松花江水质评价的模式识别方法

高洪琢　　许志义
（中国科学院长春应用化学研究所）

利用模式识别技术的逐步判别法与非线性映射法（NLM），对松花江流域某些地点的水质进行了评价，得到的结果与模糊数学方法基本上完全一致．从而，为水质评价提供了新的方法．多年来，人们提出各种环境污染评价方法，各具优缺点．一般常用综合污染指标法，此法简便，但结果粗糙．近年来有人使用模糊数学方法[1]，对样本进行聚类分析，使分类趋于准确．我们试用了模式识别方法，效果很好，不但解决了分类问题，还能得到更多有用的信息．计算是在 CROMEMCO CS-2 计算机上用 FORTRAN 语言进行的．

一、实验样本选择

中国科学院长春应用化学研究所与长春地理研究所于 1984 年在松花江沿岸 26 个地点进行水质采样与分析，得到了各地点的物理值、某些无机元素和有机成份含量，共计 13 种．其中有 pH，DO，C_6H_5Cl，$C_6H_4NO_2$，Cd，总 Cr，Cu，As，CN^-，苯酚，苯胺，COD，氨态氮．在模式识别技术中，26 个地点被看作是 26 个样本，每个地点的 13 个数据被看作是 13 个特征．这相当于 13 维空间中 26 个点组成的一个点集，表示为：

$$X_k = (x_{k1}, x_{k2}, \cdots, x_{km})^T \quad k = 1,2,3,\cdots,n \tag{1}$$

其中，k 表示样本号，n 表示样本总数，m 表示特征总数，T 表示矩阵转置．

二、方　法

采用两种方法：逐步判别法与非线性映射法．

1．逐步判别法[2]

在逐步判别过程中，根据各个特征对分类的贡献大小，得到各特征的加权值，由此选取主要特征，得到判别函数．具体做法是，考虑全部特征建立组内与总离差矩阵，计算各个特征的判别能力，将最大的一个引进判别函数．重复计算，不断引进新的特征，同时剔除变为不重要的特征，直到再没有新的特征引入与剔除为止．得到判别函数如下：

$$y_g(X_k) = \ln q_g + C_{0g} + \sum_{j=1}^{m} C_{jg} X_{kj} \tag{2}$$

其中 $j = 1,2,\cdots,m$

$g = 1,2,\cdots,G$

$$y_i^*(X_k) = \max\{y_g(X_k)\} \tag{3}$$

这里 g 是分类序号，G 是分类总数，q_g 是先验概率，$Y_g(X_k)$ 是样本 X_k 的各类计算结果．若其中最大的一个是 $y_g^*(X_k)$，则此样本属于 g 类．以上就是逐步判别法的分类过程，是基于 Bayes 准则的方法．

应用此法，每个样本必须有可供参考的分类数据．样本特征数可以很多，不能遗漏主要的，而且不必考虑相互之间的相关关系．应用此法可以得到下面几个结果：

（1）各个特征的加权因子．体现了各个特征对分类的影响大小，这对实际问题是很有意义的．

（2）判别函数．可用于对未知样本进行分类，体现了计算机学习人的经验的功能．

（3）实验样本的再分类与它们的后验概率．可对参考分类数据与逐步判别法分类结果进行对比，后验概率则表示两种方法分类的符合程度．

因为此法能选取重要特征，相当特征提取作用，所以可作为其它模式识别方法的数据预处理．

2．非线性映射法（NLM）[3,4]

将高维空间中点分布状况，在尽可能保持各点间距离不变情况下，映射到二维空间中．这样，人可以清楚地"看到"各点分布情况，便于分类处理．具体做法是利用逐步判别法得到的结果，将高维空间数据进行降维与加权处理，得到新的点集．然后用雅可比方法进行主成份分析，把点集映射到由最大和次大特征值对应的两个单位特征向量张成的二维平面上．并且以此为初值，逐步优化，以使误差函数

$$E = \frac{1}{\sum_{k<i} d_{ki}^*} \sum_{k<i}^{n} \frac{(d_{ki}^* - d_{ki})}{d_{ki}^*} \tag{4}$$

最小．这里 d_{ki}^*，d_{ki} 分别是高维空间与二维空间中第 k 点与第 i 点之间的距离．此法优点是可以看到每个样本的具体位置，即在本类中是偏上还是偏下，进而得到更多的信息．但是程序复杂，用到的数学方法

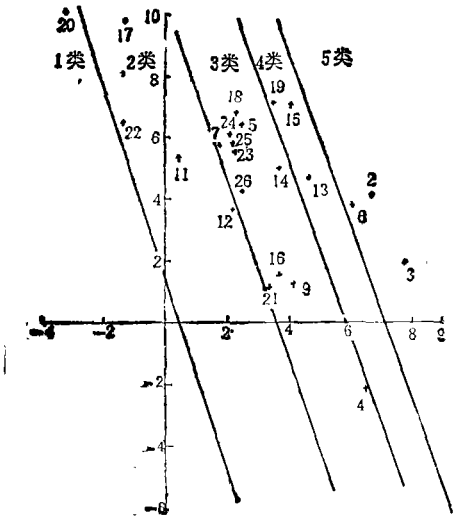

图 1　松花江水质的非线性映射

较多．使用此法之前必须先作特征选择，否则特征数太多，效果反而不好．

值得说明的是，某些污染数据严重超标，造成各数据间差别很大，使分类困难．实验中对超标数据采用指数压缩方法，较好地解决了这个问题．

三、结果与讨论

1．逐步判别法

（1）选取的主要特征与加权因子示于表 1．由表可知，氨态氮是造成污染的最主要因素，如果对其稍加控制，就可以使污染明显改善．其次是苯酚，其它因素作用较小．

（2）分类结果与模糊聚类方法对照示于表 2．由表 2 可知两种分类方法结果 100% 符合，后验概率除一点为 64.5% 以外，其它都在 92.8% 以上．所以，逐步判别法是完全可以信赖的．

（3）判别系数示于表 3．将判别系数代入公式（2），形成判别函数．与上述 26 个样测试条件相同

表 1　选中的主要特征及其加权因子

特　征	pH	总 Cr	As	苯　酚	COD	氨态氮
加权因子	0.037	0.083	0.035	0.244	0.134	1.00

表 2　模糊聚类方法与逐步判别方法分类结果对照表

样本号	模糊聚类方法	逐步判别方法	后验概率	样本号	模糊聚类方法	逐步判别方法	后验概率
1	5	5	1.000	14	3	3	0.645
2	5	5	0.991	15	4	4	0.999
3	5	5	1.000	16	4	4	0.928
4	4	4	1.000	17	2	2	0.999
5	3	3	1.000	18	3	3	1.000
6	5	5	0.986	19	4	4	1.000
7	3	3	1.000	20	1	1	0.958
8	2	2	1.000	21	3	3	0.999
9	3	3	0.999	22	2	2	1.000
10	1	1	1.000	23	3	3	1.000
11	2	2	1.000	24	3	3	0.994
12	2	2	0.992	25	3	3	0.999
13	4	4	0.999	26	3	3	1.000

表 3　判　别　系　数

g	q_g	c_{0g}	c_{1g}	c_{6g}	c_{8g}	c_{10g}	c_{12g}
1	−2.5649	−199.0693	−8.1143	29.1261	−89.5117	−102.379[illegible]	59.0955
2	−1.6487	−138.7579	−11.2608	22.8540	−73.4513	−104.701[illegible]	56.1329
3	−0.9555	−77.8007	−9.7780	18.6559	−52.6520	−78.1773	41.3521
4	−1.6487	−39.8406	−1.8538	11.2319	−39.9416	−52.9395	28.7530
5	−1.8718	−9.4207	2.1659	4.6881	−18.9634	−22.3742	12.3460

表 4　模式识别方法与模糊聚类方法对松花江流域 26 个地点污染分类结果对照表

污染分类	模糊聚类方法	模式识别方法	
		逐步判别法	非线性映射法
1	饮马河口，桦林大桥	饮马河口，桦林大桥	饮马河口，桦林大桥
2	哨口，新民，肇源，白石，汤旺河口	哨口，新民，肇源，白石，汤旺河口	哨口，新民，肇源，白石，汤旺河口
3	大安，哈达湾，红旗，四方台，依兰，牡丹江口，大来，桦川，富锦，同江	大安，哈达湾，红旗，四方台，依兰，牡丹江口，大来，桦川，富锦，同江	大安，哈达湾，红旗，四方台，依兰，牡丹江口，大来，桦川，富锦，同江，呼兰河口
4	月亮泡，拉林河口，哈下大桥下，宁安大桥，呼兰河口	月亮泡，拉林河口，哈下大桥下，宁安大桥，呼兰河口	月亮泡，拉林河口，哈下大桥下，宁安大桥
5	刘园，大民，迎风急，松花湖	刘园，大民，迎风急，松花湖	刘园，大民，迎风急，松花湖

的未知样本，代入判别函数计算，可按公式(3)方法分类.

2.非线性映射法

非线性映射见图 1. 在图上进行聚类处理，画出 4 条判别线. 分类结果与模糊聚类方法比较，除 16号样本在相邻两类间错分之外，其它完全一致，符合率达到 96.2%. 在图上可以看到一个规律，由右上角往左下角，是污染由轻到重的方向. 在图上，同类中的各点也能看出污染程度大小，如第 2 类中 22 号比 12 号污染严重，所以能得到较多的有用信息.

最后，模式识别的两种方法与模糊聚类方法对 26 个地点的污染分类结果示于表 4.

综上所述，可以认为模式识别技术是进行水质评价的一种有效方法. 它与其它方法相比，有很多特色. 当然，也可以用于其它方面的评价.

参 考 文 献

[1] 于连生，中国环境科学，6，33(1982).

[2] 中国科学院计算中心概率统计组，概率统计计算，196 页，科学出版社，北京，1983 年.

[3] Kowalski, B.R. and Bender, C. F., *J. A. C.S.*, **94**, 5632(1972).

[4] 江乃雄等，分子科学学报，1(1)，115(1981).

底栖动物在南洞庭湖岸边污染带水质评价中的应用*

邵　国　生

(湖南省洞庭湖环境监测站)

南洞庭湖是洞庭湖目前存水面积较大的湖泊，现有水面 917km²，它主要受纳澧水、沅水及长江水的一部分，然后注入长江. 近年来，由于沿岸造纸、食品、麻纺等乡镇企业的高速发展，大量有机污水排入沿岸水域，如此长期持续排污必使其水生态系统受到严重破坏，直接影响到洞庭湖水体的综合开发利用.

根据水体大型底栖无脊椎动物的特点，利用其群落结构的变化评价水质污染具有较好的效果，能够比较客观地反映污染现状[1,2,3]. 本站 1981—1985 年先后对南洞庭湖的东南洞、万子湖等几个子湖的断面进行了底栖动物调查，并结合理化监测资料从生态学角度进行了水质评价，为制订区域性废水排放标准提供科学依据.

一、工 作 方 法

本文调查的南洞庭湖主要污染水域位于湖南沅江县塞南湖至沈家湾右岸(以下称沅江岸边带). 该水域全长约 10km，采样点平均水深 2.5m，水流速 0.32m/s. 沅江岸边带地形和采样点的分布见图 1.

* 参加工作的还有刘靖涛、刘齐德、黄止其、张建波和李利强等同志.