

基于投影寻踪动态聚类法的湖泊水质评价模型

李大鹏¹ ,段沛霞¹ ,倪长健^{1 2}

(1. 成都信息工程学院环境工程系 ,四川 成都 610225 ; 2. 中国科学院成都山地灾害与环境研究所 ,四川 成都 610041)

摘要 将投影寻踪动态聚类法应用到湖泊水质综合评价中 ,并与现有的水质综合评价模型的评价结果做了对比。结果表明 ,投影寻踪聚类法能有效地克服现有模型的缺点 ,是高维数据样本聚类分析的一种有效的统计方法 ,在整个运算过程中不需人为给定参数 ,聚类结果合理、客观 ,具有稳定性好、分类结果明确、操作简便等特点 ,为多因素聚类分析提供了一种新方法。

关键词 投影寻踪 ,动态聚类 ,水质综合评价

中图分类号 :X524 文献标识码 :A 文章编号 :1004-693X(2007)S2-0085-03

水质综合评价 ,就是根据某些水质指标值 ,通过所建立的数学模型 ,对各水体样本的水质等级进行综合评判。由于实际水体中各单项水质指标的评价结果往往是不相容的 ,直接利用水质评价标准表进行水体质量评价缺乏实用性 ,因此 ,相继提出了灰色聚类法、模糊综合评价法、(模糊)神经网络法等多种方法^[1]。这些模型的计算结果都是一些离散的水质等级 ,是半定量化的 ,等级分辨率太粗。而实际水体中各水质指标值一般是连续的实数值 ,也就是说 ,按目前常用的水质评价方法 ,即使属于同一等级的水体 ,它们对应的各水质指标值常常相差显著。为此 ,文献[2]提出了基于遗传算法的投影寻踪等级水质评价模型 ,除 PP 法^[3]本身存在的不足外 ,该方法还存在建模的复杂性、模型构造的合理性以及数据处理等值得进一步探讨的问题^[4]。本文利用投影寻踪动态聚类方法^[5-6]对这一问题进行了探讨 ,并进行了实例研究。

1 投影寻踪动态聚类模型

设第 i 个样本的第 j 个因素为 x_{ij}^0 ($i = 1, \dots, n$; $j = 1, \dots, m$; n 为样本个数 , m 为因素个数) ,投影寻踪动态聚类模型的实现步骤如下 :

步骤 1 :预处理。由于各因素数值范围可能相差较大 ,为了消除极端异常值对聚类结果造成的影响 ,找出满足下式的样本个体 ,并将其从样本集中加以剔除。

$$x_{ij}^0 > 10\bar{x}_j \quad (j = 1, 2, \dots, m) \quad (1)$$

另外 ,为了满足尺度不变性 ,采用式(2)对剔除后的评价指标值进行极差归一化处理 :

$$x_{ij} = [\frac{x_{ij}^0 - x_{j_{\min}}}{x_{j_{\max}} - x_{j_{\min}}}] \quad (2)$$

式中 $\bar{x}_j$, $x_{j_{\max}}$ 和 $x_{j_{\min}}$ 分别为样本中第 j 个指标值的平均值、最大值和最小值。

步骤 2 :线性投影。线性投影实际就是给出一投影方向 ,把多因素复杂问题转化为一维投影特征值序列 ,将高维数据投影到线性空间进行研究 ,设 a 为 m 维单位向量即 $a = (a_1, a_2, \dots, a_m)$,则 x_{ij} 的投影特征值

$$z_i = \sum_{j=1}^m a_j x_{ij} \quad (3)$$

步骤 3 :构造目标函数。对于任一投影方向 ,基于某一聚类准则的样本分类结果是确定的 ,记整个样本的投影特征值序列组成的集合为 $\Omega = \{ z_1, z_2, \dots, z_n \}$,要将它们分成 K 类 ,采用动态聚类法^[7] :

- a. 随机选取 K 个点作为 K 个聚核 ,记为 $L^0 = (A_1^0, A_2^0, \dots, A_K^0)$
- b. 根据 L^0 把 Ω 中的点分为 K 类 ,记为 $P^0 = (P_1^0, P_2^0, \dots, P_K^0)$
 $P_i^0 = \{ z \in \Omega \mid d(A_i^0 - z) \leq d(A_j^0 - z) , \forall j = 1, 2, \dots, K, j \neq i \}$
 $d(A_i^0 - z)$ 为点 A_i^0 和集合 Ω 中任一点的绝对值距离。
- c. 由 P^0 出发 ,计算新的聚核 L^1 ,
 $L^1 = (A_1^1, A_2^1, \dots, A_K^1)$

$$A_i^1 = \frac{1}{n_i} \sum_{z_i \in P_i^0} z_i$$

式中 n_i 为 p_i^0 中的样本点数目。

d. 重复以上步骤 ,由此得到一个分类结果序列

$$V^m = (L^m ,P^m) \quad m = 1 \ 2 \ \cdots$$

$$\text{记} \quad D(A_i^m ,P_i^m) = \sum_{z_l \in P_i^m} |z_l - A_i^m |$$

$$u_m = \sum_{i=1}^K D(A_i^m ,P_i^m)$$

则算法的终止判断条件是

$$\frac{|u_{m+1} - u_m|}{u_{m+1}} \leq \varepsilon$$

式中 ε 是充分小的误差允许值 ,理论证明这种算法是收敛的^[6]。定义 $d(z_i ,z_j)$ 为任意两投影特征值间的绝对距离。

类内样本的聚集度可表示为

$$d(a) = \sum_{i=1}^k \sum_{\substack{z_m \in P_i \\ z_j \in P_i}} d(z_m ,z_j)$$

$d(a)$ 愈小 ,则相似样本的聚集程度越高。投影分散度记为

$$s(a) = \sum_{\substack{z_m \in \Omega \\ z_j \in \Omega}} d(z_m ,z_j)$$

其值愈大 ,则整个样本投影特征值离散程度越高。根据类内聚集度与投影分散度的定义可将目标函数 $Q(a)$ 定义为投影分散度与类内聚集度之差 ,即

$$Q(a) = s(a) - d(a) \quad (4)$$

对此目标函数的求解就是寻求一投影方向满足 $s(a) - d(a)$ 取得最大值 ,显然 ,投影分散度越大或类内聚集度越小 ,则目标函数越大 ,这正是投影寻踪建模思想的体现。

步骤 4 :优化投影方向。模型建立的关键是找到能反映系统特征的最优投影方向 ,根据上述分析可知 ,当式 (4) 取得最大值时所对应的 a 就是最优投影方向向量。所以 ,此问题可转化为式 (5) 描述的优化问题。

$$\begin{cases} \max Q(a) \\ \|a\| = 1 \end{cases} \quad (5)$$

免疫进化算法^[8]可应用于解决上述优化问题。

2 实 例

所选取水质评价指标的标准见表 1^[9] ,在此表中各级水质等级取值范围内均匀随机产生 5 个水质等级样本 ,共同组成样本系列 ,见表 2。

将上述 25 个样本聚为 5 类 ,通过运算得最大投影指标函数值为 151.7952 ,对应的最佳投影方向

表 1 湖泊水质评价标准

级 别	$\rho(\text{TP})/$ ($\mu\text{g}\cdot\text{L}^{-1}$)	耗氧量/ ($\text{mg}\cdot\text{L}^{-1}$)	透明度/ m	$\rho(\text{TN})/$ ($\text{mg}\cdot\text{L}^{-1}$)
极贫营养(1)	<1	<0.09	>37.00	<0.02
贫营养(2)	4	0.36	12.00	0.06
中营养(3)	23	1.80	2.40	0.31
富营养(4)	110	7.10	0.55	1.20
极富营养(5)	660	27.10	0.17	4.60

表 2 湖泊水样对应的等级及 PPDC 方法计算的特征值

水样 序号	$\rho(\text{TP})/$ ($\mu\text{g}\cdot\text{L}^{-1}$)	耗氧量/ ($\text{mg}\cdot\text{L}^{-1}$)	透明 度/m	$\rho(\text{TN})/$ ($\text{mg}\cdot\text{L}^{-1}$)	水质 等级	投影 特征值
	$x(1,i)$	$x(2,i)$	$x(3,i)$	$x(4,i)$		
1	0.658	0.059	50.857	0.001	1	0.1516
2	0.559	0.051	42.650	0.015	1	0.2243
3	0.106	0.034	67.902	0.009	1	0.0008
4	0.978	0.061	42.207	0.010	1	0.2280
5	0.646	0.051	47.659	0.015	1	0.1796
6	1.233	0.259	36.899	0.040	2	0.2788
7	3.225	0.349	24.079	0.047	2	0.3957
8	2.821	0.133	23.709	0.042	2	0.3943
9	1.426	0.176	16.260	0.020	2	0.4586
10	1.947	0.098	19.699	0.035	2	0.4276
11	15.338	0.918	3.359	0.185	3	0.6044
12	22.127	1.779	4.942	0.240	3	0.6159
13	20.269	1.681	2.606	0.281	3	0.6321
14	9.171	1.479	3.252	0.145	3	0.6079
15	16.739	1.308	8.979	0.290	3	0.5644
16	24.736	2.621	1.530	1.182	4	0.7444
17	91.674	3.710	2.106	0.627	4	0.7693
18	105.427	6.403	1.176	0.661	4	0.8470
19	106.776	5.898	0.993	1.145	4	0.8407
20	93.002	5.341	1.038	0.606	4	0.8116
21	627.669	20.039	0.409	2.627	5	1.8023
22	297.816	13.171	0.191	3.050	5	1.2384
23	370.796	21.733	0.460	3.068	5	1.4956
24	279.420	17.600	0.347	3.850	5	1.2973
25	281.502	22.646	0.317	1.464	5	1.3966

$a'=(0.4196,0.5172,0.5947,0.4503)$ 。由此得到各样本的投影特征值(表 2);同时得到相应的湖泊水质等级综合评价的分级标准(表 3)。投影特征值 $z'(i)$ 越大 ,表示湖泊水质越差。

表 3 基于 PPDC 方法的湖泊水质评价模型的分级标准

投影特征值	水质等级	投影特征值	水质等级
[0,0.2756)	1	[0.6544,0.9143)	4
[0.2756,0.5061)	2	[0.9143,1.9818)	5
[0.5061,0.6544)	3		

针对文献[2]提供的 5 个湖泊水质的实测数据 ,对各湖泊水质富营养化等级进行评判 ,结果见表 4 ,表中同时也列出了模糊神经网络方法^[9]、AGA 算法^[10]和 IEA 算法^[11]的评价结果。根据表 3 和表 4 ,青海湖的水质评定等级为 3 级 ,西湖的水质评定为 5 级 ,此结果与文献[2]的结论一致。东湖的总磷指标值已接近 4 级(富营养)水质标准的上限 ,其余指

表 4 基于 PPDC 的水质评价模型的实例分析

湖泊	$\rho(\text{TP})/(\mu\text{g}\cdot\text{L}^{-1})$	耗氧量/ ($\text{gm}\cdot\text{L}^{-1}$)	透明度/m	$\rho(\text{TN})/(\text{mg}\cdot\text{L}^{-1})$	PPDC 方法	神经网络 方法特征值	AGA 算法 特征值	IEA 算法 特征值
杭州西湖	130	10.30	0.35	2.76	1.1425	5	4.92	4.85
武汉东湖	105	10.70	0.40	2.00	1.0594	4	4.80	4.72
青海湖	20	1.40	4.50	0.22	0.6177	3	3.09	2.99
巢湖	30	6.26	0.25	1.67	0.8960	4	4.59	4.41
滇池	20	10.13	0.50	0.23	0.8204	4	3.67	3.81

标值都在 5 级(极富营养)水质范围内,评定结果为 5 级,而模糊神经网络法认为其水质等级为 4 级。对于巢湖和滇池,它们总磷指标值分别落在 4 级和 3 级水质范围内,耗氧量指标分别落在 4 级和 5 级范围内、且都与 4 级的上限值相近,透明度指标值都落在 5 级水质范围内,巢湖接近 5 级上限而滇池接近 4 级上限值,总氮指标值分别落在 5 级和 3 级水质范围内,评定结果为 4 级,此与文献 2 的结果一致,从两者的投影特征值来看,巢湖的水质比滇池的水质要差些。

最佳投影方向各分量绝对值的大小实质上反映了各水质指标对水质等级的影响程度,各分量绝对值越大则对应的水质指标对水质等级的影响程度就越大,由最佳投影方向 $a'=(0.4196,0.5172,0.5947,0.4503)$ 可见,透明度、耗氧量、总氮、总磷对水质等级的影响程度依次减小,这与水污染管理的经验相一致,这反过来验证了水质评价标准是合理的。

3 结 语

由上可见,基于 PPDC 方法的湖泊环境质量评价模型能有效地克服现有模型的缺点,从应用的角度来看是可行的,该模型具有简单、客观、通用等特点,有着广阔的应用前景。

(上接第 70 页)

4 结 语

阜阳港在运营阶段通过落实完善的事故防范措施,可以将溢油风险事故发生率控制在很低水平,乃至杜绝事故发生。一旦发生装卸溢油事故,虽然将对下游水环境构成明显影响,但通过紧急启动事故应急预案,可以减缓并最终消除对水环境的污染影响,保证下游水体正常的水环境功能。

参考文献:

[1] 金海明.防止船舶溢油污染宁波港的应对措施[J].中国

参考文献:

[1] 金菊良.水资源系统工程[M].成都:四川科学技术出版社,2002.
[2] 金菊良,丁晶.水质综合评价的投影寻踪模型[J].环境科学学报,2001,21(4):431-434.
[3] FRIEDMAN J H ,TURKEY J W. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Trans on Computer , 1974,23(9):881-890.
[4] 赵小勇,付强,刑贞相,等.投影寻踪模型的改进及其在生态农业建设综合评价中的应用[J].农业工程学报,2006,22(5):222-225.
[5] 王顺久,张欣莉,丁晶,等.投影寻踪聚类模型及其应用[J].长江科学院院报,2002,19(6):53-55.
[6] 倪长健,崔鹏.区域泥石流危险度评价的投影寻踪动态聚类方法[J].山地学报,2006,24(4):442-445.
[7] 任若恩,王惠文.多元统计数据分析——理论、方法、实例[M].北京:国防工业出版社,2000:148-151.
[8] 倪长健.免疫进化算法[J].西南交通大学学报,2003,38(1):87-91.
[9] 胡明星,郭达志.湖泊水质富营养化评价的模型神经网络方法[J].环境科学研究,1998,11(4):40-42.
[10] 金菊良,丁晶.遗传算法及其在水科学中的应用[M].成都:四川大学出版社,2000:118-122.
[11] 苏鹏程,倪长健.基于免疫净化算法的逻辑斯谛曲线水环境质量综合评价模型[J].山地学报,2004,22(4):439-444.

(收稿日期:2007-04-06 编辑:高渭文)

水运,2006,24(4):52-53.

[2] 杨柳俊.阳鸿石油码头水环境风险事故影响评价[J].水资源保护,2005,21(5):46-48.
[3] 李永华,刘华祥,李帅.化学品仓储项目水环境风险评价及防范[J].广东化工,2005,32(7):95-96.
[4] 王泽申.安全分析和事故预测[M].北京:北京经济学院出版社,1990:112-114.
[5] 毛文永.环境影响评价技术方法[M].北京:中国环境科学出版社,2006:263-265.

(收稿日期:2007-11-30 编辑:徐娟)