分类与集成方法在降雨预报中的应用

曹晓钟[1] 闵晶晶[2] 刘还珠[3] 赵声蓉[3] 王式功[2]

(1. 中国气象局培训中心,北京 100081; 2. 兰州大学大气科学学院; 3. 国家气象中心)

提 要: 介绍一种利用数值预报产品进行降雨预报的方法。该方法按照人工智能分类与集成的思想,利用前馈神经网络将T213、日本、德国的数值预报产品集成在一起,构成一个集成型的预报系统。在此基础上,利用高度场的天气形势和预报区域近低层流场和温湿条件,采用自组织神经网络进行天气分型,并针对不同的天气类型选用不同的预报因子,建立不同的预报模型。按照上述方法,选用江淮流域68个站点2003—2005年的5—9月数据,逐站建模,用2006—2007年5—9月的数据进行分级降水试报。各级降水预报结果表明,集成多家数值预报信息好于仅用单一模式的信息,采用天气分型建模优于不分型的建模。因此,多模式(型)预报结果的综合集成方法的研究,是数值预报解释应用中很值得探索的方向。

关键词: 数值预报 降水 聚类分型 神经网络 集成

Application of Classification and Integration to Rainfall Forecast

Cao Xiaozhong[1] Min Jingjing[2] Liu Huanzhu[3] Zhao Shengrong[3] Wang Shigong[2]

(1. Training Center, China Meteorological Administration, Beijing 100081;
2. Lanzhou University; 3. National Meteorological Center)

Abstract: A method of forecasting rainfall based on numerical prediction products is presented. According to the idea of artificial intelligence classification and integration, numerical prediction products of T213, Japanese and German models are integrated together by using the Back-Propagation neural network, and it will contribute advantages of various means and form an integrated forecast system. On this basis, self-organizing neural network is used to classify the weather type according to the situation of height field and temperature and humidity on the surface layer in forecasting area. Then different forecast elements are selected and different forecast models are established for different weather types. Using the method mentioned above, the forecast model is built by using the data from May to September in

资助项目: 本文得到中国气象局重点课题:"天气要素精细预报业务系统建设与改进"的资助
收稿日期: 2008年3月4日; 修定稿日期: 2008年3月17日

2003 to 2005, and is tested by forecasting rain-fall from May to September in 2006 to 2007 at 68 stations in the Changjiang-Huaihe River basin. The result shows that the method is quite practicable.
Key Words: numerical weather prediction rainfall classification neural network integration

引 言

近20年来,在数值预报产品提供的大量信息基础上,预报员根据前期实况观测及其演变所蕴涵的天气动力学特点对数值预报产品进行修正、解释,使之达到人们对气象要素预报的要求。很多气象工作者利用统计方法[1-3]和一些动力诊断方法[4]对数值预报产品进行释用,使定时、定点、定量的要素客观预报无论在预报种类或是在预报时效上都上了一个台阶,预报质量也得到了较大的提高。特别是对具有连续性函数特点的要素,如最高、最低温度和相对湿度的预报效果较好,但是对于具有非线性特点的降水定量预报,效果尚不够理想。由于某地发生的降雨是大尺度环流与中小尺度系统相互作用的综合结果,同时也是本地流场和热力场与当地的地形、地貌相结合的产物,正是由于存在这样一系列复杂的物理过程,因此,目前对降雨的定量预报除了依赖数值预报模式质量提高以外,对数值模式产品的释用技术也提出了更高的要求。

近年来,对降雨预报,不少气象工作者在预报方法和预报技术上的各环节都进行着不断地改进[5-6],使降水预报得到了提高。但这些工作虽在数值降水预报的释用方面有了一定的进展,但仍存在不足。如对不同原因引起的降水都笼统运用同样的预报因子,或只进行主观的天气分型,使降水预报的效果受到一定的影响。事实上,日常天气预报中,不同环流形势下,预报的着眼点是有差异的,考虑天气变化的因素也是不相同的。由于不同的天气类型所产生的降雨机理不同,所涉及到的预报因子和预报模型也会有区别。在一般天气分型中,多数是仅使用1~2个物理场,如500hPa或850hPa高度场、温度场等来简单地划分类型,而对于降水预报来说,同样的天气形势下,不同的温湿条件,产生的天气会有很大的差别。因此既要能综合考虑高度场天气形势和预报区域近低层温湿条件,又要能客观地进行天气分型,寻求具有这样功能的分类方法是当前重点考虑的一个方面。

此外,预报员每日都能收集大量的气象预报信息,如天气在线网上德国的温度、降水和风等要素预报,还有日本及我国数值模式(T213)输出的各种要素预报,在释用这些数值预报产品时,如何从大量的信息中取其精华、弃其糟粕,综合集成各种预报产品的优势,得到最优的结果,是实际预报中重点考虑的另一个方面的问题。

在解决了上述两个问题之后,该文利用T213数值预报产品,制作定量降水预报试验。由于年代和样本数的限制,并且数值预报对暴雨以上的定量降水质量较差,因此,这里仅试报了有无降水的晴雨预报、≥10mm和≥25mm的降水预报。

1 天气分型与预报集成论述

1.1 天气分型

天气变化过程非常复杂,天气类型多种多样,在某个天气条件下,或受某种异常气候因素的影响,其天气变化规律与平时天气变化的特点迥然不同。不同天气形势下,预报因子可能不同,每一个预报因子在不同天气情况下,所起的作用也会不同,很难用一个天气模型来描述变化规律差别很大的天气过

程。因此,有必要对不同变化规律的天气过程用不同的模型来描述,即可采用多个模型预报的方法,如图 1 所示。

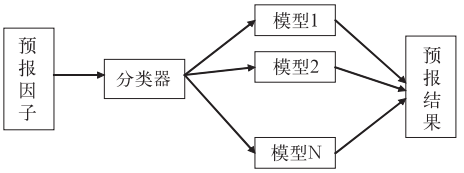

图 1　天气分型示意图

图中分类器是根据预报因子的特征值,把天气划分为不同的类型,对每种不同类型的天气用不同的模型去描述,每一个模型描述一种类型的天气,将使模型的描述更加精确、客观。分类器的选择是确定模型划分的一个重要因素。在实际应用中,分类器既可由有丰富经验的专家来完成,也可采用某种有效的客观分类方法。相关匹配法是一种有效的客观分类方法,即通过在预报因子的特征空间中计算输入特征向量和各模板特征向量之间的距离来进行分类。若将模式样本看作 N 维特征空间中的点,那么类别相同或某些特征相似的模式在 N 维空间中也比较靠近。例如同属于 A 类模式的点之间的距离比起它们中的任何一个与属于 B 类模式的点之间的距离要小。而各类型模板特征向量的值可通过样本的无监督学习的方式进行聚类划分得到。

1.2　多方法的综合集成

1980 年代末,著名科学家钱学森提出综合集成法(metasynthesis)[7],其核心思想是将多种方法包括人的经验和判断,通过计算机集成在一起,充分发挥各方面的优势,以得到一个更优的结果,这一方法对我国人工智能的研究影响较大。

在实际应用中,每一种方法都有其各自的优点、缺陷和不同的适用范围。一种方法之所以能够得到应用并得到发展,必然有其不可替代的优势,而这种方法又不能占有绝对优势而排斥其它方法,因此研究如何将不同的方法有机地结合起来,以充分发挥各自的优势,克服单个方法的缺陷,从而构成集成型的预报方法。

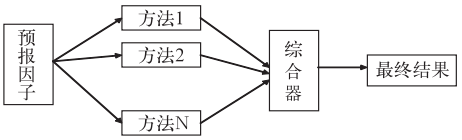

图 2　集成方法综合器的结构框图

集成方法综合器如图 2 所示,是把各种方法的输出结果进行综合集成,其目的是通过对多个互补结果的集成,得到一个优于各种方法的最终结果。常用的一种综合器就是“投票”,即从各种方法中取大多数的“意见”。但有的时候,“真理掌握在少数人手中”,按投票的原则就不能得到正确的结果。其实,从数学上看,集成器可以看作是一个非线性函数,各种模式方法输出的结果是这个非线性函数的输入,该函数的输出就是最终集成的结果。但往往这种集成器的特性难以描述清楚,因此可以用一个具有非线性学习能力的神经网络来完成,通过已有样本的学习,就可“了解”和“掌握”各种方法的特长。综合器也可看作一个专家系统,根据以往的“经验”,按照各种方法的特点对其结果进行集成,以得到一个最优的结果。

2　实现方案

2.1　资料准备

本文以江淮流域的降雨为研究背景,采用 2003—2007 年 5—9 月江淮流域 68 个站点的数据资料,其中用 2003—2005 年 5—9 月的数据进行分类、训练建模,将 2006—2007 年 5—9 月的数据用于预报试验。

在资料选取上，首先利用国家气象中心2003—2007年每年5—9月逐日的T213数值预报产品作为基本因子资料。所使用的T213数值预报产品包括15层7个预报时效(00、12、24、36、48、60、72小时)格点场中的14个基本气象要素，包括：温度、高度、纬向风、经向风、垂直速度、比湿、相对湿度、海平面气压、地面温度、地面气压、10米纬向风、10米经向风、2米温度、2米相对湿度。利用这些基本气象要素，通过动力诊断得出反映降水的100多个气象物理量，如涡度、散度、位温等，以及平流项物理量和梯度项物理量，如涡度、温度等。此外还有从地面到某层的垂直累积上升速度、水汽通量、水汽通量散度和一些时间累积的物理量，然后利用双线形插值的方法将这些基本要素和扩充物理量插值到对应的站点上，建立起所需要的站点因子库。

实况数据集是采用MEOFIS系统[8]中的历史实况库，取2003—2007年逐日08时到次日08时的24小时降水量。

2.2 因子的选择

因子选取的好坏直接影响着预报的效果，在众多的因子中客观地选出较好的因子可以提高预报的质量。首先确定某站点的预报对象，然后计算该站预报时效(可跨前后1～2个时效)所对应的预报因子与相应预报对象之间的相关，从而挑选出一批与实况降水量相关系数较大的不同层次的各种因子。然后再通过逐步回归的方法，利用F检验在这批因子中选取其中相关最好的10～20个左右的因子，形成该站点的预报因子集。对于不同的站点、不同的时效以及不同的预报对象来说，所选出来的因子和因子个数是不一样的。

另外，由于预报因子之间的量级存在着差异，在建模之前，使用公式(1)对全部样本的每一个因子分别做归一化处理，使其归一化到[0,1]之间。

$$X'_{ij}=\frac{X_{ij}-\mathrm{Min}(X_j)}{\mathrm{Max}(X_j)-\mathrm{Min}(X_j)} \qquad (1)$$

式中X'_{ij}为标准化后因子值；X_{ij}为标准化前的因子值；$\mathrm{Min}(X_j)$和$\mathrm{Max}(X_j)$分别表示第j个因子的所有样本中的最小值和最大值。其中$i=1,2,\cdots,m$；$j=1,2,\cdots,n$；m为样本总数，n为因子总数。

按上述方法所选的因子，对江淮流域68个站点逐一对未来08时的24小时降水量进行24、48、72小时3个预报时效的神经元网络建模试验，选用的数据为2003—2005年5—9月的数据，并用建好的模型对2006—2007年5—9月进行晴雨和≥10mm、≥25mm(以下分别用0.1mm、10mm和25mm表示)3个级别的降水预报。

2.3 数值预报产品的集成

一般，预报员对国外，特别是对日本的降水要素预报比较信赖，李勇[9]从天气学角度，针对2007年6—8月过程，将T213模式与欧洲中期预报中心(ECMWF)模式及日本模式进行了比较。认为3种模式对亚洲中高纬环流形势的调整演变具有较好的预报性能。总的来看，ECMWF模式对各系统及要素的预报误差最小，ECMWF模式及日本模式预报比T213准确。这里比较2006—2007年5—9月日本、德国和T213的逐日68个站的24、48和72小时08时的24小时降水量预报的检验评分(图3)，可以看出，三家的0.1mm预报TS评分比较接近，日本的3个预报时效分别为0.48、0.48和0.45，德国和T213的24小时预报略高于日本，其余两个时效与日本接近，都具有空报多于漏报的特点(图略)。比较而言，德国的空报略低些。而对10mm降水预报，日本显然略胜一筹，3个预报时效的评分分别是0.38、0.33和0.25，高于T213评分0.7～0.11个百分点，

比较起来,德国的预报更低一些,主要是空报较多。对于 25mm 降水也有类似的特点,日本最好,3 个时效的评分分别是 0.23、0.20 和 0.13,T213 次之,德国较差。具体来看每个站点的预报,总体上具有上述特点,但不同的站表现也有差异,如 24 小时晴雨预报,德国对杭州(58457)3 个时效的预报都较其他两个模式好,而 T213 对南京(58238)和合肥(58321)24 小时报得较好。另外,德国对南京 24 小时的 25mm 降水预报也好于日本。由此可见,不同的数值预报具有不同的性能,较好的模式也不见得时时处处都预报得好,差一些的产品也有其优胜之处。如将这些信息包含到降水预报的建模中,可望对提高定量(级)降水预报有所帮助。

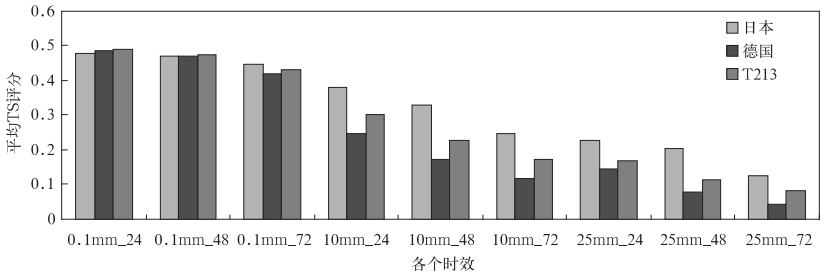

图 3　2006—2007 年 5—9 月日本、德国和 T213 的逐日江淮流域 68 个站点的 24、48 和 72 小时各降水级别预报的平均 TS 评分

因此,按照上述方法构造了综合集成预报系统(见图 4)。但在此系统中,虽然在不同的区域、不同的天气情况下、不同的数值预报产品会有不同的预报性能,可在何区域、何种条件下、以何种或几种数值预报产品的结果为主要依据,是难以说清楚的。图 4 中的集成器,在数学上,可以将它看作一个黑箱式的映射,它的特性很难用一个数学模型或一组规则描述清楚,用传统的方法很难构造一个合适的集成器,使其得到的结果能优于集成前的结果。

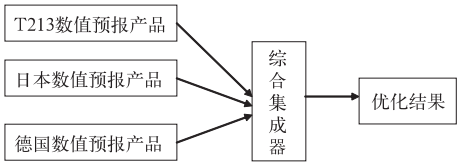

图 4　综合集成预报系统结构图

神经网络由于具有非线性的自学习能力,它能通过对历史样本的学习,将那些难以用数学模型描述的“黑箱”用神经网络来表达。因此这里在使用 T213 数值预报因子建模之后,又将日本、德国、T213 的降水预报信息也作为输入因子,与实况的观测数据作为训练的期望值,用 2003—2005 年 5—9 月的样本数据参加预报模型训练。这样就综合了德国、日本和 T213 多家的数值预报信息,形成多因子、多模式产品的集成预报数据集。在此基础上,用前述神经元网络方法建模,也对 2006—2007 年 5—9 月进行晴雨和≥10mm、≥25mm 降水预报。

2.4　天气分型

夏季江淮流域的天气受季风影响,对流层中低层西南季风和北方西风槽附近的偏北气流相遇于江淮流域,形成梅雨锋区,因而这一带多低涡、切变线活动,产生较大降水。一般,江淮流域的降水从影响系统来说,主要分为梅雨锋低涡切变降水、西风槽与南风暖湿气流之间的锋面降水、槽前暖切变降水及中小尺度对流性降水等,为了客观预报这一带的降水,在天气分型中,着重考虑 850hPa 的

风场和温湿场(用假相当位温表示),同时考虑700hPa的高度场,这样就比较全面地考虑了对降水最能产生影响的几个物理量场。

将上述4个要素场,每个格点都作为一个输入因子,通过聚类,得到若干类型的输出,这里根据天气预报经验和聚类的样本长度将夏季江淮流域的降水形势定为4种天气型。

聚类分型是一种数据分析方法,其目的是把大量数据点的集合分成若干类,使得每个类中的数据之间最大程度地相似,而不同类中的数据最大程度地不同。Self-Organizing Map(SON)网络是一种有效的聚类方法,它能将任意维输入模式在输出层映射成一维或二维图形,并保持其拓扑结构不变;网络通过对输入模式的反复学习可以使权重向量空间与输入模式的概率分布趋于一致。网络的竞争层各神经元竞争对输入模式的响应机会,获胜神经元相关的各权重朝着更有利于它竞争的方向调整,"即以获胜神经元为圆心,对近邻的神经元表现出"兴奋性侧反馈",而对远邻的神经元表现出"抑制性侧反馈"。通过近邻者相互激励,远邻者相互抑制,从而完成对数据的分类。具体算法[10]如下:

(1) 设 W_{ij} 为输入层第 i 个神经元节点到输出层第 j 个神经元节点的连接权值,对各 W_{ij} 随机赋初值并进行归一化处理,得到初始权向量$[W_{ij}]$, 其中 $i=1,2,\cdots,p$; $j=1,2,\cdots,m$;p 为输入层神经元个数, m 为输出神经元个数;建立初始优胜邻域 $NE_j(0)$;学习率为 η,赋初始值为0.1。

(2) 从 n 个样本组成的训练集中随机选取一组作为输入,记为$[X_k]$, $k=1,2,\cdots, n$。

(3) 计算与 W_{ij} 的点积,从中选出点积最大的作为获胜节点,记为 j^*。

(4) 以 j^* 为中心确定 t 时刻的优胜邻域 $NE_{j^*}(t)$,设置初始领域后,对域内权值进行训练,训练过程中, $NE_{j^*}(t)$ 随训练时间而逐渐收缩。

(5) 对优胜邻域 $NE_{j^*}(t)$ 外的节点权值保持不变,对优胜邻域 $NE_{j^*}(t)$ 内所有的节点调整权值:

$$W_{ij}(t+1)=W_{ij}(t)+\eta(t,N)[X_k-W_{ij}(t)]$$

$i=1,2,\cdots,p$; $k=1,2,\cdots, n$; $j\in NE_{j^*}(t)$,

其中 $\eta(t,N)$ 是训练时间 t 和优胜邻域内第 k 个神经元与获胜神经元 j^* 之间的拓扑距离 N 的函数。

通过SON算法,输出层存在一个权值与输入的节点最接近的节点,该节点就是此次迭代中竞争获胜的节点。随着邻域在迭代过程中线性减小,最终在对该输入产生最大响应的附近形成一个"聚类区",由此可分为几种不同的类型。通过分型,将各个类型的样本分别对4个物理量场进行平均,可将淮河流域5—9月的天气分为江淮暖湿强南风型、江淮暖湿切变低涡型、西低东高锋区南压型和槽前江淮切变型4种天气类型(图5)。从图5可以看出,通过自组织神经网络的分类方法划分出的4类天气类型,与江淮流域夏季天气的类型颇为一致,从天气学的观点再一次证明了自组织神经网络分类方法的正确和有效。

针对上述每种天气类型,再按2.3选择因子的方法和综合集成的数据集,对不同的天气类型选出10～20个与预报对象(如≥10mm降水)相关好的预报因子,并与上述的三家降水预报信息组成天气分型下的训练数据集。最后,运用神经元网络方法对不同的天气类型分别进行建模和试报。

3 试验评分结果

按上述3种实验方案(分别用"非集成"、"集成"和"分型"表示)的预报结果分别计算逐站、逐时效的各降水预报级别(晴雨、≥10mm和≥25mm)的TS评分、空报率和漏

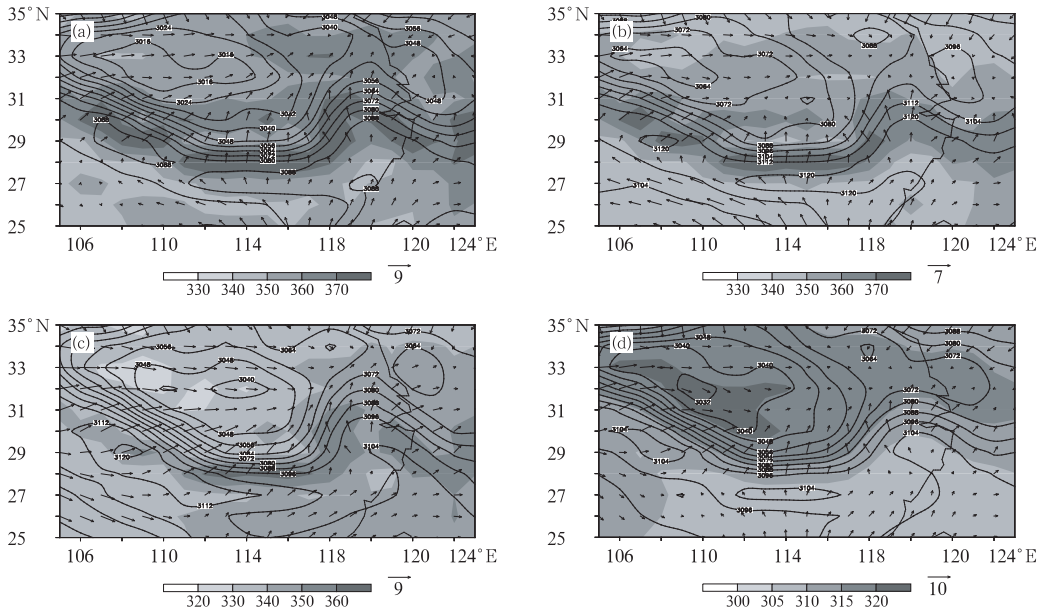

图 5　由 850hPa 风场、假相当位温场和 700hPa 高度场用自组织神经网络聚类分析得到的江淮流域夏季 5—9 月天气类型

(a) 江淮暖湿强南风型　(b) 江淮暖湿切变低涡型　(c) 西低东高锋区南压型　(d) 槽前江淮切变型

实线为 700hPa 等高线(单位:dgpm);流线为 850hPa 风(单位:m·s^{-1});阴影为假相当位温(单位:K)

报率,并将这些站点的评分结果进行总体平均,得到整个江淮流域降水预报的检验结果(见图 6)。由图 6 不难看出,与一般检验评分结果的规律一样,随着预报时效的延长和预报降水级别的提高,预报的 TS 评分逐渐降低,空报率和漏报率逐渐增大。比较 3 种试验结果可以看出,无论是哪个时效,无论是哪个降水级别,非集成模式的预报 TS 评分都是最低的,加入了日本、德国和 T213 降水的信息,集成预报的结果 TS 评分有了提高,除 24 小时 25mm 预报外,空报率都下降了。同时,72 小时和 48 小时的预报比 24 小时提高的更明显。天气分型后,预报的效果更好些,特别是 10mm 和 25mm 降水预报的漏报率有较大幅度的下降,仅 72 小时晴雨预报比不分型的预报稍大些。总体的预报效果是天气分型好于不分型,集成好于非集成的。

以上是天气分型总体的评分情况,由于各类型在 2006—2007 年 5—9 月间样本数不同,有的天气型较多,有的又较少,因此各类型预报的效果有较大的差异。图 7 给出各种天气类型下,集成方法对各降水级别预报的 TS 评分结果。由前述可知,分型以后各级降水预报都比不分型的效果好,因此各天气类型下,各级降水预报也大多数好于不分型的,但是,唯有第二型在晴雨预报和 10mm 降水预报中 TS 评分低于分型的平均,甚至还低于不分型的情况。由于这期间第二型(江淮暖湿切变低涡型)的天气个例较多,情况比较复杂,且中尺度对流性天气活动较多。而 T213 模式对这类天气预报能力较差,选取的预报因子反映这类天气形势下所产生降水的信息不够理想。如果将第二种天气类型再进一步细化,也可能使预报效果会提高一些。这说明客观天气分型中还有许多科学问题,对于预报水平的提高还有潜力可挖。

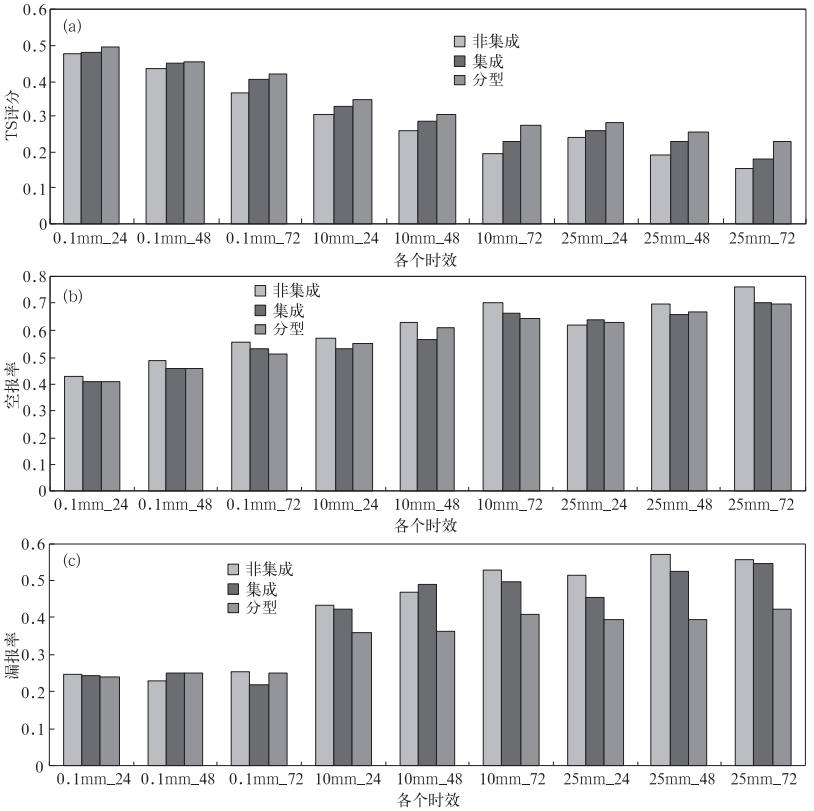

图 6 三种预报试验在晴雨≥10mm 和≥25mm 的 24、48、72 小时江淮流域 68 个站点预报平均评分结果

(a)TS 评分;(b)空报率;(c)漏报率

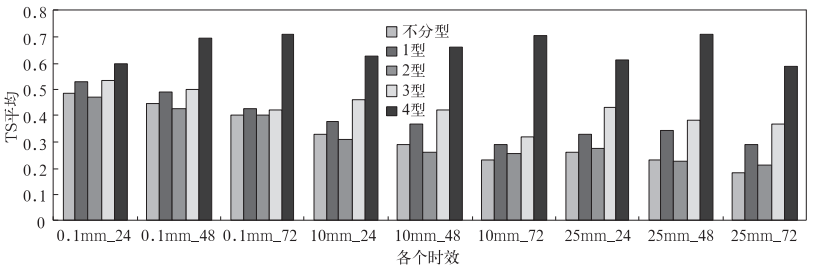

图 7 天气不分型和各天气类型下江淮流域 68 个站点的 24、48 和 72 小时晴雨预报及≥10mm 和≥25mm 的降水预报的平均 TS 评分

4 结论与讨论

(1) 采用 SON 神经网络方法进行聚类分型,选取了 850hPa 的风场和温湿场以及 700hPa 的高度场作为天气分型的依据,这样就比较全面地考虑了对降水产生影响较大的物理量场。试验表明,聚类分型的结果与实际天气类型较为一致,从天气学的角度证明了聚类分型的正确和有效。有些天气类型反

映的天气比较复杂,影响的因素较多,分型的效果不够理想,还需要进一步细化。所以在天气分型时用何种物理量场,分几种类型以及预报因子的选取等问题还须进一步研究,其中还有许多潜力可挖。

(2) 采用前馈神经网络建立的模型对江淮流域 68 个站点 2006—2007 年 5—9 月的降水进行试报的结果表明,用不同的模型预报不同天气类型的降水,其效果明显好于用一个模型预报所有天气类型的降水,表明了文中提出的对降水进行分型预报的可行性。同时,也进一步说明了降水预报的方法和手段应该向更精细化和更深入的方向发展。

(3) 对降水进行集成与非集成的预报结果显示,集成了不同数值预报产品的结果要好于没有集成的,将多种数值预报产品的更多信息应用到业务预报中,有望使预报水平得到进一步地提高。表明了多模式(型)预报结果的综合集成方法的研究,应是数值预报解释应用中很值得探索的方向。

参考文献

[1] 刘还珠,赵声蓉,赵翠光,等. 国家气象中心气象要素的客观预报－MOS 系统[J]. 应用气象学报,2004,15(2):181-191.

[2] 陆如华,何于班. 卡尔曼滤波方法在天气预报中的应用[J]. 气象,1994,9,20(9):41-46.

[3] 赵声蓉,曹晓钟. 神经元网络的降水预报. 暴雨落区预报实用方法[M]. 北京:气象出版社,2000:137-139.

[4] 姚明明,刘还珠,王淑静. 大降水预报动力诊断方法[M]. 北京: 气象出版社,2000:103-107.

[5] 陈力强,韩秀君,张立祥. 基于 MM5 模式的站点降水预报释用方法研究[J]. 气象科技,2005,10,31(5):268-272.

[6] 曾晓青,邵明轩,刘还珠,等. 基于交叉验证技术的 KNN 方法在降水预报中的试验[J]. 应用气象学报(待发表).

[7] 钱学森,于景元,戴汝为. 一个科学新领域——开放的复杂巨系统及其方法论[J]. 自然杂志,1990,13(1):3-10.

[8] 车军辉,李德生,李玉华. 数值预报产品释用业务系统历史数据存储与检索[J]. 应用气象学报,2006,17增刊:152-156.

[9] 李勇. 2007 年 6—8 月 T213 与 ECMWF 及日本模式中期预报性能检验[J]. 气象, 2007,33(11):93-100.

[10] Kohonen T. Self-Organizing Maps[M]. Springer Series in Information Science, 2001. 30.