第14卷 第3期 大 气 科 学 Vol.14, No.3
1990年9月 SCIENTIA ATMOSPHERICA SINICA Sept., 1990

辗转坐标筛选*

贡九鼎 李慧武
(浙江省气象科学研究所)

提 要

在只有一个主要因子的前提下，给出一个不涉及到因子权重而输出唯一相似样品的算法．算法里体现了因子间的相互作用．

关键词： 稠密集；有限集；辗转坐标筛选．

一、简单的数学框架

1. 最近距离在有限集里的多值性

预报中要求在已知样品曲线 f_1，f_2，…，f_m 中找出一条样品曲线 f_e，使 f_e 与待报样品 f_o 最接近．该问题即是在函数空间里求点列 $\{f_i\}$ 以 f_o 为极限的收敛点 f_e．

距离空间 R^n 的点列 $\{f_i\}=\{(x_{i1}, x_{i2}, \cdots, x_{in})\}$ 收敛于点 $f_o=(x_{o1}, x_{o2}, \cdots, x_{on})$ 的充要条件是 f_i 的每个坐标都收敛于 f_o 的每个相应坐标[1]．即在欧氏距离 $d(f_i, f_o)=\left(\sum_{k=1}^{n}|x_{ik}-x_{ok}|^2\right)^{1/2}$ 中，当 $f_i\to f_o$ 时有 $|x_{ik}-x_{ok}|<\varepsilon$ 对一切 $k=1, 2, \cdots, n$ 都成立．从而也有 $d(f_i, f_o)<\varepsilon$，ε 为无穷小量．

在有限集里，除非 f_o 本身或者恰有 $f_e=f_o$，对任何 $\{f_i\}$ 中的元素 f_i，对任何坐标 k，不等式 $|x_{ik}-x_{ok}|<\varepsilon_{kl}$ 中的 ε_{kl} 是一个有限正数．ε_{kl} 是等差递降序列 $\{\varepsilon_{kl}\}$ 的一般项（详见本文第三节第三段）．距离 $d(f_i, f_o)\leqslant\left(\sum_{k=1}^{n}\varepsilon_{kl}^2\right)^{1/2}=L$ 也是一个有限正数．因为同样的 L 可由无穷多组不尽相同的 ε_{kl} 生成．即以点 f_o 为圆心，以 L 为半径的开球内的所有点都满足 $d(f_i, f_o)\leqslant L$．这是一个多值问题，也是产生“值相似”而“形不相似”的原因[2]．

样品曲线 $f_i(x_{i1}, x_{i2}, \cdots, x_{in})$ 可表为阶梯函数．x_{ik} 表示第 i 个样品的第 k 个坐标值．

$$f_j=\sum_{k=1}^{n}C_k\,\chi_{(x_k,\,x_{k+1})}(x),$$

当 $x_k\in[x_k, x_{k+1})$ 时，$C_k=x_{ik}$．χ 为集合的特征函数．

图 1a 为稠密集中 f_o 与 f_e 的比较，图中只列出三个坐标．图 1b 是 f_e-f_o 的图像，表现了一致收敛的情形．图 2 是有限集的情形．我们就是要在 $\varepsilon_1\fallingdotseq\varepsilon_2\fallingdotseq\cdots$ 的有限集里求

1987年5月3日收到，12月10日收到再改稿．
* 本文为国家气象局台风基金资助课题．

唯一的 f_e，或者说在该集里建立全序关系. 本文中的样品和点、因子和坐标是指同一件事.

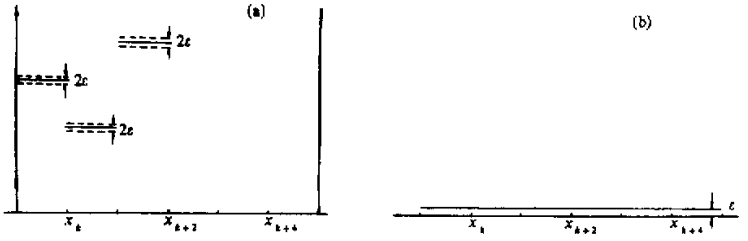

图1　稠密集中 f_o 与 f_e 的比较 (a) 及 $f_e - f_o$ 的图像 (b)
实线为 f_o，虚线为 f_e.

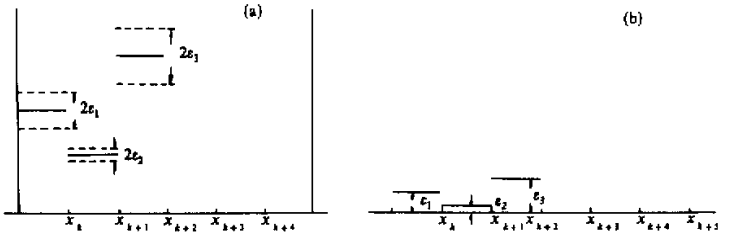

图 2　在有限集的情形
实线为 f_o，虚线为 f_e

2. 引进筛选算子 DE

点 f_i 与 f_o 在坐标 x_k 上的距离是

$$d_{io}^{k}(f_i, f_o) = |x_{ik} - x_{ok}|,$$

上述线性运算在一维空间里不产生多值情形. 如果 d_{io}^{k} 为 f_i 的像，那么原像（反函数）存在. 今以 $f_i = (d_{io}^{k})^{-1}$ 记之. 在一维空间 x_k 里，集合 $\{d_{10}^{k}, d_{20}^{k}, \cdots, d_{mo}^{k}\}$ 的上确界 $\sup_k\{d\}$ 的原像 $(\sup_k\{d\})^{-1}$ 代表了在坐标 x_k 里与 f_o 有最大距离的点. 下确界 $\inf_k\{d\}$ 的原像 $(\inf_k\{d\})^{-1}$ 代表了与 f_o 有最小距离的点. 如果集合稠密，n 个坐标上的 n 个原像：$(\inf_1\{d\})^{-1}$，$(\inf_2\{d\})^{-1}$，…，$(\inf_n\{d\})^{-1}$ 是同一个点. 即有 $(\inf_1\{d\})^{-1} = (\inf_2\{d\})^{-1} = \cdots = (\inf_n\{d\})^{-1}$. 该点就是 f_o 的收敛点 f_e.

有限集里情形远为复杂. 在某个一维空间里与 f_o 有最近距离的点，在另一个一维空间里不一定是与 f_o 有最近距离的点. 即当坐标 $i \neq j$ 时，一般有 $(\inf_i\{d\})^{-1} \neq (\inf_j\{d\})^{-1}$.

设有 m 个点，n 个坐标，可断言点 $(\sup_k\{d\})^{-1}$ 在 n 维空间里一定不是与 f_o 有最近距离的点. 不妨考察极端情形，即点 $(\sup_k\{d\})^{-1}$ 在其余 $n-1$ 个坐标里和 f_o 距离最近，那么该点就是畸点. 图 3 是 $(\sup_k\{d\})^{-1} - f_o$ 的图像. 畸点理应被筛除. 故客观上存在一个筛选算子 DE. 若将 DE 对点列 $\{f_i\}$ 在所有坐标上连续作用一次（简称 DE 作用一个周期），至少可将 n 个一阶畸点筛除.

DE 在坐标 k 上对 $\{f_i\}$ 作用一次，相当于将点列 $\{f_i\}$ 在坐标 k 上按对 f_o 的一维距离 $|x_{ik} - x_{ok}|$ 由大到小排列趋向于 f_o，并筛去序列首项.

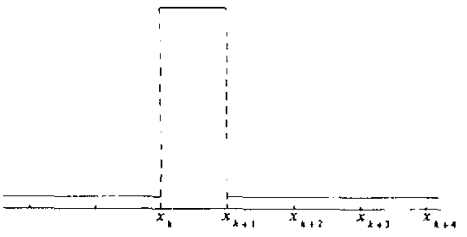

图3 $(\sup_k\{d\})^{-1}-f_o$ 的图像

算子 DE 作用的第一个周期可表为

$$DE^{(1)}(\{f_i\})=\{f_1^1, f_2^1, \cdots, f_{m-1}^1\}=\{f_i^1\},$$

$$DE^{(2)}(\{f_i^1\})=\{f_1^2, f_2^2, \cdots, f_{m-2}^2\}=\{f_i^2\},$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$DE^{(n)}(\{f_i^{n-1}\})=\{f_1^n, f_2^n, \cdots, f_{m-n}^n\}=\{f_i^n\}.$$

$\{f_i\}$是未经 DE 作用的原序列. $\{f_i^j\}$为子列$\{f_i^{j-1}\}$经DE 作用后的子列. 每个子列的项数比上一个子列少 1.

DE 作用的第二个周期为

$$DE^{(n+1)}(\{f_i^n\})=\{f_1^{n+1}, f_2^{n+1}, \cdots, f_{m-n-1}^{n+1}\}=\{f_i^{n+1}\},$$

$$DE^{(n+2)}(\{f_i^{n+1}\})=\{f_1^{n+2}, f_2^{n+2}, \cdots, f_{m-n-2}^{n+2}\}=\{f_i^{n+2}\},$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$DE^{(2n)}(\{f_i^{2n-1}\})=\{f_1^{2n}, f_2^{2n}, \cdots, f_{m-2n}^{2n}\}=\{f_i^{2n}\}.$$

$\{f_i\}$ 经 DE 连续作用第二个周期后，至少将$\{f_i\}$中的 n 个二阶畸点筛除. 自然还可有第三个周期、第四个周期…，这就叫辗转坐标筛选.

3. 物理上的最近点

稠密集里将 DE 一直作用下去就能输出收敛点 f_o. 但最近距离在有限集里的多值性将导致下面"荒谬"情形. 在 DE 作用的后几个周期里，不仅 $(\sup_i\{d\})^{-1}\neq(\sup_j\{d\})^{-1}$ 的现象早就产生，而且也会产生 $(\sup_i\{d\})^{-1}=(\inf_j\{d\})^{-1}$ 的现象. 这就是说，在第 i 个坐标上与 f_o 有最大距离的点，在第 j 个坐标上却是和 f_o 有最近距离的点. 数学上的坐标毕竟不同于物理上的因子. 如果坐标 j 恰是最大权重的因子，则点 $(\sup_i\{d\})^{-1}=(\inf_i\{d\})^{-1}$ 有可能是 f_o 在物理上的最近点，那么算子 DE 在先行坐标 i 上就将该点筛除，后续坐标 j 上该点已不复存在，这显然有悖实际情形.

以下两个条件导致了物理上的唯一最近点的输出. 第一，用实验确定 DE 的最大作用周期 t 以保证物理上的最近点不被筛除. 第二，确定物理上最重要的一个因子，即输出坐标. 不失一般性，可设第一个坐标 x_1 是主要因子. 令 DE 作用 t 周期后，在 x_1 上

输出子序列，该子序列的末项 $f_{m-1}^{[nt+1]}$ 即为有限集里的最近点（在物理上）f_e.

二、算法特点

1. 再谈值相似和形相似 [2]

稠密集里几何上的最近点（收敛点）就是物理上的最近点，因而值相似和形相似是完全统一的. 由于多值情形，这种统一性在有限集里不复存在. 气象上的样品曲线不同于几何图形，离开了值相似的形相似是没有意义的，如文献 [2] 图 1 中的样品曲线 1 与 3. 离开了形相似的值相似是不确定的（多值性），如文献 [2] 图 2 中的曲线 1 与 4，1 与 2. 但是值相似应该是本质的.

算子 DE 作用在某个坐标上，就是求 f_0 在该坐标上的值相似. 算子 DE 作用一个周期，就是在值相似的基础上求 f_0 的形相似. 最大作用周期 t 越长，则筛去畸点的阶数就越高，形相似的精度也越高. 这是指立足于值相似的形相似. 由于第一节第三段里的情形，求形相似只能到此为止，其精度是筛去了 $m-nt$ 个畸点. 下一步是在输出坐标上求值相似. 该值相似又是在 nt 阶形相似的基础上求得的.

2. 权重问题

当预报量是诸如降水分布这样复杂的对象时，就很难用一个数来表示，从而给确定预报因子的权重带来困难. 即使权重可定，则该权重一经确定后就不能改变. 这对新样品来说，无疑是一个大弊病. 在本算法中，每个因子的权重在该因子所在的坐标上为 1，同时其余因子的权重皆为零. 在 DE 作用的一个周期中，每个因子都有一次成为最大的权重因子. 在欲输出结果的时候，空间元素已由 m 个减少到 $m-nt$ 个. 即在经过 $m-nt$ 次筛选变换后的空间里，输出坐标的因子取得了权重 1，其余 $n-1$ 个因子的权重一律降为零. 从表面上看，本算法避开了确定因子权重这么一个棘手问题. 实际上在 DE 的每次作用里都把权重处理成 1 或零；但在 DE 作用的全过程中，权重又是在改变的. 这有点象求积分时，在每个小区间上将被积函数视为常数的处理方法.

由此可知，要求有一个且只有一个主要因子作输出坐标是本算法的前提条件.

3. 非线性性

由第一节第二段 DE 的作用形式可见，$k+1$ 坐标的输入就是 k 坐标的输出. $k+1$ 坐标的输出经过一个周期后又反馈给 k 坐标. 这种相互作用关系随 DE 作用次数的增加越来越密切，传递也越来越远. 这实际上机械地体现了因子之间的相互作用，也即体现了因子对预报量的非线性作用.

三、实际应用步骤

结合台风总降水量分布预报加以概述.

1. 构造样品空间

规定以温州为圆心，以冲绳岛为半径的 ± 50 km 的环带为起报区. 将 1956 — 1984 年进入该起报区的台风个例共 135 个作为样品空间. 即 $m=135$. 样品序号按日期先后从 *1* 排到 135（自然数）. 由此组成序列 $\{f_i\}(i=1, 2, \cdots\cdots, 135)$.

2. 确定因子（坐标）

上述筛选法的思想萌发于对台风降水预报的多年探索. 台风路径被公认为台风降水的最重要因子，这就解决了使用上述算法的前提条件. 我们也理所当然地将移动方向作为输出坐标. 但是台风降水的复杂性显然也不能由路径唯一确定. 另外几个与台风降水直接有关的因子是：台风中心强度；台风中心强度变化；台风位置；台风风圈范围. 台风降水应视为台风和中纬度环境流场相互作用的结果. 环境流场的因子是从 500 hPa 上将指标站的风矢进行分解，按区域相加得出. 它们是：

① 福州、温州、衢州、赣州、南昌 SE 分速之和；
② 上述站 SW 分速之和；
③ 银川、延安、西安、太原、北京、郑州、宜昌、济南 NW 分速之和；
④ 上述站 SW 分速之和；
⑤ 石桓、冲绳、庵美 NE 分速之和；
⑥ 上述站 *NW* 分速之和；
⑦ 女岛、济州岛、射阳 NE 分速之和；
⑧ 上述站 *NW* 分速之和.

其中奇数序号者是对台风降水有正贡献的因子，偶数序号者是对台风降水有负贡献的因子. 还有一个因子是本省出现不稳定降水的站点数. 一共是 9 个环境流场因子.

为取得比较稳定的台风移动方向，将台风 6 h、12 h、18 h 的移动方向均作为因子，这样与台风有关的因子有 7 个. 总共有 16 个因子，即 $n=16$.

3. 构造筛选网格 $\{\varepsilon_{kl}\}$

本筛选法其实也简单，无非是对每个样品、每个因子多层次地用不等式 $|x_{ik}-x_{ok}|<\varepsilon_{kl}$ 加以判别排序. 故首先要在已知样品空间里对每个因子设计出筛选网格序列 $\varepsilon_{k1}>\varepsilon_{k2}>\cdots$. 等差序列 $\{\varepsilon_{kl}\}$ 为正有理数，我们取因子数 $k=16$，序列项数 $l=31$. 序列首项 ε_{k1} 和公差由因子值在样品空间里的离散程度统计得出. 例如，描述台风位置的经纬度，离散度小，网格就较细，它的 $\{\varepsilon_{2l}\}$ 的首项 $\varepsilon_{21}=1.50$，公差为 0.05；描述环境流场的因子③，离散度大，网格就粗，它的 $\{\varepsilon_{7l}\}$ 的首项 $\varepsilon_{71}=90.00$，公差为 3.00.

如第一段所述，样品空间 $\{f_i\}$ 是以日期先后为序号的全序集. 依据每个因子的 $\{\varepsilon_{kl}\}$，连续进行 DE 变换，可由空间 $\{f_i\}$ 得出空间 $\{f_i^{tn+1}\}$. 该空间则是以距离大小为序号的全序集. 最后一个序号 f_{m-tn-1}^{tn+1} 代表了与待报样品距离最近的历史样品. 应该强调的是，我们得到的是一个历史样品和待报样品的全序关系，而不是它们之间距离的值. $\{\varepsilon_{kl}\}$ 的作用只在于建立全序关系. 图 4 的棱锥通道形象地表示了 $\{\varepsilon_{kl}\}$ 的作用. $|x_{ik}-x_{ok}|$ 好比一只球，它从 ε_{k1}^2 的通道大口处滚入，当球的直径大于或等于通道边长时就停止. 依据停止时的 l 值，将所有历史样品在第 k 个坐标上排序. 这就是

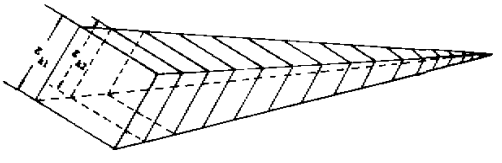

图4 $\{\varepsilon_{kl}\}$ 作用的形象表示

DE $(\{f_{\cdot\cdot}^{k-1}\})=\{f_{\cdot\cdot}^{k}\}$ 的直观表述.

将 16 组的 $\{\varepsilon_{kl}\}$ 连同历史样品因子值作为数据文件存入盘中. 当待报台风进入起报区时，即从台风情报电码和天气图上读数，将数据从键盘上输入 PDP－11 或 IBM－PC（编译）. 从风矢的预处理到输出最近历史样品约 5 — 6 分钟.

四、效果初析

经试验，取最大作用周期 $t=5$.

所谓最近样品是与待报台风 f_o 在后期路径和降水分布上最接近的历史台风 f_e. 本筛选法正确与否，就看其输出序列的末项 $f_{m-tn-1}^{tn+1}=f_{135-5\times16-1}^{5\times16+1}=f_{54}^{81}$ 是否为 f_e.

1. 拟合情形

对 135 个历史样品中的每一个进行筛选. 由于欲拟合者处在样品空间中，故输出序列的末项 f_{54}^{81} 必为其本身，序列的最后第二项 f_{53}^{81} 才是筛选结果. 该结果中有 119 个确系 f_e，还有 16 个不是 f_e. 换句话说，在历史样品中还能找到更接近于 f_o 的样品. 故本筛选法对 16 个样品是失败的，拟合率大于 96%.

2. 试报情形

本筛选法对 1985 — 1986 年 5 个台风进行了业务预报，其中 4 个找到了历史上的最近台风. 8617 找到的是 7410，而 8617 的 f_e 却是 6721，故本筛选法对本次台风预报失败（见表 1）. 本筛选法对 8510 台风的预报优于其它预报方法. 图 5、图 6、图 7 为 8510 与其最近点 7503 的比较.

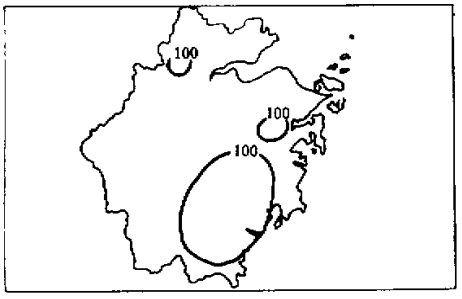

图5　7503 台风浙江省过程降水分布

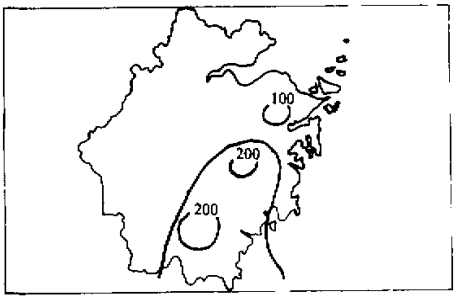

图 6　8510 台风浙江省过程降水分布

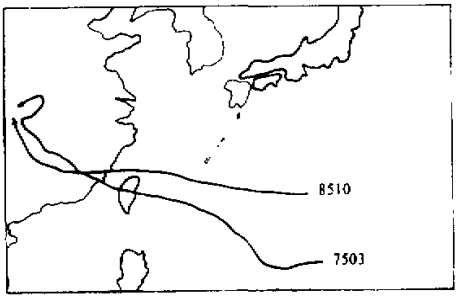

图 7　8510 与 7503 台风的路径

3. 与多维点聚法的比较

表 1 辗转坐标筛选与多维点聚法的比较

待报台风		8506	8510	8605	8615	8617
与待报台风最接近的样品 f_e		6126	7503	6506	7615	6721
输出结果	辗转坐标筛选下的末项 f_{54}^{81}	6126	7503	6506	7615	7410
	多维点聚下的最小欧氏距离	7805	7806	5905	7910	6120

多维点聚法[3]实际上是求欧氏距离. 欧氏距离最小的点即为欧氏距离下的最近点.该点是否为 f_e 呢? 经对同一样本空间的统计, 拟合率仅为 47%. 如表 1 所示, 5 次试报台风中用多维点聚法所找到的历史台风竟没有一个是 f_e. 这并不奇怪, 单用欧氏距离过于简单化, 其误差随维数的增多而急剧加大. 且不谈本筛选法所具有的特点, 至少本筛选法与多维点聚法比起来是好的.

参 考 文 献

[1] 郑维行、王声望, 1985, 实变函数与泛函分析概要, 第五章、第六章, 高等教育出版社.

[2] 李开乐, 1986, 相似离度及其使用技术, 气象学报, 44, 第 2 期, 174 — 183.

[3] 谭冠日, 1980, 气象站数理统计预报方法, 第六章, 科学出版社.