基于聚类天气分型的 KNN 方法在风预报中的应用*

陈豫英[1)]　刘还珠[2)]　陈　楠[1)]　曾晓青[3)]　马金仁[1)]　刘迁迁[1)]　马筛艳[1)]

[1)](宁夏气象防灾减灾重点实验室，银川 750002)

[2)](国家气象中心，北京 100081)　[3)](兰州大学大气科学学院，兰州 730000)

摘　要

以模式识别和相似预报思想为基础，建立基于自组织神经网络(SOM)的聚类天气分型和交叉验证的 K 最近邻域非参数估计仿真模型(KNN)。该模型首先以自组织神经网络技术对西北地区的高空流场和高度场进行聚类分型，针对不同天气形势下的历史样本，通过交叉检验，分别寻求各类天气型下的最佳 K 组合。为了验证聚类天气分型对 KNN 方法的影响，使用 2003—2006 年冬半年 T213 数值预报产品和宁夏日最大风速资料，同时建立了宁夏冬半年日最大风速≥6 m/s 天气分型和未分型的 KNN 预报模型，并对 2007 年 1—5 月进行了预报试验，预报评估结果表明：天气分型后的预报模型总体上降低了预报空报率，提高了预报准确率，特别是某些类天气型，提高幅度更大，为分类相似预报开拓了思路。

关键词：自组织神经网络；聚类天气分型；交叉验证；K 最邻近域；日最大风速预报

引　言

在我国北方及沿海地区，平均风速≥12 m/s 的大风就会造成严重影响，甚至灾害。因此，大风预报在天气预报服务中占有相当重要的地位。

风是人们能直接感觉到的气象要素之一，它不仅受大中尺度风压定律所制约，也受边界层动力和热力湍流作用的影响，同时局地地形、地貌对当地地面风的影响也不可忽视，这一切给各地地面风预报带来了较大困难。不少台站利用多年积累风资料，寻找相关因子，并根据天气形势，用统计方法制作本地日常风的预报[1]。国家气象中心利用 T213 数值预报产品建立了日最大风速的 MOS 预报方法[2]。而对于大风预报，主要是根据历史形势场资料建立大风预报模型，然后根据数值预报的形势场作为预报因子来确定是否可能有大风出现，如杨忠恩等[3]、林良勋等[4]、胡波等[5]分别用人工神经网络和完全预报(PP)方法建立了大风预报模型，取得了一定的预报效果。但是，由于风要素的地域性和瞬变性强，特别是作为小概率事件的大风，使用这些客观方法的预报效果并不是十分理想。

KNN(K-nearest neighbor)非参数估计技术[6]是近几年来在数值预报释用中颇为重要的一种方法，它是基于范例进行推理的人工智能领域中发展较快的一种求解问题技术，利用过去的范例或经验来解决当前问题的类比推理方法，亦称为相似方法。由于气象要素样本较长，并且获取资料较为及时，这使得 KNN 技术得以在天气预报，特别是在定性要素的判别中发挥作用。该方法不需要建立预报方程，直接根据训练数据(历史天气样本)建立概率天气预报的 K 近邻非参数估计仿真模型，利用训练数据中蕴含的输入输出关系进行预报，可以避免统计方法的一些弊病和概率密度估计误差的影响。翟宇梅等[7]、邵明轩等[8]利用该方法制作云量及全国 600 多站点的风和降水预报，均有较好的效果。

某地某时的风是在天气系统的风压定律制约下，由当地地形、地貌影响以及边界层局地热力和动力条件相结合而激发的最终产物。在目前尚无法精确描述这些物理过程情况下，用 KNN 相似预报方法，认为相似条件下发生的“行动”会产生相似的结果，因此对于风的预报，应用相似方法是合理的。

曾晓青等[9]提出基于交叉验证的 KNN 方法，该方法在搜索 K 邻近域过程中，考虑天气事件出现的概率不同，利用交叉验证方法，分别求取有天气事件的正样本 K^+ 值和无天气事件的负样本 K^- 值作

* 中国气象局轨道建设项目“精细化气象要素预报业务系统(一期)”资助。
2007-09-11 收到，2008-07-28 收到再改稿。

为最佳邻近域的组合,并利用这一方法对我国不同代表站点的晴雨和不小于 10 mm 的降水预报进行试验,使降水预报评分得到提高。但这一工作没有考虑天气类型对降水的影响,即无论是何类环流背景、何种影响系统都提取同一组预报因子,显然会影响降水预报效果。本文在改进的 KNN 方法基础上,尝试从模式识别思想出发,针对风的预报提出一个基于自组织神经网络的聚类天气分型的 KNN 方法,以克服上述方法的不足。

1　资料及其因子处理

本文着重考虑我国西北地区的天气影响系统,研究区域为 35°～50°N,90°～115°E,所用资料包括 2003—2006 年冬半年的 1—5 月和 10—12 月及 2007 年 1—5 月 T213 数值预报格点场资料及宁夏 24 个测站日最大风速实况资料。其中,2003—2006 年样本作为训练数据(经过质量控制,剔除错误数据),2007 年样本作为预测检验数据。本文选定日最大风速≥6 m/s 为预报对象,一方面是考虑 4 级以上风对人们日常活动已经造成影响,另一方面日最大风速≥6 m/s 的样本数不至于太少。

T213 数值预报产品与文献[9]相同,包括 14 个基本要素 15 层格点场资料,利用这些基本要素通过动力诊断得到一些反映热量、能量、对流不稳定等的热力、动力因子以及一些时间累积因子,共有 888 个扩充因子。然后将这些因子通过双线性插值方法插值到所预报的站点上,建立站点因子库。

由基本因子和扩充因子构成的 T213 因子库种类繁多、数量庞大,在预报对象与预报因子单点相关普查的基础上,选取相关系数大而且相互独立的高相关因子,按不同站点不同时效建立 KNN 方法的基本因子库,通过逐步回归方法,经过 F 检验对因子进行排序筛选,剔除了一些与预报量相关不大而且物理意义不明显的因子,将最后入选的 10～20 个相关系数最高的因子组成站点预报因子集。这些因子在计算前均作了归一化处理。

2　预报方法

2.1　SOM 聚类的天气分型

聚类分析是数据挖掘中的一类重要技术,是分析数据并从中发现有用信息的一种有效手段。它将数据对象分组成为多个类或簇,使得在同一个簇中的对象之间具有较高相似度,而不同簇中的对象差别很大。聚类算法有多种,本文采用自组织特征映射网络算法 SOM(Self-Organizing Feature Map)聚类分析。该算法是由 Kohonen[10] 提出的,其学习过程可分为两步[11]:① 神经元竞争学习过程。对于每一个输入向量,通过输入向量值 x_i 与权重值 w_j 之间的比较,在神经元之间产生竞争。权重向量与输入模式最相近的神经元被认为对输入模式反映最为强烈,将其标定为获胜的神经元,并称此神经元为输入模式的“像”,相同的输入向量会在输出层产生相同的“像”,即为同一种类型。② 神经元侧反馈过程。应用侧反馈原理在每个获胜神经元附近形成一个“聚类区”。学习的结果总是使聚类区内各神经元的权重向量保持向输入向量逼近的趋势,从而使具有相近特性的输入向量聚集在一起,这个过程被称为自组织。

与传统的模式聚类方法相比,SOM 是一种模式分离方法,这种聚类毋须知道样本的属性,而是通过自组织的方法将输入样本在指定的相似测度下,按样本间的相似程度将其映射到输出层的某个节点中[12-13]。图 1 为 Kohonen 自组织特征映射神经网络结构示意图,由图 1 可知,SOM 分为输入和输出两层,输入层用于接受输入样本,而输出层完成对输入样本的分类。

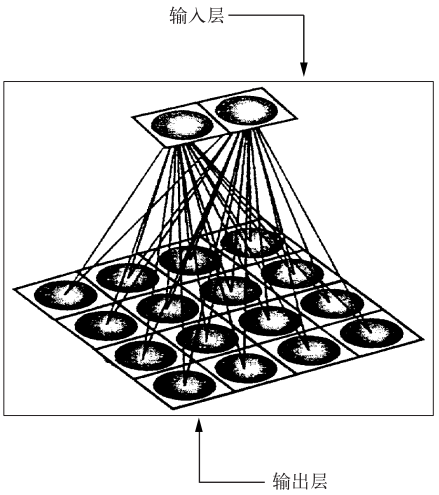

图 1　Kohonen 自组织特征映射神经网络(SOM)[8]

Fig. 1　Kohonen Self-Organizing feature Map structure[8]

设网络的输入模式为 $\boldsymbol{P}_k=(\boldsymbol{P}_1^k,\boldsymbol{P}_2^k,\cdots,\boldsymbol{P}_n^k)$,$k$ 表示第几个输入模式,$k=1,2,\cdots,q$;n 是输入向量的维数。竞争层神经元 j 与输入层神经元之间的连接权矢量为:

$\boldsymbol{W}_j=(\boldsymbol{W}_{j1},\boldsymbol{W}_{j2},\cdots,\boldsymbol{W}_{jn})$，

$i=1,2,\cdots,n;\quad j=1,2,\cdots,M$

输出层分布着网络的 M 个神经元。SOM 的具体算法详见文献[14]。

刘还珠等[15]应用该方法在多模型气象综合预报中针对样本选择问题，提出了先用 SOM 方法对样本聚类，再对多层前向网络进行训练的方法，实现了 SOM 串行性多层前向网络的综合预报系统。

SOM 方法应用在天气分型中，可将物理量（如高度场、风场等）格点场中每个格点值视为输入层的一个节点，然后根据聚类目的而确定分类数目，通过 SOM 算法，输出层的节点存在一个权值与输入的节点最接近，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，最终对该输入产生最大的响应附近形成一个聚类区，由此可将物理量场分为几种不同类型。在这一思路指导下，黄卓等[16]曾经对我国除西北地区外的五大区域高度场和风场进行聚类分型，进行逐日降水等级相似预报试验。

为了预报宁夏各站日最大风，考虑到风与大形势下的高度场和风场关系密切，由于宁夏地势较高，经过反复对比试验，最后确定 700 hPa 高度场及 u，v 风场作为聚类的基本背景场。反映物理量格点场之间的相似，一要考虑两个场之间数值的差异，二要考虑格点场分析出的等值线形状之间的差异，也就是说，既考虑值的相似又要考虑形的相似，根据预报经验，高度场形的相似更为重要。而高度距平场为该场各格点高度值都减去该场的平均高度值，即 $X_{i,k}=x_{i,k}-E_k$，其中 $E_k=\frac{1}{n}\sum_{i=1}^{n}x_{i,k}$，$n$ 为全场格点总数[15]。这样得到的空间距平高度场就可以反映高度场的槽脊位置（负值为低槽，正值为高脊），因而取距平值聚类就是将形相似的物理格点场聚为一类。对同一类型样本的高度场和风场分别进行平均，得到 4 种不同类型的平均高度场和风场，分类结果如图 2 所示。

从图 2 的天气分型结果来看，冬半年影响西北地区的天气形势主要有 4 种：第一类为平直气流型，中高纬度地区气流较平，青藏高原到河套南部有小股弱冷空气活动；第二类为西高东低型，河套以西高度场较高，为反气旋控制，河套东部有一明显的低涡；第三类为强西北气流型，整个中高纬度地区处在乌拉尔山脊前西北气流控制；第四类为蒙古高脊型，95°E 到河套东部被蒙古高压脊控制，从新疆低槽底

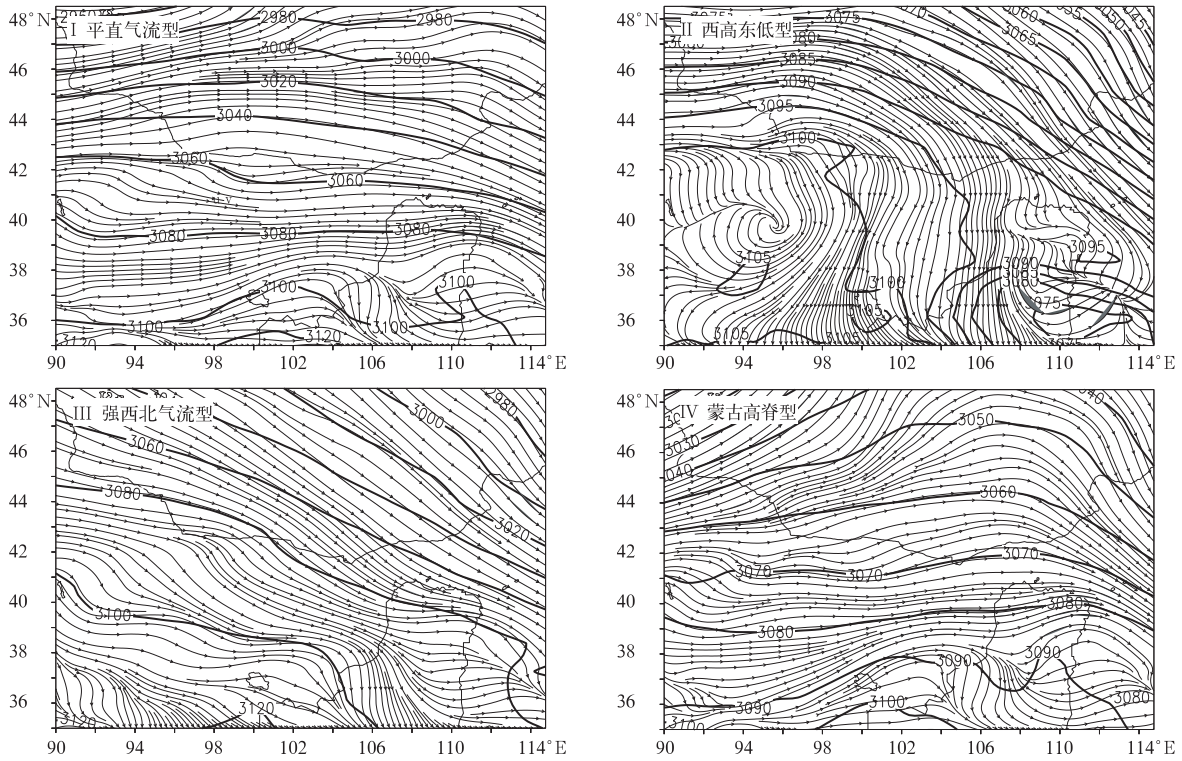

图 2 SOM 聚类分析的 4 种天气型

（粗黑线为 700 hPa 等高线，单位：gpm；细黑带箭头线为 700 hPa u，v 风场合成的流线）

Fig. 2 Four weather patterns of SOM cluster analysis

(black bold lines are contours at 700 hPa, unit: gpm; arrow lines are stream lines at 700 hPa)

部有扩散的冷空气影响西北地区。4 种天气形势对宁夏日最大风速的影响不同,其中在第二、三类天气形势下,宁夏日最大风速≥6 m/s 的正样本比例超过 60%,而在第四类天气形势下不到 50%。可见,冬半年影响宁夏大风的环流形势以西高东低型和强西北气流型为主。

2.2　改进的 KNN 方法

引用文献[9]改进的 KNN 方法,对于不同站点,采用欧式距离作为相似判据,根据正负样本数在总体样本中的比例分类求取 K 值:

$$K^+=\frac{N^-}{N^++N^-}K,\quad K^-=\frac{N^+}{N^++N^-}K \tag{1}$$

式(1)中,N^+ 为训练样本中的正样本数,N^- 为训练样本中的负样本数,K^+ 为正样本的 K 值,K^- 为负样本的 K 值。其中负样本代表没有出现的天气事件,正样本代表出现了天气事件,它们都来自历史实况资料库。

利用交叉验证的方法,取一部分样本作为预报测试集,将剩余部分作为训练样本集,通过不断的交叉更换预报测试样本,直到遍历整个样本集为止。将每次预报的结果汇集并进行检验,得到一组评分结果,再改变 K 值,得到另外一组评分结果,不断重复,直到 K 值试验完毕。

为了使准确率和正样本的概括率都达到相对最好,既要考虑总体样本的准确率,又要考虑正样本的概括率和 TS 评分,这里为了减少漏报率,提高预报准确率和正样本的概括率,用如下 K 值选择公式:

$$\begin{aligned}K_S{}^+&=K^+\left[\mathrm{Min}((1-\text{准确率})+\right.\\&\qquad\left.(1-\text{正样本的概括率}))\right]\\K_S{}^-&=K^-\left[\mathrm{Min}((1-\text{准确率})+\right.\\&\qquad\left.(1-\text{正样本的概括率}))\right]\end{aligned}\tag{2}$$

式(2)中,

$$\text{准确率}=\frac{\text{预报正确的样本}}{\text{所有样本数}}$$

$$\text{正样本的概括率}=\frac{\text{预报正确的正样本数}}{\text{所有正样本数}}\tag{3}$$

通过 L 次交叉验证后,不断调整 K^+,K^- 两个值,反复比较上述预报试验的预报评分结果,选出准确率和正样本的概括率都达到相对最优的组合作为最终结果。实际预报中,将某站点实时预报因子,依据上述确定的 $K_S{}^+$,$K_S{}^-$,从历史样本中选取最邻近域,用以下预报判别方法得到预报结果:

$$\text{预报判别}=\frac{\text{小于欧式距离的样本数}}{\text{小于欧式距离的正样本数}+\text{小于欧式距离的负样本数}}\tag{4}$$

预报判别值是通过历史资料的试预报结果比较判断给出。制作预报时,根据式(4)计算出的预报判别值大于历史的预报判别值时,则认为有天气事件发生,反之无。

2.3　实现改进后的 KNN 方法步骤

改进后的 KNN 方法具体计算步骤可归纳为:① 对经过质量控制后的历史样本的 700 hPa 高度距平场(每个格点值逐一求距平值)、u 及 v 风场进行归一化处理;运用 SOM 方法求得 4 种类型的样本群;对每种样本群的 700 hPa 高度场、u 及 v 风场分别合成,得到 4 种不同类的典型天气型;② 预报因子归一化,对聚类分析的每一种天气型的样本群所对应的因子与某站点的预报对象之间求相关系数,利用逐步回归对因子进行排序筛选,最终选取 10～20 个因子作为该站点 KNN 的预报因子集;③ 循环 K 值(比如从 2～50),利用式(1)分别得到 K^+,K^-;④ 在每一次 K 值试验中,将模型样本分为与预报测试样本和训练集样本,通过不断的交叉提取,计算每个预报测试样本中正负样本的欧式距离,做出每一个子样本的预报,将交叉所得的预报结果汇集起来,而后得出该 K^+,K^- 值所对应的预报评分;⑤ 最后通过式(2)和式(3)选出预报评分最优 K^+,K^- 值;⑥ 对由历史资料确定的 K^+,K^- 值,寻求最邻近域,利用式(4)确定每一种天气型对应预报判别值。

实际预报时,先判别该日属于何类天气型,即用该日 T213 数值预报的 24 h,48 h 预报的700 hPa高度距平场、u 及 v 风场与已分型后得到的平均 700 hPa高度距平场、u 及 v 风场分别进行比较(求对应的各格点的欧式距离),最小欧式距离所对应的天气型应为当日 24 h,48 h 预报的天气类型(24 h,48 h 也可能类型不相同),再用不同时效这种天气型已确定的 K^+,K^- 值,计算实时预报因子与历史样本对应预报因子的最邻近域,并以预报判别值为标准,最后给出预报结论。逐站按此步骤进行,即得到宁夏全区各站点未来 24 h,48 h 有无日最大风速≥6 m/s 的预报结论。

3　预报试验评估

由于文献[9]通过降水预报已验证:分类交叉验证求取 K 值较传统 KNN 方法的预报结果有明显提高。因此本文仅就聚类天气分型对 KNN 预报的影响,分别利用不分型和分为 4 种天气型的 KNN

方法，以日最大风速≥6 m/s 作为预报对象，以2003—2006 年冬半年的 1—5 月和 10—12 月经加工(如第 1 章所述)的 T213 数值预报产品作为建模样本(如某日最大风速≥6 m/s，则该日为正样本)，建立了宁夏 24 个测站冬半年 24 h 和 48 h 日最大风速≥6 m/s的预报模型，并对 2007 年 1—5 月进行了预报试验，各站评估结果如图 3，图 4 所示。为了对该模型的预报效果有一个总体评价，同时给出表 1 的

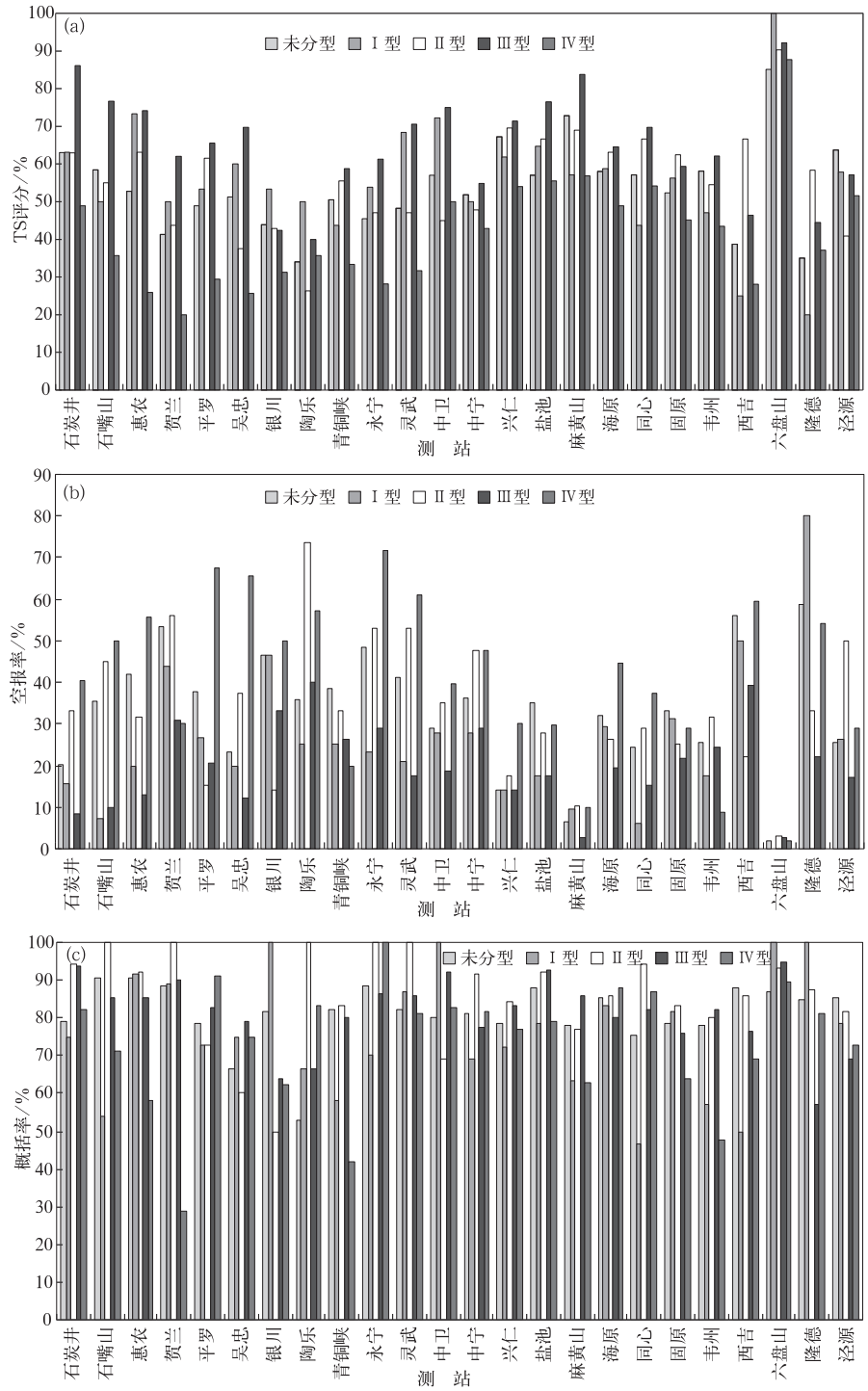

图 3 2007 年 1—5 月宁夏各站 24 h 日最大风速≥6 m/s 预报的 TS 评分(a)、空报率(b)和概括率(c)

Fig. 3 TS(a), absent forecast quotiety (b), general probability (c) of 24-hour forecast for weather stations with daily maximum velocity≥6 m/s from Jan to May in 2007 of Ningxia

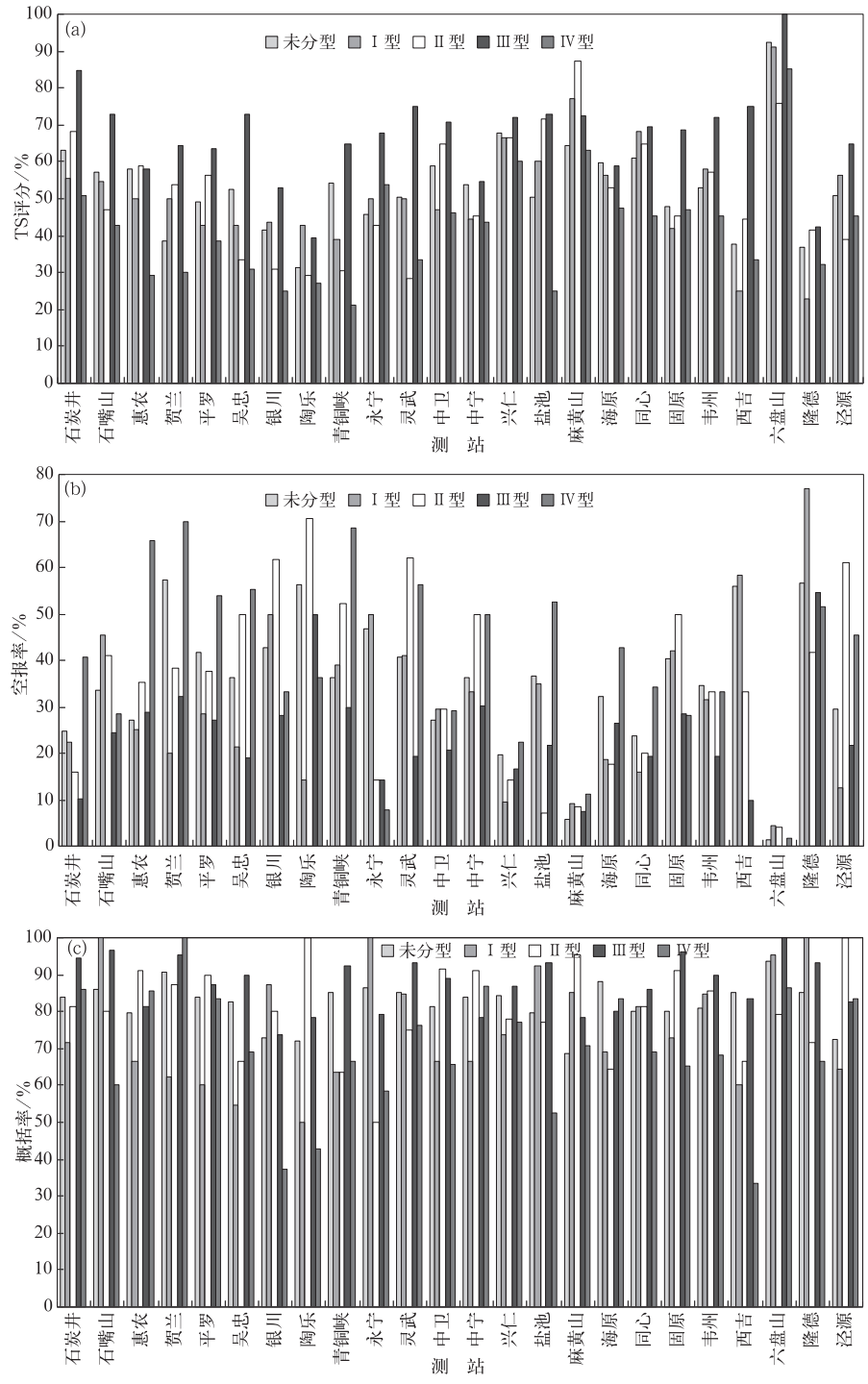

图 4　2007 年 1—5 月宁夏各站 48 h 日最大风速≥6 m/s 预报的 TS 评分(a)、空报率(b)和概括率(c)

Fig. 4　TS(a), absent forecast quotiety (b), general probability (c) of 48-hour forecast for weather stations with daily maximum velocity≥6 m/s from Jan to May in 2007 of Ningxia

评估结果。

从图 3,图 4 及表 1,可以看出:

① 从总体预报 TS 评分看,除第四类天气型,24 h 其他 3 种天气型的预报效果总体高于未分型,48 h 第三类天气型 TS 评分超过未分型,而且无论是 24 h 或 48 h,第三类天气型都远远高于未分型和

表 1 2007 年 1—5 月 24 h 和 48 h 宁夏日最大风速≥6 m/s 预报评估结果

Table 1 Average skills of 24-hour and 48-hour forecast of daily maximum velocity≥6 m/s from Jan to May in 2007 of Ningxia

预报评价	预报时效	天气分型					未分型
		Ⅰ型	Ⅱ型	Ⅲ型	Ⅳ型	平均	
TS 评分/%	24 h	55.6	56.0	65.2	41.7	54.6	53.8
	48 h	51.5	51.6	67.0	41.8	53.0	53.2
空报率/%	24 h	25.5	33.6	20.2	41.3	30.1	33.4
	48 h	30.6	35.4	23.3	38.3	31.9	35.2
准确率/%	24 h	69.3	73.9	71.7	67.5	70.6	68.8
	48 h	64.0	66.2	72.0	69.1	67.8	67.4
正样本概括率/%	24 h	75.8	85.8	81.2	73.2	79.0	81.2
	48 h	75.5	80.7	87.3	69.8	78.3	82.1

其他 3 种天气型的预报评分;从单站预报效果看,除了 24 h 的泾源和 48 h 的海原,其他测站至少有一类天气型的 TS 评分高于未分型,其中 24 h 的六盘山 4 种天气型的 TS 评分均高于未分型,另外除了 24 h 的银川和泾源及 48 h 的海原,其他测站第三类天气型的预报评分都高于未分型。这可能与这些站历史上出现日最大风速≥6 m/s 的气候概率有关。

② 空报率无论 24 h 或 48 h,加入 SOM 聚类天气分型后的 KNN 总体低于未分型的 KNN,其中第一类天气型低于未分型 5%～8%,尤其是第三类天气型空报率低于未分型达 12%～13%;单站空报率,第一类天气型 24 h 除了银川、麻黄山、隆德和泾源,其他测站都低于未分型,第三类天气型除了 24 h 的陶乐和 48 h 的麻黄山,其他测站都低于未分型,其中六盘山站第一类 24 h 和第三类 48 h 的空报率都为 0。相比空报率,漏报率总体上分型高于未分型。从实际预报效果看,虽然天气分型后漏报次数相对增加导致正样本概括率下降,但减少了空报次数,从而提高了预报准确次数,如第三类天气型总体上 24 h 漏报率较未分型高 1.8%,但空报率却下降了 13.2%,结果是预报准确率较未分型提高了 1.4%。可见,天气分型后,总体上降低了空报次数,有利于预报效果的提高。

③ 正样本的概括率总体上天气分型后的 KNN 低于未分型的 KNN,只有 24 h 的第二类天气型和 48 h 的第三类天气型高于未分型,24 h 第三类天气型与未分型的概括率基本持平;单站的概括率 24 h 第一类有 17 个测站高于未分型,48 h 第三类有 23 个测站高于未分型,其他类天气型的概括率有超过 60%的站低于未分型。

④ 总体上,天气分型后预报准确率高于未分型的,24 h 表现得尤为突出。

分析结果可见,加入 SOM 聚类天气分型后的 KNN 方法虽然降低了正样本的概括率,但克服了预报空报偏多的现象,总体上提高了预报效果,尤其是第三类天气型的预报效果较未分型提高更为显著,说明改进后的 KNN 预报模型挑选的预报因子总体上能够反映宁夏大风的预报信息。

4 历史样本和预报因子分析

本文使用历史样本共 921 个,各站正样本(日最大风速≥6 m/s)数约占总样本的 56%,六盘山、麻黄山、石炭井、通信、兴仁正样本比例超过 70%,其中六盘山达 97%;青铜峡、贺兰、平罗、西吉、银川达 30%～40%。对比这些站的 TS 评分,正样本大的测站 TS 评分也高,反之亦然。天气分型后也有同样情况。可见,历史正样本的多少对预报模型的建立及预报效果的好坏有至关重要的作用。从历史样本中寻求相似,这种"相似"也仅仅是相对而言,如果历史正样本多,这种相似的可靠性更高,因而预报的效果会更好;反之,历史正样本少,相似可靠性低,因而预报效果差。

分析两种预报模型挑选的预报因子,绝大部分测站预报因子的相关系数不超过 0.45,逐步回归使用的 F 检验值在 0.25 左右,其中未分型选择的预报因子相关系数基本在 0.15～0.3 之间,分型后的预报因子相关系数在 0.25～0.45 之间;所选的预报因子基本集中在中低层风速、纬向风、经向风、风切变及反映温度、气压梯度等方面,未分型的预报因子

相对分散,天气分型后的预报因子相对集中,主要反映与本天气型影响系统物理意义明确的因子。这说明大风的形成既与大气环流背景有关,又与中小尺度天气系统密切相联,相对海上大风[17-18],陆地受地形、地貌的影响很大,所以天气分型所选的预报因子应能更好地反映这些影响因素,建立的预报模型更符合当地实际情况。

5　结果与讨论

上述分析表明,同时利用 SOM 聚类天气分型和交叉验证的 KNN 方法对宁夏冬半年日最大风速≥6 m/s 的预报效果更好。在天气分型的前提下,针对不同天气环流背景选择的预报因子更有代表性,物理意义更明确,也更有助于邻近域的选择和预报模型的建立,从而提高预报准确率。特别是对前 3 种类型,均好于未分型的。而对于西低东高有利于降水的第四种类型,预报时,须用其他方法给予进一步关注。值得指出的是,本文所用历史数据库(范例库)容量只有 4 年 921 个样本,分型后,每种天气型样本数大大减少,大约是原未分型的四分之一,这在一定程度上影响了分型的预报效果。对那些正样本多的测站预报效果显著,如六盘山站;而正样本相对少的测站预报准确率相对较低,第四种类型正样本数更少,预报效果较差,大概也是这一原因。但大部分测站的预报准确率已基本达到业务预报要求,如果拥有能反映系统状态变化范围的较长历史数据,预报准确率将会逐步提高,因此该预报模型具有业务应用价值。

总体来说,基于 SOM 聚类天气分型和交叉验证的 KNN 近邻非参数估计仿真模型是一种实用的概率天气预报制作方法,可以根据天气学原理和天气预报经验进行天气分析和预报技术的研制。由于该方法对预报因子和预报量对象均不需加任何限制,不需有关于模拟过程的先验知识,仅用足够多的历史数据来建立系统输入和输出之间的内在关系,因此利用该技术可实现多种要素或天气现象的同时预报。同时该方法原理清晰、计算方便,随着历史样本数的不断累积,该方法将取得更好的应用前景。

致　谢:感谢中国气象局培训中心曹晓钟老师对本文的指导和修正!

参 考 文 献

[1] 陈豫英,陈晓光,马金仁,等. 风的精细化 MOS 预报方法研究. 气象科学,2006,26(2): 210-216.

[2] 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. 应用气象学报,2004,1(2):181-191.

[3] 杨忠恩,陈淑琴,黄辉. 舟山群岛冬半年灾害性大风的成因与预报. 应用气象学报,2007,18(2):80-85.

[4] 林良勋,程正泉,张兵,等. 完全预报方法在广东冬半年海面强风业务预报中的应用. 应用气象学报,2004,15(4):485-490.

[5] 胡波,杜惠良. 浙江省沿海海面日极大风预报. 海洋预报,2006,23(增刊):64-67.

[6] Cover T M, Hart P E. Nearest neighbor pattern classification. *IEEE Trans on Inf Theory*,1967,IT-13:21-27.

[7] 翟宇梅,赵瑞星. 概率天气预报的 K 近邻非参数估计仿真模型. 系统仿真学报,2005,17(4):786-788.

[8] 邵明轩,刘还珠,窦以文. 用非参数估计技术预报风的研究. 应用气象学报,2006,17(增刊):125-129.

[9] 曾晓青,邵明轩,王式功,等. 基于交叉验证技术的 KNN 方法在降水预报中的试验. 应用气象学报,2008,19(4):471-478.

[10] Kohonen T. Self-organizing Maps. Berlin: Springer-Verlag,1998,21:1-6.

[11] 许文杰,刘希玉. 基于无监督神经网络聚类算法的研究. 信息技术和信息化,2006,(6):85-88.

[12] 孙世霞,杨建池,邱晓刚,等. 基于 BP 网络的 LSCS 仿真可信性评估方法. 系统仿真学报,2006,18(7):2037-2041.

[13] 王青,祝世虎,董朝阳. 自学习智能决策支持系统. 系统仿真学报,2006,18(4):924-926.

[14] 夏文文,王士同. 基于 Voronoi 距离的鲁棒的双自组织特征映射网络. 计算机应用,2007,27(5):1109-1112.

[15] 刘还珠,郝为,林孔元,等. 基于智能计算的多模型气象综合预报//刘还珠,汤桂生. 暴雨落区预报实用方法. 北京:气象出版社,2000:30-37.

[16] 黄卓,杨洪敏,郝为,等. 基于智能聚类的综合相似预报//刘还珠,汤桂生. 暴雨落区预报实用方法. 北京:气象出版社,2000:53-59.

[17] 廖木星. 海面风场预报的技术研究报告. 青岛远洋船员学院学报,2003, 24(2):6-10.

[18] 颜梅,范宝东,满柯,等. 黄渤海大风的客观相似预报. 气象科技,2004,32(6):467-470.

Application of KNN to Wind Forecast Based on Clustering Synoptic Patterns

Chen Yuying[1)] Liu Huanzhu[2)] Chen Nan[1)] Zeng Xiaoqing[3)]
Ma Jinren[1)] Liu Qianqian[1)] Ma Shaiyan[1)]

[1)] (*Key Laboratory for Meteorological Disaster Prevention and Reduction of Ningxia*, *Yinchuan* 750002)

[2)] (*National Meteorological Center*, *Beijing* 100081)

[3)] (*College of Atmospheric Sciences*, *Lanzhou University*, *Lanzhou* 730000)

Abstract

Based on the model identification and an analogue forecasting, a new approach based on Self-Organizing feature Map(SOM)and cross validation is constructed, which is called *K*-nearest neighbor nonparametric estimation bootstrap model(KNN). 500 hPa geopotential height and 700 hPa *u*, *v* wind field over Northwest China are analyzed by the model clusterings at first, then the optimal *K* combination is sought using cross validation aiming at past samples under different weather patterns. Forecasting identification value of each synoptic pattern is determined by *K*-data, according to historical record. When forecasting in real time, what kind of synoptic pattern is to be known first, then *K*-data of different time is used to compute the nearest neighbor of real forecasting predictor to historical material predictor. Finally forecasting conclusion is obtained by using the standard of forecasting identification value. In order to validate the effect on cluster synoptic pattern to KNN, T213 NWP products from 2003 to 2006 in winter half year and the data of daily maximum velocity in Ningxia are used to construct prediction models of daily maximum velocity≥6 m/s pattern in Ningxia under synoptic and non-synoptic patterns at one time, data from Jan to May in 2007 is used for forecast experiments. The forecast evaluation results show that although the probability of original sample is reduced when adding the Self-Organizing feature Map of KNN, more false alarms in forecasting are avoided, so that the effect of forecasting is improved in general, especially the forecasting effects of some synoptic patterns compared with those that aren't patterned. The result is that the forecasting information of Ningxia high wind can be reflected by improved KNN. What's worth pointing out is that, the number of synoptic patterns is reduced when patterned, so the forecasting will be effected to some extent. It has a good effect for meteorological observing station which has more original samples, but it is not good for the ones that have less original samples. Therefore if there are more historical data which can reflect the wide range of system changing, the forecast accuracy will be improved significantly and it has a great value for operational usage. Classification analogue prediction thinking can be expanded by these results.

Key words: Self-Organizing feature Map; clustering synoptic patterns; cross validation; *K*-nearest neighbor; daily maximum velocity forecast