k-nn 方法在热带气旋路径预报中的应用

陈　见　杨宇红　张诚忠　郑宏翔

（广西气象台，南宁　530022）

提　要

引入 k-nn 方法，结合经过预报实践检验、应用证明效果较好的预报因子，制作热带气旋路径预报系统，运行效果良好，可以应用到日常预报业务中。

关键词： 热带气旋　路径　k-nn 方法

引　言

在目前的天气预报业务中，热带气旋路径的预报仍多采用相似预报法。其基本原理是：把当前的大气状况和过程与历史上曾经出现过的大气状况和过程相比较，从中找出一个或几个与当前大气状况和过程最为相似的个例，假定相似原因将产生相似的结果，那么就可以按照历史上曾经出现过的最相似的个例来预测今后将要出现的天气过程和天气现象。k-nn 方法正是基于这一原理而适用于气象预报的，这是一种非线性预报方法，因为外推是按照大气本身的演变规律进行的，它包含了大气本身所有的非线性变化。

1　方　法

k-nn（k：k 个样本，n：nearest 最近的，n：neighbour 邻居，k：近邻权函数）方法是一种非参数回归方法，其优点是可以较好地避开由于模型假设与实际情况出现重大偏差而带来的错误信息，具有较强的稳定性，因此可以应用到有一定统计规律性的热带气旋路径预报中。当热带气旋进入警戒区海面某点时，可以利用历史上曾经通过该点的几个或多个热带气旋的信息来判断其移向、移速，从而达到对其未来位置作出预报的目的。

①设 X 是自变量，Y 是因变量，X、Y 间并无确定的关系（都是随机的），但（X，Y）有一定的概率分布，因分布的数学形式是未知的，则问题的模型是非参数性的，为了刻画 Y 对 X 的依赖关系，就要考虑 $X = x$（x 为指定点）时，Y 取 A 内值的条件分布

$$P^y(A \mid_x) = P(Y \in A \mid_{X=x})$$

A 为 Y 值域中的任意子集。

知道了这个分布，从统计意义上讲，就知道了 Y 对 X 依赖的全貌，就可以计算回归函数，即求出 $X = x$ 时 Y 的条件期望值 $E(Y \mid_x)$，对预报问题而言就是预报值。

②求 $P^y(A \mid_x)$ 的原始方法是以频率估计概率，但这样就没有充分利用整个样本的信息，例如某样本（X_i，Y_i），其中 X_i 很接近 x，但不等于 x，则这个样本应包含关于 $P^y(A \mid_x)$ 的信息，但如果以频率估计概率就未加利用。权函数方法就是着眼于这个事实，以便在估计 $P^y(A \mid_x)$ 时，整个样本都用上。每个样本值（X_i，Y_i）的重要程度，要看 X_i 与指定点 x 的“距离”而定，距离越小，重要性越大。样本 X_i 与 x 标准化后的欧氏距离

$$d_n = \left(\sum_{j=1}^{m}(x^{(j)} - x_i^{(j)})^2\right)^{1/2} \quad (1)$$

（m 为所选取的样本数）

③有了距离，引进权函数表示样本（X_i，Y_i）的重要程度。用 $W_{ni} = W_{ni}(x_1, x_2, x_3, \cdots\cdots x_n)$ 表示权函数，$i = 1,2,3,\cdots,n$，要求权函数 $W_{ni} \geqslant 0$，则要求权函数是非负的，正则的，称为概率权函数。它应该把主要的权给与那些与指定点 X 靠近的点上，即“距离”x 近邻的权重大，反之则小，这就是近邻权函数的由来。

$$CN_i = \begin{cases}(k - I + 1)/b_k & 1 \leqslant I \leqslant k \\ 0 & k + 1 \leqslant I \leqslant N\end{cases} \quad (2)$$

此处 $b_k = k(k+1)/2 \quad K = 3\sqrt{n}$

④把距离 d_n 按从小到大排列，排出全距，然后定出 K 值（根据经验 k 值取 $3\sqrt{n}$，有了 k 值，算出 CN_i，把 CN_i 值从大到小排列，与 d_n 相对应，根据同一组内样本距离基本相同，不同组之间则不同的原则，反 d_n 和相对应的 CN_i 分成 g 组，每组个数设为 $m_1, m_2, \cdots, m_g$，各组 CN_i 的平均值为 $u_1, u_2, \cdots, u_g$，可知

$$\sum m_i u_i = 1$$

当 $1 \leqslant i \leqslant k$ 时，$W_{ni}(x) = U_x$

当 $i \geqslant k + 1$ 时，$W_{ni}(x) = 0$

⑤选定权数后，计算回归函数，即 $E(Y|_x)$ 的估计值

$$E(Y|_x) = \sum W_{ni}(x) Y_i$$

这样就得到了在指定点 $X = x$ 的条件下，Y 的期望值，也就是预报值。

2 系统研制

2.1 预报因子的选定

热带气旋的移动，主要受其自身内力和环境力的综合影响，下垫面的影响只在登陆前后才显著起来。因此，在选定预报因子时，主要考虑 500hPa 副热带高压的引导作用。

根据经验，我们挑选了日常预报业务中用得较多的预报因子进行相关普查，同时根据 k－nn 方法的运算特点，从中选出两个最常用的预报因子：

①气旋中心到副高脊线的距离 d；②气旋中心与副高西脊点连线到气旋中心所在纬线的夹角 θ。这两个预报因子无论在快速、中速或者移动缓慢的热带气旋路径预报中，都是必须考虑的，如图 1 所示。

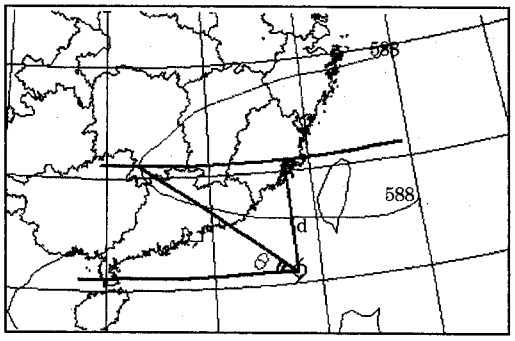

图 1 预报因子示意图

2.2 建立样本库

根据预报因子选定的思路，以及便于资料查询等角度的考虑，我们共收集了 1961～1998 年 8 月份进入 25°N 以南警戒区海面的热带气旋资料，包括气旋名称、时间、经度、纬度、预报因子①，预报因子②等，并按一定的次序排列，生成一文本文件（表 1）。

2.3 运行模块

系统采用基于 32 位 Windows 平台的 Delphi5 系统进行设计。考虑到一些因子的不确定性，系统以人机对话为主；界面简洁明

了,操作方便简单,计算出的预报结果直接在界面显示,并以文件形式保存。系统的流程图如图2。

表1

气旋名	时间	纬度	经度	d	θ
7005	700804	18.0	133.7	12	35
7005	700805	16.9	129.7	14	29
…	…	…	…	…	…

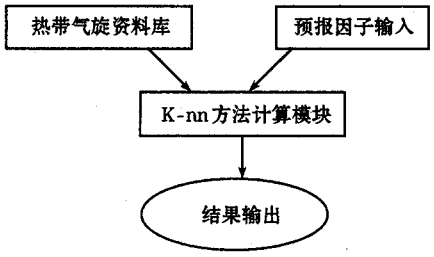

图2 流程示意图

表2 2000年0010号热带风暴"碧利斯"预报结果

时间	实况		24小时预报		48小时预报	
日.时	纬度/°N	经度/°E	纬度/°N	经度/°E	纬度/°N	经度/°E
20.20	17.6	130.8	18.2	128.2	18.4	125.5
21.08	18.8	128.2	19.7	123.7	20.2	119.3
21.20	19.7	126.1	20.6	121.6	21.2	117.3
22.08	20.8	123.9	23.1	118.3	21.0	117.4
22.20	22.5	122.0	22.6	117.3	21.8	113.9
23.08	24.3	118.9	25.0	114.0	24.4	112.6

3 业务运行技术指标完成情况

系统于2000年8月份研制完成并开始投入业务使用,表2和表3是0010号热带风暴"碧利斯"的预报运行情况。

排名前3个相似样本为:780821 780822 780823

表3 预报点与实况平均误差

时间	距离误差/km		移向误差/度	
	24小时	48小时	24小时	48小时
08时	175	412	15.0	25.0
20时	242	476	28.3	30.0

4 讨论

上述运算结果表明,目前k-nn方法在热带气旋路径预报中有一定的效果,但预报结果仍不够理想,主要问题是预报距离误差过大。经分析,主要原因有:

①系统刚刚投入运行,可能还存在着设计上的问题;②模式中预报因子的选定,是假设副高是稳定的,没有考虑到副高的变化;③预报因子的选取过多地考虑针对广西本地,而0010号气旋"碧利斯"没有影响到广西;④样本数量偏少,仍须充实。这些问题有待在今后的工作中进一步完善,使k-nn方法切实能够在预报业务中发挥作用。

Application of k-nn Method to Forecasting of Tropical Cyclone Track

Chen Jian Yang Yuhong Zhang Chengzhong Zheng Hongxiang
(Guangxi Meteorological Observatory, Nanning 530022)

Abstract

Combined with prediction factors which are turned out to be preferable by practice forecasting, a forecasting system of tropical cyclone track is developed by the k-nn method. The results show that it is satisfying, and can apply to routine operational forecasting as well.

Key Word: tropical cyclone track K-nn method