# 非线性判别函数中预报因子的逐步筛选方案

姚棣荣 刘月贞
（杭 州 大 学）（福建省气象科学研究所）

**提 要**

本文针对天气分类预报中采用非线性判别函数所遇到的问题，提出了一个非线性判别函数中预报因子的逐步筛选方案，取得了较为满意的结果. 这对目前天气分类预报中常用的线性判别分析方法是一个改进.

## 一、引 言

目前在天气分类预报中普遍采用的判别分析[1,2]，实质上是一种线性判别，可是在实际上我们所遇到的问题往往是非线性的. 这样，如果仍然采用线性判别分析来解决这种问题，必然影响预报效果，因而考虑非线性判别是必要的. Clark 等[3]在研究湍流扩散时指出，要提高判别效果，应当采用非线性判别函数，但没有给出具体计算方案. Тер-Микртичн 等[4]在雨淞预报中给出了采用线性判别函数和非线性判别函数的计算结果，表明非线性判别函数的预报效果比线性判别函数有所提高，但他们所采用的计算公式相当复杂. Lorenz[5] 也特别强调了进行非线性统计天气预报研究的重要性. 最近，姚棣荣和刘月贞[6]讨论了二级的非线性判别分析，采用了不包含逆矩阵运算的计算方案，取得了更加符合实际的较为满意的结果，显示了非线性判别函数的优越性和合理性.

但是，上述[4]和[6]的工作，都存在这样一个问题，即在非线性判别函数中由于非线性项的出现，会使非线性判别函数的项数随着预报因子数的增多而大大增加，这不仅增加了计算量，而且也无法比较它与线性判别函数的预报效果. 本文试图采用逐步判别分析的思想[7]来解决这一问题，提出了一个非线性判别函数中预报因子的筛选方案，取得了较为理想的结果.

## 二、非线性判别函数

设由 $p$ 个变量 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)'$ 组成的每一个个体分别来自 $G$ 个母体 $A_1, A_2, \cdots A_G$ 中的一个，则由 Bayes 后验概率

1983 年 3 月 18 日收到，6 月 15 日收到修改稿.

$$P(g|\boldsymbol{X})=\frac{Q_g f_g(\boldsymbol{X})}{\sum_{h=1}^{G}Q_h f_h(\boldsymbol{X})}\qquad g=1,2,\cdots G \tag{1}$$

我们可以根据一组预报因子向量 $\boldsymbol{X}$ 来判别预报对象属于哪一个母体.

若设 $G$ 个母体服从多元正态分布 $N(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$, $g=1,2,\cdots G$, 而且第 $g$ 个母体的密度函数为

$$f_g(\boldsymbol{X})=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_g)\right]$$
$$g=1,2,\cdots G \tag{2}$$

式中 $\boldsymbol{\mu}_g$ 是第 $g$ 个母体($g=1,2,\cdots G$)的期望向量,即 $\boldsymbol{\mu}_g=(\mu_{1g},\mu_{2g},\cdots\mu_{pg})'$,而 $\boldsymbol{\Sigma}_g$ 是第 $g$ 个母体 ($g=1,2,\cdots G$) 的协方差矩阵,即

$$\boldsymbol{\Sigma}_g=(\sigma_{ij})_g=\begin{pmatrix}\sigma_{11} & \sigma_{12}\cdots\sigma_{1p}\\ \sigma_{21} & \sigma_{22}\cdots\sigma_{2p}\\ \cdots\cdots & \cdots\cdots\\ \sigma_{p1} & \sigma_{p2}\cdots\sigma_{pp}\end{pmatrix}_g$$

在各个母体的协方差矩阵不相等的假定下, 对 $Q_g f_g(\boldsymbol{X})$ 作对数运算,并以 (2) 式代入,略去与 $g$ 无关的项,则得非线性判别函数为

$$U_g(\boldsymbol{X})=-\frac{1}{2}\boldsymbol{X}'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{X}+\boldsymbol{\mu}_g'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{X}-\frac{1}{2}\boldsymbol{\mu}_g'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\mu}_g$$
$$-\frac{1}{2}\ln|\boldsymbol{\Sigma}_g|+\ln Q_g\qquad g=1,2,\cdots G \tag{3}$$

式中 $Q_g$ 为第 $g$ 个母体 ($g=1,2,\cdots G$) 的先验概率.

实际上,总体的参数 $\boldsymbol{\mu}_g$ 和 $\boldsymbol{\Sigma}_g(g=1,2,\cdots G)$ 常常是未知的,需要利用样本资料对这些参数作出估计,即

$$\boldsymbol{\mu}_g\approx\bar{\boldsymbol{X}}_g=(\bar{x}_{1g},\bar{x}_{2g},\cdots\bar{x}_{pg})'\qquad g=1,2,\cdots G \tag{4}$$
$$\boldsymbol{\Sigma}_g\approx\boldsymbol{S}_g=(s_{gij})\qquad i,j=1,2,\cdots p\quad g=1,2,\cdots G$$

式中

$$\bar{x}_{ig}=\frac{1}{n_g}\sum_{k=1}^{n_g}x_{igk}\qquad i,j=1,2,\cdots p\quad g=1,2,\cdots G \tag{5}$$

$$s_{gij}=\frac{1}{n_g-1}\sum_{k=1}^{n_g}(x_{igk}-\bar{x}_{ig})(x_{igk}-\bar{x}_{jg})$$

其中 $x_{igk}$ 为第 $i$ 个变量在第 $g$ 个母体中第 $k$ 次观测值, $\bar{x}_{ig}$ 为第 $i$ 个变量第 $g$ 类的平均值, $n_g$ 为第 $g$ 个母体的观测次数. 若记$N$为观测的总次数,则有

$$N=\sum_{g=1}^{G}n_g$$

于是(3)式变成

$$W_g(\boldsymbol{X})=-\frac{1}{2}\boldsymbol{X}'\boldsymbol{S}_g^{-1}\boldsymbol{X}+\bar{\boldsymbol{X}}_g'\boldsymbol{S}_g^{-1}\boldsymbol{X}-\frac{1}{2}\bar{\boldsymbol{X}}_g'\boldsymbol{S}_g^{-1}\bar{\boldsymbol{X}}_g$$

$$-\frac{1}{2}\ln|\boldsymbol{S}_g| + \ln Q_g \qquad g = 1, 2, \cdots G \tag{6}$$

若令

$$\begin{pmatrix} c_{g11} & c_{g12}\cdots c_{g1p} \\ c_{g21} & c_{g22}\cdots c_{g2p} \\ \cdots\cdots & \cdots\cdots\cdots \\ c_{gp1} & c_{gp2}\cdots c_{gpp} \end{pmatrix} = -\frac{1}{2}\boldsymbol{S}_g^{-1} \qquad g = 1, 2, \cdots G \tag{7}$$

$$(c_{g1}, c_{g2}, \cdots c_{gp}) = \bar{\boldsymbol{X}}_g'\boldsymbol{S}_g^{-1} \qquad g = 1, 2, \cdots G \tag{8}$$

$$c_{g0} = -\frac{1}{2}\bar{\boldsymbol{X}}_g'\boldsymbol{S}_g^{-1}\bar{\boldsymbol{X}}_g - \frac{1}{2}\ln|\boldsymbol{S}_g| + \ln Q_g \qquad g = 1, 2, \cdots G \tag{9}$$

则(6)式可以写成

$$W_g(\boldsymbol{X}) = \sum_{i=1}^{p}\sum_{j=1}^{p} c_{gij}x_i x_j + \sum_{i=1}^{p} c_{gi}x_i + c_{g0} \qquad g = 1, 2, \cdots G \tag{10}$$

对于给定的个体 $\boldsymbol{X}_0$，利用(10)式可得 $G$ 个非线性判别函数值 $W_g(\boldsymbol{X}_0)$，$g = 1, 2, \cdots G$。若

$$W_g^*(\boldsymbol{X}_0) = \operatorname*{Max}_{1\le g\le G}\{W_g(\boldsymbol{X}_0)\} \tag{11}$$

则把个体 $\boldsymbol{X}_0$ 划归为母体 $A_g^*$。

如果假定各个母体的协方差矩阵相等，则得线性判别函数为[1]

$$W_g(\boldsymbol{X}) = \bar{\boldsymbol{X}}_g'\boldsymbol{S}_g^{-1}\boldsymbol{X} - \frac{1}{2}\bar{\boldsymbol{X}}_g'\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}_g + \ln Q_g \qquad g = 1, 2, \cdots G \tag{12}$$

式中

$$\boldsymbol{S} = \frac{\boldsymbol{A}}{N-G} \tag{13}$$

其中 $\boldsymbol{A} = (a_{ij})$，$i, j = 1, 2, \cdots p$，是一个组内离差矩阵，其元素为

$$a_{ij} = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{igk} - \bar{x}_{ig})(x_{jgk} - \bar{x}_{jg}) \qquad i, j = 1, 2, \cdots p \tag{14}$$

若记

$$(c_{g1}, c_{g2}, \cdots c_{gp}) = \bar{\boldsymbol{X}}_g'\boldsymbol{S}^{-1} \qquad g = 1, 2, \cdots G \tag{15}$$

$$c_{g0} = -\frac{1}{2}\bar{\boldsymbol{X}}_g'\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}_g + \ln Q_g \qquad g = 1, 2, \cdots G \tag{16}$$

则(12)式可以写成

$$W_g(\boldsymbol{X}) = \sum_{i=1}^{p} c_{gi}x_i + c_{g0} \qquad g = 1, 2, \cdots G \tag{17}$$

对应于(17)式的判别规则同上。

## 三、预报因子的筛选方案

如果给定 $p$ 个预报因子，则对于线性判别函数(17)式共有 $p+1$ 项，而对于非线性判

别函数(10)式，由于非线项的出现，则共有 $2p + c_p^2 + 1$ 项．这样，在预报因子很多时，(10)式中的项数是相当可观的，这不仅使计算量大大增加，而且与线性判别函数相比，由于项数的大大增加而使判别效果的提高不足以证明非线性判别函数比线性判别函数优越．事实上，判别函数中的每一项对区分不同类别天气的能力大小是不同的，为此，对线性判别函数已普遍地采用了逐步判别分析进行预报因子的筛选[7]，而且收到了一定的成效．我们认为，对于非线性判别函数也可以采用逐步判别的思想进行预报因子的筛选，从而建立较佳的非线性判别函数．

关于逐步判别分析的基本原理和计算过程，已有不少论述，具体可参阅文献[2,7]，这里不再重复．下面就建立非线性判别函数中采用逐步判别分析进行预报因子的筛选方案作一简要的说明和介绍．

首先，由给出的 $p$ 个预报因子组合成 $2p + c_p^2$ 个因子，即原有的 $p$ 个预报因子 $x_i(i = 1, 2, \cdots p)$ 加上 $x_ix_j(i, j = 1, 2, \cdots p;\ i \leqslant j)$，然后进行逐步判别分析．若 $p$ 较小时，则直接由 $2p + c_p^2$ 个预报因子输入逐步判别程序进行计算；若 $p$ 较大或者由于计算机容量的限制，则可采用两段筛选方法，即先对 $2p + c_p^2$ 个预报因子计算单因子判别能力大小的 Wilks $\Lambda$ 量

$$U_i = \frac{a_{ii}}{b_{ii}} \qquad i = 1, 2, \cdots 2p + c_p^2 \tag{18}$$

其中

$$a_{ii} = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (x_{igk} - \bar{x}_{ig})^2 \qquad i = 1, 2, \cdots 2p + c_p^2 \tag{19}$$

为变量 $x_i$ 的“但内离差平方和”，而

$$b_{ii} = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (x_{igk} - \bar{x}_i)^2 \qquad i = 1, 2, \cdots 2p + c_p^2 \tag{20}$$

为变量 $x_i$ 的“总的离差平方和”．

式中 $\bar{x}_i$ 为变量 $x_i$ 的总平均值，即

$$\bar{x}_i = \frac{1}{N} \sum_{g=1}^{G} \sum_{k=1}^{n_g} x_{igk} \qquad i = 1, 2, \cdots 2p + c_p^2 \tag{21}$$

同时，利用统计量

$$F_i(G - 1, N - G) = \frac{1 - U_i}{U_i} \frac{N - G}{G - 1} \qquad i = 1, 2, \cdots 2p + c_p^2 \tag{22}$$

对 Wilks $\Lambda$ 量 $U_i$ 进行 $F$ 检验，选出 $U_i$ 较小(即 $F_i$ 较大)的若干个预报因子，然后再输入逐步判别程序进行计算，建立较佳的非线性判别函数．

## 四、计 算 实 例

现以文献[6]的资料说明上述方案的实施情况和效果．

按某气象站1957—1972年汛期(5—6月)总降水量大于700毫米和小于700毫米分成两

类，每类样本数 $n_1 = n_2 = 8$，总样本数 $N = n_1 + n_2 = 16$．选取三个预报因子，即

$x_1$：1月上旬平均温度．

$x_2$：终霜日期(以1月30日为1，依次编号，如2月23日偏为25，余此类推)。

$x_3$：上年8月下旬雨量．

原始资料从略．

### (1) 单因子的判别能力

这里 $p = 3$，则 $2p + c_p^2 = 9$．利用(18)式和(22)式算得9个因子的 Wilks Λ 量以及 $F$ 值(见表1)．

表 1

| | $x_1$ | $x_2$ | $x_3$ | $x_1^2$ | $x_2^2$ | $x_3^2$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ |
|---|---|---|---|---|---|---|---|---|---|
| $U_i$ | 0.736 | 0.796 | 0.945 | 0.822 | 0.664 | 0.999 | 0.999 | 0.986 | 0.857 |
| $F_i$ | 5.03 | 3.59 | 0.82 | 3.04 | 7.09 | 0.01 | 0.01 | 0.20 | 2.34 |

### (2) 计算结果

1. 非线性判别函数

利用表1给出的九个因子，采用逐步判别分析的计算结果见表2．

表 2

| $F$水平 | 入选因子数 | $D_{12}^2$ | $F_{12}$ | $U$ | $x^2$ | 历史拟合率 |
|---|---|---|---|---|---|---|
| 0.0 | 9 | 13.85 | 2.64<br>$F_{0.20}(9,6)$<br>2.05 | 0.20 | 15.21<br>$x_{0.10}^2(9)$<br>14.684 | 94% |
| 2.0 | 3 | 7.29 | 8.33<br>$F_{0.01}(3,12)$<br>5.95 | 0.32 | 14.07<br>$x_{0.01}^2(3)$<br>11.345 | 94% |

表 3

| $F$水平 | 入选因子数 | $D_{12}^2$ | $F_{12}$ | $U$ | $x^2$ | 历史拟合率 |
|---|---|---|---|---|---|---|
| 0.0 | 5 | 8.99 | 5.14<br>$F_{0.05}(5,10)$<br>3.33 | 0.28 | 14.63<br>$x_{0.05}^2(5)$<br>11.071 | 94% |
| 2.0 | 3 | 7.29 | 8.33<br>$F_{0.01}(3,12)$<br>5.95 | 0.32 | 14.07<br>$x_{0.01}^2(3)$<br>11.345 | 94% |

表中 $D_{12}^2$ 和 $U$ 分别为检验判别效果的 Mahalanobis 距离和 Wilks Λ量，而 $F_{12}$ 和 $x^2$ 分

别为相应的 $F$ 检验和 $x^2$ 检验所采用的统计量，$F_\alpha(n_1,n_2)$为自由度是 $n_1$ 和 $n_2$、显著性水平是 $\alpha$ 时 $F$ 统计量的临界值，$x^2_\alpha(n)$ 为自由度是 $n$、显著性水平是 $\alpha$ 时 $x^2$ 统计量的临界值．

若采用两段筛选方法，先由表 1 按 $F_i$ 的大小选取五个因子(即 $x_2^2, x_1, x_2, x_1^2, x_1x_3$)，再进行逐步判别分析，其结果见表 3．

由表 2 和表 3 可见，Mahalanobis 距离 $D^2_{12}$ 随入选因子数的减小而减小，而 Wilks Λ 量 $U$ 则随入选因子数的减小而增大，这是必然的．但是从判别效果的显著性检验来看，虽然 $D^2_{12}$ 值在采用九个因子时与选取五个因子和三个因子时有较大的差异，而在取五个因子与取三个因子时则差异不太大，可是，通过显著性检验的显著性水平 $\alpha$ 却分别为 0.20，0.05 和 0.01；$U$ 值在采用九个因子、五个因子和三个因子时差异并不大，而与之相对应的通过显著性检验的显著性水平 $\alpha$ 则分别为 0.10，0.05 和 0.01；另外，三者的历史拟合率都是 94%．综上所述，我们可以说在三个非线性判别函数中以选取三个因子的非线性判别函数的判别效果最显著．于是有

$$\begin{aligned} W_1(\boldsymbol{X}) &= -0.0073x_2^2 + 0.8943x_1 + 0.6410x_2 - 7.2540 \\ W_2(\boldsymbol{X}) &= -0.0135x_2^2 + 2.1498x_1 + 0.9416x_2 - 14.0275 \end{aligned} \tag{23}$$

2. 线性判别函数

为了便于比较，我们同时由给出的三个预报因子 $(x_1, x_2, x_3)$，建立线性判别函数，其结果见表 4．

表 4

| $F$ 水平 | 入选因子数 | $D^2_{12}$ | $F_{12}$ | $U$ | $x^2$ | 历史拟合率 |
|---|---|---|---|---|---|---|
| 0.0 | 3 | 3.55 | 4.06<br>$F_{0.05}(3,12)$<br>3.49 | 0.50 | 8.75<br>$x^2_{0.05}(3)$<br>7.815 | 75% |
| 2.0 | 2 | 3.52 | 6.54<br>$F_{0.05}(2,13)$<br>3.81 | 0.50 | 9.05<br>$x^2_{0.05}(2)$<br>5.991 | 81% |

可见，在选取三个因子与二个因子时，$D^2_{12}$ 差异很小，$U$ 值却是相同的，能通过显著性检验的 $\alpha$ 水平也均为 0.05，但在选取二个因子时的 $F_{12}$ 和 $x^2$ 与 $\alpha$=0.01 时临界值 $F_{0.01}(2, 13) = 6.70$ 和 $x^2_{0.01}(2) = 9.210$ 比较接近，而在取三个因子时差异却很大；另外，历史拟合率也是取二因子时比取三因子时要高．因此，取二个因子的线性判别函数的判别效果更为显著，则有

$$\begin{aligned} W_1(\boldsymbol{X}) &= 0.5292x_1 + 0.1638x_2 - 4.6583 \\ W_2(\boldsymbol{X}) &= 1.4738x_1 + 0.0580x_2 - 5.1271 \end{aligned} \tag{24}$$

为了检验上述方案的预报效果，我们取 1973—1982 年十年的资料作为独立样本，利用(23)式和(24)式进行试报，得到线性判别函数的预报准确率为 70%，而非线性判别函数的预报准确率达 90%．可见，采用非线性判别函数的预报效果是令人相当满意的．

综合上述计算结果，我们可以看到，采用逐步筛选因子的方案所得到的三个因子的非

线性判别函数，无论从历史拟合率、试报效果，还是从 Mahalanobis 距离 $D_{12}^2$ · Wilks Λ 量 $U$ 的数值以及显著性检验来看，都比选用三个因子或二个因子的线性判别函数优越．因为非线性判别函数中的因子数与线性判别函数中的因子数相当，这说明这种非线性判别函数的判别效果比线性判别函数确有提高，而且它比较符合实际，也比较合理，因而本文提出的非线性判别函数中预报因子的逐步筛选方案是有实际使用价值的．

## 五、结 语

1．通过以上分析表明，本文所提供的非线性判别函数中预报因子的筛选方案是可行的，采用这一方案可以得出较佳的具有实际使用价值的非线性判别函数，而且这一方案已经在电子计算机上实现，可以进行大量的试验和计算．

2．本文所提供的方案对于多个母体的分类预报问题同样适用．

3．本文的工作是初步的，对于非线性判别函数以及本方案的实际应用，尚需进行深入、系统的研究．

本工作曾得到 李麦村 同志的热情鼓励和指导，徐钟济先生审阅过本文的初稿，并提出了宝贵的意见，谨此致谢．

## 参 考 文 献

[1] 王宗皓、李麦村等，天气预报中的概率统计方法，科学出版社，102—115，1974．
[2] 李麦村、姚棣荣，杭州大学学报(自然科学版)，第1期，51—72，1977．
[3] Clark, T. L., Scoggins, J. R. and Cox, R. E., *Mon. Wea. Rev.*, **103**, 514—520, 1975.
[4] Тер-Микртчян, М. Г., Снитковский, А. И., Лукиянова, Л. Е., Труды Гидрометцентра СССР, выпю 90, 3—39, 1971.
[5] Lorenz, E. N., *The collection of papers presented at the WMO Symposium on probabilistic and statistical methods in weather forecasting*, Nice, 8—12 September 3—8, 1980.
[6] 姚棣荣、刘月贞，杭州大学学报(自然科学版)，10，127—138，1983．
[7] 李麦村、姚棣荣、杨自强，应用数学学报，第4期，58—73，1977．

# THE SCREENING OF PREDICTORS IN NONLINEAR DISCRIMINANT FUNCTION

Yao Dirong
*(Hangzhou University)*

Liu Yuezheng
*(Institute of Meteorology, Fujian Province)*

## Abstract

This article analyses the problems in adopting the nonlinear discriminant function in weather typing prediction, and provides a stepwise screen way of predictors in nonlinear discriminant function. Through the calculation of a case, we have obtained good results in application of this method. Its forecast effect is, to some extent, better than that of the linear and the original nonlinear discriminant functions. This method is an improvement for that of linear discriminant analysis using in weather typing prediction at present.