

用模糊聚类分析方法作春季连阴雨预报*

邹 浩 陆善峻 楼世博 陈化成

(上海市气象局)

(上海铁道学院)

为了利用预报因子的模糊信息,进行综合分析,以提高春季连阴雨的预报水平,我们通过对 1959—1978 年 3 月 16 日—5 月 15 日上海连阴雨过程的分析,选取预报因子,试用模糊聚类分析方法,作春季连阴雨逐日短期预报,取得了较好的效果。

1. 春季连阴雨的时段和预报因子

1) 连阴雨的时段

考虑到逐日雨量,日照时数的不同组合,对农作物有不同程度的影响。我们拟定了一个指数,以定义一个“阴雨天”,凡“阴雨天”连续四天以上,过程总雨量大于 17 毫米,即作为一个连阴雨过程(具体规定略)。根据嘉定县气象站资料(上海地区基本一致),划分了各年连阴雨的时段,如表 1:

表 1 上海春季连阴雨时段

样本 序号	年 份	日 期	持续 天数	总雨 量 (mm)	平均 日照 时数	样本 序号	年 份	日 期	持续 天数	总雨量 (mm)	平均日照 时 数
1	1959	5.5—5.17	13	107.9	1.2	*	1968	5.4—5.11	8	44.1	0.9
2	1960	4.9—4.15	7	61.5	1.5	16	1969	4.15—4.21	7	39.2	1.6
3	1960	5.3—5.10	8	58.7	1.7	17	1969	4.24—4.28	5	30.7	1.1
4	1962	4.29—5.4	6	41.4	0.4	18	1970	5.1—5.6	6	81.1	0.3
5	1963	4.18—4.24	7	38.4	0.3	19	1972	3.15—3.20	6	28.9	0.1
6	1963	4.27—5.4	8	62.9	1.3	20	1973	3.29—4.3	6	33.5	0.9
7	1963	5.7—5.16	10	84.4	1.5	21	1973	4.8—4.11	4	52.9	0.0
8	1964	4.3—4.15	13	85.7	2.3	22	1975	4.15—4.19	5	31.8	0.0
9	1965	3.20—3.23	4	34.3	0.3	23	1976	4.17—4.23	7	34.8	3.0
10	1966	4.7—4.10	4	43.4	0.1	24	1977	3.15—3.18	4	24.5	0.8
11	1966	4.21—4.25	5	42.7	0.7	25	1977	4.4—4.9	6	36.7	1.1
12	1967	3.25—3.30	6	64.2	2.7	26	1977	5.1—5.11	11	118.9	1.6
13	1967	4.6—4.10	5	81.9	0.2	27	1978	3.26—3.30	5	26.6	1.5
14	1967	4.14—4.19	6	66.0	1.7	28	1978	4.15—4.23	9	17.9	1.2
15	1968	5.4—5.13	10	58.0	1.1						

* 该段连阴雨受资料限制,暂不作为聚类样本。

2) 预报因子

根据天气学分析,我们选取了十个因子(均为各次连阴雨开始前两天 20 点钟的资料),它们是:

* 本文于 1981 年 1 月 8 日收到,1981 年 11 月 4 日收到修改稿。

x_1 : 500 mb 图上 $45^\circ-65^\circ\text{N}$, $25^\circ-65^\circ\text{E}$ 区域内有无 24 小时移速 ≤ 5 个经度的高压中心或高压脊。如有, x_1 取值为 1, 否则为 0。

x_2 : 500 mb 图上 $30^\circ-45^\circ\text{N}$, $45^\circ-65^\circ\text{E}$ 区域内有无低压或低压槽(风向切变)。如有, x_2 取值为 1, 否则为 0。

x_3 : 500 mb 图上, 浙江省衢县的风向, 如为 W-S 风, x_3 取值为 1, 否则为 0。

x_4 : 500 mb 图上, 南昌、衢县、大陈三个站, 如出现 $H \geq 576$, $\Delta T_{24} > 0$, $\Delta H_{24} > 0$ 三个条件, 则 x_4 取值为 1, 否则为 0。

x_5 : 地面天气图上, $40^\circ-55^\circ\text{N}$, $75^\circ-100^\circ\text{E}$ 区域内有无冷高压中心, 有则 x_5 取值为 1, 否则为 0。

x_6 : 700 mb 图上, $40^\circ-55^\circ\text{N}$, $75^\circ-100^\circ\text{E}$ 区域内, 每 5 个纬距间的温差极大值。

x_7 : 700 mb 上海(ΔT_d)₂₄。

x_8 : 700 mb 南昌(ΔT_d)₂₄。

x_9 : 700 mb 芷江(ΔT_d)₂₄。

x_{10} : 地面天气图上, 上海减成都的气压差。

上述因子中, x_1, x_2 是对上游阻塞高压与切断低压活动情况的确定, 一般说, 阻塞或切断系统都比较稳定, 对下游环流形势有重要影响。为了进一步明确东亚环流是否平直, 又选取 x_3 , 即地处 29°N 附近的衢县高空风向, 以描述东亚大槽是否建立。这三个因子, 大致可以反映东亚高空环流的基本特点。选取 x_4 的目的, 是要反映江南暖高压脊的活动情况。 x_5, x_6 则说明北方冷空气强度。 x_7, x_8, x_9 是利用西南气流中三个不同站的露点变化, 以表示江南水汽输送情况。 x_{10} 是根据预报经验选取的, 它大致可反映 850 mb 以下的低层气压场形势。

为了进行模糊聚类分析, $x_6, x_7, x_8, x_9, x_{10}$ 诸因子需要用下式进行标准化处理:

$\bar{X}_i = \frac{x_i - x_{im}}{x_{iM} - x_{im}}$, 式中 x_{iM}, x_{im} 分别为 x_i 可能取值的上界和下界。这样, 显然有: $0 \leq \bar{X}_i \leq 1$ 。

于是, 我们准备分析的 28 个样本, 每个样本 X_k 均为一个 10 维向量, 且每个分量的值为 $0 \leq x_{ki} \leq 1, i=1, 2, \dots, 10$ 。

2. 春季连阴雨的两模糊聚类分析方法

28 个连阴雨样本之间的相似程度是不同的, 区分的界限是不清晰的, 用模糊聚类分析方法, 按其亲近程度进行分类, 能比较恰当地反映它们本来的实际情况。

1) 基于模糊等价关系的聚类方法

在[1]中已介绍了这种方法。我们计算不同样本向量间的夹角余弦, 由此构成 28×28 阶的模糊关系矩阵。用不同的 λ 水平来截此矩阵, 就得到不同的分类。

我们将 1977 年与 1978 年春季的每一天均作为样本, 用上述办法进行分类, 在尽可能区分连阴雨和非连阴雨的前提下, 取 $\lambda = 0.94$, 用此 λ 将 28 个样本进行分类, 结果分得 13 类(见表 2)。这表明有 13 种不同类型的春季连阴雨模式。

2) ISODATA 模糊聚类方法

设待分类样本集为:

表 2 Fuzzy 等价关系的聚类结果

类 别	样 本 序 号
I	1, 11, 14, 17, 23, 24, 26;
II	2, 3;
III	4, 12, 25;
IV	5, 20;
V	6;
VI	7;
VII	8;
VIII	9, 28;
IX	10, 13, 16, 21;
X	15;
XI	18, 19;
XII	22;
XIII	27;

$$X = \{X_1, X_2, X_3, \dots, X_n\}$$

每个样本有 S 个因子, 故可写成: $X_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{is})$

现要将样本集 X 分为 C 类, 和通常分类不同, 在模糊分类中, 每个样本不是清晰地划分到某类, 而是按不同的隶属程度隶属于各类。用 u_{ij} 表示第 j 个样本从属于第 i 类的隶属程度。于是我们用模糊分类矩阵 $U = [u_{ij}]_{c \times n}$ 来表示 X 上的一个模糊分类。 U 应满足如下条件:

- (1) $u_{ij} \in [0, 1]$ $0 \leq \text{隶属度} \leq 1$ 。
- (2) $\sum_{i=1}^c u_{ij} = 1$ 样本隶属度之和为 1。
- (3) $\sum_{j=1}^n u_{ij} > 0$ 每一类不空。

满足上述三个条件的模糊分类矩阵的全体, 称之为模糊划分空间, 用 M_{fc} 表示:

$$M_{fc} = \{U | u_{ij} \in [0, 1] \forall i, j; \sum_{i=1}^c u_{ij} = 1 \forall j; \sum_{j=1}^n u_{ij} > 0 \forall i;\}$$

为了使模糊划分 U 是合理的, 要使泛函:

$$J_m(U, V) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m \|X_k - V_i\|^2 (m > 1)$$

达到最小, 式中 $U \in M_{fc}$, $V_i (i=1, 2, \dots, c)$ 是第 i 类的聚类中心, $\|\cdot\|$ 是 R^s 空间中任一内积所导出的范数, m 是大于 1 的某一实数。 $J_m(U, V)$, 称为目标泛函。为了求出对应的 U 与 V 使 J_m 达到最小, 可用如下公式及算法:

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{\|X_k - V_j\|}{\|X_k - V_i\|} \right)^{\frac{2}{m-1}}} \quad \forall i, k \quad (1)$$

$$V_i = \frac{\sum_{k=1}^n (u_{ik})^m X_k}{\sum_{k=1}^n (u_{ik})^m} \quad \forall i \quad (2)$$

算法如下:

(1) 任选一组初始分划矩阵 $U \in M_{fc}$ 。

(2) 据(2)式计算出 V_i 。

(3) 据(1)式计算出 $U^* = (u_{ik}^*)$

4° 若 $\max_{1 \leq i \leq c} \{ |u_{ik}^* - u_{ik}| \} \leq \varepsilon$, 则 U^* 即为所求分划矩阵, 否则用 U^* 代替 U 回到(2)°,

进行再一次迭代。 ε 为精度控制数, ε 愈小, 所得结果愈精确(通常 ε 取 10^{-3} 、 10^{-4} 等)。

对于已得到的 U , 我们将 U 的每一列中隶属度最大的元素取成 1, 其余均取为 0, 则得到样本集 X 的一个清晰分类。

在本问题中 $n=28, s=10$, 取 $m=2, c=13$ (即有 13 种类型连阴雨天气模式), 范数 $\|\cdot\|$ 由下式确定:

$$\|X_k - V_i\|^2 = \sum_{l=1}^{10} (x_{kl} - v_{il})^2$$

并取模糊等价关系聚类方法所分 13 类作为本方法的初始分类, 进行运算, 算得的 U 如表 3, V_i (略)。

表 3 ISODATA 模糊聚类所得到的 U 矩阵(矩阵中每个元素均 $\times 100$)

77	6	5	4	4	0	0	0	0	2	91	7	15	90	0	2	64	0	0	4	2	0	63	63	4	86	0	0
6	80	83	3	2	0	0	0	1	1	2	5	10	3	0	1	8	1	1	2	1	0	9	10	3	3	0	1
3	2	2	79	1	0	0	0	0	2	1	61	15	1	0	2	5	0	0	2	2	0	5	5	78	2	0	1
5	3	3	3	86	0	0	0	1	1	2	5	0	2	0	1	8	1	1	86	1	0	9	8	3	3	0	1
0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	3	3	3	2	0	0	0	96	1	1	5	9	1	0	1	3	1	1	2	1	0	3	3	3	1	0	96
6	3	3	7	2	0	0	0	1	92	2	12	47	3	0	92	9	0	0	2	91	0	8	9	7	3	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	
2	3	2	2	2	0	0	0	1	1	1	5	5	1	0	1	3	98	97	2	1	0	3	3	3	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	

令 $r_i = \max_{[X_k \in \text{第 } i \text{ 类}]} (\|X_k - V_i\|)$

称 r_i 为第 i 类的聚类半径。

为了运算方便, 我们用聚类半径的平方 r_i^2 作为检算 X_k 是否属于第 i 类的标准, 即若:

$$\|X_k - V_i\|^2 \leq r_i^2 \quad \text{则}$$

X_k 属于第 i 类, 反之, 则 X_k 不属于第 i 类。

此外, 由于历史上的春季连阴雨样本不够多, 所得 r_i^2 不能准确的描述该类样本的变

化范围。为此,我们对 r_i^2 作了一些适当的调整,其值如表 4 所示。

表 4 各类半径平方值

类 别	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
r_i^2	0.352	0.352	0.352	0.352	0.087	0.087	0.087	0.352	0.607	0.087	0.352	0.087	0.087

3. 预报方法和效果检验

1) 预报方法:

首先将待预报的样本用 x_{10} 进行筛选,如 $x_{10} < 0$ 则该样本不参加聚类,可直接预报未来 48 小时无连阴雨出现。但若 $x_{10} < 0$ 是由于 $27^\circ-33^\circ\text{N}$, $118^\circ-122^\circ\text{E}$ 及 $27^\circ-29^\circ\text{N}$, $122^\circ-128^\circ\text{E}$ 区域内出现的低压或波动而造成的,则该样本仍需参加聚类。

用模糊等价关系的聚类方法预报连阴雨:将待预报的样本和原连阴雨样本放在一起进行聚类,经计算机运算得出模糊等价关系矩阵,再用 $\lambda=0.94$ 来进行分类,若待预报样本聚入了某一类连阴雨,就预报有连阴雨,否则预报无连阴雨。

用 ISODATA 模糊聚类方法预报连阴雨,先求出待预报样本 X^* 与每个聚类中心的距离,若 X^* 对某个 V_i 为 $\|X^*-V_i\|^2 \leq r_i^2$,则预报未来 48 小时前后将开始连阴雨,反之,则预报未来 48 小时无连阴雨。

以上第一种方法,使用时要上机计算,第二种方法,用台式计算机或手算即可进行。

2) 效果检验:

我们将 1979 年 3 月 16 日—5 月 15 日逐日作为独立样本进行试报,结果 3 月 29 日—4 月 1 日和 4 月 29 日—5 月 15 日两段连阴雨,都在前 48 小时前后报出,用逐日评定办法,分别计算预报有连阴雨的准确率和预报无连阴雨的准确率,评定结果是:

模糊等价关系方法,预报有连阴雨的准确率为 63%,预报无连阴雨的准确率为 76%。

ISODATA 方法,预报有连阴雨的准确率为 80%,预报无连阴雨的准确率为 81%。

1980 年 3 月 16 日—5 月 15 日,用 ISODATA 方法实际进行了逐日预报,结果预报有连阴雨的准确率为 88%,预报无连阴雨的准确率为 85%,其中 3 月 28 日—3 月 31 日和 4 月 24 日—4 月 29 日两段连阴雨,都在前 48 小时预报出来了。

模糊等价关系方法在 1980 年 3 月 16 日—5 月 15 日中检验效果比 ISODATA 方法差。

参 考 文 献

- [1] 曹鸿兴,陈国范,天气过程的模糊划分,科学通报,1980 年 10 期。
- [2] Bezdek, T. C. P. F. Castela, Prototype Classification and Feature Selection with Fuzzy Set, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol SMC-7, No. 2 Feb 1977.