

文章编号: 1004- 4574(2010) 04- 0129- 06

基于模糊聚类算法的洪水预报研究

钱镜林^{1,2}, 张 晔¹, 刘国华²

(1. 浙江省水利河口研究院, 浙江 杭州 310020 2. 浙江大学建筑工程学院, 浙江 杭州 310027)

摘 要: 考虑降雨时空分布、降雨平均强度和土壤初始含水量对洪水过程的影响, 将这些影响因子分为数值特征影响因子和属性特征影响因子, 运用模糊 C 均值算法对历史洪水样本模糊聚类, 然后根据不同的类建立不同的洪水预报模型。实例验证表明, 该方法可以有效提高洪水预报的精度。

关键词: 降水; 洪水预报; 模糊聚类

中图分类号: P457. 6

文献标识码: A

Flood forecast research based on fuzzy cluster algorithm

QIAN Jing-lin^{1,2}, ZHANG Ye¹, LIU Guo-hua²

(1. Zhejiang Institute of Hydraulics & Estuary, Hangzhou 310020, China 2. Zhejiang University, Hangzhou 310027, China)

Abstract The flood characteristic influence factors were divided into the value characteristic influence factors and the attribute characteristic influence factors in this paper. In the identical kind of flood, the influence on the flood of these two kind of characteristics influence factors is the same. Thus, the characteristic weighted fuzzy clustering algorithm (FCM) was used to classify the history flood sample, and then the parameters of each kind of the Xin'an River model were optimized with each kind of sample separately. The application shows that the method presented in the article may improve the precision of the flood forecast effectively.

Key words precipitation; flood forecast; fuzzy cluster

洪水预报是流域防洪的重要内容之一, 从洪水预报技术发展情况来看, 多因素经验相关、径流成因分析和流域水文数学模型, 构成了 3 个主要阶段, 它们是洪水预报方法从经验估计达到理论预测的进步的表现, 尤其是电子计算机的普及, 使得原来只能做定性讨论的径流形成机制可以深化为实用的数学模型。

然而, 目前洪水预报的精度仍难以满足实际要求, 主要原因是目前使用的模型大部分为概念性模型, 这种模型的特点为输入集中、参数集总, 与分散输入、分布参数的流域客体之间存在着矛盾^[1]。即水文要素在空间平均后, 引起了表现水文规律的歧变, 模拟的水文状态不能在流域中找到真实的对应, 进而导致模拟误差。然而这种集总型模型又是必须的, 因为即使现代计算技术有可能运用分布式模型体系来进行流域模拟, 也会由于其中的分布参数难以通过分析和观测手段实时量测得到, 或无法从地质、地形条件资料中推出, 从而失去应用的客观性, 这种参数确定上的困难在实际建模中比设计理想的分布式模型困难更大, 使得运用分布式模型的努力难以完全实现。

洪水过程是流域上各点净雨通过流域汇流所形成的结果, 往往受到降雨时空分布、降雨强度、土壤初始含水量以及下垫面条件的影响。对于某一特定流域, 由于洪水观测数据做了一致性处理, 下垫面的影响可以

收稿日期: 2008- 10- 12 修订日期: 2009- 09- 23

基金项目: 浙江省自然科学基金资助项目 (Y507710, Y507046)

作者简介: 钱镜林 (1982-), 男, 博士, 主要从事洪水预报研究。E-mail: qjln@yahoo.com.cn

不做考虑。因此,本文仅考虑降雨的时空分布、降雨平均强度、土壤初始含水量对洪水过程的影响,并将其归纳为两类,命名为数值特征影响因子和属性特征影响因子。其中数值特征影响因子包括平均雨强、流域土壤初始含水量;属性特征影响因子则为暴雨中心所在位置。以每场洪水为研究对象,将这些影响因子作为洪水样本特征,运用模糊 C 均值算法将洪水样本分类,对每一类洪水分别建模分析。

1 模糊 C 均值聚类^[2-4]

模糊 C 均值聚类 (FCM) 是用隶属度确定每个数据点属于某类的程度的一种聚类算法。1973 年 Bezdek 提出了该算法,作为早期硬 C 均值聚类 (HCM) 方法的一种改进。

FCM 把 n 个向量分为 c 个模糊组,并求每组的聚类中心,使得非相似性指标的价值函数达到最小。FCM 的价值函数 (或目标函数) 一般形式为:

$$J(U, c_1, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m d_{ij}^2 \tag{1}$$

其中, $\sum_{i=1}^c \mu_{ij} = 1, j=1, \dots, n$ (2)

式中 $\mu_{ij} \in [0, 1]$ 表示第 j 个数据点属于第 i 个聚类的程度, $d_{ij} = \|c_i - x_j\|$ 为第 i 个聚类中心与第 j 个数据点间的欧氏距离, c_i 为模糊组 i 的聚类中心, $m \in [1, \infty)$ 为加权指数。

构造如下新的目标函数,可求得使 (1) 式达到最小值的必要条件。

$$\begin{aligned} \bar{J}(U, c_1, \dots, c_c, \lambda_1, \dots, \lambda_n) &= J(U, c_1, \dots, c_c) + \sum_{j=1}^n \lambda_j (\sum_{i=1}^c \mu_{ij} - 1) = \\ &= \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m d_{ij}^2 + \sum_{j=1}^n \lambda_j (\sum_{i=1}^c \mu_{ij} - 1). \end{aligned} \tag{3}$$

这里, $\lambda_j (j=1, \dots, n)$ 是式 (3) 的 n 个约束式的拉格朗日乘子。对所有输入参量求导,使式 (1) 达到最小的必要条件为:

$$c_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j}{\sum_{j=1}^n \mu_{ij}^m}, \tag{4}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}. \tag{5}$$

由上述两个必要条件,得到模糊 C 均值聚类算法的简单迭代过程:

- 1) 生成 0~1 之间的随机数初始化隶属度矩阵 U , 使其满足式 (2) 中的约束条件;
 - 2) 用式 (4) 计算 c 个聚类中心 $c_i (i=1, \dots, c)$;
 - 3) 根据式 (1) 计算目标函数。如果它小于某个确定的阈值, 或它相对于上次目标函数数值的改变量小于某个阈值, 则算法停止;
 - 4) 用式 (5) 计算新的 U 阵, 返回步骤 2)。
- 当算法收敛时, 就得到了各类的聚类中心和各个样本对于各类的隶属度值, 从而完成了模糊聚类划分。

2 模糊 C 均值聚类分析在洪水预报中的应用

2.1 应用实例介绍

某预报站以上集雨面积 2480km², 该站以上流域内有两座大型水库, 在洪水预报中, 两座大型水库采用实时下泄流量为模型的输入条件, 根据马斯京根法推算至出口断面, 剩余流域面积洪水采用新安江三水源模型进行预报。

2.2 洪水样本特征提取

在模糊聚类分析计算中, 数值特征影响因子根据实测值计算得到, 土壤初始含水量是指预报计算时段初的土壤含水量, 主要通过产汇流日模型计算而得; 平均雨强是指预报时段以来的平均雨强。属性特征影响因子则根据表 1 进行赋值,

即将整个流域从上游往下游均分成 3 等份, 如暴雨中心位于上游属性值为 1, 中游为 2 下游为 3。

最终, 根据数值特征影响因子和属性特征影响因子组成的流域历史洪水的特征样本系列见表 2 所示。

表 1 属性特征赋值

Table 1 Value assignment of attribute character

特征描述	上游	中游	下游
属性值	1	2	3

表 2 历史洪水样本特征

Table 2 Characteristic values of historical flood samples

洪号	序号	暴雨中心	平均雨强 / (mm · h ⁻¹)	土壤初始含水量 / mm	洪号	序号	暴雨中心	平均雨强 / (mm · h ⁻¹)	土壤初始含水量 / mm
860410	1	2	12.6	89.1	900904	21	2	6.2	90
860708	2	2	15.7	85	900909	22	1	10.7	80.1
860905	3	2	19.7	69.8	900912	23	1	9.2	58.6
860919	4	2	7.2	68.6	910417	24	2	18.5	89.8
870620	5	1	7.2	68.7	910619	25	1	10.8	70.8
870722	6	2	7	43.2	910812	26	2	9.7	40.4
870727	7	2	16.6	90	920622	27	2	5.7	54.7
870908	8	2	13.2	59.1	920624	28	1	6.9	85
880617	9	2	14.9	74.3	920701	29	3	4.8	89.6
880807	10	2	32	28.5	920829	30	3	8.4	89.3
890411	11	3	5.3	51.8	920922	31	3	19.6	69.6
890522	12	2	8.7	81.6	930501	32	2	10.1	62.4
890616	13	2	12.9	42	930618	33	2	9.9	90
890630	14	2	17.8	90	930630	34	2	9.6	90
890723	15	2	18.4	38.3	930703	35	2	11.8	85.2
890726	16	2	7.2	61.7	950428	36	1	7.6	83.9
890915	17	1	23.6	66.2	950530	37	1	18.5	55.9
900421	18	1	8.4	83.1	950625	38	1	17.9	80.5
900531	19	1	6.8	66.8	950702	39	1	11.2	62.4
900830	20	2	19.4	31.1	960630	40	2	13	60.3

由于预报站历史洪水样本个数相对较少, 根据历史洪水特征和流域模型预报经验, 洪水样本可以分为 5 大类, 隶属度计算结果见表 3 所示。

表 3 模糊聚类隶属度表

Table 3 Membership degree of fuzzy cluster

序号	样本代码	各类别隶属度				
		1	2	3	4	5
1	860410	0.31	0.13	0.17	0.33	0.06
2	860708	0.47	0.18	0.11	0.17	0.07
3	860905	0.10	0.05	0.06	0.14	0.64
4	860919	0.05	0.89	0.02	0.03	0.01
5	870620	0.22	0.13	0.08	0.50	0.06
6	870722	0.15	0.52	0.11	0.18	0.04
7	870727	0.24	0.10	0.34	0.13	0.18
8	870908	0.24	0.16	0.10	0.46	0.05

续表 1

序号	样本代码	各类别隶属度				
		1	2	3	4	5
9	880617	0.17	0.08	0.06	0.64	0.05
10	880807	0.14	0.11	0.12	0.15	0.48
11	890411	0.11	0.69	0.05	0.12	0.03
12	890522	0.11	0.08	0.68	0.08	0.04
13	890616	0.15	0.35	0.09	0.35	0.06
14	890630	0.20	0.08	0.22	0.14	0.35
15	890723	0.17	0.15	0.10	0.36	0.22
16	890726	0.08	0.82	0.03	0.06	0.02
17	890915	0.14	0.10	0.15	0.14	0.47
18	900421	0.77	0.07	0.05	0.09	0.02
19	900531	0.03	0.94	0.01	0.02	0.00
20	900830	0.15	0.14	0.12	0.36	0.23
21	900904	0.30	0.30	0.23	0.13	0.04
22	900909	0.16	0.10	0.49	0.12	0.12
23	900912	0.21	0.27	0.22	0.25	0.04
24	910417	0.34	0.11	0.18	0.17	0.20
25	910619	0.14	0.71	0.05	0.09	0.02
26	910812	0.21	0.30	0.08	0.31	0.10
27	920622	0.17	0.40	0.11	0.28	0.05
28	920624	0.39	0.36	0.10	0.11	0.03
29	920701	0.40	0.33	0.10	0.14	0.04
30	920829	0.62	0.12	0.09	0.12	0.05
31	920922	0.06	0.03	0.04	0.07	0.80
32	930501	0.08	0.06	0.03	0.80	0.02
33	930618	0.63	0.17	0.08	0.10	0.03
34	930630	0.28	0.15	0.15	0.35	0.07
35	930703	0.06	0.03	0.85	0.04	0.02
36	950428	0.20	0.11	0.55	0.11	0.03
37	950530	0.06	0.03	0.04	0.07	0.80
38	950625	0.17	0.09	0.55	0.12	0.08
39	950702	0.38	0.14	0.11	0.27	0.10
40	960630	0.10	0.08	0.05	0.73	0.04

根据隶属度值进行模糊聚类划分, 最终结果见表 4

表 4 模糊 C 均值聚类分析的分类结果

Table 4 Classification result from fuzzy cluster analysis

类号	各类洪水的序号	类号	各类洪水的序号
1	2, 18, 24, 28, 29, 30, 33, 39	4	1, 5, 8, 9, 13, 15, 20, 26, 32, 34, 40
2	4, 6, 11, 16, 19, 21, 23, 25, 27	5	3, 10, 14, 17, 31, 37
3	7, 12, 22, 35, 36, 38		

2.3 模型率定结果

表 5 列出了根据 5 大类历史洪水样本优选出的三水源新安江模型的参数。

表 5 预报站各类洪水模型参数率定结果表

Table 5 Flood model's parameters optimized from 5 kinds of historical flood samples

类号	K	WM	WUM	WIM	B	C	SM	EX	KI	KG	CS	CI	CG	XE
1	0.86	90	25	50	0.36	0.16	7	1.3	0.15	0.72	0.77	0.7	0.996	0.4
2	0.80	90	25	50	0.33	0.16	7	1.7	0.13	0.72	0.774	0.79	0.92	0.5
3	0.88	90	25	50	0.36	0.16	7	1.4	0.16	0.74	0.766	0.79	0.996	0.4
4	0.82	90	25	50	0.32	0.16	10	1.6	0.11	0.73	0.75	0.72	0.993	0.5
5	0.89	90	25	50	0.36	0.16	10	1.3	0.15	0.75	0.743	0.78	0.997	0.4

分别对聚类前后的率定成果进行综合比较 (见表 6), 通过对比分析, 应用聚类分析方法的新安江模型率定精度有明显的提高。

表 6 聚类前后历史洪水样本的率定成果统计表

Table 6 Comparison of determined result of historical flood samples between before and after cluster

类号	平均径流深		洪峰流量		平均确定性系数	
	误差	合格率 %	误差	合格率 %		
聚类后	1	5.32	85.7	7.45	85.7	0.91
	2	4.24	100.0	3.92	100.0	0.91
	3	4.10	100.0	5.08	100.0	0.94
	4	3.78	100.0	3.12	100.0	0.94
	5	5.97	83.3	4.29	100.0	0.91
聚类前	5.86	82.5	5.92	85.0	0.90	

2.4 模型验证分析

本文选择了 1997-2005 年 15 场洪水对模型进行验证, 类似于实时洪水预报方法, 采取逐时段递推预报。在预报作业时, 首先根据当前获得的降雨特征确定实时洪水相对于率定各类的隶属情况, 进行相应分类。然后提取所在类的模型参数值进行洪水预报。因此, 每做一次预报均要进行一次聚类分析。验证洪水最终聚类结果见表 7 所示。

2.5 洪水预报精度评定

根据《水文情报预报规范》, 洪水预报精度评定项目包括洪峰流量、峰现时间、洪量和洪水过程等。当一次预报误差小于许可误差时, 为合格预报。合格预报的次數与总预报次数之比的百分数为合格率, 当合格率达到 85% 以上, 则预报精度为甲级。

降雨径流预报以实测洪峰流量的 20% 作为许可误差, 峰现时间以预报根据时间至实测洪峰出现时间之间时距的 30% 作为许可误差, 径流深预报以实测值的 20% 作为许可误差。

洪水预报过程与实测过程之间的吻合程度用确定性系数作为指标, 当 $C_D > 0.9$ 预报精度为甲级。

C_D 按下式计算:

$$C_D = 1 - \frac{\sum_{i=1}^n [y_c(i) - y_0(i)]^2}{\sum_{i=1}^n [y_c(i) - \bar{y}_0]^2} \quad (6)$$

式中: C_D 为确定性系数, $y_c(i)$ 为预报值, $y_0(i)$ 为实测值, \bar{y}_0 为实测值均值, n 为资料序列长度。

验证洪水预报结果见表 8 所示。从表 8 可以看出, 模拟预报的精度可以达到水文洪水预报规范的甲级要求, 精度较高。

表 7 验证洪水样本隶属情况分析表

Table 7 Verification of membership situation of flood samples

类号	样本洪号
1	19990616, 20010623
2	19980623, 20000619, 20000623
3	19980618, 19990523, 20010625
4	19970707, 19990416, 20000608
5	19970713, 19980919, 20020627, 20050911

表 8 洪水预报结果验证

Table 8 Verification of flood- forecasting results

洪号	径流深误差 %	洪峰误差 %	峰现时差	确定性系数	聚类类别
19970713	2.96	0.07	0	0.92	5
19980919	6.05	0.39	0	0.85	5
20020627	5.39	1.45	0	0.95	5
19970707	5.28	5.48	0	0.91	4
19990416	2.63	2.24	0	0.95	4
20000608	1.93	0.27	0	0.96	4
19980618	0.99	4.15	3	0.94	3
19990523	0.43	0.01	0	0.95	3
20010625	4.60	5.62	0	0.95	3
20000619	8.61	2.13	0	0.93	2
20000623	0.87	2.93	0	0.95	2
19980623	0.11	0.12	0	0.93	2
19990616	9.74	6.32	0	0.94	1
20010623	9.51	2.71	0	0.92	1
20050911	5.93	0.68	0	0.91	1
平均值	4.34	2.30		0.93	

3 结论

(1) 本文从考虑影响洪水的影响因子出发, 将洪水样本影响因子分为数值特征影响因子和属性特征影响因子两大类, 建立基于这两者特征影响因子的 FCM 算法模型, 通过对历史洪水验证发现, 本方法有效的提高了洪水预报的精度。

(2) 由于历史洪水样本的获取存在一定的局限性和特殊性, 一般而言, 流域性洪水的发生概率是中小洪水居多, 大洪水偏少, 样本的收集过程中, 各类样本数量相差较大, 这种数量上的差异对样本特征选取产生的影响值得进一步研究。

参考文献:

- [1] 程根伟, 舒栋材. 水文预报的理论与数学模型 [M]. 北京: 中国水利水电出版社, 2006.
- [2] 何清. 模糊聚类分析理论与应用研究进展 [J]. 模糊系统与数学, 1998, 12(2): 89- 94.
- [3] BEZDEK JC. A physical interpretation of fuzzy ISODATA [C]. IEEE Trans SMC, 1976 6(3): 387- 390
- [4] 高新波. 模糊聚类分析及其应用 [M]. 西安: 电子科技大学出版社, 2004. 49- 60
- [5] 刘国华, 钱镜林, 汪树玉. 产汇流模型参数综合率定在洪水预报中的应用 [J]. 浙江大学学报 (工学版), 2003 5