步：计算子样标准差

$$S_X=\sqrt{\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{n_1-1}}\,,\qquad S_Y=\sqrt{\frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_2-1}}$$

第三步：计算Δ

$$\Delta=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n_1-1)S_X{}^2+(n_2-1)S_Y{}^2}{n_1+n_2-2}}}$$

第四步：计算统计量 t

$$t=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n_1-1)S_X{}^2+(n_2-1)S_Y{}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\Delta}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

第五步：查t_α表

计算自由度$f=n_1+n_2-2$，给出置信度α，查t分布表得到置信限t_α。

第六步：比较t和t_α

若$|t|\geq t_\alpha$拒绝原假设，说明$\overline{X}$与$\overline{Y}$差异显著。若$|t|<t_\alpha$接受原假设，说明$\overline{X}$与$\overline{Y}$差异不显著。

4.1.2 方差检验

自然界在不同条件下，出现的各种天气次数往往不相等，所以不等重复实验在气象上应用更加广泛。如果因素 A 有个m水平，每个水平重复次数分别为$n_1,n_2,\cdots,n_m$，试验总次数为 N，

$$N=\sum_{i=1}^{m}n_i=n_1+n_2+\cdots n_m$$

设各水平各次试验值为X_{ij}，其中i为试验水平$i=1,2,\cdots,m$；j为每个试验重复序号$j=1,2,\cdots,n_i$；

则第i水平观测值的总和记为：

$$X_{Si}=\sum_{j=1}^{n_i}X_{ij}$$

X的总和记为：

$$X_S=\sum_{j=1}^{m}X_{Si}=\sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}$$

总合计平均平方和P为：

$$P=\frac{1}{N}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}\right)^2$$

A因素各水平合计平均平方和：

$$Q=\sum_{i=1}^{n}\left(\sum_{j=1}^{n_i}X_{ij}\right)^2\Big/n_i$$

个体平方和：

$$R=\sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}^{\ 2}$$

那么，各水平之间平方和：$S_A=Q-P$

误差平方和：$S_E=R-Q$

总平方和：$S_T=R-P$

它们的自由度分别为：$f_A=m-1,\quad f_E=n-m\quad f_T=n-1$

计算 F 分布的统计量：

$$F(m-1,n-m)=\overline{S}_A/\overline{S}_E=\frac{Q-P}{R-Q}\frac{n-m}{n-1}$$

给出置信度α，查 F 分布表，得到置信限$F_\alpha(f_A,f_E)$。比较 F 和F_α，判断差异是否显著。

4.2 沙尘暴特征的统计检验

4.2.1 沙尘暴模式间的显著性检验

为分析沙尘暴样本的40个特征能否反映沙尘暴各子模式间的差异，利用192个强沙尘暴样本（10个子模式涉及的样本数）进行方差检验，表4-1为方差检验的结果（$\alpha=0.01, F_\alpha(9,182)=2.52$）。检验结果表明10个沙尘暴模式间具有显著性差异，这同时说明了聚类结果的合理性。

表 4-1　模式间的方差检验结果

特征号	F	特征号	F	特征号	F	特征号	F
1	54.35	11	16.46	21	15.11	31	18.19
2	30.12	12	30.68	22	17.04	32	18.45
3	58.81	13	23.67	23	55.90	33	34.77
4	77.52	14	13.09	24	46.43	34	20.71
5	24.35	15	26.65	25	22.78	35	26.65
6	18.14	16	21.37	26	21.03	36	25.64
7	61.59	17	16.65	27	67.38	37	41.89
8	42.59	18	17.43	28	37.16	38	32.35
9	31.07	19	35.04	29	19.16	39	41.92
10	20.40	20	19.53	30	9.02	40	106.65

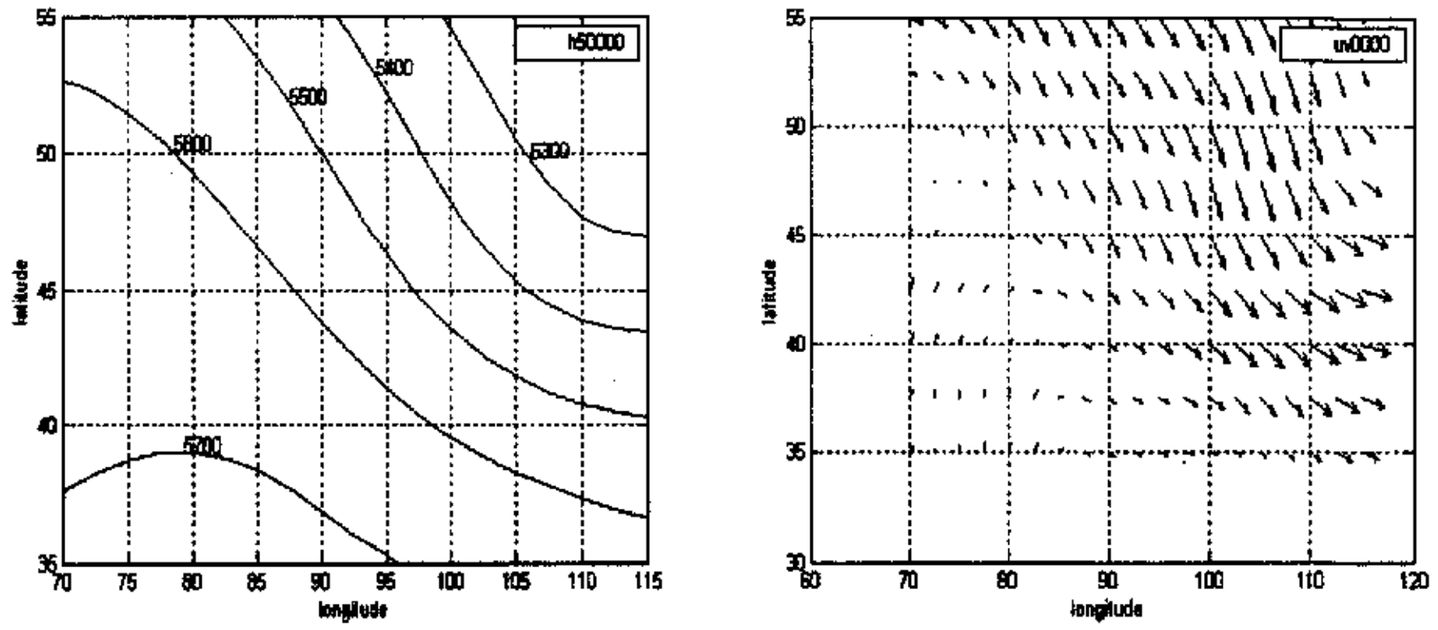
4.2.2 沙尘暴与非沙尘暴之间的显著性检验

沙尘暴样本的特征提取是建立沙尘暴预报模型的前提，能明确区分沙尘暴与非沙尘暴是 40 个特征的真正意义所在，故重要的是在这 40 个特征描述下，沙尘暴与非沙尘暴之间具有显著性差异。利用 81-89 年 1082 个样本，针对每一个特征组织两类样本集之间的方差检验，检验结果如表 4-2 所示，在α=0.01 的情况下，$F_{0.01}(1,\infty)=6.63$。其中标*的为差异不明显特征。

从表 4-2 所示结果可以看出，50%以上的特征对沙尘暴和非沙尘暴而言，不单独具有分类能力。对此可以解释如下：从确立特征的过程可以看出，每一沙尘暴子模式都具有 4 个中心场，通过计算样本的值或形的 4 个格点场与相应中心场的欧式距离得到 4 个特征，10 个模式得到 40 个特征，其中每 4 个特征出于同一个模式，应该是这 4 个特征的联合而不是单独某 1 个特征体现沙尘暴与非沙尘暴之间的差异。例如，沙尘暴日和非沙尘暴日可能具有相似的风场分布。另外，因为是四个物理场的相互作用促进了沙尘暴的形成，故出于同一模式的四个特征相互之间必然存在一定的联系，例如，风场和高度形场之间具有相似的趋势等，如图 4-1、4-2 所示。因此，将同一模式四个特征协同起来审视沙尘暴与非沙尘暴之间的显著性差异较为客观、合理。

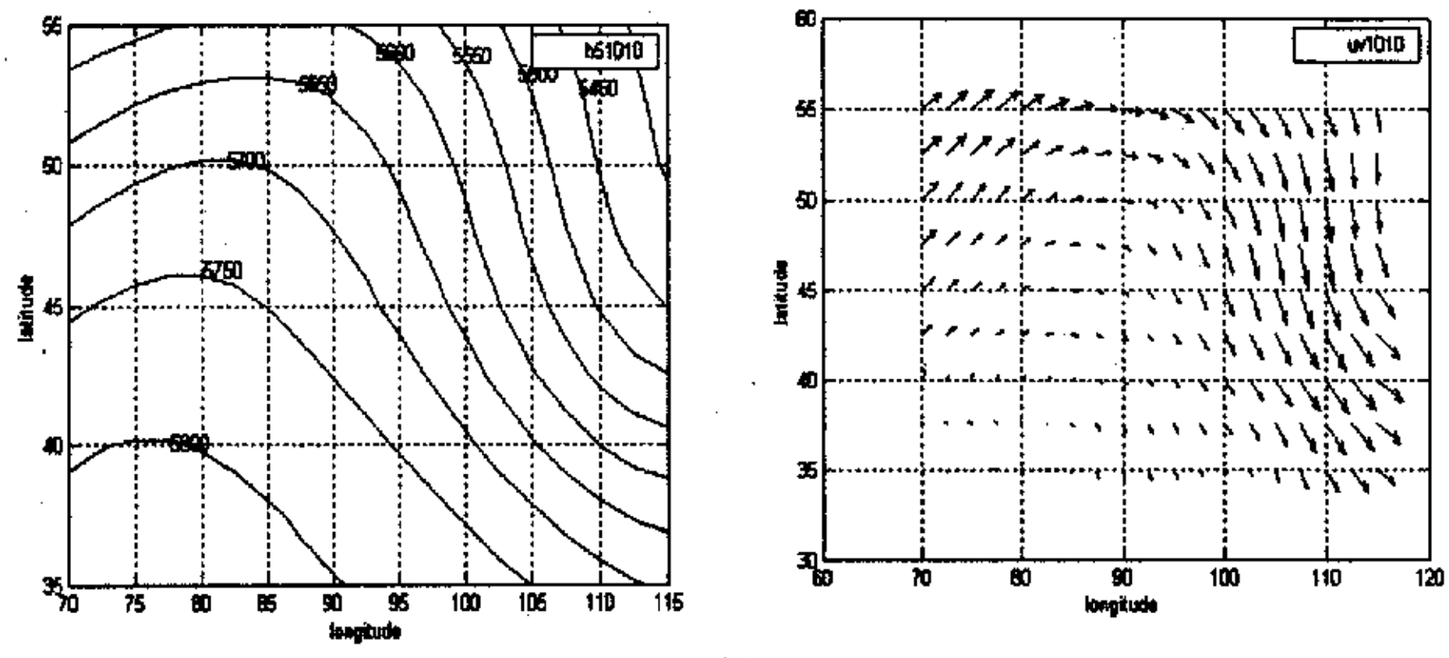
表 4-2 强沙尘暴与非沙尘暴间 40 个特征的检验结果

特征号	F	特征号	F	特征号	F	特征号	F
*1	1.14	*11	1.41	21	42.96	*31	1.07
*2	0.01	*12	0.84	22	44.00	32	10.76
*3	0.06	*13	1.20	*23	2.42	*33	0.26
*4	1.50	*14	0.16	*24	1.48	*34	0.94
*5	5.79	15	29.80	25	127.75	35	12.85
6	39.54	*16	0.04	26	78.50	*36	0.70
7	33.14	*17	4.56	27	58.98	37	8.78
8	34.58	*18	5.82	*28	0.26	*38	0.97
9	28.95	19	32.74	29	108.95	39	12.22
10	23.17	20	11.02	*30	0.01	40	12.17



(a) (b)

图 4-1 模式 1（a）高度场分布 (b)风场分布



(a) (b)

图 4-2 模式 7 (a)高度场分布 (b)风场分布

4.3 特征提取

4.3.1 代表性样本选择

从 40 维特征的形成过程看，特征是样本与典型模式中心场的距离。根据距离的大小，大体有距离近和距离远两种情况。理想情况下，沙尘暴样本应接近于某一典型模式，而非沙尘暴样本则应具有与典型模式较远的距离。距离模式 1 较近的沙尘暴样本客观上反映了其与沙尘暴典型模式 1 的联系，具有代表意义；而非沙尘暴样本则应远离沙尘暴的典型模式，即与典型模式 1 距离较远的非沙尘暴也具代表性，这些统称为模式 1 的代表性样本。相应的，10 个典型模式则能选择出 10 个代表性样本集。

一、模式中心值计算

设样本的 40 个特征为 $(x_{ij}, i=1,2,\cdots,10, j=1,2,3,4)$，其中 i 表示沙尘暴的典型

模式，j表示高度形和值场、风形场、位温形场等四个中心场，相对于这4个中心场，样本每4个特征出于一个模式，则特征x_{ij}表示出于模式i的第j个特征。对模式1的样本，选择出于模式1的特征x_{11}、x_{12}、x_{13}和x_{14}，利用公式(4-1)求得4个模式中心值，如此10个模式共得到40个模式中心值$(\bar{x}_{ij}, i=1,2,\cdots,10, j=1,2,3,4)$。

$$\bar{x}_{ij}=\frac{1}{n_i}\sum_{l=1}^{n_i}x_{ij}^{il} \tag{4-1}$$

其中，x_{ij}^{il}为模式i中第l个样本的出于模式i的第个j特征，n_i为模式i的样本个数。

二、参考均值与参考方差计算

计算10个典型模式中每个样本与对应模式中心的相似程度，相似程度用距离量来度量；

$$dis_i^{il}=\sqrt{\sum_{j=1}^{4}(x_{ij}^{il}-\bar{x}_{ij})^2} \tag{4-2}$$

其中，i为模式类，dis_i^{il}为模式i中的第l个样本与模式中心i的相似程度。

利用典型模式中样本与对应模式中心的相似程度，总结出每个典型模式与模式中心的相似度，其中均值$E_i, i=1,2,\cdots,10$反映典型模式的集中分布中心，而方差$D_i, i=1,2,\cdots,10$则又反映模式中样本对均值的偏离程度。

$$E_i=\frac{1}{n_i}\sum_{l=1}^{n_i}dis_i^{il} \tag{4-3}$$

$$D_i=\sqrt{\frac{1}{n_i}\sum_{l=1}^{n_i}(dis_i^{il}-E_i)^2} \tag{4-4}$$

式(4-3)与(4-4)中的均值与方差将作为选择代表性样本的参考值，在此将其称为参考均值与参考方差。

三、代表性样本选择

根据Δdis_i^l，将所有沙尘暴和非沙尘暴样本进行归类，当沙尘暴样本满足$\Delta dis_i^l < k * D_i$，$i=1,2,\cdots,10$或非沙尘暴样本满足$\Delta dis_i^l >= k * D_i$，$i=1,2,\cdots,10$时，则将其视为第$i$类代表性样本。

其中：　k为一常数，经试探，取$k=1.2$

$$\Delta dis_i^l = dis_i^l - E_i$$

$$dis_l^l = \sqrt{\sum_{j=1}^{4}(x_{ij}^l - \overline{x}_{ij})^2}$$

l为样本编号。

如此从所有备选样本中挑出 10 个代表性样本集，组成 10 个文件，以备特征提取之用。

4.3.2 特征提取

一、主成分分析法[41][42]

借助主成分分析法可以使高维样本通过降维得到简化，并尽量保留原变量的信息量，主成分分析有消除相关、降维的功能。具体做法如下：

1．设 n 维模式矢量组为 $X=[x_1,x_2,\cdots,x_n]^T$，则矢量组的均值可表示为：$\mu_x=\mathrm{E}[X]$，在未知概率密度函数的情况下，矢量均值可以通过离散抽样近似得出：

$$\mu_x=\frac{1}{\mathrm{M}}\sum_{k=1}^{\mathrm{M}}x_k$$

2．模式矢量组的协方差一般可定义为：

$$\sum_x=E[(x_k-\mu_x)(x_k-\mu_x)^T]$$

3．求出协方差矩阵 $\sum_x$ 的特征矢量 ϕ_i 和特征值 λ_i：

$$\sum_x\phi_i=\lambda_i\phi_i$$

4．n 维随机输入矢量可用主分量表示为：

$$Y=\phi^T X$$

分量形式为：

$$y_i=\phi_i^T x_i$$

二、对样本数据的主成分分析

分别对 10 个代表性样本集的样本数据进行主成分分析，过程如下：

1．对数据进行标准化处理

$$z_{ij}^{l}=\frac{X_{ij}^{l}-X_{ij}(\min)}{X_{ij}(\max)-X_{ij}(\min)}$$

其中，$X_{ij}(\max)=\max_{1\le l\le n}(X_{ij}^{l})$ 为极大值；$X_{ij}(\min)=\min_{1\le l\le n}(X_{ij}^{l})$ 为极小值。标准化后数据取值范围为[0，1.0]。

2．利用主成分分析法进行数据变换

$$(z_{i1}\quad z_{i2}\quad z_{i3}\quad z_{i4})^{T}\xrightarrow{\text{(主成分分析)}}y_{ij}$$

其中：$i=1,2,\cdots,10$ 为代表性样本集序号；$j=1,\cdots,4$ 为特征值序号；$y_{ij}=a_{ij}{}^{T}Z_{i}=a_{ij1}z_{i1}+\cdots+a_{ij4}z_{i4}$ 为特征值 λ_{ij} 对应的主成分。

3．方差检验

取第一主成分 y_{i1} 作为样本特征中与模式 i 相关的 4 个特征的综合特征，共得到 10 个综合特征，对这 10 个特征进行沙尘暴和强沙尘暴间的显著性差异检验。检验结果如表 4-3 所示。

在α=0.01 的情况下，$F_{0.01}(1,\infty)=6.63$，假设检验差异不明显的标以*，从表中可以看出，大约有一半以上的特征不明显。

表 4-3　沙尘暴与非沙尘暴间关于第一主成分的方差检验

y_{i1}	y_{11}^{*}	y_{21}^{*}	y_{31}	y_{41}^{*}	y_{51}	y_{61}^{*}	y_{71}	y_{81}^{*}	y_{91}	$y_{10,1}^{*}$
(F_{α})	4.9	0.7	17.4	1.2	48.0	6.4	21.6	0.2	73.9	4.8

三、各主成分的信息综合

根据特征值 λ_{ij} 能够定量反映主成分 y_{ij} 代表的分类信息的特点，求出第 1 主成分 y_{i1} 表示分类信息总量的比例，如表 4-4 所示。可以看出，利用第一主成分 y_{i1} 作为样本特征中与模式 i 相关的 4 个特征的综合，舍弃分类信息均在 52%以上。

为此设计特征综合方案如下：

将模式类 i 的 4 个主成分（特征）综合，即令

$$y_{i}=b_{i1}y_{i1}+\cdots+b_{i4}y_{i4}$$

根据主成分反映分类信息总量的比例确定上式中权系数

表 4-4 y_{ij} 所反映的分类信息比例

模式	λ_{i1}	λ_{i2}	λ_{i3}	λ_{i4}	y_{i1} 的分类信息
1	0.01521	0.00835	0.00583	0.00295	47.03%
2	0.00856	0.00508	0.00317	0.00238	44.59%
3	0.00793	0.00750	0.00597	0.00227	33.51%
4	0.01581	0.01171	0.00956	0.00369	38.79%
5	0.00895	0.00633	0.00479	0.00259	39.49%
6	0.01082	0.00654	0.00488	0.00198	44.68%
7	0.01504	0.00729	0.00666	0.00262	47.60%
8	0.01588	0.01208	0.00969	0.00438	37.78%
9	0.01403	0.01264	0.00770	0.00450	36.11%
10	0.01368	0.00860	0.00711	0.00280	42.52%

$$b_{ij} = \frac{\lambda_{ij}}{\lambda_{isum}}$$

则
$$y_i = \frac{\lambda_{i1}}{\lambda_{isum}} y_{i1} + \frac{\lambda_{i2}}{\lambda_{isum}} y_{i2} + \frac{\lambda_{i3}}{\lambda_{isum}} y_{i3} + \frac{\lambda_{i4}}{\lambda_{isum}} y_{i4} \tag{4-5}$$

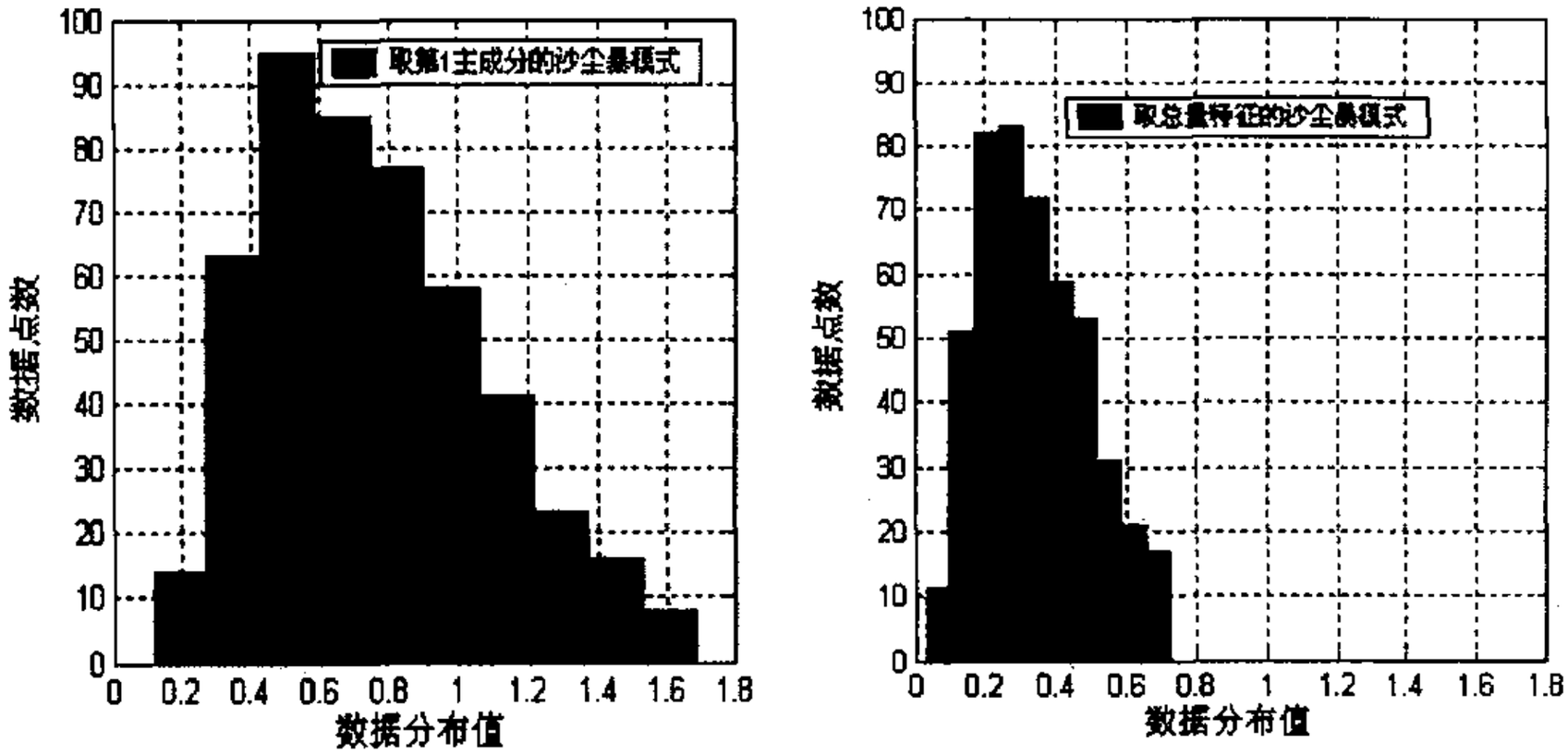
式中，$\lambda_{ij} = D(y_{ij})$ 为主成分分量 y_{ij} 所代表的分类信息；

$\lambda_{isum} = \sum_{j=1}^{4} \lambda_{ij} = \sum_{j=1}^{4} D(y_{ij}) = D(y_i)$ 为分类信息总量。并称 y_i 为模式 i 的“总量特征”。

图 4-3 分别给出了沙尘暴样本关于第 1 主成分(特征)和总量特征的分布图示例，可以看出，在总量特征下样本分布更集中、分布范围明显减小。

对这 10 个总量特征进行非沙尘暴和沙尘暴间的显著性差异检验。设 $\alpha = 0.01$，$F_{0.01}(1,\infty) = 6.63$，检验结果列于表 4-5 中。其中，符号“*”表示相应特征关于沙尘暴和非沙尘暴两类差异不明显。

从统计检验的结果看，与表 4-3 的结果相比沙尘暴和非沙尘暴的差异总体上比较显著。这表明，本文所讨论的特征提取方法是成功的，提取的特征是有效的。



(a) (b)

图 4-3 沙尘暴的样本分布 (a) 关于第 1 主成分(b) 关于总量特征

表 4-5 沙尘暴与非沙尘暴间关于总量特征的方差检验

y_i	y_1	y_2^*	y_3	y_4^*	y_5^*	y_6	y_7	y_8	y_9	y_{10}^*
(F_α)	14.3	3.3	10.5	0.2	2.4	69.8	14.5	22.1	175.2	0.002

第五章 沙尘暴预报模型的建立

5.1 基于模糊权方法的沙尘暴预报模型

5.1.1 模糊权的神经网

一、网络拓扑

网络的层数和每层节点数将影响决策曲面。对一给定任务到底需要多少隐层单元没有简单规则可循。除了训练中和训练后的性能问题外，过多的隐层神经元会产生所谓过拟和现象。网络有过多的信息处理能力时，过拟和就可能产生。它将学习训练集合中的不重要的方面[42]。

故应慎重选择神经网层数和每层节点数，利用前馈神经网的结构，整个神经网由输入层、第一隐层、第二隐层和输出层组成，经多次试验，各层节点数选择 10*25*15*1，如图 5-1 所示。

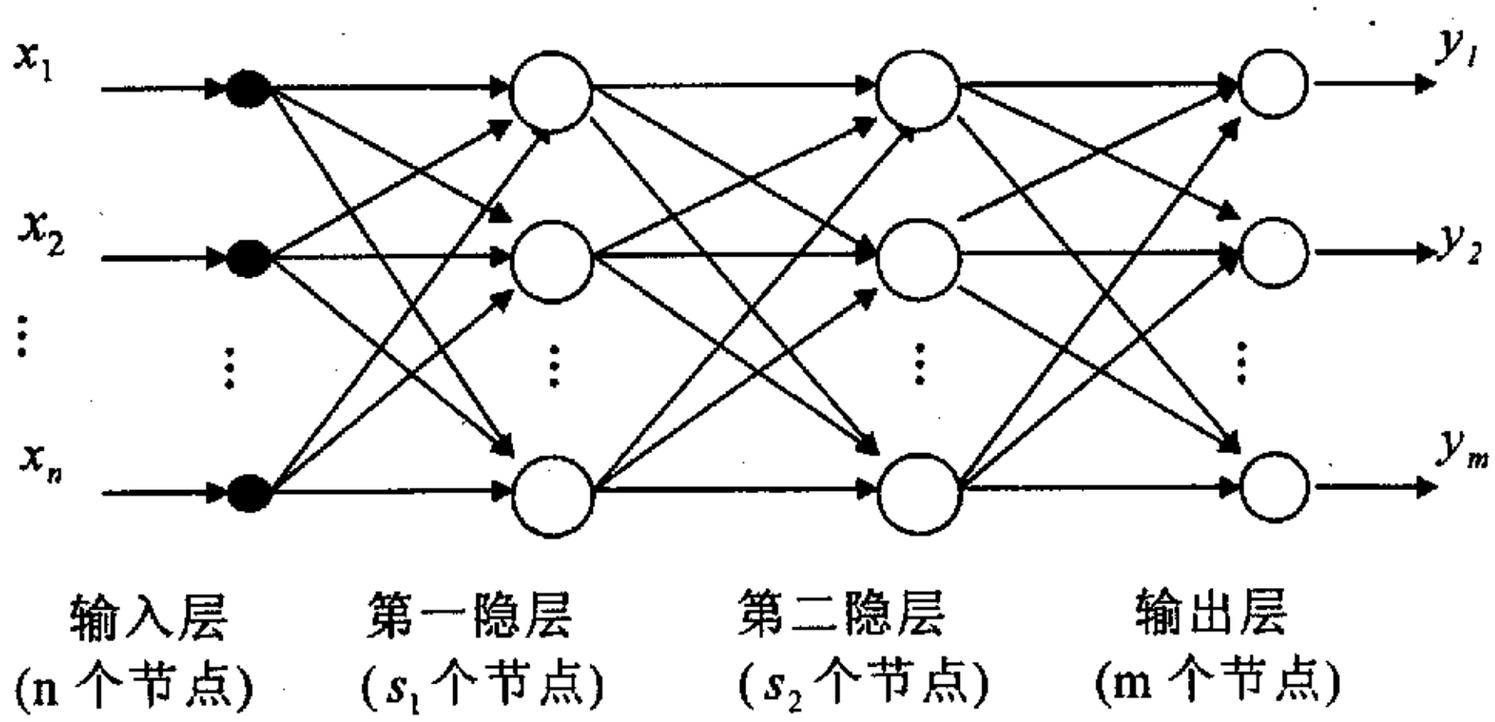


图 5-1 神经网的拓扑结构

二、判别规则

根据第三章所述的模糊权神经网的学习算法，其输出、权系数都是三角模糊量；对于 α 截集（取为 0.5），权值有 w_l 和 w_u 之分，输出有 o_l 和 o_u 之分，相应的，学习率分为 η_l,η_u，惯性系数分为 β_l,β_u。

样本属于沙尘暴的隶属度为：

$$o=\frac{o_l+o_u}{2} \tag{5-1}$$

依据式(5-1)计算出的隶属度 o，来判断样本是否属于沙尘暴类。判别规

则如下：

$$\text{样本被}\begin{cases}\text{判为沙尘暴} & \text{当} o \geq \sigma \text{时} \\ \text{判为非沙尘暴} & \text{当} o < \sigma \text{时}\end{cases}$$

其中，σ为划分沙尘暴和非沙尘暴的阈值，取为一常数（例如取σ=0.5）。

5.1.2 样本编辑

神经网络中记忆的规律，是从已知样本和已知变量组成的数据集中学习训练得到的。很明显，只有当已知样本的数据可靠且具有代表性时，学习到的规律才可能是正确的，才能得到对预报有用的数学模型。

编辑学习样本和试报样本，组织几种典型的学习方案，选择合适的训练次数，使沙尘暴和非沙尘暴的报对率均衡。其预报结果如表 5-1 所示，其中网络拓扑为 10*25*15*1，学习率为 $\eta_l = 0.5, \eta_u = 0.3$，惯性系数为 $\beta_l = 0.08, \beta_u = 0.04$。

表 5-1 几种样本编辑方案的预报结果比较

试报样本			学习样本范围及训练次数		
			81-85 年 (900 次)	81-87 年 (600 次)	81-89 年 (2100 次)
试报样本范围	86-97 年	沙尘暴报对	67.3%		
		非沙尘暴报对	69.0%		
		总报对率	68.6%		
		CSI	32.1%		
	88-97 年	沙尘暴报对	68.7%	73.0%	
		非沙尘暴报对	68.9%	71.1%	
		总报对率	68.9%	71.4%	
		CSI	29.9%	33.1%	
	90-97 年	沙尘暴报对	69.1%	73.0%	43.8%
		非沙尘暴报对	69.6%	71.2%	85.3%
		总报对率	69.5%	71.4%	77.7%
		CSI	29.5%	32.2%	26.6%

从表中看出，以 81-87 年的样本作为训练样本，训练次数少，沙尘暴（正例）和非沙尘暴（反例）的预报率均衡且预报结果较好，81-85 年学习样本

较少，当试报样本中出现一些新的典型的沙尘暴或非沙尘暴类型时，神经网因为事先没有学习不能做到很好的识别，故 81-85 年的学习方案虽能保证沙尘暴和非沙尘暴的预报率均衡，但预报结果相对较差。而 81-89 年虽然囊括了较多的样本，但由于沙尘暴和非沙尘暴数量不均衡造成沙尘暴的预报率较低。各年段沙尘暴和非沙尘暴的样本分布情况如表 5-2 所示。

沙尘暴预报问题具有小概率和样本量不均衡的特点，反例过多将造成神经网过拟和于非沙尘暴样本，另外，反例样本中也包括相当数量的非典型样本（偏离本类而与异类相似的样本），过多的非典型样本将产生干扰，对沙尘暴预报模型的训练产生不利影响。

所以，在建模过程中，对训练集合的组织既要考虑到样本的代表性又要考虑到样本的均衡性。

表 5-2　沙尘暴和非沙尘暴样本数比例

81-85 年 样本数(601 个)		81-87 年 样本数(841 个)		81-89 年 样本数(1082 个)	
沙尘暴	非沙尘暴	沙尘暴	非沙尘暴	沙尘暴	非沙尘暴
257 (42.8%)	344 (57.2%)	342 (41.7%)	499 (59.3%)	397 (36.7%)	685 (63.3%)

5.1.3 参数调整

一、学习率和惯性系数

误差反向传播算法中有两个参数η和β。步长（学习率）η对收敛性影响很大，而且对于不同的问题其最佳值相差也很大，惯性系数β影响收敛速度，合理调整这两个参数，则会使预报结果有所改善。

对模糊权的神经网络，其输出、权系数都是三角模糊量；权值有w_l和w_u之分，鉴于此，我们在调整权值时，设计如下学习率和惯性系数调整方案：

方案 1：不同的学习率$\eta_l \neq \eta_u$和惯性系数$\beta_l \neq \beta_u$；

方案 2：相同的学习率$\eta_l = \eta_u$和惯性系数$\beta_l = \beta_u$。

比较在两种方案下 88-97 年样本的预报率，如表 5-3 所示，方案 1 即对权值w_l和w_u的调整采用不同的学习率η_l, η_u和惯性系数β_l, β_u，样本的报对率较高。分析原因如下：对模糊权的神经网络，训练神经网的目标是使神经网的输出o_l和o_u分别收敛于理想输出Y_l和Y_u，由于最佳值的差别，o_l和o_u对Y_l和Y_u的拟和不能同时达到最佳值，总会有其中一组出现过拟和或拟和不够

的现象，利用并合理调节学习率 $\eta_l \neq \eta_u$ 和惯性系数 $\beta_l \neq \beta_u$，尽量使两者的拟和值同时达到或接近最佳值，将会取得较好的效果。

表 5-3 惯性系数和学习率的调整与试报效果

参数调整方案		训练样本(81-87) 试报样本(88-97)			
		沙尘暴报对	非沙尘暴报对	总报对	*CSI*
方案 1	$\eta_l = 0.5, \eta_u = 0.3$，$\beta_l = 0.08, \beta_u = 0.04$	73.0%	71.1%	71.4%	33.1%
方案 2	$\eta_l = \eta_u = 0.5$ $\beta_l = \beta_u = 0.08$	61.4%	78.3%	75%	32.2%
	$\eta_l = \eta_u = 0.3$ $\beta_l = \beta_u = 0.04$	69.5%	68.4%	68.6	30.0%

二、选择阈值

阈值 σ 是划分沙尘暴和非沙尘暴界限，通过选择合适的阈值，可以最大限度的分离沙尘暴样本和非沙尘暴样本。依据 88-97 年样本的预报结果，其预报率随阈值的变化曲线如图 5-2 所示。

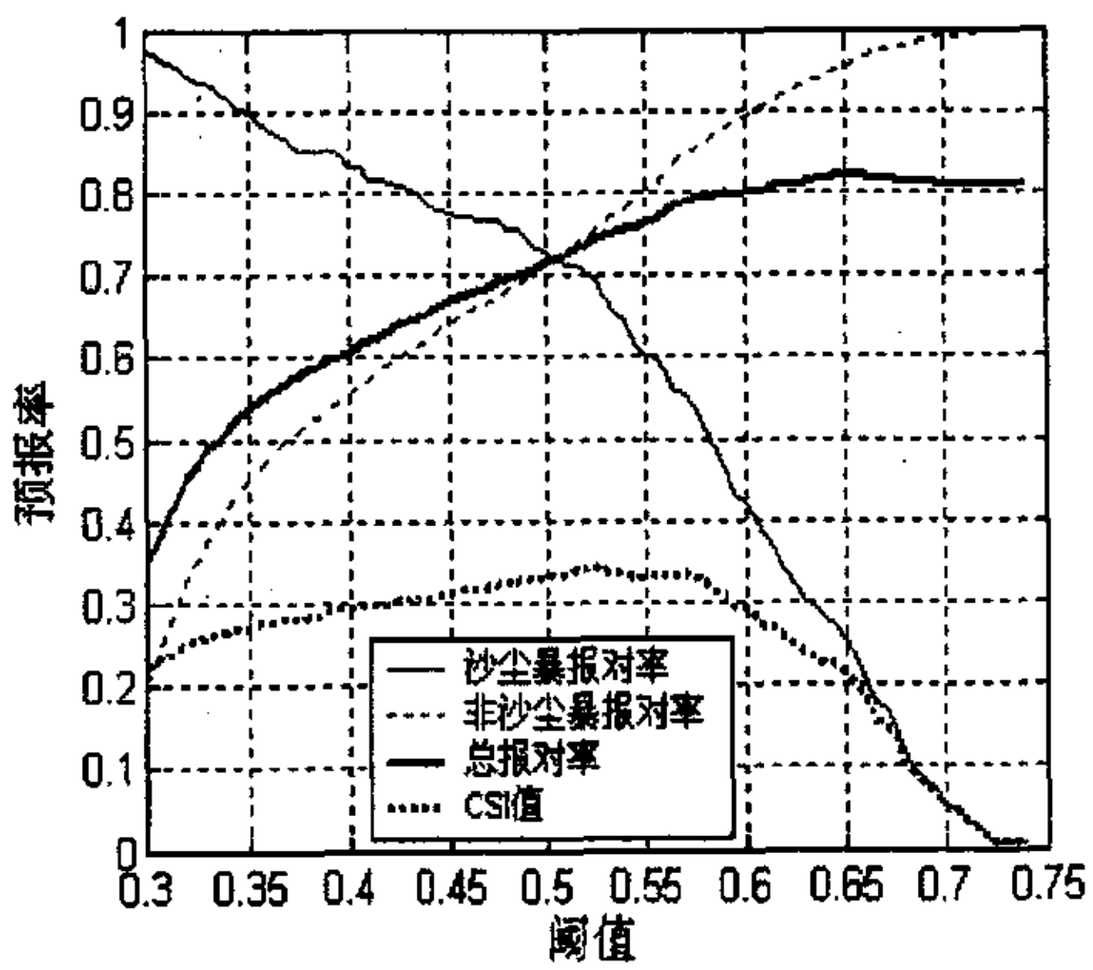


图 5-2 预报率随阈值变化的曲线

因沙尘暴属于小概率事件，虽然阈值的增加使沙尘暴报对率呈递减趋

势，但因非沙尘暴报对率的递增和其样本数量的绝对优势，使总报对率居高不下，均衡考虑沙尘暴和非沙尘暴的报对率以及 *CSI* 值，最佳阈值分布在 0.5-0.55 之间。

5.2 统计建模

由于地面情况、气压场、风场、湿度、温度等都是发生沙尘暴天气的相关因子，复杂的天气变化将使沙尘天气和非沙尘天气的某一个或几个物理场相类似，这些样本总体上表现为沙尘或非沙尘天气，但其物理场却与同类和异类物理场有相似之处或偏离本类物理场，属于非典型的沙尘暴或非沙尘暴样本，本文对使用基于模糊权方法的沙尘暴预报模型预报不中的样本进行分析，发现它们基本表现出以上特点，我们将这些样本统称为非典型建模样本，并用来进行统计建模。

利用模糊权的神经网对沙尘暴的预报可达到一定的预报效果，但由于非典型样本的干扰，对一些样本不能做到很好的识别，这就需要对样本进行非典型性分析即统计建模。建立统计模型，需参考并依赖于模糊神经网的预报结果，具体如下：

1. 模糊神经网的预报结果可作为界定非典型样本的依据。

2. 统计建模只是对样本属于沙尘暴的隶属度进行小规模的调整，其隶属度调整公式应兼顾并侧重模糊神经网预报结果。

3. 选择隶属度调整公式时也以模糊神经网预报结果为依据，例如在模糊神经网预报中报为沙尘暴与报为非沙尘暴的样本分别采用不同的隶属度调整公式。

5.2.1 非典型样本的界定

基于模糊权方法的沙尘暴预报模型的测试结果中，88-97 年的非典型样本 344 个，将其分为 88-92、93-97 两个年段，其中 88-92 年段的样本 189 个，沙尘暴 39 个，非沙尘暴 150 个，93-97 年段的样本 155 个，沙尘暴 24 个，非沙尘暴 131 个。

5.2.2 统计建模

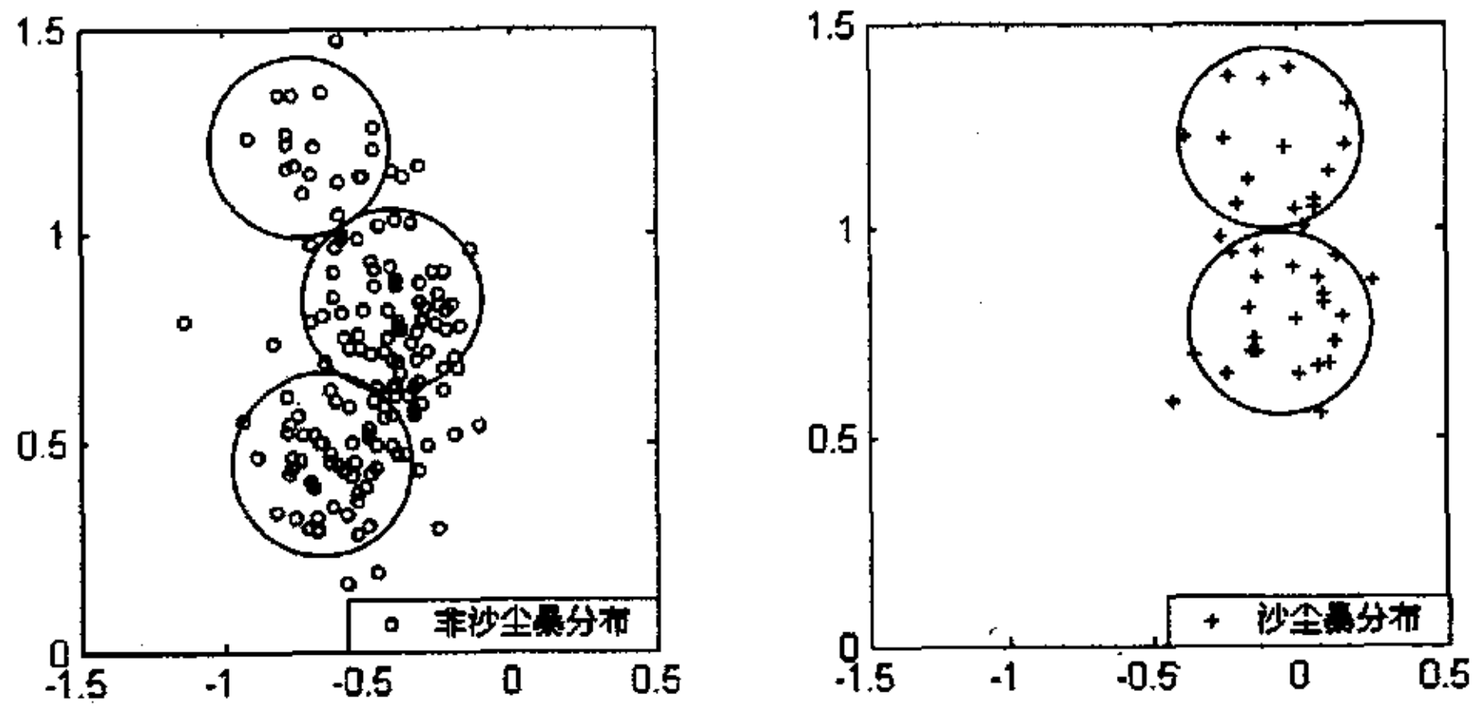
一、参考样本的聚类

将 88-92 年的非典型样本作为参考样本，将参考样本进行降维处理，形成如图 5-3 所示的参考样本分布。

依据参考样本分布，以尽量减小类内离散度为原则，形成如下聚类策略：

1．将参考沙尘暴样本聚为 2 类，其中，类 1 含样本 23 个，类 2 含 16 个。

2．参考非沙尘暴样本被聚为 3 类，其中，类 1、2、3 各含样本 70、30、11 个。



(a) (b)

图 5-3 非典型样本的分布(a)非沙尘暴样本(b)沙尘暴样本

二、计算非典型样本聚类中心

分别计算 5 类参考样本的聚类中心（$C^i=(c_1^i,\cdots,c_k^i,\cdots,c_{10}^i)$，$i=1,2,\cdots,5$）如下：

$$c_k^i=\frac{1}{n_i}\sum_{l=1}^{n_i}x_k^{il}, \tag{5-2}$$

其中 x_k^{il} 为类 i 中第个 l 个样本的第 k 个特征，

n_i 为类 i 中的样本个数，

c_k^i 为类 i 的聚类中心的第 k 个特征。

三、调整沙尘暴隶属度

1．计算最大距离 d_{max}^i

称非典型样本的分布中心为非典型类中心，而样本离其非典型类中心的距离则反映了该样本的非典型程度，其计算公式如式(5-3)所示：

$$d^{il}=\sqrt{\sum_{k=1}^{10}(x_k^{il}-c_k^i)^2} \tag{5-3}$$

其中：d^{il}为类i中的第个l个样本与其非典型类中心的距离。

而每类中都有样本离其非典型类中心的距离最远，该最大距离用d_{max}^i，$i=1,2,\cdots,5$表示，其中$d_{\max}^i=\max(d^{il}),l=1,2,\cdots,n_i$，$n_i$为类$i$中的样本个数。

2．隶属度调整

样本离非典型类中心的距离反映了该样本的非典型程度，同时也可以作为判断是否落入非典型样本区的依据，同一个样本，可能不落入非典型区或落入一个非典型区，也可能同时落入几个非典型区，由于各非典型样本区的最大距离$d_{\max}^i$不同，故利用相对距离γ^i（$\gamma^i=d^i/d_{\max}^i$，d^i为样本与第i个非典型区的距离）作为判断样本是否落入非典型区的条件。

计算待调整样本离 5 个非典型区的相对距离$\gamma^i,i=1,2,\cdots,5$，并取其中的最小值$\gamma_{\min}=\min(\gamma^i),i=1,2,\cdots,5$，相应的，相对距离最小的非典型区被称为该样本的最近非典型区，根据$\gamma_{\min}$的大小来判断该样本是否落入最近非典型区。

判别方式如下：

$$\text{样本}\begin{cases}\text{进入非典型区} & \text{当}\gamma_{\min}<\sigma_\gamma\text{时}\\ \text{未进入非典型区} & \text{当}\gamma_{\min}\geq\sigma_\gamma\text{时}\end{cases}$$

其中，σ_γ为区分典型区和非典型区的阈值。

相应的，样本的隶属度调整如下：

1）当样本未落入非典型区时，若模糊神经网报为沙尘暴，说明该样本为典型的沙尘暴模式，则应适当提高属于沙尘暴的隶属度，隶属度调整公式为(5-4)。

2）样本未落入非典型区时，若模糊神经网报为非沙尘暴，则该样本为典型的非沙尘暴模式，应降低属于沙尘暴的隶属度，见公式(5-5)。

3）样本落入沙尘暴的非典型区时，说明其类型与非典型的沙尘暴样本有所类似，应根据其与该非典型类的相似程度适当调整属于沙尘暴的隶属度。计算公式为(5-4)。

4）样本落入非沙尘暴的非典型区，说明其类型与非典型的非沙尘暴样本有所类似，应根据其与该非典型类的相似程度适当调整并减小属于沙尘暴的隶属度。计算公式为(5-5)。

式(5-4)和(5-5)中，a反映了侧重模糊神经网预报结果的程度，而b则反映了对非典型程度的侧重度。

$$result_2 = \begin{cases} a * result_1 + b * \dfrac{1}{1+e^{-|\gamma_{\min}-u|}} & (5-4) \\ a * result_1 + b * (1 - \dfrac{1}{1+e^{-|\gamma_{\min}-u|}}) & (5-5) \end{cases}$$

其中：

$u = \dfrac{1}{n}\sum_{k=1}^{n}(d_k / d_{\max})$ 为最近非典型区中参考样本的相对距离均值，

$result_1$ 为模糊神经网预报结果中的沙尘暴隶属度，

$result_2$ 为统计建模结果中的沙尘暴隶属度，

a,b 为侧重度系数，

函数 $\dfrac{1}{1+e^{-|\gamma_{\min}-u|}}$ 与 $1-\dfrac{1}{1+e^{-|\gamma_{\min}-u|}}$ 的取值范围分别为[0.5,1)与(0,0.5]。

5.2.3 统计建模预报结果

利用统计建模中的4种隶属度调整方式，对93-97年的样本进行隶属度调整，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 $a=0.9,b=0.1$。统计建模的预报结果如表5-4所示。取不同的阈值，沙尘暴和非沙尘暴的报对率都有所变化，从 *CSI* 值上来看，阈值取为0.53时，可达到38.7%，较利用40维特征的神经网建模[5]结果（*CSI* = 25.9%）来看，结果较好。

表5-4 统计建模预报结果

	阈值	沙尘暴报对率	非沙尘暴报对率	总报对率	*CSI*
统	0.5	76.2%	76.4%	76.4%	35.2%
计	0.51	75.2%	77.2%	76.9%	35.3%
建	0.52	75.2%	79.8%	79.0%	37.6%
模	0.53	73.3%	82.0%	80.5%	38.7%
备注：文献[5]中，沙尘暴报对率为60%，*CSI* = 25.9%					

在对统计模型的预报测试中，共涉及到样本601个，其中进入非典型区的样本共238个，经隶属度调整后（阈值取为0.53），进入非典型区和未进入

非典型区的报对与报错样本数量都有所变化， 25.81%的模糊神经网预报时报错样本经调整后得到了纠正，而报对样本的调整出错率仅有0.67%，如表5-5所示。

由表 5-4、5-5 的结果可以看出，经统计建模的隶属度调整后，报对样本的总量有所增加。取阈值为最佳范围（0.5-0.55），从图 5-4 中的曲线可以看出，预报率有显著的提高。

表5-5 报对与报错样本数量变化比较

		进入非典型区	未进入非典型区	合计
报错样本	模糊神经网报错总数（个）	69	86	155
	统计建模纠正数（个）	27	13	40
	纠正率（%）	39.13%	15.12%	25.81%
报对样本	模糊神经网报对总数（个）	169	279	448
	统计建模出错数（个）	1	2	3
	出错率（%）	0.59%	0.72%	0.67%

5.3 总结

本文先从 40 维特征出发，分别利用神经网、模糊神经网和集成模糊神经网进行建模，利用神经网络中记忆的规律，并采用多种学习方案，即利用不同的学习样本和测试样本以及不同的学习率和惯性系数进行尝试，取其中一种较优的方案作为最终结果，结果表明，利用 40 维特征时，预报结果不理想，主要表现为沙尘暴和非沙尘暴的报对率不均衡或报对率低。

在利用 40 维特征建模不理想的情况下，对 40 维特征进行了方差检验，针对检验结果，通过主成分分析，形成了更为合理的 10 维建模样本，实现了模糊神经网和统计建模相结合沙尘暴建模方式，并得到较为理想的结果。图 5-6 给出了本文工作的全过程。

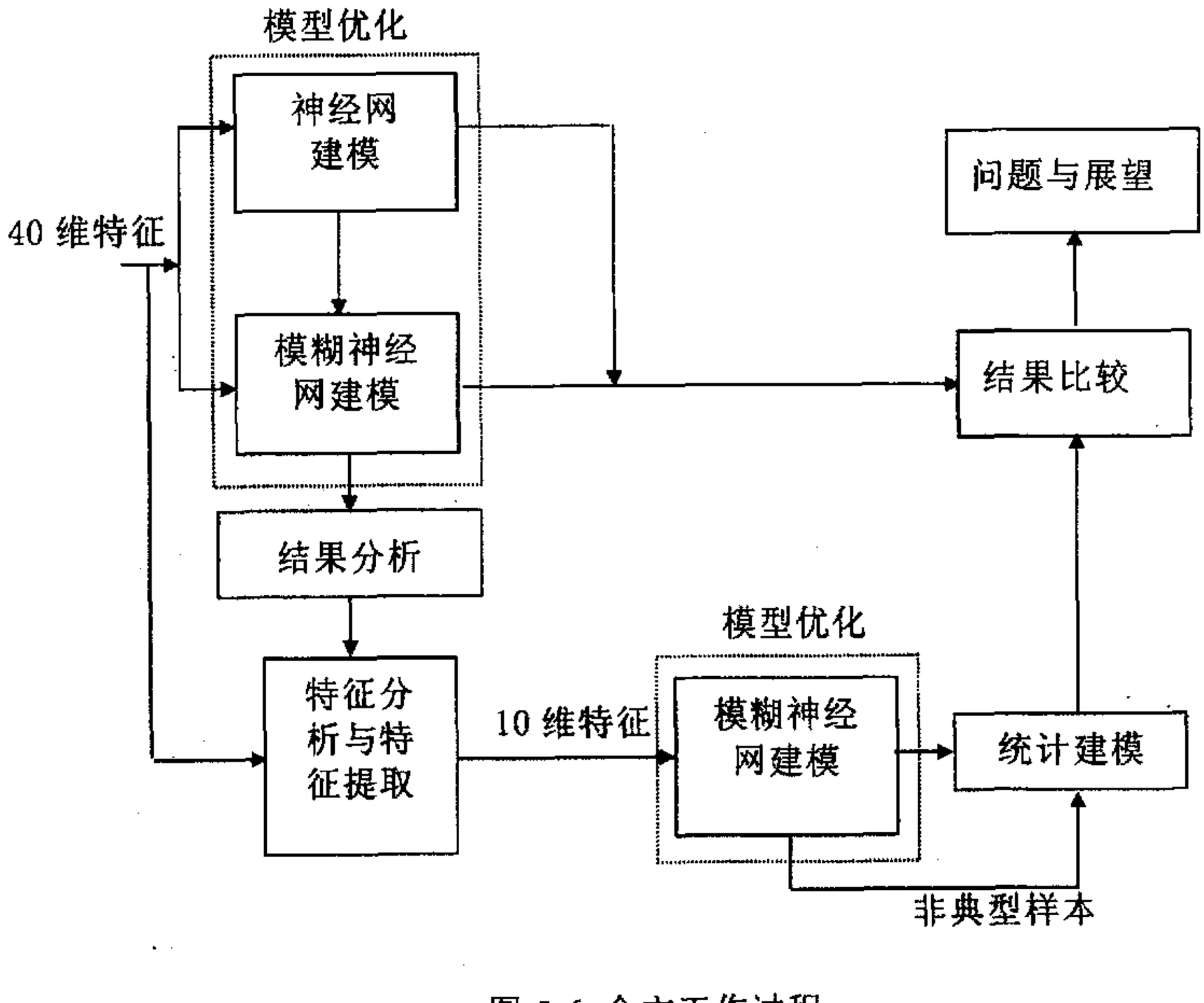

模型优化
神经网
建模
40 维特征
模糊神经
网建模
结果分析
特征分
析与特
征提取
10 维特征
模型优化
模糊神经
网建模
非典型样本
结果比较
统计建模
问题与展望

图 5-6 全文工作过程

第六章 问题与展望

6.1 问题分析

虽然本文系统已经达到一定的沙尘暴预报水平，但鉴于沙尘暴的小概率、多因子、高维、数据量大和难预测的特点，仍有很大潜力有待进一步挖掘。

一、存在的问题

1．沙尘暴原始数据是 NCEP 资料。选用的气象因子为 500hpa 的高度场、700hpa 的风场、850hpa 的位温场。样本的原始特征共计 855 维。选择 10 及 10 个站点以上的强沙尘暴，通过聚类，得到 10 个典型模式类，并利用与这 10 个典型模式类的相似度，得到样本的 40 个特征 [5]。从 40 维特征的提取过程看，在选择模式类时，只考虑了强沙尘暴典型模式，却忽略了少站点沙尘暴和非沙尘暴的典型模式；

2．BP 网络优化及样本数据库的更新都将影响沙尘暴预报模型的准确率。

二、进一步的研究

1．在建模过程中，利用 40 个特征组织沙尘暴预报模型，非沙尘暴和强沙尘暴报对率都比较理想，报错样本主要是少站点的沙尘暴样本。样本特征不能合理的描述少站点沙尘暴是预报出错的原因，特征的合理性和描述问题的能力是所有模式识别问题的重要前提，且特征的确定，需气象专家的更多经验和试验，充分重视少站点沙尘暴和非沙尘暴样本，使样本同时具备反映与强沙尘暴、少站点沙尘暴和非沙尘暴相似程度的特征将更能合理的描述沙尘暴问题。

2．模型优化的进一步研究。

3．样本分布不均衡是沙尘暴预报的一大特点，特别是最近几年，这一特点更加明显，如在 81-85、90-95、96-97 中，沙尘暴所占比例分别为 42.8%、21.2%、10.4%，从递减的沙尘暴出现率看，一些作为参考的典型沙尘暴模式不再出现或不再典型，相应的也会产生一些新的典型模式，故及时更新典型模式样本或适当调整典型模式的类型对预报结果也会产生重大的影响。

4．从建模的比较结果看，针对非典型样本，对沙尘暴预报进行统计建模为一种合理建模方案，但非典型样本的选择至关重要，另外，非典型区的选取和界定，隶属度的调整方案、以及非典型样本的及时更新问题都需要进行进一步的研究。

6.2 展望

6.2.1 特征重构

沙尘暴的预报模型，目的是进行沙尘暴和非沙尘暴之间的预报，其中，沙尘暴包括强沙尘暴和少站点沙尘暴，因为少站点沙尘暴在沙尘暴中占了相当大的比例（约 57.9%），提取少站点沙尘暴的典型模式和非沙尘暴的典型模式，使样本同时具备反映与强沙尘暴、少站点沙尘暴和非沙尘暴相似程度的特征将更具有广泛性。

6.2.2 模型优化

一、集成神经网络[43]

目前，集成神经网络已经在各个领域广泛应用，如高木等人就利用神经网络集成的模糊系统成功地预报了日本大阪湾的 COD(chemical oxygen demand)浓度。

利用集成的神经网络代替原有的单一 BP 网络，是模型优化的一个侧面。根据沙尘暴特征提取过程的特点，可组织以下集成的神经网络：

1．混合集成神经网络

集成神经网络的结构图如图 6-1 所示。

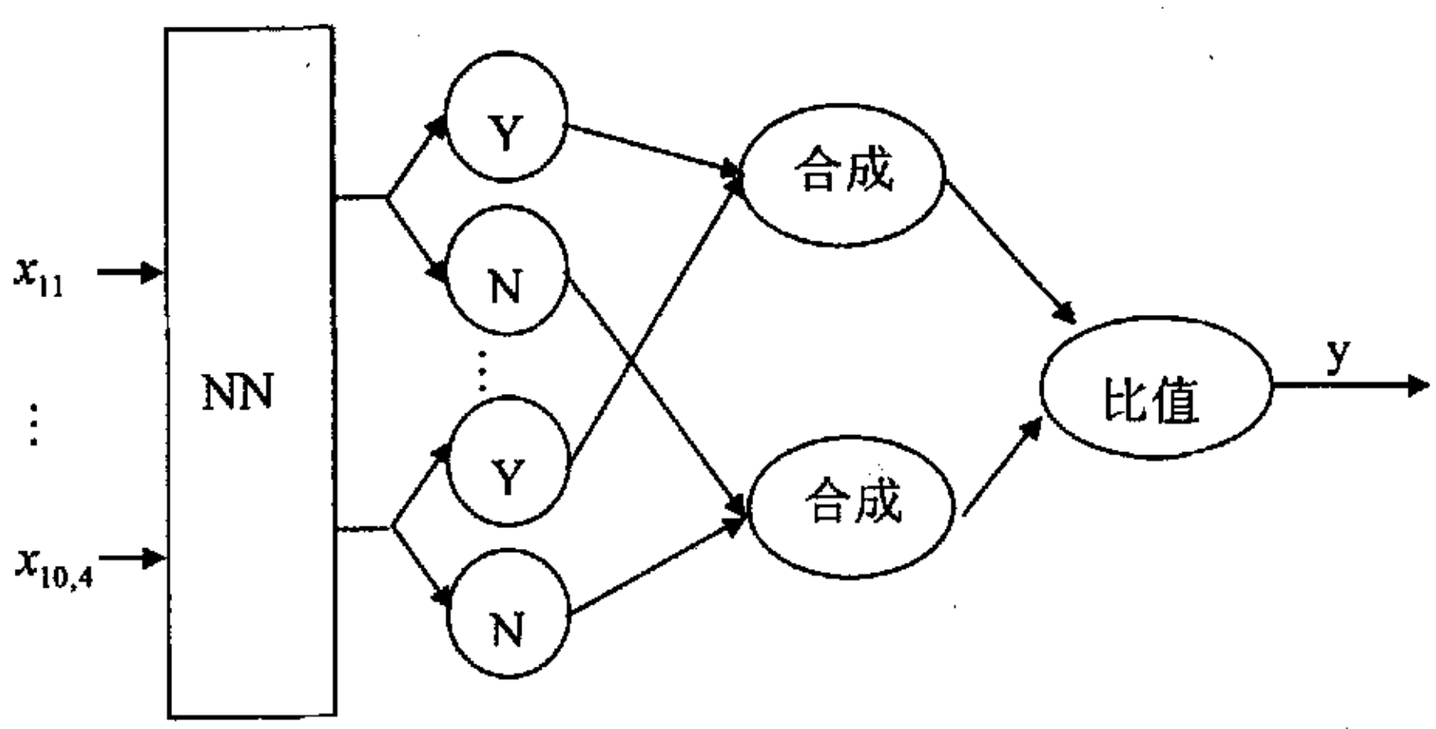


图 6-1 混合集成神经网络

1）40 维特征作为神经网 NN 的输入，同时神经网具有 12 个输出节点，其中，10 个输出节点拟和沙尘暴的典型模式，1 个拟和少站点沙尘暴，1 个拟和非

沙尘暴。

2）采用隶属度函数将神经网的输出进行模糊化，模糊化后节点输出的含义分别为属于强沙尘暴典型模式 1-10、少站点沙尘暴和非沙尘暴的程度。

3）通过合成，计算出该样本属于沙尘暴和非沙尘暴的隶属度，取其比值得到输出 y，则当 $y>=1$ 则为沙尘暴样本，反之为非沙尘暴。

2. 基于特征提取的集成神经网络

基于特征提取的集成神经网络的结构图如图 6-2 所示。

1）FNN1-FNN10 为 10 个完成特征提取的模糊神经网络，将出于同一典型模式的 4 个特征作为一模糊神经网络的输入，10 个模糊神经网提取到 10 个特征，这 10 个特征也分别代表了样本与 10 个典型模式的相似度。

2）用隶属度函数对这 10 个特征进行模糊化，模糊化后的节点输出为属于 10 个典型模式的隶属度。

3）通过求大与求小的合成，计算出该样本属于沙尘暴和非沙尘暴的隶属度，取其比值得到输出 y，则当 $y>=1$ 则为沙尘暴样本，反之为非沙尘暴。

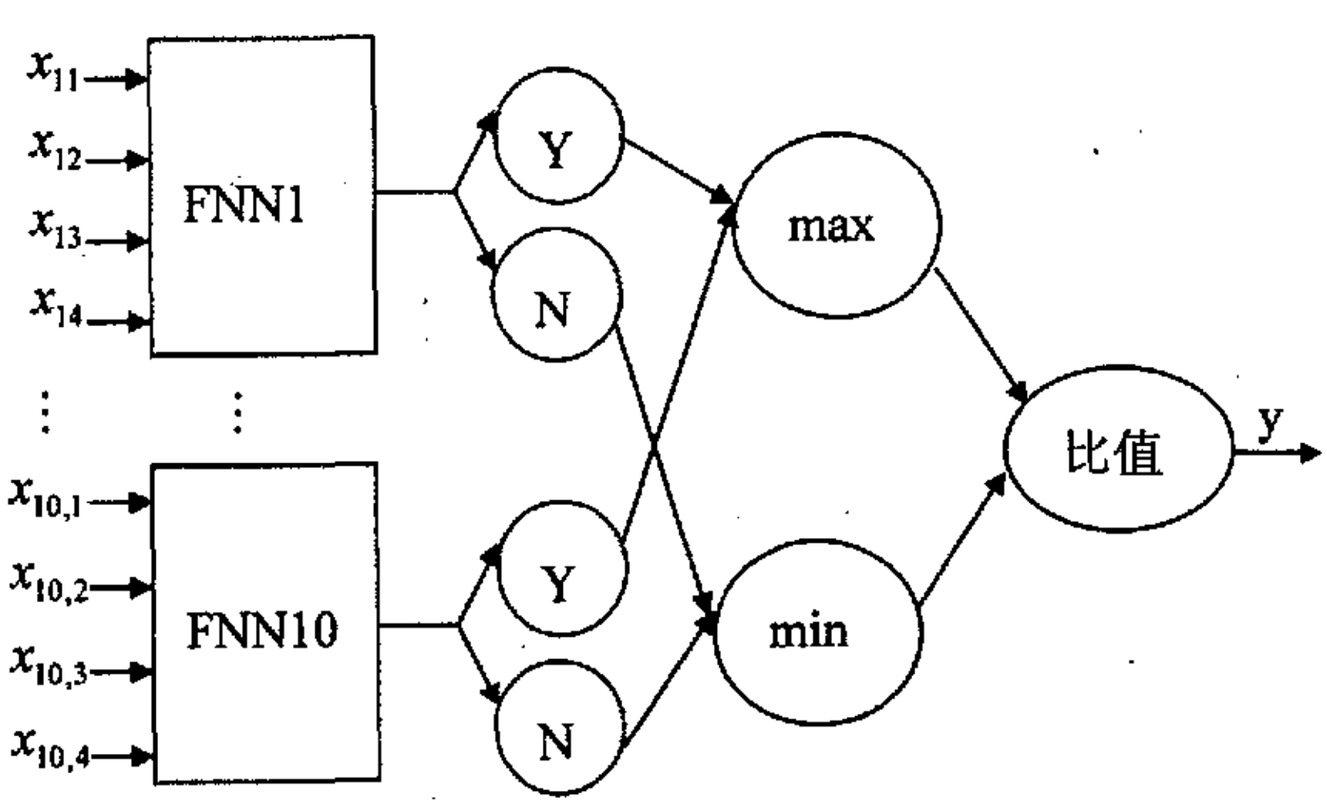


图 6-2 基于特征提取的集成神经网络

3. 基于聚类的集成神经网络

基于聚类的集成神经网络结构图如图 6-3 所示。

1）神经网 NN1-NN4 为 4 个实现聚类的神经网络，每个神经网对应一个物理场，其输入为出于该物理场的 10 个特征，输出为该物理场的类别信息。

2）神经网 NN1 对应高度值场，根据典型模式的聚类规则，从值相似角度，

高度场聚为 2 类；从形相似角度，高度场聚为 3 类，风场聚为 2 类，位温场聚为 2 类。故其输出模糊化后的节点数为 2，分别代表属于高度值场类 1 和类 2 的隶属度，同理，神经网 NN2、NN3、NN4 分别代表高度、位温和风的形场，输出模糊化后的节点数分别为 3、2、2 个。

3）计算属于沙尘暴各个典型模式的隶属度，经合成运算得到属于沙尘暴和非沙尘暴的隶属度。取其比值得到输出 y，则当 $y>=1$ 则为沙尘暴样本，反之为非沙尘暴。

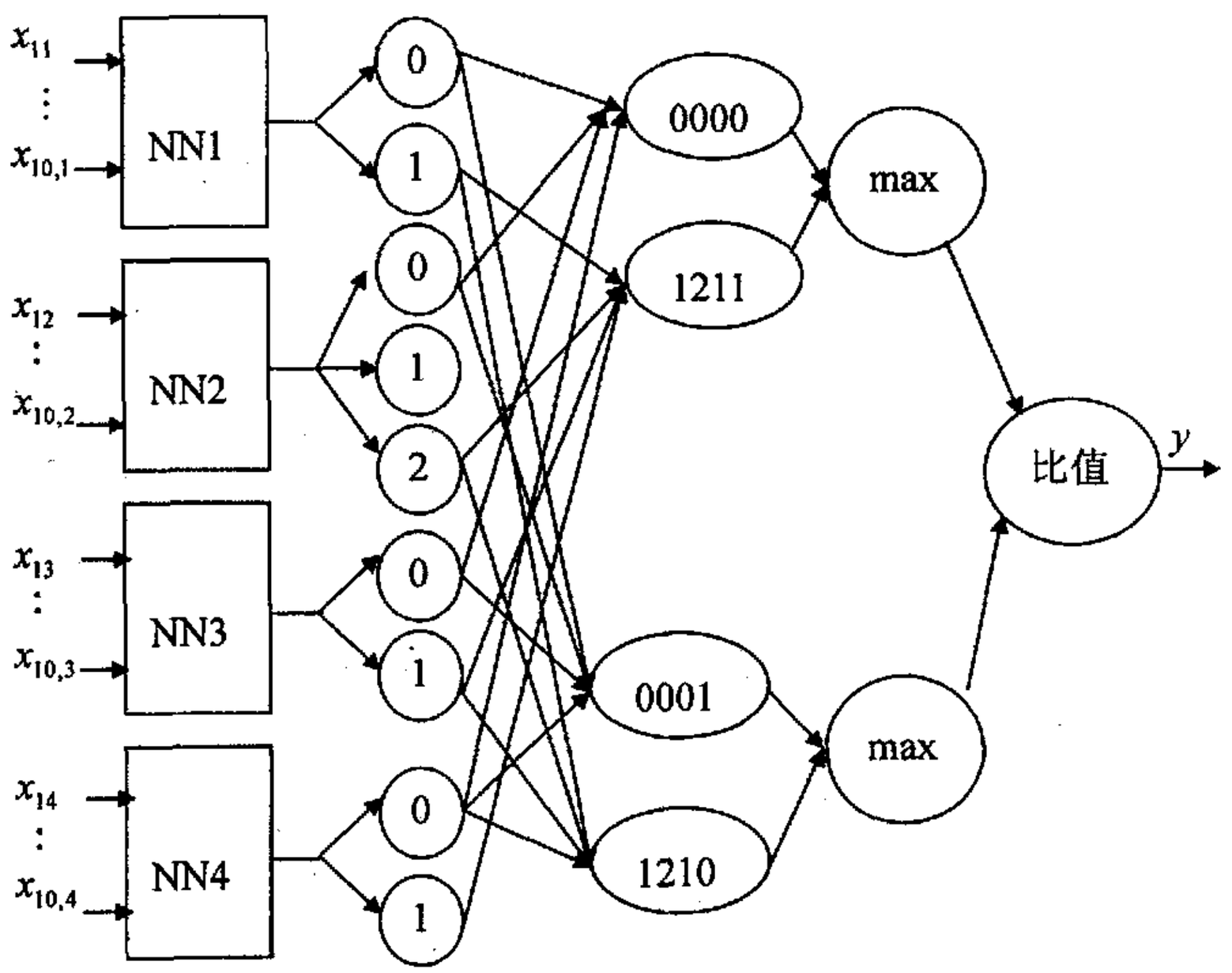


图 6-3 基于聚类的集成神经网

二、专家系统[44][45]与神经网络的结合（ES+NN）

专家系统是以专家经验性知识为基础建立的，以知识库和推理机为中心的智能软件系统。设计专家系统的基本思想是使计算机的工作过程竭尽全力的来模拟人类专家解决实际问题的工作过程，也就是模拟人类专家如何运用它的知识与经验来解决所要解决的问题的方法与步骤。

基于规则的专家系统在知识获取、并行推理、适应性学习、联想推理等方面较薄弱，而这正是人工神经网络的优势所在；人工神经网络的发展则受到系统

规模及推理过程自解释等方面的限制，但这些方面又是基于规则专家系统的特长。

将这两种方法相结合，如能达到优势互补，将能大大提高预报效果。根据侧重点不同，一般在神经网络与传统专家系统集成时，有三种模式[46]，即

1．神经网络支持专家系统。以传统的专家系统为主，以神经网络的有关技术为辅。比如对专家系统的知识和样例，通过神经网络自动获取知识，运用神经网络的并行推理技术以提高推理效率。

2．专家系统支持神经网络。以神经网络的有关技术为核心，建立相应领域的专家系统，采用专家系统的相关技术完成解释等方面的工作。

3．协同式的神经网络与专家系统。针对大的复杂问题，将其分解为若干子问题，针对每个子问题的特点，选择用神经网络或专家系统加以实现，在神经网络与专家系统之间建立一般耦合联系。

下面针对“沙尘暴预报”模型的建模过程中遇到的问题，提出以下几个改善方案。

1．串行相接法：框架如图 6-4

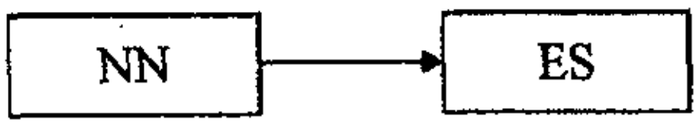


图 6-4 串行相接

系统中神经网络与专家系统各模块串联相接，独立工作，实现特定的功能。

例如，对于“沙尘暴预报”这个任务， 500hpa 高度、700hpa 风、850hpa 位温这三个因子都属于沙尘暴发生的大气背景，没有加入距离地面较近的因素，如地面气压等。用 NN 后再串接 ES 的系统来实现对沙尘暴的预报。

首先，通过组织训练样本集合的内容，使 NN 不再漏报，再把这样的输出送往存贮有其他专家知识的上层专家系统中进行再分析，排除空报。专家系统中的知识库包括地面气压、起沙速度等，若再把资料的范围扩展到卫星云图等，预报沙尘暴的多方面，并将相关经验引入知识库，预报的准确率有望得到长足的提高。

2．协同式

系统中的各个模块并列存在，相对独立工作。图 6-5 所示。

根据所用的资料性质决定进入 NN 还是 ES2 模块，然后由主专家系统 ES1 控制 NN 和 ES2 模块，并得到最终结果。

对于“沙尘暴预报”这个任务。如同串行方式，数值预报资料中的大气环流背景由 NN 归纳、记忆、判断，地面信息由 ES2 推断，其他信息根据具体情况决定进入 NN 或 ES2 模块，最后由 ES1 模块作出决策和判断。

协同式相对于串行式更为灵活。两种方式的专家系统部分都可以设置人机接口，根据系统使用的地区不同调整专家系统的知识库中的知识。

无论是协同式还是串行式，由于专家系统的加入，与神经网络相互协调配合，都使系统获得比单用神经网络更高的性能。

正如人是不断从感性认识上升到理性认识，但总是需要通过感性认识加入新鲜血液；一套 ES+NN 混合系统在应用中不断的成熟和完善，系统内的 NN 部分应逐渐转向 ES，但总会保留有部分 NN 解答 ES 难以处理的问题。例如，从神经网络中提取出规则就是 NN 转向 ES 的工作。

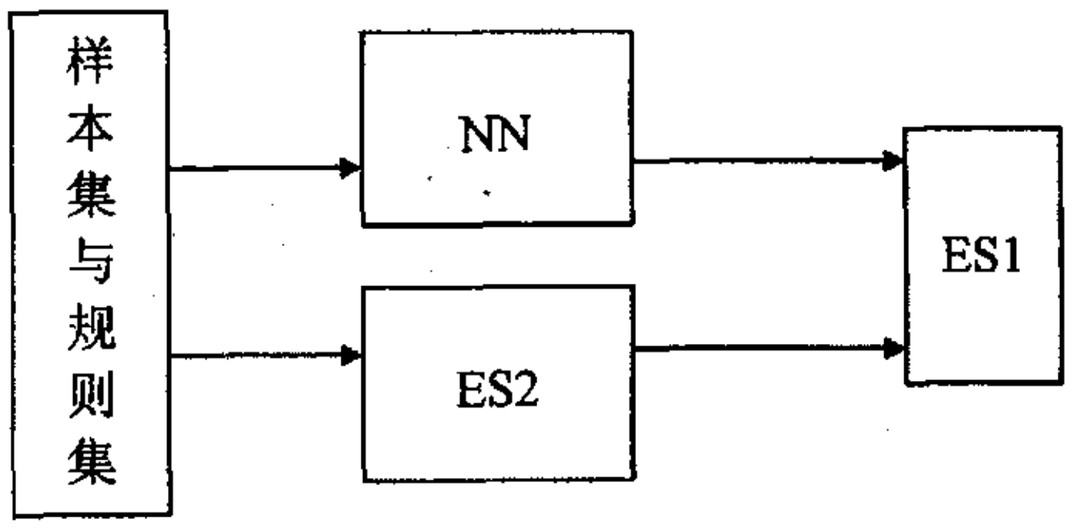


图 6-5 协同式

6.2.3 动态预报[47]

一、动态预报的模型

动态预报实现实时更新，及时让新出现的沙尘暴样本参与沙尘暴预报，并剔除陈旧的沙尘暴样本，以免对预报产生干扰，影响预报效果。沙尘暴的动态预报模型如图 6-6 所示：

其过程如下：

1．预报新出现的样本。

2，经实际天气验证分析新样本，将报对样本存入学习样本数据库，同时进行典型模式分析，存入典型模式样本库，将报错样本存入非典型样本数据库。

3．重新计算典型模式的中心场，进行特征提取并训练神经网。

4．利用神经网和非典型样本库中的样本进行下一轮的预报。

5．分析非典型样本库中的样本，若出现典型样本，则从非典型样本数据库

中删除，进入学习样本数据库及典型模式数据库。

6. 及时剔除陈旧样本。

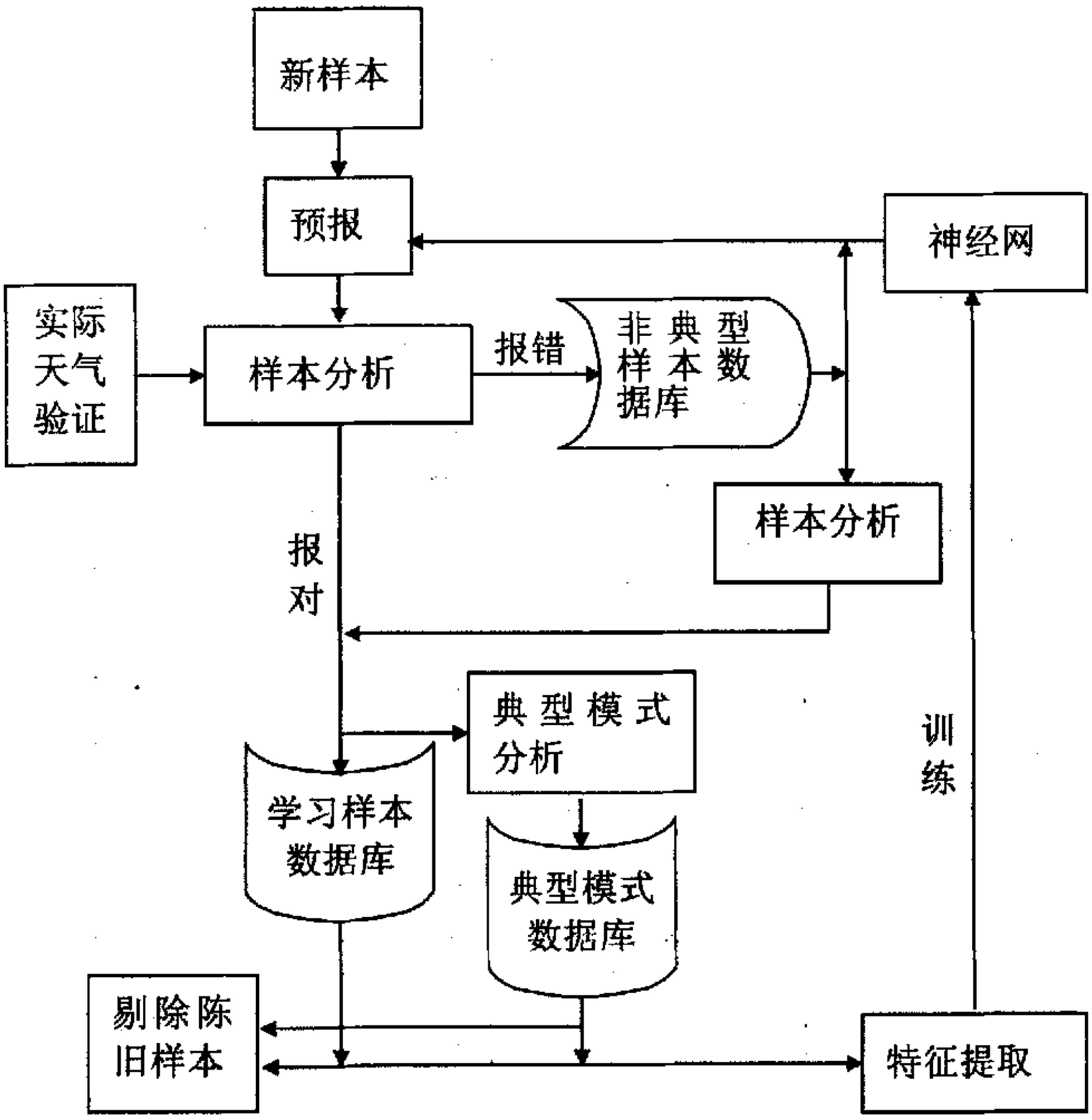


图 6-6 动态预报模型

二、数据库的更新

动态预报中，主要涉及三个数据库的更新，即典型模式数据库、学习样本数据库和非典型样本数据库。

1. 典型模式数据库的更新

1）支持向量机[48][49][50][51][52]

支持向量机是 Vapnik 等人根据统计学习理论提出的一种新的机器学习方法，它以结构风险最小化原则为理论基础，通过适当选择函数子集及该子集中的判别函数使学习机的实际风险达到最小，保证了通过有限训练样本得到的小误差分类器对独立测试集的测试误差仍然小，得到一个具有最优分类能力和推

广泛化能力的学习机。

2）典型模式分类

沙尘暴属于小概率事件，站点为 10 及 10 以上的强沙尘暴更少，其中沙尘暴典型模式 2 中的样本总量仅为 9 个，而样本数最多的模式 10 也仅有 37 个，支持向量机方法对小样本、非线性和高维特征具有很好的分类性能。特别适合于沙尘暴模式类的分类问题。

3）典型模式类数据库的更新

典型模式的沙尘暴样本是计算中心场的依据，也是样本 40 个特征的依据。随着时间的推移，一些作为参考的典型沙尘暴模式或许不再出现，相应的，新的典型模式或许产生，故及时更新典型模式样本或适当调整典型模式的类型对中心场的计算、样本特征的提取及对沙尘暴的预报结果将会产生重大的影响。

2. 学习样本数据库的更新

神经网络中记忆的规律，是从已知样本和已知变量组成的数据集中学习训练得到的。学习样本的代表性和全面性对提高神经网的识别能力极为重要，一些陈旧的沙尘暴或非沙尘暴类型，对模型的贡献率下降，影响神经网识别能力的提高，相应地，也应使神经网对一些新出现的类型具有识别能力，及时更新学习样本数据库，使神经网始终具有最新类型的识别能力，对沙尘暴的预报率也会产生很大的影响。

3. 非典型样本数据库的更新

非典型样本的隶属度调整方法在提高预报率中起了一定的作用，而随着时间的变化将会出现一些新的非典型样本，或非典型样本中，一些样本的类型随着出现个例的逐渐增加，将不再属于非典型样本，这就需要及时更新非典型样本数据库。

总之，复杂的天气变化使对沙尘暴的预报含有许多不确定的因素，目前还没有明确的规律可循，这就需要全方位的试探和研究，通过特征重构、模型优化和动态预报，只要特征合理、结构最优、样本具有代表性，定能进一步挖掘整个预报系统的潜力，实现沙尘暴的准确预报。

结束语

天气系统是一个包含有多种空间尺度、时间尺度和多种天气要素的多维复杂的系统，气象预报是一个很复杂的问题。近年来，人工智能和模式识别技术越来越多的应用于气象预报问题。本文主要运用模糊神经网络技术就“沙尘暴预报”的客观建模问题展开研究。

本文利用统计检验技术系统分析了沙尘暴样本的 40 个特征，并从沙尘暴和非沙尘暴样本在四个基本物理场上反映出的特点出发，选出用于建模的代表性样本，通过对主成分分析结果的研究，设计出特征综合方案，形成兼顾各个主成分的“总量特征”。利用这种特征综合方法，实现了对 40 维样本进行再次的特征提取，最终形成 10 维更为合理的建模样本。然后，展开基于模糊神经网络的沙尘暴建模方案研究，并从网络的拓扑、训练参数、样本集合等方面对模型进行优化，使对沙尘暴的预报达到了一定的预报效果。

本文对基于模糊神经网的预报结果展开进一步研究，指出影响预报准确率的主要因素是训练样本中含有为数较多的非典型性样本。于是，通过聚类建立非典型样本区，再构建基于非典型样本的统计模型，并设计出一种兼顾模糊神经网预报结果和样本非典型程度的沙尘暴隶属度调整方案，使建立在模糊神经网预报（1 级）结果之上的统计模型再预报（2 级），在基本不影响 1 级报对率的前提下，纠正了相当比例的报错样本，与文献[5]的基于 40 个特征的神经网络相比，本文提出的模糊神经网与统计模型的联合预报方案，使沙尘暴的报对率从 60%提高到 73.3%以上，*CSI* 值也由 25.9%提高到 38.7%，预报效果得到明显改善。

最后对建立预报系统中遇到的问题作了总结，并提出了特征重构、集成神经网络、神经网络与专家系统结合的解决方案，同时就沙尘暴的动态预报问题提出了实现框架。

沙尘暴预报是气象领域的一个崭新课题，实现对沙尘暴的有效预报需要多方的不懈努力和探索。消除沙尘暴危害最根本的办法还是有效的防止沙尘暴的发生，这需要全国，乃至全球人民的共同努力。

随着研究的深入，越发感觉到自己知识的浅薄，所以整个研究过程也是我不断学习和充实自己的过程。因时间和水平所限，特别是气象知识的不足，论文中尚有许多待完善的地方，希望大家提出宝贵意见。

参考文献

[1] 赵兴梁,甘肃特大沙尘暴的危害与对策,中国沙漠，1993，13(3):1-7

[2] 王式功 董光荣，沙尘暴研究的进展，中国沙漠，2000，20(4): 349~356

[3] 夏训诚 杨根生，中国西北地区沙尘暴灾害及防治，气象，1996.10

[4] 许东蓓，西北地区 4.18 强沙尘暴、浮尘天气成因分析，甘肃气象，1999.2

[5] 岳斌，基于神经网络的沙尘暴预报模型的研究与应用：[硕士学位论文]，天津；天津大学，2002

[6] 高庆先 李令军，我国春季沙尘暴的研究，中国环境科学，2000，20(6):495~500

[7] Zhang Guoping,Liu Jiyuan,Zhang Zenxiang, Remote sensing investigation of the main sand-supplying areas of dust storm hitting northern China, Geoscience and Remote Sensing Symposium,2001，2001,5: 2097~2099

[8] 张承福，人工神经网络在天气预报中的应用研究，气象人工智能专辑（二），1994

[9] 傅京孙，模式识别应用(程民德，石青云译)，北京：北京大学出版社，1987

[10]边肇祺 张学工，模式识别，北京：清华大学出版社，1999

[11]王众托，系统工程引论，北京：电子工业出版社，1991

[12]Castillo O.,Melin P.,Simulation and forecasting complex economic time series using neural networks and fuzzy logic,NeuralNetworks, 2001.Proceedings.IJCNN'01. International Joint Conference on,2001, 3:1805-1810

[13]Meesad P.,Yen G.G.,A neuro fuzzy network and its application to machine health monitoring Neural Networks,2001.Proceedings. IJCNN'01.International Joint Conference on , 2001,3:2298-2303

[14]Garliauskas A.,Fuzzy and chaotic neuro-network modeling, Fuzzy Systems, 2001. The 10th IEEE International Conference on , 2001,1: 428 -431

[15]邵军力，人工智能基础，北京： 电子工业出版社，1993

[16]冯健翔，人工智能及其航天应用概论（上）-广义人工智能基础研究，北京：宇航出版社 1999

[17]郝为，基于计算智能的多模型气象综合预报：[硕士学位论文]，天津：天津大学，2000

[18]丛爽，典型人工神经网络的结构、功能及其在智能系统中的应用，信息与控制，2001，30(2):97-103

[19]汪镭 周围兴 吴启迪，人工神经网络理论在控制领域中的应用综述，同济大学学报，2001，29(3):357-361

[20]刘良江 侯拥和，模糊神经网络技术的发展与应用，矿冶工程，2002，22(1):66-68

[21]黄德双，神经网络模式识别系统理论，北京：电子工业出版社，1996.5

[22]Mitra S.,Mitra P.,Pal S.K.,Data mining in soft computing framework: a survey,Neural Networks, IEEE Transactions on , 2002,13: 3-14

[23]阎平凡，自动化学报，对多层前馈神经网络研究的几点看法，1997，23(1):129-135

[24]Hawley A., Fuzzy nets: fuzzy logic and neural networks , Neural Networks,2001.Proceedings.IJCNN'01. International Joint Conference on,2001,1: 544 -546

[25]D.Rutkowska,Y.Hayashi, Neuro-Fuzzy Systems Approaches, Journal of advanced Computational Intelligence,1999, 3(3): 177-185

[26]Zadeh L.A, The Concept of a Linguistic Variable and its Application to Approximate Reasoning,Information Sciences,1975(8):43-80

[27]E.H.Mamdani,S.Assi1ian,An experiment in linguistic synthesis with a fuzzy logic controller,Int J Manmachine Studies,1975,7(1):1-13

[28]Takagi T.,Sugeno M.,Fuzzy identification of systems and its applications to modeling and control,IEEE Trans on SMC,1985,15(1):116-132

[29]Ying-Chung Wang,Chiang-Ju Chien,Ching-Cheng Teng, Takagi-Sugeno recurrent fuzzy neural networks for identification and control of dynamic systems,Fuzzy Systems, 2001. The 10th IEEE International Conference on, 2001,1:537-540

[30]Wan-Jui Lee,Chen-Sen Ouyang,Shie-Jue Lee, Constructing neuro-fuzzy systems with TSK fuzzy rules and hybrid SVD-based learning ,Fuzzy Systems,2002.FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on , 2002,2: 1174 –1179

[31]朱文彪 孙增析 常佑，模糊推理与复杂过程建模，1999 年中国智能自动化学术会议论文集（上册），北京：清华大学出版社，1999，455-450

[32]张秀玲，神经网络自适应控制的研究进展及展望，工业仪表与自动化装置，2002，(1):10-14

[33]李学桥 马莉，神经网络.工程应用，重庆：重庆大学出版社 1996.8

[34]谢政 刘卫华，赋模糊权的拟强连通有向网络中的最佳树形图，模糊系统与数学，1997，11(4):86-90

[35]H.Ishibuchi K.kwan, H.tanaka, A learning algorithm of fuzzy neural networks with training fuzzy weights. Fuzzy Sets and System ,1995:217-293

[36]Hongli Lei,Jianbang Zhang,Dianzhi Zhang,A new algorithm of fuzzy neural networks with multiple form fuzzy weights,Intelligent Control and Automation, 2002. Proceedings of the 4th World Congress on ,2002,4:3252-3255

[37]乔志骏 刘其真 易维列等， 一个基于模糊神经网络的数据逼近和泛化建模方法，模式识别与人工智能，2001，14(2):253-256

[38]Wan-Jui Lee,Chen-Sen Ouyang, Shie-Jue Lee, Constructing neuro-fuzzy systems with TSK fuzzy rules and hybrid SVD-based learning Fuzzy Systems, 2002. FUZZ-IEEE'02.Proceedings of the 2002 IEEE International Conference on,2002,2: 1174 -1179

[39]邢松寅 王士同，基于 Pi-Sigma 神经网络的高木-关野模糊系统用于数据关联计算的建模，电子科学学刊，1999，21(1):72-77

[40]丁士晟，多元分析方法及其应用，吉林：吉林人民出版社，1977

[41]白雪梅 赵松山,对主成分分析综合评价方法若干问题的探讨，统计研究，1995，(6):47~51

[42]Abhijit S. Pandya,Robert B.Macy,神经网络模式识别及其实现（徐勇 荆涛等译），北京：电子工业出版社，1999.6

[43]Yan-Qing Zhang,F'u-lai Chung,A fuzzy neural network tree with heuristic backpropagation learning,Neural Networks, 2002.IJCNN '02.Proceedings of the 2002 International Joint Conference on,2001,1: 553 –558

[44]田胜丰 黄厚宽，人工智能与知识工程，北京：中国铁道出版社，1999

[45]赵瑞清，专家系统原理，北京：气象出版社，1987

[46]陈肇乾，基于神经网络的混合型天气预报系统，中国人工智能学会第八届年会论文，1994.11

[47]Lewis H.W.,Intelligent hybrid load forecasting system for an electric power company, Soft Computing in Industrial Applications,2001.SMCia/01. Proceedings of the 2001 IEEE Mountain Workshop on , 2001: 23 -27

[48]都云琪 肖诗斌，基于支持向量机的中文文本自动分类研究，计算机工程，2002，28(11):137-139

[49]任力安 何清 史忠植,HSC 分类法及其在海量数据分类中的应用，电子学报,2002(12):1870-1872

[50]Challa S.,Palaniswami M.,Shilton A.,Distributed data fusion using support vector machines,Information Fusion, 2002.Proceedings of the Fifth International Conference on , 2002,2:881 -885

[51]Xin-Wei Fan,Shu-Xin Du,Tie-Jun Wu,Noise-immune SVM classifier with uneven class sizes in wastewater treatment process,Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on ,2002,3: 1189-1193

[52]Chapelle O.,Haffner P.,Vapnik V.N.,Support vector machines for histogram-based image classification,Neural Networks,IEEE Transactions on ,1999,10: 1055 –1064

发表论文和参加科研情况说明

一、论文发表及录用情况

1.《一种大规模样本数据的特征提取方法》　　天津轻工业学院学报

针对沙尘暴样本数据的特点，成功地在每个样本的几百维数据中提取到 40 个特征。同时，提出了一种建立在主成分分析基础上的特征综合方法，有效的降维和消除特征间的相关性，并协助完成关于多维特征的类间差异检验。

2.《基于格点场数据源的沙尘暴样本的特征提取》　　计算机应用研究

沙尘暴样本的特征提取是建立沙尘暴预报模型的前提。从数据源特点出发，根据专家经验依次通过聚类分析、建立典型模式类、计算中心场，再以样本与中心场的距离作为样本的特征。成功地在每个样本的几百个数据中提取到 40 个特征。通过对模式类的类型特点分析及对提取特征的统计检验证明了该方法的可行性和提取特征的有效性。

二、参加科研情况

1．参与研制教师工作量统计系统，利用 PowerBuilder 数据库知识，实现教师本科授课与实验教学、研究生授课与指导研究生工作量的统计，同时具有个人基本信息的维护及数据库整体维护的功能。

2．基于模糊神经网的沙尘暴预报，利用模式识别、人工智能、模糊理论等，在 C++编程平台上建立沙尘暴预报模型，成功地解决了数据降维、特征提取及综合等较难的课题，模型具有较高预报准确率。

致谢

本论文是在导师王萍教授和林孔元教授的悉心指导下完成的，在两年多的研究生学习过程中，始终得到两位老师的热切关怀与悉心指导，二位老师对模式识别与智能系统学科深刻的认识，王老师对待事业的认真、严谨的敬业精神，林老师在大方向的指引和开放的思维方式，二位导师的言传身授，都使我获益非浅，是我终生的精神财富。在此对王老师、林老师表示衷心的感谢。

感谢课题组的刘正光教授、路志英副教授、杨正瓴副教授等给我的热心指点和帮助。

最后还要向帮助过我的课题组的各位同学表示真诚的谢意。