天津大学

硕士学位论文

基于神经网络的沙尘暴预报模型的研究与应用

姓名：岳斌

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

2002.1.1

摘　要

随着人工智能和模式识别技术的日益发展，它们越来越多的应用到各个具体领域来解决各种各样的问题。本文所作的研究工作，正是围绕“沙尘暴预报”这一具体问题展开的。

本文在讨论模式建模方法的同时，对人工神经网络技术做了重点介绍。

本文在分析和处理原始资料的基础上，通过聚类分析并参考专家经验及先验知识，成功地构建了建模样本特征，使借助 855 维数据描述的原始样本转变为由 40 个有效特征描述的建模样本，为模式建模工作奠定了坚实的基础；本文以我国近 20 年沙尘暴多发季节的正反两类天气情况作为建模样本，选用多层前向神经网络，训练沙尘暴预报模型，通过对 BP 网络预报模型在网络的拓扑、训练参数、样本集合等几个方面的优化工作，有效提高了模型的预报准确率，并进一步探讨了网络权值中隐含的信息可挖掘性。

本文最后对建立预报系统中遇到的问题作了总结，并提出了多网络决策系统、专家系统和神经网络结合的解决方案，对预报系统的业务化提出了实现框架。

关键词：模式建模 神经网络 沙尘暴预报 专家系统

ABSTRACT

With Artificial Intelligence and Pattern Recognition's development, they have been widely applied in various fields. The implementation of a dust storm forecast system is presented by those techniques in this paper.

While introducing Pattern Modeling, Artificial Neural Network (ANN) is discussed in detail.

The complete forecast system is consisted of data analyzing and processing, sample feature constructing, modeling and model optimizing. Data is analyzed and processed with the method of " clustering first, and then constructing feature", which changes the original samples described by 855 dimension data as the modeling samples described by 40 dimension features. Using the multi-layer forward ANN as the forecasting model, the model optimizing is discussed in several aspects such as ANN topology, parameters training and compilation of the samples, which improves effectively the exact rate of the forecasting model. The possibility of the mining of information concealed in the weights of network is also discussed.

At the end of the paper, the problems in the forecast system are analyzed. A multi-NN resolution and an ES-NN resolution are put forward and the operable realization frame of forecast system is presented.

Key words:

Pattern modeling

Artificial Neural Network

Dust storm forecast

Expert System

独创性声明

本人论文是我个人在导师的指导下取得的研究成果，尽我所知，除特别的加以标注的和致谢的地方外，论文中不包含其他人已经发表或已撰写过的研究成果，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作的任何贡献均已在论文中作了明确的说明并表示谢意。

签名：　　日期：

关于与论文相关研究成果的归属权和论文使用授权的说明

本人完全了解天津大学有关与论文相关研究成果的归属权以及保留、使用学位论文的规定，即：本论文的所有相关成果的归属权由天津大学界定，无论在学期间或毕业后，在发表时需将天津大学冠于首位;学校有权保留送交论文的复印件，允许论文被查阅或借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他的复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：　　导师签名：　　日期：

第一章 绪论

§1.1 人工智能与天气预报

“人类正是有着一种对万物的控制欲望，才是前进、发展着的”。[9]纳米技术的发展，试管婴儿的诞生，克隆羊的培育成功，气象部门天气预测准确率的提高，……，这些都是近年来人类对自身以及周围的万物的控制上取得的一步步进展，科学技术在发展、社会在进步。

人工智能正是随着人类对其本身的深入了解而产生并随之发展的学科。

人工智能是研究怎样使计算机来模仿人脑所从事的推理、学习、思考、规划等思维活动，来解决需人类专家才能处理的复杂问题。如医疗诊断、气象预报等。[1] 近年来随着人工神经网络、专家系统、模式识别、信息科学、认知科学、计算机科学、人脑神经网络结构和模糊逻辑等有关理论的发展，人工智能技术在各领域得到了越来越多的应用。人工智能技术借助的主要工具是计算机，它能实现人的某些功能，并因计算机自身的特点（如：计算能力强、存储量大、耐疲劳等），某些工作比人亲手来做更为胜任。

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巨系统，气象预报问题是一个很复杂的问题。在本世纪四十年代，N.维纳在其名著《控制论》中对天气系统的复杂性有过这样的生动描述：“多少世纪以来，人们就能够预测比较常见的天文现象，但是要精确地预测明天的天气一般是不容易的……”“在气象学方面……，要准确地记录它们的初始位置和初始速度是绝对不可能的……，所谓‘云’、‘温度’等术语都不是指的个别的物理状态，而是指的许多可能状态的分布。”“运用牛顿定律或其他因果体系，我们对未来任何时刻所能做的预测只是系统中若干参数的几率分布，甚至这种预测性还会随着时间的增长而消失。”如此精辟地说明了天气现象的随机性特征，它表明，对天气系统这样复杂的系统的行为预报，需要寻求更有效的途径。另外，随着气象科技的日益进步，气象信息的数量与种类迅速增长，如何存储海量的数据，并对其实施有效的管理，进而从中有效地检索和利用信息，也需要有新的工具支撑。

人工智能以认知科学为基础，以其独特的角度给气象预报提供新的途径和方

法，这便产生了气象人工智能这门边缘交叉学科。

人工智能技术引入气象领域已有近 20 年的历史，这期间专家系统技术用于气象灾害的预测，神经网络系统在暴雨预报中的试验，利用数学形态学和模式识别技术对卫星数值云图的分析、识别等都取得了一定的进展。[11][35]

天津大学模式识别与智能系统学科组近年来致力于人工智能技术引入气象领域的研究，在“集成化智能信息处理系统”[10]、“天气过程理解系统”[48][50]、“卫星云图处理与识别”[52][53][54][55]、“气象智能数据库系统”[49][51]等项目的研究上取得了重要成果。

但对于基于数值预报资料的沙尘暴的预测，在我国的气象领域，还是个崭新的课题。

§1.2 关于沙尘暴

沙尘暴的危害

近年来，在我国沙尘暴、扬沙和浮尘天气频繁的发生，严重干扰和影响着人们的正常生活，对社会经济和环境均造成一定程度的危害。

在我国，沙尘暴多发生在北方地区的冬春季节，我国西部有些地区由于乱垦乱伐，过度放牧，使土地植被日益稀少，沙土裸露，荒漠化问题日益严重，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，有时甚至飘洋越海，使韩国和日本也不能幸免。当前正值中国刚加入 WTO（世界贸易组织）和祖国正在实施西部大开发的良好时机，世界关注中国，中国比以往更加重视西部，因而沙尘暴问题受到了不仅是气象部门，更是引起了各级政府和广大群众的高度重视。为此，气象部门把沙尘暴列为重点的预报和服务对象。[21]

沙尘暴的成因及沙尘暴的可预报性

专家研究指出，沙尘暴形成的基本条件，一是大风，二是地面上裸露沙尘物质，三是不稳定的空气，即强风因子、沙源因子和热力不稳定因子。三者同步出现，便可能产生沙尘暴。三因素中强风是起沙尘的动力，丰富的沙尘源是形成沙尘暴的物质基础，而不稳定的空气使沙尘卷扬的更高，乃是非常重要的热力条件。

因此可以说沙尘暴是特定的气象和地理条件结合的产物。[26][27]

在对沙尘暴进行预报时，必须综合考虑上述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，只要冷空气经过的下垫面有丰富的沙源，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。

沙尘暴的预报方法

现在气象领域对沙尘暴研究取得的进展主要是在对沙尘暴发生时基本特征观测和分析，沙尘暴成因的分析，沙尘暴的数值模拟与沙尘输送这几个方面。[22]

对沙尘暴的基本特征观测和分析包括气象要素变化[28]、卫星云图与光学特性[24][25]、沙尘气溶胶的物理化学及辐射特性。

沙尘暴的数值模拟，是研究沙尘暴形成的物理机制的重要手段之一。

对沙尘暴成因的研究主要在宏观条件、主要环流形势和影响系统、经纬向环流调整、冷锋活动、低空东风急流、中尺度系统几方面。

所做的这些工作对沙尘暴的预报都是必须且有益的，但是“现阶段中央气象台还没有一个行之有效的预报方法和技术，主要仍是基于传统的天气学分析和个人对数值预报产品的理解上，凭预报员的经验来进行预报，急需建立一套新的沙尘暴预报方法和预报流程。各地方台站也都正在对此作着研究。”[21]

在此基础上国家气象中心提出沙尘暴短期预报思路：首先是天气学分析，在确定出现有利于沙尘暴发生的形式的情况下，通过数值模式求算沙尘暴指数，进而判定沙尘暴发生的可能性。

沙尘暴的准确预报是我国迫切需要的。特别是能够及时准确预报出来强沙尘暴，可以使人们在沙尘暴到来时能够采取一定的防范措施，避免出现象“甘肃93·5·5黑风暴”死亡85人，失踪、受伤数百人，直接经济损失5.6亿元的惨剧，把沙尘暴的危害降低到最小。当然，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，乃至全球人民的长期共同努力。

§1.3 沙尘暴预报问题分析

天气预报是根据过去或现在观测得到的某些物理量的信息，对未来的天气形

势做出判断。

沙尘暴天气的预报就是针对现在某区域的天气特征，根据观测、总结得到的经验，对某区域未来某时段内有无沙尘暴乃至沙尘暴的可能程度做出判断。

气象部门多年来收集、整理了大量的沙尘暴日和非沙尘暴日的历史资料，以期在这大量的资料中挖掘出可用的信息。借此反映沙尘暴与非沙尘暴的区别（即沙尘暴的特征），用于判断（预报）沙尘暴是否发生。

预报研究过程如图 1-1：

图 1-1

“一般来说，模式识别问题指的是对一系列过程或事件的分类与描述。要加以分类的一系列过程或事件可以是一系列物理的对象，或者是一些比较抽象的如心理状态等。具有某些相类似的性质的过程或事件就分为一类。”[3] 所有的沙尘暴日看作一个总体，称做沙尘暴类，所有的非沙尘暴日则为另一类。沙尘暴天气的预报实际上就是根据观测到的信息，判断要预报时刻和地区可能发生的天气事件应归入沙尘暴类还是非沙尘暴类。

可见沙尘暴天气的预报是典型的模式识别过程。

§1.4 本文的主要工作

本文所作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，本文任务是研制基于数值预报场的沙尘暴短期预报模型，目标是建立的沙尘暴短期预报模型具有一定实用价值。

本文首先探讨了模式建模方法，并对本文中使用的人工神经网络技术做了重点介绍。

第三章介绍了沙尘暴预报模型建模样本的形成过程，通过对原始资料的信息提取和聚类分析，构建样本特征，最终形成可用的建模样本。

第四章详细介绍了所选用的建模工具——多层前向神经网络的原理、几何解释，以及训练过程选用的 BP 学习算法，形成了沙尘暴预报模型建模框架。

第五章介绍了围绕提高模型的预报效果作的工作，从网络的拓扑、训练参数、样本集合等几个角度详细介绍了 BP 网络模型的优化过程，并进一步探讨了网络权值中隐含信息的可挖掘性。

第六章总结预报模型研制全过程，对遇到的问题作了进一步探讨，提出多网络决策系统、专家系统和神经网络结合的解决方案，并对系统的业务化提出了实现框架。

第二章 模式建模

模式建模：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。

在很多时候，特别是在气象问题中，运用模式识别的观点去分析问题时，可定义具体的过程或事件为相应的模式类，而各模式类类间的差异规律常含在大量的样本之中，隐藏在海量的数据之中，需要挖掘出来。

§2.1 传统模式建模方法

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的结果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。[3][4]

主要方法：

贝叶斯（Bayes）准则

用这种方法进行分类时要求：

（1） 各类别总体的概率分布是已知的；

（2） 要决策分类的类别数是一定的。

在连续情况下，假设对要识别的物理对象有 d 种特征观察量 $x_1, x_2, \cdots, x_d$，这些特征的所有可能的取值范围构成了 d 维特征空间，称 $x=[x_1, x_2, \cdots, x_d]^T$ 为 d 维特征向量。要研究的分类问题有 c 个类别，各类别状态用 ω_i 来表示，$i=1,2,\cdots,c$；对应于各个类别 ω_i 出现的先验概率 $P(\omega_i)$ 及类条件概率密度函数 $p(x|\omega_i)$ 是已知的或可根据样本统计估计出来。后验概率可由贝叶斯公式确定，即

$$P(\omega_j|x)=\frac{p(x|\omega_j)P(\omega_j)}{p(x)}=\frac{p(x|\omega_j)P(\omega_j)}{\sum_{j=1}^{c}p(x|\omega_j)P(\omega_j)} \qquad 2\text{-}1\text{-}1$$

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。

由式 2-1-1 和具体问题选用的判别规则结合就可进行分类判决。

费歇（Fisher）准则 Fisher 提出的判别分析方法并不对样本 p 维空间 R 直接划分，而是把 p 维空间中的样本投影在较低维空间中去，再在低维空间中进行分类。所以，Fisher 判别分析的中心思想是找出最有利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。

结构建模法[3]

在一些模式分类问题中，描述每一具体模式的结构信息是十分重要的。例如图片、语音、景物等问题的识别。

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。

为了表示每一模式的层次结构，可以把对模式结构的描述类比于形式语言中的语句。由基元、子模式以不同方式构成的模式的描述，恰似由字构成词，由词构成句子的过程。原则上应提取远比模式本身简单的子模式，即“模式基元”。用作模式的结构描述的语言包括两部分：即模式基元和对基元的合成操作规则，这种语言就被称为“模式描述语言”。对基元作合成操作以构成模式的规则，就叫做语法。当模式中的每一基元被辨认以后，识别过程就可通过语法分析来实现，亦即对描述待识模式的“句子”进行分析，看它是否遵守指定的语法。与此同时，语法分析还产生出该句子的结构描述，通常是树状的结构描述。在语法分析阶段，系统应对已完成描述的模式作语法检查，以判定它是按何种语法结合成的，从而完成待识模式的分类。参见图 2-1。

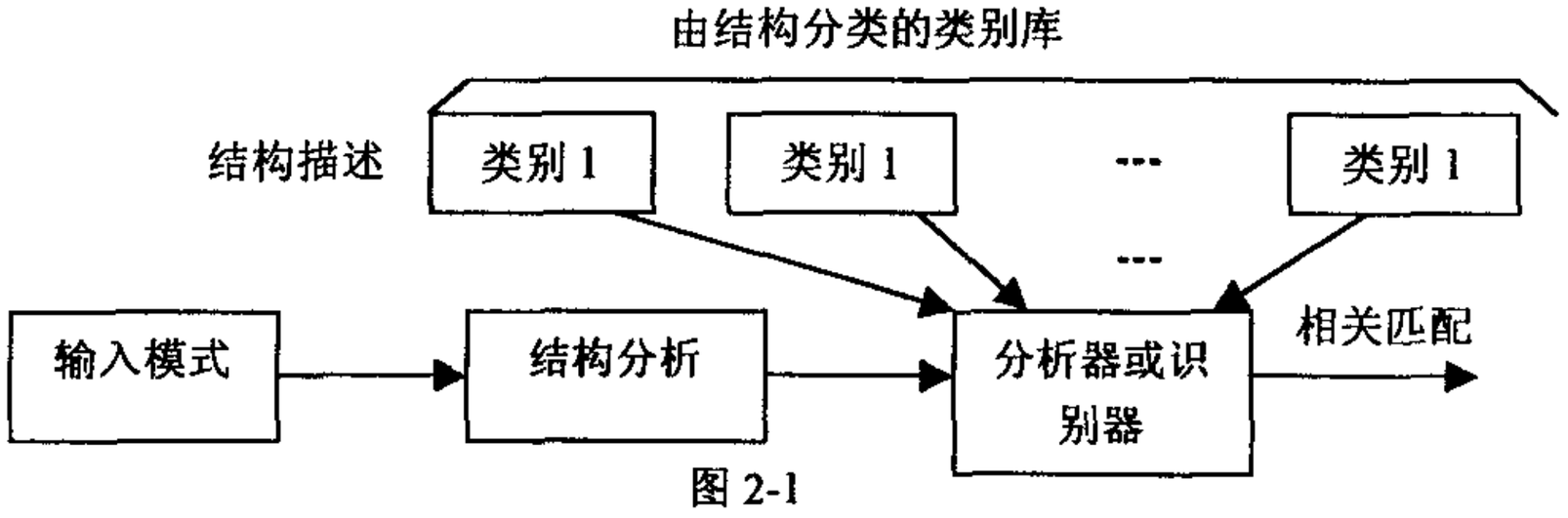


图 2-1

传统的模式建模过程

模式识别系统都由两个过程所组成，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）确定出一套分类判别规则，使得按这套分类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。

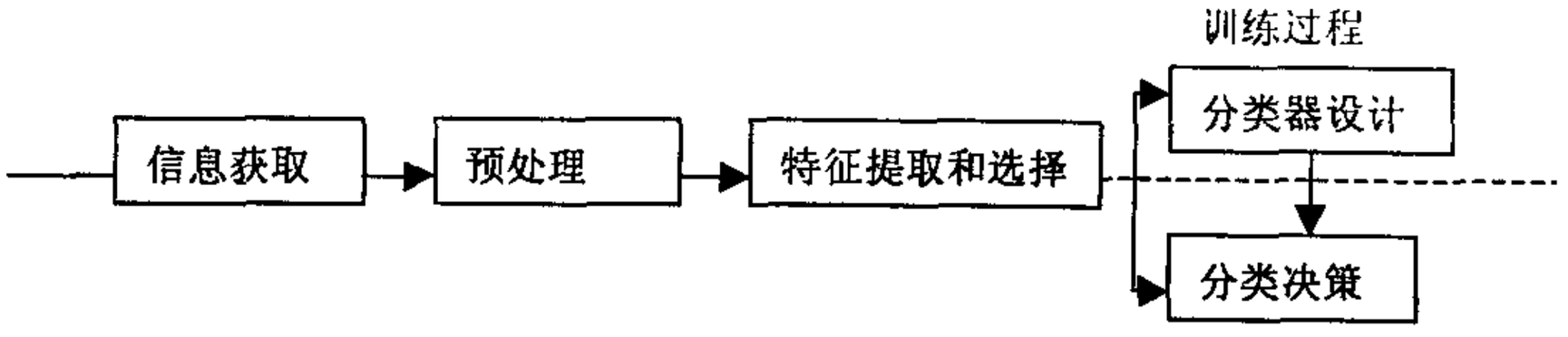


图 2-2　模式识别系统基本构成

①数据获取。获取研究对象已知模式和待识别样本的特征信息，使计算机对研究对象进行分类识别，将所研究对象的特征按计算机所能接受的标志形式表示并输入到计算机，通常输入对象的信息包括：

（1） 二维图像，如遥感图像、文字、地图等；

（2） 一维波形，如地磁波，谱图，脑电图等各种记录曲线；

（3） 物理、化学参量和逻辑值，如地球物理观测数据、地形等高值、地温等。

通过测量、采样和量化，可以用矩阵或向量表示二维图像或一维波形。这就是数据获取的过程。

②预处理。预处理的目的是取出噪声，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。由图像或波形所获得的数据量是相当大的。为了有效的实现分类识别，就要对原始数据进行变幻，得到最能反映分类本质的特征。这就是特征提取和选择的过程。一般我们把原始数据组成的空间叫测量空间，把分类识别赖以进行的空间叫特征空间，通过变换，可把在维数较高的测量空间中表示的

模式变为在维数较低的特征空间中表示的模式。

④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为某一类别。基本作法是在样本训练集基础上确定某个判决规则，使按这种判决规则对被识别对象进行分类造成的错误识别率最小或引起的损失最小。

§2.2 智能模式建模方法

近年来对人工神经网络法、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。

一、人工神经网络法[2][4][14][15][17]

人工神经网络的由来

人工神经网络（ANN）使用自底向上的综合方法，从生理途径直接模拟人的大脑。

人的神经元的结构：由细胞体、树突、轴突和突触等组成。细胞体由细胞核、细胞质和细胞膜组成；树突是由细胞体向外伸出的较短的分支，为信息的输入端；轴突是由细胞体向外伸出的最长的一条分支，为信息的输出端；突触为其它神经元的轴突与该神经元的树突的接口。

神经元是大脑控制与信息处理的基本单元。人脑是由大约 10^{11} 个神经元组成的巨型神经网络。人的认识过程就是大量神经元的整体活动。

人工神经网络就是抽象、简化与模拟上述大脑生物结构的计算模型，又称为联接主义或并行分布处理（PDP）模型。

人工神经网络基本原理

人工神经网络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。

人工神经元

人工神经元是对人脑神经细胞功能的抽象、简化与模拟。是神经网络的基本计算单元，通常认为是多个输入和输出的非线性单元，由内部状态反馈信息和表示神经元活动的特性函数将输入和输出连接起来。常用的特性函数有线性函数、阈值函数、Sigmoid 函数、双曲正切函数等。

网络拓扑

单个的人工神经元只能完成局部的、简单的处理，神经系统之所以具有强大的功能，就在于众多的人工神经元间复杂的连接组合。因此，人工神经网的拓扑即连接结构是极为重要的。这些结构有若干分类方法。按神经元排列层次分，有单层结构与多层结构，按信息传递的方向分，有前馈结构和反馈结构。如下详述

- 单层结构 只有一个输入层和输出层构成。在这种神经网中只有输出层这一个神经元处理层。
- 多层结构 比单层结构有更强的处理能力。不仅有一个输入层和输出层构成，它的神经元处理层是多个。当多层网中的激励函数选为线性函数时，便失去多层的意义。应取用非线性激励函数。
- 前馈结构 其信息传输是向前传而且不回传（无环路）的。
- 反馈结构 在基于反馈结构的神经网中，一个神经元的输出可能反送到它的输入端或者同层或前面各层其它神经元的输入端。反馈网中，输出不但受当前输入影响，同时也可能受某些低层的输入影响。因此它在运行时能获得短期记忆功能。反馈结构有很多种，典型的有层内反馈、层间反馈、纵横反馈结构等。

学习规则

为了使神经网络具有解决具体问题的能力，它的权值矩阵常常不能事先固定，而必须能够随环境（输入）情况的变化或者随应用目的不同而有所变化。设置和修改权值矩阵的过程即是学习过程。网络通过学习从输入信息中获取某些历史的或实际的知识。

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。

有导师的学习规则

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，最常用的偏差衡量准则为最小均方误差（输出层各节点的误差平方和最小）准则，在此准则的指导下，1985 年由 Rumelhart 等提出著名的多层前馈神经网络的误差反传学习算法。其它著名的还有 1987 年由 Hinton 提出的互熵函数最小准则，1990 年 Hampshire 提出

的分类质量因素准则，1994 年 Telfer 提出的最小分类误差准则等。

无导师学习规则

无导师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来重新组织样本空间，使得在输出空间里，相似的样本自动聚在一起。

与有导师学习不同，无导师学习网络的学习准则不是基于误差代价的函数，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，BAM 网络学习的 Hebbian 学习规则和常用于 Kohonen 自组织特征映射（SON）的竞争学习规则。

人工神经网络的基本功能：①联想记忆与回忆 ②分类 ③优化决策与计算。

人工神经网络用于分类问题时，做为分类器使用，随着神经网络的训练完成，标志智能模式模型的形成。

人工神经网络模式建模过程

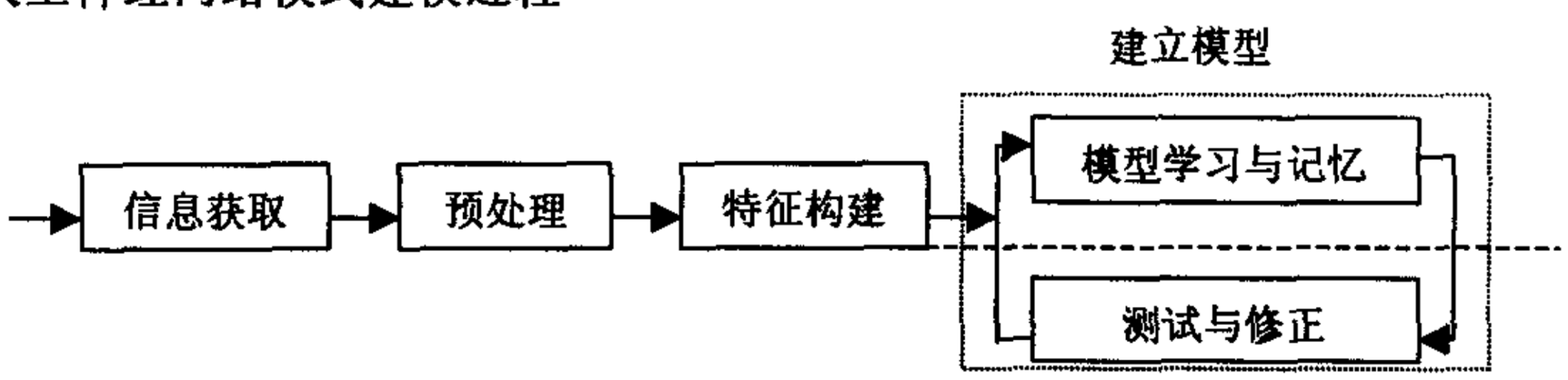


图 2-3 人工神经网络用于模式识别一般过程

图 2-3 是人工神经网络模式建模的一般过程，与传统的模式建模过程是基本一致的。而因人工神经网络作为模式分类器，有它自身的特点，与传统的模式建模过程又略有不同。

“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。”[60]所以人工神经网络只需经过特征构建，一般不需要特征提取和选择就可形成用于建模的样本。

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。正因为它的功能过于强大，在建模初期模型就具有更多的不确定因素，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，从而建立更优的模型。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个认识、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，这点是它与传统的建模方法有区别的。

二、模糊识别法[4]

在模式识别中引入模糊集的概念，先建立一组标准模式，求出待识别样本与每一标准类的隶属度，对应最大隶属度的标准类即为判别归属类——最大隶属原则法。[14]

模糊识别法做出的判断与人们现实生活中的习惯更为接近，如：人们常用“今天天气很热”、“车速过高，需要适当踩刹车”等这些含义确定但又不准确的语言表述，模糊数学能够较好的表达。

§2.3 数据挖掘

数据挖掘（Data Mining），也叫数据开采，数据采掘等，是按照既定的任务目标从海量的数据中提取出潜在的、有效并能被人理解的模式的高级处理过程。数据挖掘是数据库中的知识发现（KDD，Knowledge Discovery in Database）的核心部分，一般把 KDD 等同于数据挖掘。[43]

一、 数据挖掘的步骤及其模式

数据挖掘是处理海量数据时一有效的工具，它有着自己独特的模式。各种模式在不同阶段，执行不同的任务，发挥各自特殊的作用。

数据挖掘包括如下步骤（参见图 2-4）：

1 数据准备：包括数据的选择、预处理、转换。

2 数据挖掘所采用的技术有：决策树、分类、聚类、粗糙集、关联规则、神经网络、遗传算法等。数据挖掘根据 KDD 的目标，选取相应算法的参数，得到可能形成知识的模式模型。

3 评估、解释模式模型：上面得到的模式模型，需要评估以确定模式的有效性。

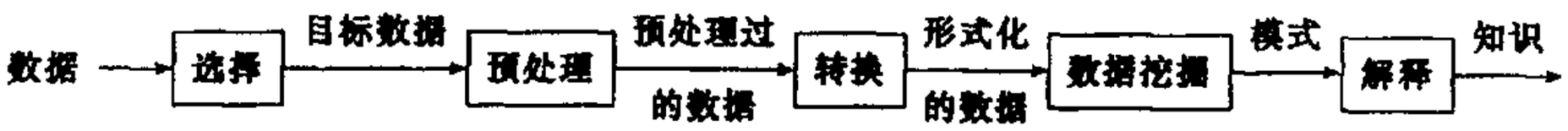


图 2-4 数据库中的知识发现过程

4 巩固知识。

5 运用知识，包括 A：只需看知识本身所描述的结果就可以对决策提供支持；B：要求对新的知识运用知识，由此可能产生新的问题，而需要对知识做进一步的优化。

数据挖掘模式可分为以下 6 种：

分类模式 它是一分类函数，能够把数据集中的数据相应映射到某个给定的类上。

回归模式 回归模式与分类模式相似，它们的差别在于分类模式的预测值是离散的，回归模式的预测值是连续的。

时间序列模式 根据数据随时间变化的趋势预测将来的值。

聚类模式 把数据分到不同的组中，组间差别尽可能大，组内差别尽可能小。

关联模式 通过量化的数据描述事物之间的相互影响。

序列模式 序列模式和回归模式相仿，但此模式把数据之间的关系与时间联系起来。

二、数据挖掘方法

现在已有很多数据挖掘方法，常用的有：

1 关联规则

关联规则(association rules)是指描述在一个事务中物品之间同时出现的规律，或者说是通过量化的数据描述物品甲的出现对物品乙的出现有多大的影响。一般用四个参数来描述关联规则的属性：

（1） 可信度：在出现了物品集 A 的事物 T 中，物品集 B 也同时出现的概率有多大。

（2） 支持度：描述了 A、B 这两个物品集的并集 C 在所有事务中出现的概率有多大。

（3） 期望可信度：在没有任何条件影响时物品集 B 在所有的事务中所出现的概率。

（4） 作用度：是可信度与期望可信度的比值。它描述了物品集 A 的出现对物品集 B 的出现有多大的影响。可信度是对关联规则准确度的衡量，支持度是对关联规则重要性的衡量。

2 决策树方法

决策树是一基于实例的指导学习方法。学习的质量取决于分类准确度和树的规模大小。通过训练例，构造决策树，如果此树不能完全划分所有的实例，则选择树节点进行再划分，直到所有训练例均被划分为止。其中最典型的决策树算法为 Quiulan 的 ID3。[44]

3 神经网络法

模拟人脑神经元方法，以 MP 模型和 HEBB 学习规则为基础，建立了三大类多种神经网络模型：前馈式网络、反馈式网络、自组织式网络，它是一种通过训练来学习的非线性预测模型，可以完成分类、聚类、特征挖掘等多种数据挖掘任务。

§2.4 沙尘暴模式建模

沙尘暴预报问题的特点

沙尘暴是小概率事件，本文研究的区域范围，2 月中旬至 6 月中旬时间段内，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。

沙尘暴与非沙尘暴两类样本的因子在高维空间中错综交缠，结构复杂，加上气象变化多端，存在很多不确定因素，因此单纯的数学模型很难很好的解决。

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。

一些对神经网络有研究的学者对 BP 网络有下面的评价：“从哲理的角度来看，那些有明确因果关系表达式的问题（例如矩阵变换、Fourier 分析等），是不必要通过神经网络学习来解决的；那些毫无规律可循的问题（如完全随机的预报

问题）也是不可能用神经网络学习来解决的。BP 学习网络所擅长的，是处理那种规律隐含在一大堆数据中的映射逼近问题，特别是需要通过学习自适应可调的实时性问题，例如模式识别，自适应控制和模糊决策等。”[58]

本文选用 BP 网络作为沙尘暴预报模型。

沙尘暴模式建模与数据挖掘

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。

首先，所有的信息都是隐藏在大量的沙尘暴相关数据资料中。对其进行分析，从中挖掘出有用的信息。这个过程需要使用聚类分析等数据挖掘工具。

建立沙尘暴预报模型的所有信息也都来自于大量的样本资料，依靠人工神经网络的数据挖掘能力，学习、总结数据规律，并分布式记忆在网络的权值中。

综上所述，确定建立沙尘暴预报模型方案大体上分三个阶段：资料的分析与预处理、预报模型的建立、模型优化。

（1） 资料分析与预处理阶段：具体运用自组织神经网络和 K 均值等聚类方法为工具分析资料，通过特征提取和选择形成建模样本。

（2） 建立模型选用多层前向神经网络。

（3） 模型优化，优化的目的是使模型预报效果更好，其中包括神经网络本身的优化、样本集合的优化等。

第三章 建模样本的形成

§3.1 原始资料

目前的气象数据有两大来源，一种数据利用直接探测技术和遥感技术获取，它的精度和误差基本上取决于探测手段。另一种是加工数据,也称再分析资料，此类数据在一定的气象知识指导下，通过对原始数据进行加工后得到。通常指物理量诊断场、天气系统情况及数值预报产品等。[41]

根据气象专家的气象知识和预报经验，本文选用美国环境气象中心（NCEP）资料作为沙尘暴模式建模的的原始数据。NCEP 资料是数值预报产品，属于再分析资料。

1、资料范围

鉴于我国沙尘暴发生的时间和地区特点,选择我国西北部从初春到初夏自1981 年到 1997 年 2 月 11 日~6 月 10 日资料。如表 3-1 所示：

表 3-1 圈定时间和地域的 NCEP 资料

时间	年份	1981~1997	17 年	
	季节	2 月 11 日~6 月 10 日	初春——初夏	
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°	

2、原始资料的数据格式

NCEP 资料为 2.5%2.5（每格 2.5 度）的格点场资料，包括物理量、时间、空间（经度、纬度、层次）这五个自由量。其中选用资料的时间、经度和纬度范围如表 3-1。

根据专家经验，选取如下层次的各物理量：

500hpa 的高度

700hpa 的东南风（u）

700hpa 的西北风（v）

850hpa 的温度

850hpa 的比湿

地面气压（后因聚类效果不明显，在本系统中未被采用。）

我们用 850hpa 的温度和 850hpa 的比湿推导出 850hpa 的位温，再将东南风和西北风合到一起考虑，进而得到 500hpa 的高度（Hgt）、700hpa 的风（UV）和 850hpa 的位温（θ_{se}）。

物理量 1（500hpa 的高度）年月日

格点值 … …

物理量 2（700hpa 的风）年月日

格点值 … …

物理量 3（850hpa 的位温）年月日

格点值 … …

完整的原始数据样式参见附录。

3、资料分析

根据选定的区域及每 2.5 度对应 1 格的记录方式，易知东西向跨 45°分 18 格、南北向跨 20°分 8 格，对于与沙尘暴天气相关联的物理量 500hpa Hgt、700hpaUV 和 850hpa θ_{se}，各有格点场数据 19*9=171 个。

设三个物理量场的格点数据包含了区分沙尘暴天气和非沙尘暴天气的信息，它们隐含在一个天气样本中的 171*4=684 个格点数据之中。

按照统计模式识别描述样本的惯例

一个样本数据量=样本特征维数

显然，直接将三个物理量场作为建模样本，会由于极高的特征维数，使得建模工作难以进行。

§3.2 信息提取

距平是气象上常用的量，定义为观测值对平均值的差值。[20]

设样本 $X=(x_1,x_2,\ldots,x_n)$，其自身的平均值用 $\bar{x}$ 表示，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

自距平用 X_d 表示，$X_d=(x_1-\bar{x},x_2-\bar{x},\ldots,x_n-\bar{x})=(x_{d1},x_{d2},\ldots,x_{dn})$

当借助原始数值比较样本间的相近程度时，可得到样本间数值相似（简称值

相似）的程度。而借助样本的自距平值比较样本间的相近程度时，可得到样本间等值线形状相似（简称形相似）的程度。

根据专家的建议，我们借助 500hpaHgt 值的相似、500hpaHgt 形的相似，700hpaUV 形的相似和 850hpaθ_{se}形的相似来获取沙尘暴天气信息。由此以来，每个样本的数据量多达 171*2+342+171=855 个。

由上述可知一个样本共考虑三个天气因子的四种相似，致使样本具有因子多、数据量大的特点，这为直接观察样本信息带来困难。

为从大量的样本数据中抽出关于沙尘暴预报的有用信息，采用模式识别中的聚类技术。对于用多因子场共同描述的待聚类对象，存在如下聚类策略：

1. 筛选因子聚类法

A.I.D.方法

A.I.D.方法是从m个因子中选中一个因子，是它对因变量 Y 的最优二分割最好。然后对二类因变量 Y 分别各自去选一个因子，使它对因变量 Y 的第一类、第二类再进行最优二分割，也就是把 Y 分为四类，要使他们分割最好。每次分割均要做 F 检验，如果 F 检验通过，表示分割是有效的，否则是无效的。这样可以从m个因子中挑选出若干个因子对 Y 进行分类。当我们已知因子时，就可以对因变量 Y 做出判断。[47]

2. 组合因子聚类法，即数学上的组合思想。

基于上述两种思想的综合，对强沙尘暴日（5 站点以上）样本集合进行了聚类分析。即

（1）首先基于 A.I.D.方法思想，根据 hgt 的值相似将样本集聚为两类，再对上述两类根据 hgt 的形相似各分 3 类，这样针对 hgt 这个因子，整个样本集就分为 2*3=6 类。

（2）根据 UV 形相似整个样本集聚 2 类。

（3）根据θ_{se}形相似整个样本集聚 2 类。

以上每种聚类，都是从不同的角度对强沙尘暴日样本集合的认知过程。

这样综合考虑上述 3 个因子组合，可得到 6*2*2=24 种类型强沙尘暴日。

§3.3 聚类分析

人们认识自然界很重要的一种方法是对事物进行分类。近几十年来，由于电子计算机技术的发展和多元统计的发展，数值分类学逐渐形成一个新的分枝，称为聚类分析，也有的叫群分析、群落分析、点群分析、分枝分析。[47]

气象领域的资料大多具有内容多、数据量大的特点，随着电子计算机日益成为气象领域不可或缺的工具，聚类分析方法在气象上得到越来越广泛的应用。

§3.3.1 聚类分析原理[2][8][9][20][22][26]

一、聚类统计量[19][47]

聚类分析是根据样本在特征空间中的距离远近，进行客观分类的一种统计方法。

度量样本间相似程度统计量称为相似测度，通常有样本间距离、样本间的相似系数等。设样本集含 N 个样本，由 p 个特征描述，第 i 个样本

$$X_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}, \quad i = 1,2,\cdots,N$$

（1） 样本 X_i 和样本 X_j 间的欧式距离定义为

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad 3\text{-}3\text{-}1$$

（2） 样本 X_i 和样本 X_j 间的平均距离定义为

$$d_{ij} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad 3\text{-}3\text{-}2$$

（3） 样本 X_i 和样本 X_j 间的域块距离定义为

$$d_{ij} = \sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right| \qquad 3\text{-}3\text{-}3$$

（4） 样本 X_i 和样本 X_j 间的相似系数定义为

$$a_{ij} = \arccos s_{ij} \qquad 3\text{-}3\text{-}4$$

其中

$$S_{ij}=\frac{\sum_{k=1}^{p}x_{ij}x_{jk}}{\sqrt{\sum_{k=1}^{p}x_{ik}^{2}\sum_{k=1}^{p}x_{jk}^{2}}} \qquad 3\text{-}3\text{-}5$$

二、聚类方法[3][4][7]

常用的聚类方法有系统聚类法、动态聚类法和模糊聚类法。

系统聚类法

1、计算出样本之间的相似测度，列出相似测度矩阵，例如某一距离

$D=(d_{ij}),i,j=1,2,\ldots\ldots N$

2、将相距最近的两个（类）样本归为一类 ；

3、计算新类与类的距离，列出距离矩阵，转向 2；

终止条件一般有两个：

1）类间距大于设定的类内距阈值 μ

2）类别数等于给定值 k，k 可由先验知识确定。

动态聚类法

动态聚类法先选择若干样品作为聚类重心，再按某种聚类准则，例如最小距离准则使其余样品向各重心聚集，从而得到初始分类，然后判断初始分类是否合理，若不合理就修改分类。如此反复迭代，直到获得合理的分类。动态聚类方法一般具有以下三个要点：

1）选定某种距离度量作为样本间的相似度量；

2）确定某个评价聚类结果质量的准则函数；

3）给定某个初始分类，然后用迭代算法找出使准则函数取极值的最好聚类结果。

常用的动态聚类算法有 K 均值算法，迭代自组织数据分析算法，自组织神经网络法等。

模糊聚类法

通常的类别区别，类与类之间总有一个比较明确清晰的界限，然而有时各类别间，很难找到这样一个界限能够使所有样本均实现比较满意的聚类，在这种情况下，引入模糊数学的方法，根据隶属度最大原则来聚类，也就是模糊聚类。算法过程在这里不再赘述。

三、检验

“物以类聚”作为聚类分析的一种朴素的思想依据，一个群体总是可以根据在某方面的相似程度再分为若干个子群体，但是聚类分析方法没有严格的数学证明支撑，利用其总能得到分类结果，只是不知道该结果是否可靠，所以聚类后还需要对分类结果做显著性检验。

常用的两种检验手段：专家检验和统计检验

专家检验：就是气象专家根据自身多年积累的经验及气象工具对分类结果的合理性做出分析、判断。

统计检验[20][47]

常用x^2检验和F检验于分类结果的检验。x^2检验比较简单，但要求在大样本时，N至少要大于 30，最好N>100 时，用x^2检验。F检验比较复杂，但精度高，特别对于因子少，分类少也有较好的效果。

§3.3.2 K 均值与自组织特征映射网络聚类效果比较

在本文中对强沙尘暴日的样本集用 k 均值法和自组织特征映射网络两种典型的动态聚类法进行了聚类分析。聚类统计量选用欧式距离。

一、k 均值法 [4][6]

K 平均算法建立在误差平方和准则基础之上，其主要步骤如下：设样本集$\{X_1, X_2, \cdots, X_N\}$，含$k$个类型，特征空间$R = S_1 \cup S_2 \cup \cdots \cup S_K$

第一步 任意选择k个初始聚类中心：z_1^1，z_2^1，……，z_k^1（上脚标为寻找聚类中的迭代源算次数）。

第二步 取样本 X_i，若有$|X_i - z_j^m| < |X_i - z_l^m|$（$i = 1,2,\cdots,N$，$j,l = 1,2,\cdots,k$，m 是迭代次数），则 $X_i \in S_j^m$，S_j^m是聚类中心为z_j^m的样本子集S_j^m。

第三步 计算新的聚类中心$z_j^{m+1} = \frac{1}{n_j}\sum_{x \in s_j} X$（j=1，2，……，k），$n_j$为该类中所包含的样本数。

第四步 若$z_j^{m+1} = z_j^m$，j=1，2，……，k，程序结束。否则转到第二步。

聚类中心数 k、初始聚类中心的选择、样本输入的次序，以及数据的几何特性等均影响 k 平均算法的进行过程。对这种算法虽然目前尚无法证明其收敛性，但当各类之间彼此远离时这个算法所得的结果是令人满意的。

二、自组织特征映射网络[7][17]

自组织特征映射网络（简称为 SOFM）是一种无教师学习网络，可以自动地向环境学习，从而具有较强的自适应学习能力。SOFM 的学习则是使得网络节点有选择地接受外界刺激模式的不同特性，从而提供了基于检测特性空间活动规律的性能描述。SOFM 的学习特点是在某个学习准则的指导下，逐步优化网络参数的过程。因此，学习准则是影响 SOFM 学习性能的关键因素之一。

SOFM 网络用作自联想最近邻分类器，能将任意连续值模式分成 P 个类别。网络用竞争学习规则离线学习，按离散时间方式运行。SOFM 网络通过寻求最优参考向量集合来对输入模式集合进行分类，每个参考向量为一输出单元对应的连接权向量。

三、原始样本集的聚类处理：

用 K 均值法和 SOFM 对 1981 年—1997 年样本集中 242 个强沙尘暴日的 500hpa 高度场格点资料分别进行聚类分析。

值相似分 2 类。用下面三种方式聚类分析，结果比较于表 3-2。

1．用 k 均值法，选样本集中前两个样本作为初始聚类中心；

2．用 k 均值法，选样本集中第 20、21 个样本作为初始聚类中心；

3．用 SOFM 聚类

表 3-2

聚类方法 \ 不同个数		k 均值		SOFM	迭代次数
		$z_1^0=X_1\ z_2^0=X_2$	$z_1^0=X_{20}\ z_2^0=X_{21}$		
K 均值	$z_1^0=X_1$，$z_2^0=X_2$		23	21	10
	$z_1^0=X_{20}$，$z_2^0=X_{21}$	23		3	10
SOFM		21	3		1000

在表 3-2 的方式 2 与方式 3 所得结果中，只有 19890410、19930503、19940408

三个样本分得类别不同，这对于共有 242 个样本的集合来说，可以认为两种方法效果基本相同。方式 1 与方式 2、方式 1 与方式 3 相比都有 20 多个样本划分的类别不同，约占整个样本集个数的十分之一，而方式 1 与方式 2 同属 K 均值法，只是初始聚类中心不同，结果说明，初始聚类中心的选择对 K 均值算法的进行过程及结果影响很大。一般来讲，聚类中心数 k、样本输入的次序，以及数据的几何特性等均影响 k 平均算法的进程和结果。

使用 SOFM 则不受初始聚类中心的选择、聚类中心数 k、样本输入的次序，以及数据的几何特性的影响。但用 SOFM 法的缺陷是学习算法的收敛速率很慢，需要上千次的迭代。

四、用密度中心法改进初始聚类

为了克服初始聚类中心的选择对聚类结果的影响，用密度中心法对初始聚类中心的选择作了改善。

密度中心法的思想就是首先计算每个样本的密度，并将高密度样本列为初始聚类中心的候选样本。具体做法如下：

1. 计算各样本间最大欧氏距离 $D_{\max}$，选定半径 $r = D_{max}/m$。$m \in \{2,3,\cdots,N\}$，m 的选取与样本的分布和聚类类别数有关。本文此处用的 m=4。
2. 以各样本为中心，以 r 为半径圈定区域，计算圈入区域的样本数目 n，n 在一定程度上反映了样本周围的聚集程度，简称密度。
3. 按密度 n 由大到小排序，选密度最大的样本为第一个类中心；计算密度次大的样本与第一个类中心的欧氏距离，如大于 r 则证明与第一类不在同一区域，选其为第二个类中心；否则计算下一个高密度样本与第一个类中心的欧氏距离 D，看是否大于 r；依此直至找到指定个数的类中心。

用密度法找到初始类中心之后，再用 k 均值法聚类，经试验结果与方式 3 比较，只有 2 个样本归属不同（是 19890410、19940408）。可见，用密度中心改进 k 均值法进行聚类，减小了由于初始聚类中心选择对聚类进程和结果的影响，同时又由聚类结果的不同发现了一些特殊的样本，这些样本处于类的边缘。

SOFM 聚类进程不受初始聚类中心的选择、聚类中心数 k、样本输入的次序，以及数据的几何特性的影响，但收敛速率慢。而 K 均值法收敛快，用密度中心改进的 K 均值法能够减小初始聚类中心选择不同对聚类进程造成的影响，但不

能消除样本输入的次序，以及数据的几何特性的影响。由此可见，不同的聚类方法各有特点，根据具体问题的侧重，选择聚类方法。在本文中，对资料的分析和处理是全部工作的基础，可靠性是最重要的，所以选用 SOFM 对样本集合进行聚类分析。

§3.3.3 多因子聚类结果综合分析

对强沙尘暴日综合分析的目的有两个：一是了解沙尘暴日的类型信息；二是找到沙尘暴日典型模式。

分别对强沙尘暴日的 Hgt、UV、θ_{SE} 三个因子的样本集合聚类分析，可各分得多个类，三个因子综合考虑有多种组合情况。

根据 Hgt 值相似将样本集聚两类，再对上述两类根据 Hgt 形相似各分 3 类，这样针对 Hgt 因子，整个样本集分为 2*3=6 类；根据 UV 形相似整个样本集聚 2 类；根据θ_{se}形相似整个样本集聚 2 类。综合考虑 Hgt、UV、θ_{SE} 三个因子，共有 6*2*2=24 种类型，简称 6_2_2 组合。经专家经验的指导和本文多种组合试验的结果，这种组合对于分析强沙尘暴日样本集合来说，其类别的规模最为合适。表 3-3 为对 24 种类型的分析情况。

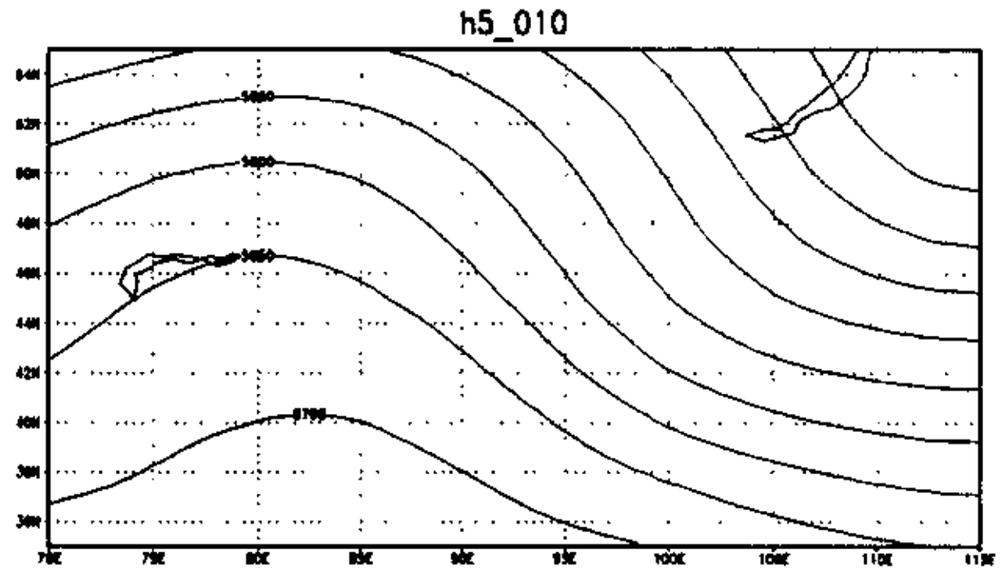
表 3-3 中数字 000、…、1211 代表相应组合所属类别，如 1211 中的 12 表明它聚类结果为先是根据 Hgt 值聚类在 1 类中，然后 Hgt 形聚类在 2 类中，第三个数字 1 表明它 UV 形聚类在 1 类中，第四个数字 1 表明它θ_{se}形聚类在 1 类中。说明风场特点是用符号表示的，E 代表东、S 代表南、W 代表西、N 代表北，如 uv：NW W 表示该风场的特点是最强的风由西北风转为西风。

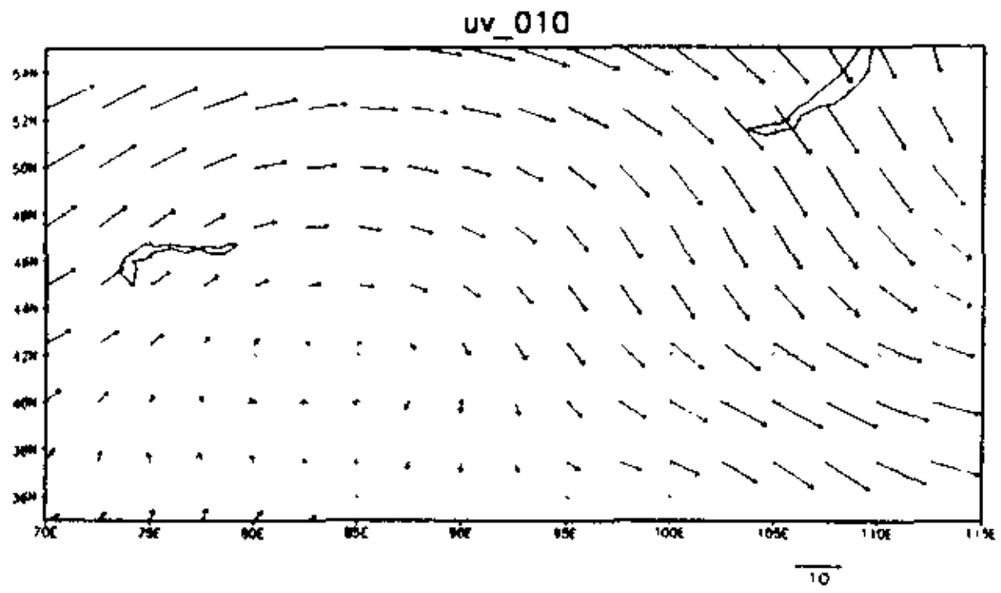
表 3-3

类别号：类内样本数	该模式三因子场特点	
000：35	hgt：一脊一槽 uv：NW W 位温：经向度大	模式 1
001：4		
010：9	hgt：一脊一槽（槽宽） uv：NW W 位温：经向度大	模式 2
011：0		
100：8		
101：17	hgt：两脊一槽 uv：W 位温：温度低	模式 3
110：6		
111：0		
200：0		

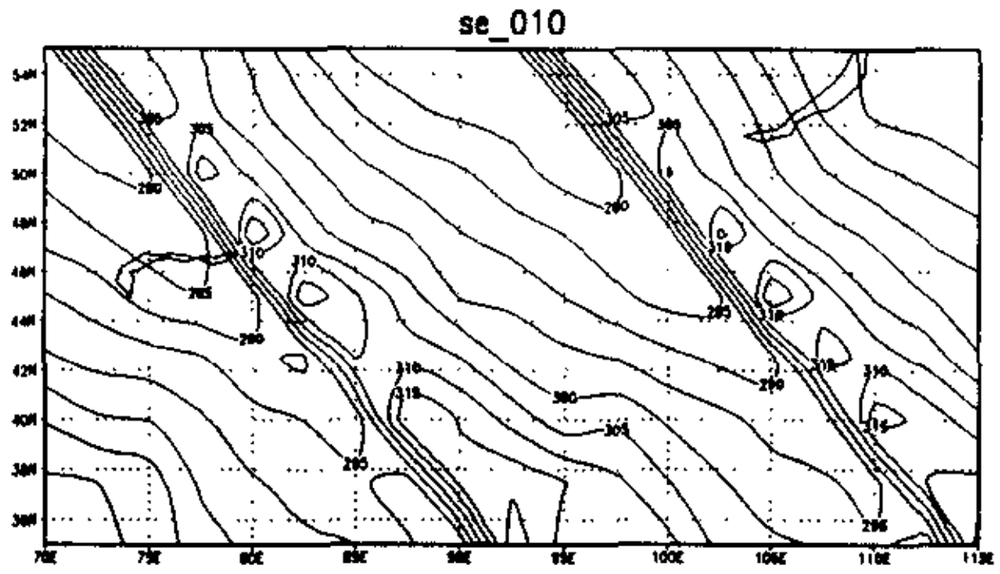
201：15	uv：平直 W 位温：更低	模式 4
210：17	hgt：两槽一脊 uv：SW NW SW 位温：低	模式 5
211：12		
1000：16	hgt：一脊一槽 uv：NW 位温：经向度大	模式 6
1001：8		
1010：20	hgt：两脊一槽 uv：NW SW 位温：高	模式 7
1011：1		
1100：0		
1101：11	hgt：平直 uv：W 位温：	模式 8
1110：15	hgt：两槽一脊 uv：SW NW 位温：经向度小	模式 9
1111：9		
1200：0		
1210：1		
1211：37	hgt：一（二）槽一脊 uv：SW NW 位温：高	模式 10

表 3-3 中 24 种沙尘暴类型，有的类型样本数为 0，因为三个因子是互相联系的，这种类型事实上不会出现，例如 111 类型在 hgt：两脊一槽 位温：温度低情况下，不会出现在 uv：NW SW 或 SE NW 下的沙尘暴；有的类型样本数少，这种类型发生的概率较低，例如 110 类型在 hgt：两脊一槽 位温：温度高情况下，出现在 uv：NW SW 或 SE NW 的沙尘暴概率较低。典型的含义一是这种类型的样本数量多，二是历史上特别严重的沙尘暴日（即典型的沙尘暴日）不要漏掉。基于这两点考虑从 24 种沙尘暴日的类型中，挑选出以上 10 种作为典型类型，计算这 10 种类型的各因子中心场，并称为典型沙尘暴模式。图 3-1 和图 3-2 为其中两种典型沙尘暴模式的可视化描述。



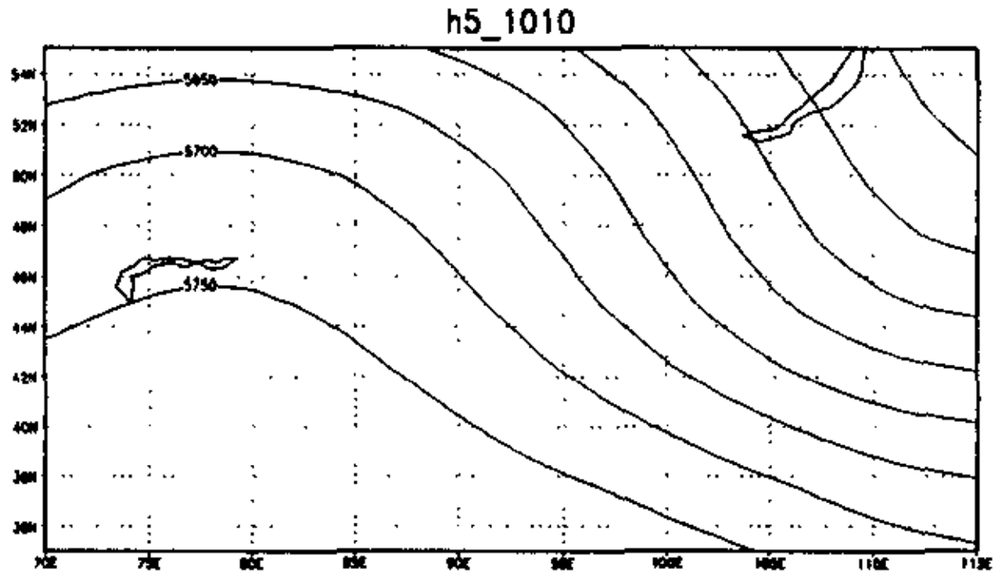


(b)

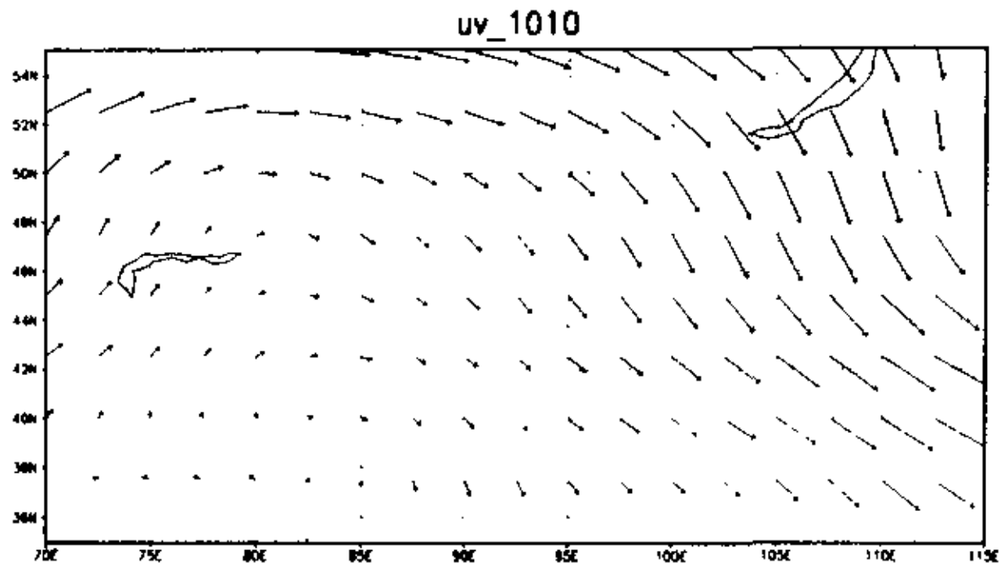


(c)

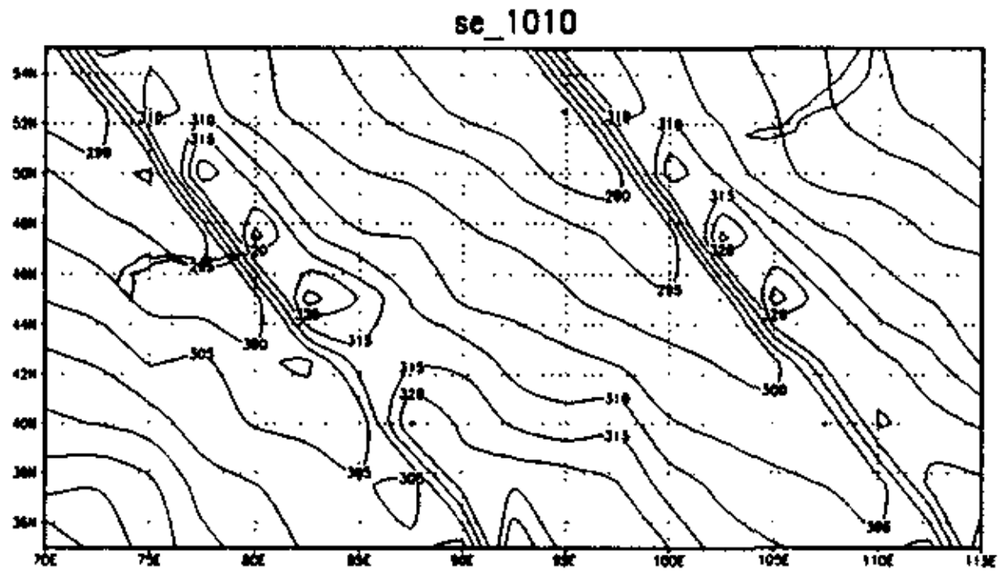
图 3-1 典型模式 2 (a) hgt：一脊一槽（槽宽）(b) uv：NW W (c) 位温：经向度大



(a)



(b)



(c)

图 3-2 典型模式 7 (a) hgt：两脊一槽 (b) uv：NW SW (c) 位温：高

§3.4 样本特征的构建

用计算机代替人类的识别行为，建筑在用以描述样本的特征的有效性和代表性之上。

本文用的原始数据是 NCEP 资料。选用的气象因子为 500hpaHgt、700hpaUV、850hpa θ_{se}。并提取了高度场 Hgt 值的信息、高度场 Hgt 形的信息，风场 UV 形的信息，位温场 θ_{se} 形的信息。

借此构建样本特征如下：

1. 计算上节聚类结果综合分析后得到的沙尘暴日 10 个典型模式 Hgt 值和形的信息场，UV 形的信息场，θ_{se} 形的信息场的共计 40 个中心场。

2. 计算各样本与这 10 个典型模式 40 个中心场的距离，得到 40 个距离量。

3. 将 40 个距离量作为描述样本 40 个特征，依次为 hz_1、fx_1、sx_1、hx_1---hz_i、fx_i、sx_i、hx_i---hz_{10}、fx_{10}、sx_{10}、hx_{10}。

hz_i、fx_i、sx_i、hx_i 分别代表高度值的第 i 个特征、风形的第 i 个特征、位温形的第 i 个特征、高度形的第 i 个特征，分别表示某样本与第 i 个典型沙尘暴模式的高度值、风形、位温形和高度形的相似程度。

为进一步分析这 40 个特征对于分类的贡献大小，引入 Fisher 比率和主成分进行分析。

Fisher 比率法[13]：特征参数 i 的 Fisher 比率 F_i 为

$$F_i = \frac{(\bar{x}_{i1} - \bar{x}_{i2})^2}{v_{i1}^2 + v_{i2}^2} \qquad \textbf{3-4-1}$$

式中 $\bar{x}_{i1}$ 和 $\bar{x}_{i2}$ 分别为一类样本和二类样本的变量 i 的平均值；v_{i1} 和 v_{i2} 分别为一类样本和二类样本变量 i 的标准偏差

$$v_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2} \qquad \textbf{3-4-2}$$

计算出这 40 个特征的比率 F_i 后，按由小到大排序，F_i 小的特征对于分类的贡献小，包含的分类信息少。

主成分分析：属于线性映射方法，利用坐标变换从原有特征得到一批个数相同的新特征，这些新特征中的前几个包含了原有特征中的主要信息，当前几个新特征的信息量已经足够大时，便可以舍弃其余的新特征，从而达到减少特征个数的目的。主成分法使变换后的特征相互正交，每个特征反映的有用信息可以度量。主成分分析的详细步骤参见文献[59]。

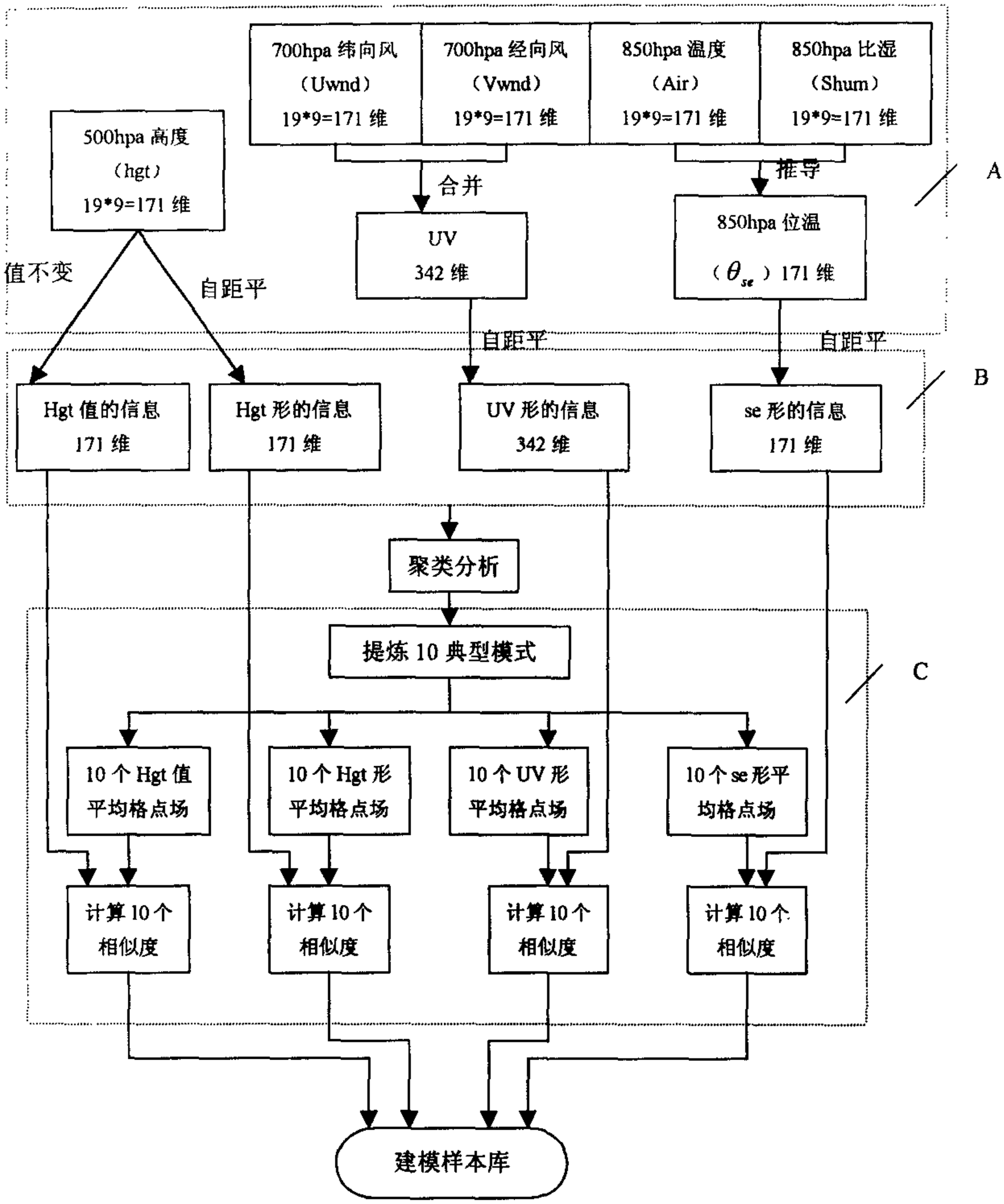
特征选择和提取的意义在于通过特征维数降低乃至特征间正交使后面的建模复杂程度降低，但通常以损失一定的建模信息为代价。本文选用 BP 网络建模，去掉 1、2 个输入特征对于网络的复杂程度没有大的下降，只有大量的筛去输入数量，才能使网络拓扑结构得到明显简化；但输入数量的大量筛除，必然损失较多的有用信息，网络规模降低是以牺牲一些建模信息作为代价的，这又互相牵制，

只有看最后的综合效果决定取舍。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，网络的隐层本身就具有特征提取的功能。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，当然代价是网络的拓扑没有简化、训练时间仍很长。

在这种情况下综合考虑，决定上述 40 个特征不再作选择和提取，直接用作建模样本的特征。至此，建模样本形成。

§3.5 建模样本描述

通过资料的分析→信息的提取和特征的构建，使同时考虑值和形相似的样本的格点场描述 $X_0=(x_1,x_2,\cdots,x_{855})$（171（高度场值）+171（高度场形）+342（风场形）+171（位温场形）=855）改变为样本与典型沙尘暴模式值和形相似度的描述 $X=(x_1,x_2,\cdots,x_{40})$，并称为建立人工神经网络模型样本。过程如图 3-3 所示。



A：原始数据　B：信息提取　C：特征构建

图 3-3 建模样本形成过程示意图

第四章 基于人工神经网络的沙尘暴预报模型

§4.1 多层前馈神经网络

神经网络的三要素是网络拓扑、神经元特性函数、学习算法。

一、网络拓扑

多层前馈神经网络是根据它的网络拓扑命名的（多层和前馈的含义见第二章），如图 4-1 所示。

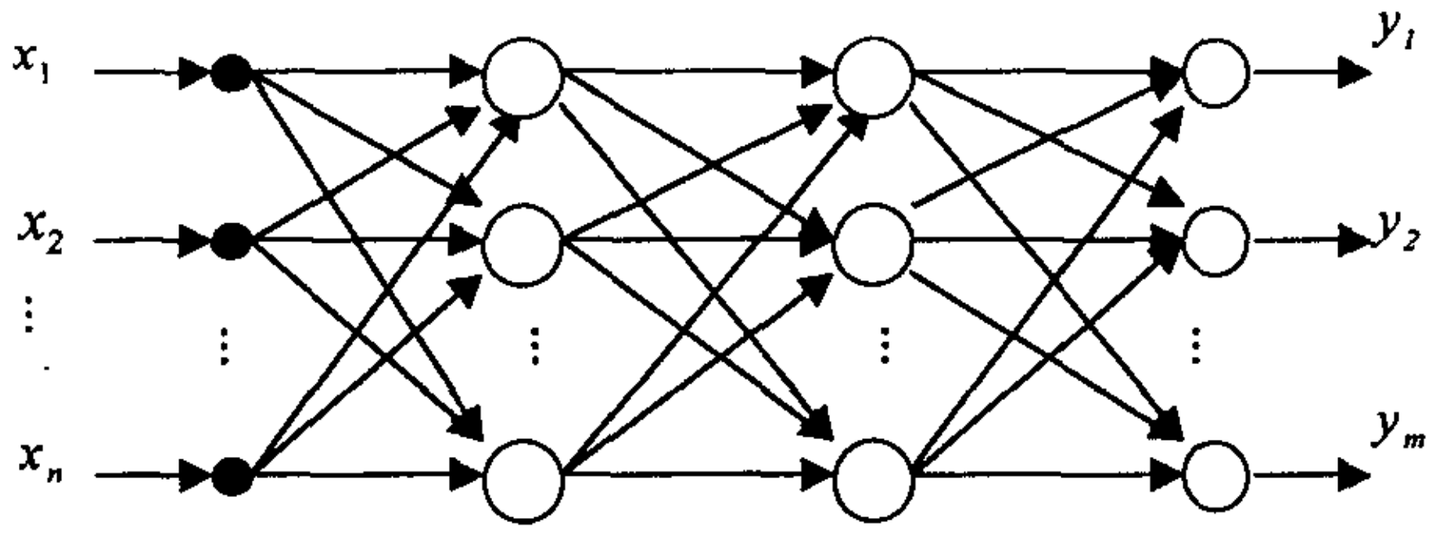


图 4-1 多层前馈神经网络

二、神经元特性函数

多层网络的特性函数通常取非线性函数，最常用的是 Sigmoid 函数。它对每个处理层的形式化表示为：

$$y_i = \frac{1 - e^{-g(x)}}{1 + e^{-g(x)}}, g(x) = \sum_{k=1}^{s} \omega_{ik} z_k + \omega_0 \qquad 4\text{-}1\text{-}1$$

式中 s 为当前处理层前一层的节点个数，z_k 为前一层节点的输出，ω_0 为节点的阈值，ω_{ik} 为节点 i 与节点 k 的连接权值。当处理层为第一隐层时，s=n，z=x。

整个网络的输入与输出之间的映射关系为：

$$F : \Re^n \to \Re^m; \qquad F(x_1, \ldots, x_n) = (y_1, \ldots, y_m) \qquad 4\text{-}1\text{-}2$$

三、学习算法

最常用的是误差反传训练算法（BP）。

BP 算法是一种有指导的训练，根据各训练组网络的实际输出与期望输出之间的偏差来调整网络中的各权值，这些训练组由输入输出对$\{x_i^k, d_i^k\}$组成。

BP 学习算法的核心是通过一边向后传播误差，一边修正误差的方法来不断调节网络参数（权，阈值），以实现或逼近所希望的输入输出映射关系。它对每一个训练进行两趟传播计算：

- 前向计算——从输入层开始向后逐层计算输出，产生最终输出，并计算实际输出与目标输出间的误差。
- 反向计算——从输出层开始向前逐层传播误差信号，修正权值，直到误差小于给定值。

步骤如下：

（1） 初始化权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$d_0, d_1, \cdots\cdots d_{n-1}$

（3） 从第一隐层开始，逐层计算输出矢量$y_0, y_1, \cdots\cdots y_{n-1}$。

（4） 按下式修正权值：

$$\omega_{ij}(t+1) = \omega_{ij} + \eta\delta_i X_i' \qquad 4\text{-}1\text{-}3$$

其中：η>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：

$$\delta_i = y_i(1-y_i)(d_i - y_i) \qquad \text{（对输出层）} \qquad 4\text{-}1\text{-}4$$

$$\delta_i = x_i'(1-x_i')\sum\delta_k\omega_{ik} \qquad \text{（对隐层）} \qquad 4\text{-}1\text{-}5$$

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。

BP 学习算法流程图如下：

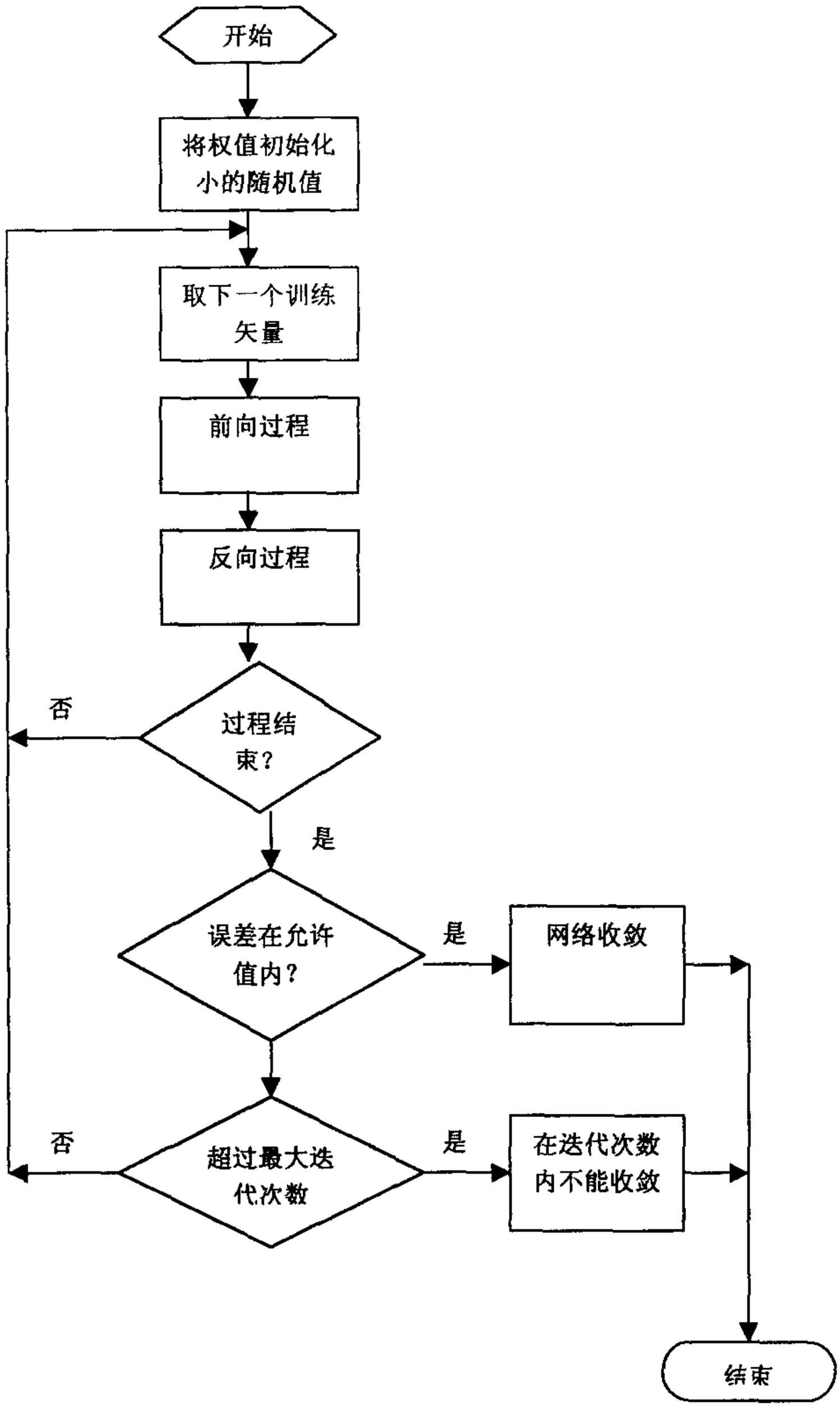


图 4-2 BP 学习算法流程图

§4.2 模式分类与多层前馈神经网络功能解析

根据多层前馈神经网络的构造和机理，可以进一步理解其模式识别能力。

1．线性判别函数、感知器与线性模式分类[4]

线性判别函数

在传统模式识别建模方法中线性判别函数是一类较为简单的判别函数。它的一般表达式为

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \qquad 4\text{-}2\text{-}1$$

式中 **x** 是 **d** 维特征向量，又称样本向量，**w** 称为权向量，分别表示为

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

w_0 是个常向量，称为阈值权向量。对于两类问题，线性分类器可以采用 Bayes 决策规则，即令

$$g(x) = g_1(x) - g_2(x)$$

如果样本 $x \in \omega_1$，有 $g(x) > 0$，则

$$\begin{cases} g(\mathbf{x}) > 0, \text{则决策} \mathbf{x} \in \omega_1 \\ g(\mathbf{x}) < 0, \text{则决策} \mathbf{x} \in \omega_2 \\ g(x) = 0, \text{可将} x \text{任意分到某一类，或拒绝} \end{cases} \qquad 4\text{-}2\text{-}2$$

方程 $g(\mathbf{x}) = 0$ 定义了一个决策面，它把归类于 ω_1 类的点与归类于 ω_2 类的点分割开来。当 $g(\mathbf{x})$ 为线性函数时，这个决策面便是超平面。

利用线性判别函数进行决策，就是用一个超平面把特征空间分割成两个决策区域。超平面的方向由权向量 **w** 确定，它的位置由阈值权 w_0 确定。判别函数 $g(\mathbf{x})$ 正比于 **x** 点到超平面的代数距离（带正负号）。当 **x** 在 H 正侧时，$g(\mathbf{x}) > 0$；在负侧时，$g(\mathbf{x}) < 0$。

感知器

感知器是构成前向神经网络的基本单元。

感知器（Pereceptron）是一种双层神经网络模型，一层为输入层，另一层具有计算单元，可以通过监督学习建立模式判别的能力，如图 4-3 所示。

学习的目标是通过改变权值使神经网络由给定的输入得到给定的输出。作为分类器，可以用已知类别的样本集合作为训练集，当输入为属于该类的样本 **x** 时，应使对应于该类的输出等于 1；而输入不属于该类的样本 **x** 时，输出则为 0（或-1）。设理想的输出为 y，实际的输出为 $\hat{y}$。为了使实际的输出逼近理想输出，可以反复依次输入训练集中的样本 **x**，并计算出实际的输出 $\hat{y}$，对权值作如下的修改：

$$\omega_i(t+1) = \omega_i(t) + \Delta\omega_i(t) \qquad 4\text{-}2\text{-}3$$

其中

$$\Delta\omega_i = \eta(y - \hat{y})x_i \qquad 4\text{-}2\text{-}4$$

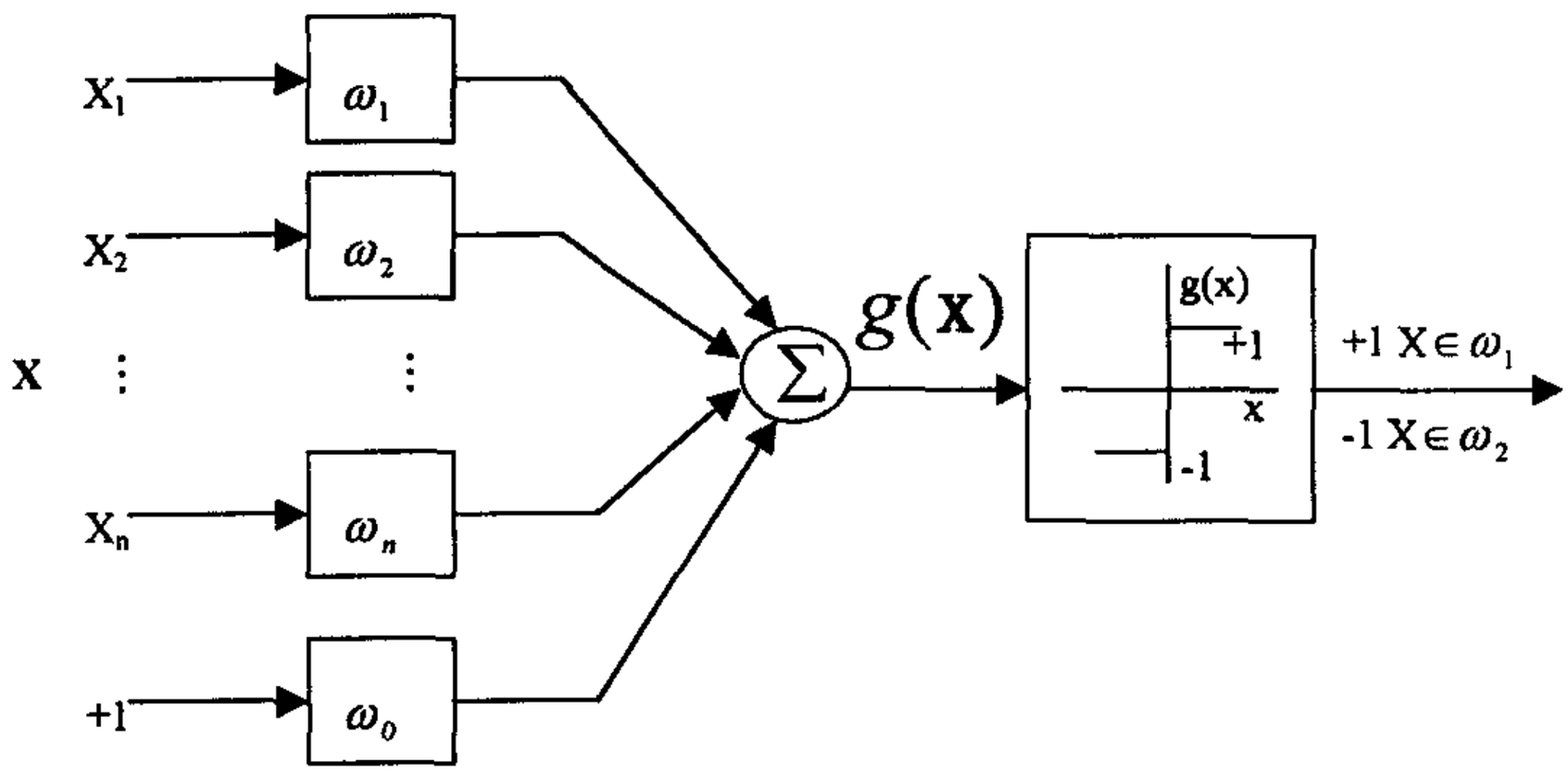


图 4-3 感知器

感知器的学习过程与求取线性判别函数的过程是等价的，感知器的特性：两层感知器只能用于解决线性可分问题；学习过程收敛很快，且与初始值无关。

2. 分段线性划分

分段线性判别函数是一种特殊的非线性判别函数，它确定的决策面是由若干超平面组成的。它能逼近各种形状的超曲面，具有很强的适应能力。由于它的基本组成仍然是超平面，因此，与一般超曲面相比，仍然是简单的；分段线性判别函数比一般的线性判别函数错误率要小，又比一般的非线性判别函数简单。

3．前馈神经网络的分解[4][5][60]

如图 4-1 所示为一个一般的前馈神经网络，它有两个隐层。设它的输入有 n 个，分别代表选择的 n 个特征。训练样本是 n 维模式向量，它们依次送到神经网络的输入层，用来训练神经网络的参数（权值）。输入层各节点 x 以一定权值送到第一隐层各节点，第一隐层各节点输出再以一定权值送到第二隐层各节点。最后，第二隐层节点分别送到输出层，得到运算结果。

观察第一隐层诸神经元，它们有共同的输入和不同的输出。每个这样的神经元都是前面所说的感知器，它决定在 n 维空间的一个超平面。这些是并列的神经元（设共有 s_1 个）。一般说来，这 s_1 个并列神经元的输入权值集是不同的。由于权值集的不同使它们的输出决定了 n 维空间的不同的超平面。每个超平面将 n 维空间分割成正负两个半空间。从而，这 s_1 个并列的神经元在第一隐层的输出提供了一组 s_1 个超平面。

如果样本的训练得到了一个正确的结果，则这个结果给出的超平面通过以后的联接和组合构成一个正确的分段线性划分，这个分段线性划分能拟和对这些类模式能正确分类的超曲面。可见，前馈神经网络第一隐层的作用是提供正确分段线性划分的超平面。如果能做到这一点，第二隐层以后的各层的作用则是正确连接和组合这些超平面，以保证得到能正确分类的分段线性划分。这就把模式识别中的分段线性划分与前馈神经网络联系了起来。应该指出，无论采用什么样的前馈神经网络训练方法，以上的分析对于一个前馈神经网络总是成立的，即总可以归结为某个分段线性划分。这正体现了神经网络的模式划分能力。

第二隐层及以后的各层的神经元也都是并列的感知器，每个神经元在相应的空间同样决定一个超平面。

以上讨论的是前馈神经网络构造良好的情况，即第一隐层提供了正确的分段线性划分，第二隐层的作用只是按照任务正确组织第一隐层的输出。在这种情况下只要两个隐层已经完全可能对任意样本空间分布做正确的分段线性划分。

在一般前馈神经网络的应用中，首先需凭经验来决定神经网络的结构，即所需的层次和各层的节点数。这是一项困难的工作，因为对许多算法来说，网络内部工作机制是不透明的，为了保证神经网络的正确工作，通常多设一些节点，这

往往导致增加冗余的节点和隐层。例如，设一类样本分布需要 5 个超平面来正确分割，在盲目设置第一隐层为 4 个节点的情况下，第一隐层不可能对样本做完全正确的分割。这就需要另一隐层来继续完成第一隐层未完成的任务。

可见，不考虑神经网络内部工作机制的盲目构造神经网络，会出现以下结果：或是得到正确分类，但网络中包含许多冗余的节点和隐层；或是得不到良好的分类结果。

一个前馈神经网络可以看作是一个分段线性划分器，它的第一隐层决定一组超平面，其余各层决定这些超平面的组合和联接方式。因此，适当选择隐层数和每个隐层的神经元数，就能提高网络的正确识别率，这是其性能会远超过单层的简单感知器分类器。

神经元的非线性环节

在前馈神经网络中必须采用非线性环节，即它的输入和输出的关系是非线性的，否则，无论多么复杂的联结总可用一个等价的感知器来代替，就失去了使用多层网络的意义。非线性环节可以是阶跃函数，即

$$y=\begin{cases}1, if g(x)=\sum_{k=1}^{s}\omega_k x_k+\omega_0>0;\\ 0, if g(x)=\sum_{k=1}^{s}\omega_k x_k+\omega_0\le 0.\end{cases} \qquad 4\text{-}2\text{-}5$$

如在感知器模型中就是如此。而在一般神经网络中采用 s 函数，便有以下关系：

$$y=\frac{1}{1+e^{-g(x)}}, g(x)=\sum_{k=1}^{s}\omega_k x_k+\omega_0 \qquad 4\text{-}2\text{-}6$$

式中 $g(x)$ 实际上是线性判别函数。从线性判别函数分析中我们知道，判别函数值正比于样本到分界面的距离，这个式子表明，当样本位于分界面正面无穷远处时 $g(x)=\infty, y=1$，即样本非常“可靠”地属于该类。当样本位于分界面正面且离分界面很近时，$g(x)\approx 0, y\approx 1/2$，即样本不十分“可靠”地属于该类；当样本位于分界面反面无穷远处时 $g(x)=-\infty, y=0$，即分类完全不正确。

在第一隐层的输出，可以得到模式空间的这样一个划分，各个超平面所围成的区域内包含同一类样本，这些区域是这些超平面所围成超多面体。设有 s 个超平面，若每个超平面决定的正半空间用 1 表示，而负半空间用-1 表示，则每个子

空间可用 s 维列向量来表示，即：

$$Y=(a_1,a_2,\cdots,a_s)^T, a_i\in\{-1,1\} i=1,2,\cdots,s \qquad 4\text{-}2\text{-}7$$

每个区域都可用这些超平面的正反面特性唯一的决定。

图 4-4a 是一个简单的二维例子。图中分段线性边界的两条边界直线 H_1,H_2 将二维二类模式空间分成四个部分。其中阴影区域为第一类 ω_1，它对应于第一个和第二个判别函数的正面，其区域特征为（1，1），其余区域为第二类。图 4-4b 是上例在分类空间中表示，可用边长为 1 的正方形顶点来表示。顶点（1，1）即为第一类区域，第二类区域占据其他三个顶点（0，0），（1，0），（0，1）。

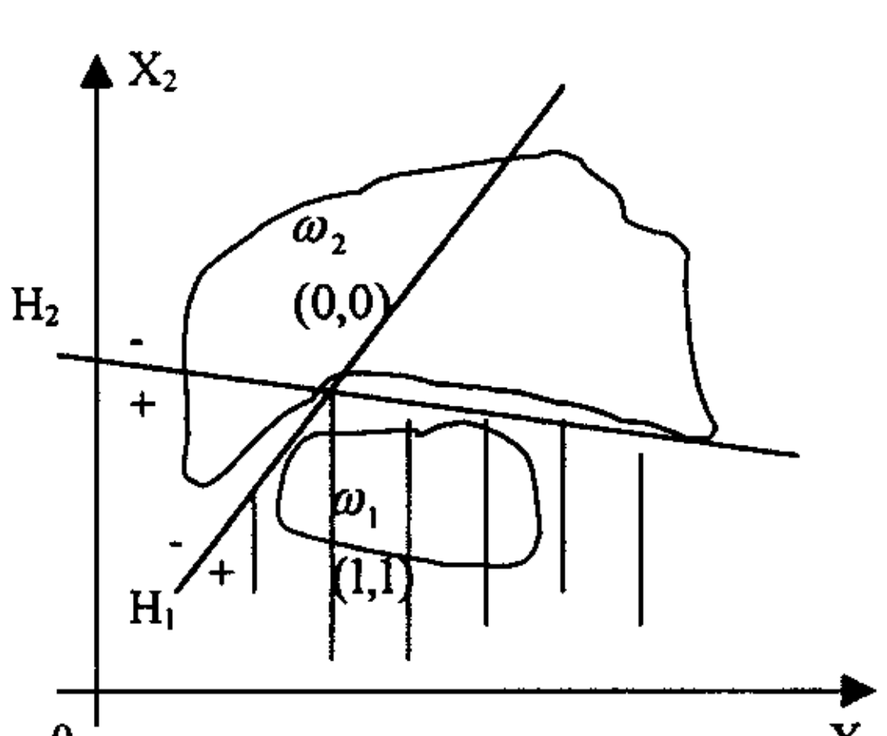


图 4-4a

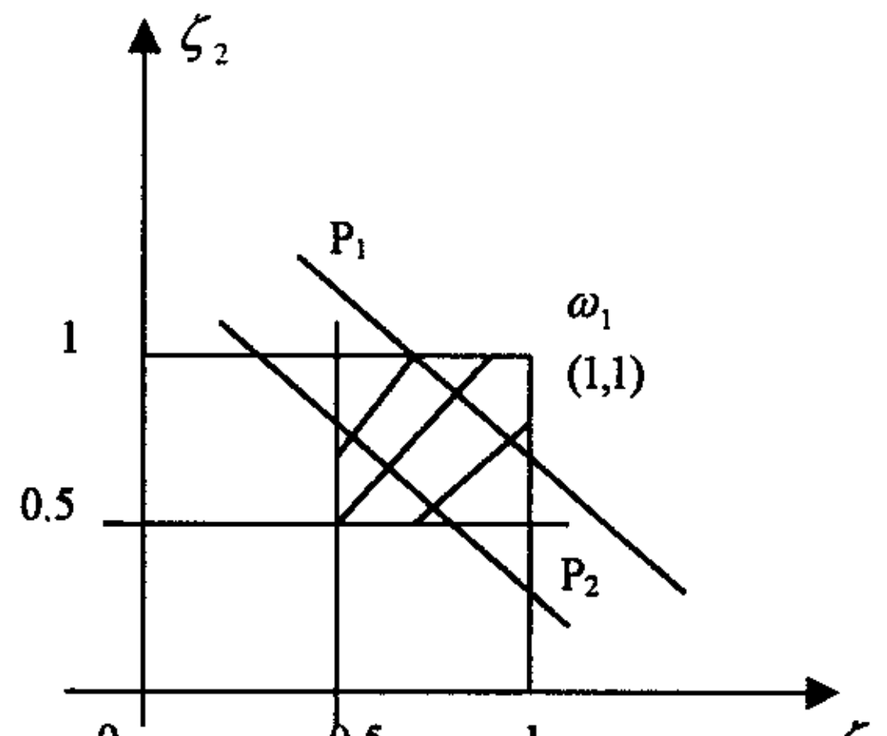


图 4-4b

当应用阶跃函数时，每类模式在分类空间都位于超立方体的顶点上，用一个超平面 P_1 即可将 ω_1 与其他类模式分离出来，得到正确的分类结果。而当采用 S 函数时，属于 ω_1 类的模式样本占据了在两个坐标轴方向的(0.5,1]的阴影区域。这样，在第二隐层用超平面 P_2 将 ω_1 与其他类模式分开时，就会产生一定的混淆。从这个观点看，对于前馈神经网络，采用 S 函数对分类是不利的。

在一般 BP 算法中采用 S 函数并不会使分类边界成为非线性的，因为非线性环节仅仅是改变样本到分界超平面的距离而已。不管输出采用何种非线性环节，第一隐层提供的总是一个线性划分。

§4.3 沙尘暴预报模型

由以上对于神经网络的讨论，确定沙尘暴预报模型建模框架如下：

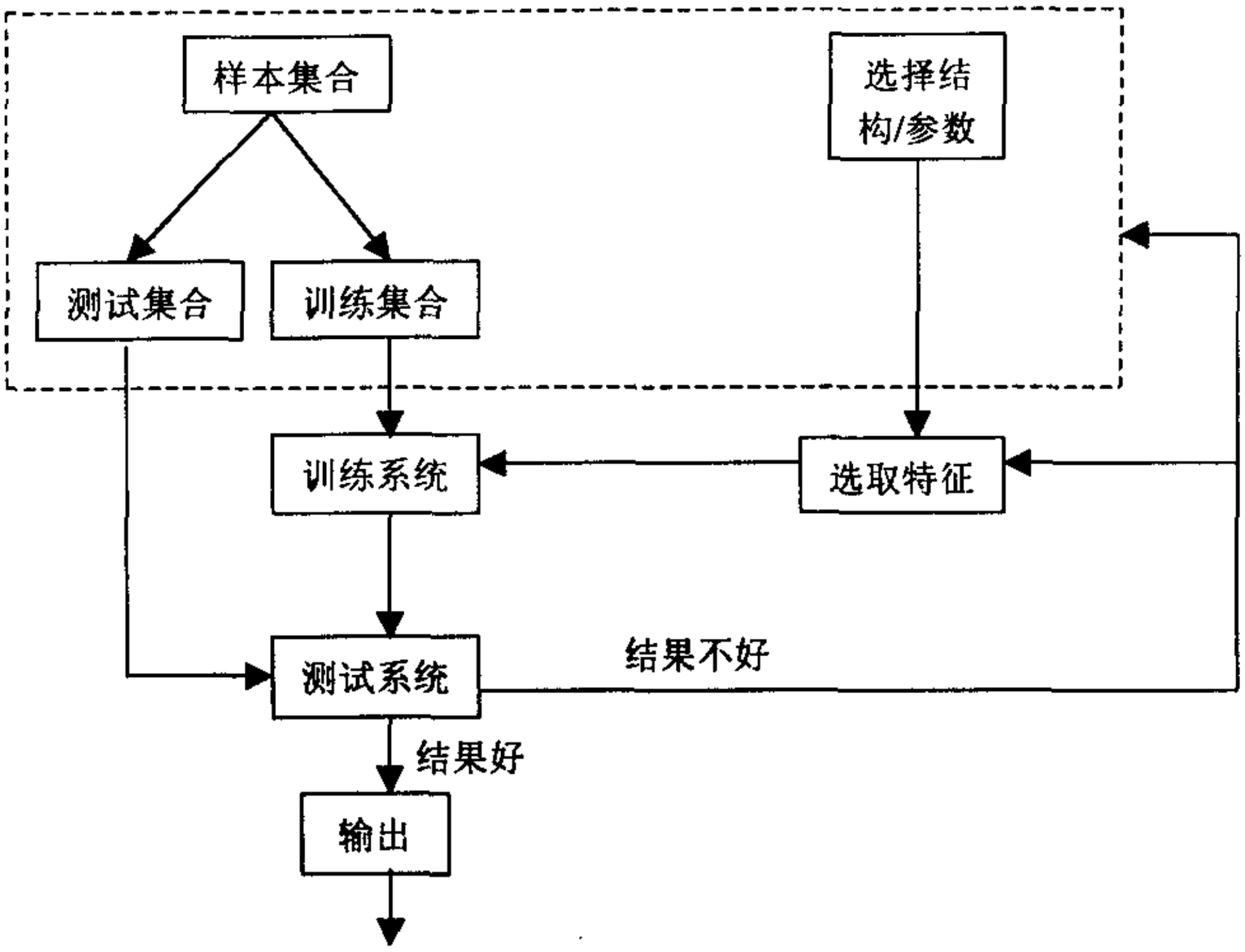


图 4-5 建模过程流程图

建模过程是个反复训练和测试的过程，目的是为了训练好的模型能够达到最好的预报效果。对于多层前馈神经网络来说，模型的好坏，主要取决于样本集合的大小和特征、网络的拓扑、训练参数这几个方面。[12]

在图 4-5 中，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，再训练、测试，反复试探，直至测试效果满意为止。

§4.4 问题分析

在本课题中，在选用 BP 网络训练沙尘暴预报模型时，神经元特性函数选为

$$y=\frac{1-e^{-g(x)}}{1+e^{-g(x)}},g(x)=\sum_{k=1}^{s}\omega_k x_k+\omega_0 \qquad 4\text{-}4\text{-}1$$

成功界限指数 CSI 可作为衡量模型质量的指数。气象上常用 CSI 代表预报水

平

$$CSI = \frac{c_f}{c_f + w_f} \times 100\% \qquad 4\text{-}4\text{-}2$$

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。[39]

下面是本文早先所建的模型 0 和用模型 0 拟和、试报的结果。

模型 0 如下：

网络拓扑：40*40*20*1；**训练参数：**学习率 0.4，惯性系数 0.02，迭代次数 1000；**训练样本集合：**90~95 年样本集，共有样本 721 个，其中沙尘暴日 57 个，非沙尘暴日 664 个；**测试样本集合：**96~97 年样本集，共有样本 239 个，其中沙尘暴日 25 个，非沙尘暴日 214 个。拟和及试报结果如表 4-1 所示。

表 4-1

模型 0	结果				
	空报	漏报	正确报有	正确报无	CSI
拟和	0	0	153（100%）	562（100%）	100%
试报	30（14%）	18（72%）	7（28%）	184（86%）	12.7%

上面拟和达到了 100%，试报 CSI 值为 12.7%，试报结果 CSI 值偏低，为了达到更好的预报效果，需从样本集合、网络的拓扑、训练参数这几个方面，对模型进一步优化。下一章就此作更详细的介绍。

第五章 模型优化

§5.1 网络拓扑优化

神经网络网络拓扑的优化系指网络的层数和每层节点数的优化。

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不重要的方面。[12]

希望得到的网络具有最好的性能，这样的网络应具有尽量少的隐层单元。确定网络的拓扑结构通常用 PLD 算法或试探法：

- **PLD 算法思想：**

根据第四章对神经网络的解析，可以通过对样本集合分段线性划分来构造神经网络。即对于分好类的样本集，先用分段线性划分算法得到必须的一组超平面以及它们的参数，这些超平面的数目就是第一隐层的神经元个数，而这组超平面的参数就是输入层与第一隐层联系的权集；再构造第二隐层来正确组织第一隐层的输出。[5]

- 试探法大致有以下几种做法：

1) 自底向上法：

由较少的隐层节点数开始，最好是低于解决此问题所需要的节点数。如果网络没有令人满意地收敛，确认网络分化样本的能力不够，于是可增加隐层大小。这个过程重复进行，直到获得解决问题的最小网络为止。

2) 自顶向下法：

从足够大的网络开始，然后逐渐删除网络隐层节点。这是定义网络拓扑的另一方法。如 Lee 等[1992]提出一种删除方法，在它的方法中可被删除的节点是：

a）对于所有训练模式来说输出是常量（在限度内）。

b）对于权值矢量的幅度比其它隐节点和输出层的模相对小。

3）离散优化法：

即将神经元网络的学习和结构的优化同时进行，给每个网络结构分配一个评

估值，此类方法中最典型的是利用遗传算法来进化神经元网络。[17]

由于网络结构由节点个数及其关联数量确定，而网络的节点数及其关联数可能无限多，如果再考虑网络参数变化的影响，人工试探需要耗费人们大量的时间和精力，为此试探法程序的自动实现是目前的一个重要研究方法。[17]

本文也对程序自动化作了试验，但因为程序中使用了随机技术避免网络在学习过程中陷入局部最小，使学习过程具有一定的不可重现性。随机函数严格上来说并不是完全随机的，当每次重新启动一个进程所用的随机函数过程还是相同的，也正因为这一点，使程序过程是可以重现的，一个不可重现的程序，对于一个研究过程来说意义大为降低。

反向传播算法有可能陷入局部最小，如何避免呢？有两种避免局部最小的方法，它们分别是模拟退火算法和遗传算法。而避免局部最小的另一项技术是在训练模式中注入噪声，这种数据适应技术的随机性质在避免局部最小问题上是很有用的。程序中两个地方使用了随机函数，权值初始化是随机的，每次迭代时训练样本进入的顺序也是随机的。

本文采用手工试探法寻找最优网络拓扑。

表 5-1 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

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。表中只是列出了其中很少一部分的试探结果。通过比较各种拓扑结构下试报 CSI 值的大小，选择合适的网络拓扑。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。

网络拓扑的确定无论是采用试探法还是 PLD 算法，它还要取决于训练网络所用的样本集合的性质，下面就是为研究样本集合的影响所做的工作。

§5.2 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。

再分析资料 NCEP，它的原始数据可靠性已得到气象方面的验证，本文在数据的预处理、特征构建等过程中，每步都要进行检验，也是为了保证数据的可靠性。

那么数据如何组织才能具有代表性呢？训练集合应该取多大最合适呢？

训练集合的内容：

训练集合必须足够大而又具有代表性，以便充分地描述问题域。尽管有时其内容不一定准确，但被识别的每类必须是存在的。同时还需考虑到，一些不能被识别的反例也可能需要。在每一类型中，必须存在大量的例子以反映此类型的现实世界变化。

训练集合的大小：

网络的拓扑越复杂，它越可能过拟和。解决此问题在于，必须有大量丰富的训练数据。

对于一定的拓扑，取多少训练样本矢量才是足够的？以往的经验是，假如在第一隐层有 L_0 个输入单元和 L_1 隐单元，则输入层和第一隐层间的权值数是 $N=(L_0+1)*L_1$，这里包括偏畸权值。其中输入层同其他层相比较大，这些层构成了系统中的主要自由参量。这种情况下，至少使用 2N，或最好使用 4N 个训练样本。[12]

在建模过程中，本文对训练集合的组织作了很多尝试，下面介绍一下其中两种情况的效果比较。

测试集合：首先定下选用 96~97 年共 2 年的样本作为测试集合，共有 239 个样本，其中含沙尘暴日样本有 25 个，占总数的 10.5%

训练集合：

1）最初选用 90-95 年共 6 年的样本作为训练集合，是考虑随着时间的变化整个大的气候环境有所变化，时间越接近的大环境应该越相似。

90~95 年作为训练集合，共有 715 个样本，其中含沙尘暴日样本有 153 个，占总数的 21.4%。

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

表 5-2

训练集合年限	样本个数	网络拓扑	2N	试报 CSI%
90~95 年	715	40*20*10*1	2*820=1640	22
△ 81~95 年	1788	40*20*9*1	2*820=1640	25.9

此为两网络均为学习率 0.4，惯性常数 0.02，迭代 1000 次结果，此时 90~95 训练的网络拟和 CSI 值为 100%，81~95 训练的网络拟和 CSI 值为 68.4%。

表 5-2 说明用 90~95 年样本集合作为训练集合未完全包含所有沙尘暴类型，通过增加训练集的样本数目，提高了模型预报水平。下面采用 KNN 法对模型试报结果展开进一步的分析和讨论。

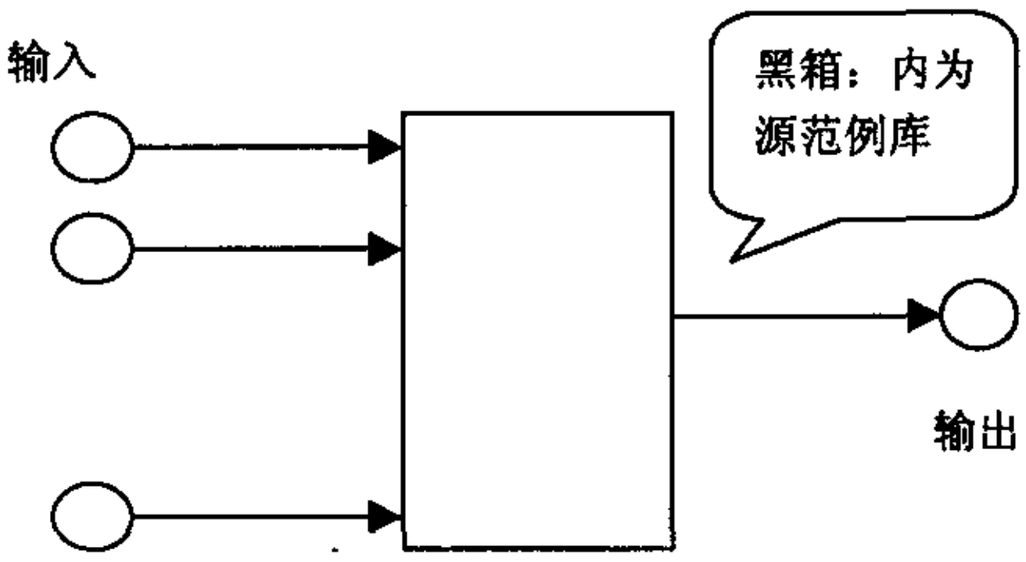


图 5-1 KNN 法示意图

K 近邻（KNN）法 简单来说就是取未知样本 x 的 k 个近邻，看这 k 个近邻中多数属于哪一类，就把 x 归为哪一类。

KNN 和神经网络都是基于相似建模的思想。KNN 可看作一个只有两层的前馈神经网络，输入层和输出层，只是中间经过一个源范例库的环节，从源范例库中搜索相似范例，根据搜到的相似范例结果进行综合，从而做出预报。而神经网络则是通过用源范例训练网络的方式，对源范例进行了规律性的总结，并以权值的方式记录下来。

KNN 和神经网络的区别：

1．KNN 输入层与下一层的联接权值均为 1，而神经网络输入层的各输入节点与

下一层的联接权重经训练后各不相同。

2. KNN 在每次预报时直接利用源范例库，没有记忆和归纳能力，神经网络则不然。

3. KNN 的结果比较直观，可以估计与预报内容相似的源范例集的相似程度，而神经网络则具有不透明性。所以在训练神经网络的过程中，适当的结合 KNN 法，可增加网络的透明度，提高对训练过程和结果的把握和理解。

首先，利用 KNN 法能够分析并解释被神经网络报错的样本。

测试样本集合中的每个样本，都可以用 KNN 法在源范例库（训练集合）中找到 N 个和它最相似的样本，并按照相似程度排序。用这个信息观察 BP 网络试报结果中总是报错的那些样本，发现这些样本报错的原因是很多与它相似的源范例与它的类型相反，当然报的结果也就是与它的实际类型相反的了。这反而更进一步证明了神经网络确实有着相似预报的能力，同时也证明了在现有特征下有一小部分样本交错纠缠在一起很难区分。

经总结，测试样本集合中有两种类型的沙尘暴日样本，其中约一半的样本总被报对，其他则容易报错。与测试集合情况类似，训练集合中的样本也有两种类型，一部分拟和时总易报错，大部分还是拟和的很好。

其次，利用 KNN 法可以细化预报结果。

神经网络一般只是能够定性的报出沙尘暴的有无，而不易报出沙尘暴的程度，结合 KNN 法，就可以根据历史上相似的样本对要预报的实况程度做出概率判断。

§5.3 训练参数的研究

学习率

由第四章第一节的式 4-1-3 可知，学习率的大小影响训练过程中权值的调整幅度。

学习速率的适应调整，是从一个递归到下一个递归的递推过程。反向传播算法收敛减慢的可能原因之一，是在一给定的误差表面区域中，给定的权值维数不能接受单一的学习率。如果误差函数对于已知权值的微分在连续的递推过程中符号不变，可试着增加它的学习速率。相反，如果符号变化，则应减少学习速率。

[12]

在训练网络的过程中，可通过平均误差观察学习率对训练进程的影响。

过程跟踪：

平均误差曲线图

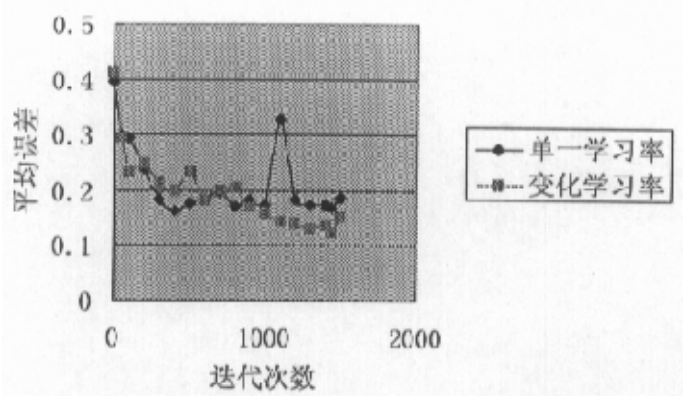


图 5-2 平均误差曲线

从图中可以看到，单一学习率的平均误差在迭代 400 次左右，即不再减小而是保持在一个水平上波动，其中偶有异常点，是因训练中采用了随机策略导致的；变化学习率的平均误差则一直保持着下降的趋势，中间有些点还出现波动证明学习率减小的速度应该更快些。

效果检验：

CSI拟和曲线

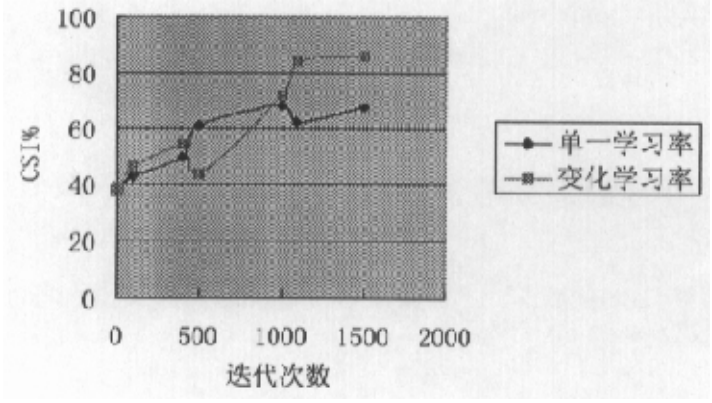


图 5-3 CSI 拟和曲线

从图中可看到，迭代达到一定次数后，单一学习率的拟和 CSI 值不再有明显提高，变化学

习率的拟和 CSI 值则保持着提高的趋势。

上面的平均误差曲线和 CSI 拟和曲线，对比了在单一学习率 0.4 情况下和在初始学习率为 0.5 每 500 次学习率降低 0.1 情况下的学习过程和效果变化。训练集合用 81~95 年样本集、惯性系数 0.02、网络拓扑 40*20*9*1。

迭代次数

学习的根本目的是逼近规律，有推广性，而不是单纯追求学习正确率。实际上，网络开始记忆训练集合细节的同时，已失去了它的推广能力。所以，学习应“适时而止”。我们的任务就是找到这个最佳时刻。

下面的平均误差曲线和 CSI 拟和曲线，对比了在变化学习率 0.5/0.4/0.3 情况下，拟和与试报随着迭代次数的改变，平均误差和 CSI 值的变化。训练集合用 81~95 年样本集、惯性系数 0.02、网络拓扑 40*20*9*1。

过程跟踪：

平均误差曲线图

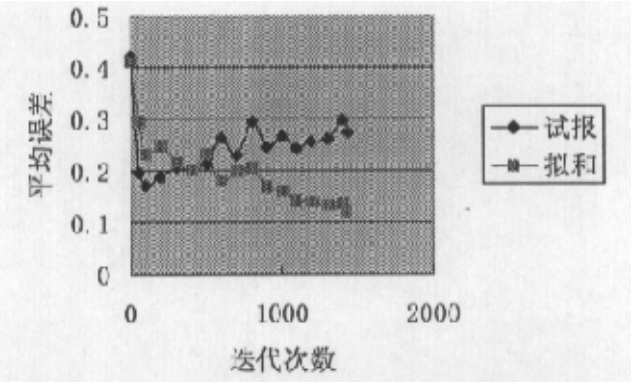


图 5-4 过度训练曲线

图中随着迭代次数增加，拟和曲线的平均误差不断降低，而试报在迭代一定次数后平均误差开始上升，证明此时网络推广能力下降。

综合考虑学习率和迭代次数的试报效果：

CSI测试曲线

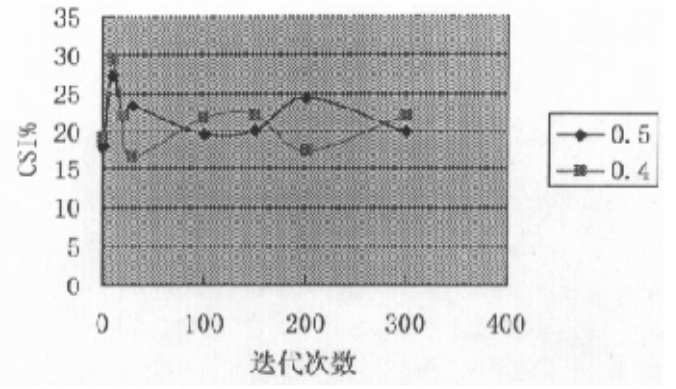


图 5-5 CSI 测试曲线图

在迭代 10 次都达到一个最高的 CSI 值，但此时推广能力较强，拟和较差。200 次左右学习率 0.5 的 CSI 值达到较高的水平。

综合考虑网络的拓扑结构、训练参数、样本集合，经多次试探得到以下几种较优的模型：

模型 1：网络拓扑 40*20*9*1，学习率 0.4、惯性系数 0.02、迭代 10 次，训练集合用 81~95 年样本集，试报集合用 96~97 年样本集；

模型 2：网络拓扑 40*20*9*1，学习率 0.5、惯性系数 0.02、迭代 200 次，训练集合用 81~95 年样本集，试报集合用 96~97 年样本集；

模型 3：网络拓扑 40*20*9*1，学习率 0.4、惯性系数 0.02、迭代 1000 次，训练集合用 81~95 年样本集，试报集合用 96~97 年样本集。

表 5-3

模型	试报结果				
	空报	漏报	正确报有	正确报无	CSI
模型 0	30（14%）	18（72%）	7（28%）	184（86%）	12.7%
模型 1	26（12%）	10（40%）	15（60%）	188（88%）	29.4%
模型 2	20（ 9%）	14（56%）	11（44%）	194（91%）	24.4%
模型 3	33（15%）	10（40%）	15（60%）	181（85%）	25.9%

表 5-3 的测试结果表明相对于最初建立的模型 0（见第四章第四节），模型 1、2、3 的预报准确率都有一定程度的提高，模型优化取得了一定效果。但这 3 个模型

是用试探法找到的相对较优的模型，无法证明他们就是最优的。如用于业务时，使用哪一个模型，需要在实用中根据效果决定，并进一步完善。

§5.4 权值信息挖掘

神经网络具有不透明性

神经网络方法采用的是“黑箱”策略，通过样本的多次训练，使神经网络这个“黑箱”学习得到模式类别间差异的规律，并被存储于神经网络的内部节点和权值中。

由于神经网络的“黑箱”策略，和其它方法相比能够采用更简单更直接的方式从特定领域中获取知识，但同样带来的问题是得到的知识难以理解。这就妨碍了它在更多的领域发挥更大的作用。

根据神经网络的构造与内部运行机理，知道从训练样本集合中总结得到的规律分布式记忆在网络的权值中，那么权值中记忆的信息规则是什么呢，怎样才能够发现这些规则呢？

基于上面的问题，就“沙尘暴预报”这一具体问题训练得到的网络模型，进行内部权值的信息抽取的初步探讨如下。

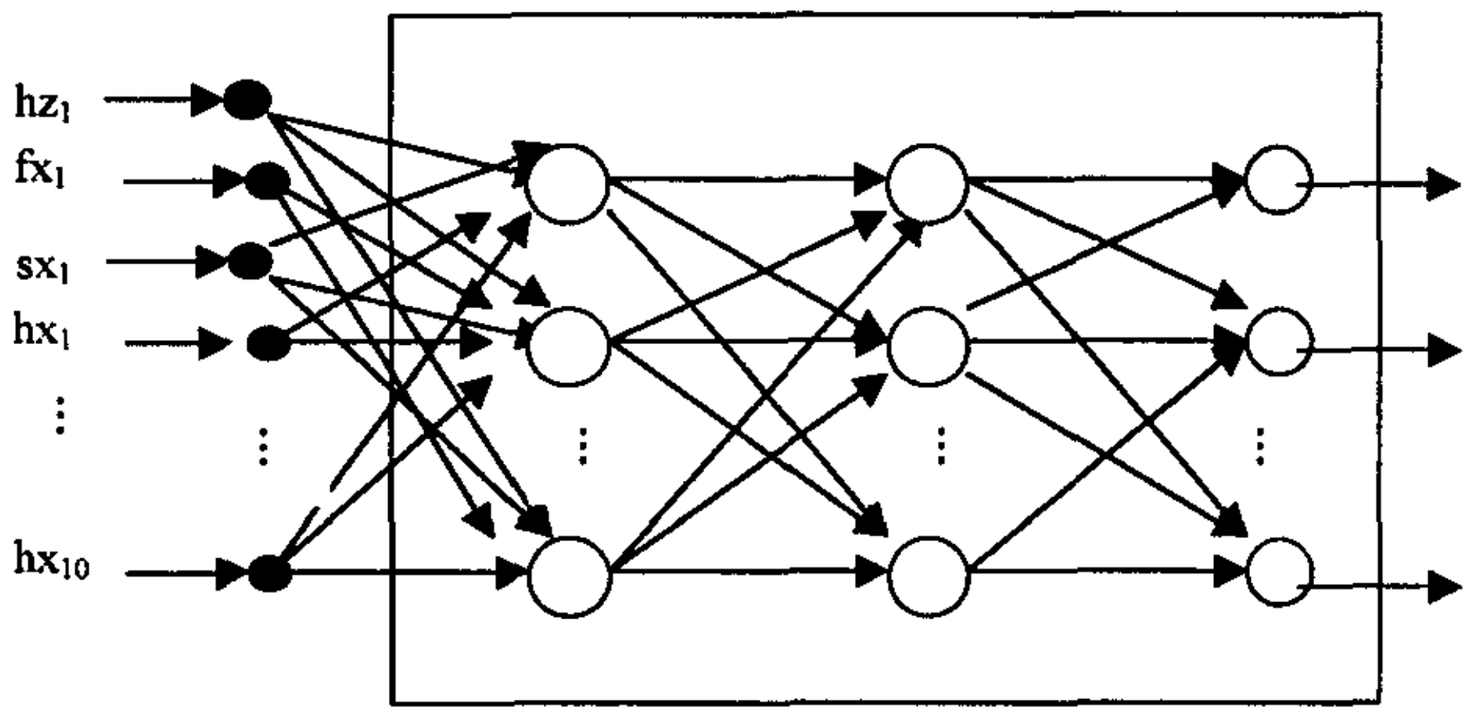


图 5-6

本文建立的“沙尘暴预报”最终模型如图 5-6 所示，网络拓扑为 40*20*9*1。输入层 40 个输入节点依次为 hz_1、fx_1、sx_1、hx_1---hz_i、fx_i、sx_i、hx_i---hz_{10}、fx_{10}、

sx_{10}、hx_{10}。hz_i、fx_i、sx_i、hx_i分别代表高度值的第 i 个特征、风形的第 i 个特征、位温形的第 i 个特征、高度形的第 i 个特征。

输入层各节点与第一隐层各节点间的权值大小，代表了输入层各节点与下一层节点联系的强弱，把所关心的代表相同特征的各节点的权重统计到一起，就大致可以得到该特征对最终结构影响的强度。实际上第一隐层各节点与下一层节点联系的强弱各不相同，第二隐层各节点与输出层节点联系同样有强有弱，因此输入层节点对最终输出的影响应是它与第一隐层节点间的权重再乘以一个比例系数，而又因为中间神经元存在非线性环节，所以这个比例系数是得不到的。在这里使用简单化的思路，把后面的网络结构当作一个黑箱，并视为最终的输出节点。

按照上述想法，帮助分析特征，希望能够得到输入特征与输出定性的关系。

本文建立的模型，输入层与第一隐层间共有 40*20=800 个联接，把代表相同特征的权值统计在一起，与每个特征相关的就各有 200 个联接，如下式：

$$\sum \omega_{HZ} = (|\omega_{HZ11}| + |\omega_{HZ12}| \cdots + |\omega_{HZ120}|) + \cdots() + (|\omega_{HZ401}| + |\omega_{HZ402}| \cdots + |\omega_{HZ4020}|) \quad 5\text{-}4\text{-}1$$

共计 200 条权值的和，同样可得到$\sum \omega_{FX}$、$\sum \omega_{SX}$、$\sum \omega_{HX}$。结果如图：

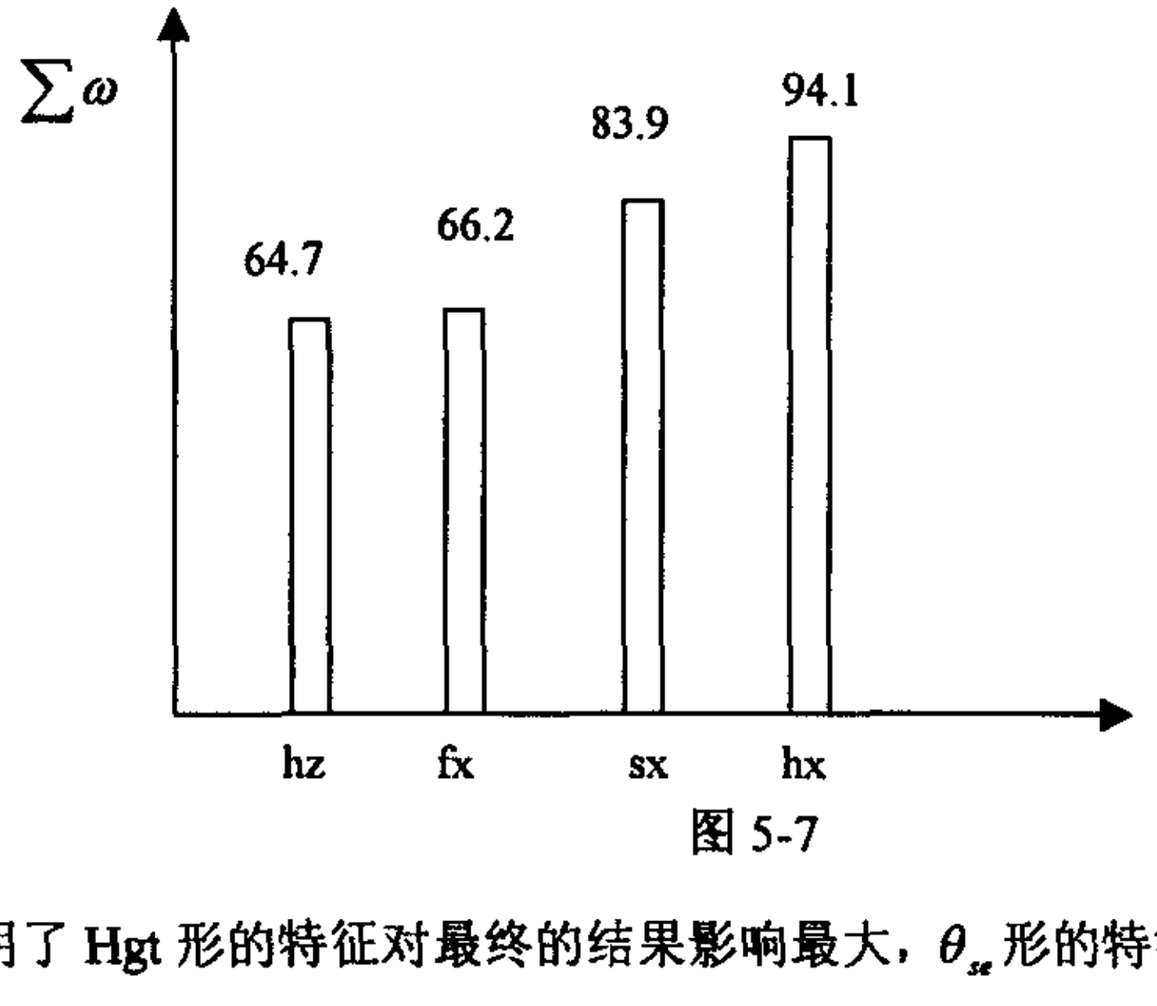


图 5-7

它说明了 Hgt 形的特征对最终的结果影响最大，θ_{se}形的特征其次，UV 形的特征和 Hgt 值的特征相差不大。

当然这个结果，在这个问题上不一定是精确的，还需要验证和进一步深入研究。它只是提供了一个从权值中挖掘有用信息的思路。

现在随着对神经网络研究的深入，神经网络已变得越来越透明了，因为神经

网络具有归纳能力，人们利用神经网络总结和记忆规律，当能将神经网络中记忆的精确的规则变为显式（符号或公式）时，或许利用它就能对一个未知的问题得到象经典的方程那样的精确的关系描述，从这个意义上讲神经网络具有很大的数据挖掘方面的潜力。

具体到气象领域，我想如能把神经网络和传统的动力方式结合起来研究，一定会互有裨益的。

本章针对沙尘暴建模过程中，网络拓扑、样本集合、学习参数对网络性能的影响作了详细的探索。发现这几方面对网络性能的影响不是独立的，他们互相影响、环环相扣，一个因素变了，其它因素也要相应做出调整。

目前网络拓扑、样本集合、学习参数等的选择，这在 BP 网络的使用中还没有规律可循，一般是具体问题具体分析。只有对神经网络的机理更为了解，才能对上述几方面作出的最优的选择。

第六章 问题与展望

§6.1 结果分析

第五章表 5-3 的测试结果表明，通过模型优化，预报模型的试报准确率有所提高，人工神经网络用于沙尘暴的短时预报能够达到一定的准确率，但还不十分理想。这一方面是由于沙尘暴具有小概率、多因子、高维、数据量大，从而难预测的特点；另一方面整个预报系统也还有很大提高的潜力待进一步挖掘。

图 6-1 给出了本文工作的全过程。

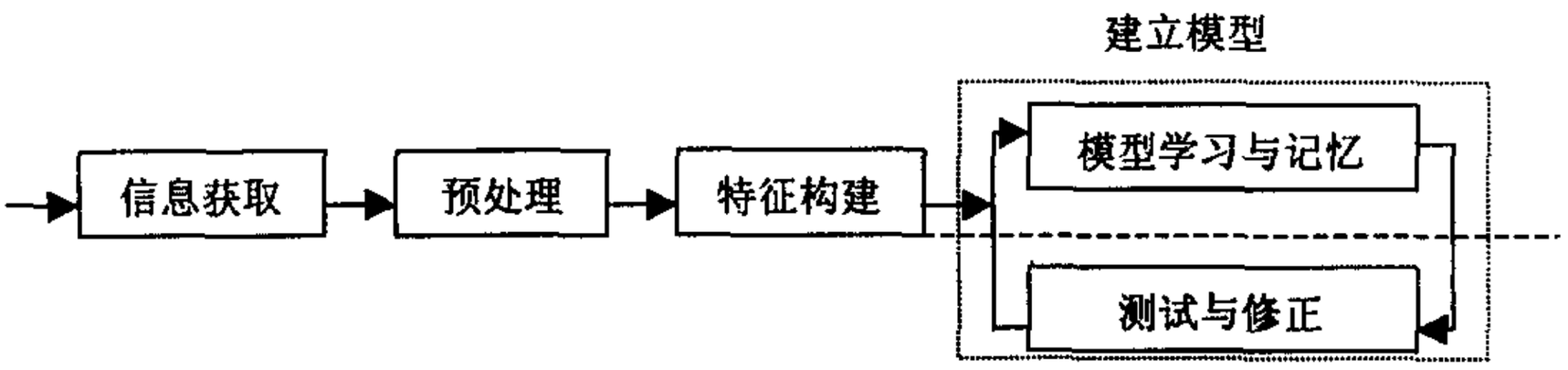


图 6-1

本文用的原始数据是 NCEP 资料。选用的气象因子为 500hpa 的高度场、700hpa 的风场（u 和 v）、850hpa 的位温场。样本的原始特征共计 855 维。这样高维的数据直接用来建模是不现实的，经重新构建得到 40 维特征，形成建模样本。预报模型用的是 BP 网络。通过反复的训练、测试，得到一个相对较好的模型，达到了一定的预报效果。

存在的问题是：①在现有特征下部分不同类样本交错纠缠在一起；②BP 网络优化问题。这些都影响着本模型的预报准确率，因此有待在以下方面展开进一步的研究。

1．特征的确定，需气象专家的更多经验和试验。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。

2．模型优化的进一步研究。

另：对于沙尘暴的短时预报，还作了计及站数的预报试验，即训练神经网络的样本输入量不变，理想输出量为当日报出沙尘暴的站的个数，对于神经网络来说就变成了输出值为连续值的形式。经试验，拟和和试报效果均不是很理想。我想原因可能是站数相同的沙尘暴日天气形势不一定就相似，单一的站数还不能作为衡量沙尘暴强度的标准，还要进一步结合具体地

区、具体情况对资料作更为详细的分析。

§6.2 进一步讨论

一、多网络决策系统

在第五章第 1 节提到，由于有部分样本在现有特征向量下两类交错纠缠在一起造成了预报准确度的下降。就此问题，本文提出多网络组合决策的方法。整体框架见图 6-2。

图 6-2

资料预处理可使用聚类分析的方法，但是分析的目的是为了分清全体沙尘暴日的类型，所以聚类对象扩展为全体沙尘暴日（包括小站点），综合各气象因子共分得 K 个模式。

同时用两层网络代替原来的一个 BP 网络，其方案如下图所示：

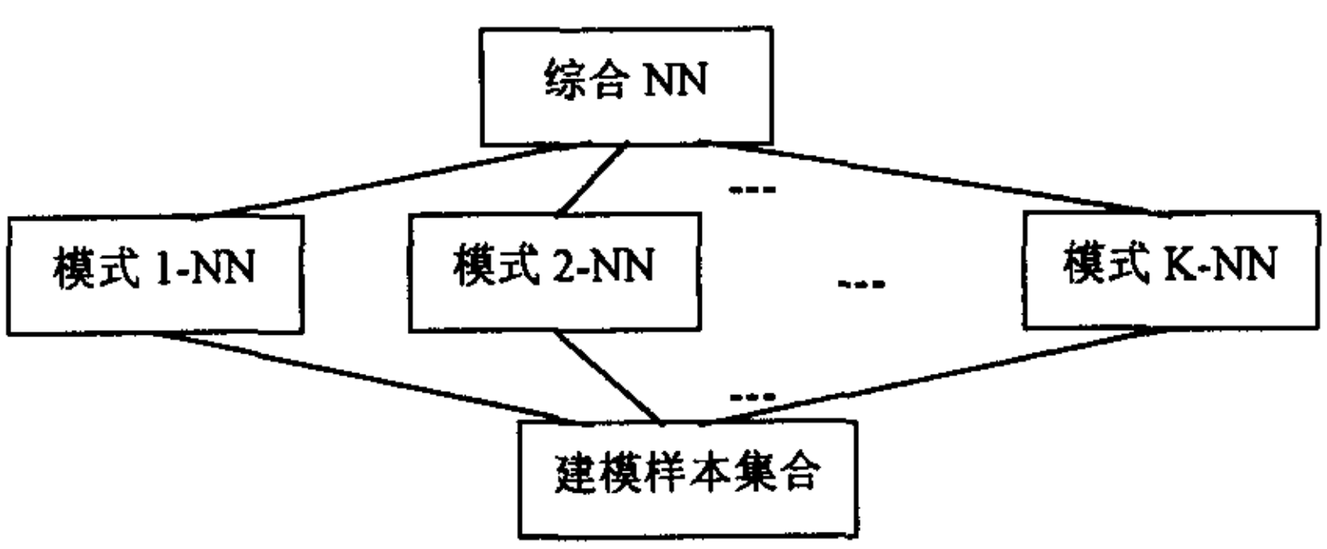


图 6-3 多网络决策系统

在图 6-3 中，第一层共有 K 个子网络，每个网络对应一个模式，判断是否产生了该模式的沙尘暴，而后提供给第二层的综合网作判断并产生相应的输出。

这个方案与单个神经网络模型相比，有以下几点优于原模型：

（1） 有效利用学习样本。对第一层的各个子网络来说，一个样本对其中一个网络是正例，对其他网络则为反例，因此相对增加了每个子网络的训练样本数。

（2） 样本的特征有所改善。如仍使用第四章所讲的特征构建方法，对于每个子网络属于该模式的样本和非该模式的样本特征差异，较原来的沙尘暴日

与非沙尘暴日的两类特征差异明显了许多。

（3） 缩小了网络的规模和学习时间。由于样本的特征的改善，对于单个子网络来说它的训练样本集合复杂程度大为降低，由第四章关于神经网络的解释可知网络的复杂程度也相应下降。

二、专家系统与神经网络的结合（ES+NN）

专家系统[45][46]

用计算机模拟人脑功能的另一个思路，就是用自顶向下的分析方法，从要实现的功能出发，将功能分解成子功能，直至设计出算法来实现这些子功能。基于这种思路的被称为“符号主义”。专家系统（Expert System）是“符号主义”目前应用最成熟的领域。

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。结构可表示为：

知识+推理=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。

表 6-1 神经网络与专家系统比较

神经网络	专家系统
NN 主要优点： 1） 信息以分布方式存储于整个网络中，即使网络局部受损，也不会对整个网络造成很大影响，还可根据不完整或模糊的信息联想出完整的信息，导出正确的输出； 2） 具有自适应、自组织、自学习能力； 3） 可通过训练样本，根据周围环境来不断改变自己的网络，并根据变化的信息，调整自身的结构； 4） 具有并行处理特征，在完成训练样本后，运行速度很快； 5） 能从训练样本中自动获取知识。	ES 其不足之处有： 1） 知识获取费时费力，成为瓶颈问题； 2） 规则提取、知识库建立需要领域专家与知识工程师的密切合作，但有时很难找到合适的能够清楚表达领域知识的专家； 3） 推理不能适应变化的环境； 4） 由于产生式的串行结构，随着知识库规模的加大，推理效率可能急剧下降，在单机上很难提高运行速度，当规则量很大时，进行并行处理也很困难； 5） 开发周期长，一般需要 8~12 个月。
NN 的主要局限有：	ES 的优点：

1） 难以表达结构化知识； 2） 需要大量实例进行训练，且训练时间可能很长，可能陷入局部最小； 3） 难以处理事先未训练的异常情况； 4）系统以“黑箱”方式运行，解释系统的结论比较困难。	1） 可以清晰可读的类自然语言方式表达无法用数学模型表达的专家知识，便于理解及知识库维护； 2） 能在特定领域内模仿专家工作，处理非常复杂的情况，包括异常情况； 3） 在已知其基本规则的情况下，无需输入大量细节数据，即可运行； 4）能对系统的结论做出解释。

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，一定能大大提高预报效果。专家系统与神经网络的结合，现在有许多人在作这方面的研究，这是今后一段时间人工智能发展的趋势。

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[40]，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。

- 串行相接法：框架如图 6-4

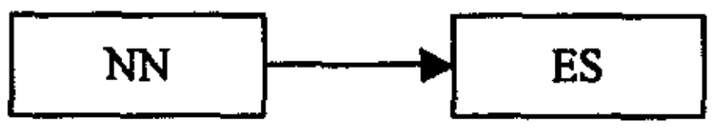


图 6-4 串行相接

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，前面选的 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。这时 NN 报出的准确率并不是很满意。

现在用 NN 后串接 ES 的系统来实现对沙尘暴的预报。

当单独用 NN 预报时，总是有些沙尘暴日被漏掉和一些非沙尘暴日被错报(见第五章)，应采取措施尽量避免之。

首先，通过组织训练样本集合的内容，使 NN 偏向于报 1，不再漏报，这时系统的输出中空报增多，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。

- 协同式

  系统中的各个模块并列存在，相对独立工作。见下图。

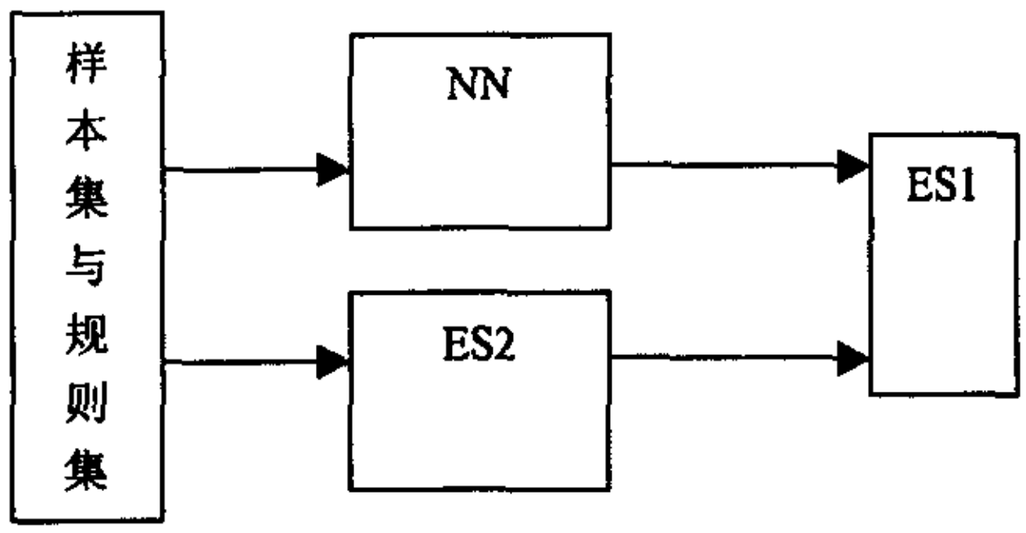


图 6-5 协同式

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存入的知识，如系统使用地区位于大区域内，神经网络作为大的天气背景就不必修正了。

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作的 NN 转向 ES 的工作。

三、业务化构想

下面对基于神经网络的沙尘暴预报模型的业务化提出一些构想。

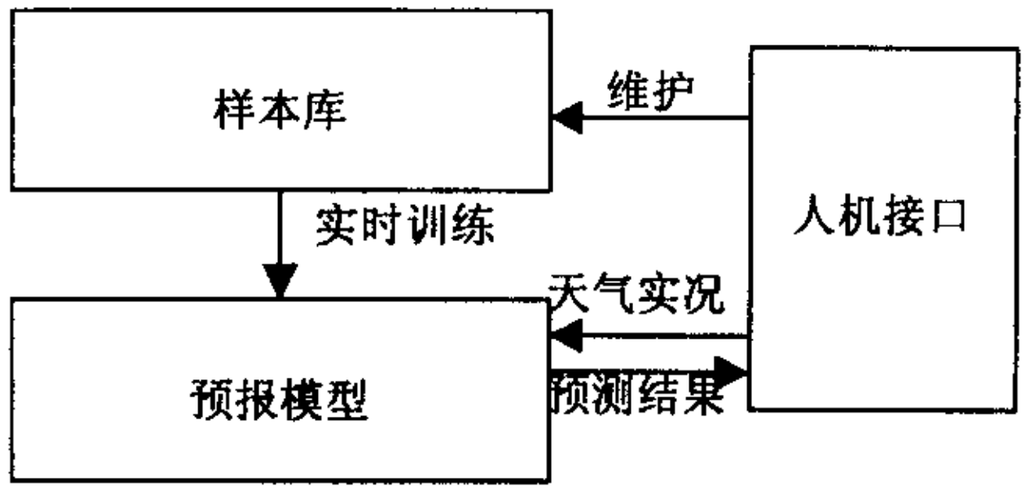


图 6-6 沙尘暴预报系统框架

沙尘暴预报模型准确率达到一定程度时，便可用于预报业务。为了方便非计算机专业人员的使用，首先需要开发一个友好的人机接口，操作人员只需输入需要的天气实况，通过接口进入训练好的神经网络，预报模型对有无沙尘暴做出判断，通过人机接口反馈给预报员。

建模样本库的维护

随着时间的推移，大的气候环境发生变化，样本库中个别历史样本对模型的贡献下降，所以建模样本库需要经常维护。

这包括两个方面：新样本的入库和库中旧样本的剔除。

因为沙尘暴属于小概率事件，样本库中正例相对较少，当新发生沙尘暴时，可让其及时入库。为了保持样本库的规模，库中年代较远的不典型样本，可适当剔除。当样本库内容变化后，触发对网络训练的操作，更新预报模型，提高模型的适应性和准确率。

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巨系统，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。本文主要就运用人工神经网络技术于“沙尘暴预报”展开研究。

本文利用聚类分析技术对沙尘暴相关资料进行分析和预处理，通过样本特征构建，形成建模样本；在此基础上，研制基于人工神经网络的沙尘暴预报模型，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，经测试，预报模型达到了一定的预报效果。最后，对进一步提高预报模型的准确率提出多网络决策系统、神经网络与专家系统结合的解决方案，并就预报系统的业务化提出实现框架。

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。

最后祝愿本学科组和国家气象中心能够合作的更为愉快。

致谢

本论文是在导师王萍副教授和林孔元教授的悉心指导下完成的，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！

在本课题的研究过程中，自始至终都得到刘还珠主任许多具体的指导，刘主任对待科研事业的执著态度，是我们许多年轻人学习的榜样。

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。

同时感谢国家气象中心数控室的王品成老师，宋振鑫博士，黄卓先生在样本资料的提取、检验上给予的帮助，以及在气象知识上给予的指导。

最后还要向帮助过我的课题组同学和国家气象中心的同志表示真诚的谢意！

附录 原始数据样式

500hpa 等压面高度原始数据 单位：m

saH5 19900211

5689.0000 5702.0000 5706.0000 5706.0000 5707.0000 5707.0000 5705.0000 5700.0000
5694.0000 5690.0000 5690.0000 5694.0000 5697.0000 5699.0000 5694.0000 5686.0000
5675.0000 5659.0000 5640.0000
5653.0000 5668.0000 5675.0000 5678.0000 5678.0000 5677.0000 5673.0000 5667.0000
5661.0000 5658.0000 5660.0000 5664.0000 5668.0000 5668.0000 5663.0000 5653.0000
5638.0000 5615.0000 5585.0000
5627.0000 5641.0000 5648.0000 5648.0000 5645.0000 5639.0000 5633.0000 5627.0000
5624.0000 5624.0000 5627.0000 5631.0000 5634.0000 5634.0000 5628.0000 5615.0000
5592.0000 5557.0000 5511.0000
5597.0000 5610.0000 5615.0000 5613.0000 5607.0000 5599.0000 5593.0000 5590.0000
5591.0000 5594.0000 5599.0000 5603.0000 5604.0000 5601.0000 5591.0000 5570.0000
5535.0000 5485.0000 5425.0000
5554.0000 5565.0000 5570.0000 5570.0000 5567.0000 5563.0000 5562.0000 5564.0000
5569.0000 5575.0000 5579.0000 5580.0000 5578.0000 5569.0000 5551.0000 5518.0000
5469.0000 5408.0000 5341.0000
5507.0000 5517.0000 5523.0000 5529.0000 5533.0000 5537.0000 5543.0000 5550.0000
5557.0000 5561.0000 5561.0000 5556.0000 5546.0000 5527.0000 5497.0000 5453.0000
5396.0000 5333.0000 5270.0000
5479.0000 5487.0000 5495.0000 5504.0000 5512.0000 5522.0000 5531.0000 5539.0000
5542.0000 5540.0000 5532.0000 5518.0000 5497.0000 5468.0000 5428.0000 5378.0000
5322.0000 5265.0000 5216.0000
5467.0000 5474.0000 5482.0000 5490.0000 5498.0000 5505.0000 5511.0000 5514.0000
5511.0000 5501.0000 5485.0000 5462.0000 5432.0000 5395.0000 5352.0000 5304.0000
5256.0000 5213.0000 5180.0000
5439.0000 5448.0000 5454.0000 5459.0000 5463.0000 5465.0000 5464.0000 5459.0000
5450.0000 5435.0000 5414.0000 5387.0000 5355.0000 5319.0000 5281.0000 5242.0000
5207.0000 5178.0000 5159.0000

700hpa 等压面纬向风速原始数据 单位：$m \cdot s^{-1}$

sau7 19900211

3.4300 1.7300 1.9600 3.6000 5.5300 6.9300 7.3500 7.1600
6.6600 5.7800 4.4500 3.0000 2.4300 3.5300 6.2500 9.4100
11.7100 12.7300 13.4300
1.8000 0.7600 1.5300 4.0000 7.0300 9.1300 9.6800 9.0100
8.0000 7.2100 6.5100 5.8100 5.5100 6.3300 8.4300 11.0800
12.9800 13.6300 13.6800
3.3800 2.6300 3.0100 4.8600 7.4300 9.4800 10.1800 9.8000
9.1800 8.9600 9.0300 8.8300 8.3800 8.3000 9.0800 10.5800
11.9600 12.6100 12.8600
7.3600 6.7500 6.0800 6.2100 7.1300 8.1100 8.6300 8.5800

8.6100 8.9600 9.4100 9.4600 8.9800 8.5300 8.8100 9.9300
11.2500 12.1800 12.6100
9.6600 9.4600 8.4800 7.5100 6.9100 6.6100 6.3500 6.1500
6.1300 6.4600 6.9100 7.1600 7.2500 7.6100 8.6000 10.2600
11.9300 12.8100 12.7600
7.7300 8.3300 8.1600 7.5500 6.7300 5.8600 4.9100 4.1100
3.5800 3.5600 3.9100 4.5800 5.6300 7.1300 9.1300 11.1800
12.6100 12.8100 11.6800
4.5300 5.6800 6.6300 7.0600 6.8800 6.1800 5.2000 4.2800
3.6600 3.6600 4.2800 5.4300 7.0300 8.8300 10.6100 11 8100
12.0300 10.9800 8.8100
5.2300 6.3600 7.5600 8.4300 8.7300 8.5300 8.0800 7.7300
7.8500 8.4100 9.4600 10.6300 11.7300 12.4800 12.6100 11.9300
10.4300 8.2300 5.6800
10.4800 11.1600 11.8000 12.2500 12.4300 12.4800 12.7100 13.2100
14.0300 15.0800 15.9800 16.5300 16.3100 15.2300 13.4100 10.9600
8.2500 5.6600 3.5100

700hpa 等压面经向风速原始数据 单位：$m \cdot s^{-1}$

sav7 19900211

11.7800 7.6600 5.5500 5.5600 5.8600 5.2300 3.8300 2.8300
2.8300 3.2800 3.5300 3.3800 3.1300 2.7800 1.6300 -1.0200
-4.8900 -8.4700 -10.1700
11.8000 7.8000 5.4300 5.1800 5.7300 5.6300 4.6800 3.7100
3.2500 3.0600 2.4600 1.3300 0.0600 -0.9700 -2.2500 -4.5900
-7.9200 -11.0500 -12.4700
10.7800 6.4800 3.1800 1.8800 2.0300 2.4300 2.5500 2.5300
2.5100 2.2500 1.3100 -0.2200 -1.8700 -3.3200 -4.9400 -7.4000
-10.6400 -13.7400 -15.1900
9.6800 5.5800 1.6300 -0.8400 -1.5500 -1.1400 -0.2500 0.6600
1.2500 1.1800 0.3800 -0.9700 -2.4400 -3.9200 -5.7200 -8.4200
-11.9000 -15.1700 -16.8700
8.6800 5.9300 2.4300 -0.3900 -1.8200 -1.8700 -1.0700 -0.0500
0.5600 0.4300 -0.4700 -1.7900 -3.1400 -4.5400 -6.3700 -8.9000
-12.0700 -14.9500 -16.4700
7.3500 6.3800 4.1800 1.8300 0.2800 -0.1700 0.1600 0.7100
0.7300 -0.0700 -1.6200 -3.5000 -5.3200 -6.9700 -8.5700 -10.3500 -12.1700
-13.6400 -14.2200
5.7300 5.6800 4.5000 2.9800 1.8000 1.3300 1.4300 1.4000
0.7500 -0.8700 -3.3200 -6.0700 -8.5700 -10.5200 -11.6900 -12.1900 -12.1400
-11.6700 -10.9000
4.4600 4.1300 3.2500 2.2800 1.5800 1.3800 1.3800 1.0000
-0.1700 -2.3900 -5.3200 -8.4400 -11.1200 -12.7700 -13.2000 -12.3900
-10.7200 -8.7400 -7.0900
3.6800 2.9100 2.1000 1.4100 1.0300 0.8600 0.5600 -0.2400

-1.7700 -4.0900 -6.7500 -9.3200 -11.1900 -11.8900 -11.3400 -9.6400
-7.2900 -5.0200 -3.3900

850hpa 等压面温度原始数据 单位：degK

saa8 19900211

278.5300 278.0300 278.2800 278.5600 278.3100 277.5300 276.8600 276.5300
276.7300 277.4100 278.3800 279.3300 279.6800 279.3300 278.6800 278.1300
277.6000 276.6000 274.8500
277.9300 276.3300 275.4300 275.2300 275.3800 275.4600 275.2800 275.0500
275.0300 275.4100 276.1300 276.8300 277.0600 276.7500 276.1000 275.4000
274.5000 273.1300 271.2300
278.1600 275.9000 273.7800 272.5800 272.3000 272.5300 272.8500 273.1300
273.4000 273.8300 274.2800 274.5800 274.5500 274.1500 273.4600 272.5300
271.2800 269.5300 267.3000
278.1800 276.2500 273.7800 271.7800 270.5500 270.1800 270.2500 270.5800
271.1000 271.7100 272.2300 272.5000 272.4300 272.0600 271.2800 270.0800
268.3600 266.0800 263.4100
276.1500 275.4600 273.8300 272.0000 270.3800 269.2100 268.5500 268.3800
268.6000 269.1800 269.8300 270.3100 270.4600 270.1600 269.2800 267.7100
265.4800 262.7300 259.8800
271.9600 272.4300 272.1600 271.2800 270.1000 268.8300 267.7300 266.9600
266.7300 267.0000 267.6300 268.2800 268.5800 268.1800 267.0000 264.9600
262.3000 259.4800 256.9800
268.3500 269.0500 269.4100 269.3800 268.9100 268.0800 267.0300 266.1000
265.6000 265.6600 266.1600 266.7100 266.8800 266.2500 264.6000 262.1000
259.2300 256.5500 254.7500
267.7100 267.8600 267.9600 267.9800 267.8000 267.3800 266.6600 265.8800
265.3100 265.1600 265.3300 265.5800 265.3800 264.3800 262.3500 259.6000
256.7100 254.3500 253.0000
268.8300 268.4300 268.1600 268.0000 267.8600 267.5800 267.0500 266.3500
265.5800 265.0000 264.6000 264.2300 263.5800 262.2500 260.1500 257.5800
254.9100 252.8000 251.6300

850hpa 等压面比湿原始数据 单位：$kg \cdot kg^{-1}$

sas8 19900211

0.0062930 0.0063810 0.0056900 0.0017360 0.0005310 0.0031300 0.0026480 0.0010150
0.0008580 0.0029750 0.0043980 0.0054630 0.0053050 0.0040850 0.0037730 0.0022410
0.0020110 0.0020230 0.0019930
0.0055430 0.0042280 0.0038780 0.0042250 0.0038430 0.0041360 0.0038660 0.0013150
0.0011160 0.0029930 0.0028630 0.0018230 0.0021610 0.0025430 0.0024330 0.0020810
0.0019260 0.0018710 0.0018580
0.0049060 0.0037830 0.0036530 0.0038630 0.0028700 0.0032030 0.0034950 0.0037830
0.0030430 0.0031300 0.0026750 0.0013730 0.0016630 0.0021800 0.0022610 0.0022310
0.0025210 0.0018130 0.0013460
0.0029480 0.0043300 0.0038030 0.0031230 0.0019600 0.0020110 0.0029980 0.0027380
0.0019430 0.0022180 0.0025530 0.0025260 0.0023300 0.0021000 0.0017960 0.0016950

0.0017760 0.0017180 0.0012410
0.0014580 0.0015750 0.0019130 0.0025680 0.0026230 0.0027230 0.0019030 0.0020930
0.0020100 0.0022410 0.0019760 0.0018210 0.0020180 0.0019600 0.0017950 0.0017580
0.0013000 0.0011780 0.0014130
0.0005880 0.0008860 0.0013060 0.0018030 0.0021130 0.0022580 0.0026980 0.0022610
0.0021110 0.0020680 0.0020310 0.0020380 0.0021480 0.0021130 0.0021280 0.0019480
0.0016280 0.0005900 0.0009030
0.0016560 0.0013930 0.0012510 0.0011280 0.0010760 0.0016330 0.0024330 0.0020010
0.0020030 0.0022800 0.0023460 0.0024680 0.0024580 0.0023580 0.0020180 0.0017510
0.0012680 0.0006950 0.0007980
0.0018130 0.0018530 0.0021830 0.0024230 0.0023860 0.0025730 0.0026800 0.0022980
0.0019250 0.0018130 0.0019450 0.0019400 0.0020760 0.0015230 0.0017210 0.0015850
0.0011780 0.0007830 0.0007480
0.0019610 0.0019150 0.0021310 0.0024380 0.0025130 0.0024030 0.0024280 0.0024930
0.0023060 0.0019210 0.0017280 0.0018750 0.0020810 0.0019480 0.0015860 0.0011330
0.0006660 0.0006680 0.0006280

参考文献

[1] 石纯一 黄昌宁等. 人工智能原理. 清华大学出版社 第 1 版 1993.10

[2] 邵军力等. 人工智能基础. 电子工业出版社

[3] 傅京孙 主编. 程民德 石青云等译. 模式识别应用. 北京大学出版社 1987

[4] 边肇祺 张学工等. 模式识别. 清华大学出版社 1999.12

[5] 杨先正 吴岷等. 模式识别. 中国科学技术大学出版社 2001

[6] 李介谷 蔡国廉等. 计算机模式识别技术. 上海交通大学出版社 1986

[7] 姚敏. 计算机模糊信息处理技术. 上海科学技术文献出版社 1999

[8] 维纳. 控制论(中译本). 郝季仁译. 科学出版社. 1952

[9] 陈翰馥 程代展. 人类对万物的驾驭术——谈控制论. 广西师范大学出版社 1999

[10] 林孔元等. 气象预报智能系统集成化问题的研究. 气象人工智能专辑（二） 1994:39-42

[11] 王耀生. 人工智能、模式识别在气象领域应用的现状与展望. 气象人工智能专辑（二） 1994:9-14

[12] [美]Abhijit S.Pandya and Robert B.Mary 著. 徐勇 荆涛等译 神经网络模式识别及其实现. 电子工业出版社 1999 年 6 月第 1 版

[13] 陈念贻. 模式识别优化技术及其应用. 科学出版社 2000

[14] 陈守余 周梅春. 人工神经网络模拟实现与应用. 中国地质大学出版社 2000

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

[16] 周远晖. 多策略学习方法的研究与应用. 清华大学博士论文 1998

[17] 郝为. 基于计算智能的多模型气象综合预报. 天津大学硕士学位论文 2000.1

[18] 杨洪敏. 智能气象预报系统中的相似建模方法研究. 天津大学硕士学位论文 2000.1

[19] 黄嘉佑. 气象统计分析与预报方法. 气象出版社 2000.3

[20] 余剑莉. 统计天气预报. 气象出版社 第 1 版 1994.12

[21] 孙军. 中央气象台的沙尘暴预报. 2001 年沙尘暴气象服务工作研讨会交流资料

[22] 王式功等. 沙尘暴研究的进展. 中国沙漠 2000 年 12 月:349-356

[23] 孙军 李泽椿. 西北地区沙尘暴预报方法的初步研究. 气象.2001.27(1)-19-24

[24] 郑新江等. 沙暴天气的云图特征分析. 气象 1995.21(2):27-31.

[25] 徐希慧. 塔里木盆地沙尘暴的卫星云图分析与研究[A]. 中国沙尘暴研究[C]. 北京:气象出版社,1997.88-91.

[26] 夏训诚 杨根生等. 中国西北地区沙尘暴灾害及防治. 气象 1996.10

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

[28] McNaughton D L. Possible connection between anomalous anticyclones and sandstorms[J]. Weather, 1987,42(1): 8-13.

[29] Carlson, T.N. and J.M. Prespero, 1972: The large scale movement of Sahara air Outbreak over the North Equatoril Atlantic J.A.M, 11,283-297.

[30] Iwasaka Y.Etal. The transport and spacial scale of Asian dust storm clouds. Acase study of the dust storm event of April 1979. Tellus B.35 189-196, 1983.

[31] WEN GAOJIYONG MA, SING LANGUAGE RECOGNITION BASED ON HAMM/ANN/DP, International journal of pattern recognition and artificial intelligence. –2001,14(5)-587-602

[32] Cho, K. Z.; Kang, Y. C. ; ; An ANN Based Approach to Improve the Speed of a Differential Equation Based Distance Relaying Algorithm. IEEE transactions an power delivery.

-1999,14(2)-349-357
[33] James C.Beidek and Sunkar K.palleds Fuzzy Models for Pattern Recognition, New York: IEEE press, 1992.
[34] Vladimir Cherkassky and Filip Mulier, Learning From Data: Concepts, Theory and Methods, New York: John Wiley &Sons, 1997.
[35] 曾晓梅. 国外人工智能技术在天气预报中的应用综述. 气象科技. 1999 年第 1 期: 4-10
[36] 倪志伟等. 神经网络专家系统及其数据裁决技术的探讨. 系统工程学报. 2001 年 2 月:61-65
[37] 陈成 K.W.Wong 等. 神经网络的自适应删剪算法及其应用. 物理学报. 2001 年 4 月:674-681
[38] 王振雷等. 神经网络过学习问题的统计学分析及改进算法. 东北大学学报(自然科学版). 2001 年 8 月:358-361
[39] 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
[40] 陈肇乾等. 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集 1994.11
[41] 王耀生. 智能预测系统设计及应用 气象出版社 1990.2
[42] 路广等. 数据仓库与数据挖掘技术在电力系统中的应用. 电网技术 2001 年 8 月:54-57
[43] 王钦军 薛林福. 数据挖掘技术及其在地学中的应用. 世界地质 2000 年 9 月:235-239
[44] 程道超. 智能气象预报系统中建模方法的研究. 天津大学硕士学位论文 1999.3
[45] 田胜丰 黄厚宽. 人工智能与知识工程. 中国铁道出版社 1999
[46] 赵瑞清. 专家系统原理. 气象出版社 1987
[47] 丁士晟. 多元分析方法及其应用. 吉林人民出版社 1981
[48] 任冬梅. 基于过程理解的多模型预报系统的研究与应用. 天津大学硕士学位论文 1999.1
[49] 刘勇. 卫星云图数据库的研究. 天津大学硕士学位论文 2000.1
[50] 吕申. 集成化天气过程理解系统的研究. 天津大学硕士学位论文 1995
[51] 程彦. 卫星云图形态特征识别系统与智能数据库的研究. 天津大学硕士学位论文 1997
[52] 郭爱民. 卫星云图处理与识别的研究. 天津大学硕士学位论文 1996
[53] 王长桥. 卫星云图/台风云图的纹理特征分析方法研究. 天津大学硕士学位论文 1997
[54] 薛俊韬. 锋面、涡旋云系的处理与识别系统. 天津大学硕士学位论文 1998
[55] 宋平. 卫星云图处理及其业务化系统实现. 天津大学硕士学位论文 1999
[56] Bruce Eckel 著. 刘宗田等译. C++编程思想. 机械工业出版社 2000.1
[57] 蒋长锦. 科学计算和 C 程序集. 中国科技大学出版社 1998
[58] 勒蕃. 神经计算智能基础原理• 方法.
[59] 王碧泉 陈祖荫. 模式识别理论、方法和应用. 地震出版社 1989
[60] 黄德双. 神经网络模式识别系统理论. 电子工业出版社 1996 年 5 月第 1 版

天津大学硕士学位论文

基于神经网络的沙尘暴预报模型的研究与应用

（A Study of Dust storm Forecast Model Based On ANN and Its Application）

（大摘要）

专　　业：模式识别与智能系统

研 究 生：岳　斌

指导教师：王萍 副教授

林孔元 教授

天津大学电气自动化及能源工程学院

2002 年 1 月

引言 人工智能和模式识别技术日益发展，它们越来越多的应用到各个具体领域来解决各种各样的问题。本文所作的研究工作，正是围绕着“沙尘暴预报”这一具体问题展开的。

一、 问题分析

近年来，在我国沙尘暴、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，对社会经济和环境均造成一定程度的危害。为此，气象部门把沙尘暴预警预报列为重点的研究对象。

专家研究指出，沙尘暴形成的基本条件，一是大风，二是地面上裸露沙尘物质，三是不稳定的空气，即强风因子、沙源因子和热力不稳定因子。三者同步出现，便可能产生沙尘暴。

国家气象中心提出沙尘暴短期预报思路：首先是天气学分析，在确定出现有利于发生沙尘暴的形势的情况下，通过数值模式求算沙尘暴指数，进而判定沙尘暴发生的可能性。

“一般来说，模式识别问题指的是对一系列过程或事件的分类与描述。要加以分类的一系列过程或事件可以是一系列物理的对象，或者是一些比较抽象的如心理状态等。具有某些相类似的性质的过程或事件就分为一类。”若将所有的沙尘暴日称作沙尘暴类，所有的非沙尘暴日称作非沙尘暴类，用科学的方法将隐含在大量历史样本中类间差异的规律总结出来并给予恰当地描述（模式建模）。再根据观测到的信息，预报未来的天气事件应该归入沙尘暴类还是非沙尘暴类，以及沙尘暴可能发生的时刻和地区，可见沙尘暴天气的预报是典型的模式识别过程。本文主要就沙尘暴预报模式建模问题展开研究。

二、 智能模式建模

传统的模式建模方法包括统计模式建模法和结构模式建模法。人工神经网络技术、模糊数学方法的引入形成了智能模式建模法。本文选择基于人工神经网络技术的智能模式建模法建立沙尘暴预报模型。

图 1 是人工神经网络模式建模的一般过程，与传统的模式建模过程基本一致。作为模式分类器使用的人工神经网络，有它自身的特点，与传统的模式建模过程又略有不同。

首先，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，一般经过样本特征构建后，不再需要特征提取和选择，可直接进入人工神经网络的训练。

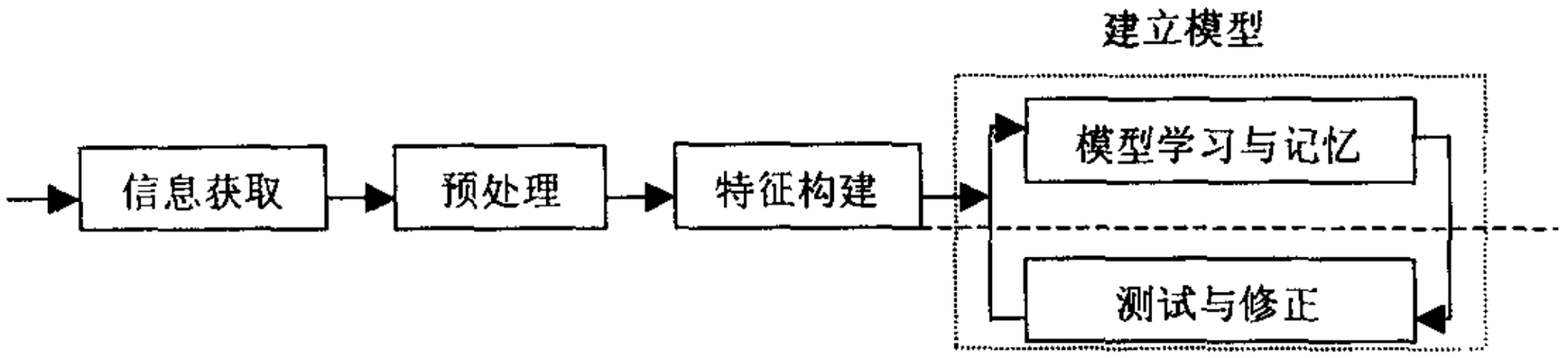


图 1 人工神经网络用于模式识别一般过程

其次，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。正因为它的功能过于强大，在建模初期模型就具有更多的不确定因素，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，从而建立更优的模型。即使其中的一个建模过程，也伴随有不断的学习和记忆，通过大量的建模样本的训练，得到正确的认识。

三、 建模样本的形成

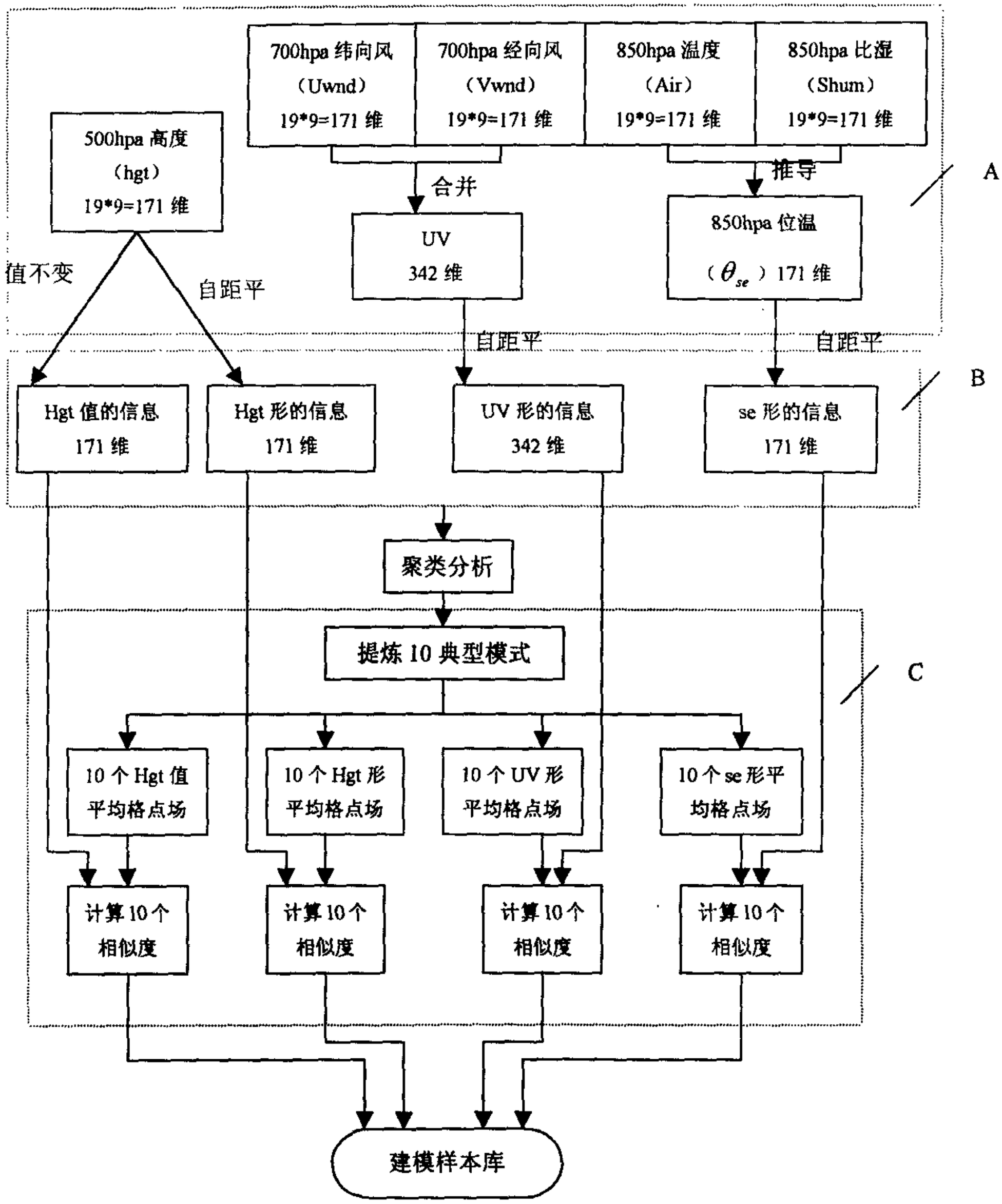
从大量的沙尘暴相关资料中挖掘出有用的建模信息，需要基于数据挖掘的思想并合理运用统计方法和模式识别技术，这是一个细致而具有艺术性的工作。

本文主要使用了聚类分析方法，经资料预处理→信息提取→分析→特征构建形成建模样本（参见图 2）。

本文用的原始数据是 NCEP 资料。选用的气象因子为 500hpaHgt、700hpaUV、850hpaθ_{se}。首先，提取出高度场 Hgt 值的信息、高度场 Hgt 形的信息、风场 UV 形的信息以及位温场θ_{se}形的信息。再综合考虑对强沙尘暴日样本集合各因子聚类分析结果，找到 10 个典型沙尘暴模式。计算各样本与 10 个典型模式的相似度构建样本特征，从而使通过格点场描述的原始样本 $X_0=(x_1,x_2,\cdots,x_{855})$（171（高度场值）+171（高度场形）+342（风场形）+171（位温场形）=855）改变为通过与典型沙尘暴模式值和形的相似度描述的建模样本 $X=(x_1,x_2,\cdots,x_{40})$。

四、 基于神经网络的沙尘暴预报模型

沙尘暴预报问题具有小概率、多因子、高维、样本数据量大的特点，目前气



A：原始数据　B：信息提取　C：特征构建

图 2 建模样本形成过程示意图

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用于沙尘暴预报。

一些对神经网络有研究的学者对 BP 网络有下面的评价：“从哲理的角度来

看，那些有明确因果关系表达式的问题（例如矩阵变换、Fourier 分析等），是不必要通过神经网络学习来解决的；那些毫无规律可循的问题（如完全随机的预报问题）也是不可能用神经网络学习来解决的。BP 学习网络所擅长的，是处理那种规律隐含在一大堆数据中的映射逼近问题，特别是需要通过学习自适应可调的实时性问题，例如模式识别，自适应控制和模糊决策等。”

本文选用 BP 网络作为沙尘暴预报模型。

建模过程是个反复训练和测试的过程。通过训练，网络归纳、总结出训练样本集合中隐含的规律，并分布记忆在网络的内部节点和连接之中；测试的目的是为了检验训练模型是否能够达到预期效果，根据测试结果进行模型优化。

五、 模型优化

网络的拓扑结构决定模型的复杂程度，性能较好的网络应具有尽量少的隐层及适当的隐层节点。常见的拓扑优化算法有 PLD 算法和试探法，本文使用试探法；样本的代表性和有效性决定模型的实用效果，本文引入 KNN 法编辑训练样本并分析试报结果；学习率的大小决定学习进程的快慢，同时影响网络拟和的精度；训练应适时而止，过多的迭代，反而可能使模型失去推广能力。

网络的拓扑、样本集合、训练参数这几方面对 BP 网络性能的影响不是独立的，他们互相影响、环环相扣，一个因素变了，其它因素也要相应做出调整。因而，需要多次的试探，这是个需要花费大量时间和精力的过程。

本文经过模型优化，确定了几个较优的模型，经测试均达到一定的预报准确率（见表 1）。

表 1

模型	试报结果				
	空报	漏报	正确报有	正确报无	CSI
模型 0(优化前)	30（14%）	18（72%）	7（28%）	184（86%）	12.7%
模型 1(优化后)	26（12%）	10（40%）	15（60%）	188（88%）	29.4%
模型 2(优化后)	20（ 9%）	14（56%）	11（44%）	194（91%）	24.4%
模型 3(优化后)	33（15%）	10（40%）	15（60%）	181（85%）	25.9%
测试样本	总量：239；沙尘暴类：25；非沙尘暴类：214				

神经网络采用“黑箱” 策略，导致结论的难以理解，对权值进行信息挖掘，发现通过对输入层和第一隐层间的连接权值进行统计，可以定性地分析并评价各

样本特征的建模能力。

六、 问题与展望

本文通过资料的分析和处理→特征构建→模型建立→模型优化（如图 1），实现了基于人工神经网络的沙尘暴预报模型。经测试，预报模型达到预期要求。为进一步提高预报准确率，本文提出如下解决方案。

1．研制多网络决策系统。该系统在原有样本特征下，充分利用学习样本，使类间差异更为突出。

2．将专家系统与神经网络相结合。专家系统与神经网络优势互补，专家系统的加入可引入规则经验，并有助于对结论的理解。

最后，本文提出系统的业务化的构想。业务化系统由人机接口、样本库、预报模型等模块构成，并通过对建模样本库的维护，预报模型的实时训练，提高模型的适应性和准确率。

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。