基于主成分分析的河流洪水系统聚类法

包为民[1,2],万新宇[1,2],荆艳东[1,2],江 鹏[1,2]

(1.河海大学水文水资源与水利工程科学国家重点实验室,江苏 南京 210098;2.河海大学水文水资源学院,江苏 南京 210098)

摘要:进行河流洪水聚类的目的是根据洪水特征的相似程度划分洪水类别,研究同类洪水的规律性以及应对措施.但是,洪水特征选择过多往往会增加计算的复杂程度,同时特征之间的相关性也使得信息大量重叠,导致计算结果失真.为此,提出基于主成分分析的河流洪水系统聚类法.该法首先将所选的洪水特征综合成少数几个不相关的主成分,然后计算出每场洪水在各主成分上的得分值并将该值作为新的洪水特征值,最后根据这些新特征值进行洪水聚类.三门峡水库入库洪水聚类实例证明了该方法的可行性.

关键词:洪水;聚类;主成分分析

中图分类号:P338 文献标识码:A 文章编号:1000-1980(2008)01-0001-05

在水文研究中,历史水文资料非常重要,因为它蕴藏着大量的有用信息.近年来,随着数据仓库和数据挖掘技术[1-2]的日趋成熟,针对历史水文资料的数据挖掘研究也在不断开展[3-4].聚类(clustering)是数据挖掘的一项重要内容,其应用主要表现在如下 3 个方面[5-10]:(a)对水文资料进行分组,寻找同类水文现象的规律;(b)对水文预报模型的参数进行率定和优化;(c)对洪水灾害等级进行划分.本文研究的是河流洪水聚类问题.通常,聚为一类的洪水之间表现出较强的相似性,反之较弱.同类洪水间的相似性可以为防洪决策提供必要的信息支持.比如,利用流域水文模型预报出未来的洪水过程后,通过洪水聚类可以在历史洪水库中检索出与其相似的历史洪水,这些相似历史洪水以及当时的决策处理方法,对于未来洪水的防洪决策都有着非常重要的参考价值和实际意义[11].在实际研究中,历史洪水常用其特征来刻画,表征洪水过程的指标(特征)较多,多沙河流洪水更甚,许多指标之间又具有相关性,这就不可避免地造成了信息的大量重叠,这种信息重叠有时甚至会抹煞事物的真正特征和内在规律,这也是其他聚类研究普遍存在的问题.为此,本文提出了基于主成分分析(principal components analysis,简称 PCA)的河流洪水系统聚类方法,并利用自由软件对多沙河流洪水进行了聚类研究.

1 基于 PCA 的河流洪水聚类法

1.1 洪水指标集的建立

洪水过程通常表现为一时间序列.聚类前,需对洪水过程进行特征提取,即进行洪水指标的选择.选择的原则主要是要选择那些对某一水文现象影响显著的特征因子.选择的方法主要有 3 种:(a)理论分析;(b)影响因子相关分析;(c)模式识别[5].

1.2 聚类步骤

首先,进行洪水主成分分析,在尽可能保留原有洪水指标所含有的信息的基础上,将原来较多的且具有相关性的洪水指标简化为少数几个新的不相关的综合性指标;其次,利用第 1 步分析的结果,根据洪水之间在几个新的综合指标上的相似性来进行洪水聚类分析.

1.2.1 主成分分析

主成分分析是一种化繁为简,尽可能压缩指标数的降维(即空间压缩)技术,其结果是把多个指标简化为少数几个新的综合指标.新的综合指标称为原指标的主成分,并且按其所含有的信息量的大小依次称为第 1

收稿日期:2007-03-08

基金项目:国家自然科学基金(50679024);国家科技支撑计划课题(2006BAC05B02)

作者简介:包为民(1956—),男,浙江东阳人,教授,博士,主要从事洪水预报和防洪管理研究.

主成分,第2主成分……新指标要求满足:(a)尽可能保留原有指标所含有的信息;(b)各指标之间不相关,即各自含有的信息不重叠.文献[12]对主成分分析方法的原理和性质作了详细介绍,本文仅给出如下主成分分析的计算步骤.

a. 输入洪水特征的观测值.设有 n 场洪水,对每场洪水测得 p 个指标(变量)的数据,这 $n \times p$ 个数据构成一个洪水特征观测矩阵

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \square & & \square \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

其中 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,p)$ 为第 i 场洪水的第 j 个指标的观测值.

b. 计算各洪水指标的均值和标准差:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} \qquad (j=1,2,\cdots,p) \tag{1}$$

c. 标准化洪水特征观测矩阵 $\boldsymbol{X}$,得标准化数据阵 $\boldsymbol{Y}=(y_{ij})_{n\times p}$,再据此计算其相关矩阵 $\boldsymbol{R}=(r_{jk})_{p\times p}$:

$$y_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j} \qquad (i=1,2,\cdots,n;j=1,2,\cdots,p) \tag{2}$$

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} y_{ij}y_{ik} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ij}-\bar{x}_j)}{s_j}\frac{(x_{ik}-\bar{x}_k)}{s_k} \qquad (j,k=1,2,\cdots,p) \tag{3}$$

d. 求 $\boldsymbol{R}$ 的特征值及特征向量.若能通过正交换 $\boldsymbol{Q}$,使 $\boldsymbol{Q}'\boldsymbol{R}\boldsymbol{Q}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$,则 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 为 $\boldsymbol{R}$ 的 p 个特征值.不妨设 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$,则 $\boldsymbol{Q}$ 的各列 $\boldsymbol{l}_j=(l_{1j},\cdots,l_{pj})'$,即为 λ_j 所对应的正则化特征向量(亦称载荷值),$j=1,2,\cdots,p$.

e. 确立主成分.按累计方差贡献率 $\sum_{j=1}^{k}\lambda_j/\sum_{j=1}^{p}\lambda_j > 85\%$(或80%)的准则,确定 k,从而取前 k 个主成分:

$$\boldsymbol{Z}_j = \boldsymbol{l}'_j\boldsymbol{Y} = l_{1j}\boldsymbol{Y}_1 + \cdots + l_{pj}\boldsymbol{Y}_p \qquad (j=1,2,\cdots,k) \tag{4}$$

其中 $\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_p$ 为标准化后的指标向量,$1\leqslant k\leqslant p$.

f. 计算前 k 个主成分的样本值(或称得分值),即 $z_{ij}=\sum_{t=1}^{p} y_{it}l_{tj}(i=1,2,\cdots,n;j=1,2,\cdots,k)$,从而可得新指标(主成分)矩阵 $\boldsymbol{Z}=(z_{ij})_{n\times k}$,以其代替原洪水特征观测矩阵 $\boldsymbol{X}=(x_{ij})_{n\times p}$ 作聚类分析,便可将问题简化.

1.2.2 系统聚类法

洪水聚类,就是对洪水样本进行分组.与一般的分类不同,聚类中的组不是预先定义的,而是根据实际洪水之间的相似性来定义的,同组中的成员是相似的,不同组中的成员是不相似的.有关聚类的方法较多,本文采用其中最常用、最基本的一种方法——系统聚类法[12].系统聚类的具体步骤如下:

a. 计算 n 个洪水样品两两间的距离 d_{ij},记作 $\boldsymbol{D}=(d_{ij})_{n\times n}(i=1,2,\cdots,n;j=1,2,\cdots,n)$.洪水样品 i 和 j 之间的距离通常可以采用绝对值距离、欧氏距离、明考斯基距离和切比雪夫距离等.这些距离与指标的量纲有关,具有一定的人为性;另一方面,这些距离均没有考虑指标间的相关性.因此,在计算前须对样本进行标准化处理,并要设法降低指标间的相关性,而主成分分析法就具有这2项功能.这里采用欧氏距离,即

$$d_{ij} = \left[\sum_{k=1}^{p}(z_{ik}-z_{jk})^2\right]^{\frac{1}{2}} \tag{5}$$

b. 构造 n 个类,每个类只包含一场洪水样品,即各场洪水各自形成最相似的一类.

c. 合并距离最近的2类为一新类.类间距离与上述洪水样品间的距离不同,常用的有最短距离法、最长距离法和类平均法等8种[12],本文采用类平均法.将2类之间的距离平方定义为这2类元素两两之间的平均平方距离,即任给2类 G_{p} 和 G_{q},得

$$D^2_{\mathrm{pq}} = \frac{1}{n_{\mathrm{p}}n_{\mathrm{q}}}\sum_{i\in G_{\mathrm{p}}}\sum_{j\in G_{\mathrm{q}}} d^2_{ij} \tag{6}$$

其中 n_{p} 和 n_{q} 分别为类 G_{p} 和 G_{q} 的洪水样本个数.假设 G_{p} 和 G_{q} 合并为某一类 G_{r},G_{r} 内有洪水 $n_{\mathrm{r}}=n_{\mathrm{p}}+n_{\mathrm{q}}$ 个,则类 G_{r} 与类 G_{k} 的类平均法聚类递推公式为

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 \tag{7}$$

式中下标 r,p,q,k 表示不同的洪水类别.

d. 计算新类与当前各类的距离.若类的个数等于 1,转至步骤 e,否则回到步骤 c.

e. 画聚类图(或称谱系图).

f. 根据求解问题的需要,确定类的个数和类.

2　应 用 实 例

本文以三门峡水库建库以来的入库潼关站 170 场洪水为例,进行多沙河流洪水聚类研究.一般地,洪水特征属性较多,对于多沙河流,还要考虑洪水中的泥沙属性,根据相似性原则选择 10 个比较有代表性的洪水特征指标,即洪水总量、洪水历时、洪峰流量、洪峰出现时间、起涨流量、涨峰段洪量、泥沙总量、沙峰含沙量、沙峰出现时间以及涨峰段沙量,分别记为 $X_1, X_2, \cdots, X_{10}$,具体数据见表 1.

表 1　洪水指标值

Table 1　Index values of floods

洪水序号	洪水总量/亿 m^3	洪水历时/h	洪峰流量/($m^3 \cdot s^{-1}$)	洪峰出现时间/h	起涨流量/($m^3 \cdot s^{-1}$)	涨峰段洪量/亿 m^3	泥沙总量/亿 t	沙峰含沙量/($kg \cdot m^{-3}$)	沙峰出现时间/h	涨峰段沙量/万 t
1	11.90	240	2210	78	1150	4.55	0.151	22.0	120	700
2	29.10	577	1760	88	830	3.52	0.271	20.0	130	200
3	14.00	431	2400	23	930	1.36	0.127	28.0	26	300
4	8.36	264	1440	116	380	3.62	0.060	12.7	120	200
5	15.00	256	3330	22	935	1.75	0.906	201.0	23	1600
6	21.70	227	5400	49	1660	5.81	0.775	59.4	50	2600
7	27.80	453	2990	25	1870	1.88	0.588	60.4	43	400
8	6.37	192	1530	27	902	1.18	0.049	11.4	28	100
9	3.75	138	1660	11	690	0.48	0.603	217.0	36	700
10	30.30	267	5600	87	2410	10.50	2.710	219.0	80	9500
11	45.30	432	4180	251	1950	28.00	2.650	144.0	227	17600
12	65.50	433	6070	87	2060	11.20	2.020	71.5	90	4300
□	□	□	□	□	□	□	□	□	□	□
170	13.00	336	1620	12	1250	0.62	0.137	15.7	239	100

计算洪水特征指标之间的相关性,相关系数见表 2.从表 2 可以看出,一些洪水特征指标间存在一定的相关性,相关系数的变化范围为 $-0.183 \sim 0.797$.相关程度最大的是 X_1 与 X_6(正相关).因此,在进行洪水聚类之前,需对该洪水特征样本进行必要的处理.

表 2　相关系数

Table 2　Correlation coefficient

变量名	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1	0.601	0.449	0.519	0.522	0.797	0.428	−0.034	0.348	0.475
X_2	0.601	1	−0.088	0.566	−0.097	0.412	0.152	−0.027	0.521	0.246
X_3	0.449	−0.088	1	−0.075	0.512	0.280	0.712	0.397	−0.114	0.399
X_4	0.519	0.566	−0.075	1	−0.052	0.771	0.171	−0.004	0.727	0.561
X_5	0.522	−0.097	0.512	−0.052	1	0.344	0.139	−0.183	−0.092	0.069
X_6	0.797	0.412	0.280	0.771	0.344	1	0.336	−0.024	0.456	0.645
X_7	0.428	0.152	0.712	0.171	0.139	0.336	1	0.748	0.066	0.749
X_8	−0.034	−0.027	0.397	−0.004	−0.183	−0.024	0.748	1	−0.106	0.586
X_9	0.348	0.521	−0.114	0.727	−0.092	0.456	0.066	−0.106	1	0.256
X_{10}	0.475	0.246	0.399	0.561	0.069	0.645	0.749	0.586	0.256	1

对洪水特征指标样本观测值进行主成分分析,提取主成分 Z_j,计算结果见表 3 ~ 5.其中,表 3 为计算得到的各成分的特征值、贡献率以及累计贡献率.从中可以看出,前 4 个成分的方差累计贡献率已经达到了 88.8%,超过了一般所要求的85%,所以本文选取 Z_1 为第1主成分、Z_2 为第2主成分、Z_3 为第3主成分

和 Z_4 为第 4 主成分,该 4 个主成分基本上保留了原来指标 $X_1, X_2, \cdots, X_{10}$ 的信息. 表 4 为原 10 个洪水指标在上述 4 个主成分上的载荷值. 载荷值反映了所取主成分与各原始指标之间的相关关系,反映了各指标对选出的主成分所起的作用. 表 5 为各场洪水在前 4 个主成分上的得分值,可以用这 4 个综合指标代替 10 个原指标,其得分值作为新指标的样本值.

表 3 各成分的特征值和贡献率

Table 3 Eigenvalue and contribution ratio of each component

主成分	特征值	贡献率	累计贡献率	主成分	特征值	贡献率	累计贡献率
Z_1	4.139	0.414	0.414	Z_6	0.262	0.026	0.963
Z_2	2.422	0.242	0.656	Z_7	0.140	0.014	0.977
Z_3	1.645	0.164	0.821	Z_8	0.109	0.011	0.988
Z_4	0.678	0.068	0.888	Z_9	0.074	0.007	0.996
Z_5	0.487	0.049	0.937	Z_{10}	0.044	0.004	1.000

表 4 4 个主成分的荷载值

Table 4 Loading of each flood index on 4 principal components

主成分	Z_1	Z_2	Z_3	Z_4	主成分	Z_1	Z_2	Z_3	Z_4
X_1	-0.412	0.070	-0.322	0.286	X_6	-0.422	0.148	-0.166	-0.339
X_2	-0.278	0.321	0.112	0.744	X_7	-0.334	-0.408	0.182	0.149
X_3	-0.241	-0.446	-0.250	0.117	X_8	-0.157	-0.432	0.459	0.071
X_4	-0.363	0.342	0.159	-0.310	X_9	-0.259	0.384	0.158	-0.073
X_5	-0.147	-0.151	-0.664	-0.090	X_{10}	-0.400	-0.175	0.244	-0.315

表 5 各场洪水在主成分上的得分值

Table 5 Score of each flood on 4 principal components

洪水序号	Z_1	Z_2	Z_3	Z_4	洪水序号	Z_1	Z_2	Z_3	Z_4
1	-0.014	2.054	-0.101	-0.165	86	-0.014	2.054	-0.101	-0.165
2	-1.517	3.493	0.170	2.789	87	-1.517	3.493	0.170	2.789
3	0.753	1.294	-0.255	2.123	88	0.753	1.294	-0.255	2.123
4	0.107	2.744	0.828	-0.224	89	0.107	2.744	0.828	-0.224
5	0.567	-0.273	0.118	0.983	90	0.567	-0.273	0.118	0.983
6	-0.581	0.152	-1.228	0.294	91	-0.581	0.152	-1.228	0.294
7	-0.387	1.094	-1.217	2.552	92	-0.387	1.094	-1.217	2.552
8	1.810	0.783	-0.231	0.113	93	1.810	0.783	-0.231	0.113
9	1.791	-0.270	0.681	-0.048	94	1.791	-0.270	0.681	-0.048
10	-3.437	-0.563	-0.768	-0.090	95	-3.437	-0.563	-0.768	-0.090
11	-8.648	3.455	0.311	-1.766	96	-8.648	3.455	0.311	-1.766
12	-4.459	1.156	-2.210	2.255	97	-4.459	1.156	-2.210	2.255
□	□	□	□	□	□	□	□	□	□
85	-2.280	2.404	-1.241	0.721	170	-0.058	2.955	0.231	1.095

根据表 5 中主成分新样本值进行洪水聚类研究. 本文仅给出前 40 场洪水谱系图,如图 1 所示,其中纵坐标 H 表示将要合并的 2 类之间的距离,类间距离最小的,最先合并成一类,比如 14 号和 30 号洪水,其各项洪水指标都非常接近. 根据不同的距离阈值(H 值),可以将洪水集划分为不同的洪水类进行研究. 由于研究目的的不同,划分洪水类别的距离阈值也不尽相同,目前主要依靠专家经验和问题求解的精度[13]来确定. 在 170 场洪水聚类谱系图上,聚集密度较大(洪水数大于 5)的有 9 处,其余的聚集密度较小,有部分洪水独自成类,原因主要在于水量、沙量、洪峰、沙峰以及水沙搭配的不同.

图 1 洪水聚类谱系图

Fig.1 Sketch of flood clustering

3 结　　语

洪水聚类,是水文数据挖掘中的一项重要内容.本文以三门峡水库入库洪水为例对多沙河流洪水进行了聚类研究,并对洪水特征指标之间的相关性问题作了必要的处理,结果表明主成分分析方法在处理此类问题上是简单有效的.

本文研究的最终目的不是仅仅要得到洪水聚类的结果,而是要考察同类洪水之间的相似性,研究相似洪水形成的原因以及相应的处理方法,为防洪决策提供决策参考.本研究只是初步的,以下 3 个方面的问题还需进一步研究:

a. 洪水特征指标的选择.在不同研究中,洪水考察的侧重点不同,其特征指标的选择也会不同.例如,在研究黄河调水调沙时,还需考虑洪水的泥沙特征.

b. 洪水特征指标权值的确定.本文对于洪水间距离的定义还只是一般性的定义,洪水特征指标(或者主成分)所起的作用相同.事实上,各特征指标对一场洪水整体上的相似性有不同的贡献,因而在进行洪水聚类时还需考虑特征指标的权值.

c. 洪水聚类结果的合理性分析.阈值(H 值)的选择不同,所得到的聚类结果也不尽相同,这就需要根据洪水的形成机制和特征对洪水聚类结果的合理性进行分析.

参考文献:

[1] 史忠植.知识发现[M].北京:清华大学出版社,2002:1-20.

[2] 苏新宁,杨建林,江念南,等.数据仓库和数据挖掘[M].北京:清华大学出版社,2006:1-12.

[3] 艾萍,王志坚,索丽生,等.水文数据在线分析与知识发现系统模型研究[J].水利学报,2001(11):15-19.

[4] 张弛.数据挖掘技术在水文预报与水库调度中的应用研究[D].大连:大连理工大学,2006.

[5] 陈小红,叶锦昭,杨德林.洪水模型识别与模拟:长江三峡工程设计洪水的推求[J].水利学报,1995(9):82-88.

[6] 马寅午,周晓阳,尚金成,等.防洪系统洪水分类预测优化调度方法[J].水利学报,1997(4):1-9.

[7] LUCHETTA A. A real time hydrologic forecasting system using a fuzzy clustering approach[J]. Computers & Geosciences,2003,29:1111-1117.

[8] SEE L,OPENSHAW S. Applying soft computing approaches to river level forecasting[J]. Hydrological Sciences Journal,1999,44(5):763-778.

[9] 于雪峰,陈守煜.模糊聚类迭代模型在洪水灾害度划分中应用[J].大连理工大学学报,2005,45(1):128-131.

[10] 陈亚宁,杨思全.灰色聚类在洪水灾害等级划分中的应用[J].干旱区地理,1999,22(3):37-42.

[11] 包为民.洪水预报信息利用问题研究与讨论[J].水文,2006,26(2):18-21.

[12] 张济世,刘立昱,程中山,等.统计水文学[M].郑州:黄河水利出版社,2006:101-125.

[13] DUNHAM M H. 数据挖掘教程[M].郭崇慧,田凤占,靳晓明,等,译.北京:清华大学出版社,2005:107-140.

Flood clustering method based on principal component analysis

BAO Wei-min[1,2], WAN Xin-yu[1,2], JING Yan-dong[1,2], JIANG Peng[1,2]

(1. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing* 210098, *China*;
2. *College of Hydrology and Water Resources, Hohai University, Nanjing* 210098, *China*)

Abstract: The aim of flood clustering is to classify the flood category by the similarity of flood characteristics, to study the law of similar floods and to find countermeasures, however the selection of too many flood characteristics always makes the calculation more complicated, and the correlation between the characteristics leads to a great amount of information overlapping and incorrect result of calculation. Therefore, the clustering method for river floods based on principal component analysis was proposed. In this method, the flood characteristics selected were classified into a few uncorrelated principal components, then the scores of each flood on each principal component were calculated and taken as the new characteristic values of flood. Finally flood clustering was performed according to these new characteristic values. The feasibility of the present method was verified by case study on inflow flood of Sanmenxia Reservoir.

Key words: flood; clustering; principal component analysis