

贝叶斯判别分析在中长期径流预报中的应用研究

张锐^{1,2}, 王本德¹, 张双虎², 王浩²

(1.大连理工大学 建设工程学部, 辽宁 大连 116024; 2.中国水利水电科学研究院 水资源研究所, 北京 100038)

摘要: 基于贝叶斯判别分析原理, 考虑径流形成的特点及物理成因, 选择研究区碧流河水库 1959~2011 年的实测资料, 且将径流分为 5 个级别, 用相关系数模型, 考虑因子的不确定性, 筛选出相关性高、彼此间独立性强的 10 个判别因子, 分别构建基于朴素贝叶斯分类器(N-Bayes)和贝叶斯判别准则(D-Bayes)的年径流属性级别预报模型, 并利用研究区 2001~2011 年资料进行了试报和检验。预报结果表明: 两种模型的预报结果相近, 准确率均超过 70%, D-Bayes 的预报效果略优于 N-Bayes, 初步说明: 贝叶斯判别分析原理在水文预报中的应用, 有着较好的前景。

关键词: 中长期; 贝叶斯判别分析; 判别因子; 碧流河水库

中图分类号: P338

文献标识码: A

文章编号: 1000-0852(2015)05-0001-05

引言

目前, 就国内外的研究现状而言, 长期预报准确率的统计上限可达 80%~85%, 而各国的实际水平多数在 60%~70%。在不同地区和各年、季、月的实际预测中, 其准确性也是时高时低, 难于满足指导生产实践的要求。为此, 相关领域科研工作者, 曾按一定的原则将预报对象划分为多个属性级别, 并将不同属性级别视为不同模式, 根据合理的预报因子运用诸如神经网络、模糊推理等的模式识别或推理的方法对预报对象的属性级别进行判别, 在水资源高效利用中发挥着非常重要的作用。近年来, 又在径流分级预报的基础上, 考虑水文水资源行为的动态性、不稳定性、非平衡性, 寻求同类型水文现象规律和物理机制等方面做了大量研究。周惠成等^[1], 针对径流多因素分级预报中因子的确定问题, 构建了粗集-模糊推理径流分级预报模型; 周晓阳等^[2], 用系统聚类方法将历史洪水数据分成若干类型, 构建模糊概率预测的隶属度回归方法, 对未来洪水进行分类概率预测; 王本德等^[3], 针对暴雨洪水形成机制的随机性、确定性的问题, 归纳出考虑暴雨成因的大伙房水库流域洪水过程分类方法等。鉴此, 本文拟建立一个中长期径流的多因素分级预报模型。基本的思路

是: 将碧流河水库 1959~2011 年的年径流资料进行聚类并划分为 5 级, 基于成因分析, 提取显著的相关预报因子, 以贝叶斯判别分析原理为主导思想, 定性判别碧流河水库年径流属性级别, 其预报结果可为防汛减灾非工程措施提供参考信息。

1 贝叶斯判别分析原理

所谓判别分析是在数据分类情况已知的前提下, 通过分析现有训练数据集, 总结分类规则、描述和区分数据类型或概念的模型, 对类别标号未知的对象进行归类的过程^[4]。贝叶斯判别分析理论是当前判别分析算法中较为常用的一种。其基本思想是根据最大后验准则找到最可能的分类。基本原理是:

设观测样本集的分类特征向量 $C=\{C_1, \dots, C_m\}$, 待分析对象为 $X=\{x_1, x_2, \dots, x_n\}$, 其中 m 为观测样本集的属性类别数, 向量 X 表示属性变量 A_1, A_2, \dots, A_n 样本的 n 个度量。则已知 X 取值条件下, 归属于各属性级别 C_k 的概率 $P(C_k|X)$, 其中 $k \in (1, 2, \dots, m)$, 由贝叶斯定理^[5]表示为:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad (1)$$

式中: $P(C_k)$ 为属性类别发生的先验概率; $P(X)$ 为属性

收稿日期: 2013-09-20

基金项目: 国家重点基础研究发展计划(973 计划)项目(2013CB036406)

作者简介: 张锐 (1987-), 女, 陕西西安人, 博士研究生, 主要从事水文学与水资源方面的研究。E-mail: zr0703@163.com

变量 $\{x_1, x_2, \dots, x_n\}$ 同时出现的概率; $P(X|C_k)$ 为分析对象 X 在属性类别 C_k 中出现的条件概率。

那么,依据最大后验准则,找出 X 最有可能的属性类别,即 X 的类标签为 $\max[P(C_k|X)]$ 的类 C_k 。

1.1 朴素贝叶斯网络分类器

朴素贝叶斯分类器(Naive Bayesian Classifier,简称N-Bayes分类器)假定数据对象的属性对于判别分析的影响和其他属性相对独立^[5]。通常,N-Bayes分类器用于处理各属性值间关联关系比较低的数据分类问题,以达到简化计算过程,降低复杂度的目的。基于此,对式(1)做进一步分析:

因 $P(X)$ 对于所有属性类别为常数,故根据最大后验概率准则,只需要使 $P(X|C_k)P(C_k)$ 最大即可。 $P(C_k)$ 可通过下述方法得到:①当类的先验概率未知,则通常假定各属性类别的先验概率为等概率事件,即 $P(C_1)=P(C_2)=\dots=P(C_m)$;②当观测样本具有一定的代表性, $P(C_k)=\frac{T_k}{T}$,其中 T_k 为属性类别 C_k 中的样本个数, T 为样本总数。在此基础上,假设各属性间不存在依赖关系,那么

$$P(X|C_k)=\prod_{i=1}^n P(x_i|C_k) \quad (2)$$

式中: $P(x_i|C_k)$ 可由样本估计得到, x_i 表示待分析对象 X 的属性变量 A_i 的值,需要注意的是:①当 A_i 是离散的,则 $P(x_i|C_k)=\frac{T_{ki}}{T_k}$,其中 T_{ki} 为属性变量 A_i 上具有 x_i 的属性类别 C_k 的样本数。②当 A_i 是连续的,则通常假定连续值属性服从均值 μ ,标准差 σ 的高斯分布 $g(x_i, \mu, \sigma)$,如式(3):

$$P(x_i|C_k)=g(x_i, \mu, \sigma)=\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \quad (3)$$

综上,可由式(1)计算出的最大后验概率的类别,并最终确定其类标签。

1.2 贝叶斯判别准则

贝叶斯判别准则(Bayesian Discriminance,简称D-Bayes)的主导思想是根据已掌握的每个类别的若干样本的数据信息,总结出客观事物分类的规律性建立判别函数,进而确定判别新样本所属类别。其原理可描述为:设观测样本集的属性类别总体为 (C_1, C_2, \dots, C_m) ,且类别 $C_k(k=1, 2, \dots, m)$ 为 n 元总体(考察 n 个指标),则属性类别为 C_k 的样本集矩阵的均值向量为 $\mu_k=(\mu_{1k}, \mu_{2k}, \dots, \mu_{nk})$,协方差矩阵为 Σ_k 。假定错判损失相等,

确定各类别样本集的概率密度函数及 $f_k(x)$ 其先验概率 p_k ,建立判别函数:

$$P(C_k|X)=\frac{p_k f_k(x)}{\sum_{k=1}^m p_k f_k(x)} \quad (4)$$

可知(4)式主要需计算 $p_k f_k(x)$,当 $P(C_k|X)>P(C_i|X)$, $k \neq i(i=1, 2, \dots, m)$,确定 $X \in C_k$ 。若假设属性类别为 C_k 的样本服从正态分布,即

$$f_k(x)=(2\pi)^{-\frac{n}{2}} * \Sigma_k^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1} (X-\mu_k)\right) \quad (5)$$

整理并简化式(5),可得贝叶斯判别公式

$$z_k(X)=\ln p_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + X^T \Sigma_k^{-1} \mu_k \quad (6)$$

判别时,将某一样本 X ,代入式(6)中,计算 $z_k(X)$, $(k=1, 2, \dots, m)$,若 $z_k(X)=\max_{1 \leq i \leq m} z_i(X)$,则 $X \in C_k$ 。

2 方法应用

2.1 资料收集与预报因子遴选

考虑到长期径流预报常用的预报因子为前期的降雨和径流,这两个因子的物理意义很明显,然而其对于后期径流的影响是有限的。因此,本文以1959~2011年(共53年)碧流河水库的历史实测逐月径流量,逐日降雨量和国家气候系统诊断室发布的74项大气环流特征量为基本资料,对其进行分析,从中寻找径流和气象因子之间的统计规律。其中1959~2000年样本用于模型可行性分析,2001~2011年样本用于试报检验。

本文对74项环流指数资料进行提取并以月为尺度整理,构成888个预报因子。此外,结合碧流河水库已有的大量预报经验和实测资料,从物理成因角度出发,添加直接产生径流的降雨因子和间接反映土壤含水量径流因子,并对相应时段的实测资料进行二次分析、计算,避开径流形成机理中的大量物理因素的复杂性,补充选取28项附加因子,分别是:前一年各月实测降雨量、前一年各月实测径流、前一年年降雨量、前一年年径流量、前一年汛期降雨量、前一年汛期径流量。之后,对上述整理的916项因子,按如下步骤进行因子筛选:

(1) 建立前一年的预报因子与下一年的年径流量的对应关系。

(2) 考虑气象因素(74项环流指数构成的要素场)的不确定性,对训练期1959~2000年的不同时段(1959~2000、1969~2000、1979~2000、1989~2000)进行

相关系数排序并选取相关系数排在前 25 位的因子,组成样本数为 100 的关键因子集 1。

(3)若因子集 1 中某一因子重复次数超过 3 次,则表示该因子对预报结果的影响具有稳定性,可作为精选因子。

(4)重新整理因子集 1,保证所有因子无重复出现,剔除步骤 3 中所选的因子,并添加 28 项附加因子,构成新的关键因子集 2。

(5)在步骤 4 的基础上,建立关键因子与预报对象之间的统计相关关系,选择相关系数排在前 6 位的因子,作为精选因子。

(6)由步骤 3 和步骤 5 的精选因子共同构成样本数为 10 的综合判别因子集(见表 1)。

表1 碧流河水库年径流分级判别因子及相关系数
Table1 The classification predictors and the correlation coefficient for annual runoff of the Biliu reservoir

预报因子名称	相关系数	预报因子名称	相关系数
前一年 10 月份太阳黑子*		前一年 1 月份太平洋区涡强度指数	0.412
前一年 12 月份大西洋欧洲环流型 C*		前一年 5 月份亚洲区极涡强度指数	0.41
前一年 9 月份太平洋副高脊线*		前一年 4 月份北半球极涡面积指数	0.408
前一年 11 月份亚洲区极涡面积指数*		前一年 1 月份东亚槽位置	0.405
前一年 9 月份北半球极涡面积指数	0.417	前一年 10 月径流	0.403

注:带“*”为步骤 3 筛选稳定因子。

2.2 建立预报模型

贝叶斯判别分析原理是在具有一定代表性的属性样本数据信息基础上,总结出属性类别划分的规律性,通过概率或判别函数值的形式,判别新样本的所属类别。依据贝叶斯判别分析原理,分别建立研究区年径流分级 N-Bayes 和 D-Bayes 预报模型。其步骤如下:

步骤 1:采用《水文情报预报规范》(GB/T22482-2008)中的距平百分率作为中长期径流预报划分径流丰、平、枯的标准(见表 2)。将训练期样本集分为 5 个属性级别:丰水、偏丰水、平水、偏枯水、枯水,列出各类样本矩阵,即: X_k ,其中($k=1,2,\dots,5$)。通过 $P(C_k)=p_k=\frac{T_k}{T}$ 计算出样本属于第 k 类的先验概率,其中 T_k 和 T 同上文。

步骤 2:假定预报因子服从均值 μ , 标准差 σ 的高

斯分布,使用训练期样本 X 估计分布参数。其中各属性类别样本矩阵的均值向量 $\bar{X}_k=(\bar{x}_{1k},\bar{x}_{2k},\dots,\bar{x}_{10k})$, 方差 σ^2 和协方差矩阵 Σ_k ,且令 $\bar{X}_k=u_k=\mu$ 。

步骤 3:对朴素贝叶斯分类器模型(N-Bayes):将训练期样本数据估计的分布参数 μ, σ^2 代入式(3),可计算属性类别为 C_k 的条件下,第 i 个因子发生的概率 $P(x_i|C_k)$ 。在此基础上,求得试报样本 Y 的条件概率 $P(y_i|C_k)$,结合步骤 1 所得的先验概率 $P(C_k)$,通过式(1)(2)可得试报样本 Y 的后验概率 $P(C_k|Y)$,按照贝叶斯最大后验准则,将预报结果归入后验概率最大的 C_k 类。

步骤 4:对贝叶斯判别准则(D-Bayes):将训练期样本数据估计的分布参数 μ_k, Σ_k 及先验概率 p_k 代入式(6)得到各类别属性的贝叶斯判别函数。将试报样本 Y 代入所得的贝叶斯判别函数中,计算结果为 z_k ,将 z_k 数值中较大的类 k 为 Y 的预报结果。

表2 年径流量等级划分标准
Table2 Biliuhe reservoir annual runoff five rank partition standards

属性级别	丰水	偏丰水	平水	偏枯水	枯水
划分标准	$P>20\%$	$10\%<P\leq 10\%$	$-10\%<P\leq 10\%$	$-20\%<P\leq -10\%$	$P<-20\%$

注: $P=\frac{(x_i-\bar{x})}{\bar{x}}\times 100\%$,其中 x_i 表示第 i 年的年径流量, \bar{x} 表示多年平均径流量。

2.3 模型预报准确性评价

本文结合气象部门的统计方法,引入准确率 α 、漏报率 β 和空报率 γ 作为研究区年径流属性级别判别结果的评分标准^[6],见式(7)~(9)。

$$\alpha = \left(\frac{\sum_{k=1}^m \phi_k}{\sum_{k=1}^m T_k} \right) \times 100\% \quad (7)$$

$$\beta = \left(\frac{\sum_{k=1}^m \varphi_k}{\sum_{k=1}^m T_k} \right) \times 100\% \quad (8)$$

$$\gamma = [1 - (\alpha + \beta)] \times 100\% \quad (9)$$

式中: ϕ_k 为属性类别 C_k 判断准确的次数; φ_k 表示属性类别 C_k 漏报的次数; T_k, m 意义同前。

3 结果分析

3.1 模型可行性分析

为进一步验证两种贝叶斯判别模型的可靠性,本文选取碧流河水库 1959~2000 年 42a 作为训练样本,分别采用 N-Bayes 和 D-Bayes 进行回判验证,并计算不同模型下的预报准确率,结果见表 3。

表3 两种模型回判验证结果统计
Table3 Statistics of discrimination results with different discrimination models

序号	属性分级	样本个数	N-Bayes 回判预测对象属性					准确率	D-Bayes 回判预测对象属性					准确率
			丰水	偏丰	平水	偏枯	枯水		丰水	偏丰	平水	偏枯	枯水	
1	丰水	5	3	1	1			60.00%	4	1				80.00%
2	偏丰	10		8	1	1		80.00%		8	1	1		80.00%
3	平水	13	1	1	9		2	69.23%	1	10	1	1		76.92%
4	偏枯	10			2	7	1	70.00%		2	8			80.00%
5	枯水	4				1	3	75.00%		1	1	2		50.00%
样本总数		42						71.43%						76.19%

由表3可知,两种模型对偏丰的属性初判结果一致,准确率均为80.00%。对丰水、平水和偏枯的属性判别,D-Bayes初判结果略优于N-Bayes;而对于枯水N-Bayes的初判结果略优于D-Bayes。例如:实际属性为平水的样本数是13,按N-Bayes回判验证结果为:1个丰水,1个偏丰,9个平水,2个枯水,准确率为69.23%;按D-Bayes回判验证结果为:1个偏丰,10个平水,1个偏枯,1个枯水,准确率为76.92%。

从两种判别准则模型对5个属性类别的单独统计可得出,D-Bayes的判别结果优于N-Bayes,从总体上看,两种模型的预报结果都较好,其准确率分别为71.43%、76.19%,均高于70%。这表明,已建立的两两种预报模型对研究区年径流预报有一定的使用价值。

3.2 试报与结果检验

将碧流河水库2001~2011年11年的样本数据集代入2.2中建立的N-Bayes和D-Bayes径流属性判别模型中进行试报,结果见表4。通过式(7)~(9)对初判结果进行评价,结果见表5。

由表4给出了基于两种判别模型2001~2011年碧流河水库年径流的预报结果值,分析可知,N-Bayes对2001、2002年的预报结果分别为偏丰、偏枯与相应年份实际情况比较,均高于实际级别,属于空报。D-Bayes对2002、2003年的预报结果分别为偏枯、平水,均高于实际级别,属于空报。2008年两种判别模型的预报结果均为枯水,低于实际情况,均属漏报。由此可见,两种判别模型分别通过概率和判别函数值的形式,直观的反应预报对象与各级别的从属关系,有益于决策者将其与实际情况有效地结合起来。

表5为N-Bayes、D-Bayes的预报结果的综合效果比较,两种判别模型的整体预报结果都较好,检验期样本的判别准确率均为72.7%,并且比较样本集的漏报率和空报率均可得出同样的结果。此外,尽管训练期样本N-Bayes、D-Bayes的判别准确率分别为71.4%、76.19%,但两种判别模型训练期、检验期的漏报率、空报率相同。由此可说明,两种模型的核心原理相同,所以评价系数表现出相近的分布。

表4 两种预报模型的判别结果
Table4 Discrimination results of N-Bayes and D-Bayes

年份	年径流量	实际属性	N-Bayes 预报后验概率 P					初判属性	D-Bayes 预报判别函数值 Z					初判属性
			丰水	偏丰	平水	偏枯	枯水		丰水	偏丰	平水	偏枯	枯水	
2001	418.6	平水	0.001	0.359	0.259	0.090	0.291	偏丰	1549.9	1556.4	1556.5	1555.0	1555.5	平水
2002	109.3	枯水	0.002	0.008	0.262	0.523	0.205	偏枯	1618.9	1621.2	1625.2	1625.4	1623.8	偏枯
2003	140.3	枯水	0.001	0.058	0.327	0.044	0.570	枯水	1572.8	1578.1	1580.2	1577.8	1579.6	平水
2004	442.5	平水	0.001	0.056	0.532	0.173	0.238	平水	1559.2	1564.6	1567.3	1565.7	1565.4	平水
2005	614.8	平水	0.001	0.084	0.390	0.385	0.139	平水	1635.0	1640.3	1642.3	1641.8	1640.1	平水
2006	421.5	平水	0.000	0.045	0.271	0.318	0.366	枯水	1530.0	1537.5	1539.7	1539.4	1538.9	平水
2007	451.2	平水	0.010	0.001	0.721	0.097	0.172	平水	1604.4	1602.6	1610.0	1607.6	1607.5	平水
2008	386.2	偏枯	0.002	0.010	0.133	0.248	0.607	枯水	1580.2	1582.8	1585.8	1586.0	1586.2	枯水
2009	226.8	偏枯	0.000	0.011	0.101	0.670	0.218	偏枯	1575.8	1582.6	1585.3	1586.8	1585.0	偏枯
2010	701.4	偏丰	0.001	0.594	0.066	0.102	0.237	偏丰	1485.3	1493.1	1491.3	1491.3	1491.5	偏丰
2011	651.2	偏丰	0.001	0.439	0.090	0.051	0.419	偏丰	1543.0	1550.1	1548.9	1548.0	1549.4	偏丰

表5 碧流水库年径流属性初判结果评价系数
Table5 Evaluation result index of the annual runoff of Biliu reservoir forecasting

样本集	N-Bayes 预报效果			D-Bayes 预报效果		
	准确率	漏报率	空报率	准确率	漏报率	空报率
1959~2000	71.42%	14.29%	14.29%	76.19%	11.90%	11.9%
2001~2011	72.7%	9.1%	18.2%	72.7%	9.1%	18.2%

4 初步结论

考虑物理成因与径流序列的关系,联系大气环流的长期演变规律,采用一定方法筛选与年径流相关性高且独立性好的预报因子,通过分析、试报、检验等步骤,以贝叶斯判别分析原理为基础,构建中长期径流预报模型,应用于碧流水库年径流预报中,并结合气象统计方法,引入准确率、漏报率、空报率对预报结果进行了检验。得到以下结论:

(1)在一定的属性级别划分标准下,基于贝叶斯原理的两种判别模型的初判结果准确率均高于70%,具体而言,贝叶斯判别准则(D-Bayes)在中长期径流分级的预报效果优于朴素贝叶斯分类器(N-Bayes)。

(2)朴素贝叶斯分类器(N-Bayes)和贝叶斯判别准则(D-Bayes)的预报结果分别通过概率和判别函数值的形式直观反映径流与各属性级别的从属关系,有益于决策者将预报结果与相关防洪减灾非工程措施有效地结合起来。

(3)D-Bayes各属性级别判别函数值比较相近,干扰性较强,当遇到变量较多、数据资料较多的预报类型仍需要进一步深入考虑。今后,需要用更多流域的资料,来检验本文研究模型的稳定性与可展延性。

参考文献:

[1] Yong-Ying Zhu, Hui-Cheng Zhou (2009). Rough Fuzzy Inference

Model and its Application in Multi-factor Medium and Long-term Hydrological Forecast [J]. J Water Resour Manage 23:493-507, DOI 10.1007/s11269-008-9285-1.

- [2] 周晓阳,张勇传.洪水的分类预测及优化调度[J].水科学进展,1997,(2):27-33. (ZHOU Xiaoyang, ZHANG Yongchuan. Flood Classifying Prediction and Optimal Flood Dispatching [J]. Advances in Water Science 1997,(2):27-33. (in Chinese))
- [3] 王本德,张静.考虑暴雨成因的大伙房水库洪水分类研究[J].水文,2008,(1):15-20.(WANG Bende, ZHANG Jing. Flood Classification of Dahuofang Reservoir Based on Storm Genesis [J]. Journal of China Hydrology, 2008,(1):15-20. (in Chinese))
- [4] 黄嘉佑.气象统计分析与预报方法(第3版)[M].北京:气象出版社,2004.(HUANG Jiayou. Statistical Analysis and Forecast Methods in Meteorology [M]. Beijing: China Meteorological Press, 2004. (in Chinese))
- [5] 胡邦辉,袁野,王学忠.基于贝叶斯分类方法的雷暴预报[J].解放军理工大学学报(自然科学版),2010,(5):578-584. (HU Banghui, YUAN Ye, WANG Xuezhong et al. Thunderstorm Prediction Based on Bayesian Classification Method [J]. Journal of PLA University of Science and Technology (Natural Science Edition), 2010,(5):578-584. (in Chinese))
- [6] 王本德,郭晓亮,周惠成.基于贝叶斯定理的汛限水位动态控制风险分析[J].水力发电学报,2011,(3):34-38. (WANG Bende, GUO Xiaoliang, ZHOU Huicheng, et al. Risk Analysis on the Dynamic Control of Limit Water Level Based on Bayes Theorem [J]. Journal of Hydroelectric Engineering, 2011,(3):34-38. (in Chinese))

Application of Bayesian Discriminant Analysis Theory in Medium- and Long-term Hydrological Forecast

ZHANG Rui^{1,2}, WANG Bende¹, ZHANG Shuanghu², WANG Hao²

(1 School of Hydraulic Engineering, Dalian University of Technology, Dalian 110624, China;

2. China Institute of Water Resources and Hydropower Research, Beijing 100038, China)

Abstract: The objective of this paper is to test the application effect of Bayesian discriminant analysis theory in the annual runoff of Biliu reservoir forecasting. The proposed models are called Naive Bayesian classifier (N-Bayes) and Bayesian Discriminance criterion (D-Bayes), The structure of the paper is as follows: 1) according to the result of clustering analysis and actual conditions, the range of the runoff is divided to five levels that are deficient water, Partial deficient water, medium water, Partial abundant water and abundant water respectively; 2) selecting ten classification predictors as the correlation coefficients method and establishing runoff forecasting models of Biliu reservoir by considering the characteristic and physical genesis of runoff; 3) the forecasting models were experimented and tested by use of the data during 2001 to 2010. The result is compared with the practical attribute category shows that the accuracy of N-Bayes and D-Bayes are more than 0.7, the D-Bayes is superior to the N-Bayes in forecasting efficiency, both the two models can be a reference for reservoir operation management.

Key words: Bayesian discriminant analysis theory; classification predictors; Medium- and long-term forecast; Biliu reservoir