

基于粗集推理的大坝安全监测预报模型研究

雷 鹏,顾冲时

(河海大学水利水电工程学院,江苏 南京 210098)

摘要 针对传统大坝安全监测预报模型存在的受环境量相关性影响较大和网络训练时间长、易收敛到局部最优点等缺点,通过对原始监测信息的粗糙集预处理提取主要影响因素和决策规则集,并通过对规则集的不确定性推理建立了大坝监测的粗集预报模型.与传统方法相比,粗集预报模型不依赖于具体的数学模型,所得到的预报值是一定可信度下的区间值,并且预测值区间与实测值区间基本一致.

关键词 大坝安全监测;粗集理论;不确定性推理;区间预报

中图分类号:TV698.1

文献标识码:A

文章编号:1000-198X(2005)04-0391-04

传统的大坝安全监测预报模型通常需借助于统计模型^[1]和神经网络模型^[2]等.而在实际问题中,统计模型常受到影响量相关性的困扰,神经网络模型则往往存在训练时间长、易收敛到局部最优点等缺点,同时两者所得预报值为一确定量.对于运行多年的大坝来说,时效变形已基本稳定,而效应量主要受环境量变化的影响,即在一定的环境量的作用下,其测值应落在一定的区间内^[3],所以用某可信度下的测值变化区间来描述这种变化则更符合实际.粗糙集^[4,5]理论不需要预先给定数学模型,它通过对数据的区间划分挖掘数据中的隐藏信息,可保证分析过程的客观性.因此,本文基于粗集推理建立了大坝安全监测预报模型.

1 粗集基本原理

1.1 粗集描述和上近似集、下近似集定义^[6]

一个决策表为一个信息知识表达系统 $S, S = \langle U, R, V, f \rangle$. 其中: $U = \{x_1, x_2, \dots, x_n\}$ 是一组对象的有限集合,称为论域; R 是属性的有限集合,它又可分为条件属性集 C 和决策属性集 D 两个不相交的集合,即 $R = C \cup D$, 且 $C \cap D = \emptyset$; V 是一组属性值的值域; f 为一个信息函数, $f: U \times (C \cup D) \rightarrow V$, 它指定 U 中每一个对象 x 的属性值,即对于 $\forall x \in U, a \in C \cup D$, 有 $f(x, a) \in V_a$. 若 $B \subset R$, 且 $B \neq \emptyset$, 则 B 在 U 上的不可分辨关系(或等价关系)为

$$\text{IND}(B) = \bigcap_{b \in B} \text{IND}(\{b\}) = \{(x, y) | (x, y) \in U^2, b \in B, \forall f(x, b) = f(y, b)\} \quad (1)$$

对于 $X \subseteq U$, 根据知识 B , 引入分类正确率因子 $\beta \in (0.5, 1]$, 定义以下广义扩展粗集模型——VPRS 模型:

$$\underline{B}^\beta(X) = \text{POS}_B^\beta(X) = \left\{ x \mid \left(x \in U \wedge \frac{[x]_B \cap X}{[x]_B} \geq \beta \right) \right\} \quad (2)$$

$$\bar{B}^\beta(X) = \text{NEG}_B^\beta(X) = \left\{ x \mid \left(x \in U \wedge \frac{[x]_B \cap X}{[x]_B} \geq 1 - \beta \right) \right\} \quad (3)$$

$$\text{BN}_B^\beta(X) = \text{BOS}_B^\beta(X) = \left\{ x \mid \left(x \in U \wedge 1 - \beta < \frac{[x]_B \cap X}{[x]_B} < \beta \right) \right\} \quad (4)$$

当 $\beta = 1$ 时, 上述扩展粗糙集就转变成标准粗糙集. 由于 β 的引入, 可变精度粗糙集对数据不一致性有更大的容忍度, 并可以增强所产生的规则的鲁棒性, 提高预测精度. 其中 $[x]_B$ 表示 B 对 U 划分的不可分辨关系集合, 是论域 U 的 B 基本集的集合. 根据知识 B , $\underline{B}^\beta(X)$ 是 X 的下近似集, 又称为 X 的正域 $\text{POS}_B^\beta(X)$; $\bar{B}^\beta(X)$ 是 X 的上近似集, 又称为 X 的负域 $\text{NEG}_B^\beta(X)$; $\text{BN}_B^\beta(X)$ 是 X 的 B 边界, 又称为边界域 $\text{BOS}_B^\beta(X)$.

1.2 决策规则

如果决策表 $S = \langle U, C, D, V, f \rangle$, 则对于任意的 $x \in U$, 用 d_x 表示决策规则, 即: $d_x: \text{dex}([x]_C) \Rightarrow \text{dex}([x]_D)$, $d_x(a) = d(x)$, $a \in C \cup D$, 且 $d_x|_C, d_x|_D$ 分别称为 d_x 的条件和决策.

在一个相容知识表达系统 S 中, 对于决策规则 d_x , 有 $[x]_C \subseteq [x]_D$. $\forall r \in C$, 若 $[x]_C \setminus r \not\subseteq [x]_D$, 则 r 是 d_x 的核值属性, 为 d_x 中不可省略的; 若 $[x]_C \setminus r \subseteq [x]_D$, 则 r 不是 d_x 的核值属性, 为 d_x 中可省略的.

2 监测资料的粗集预处理

2.1 监测资料的离散化

大坝监测资料是连续的, 在进行粗集数据分析之前, 应将连续数据离散化. 本文采用基于属性重要性的离散化方法^[7,8]对监测资料进行离散. 其步骤如下:

- 对于每一个可能的荷载属性 $c_j (j = 1, 2, \dots, m)$, 计算其重要性指标 $r_C(D) - r_{C \setminus \{c_j\}}(D)$. 其中 $r_C(D) = |\text{POS}_C(D)|/|U|$ 表示条件属性集合 C 对决策属性集合 D 的依赖程度.
- 根据计算出的 $r_{C \setminus \{c_j\}}(D)$, 对条件属性 c_j 由小到大进行排序. 在属性重要性相同的条件下, 按条件属性断点个数由多到少对条件属性进行排序.
- 对于属性 c_j 中的每个断点 $a_i (j = 1, 2, \dots, m)$, 考虑其存在性. 即把信息表中与 a_i 相邻的两个属性的较小值改为较大值. 如果信息系统不引入冲突, 则将该点去掉; 否则, 把修改过的属性还原.
- 用正整数序列对每个样本 x_i 在影响因素 $c_j (j = 1, 2, \dots, k)$ 下的属性值 a_{ij} 进行编码, 得到如表 1 所示的决策信息表.

表 1 影响因素~效应量决策信息表
Table 1 Decision table of affecting factors and effect

样本	影响因素					效应量
	水位 c_1	温度 c_2	降雨 c_3	...	时效 c_k	D
x_1	a_{11}	a_{12}	a_{13}	...	a_{1k}	d_1
x_2	a_{21}	a_{22}	a_{23}	...	a_{2k}	d_2
x_n	a_{n1}	a_{n2}	a_{n3}	...	a_{nk}	d_n

2.2 决策表的约简

监测资料的粗集约简就是提取不含多余影响因素并保证分类决策正确的最小条件属性集. 这类似于回归计算中因子的剔选, 即去除影响极小的因素而保留影响较大的因素. 决策表的约简包括属性约简和属性值约简.

- 属性约简.** (a) 给定阈值 $0.5 < \beta \leq 1$, 从可能影响因素集 C 中去掉第 i 个影响因素 c_i , 计算 $\text{POS}_{C \setminus \{c_i\}}^{\beta}(D)$; (b) 若属性核 $\text{POS}_{C \setminus \{c_i\}}^{\beta}(D) = \text{POS}_C^{\beta}(D)$, 即去除该因素对效应量无影响 (或影响较小), 则属性 c_i 为冗余属性, 从影响因素集 C 中去除该属性; (c) 若属性核 $\text{POS}_{C \setminus \{c_i\}}^{\beta}(D) \neq \text{POS}_C^{\beta}(D)$, 即去除该因素会影响效应量的变化, 则因素 c_i 为必要属性, 所以在影响因素集 C 中保留该属性; (d) 对所有可能的影响因素作上述判断后, 即可得到约简的主要影响因素集 C' .
 - 属性值约简.** (a) 删除决策表中相同的行 (规则), 得到所有规则集合 d_x ; (b) 对于规则集中的每一条规则 $(d_x)_i$ 中的属性 r_j , 即 $\forall r \in (d_x)_i|C$, 如果 r_j 不是 d_x 的核值属性, 则将 r_j 从该规则中去除; (c) 循环 (b) 步, 直至所有规则中的属性值均被约简.
- 决策表约简的结果是去除了多余的属性和属性值, 得到了“影响因素~效应量”的具有较大适应度的规则集.

3 基于粗集规则推理的预报模型

通过对决策表的约简, 得到了简化的决策规则, 每条决策规则对应一个“影响因素~效应量”的因果关系. 根据规则集, 可建立大坝监测资料预报模型. 其步骤为: (a) 对于监测的环境量值 x_{C_i} , 根据 2.1 中求得的断点进行编码; (b) 利用所求属性值的编码, 根据规则表进行推理^[9], 得到简化的决策规则形式 $\text{dex}(X) \Rightarrow \text{dex}(d)|\mu, \mu$ 表示规则的可信度; (c) 设置阈值 α , 提取 (b) 中得出的决策集 $\text{dex}(d)|\mu (\mu \geq \alpha)$, 得到在多因素作用下的效应量变化区间及其可信度. 由以上过程可以看出, 粗集预报模型实际上是一个定性的区间预报模型, 它得不到具体的预测值, 而只得到效应量的区间值. 这是它与传统的定量预报模型的最大区别. 大坝监测的粗集预报模型流程如图 1 所示.

4 算 例

取某大坝 13 号坝段扬压力孔水位监测资料进行粗集模型预处理,提取数据中隐含的影响因素-效应量的决策规则,建立扬压力孔水位的预报模型,并以 Matlab6.5 为开发平台进行编程计算,得到了以下计算结果.

4.1 监测资料预处理

4.1.1 初始信息表

据 13 号坝段扬压力孔水位 1997~2001 年 5 年的监测资料(共 547 条记录),建立了初始信息表,如表 2 所示(受篇幅限制,表中只列出了部分监测信息).

表 2 初始信息表

Table 2 Initial information table					
日期	上游库水位/m	下游库水位/m	温度/℃	降雨/mm	扬压力孔水位/m
1999-05-04	68.23	47.61	21	13	55.11
1999-05-11	72.21	48.54	20	0	56.17
1999-05-18	67.93	47.65	20	26	54.98
1999-05-25	70.36	49.79	22	6	55.47
1999-06-01	70.35	50.42	23	0	56.04
1999-06-08	71.65	47.97	26	38	55.64
1999-06-15	69.52	48.15	27	0	55.16

经过计算,得断点集如下: $CUT(c_1)=\{67.22,68.45,69.95,71.18,72.53\}$; $CUT(c_2)=\{47.20,47.57,47.98,48.33\}$; $CUT(c_3)=\{15.5,22.12,26.95\}$; $CUT(c_4)=\{27\}$; $CUT(c_5)=\{\}$.由于基于属性离散化的过程也是一个决策属性约简的过程,而 c_5 没有断点,即表示时效的影响极小,是不必要属性,因此被舍弃.于是得到如表 3 所示的属性离散化区间.

表 3 属性离散化区间

编码	条件属性					决策属性: 扬压力孔 水位 D/m
	上游库水位 c_1/m	下游库水位 c_2/m	温度 $c_3/^\circ C$	降雨 c_4/mm	时间 c_5/d	
1	$(-\infty,67.22)$	$(-\infty,47.20)$	$(-\infty,15.5)$	$(-\infty,27)$		$(-\infty,51.51)$
2	$[67.22,68.45)$	$[47.20,47.57)$	$[15.5,22.12)$	$[27,+\infty)$		$[51.51,52.42)$
3	$[68.45,69.95)$	$[47.57,47.98)$	$[22.12,26.95)$			$[52.42,54.13)$
4	$[69.95,71.18)$	$[47.98,48.33)$	$[26.95,+\infty)$			$[54.13,55.83)$
5	$[71.18,72.53)$	$[48.33,+\infty)$				$[55.83,56.74)$
6	$[72.53,+\infty)$					$[56.74,+\infty)$

4.2 基于规则的预报模型

通过对主要影响因素(水位、温度、降雨)组合的规则约简,得到了约简的规则集.为了对 2002 年测值进行区间预报,首先应用 GM(1,1)灰色模型^[10]延拓环境量,求得 2002 年的环境量值,然后利用表 3 中的区间划分环境量区间,求得相应的条件属性值.设置可信度阈值 $\alpha=0.6$,提取大于阈值 α 的预测区间值,得到相应的预测值.

2002 年扬压力预报结果如表 4 所示(受篇幅限制,表 4 只列出了部分预报值),实测值区间与预测值区间对比如图 2 所示.从表 4 和图 2 可以看出,按最大置信度原则,2002 年 1 月 15 日、2 月 16 日、3 月 18 日、4 月 10 日、5 月 20 日的预测值分别为 4.5544.取 2002 年每月典型值计算,除 4 月外,其他月份预测值区间与实测值区间完全一致,其预测正确率达到 91%.这说明,该模型预测能力较强.

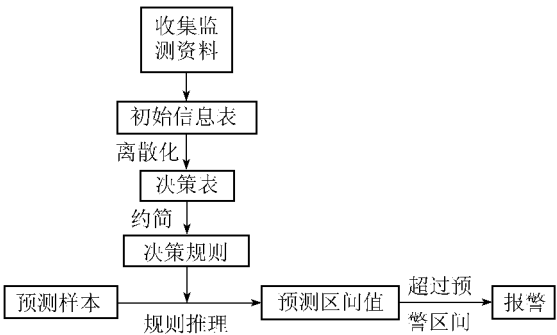


图 1 大坝安全监测粗集预报模型流程

Fig.1 Flow chart for rough set prediction model

4.1.2 区间离散化

a. 离散化决策属性(即扬压力孔水位)值.本文用数理统计方法进行计算.根据 $p_k=P(X\leq X_k)$,假定 $p_k=0.05,0.2,0.5,0.8,0.95$,得到断点集 $CUT(D)=\{51.51,52.42,54.13,55.83,56.74\}$,对孔水位测值按表 3 进行编码.当然,决策属性的划分还可采用其他方法.比如,可以选择其他概率划分点进行离散,或直接根据经验选择感兴趣的区间进行离散.

b. 采用基于属性重要性的离散化方法,对上游水位、下游水位、气温、降雨和时间等变量进行离散.

表 4 2002 年扬压力孔水位预报结果

Table 4 Predicted result of uplift pressure in 2002

日期 (月-日)	条件属性								决策属性 <i>D</i>		
	<i>c</i> ₁		<i>c</i> ₂		<i>c</i> ₃		<i>c</i> ₄		实测值/m	预测区间值	
	上游水位/m	编码	下游水位/m	编码	温度/℃	编码	降雨/mm	编码		决策 1	决策 2
01-15	66.74	1	46.98	1	17.5	2	0	1	54.25	3 0.67	4 0.95
02-16	70.63	4	47.76	3	14.0	1	0	1	56.44	4 0.70	5 1.00
03-18	70.33	4	47.59	3	16.4	2	35	2	56.03	4 0.70	5 0.90
04-10	68.95	3	47.01	1	16.7	2	9	1	55.88	4 0.83	5 0.80
05-20	66.37	4	47.27	2	23.2	3	9	1	54.86	4 1.00	

5 结 论

- a. 粗集预报模型不依赖于具体的数学模型,它通过提取数据之间的隐含规则进行测值预报,其分析更具客观性.
- b. 粗集预报模型是一个定性的区间预报模型,它得不到具体的预测值,而只得到效应量在某可信度下的区间值.这是该模型与传统定量预报模型的最大区别.
- c. 本文效应量区间是按概率进行划分的,而概率的选择具有一定的主观性.如何根据实际应用的需要选择合适的概率,还需进一步研究.

参考文献:

[1] 吴中如,阮焕祥,沈长松. 大坝安全监控理论及其应用[M]. 北京: 中国教育出版社, 2002. 44—73.

[2] 苏怀智, 吴中如, 温志萍. 遗传算法在大坝安全监控神经网络预报模型建立中的应用[J]. 水利学报, 2001 (8) :44—48.

[3] 张乾飞, 王锦国, 李雪红. 大坝安全监控中的不确定性信息初探[J]. 河海大学学报(自然科学版) 2002, 30(5) :113—117.

[4] PAWLAK Z, BUSSE J G. Rough set[J]. Communications on the ACM, 1995, 38(11) :89—95.

[5] PAWLAK Z. Rough set theory and its applications to data analysis[J]. Cybernetics and Syst, 1998, 29(7) :661—688.

[6] 王国胤. 粗糙集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001. 23—38.

[7] HU X. Knowledge discovery in databases: an attribute-oriented rough set approach[D]. Regina: University of Regina, 1995.

[8] 胡可云, 陆玉昌, 石纯一. 粗糙集理论及其应用进展[J]. 清华大学学报(自然科学版) 2001, 41(1) :64—68.

[9] 刘清. 粗糙集及 Rough 推理[M]. 北京: 科学出版社, 2003. 80—99.

[10] 刘思峰, 郭天榜, 党耀国. 灰色系统理论及其应用[M]. 北京: 科学出版社, 1999. 102—122.

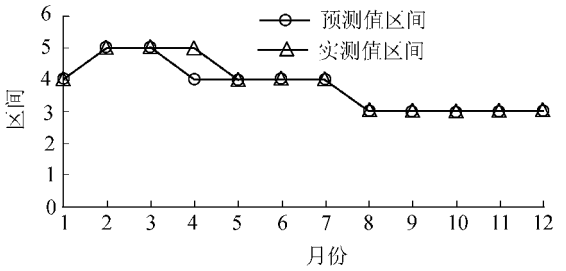


图 2 2002 年预测值区间与实测值区间对比

Fig.2 Comparison of intervals of predicted result and measured data for 2002

Prediction model for dam safety monitoring based on rough set reasoning

LEI Peng , GU Chong-shi

(College of Water Conservancy and Hydropower Engineering , Hohai University , Nanjing 210098 , China)

Abstract :In consideration of the fact that the conventional prediction model for dam safety monitoring is easily affected by the correlation of environmental factors , and the fact that it needs a long time of network training and easily converges to local optimum points , a rough set prediction model is developed by extraction of main affecting factors and decision rules from pretreatment of rough sets of original monitoring data and reasoning of decision rules under uncertainty. Compared with the conventional method , the present model is independent on a special mathematical model , and the predicted results are intervals under certain reliability , which are in accordance with measured data.

Key words :dam safety monitoring ; rough set theory ; reasoning under uncertainty ; interval prediction