Vol. 14　　高等学校化学学报　　No. 5
1993年5月　　CHEMICAL JOURNAL OF CHINESE UNIVERSITIES　　621～624

多变量判别分析用于癌症诊断研究*

朱尔一　王小如　邓志威　杨芃原　黄本立
(厦门大学化学系，厦门，361005)

摘要　用感应耦合等离子体原子发射光谱及石墨炉原子吸收光谱仪测定了正常人及癌症病人头发样品中15种元素的含量. 所得数据用多元多项式扩展增维和逐步回归变量压缩技术以及PLS方法处理，得到了病人与正常人分类极为清晰的二维判别图. 据此可将头发用作癌症临床诊断中的分析样品以取代血液样品.

关键词　ICP-AES，PLS，多元多项式展开，逐步回归，癌症诊断

癌症的临床诊断方法之一是采集病人血样进行分析. 血样的采集很容易引起交叉感染，不宜重复测定，血液的存储及运输也比较困难. 因此寻找一种取代的血液分析临床诊断方法的研究具有很重要的意义.

本文提出仅采集病人头发样品，经过化学处理后，用原子光谱分析技术对样品中15种微量元素同时测定，然后用PLS[1,2]等多变量分析方法对所得数据进行分类处理，以多元数据的综合信息为依据，建立癌症初级诊断模型，从而用于癌症的临床诊断.

实　验　部　分

1　仪器与试剂

PS-4型多道感应耦合等离子体原子发射光谱仪(美国Baird公司)，PE 3030B石墨炉原子吸收光谱仪(美国PE公司)，HP386计算机(美国HP公司). 用法国Rabit双道蠕动泵及自制手动六通阀将微量头发样品以流动注射方式引入等离子体炬中. 元素标准溶液均由1000 ng/g的储存溶液稀释制备. 硝酸、高氯酸均为优级纯.

2　样品制备

头发样品由厦门大学抗癌中心提供. 将发样剪成小于5 mm的碎段，用流水及去离子水洗涤后，加入一定量丙酮，摇荡5 min，弃去丙酮，再用去离子水漂洗，然后用5%洗洁精洗涤三遍，每次之间均用去离子水漂洗到无泡沫. 样品置于90℃烘箱中，称取干燥后样品0.3 g，加HNO_3+$HClO_4$混合酸(5∶1)2 mL，缓慢加热消化样品至溶液澄清，移入25 mL容量瓶定容. 待测样品体积在0.3～0.5 mL左右，且微量元素浓度较低，因此采用了流动注射方法，固定样品环体积为0.2 mL，用5%HNO_3载液通过六通阀直接载入等离子体炬中.

3　原子光谱分析头发样品

用ICP-AES多道光谱仪分析样品中14种微量元素的含量，采用本实验室发展的瞬间信

收稿日期：1992-12-02. 修改稿收到日期：1993-02-15. 联系人：朱尔一.
*国家自然科学基金、国家教委归国留学生启动基金及国家人事部博士后基金资助课题.

号采集软件收集测定，样品中的Se含量采用石墨炉原子吸收法直接测定，共分析样品120多个，典型结果见表1.

Table 1 Spectroscopy anlaysis of trace elements in human hair(μg/g)

No.	Element	Normal people average value	Cancer patient average value	Total average value	Mean square deviation
1	Zn	266.7	179.7	238.8	109.6
2	Pb	8.44	10.52	8.95	10.89
3	Ba	3.04	2.96	3.01	2.54
4	Ni	1.14	1.00	1.09	1.20
5	Co	0.36	0.35	0.36	0.38
6	Cd	0.50	0.39	0.45	0.40
7	Mn	1.14	1.11	1.13	1.03
8	Cr	0.37	0.92	0.54	0.55
9	Mg	82.6	59.8	75.2	45.61
10	V	0.86	0.63	0.79	0.98
11	Al	8.94	8.30	8.74	10.39
12	Ca	931.8	609.6	828.5	607.8
13	Cu	12.57	9.63	11.62	3.96
14	Fe	5.464	1.294	0.759	0.80
15	Se	0.159	0.420	0.243	0.179

多变量判别分析结果与讨论

本工作处理的问题为两类样本模式识别分类问题，其中一类为癌症病人样本而另一类为正常人样本. 通常可用最优判别平面法[3]处理，但也可用回归法处理，因为若将分类信息代入目标变量，用多元线性回归法求得的回归系数与最优判别平面法求得的第一判别矢量方向是相同的[4]. 本工作用PLS法及逐步回归法处理两类样本模式分类问题.

1 数据预处理

共取训练样本106个，将样本分为两类，第1类为正常人样本，共72个；第2类为癌症病人样本，共34个. 测得的数据在进行计算机分析前，先使各变量的均值为0，均方差为1.

2 PLS法处理结果

PLS方法的特点是：在进行正交分解时引入了目标变量(分类)信息，能较有效地确定两类样本点在多维空间中变化的总趋势，经过正交分解得到的正交分量中，第一分量包含的信息最多，其次是第二分量，因此可用这两个分量构成判别平面. 本工作使用了PLS1算法[5]，其中目标变量用分类信息代入，处理15个变量数据的结果见图1. 图中的两坐标分别为PLS法求得的第一和第二得分矢量. 图1中每个点都由原15维空间的样本点映射而来，由图可看出，两类样本点有明显的分类趋势，但两类样本点之间仍有部分样本点落在对方区域. 用另一种线性判别分析最优判别平面法处理也可得到与图1类似的结果.

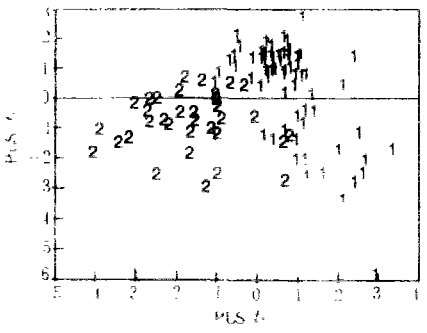

Fig. 1 Results for 15 variables by PLS method

1. Normal people, 2. Cancer patient.

3 多元多项式扩展与逐步回归结果

为了能得到分类更清晰的判别平面图，我们除了考虑各变量线性影响外，还考虑其非线

性影响，即对原各变量进行扩展，引入原各变量的平方项和所有二次交叉项，以及所有三次组合项，这样对 15 个变量进行扩展，可使数据由原来的 15 维增加至 815 维，再用逐步回归正向选择法[5]对所构成的 815 维变量进行筛选或压缩，从近千维数据中筛选出一些含信息较多的变量，对这些变量再用 PLS 法处理，得到判别平面图. 本工作中用的变量筛选判据为 PRESS 判据[5]. 根据 PRESS 判据值为最低，用逐步回归正向选择法选出了 18 个变量见表 2 (X 下标对应表 1 中的编号)，再用 PLS 法对变量筛选出的 18 维数据进行处理，结果见图 2.

Table 2 Selected nonlinear factors and model

No.	Factors	Model coefficient	No.	Factors	Model coefficient
1	X_{15}	−0.0895	11	X_7X_{15}	0.986
2	X_{14}	0.303	12	$X_5X_7X_{15}$	−0.746
3	$X_1X_{14}X_{15}$	0.00151	13	$X_8X_8X_{15}$	0.0128
4	$X_6X_{13}X_{14}$	0.00938	14	$X_2^2X_8$	0.000114
5	$X_8X_{11}X_{15}$	−0.0177	15	X_3^2	0.101
6	$X_{10}^2X_{15}$	−0.0464	16	X_6	−0.286
7	$X_{11}X_{14}X_{15}$	−0.0105	17	$X_7X_{14}X_{15}$	−0.228
8	$X_1X_6X_{15}$	−0.00371	18	X_9X_{15}	−0.0055
9	$X_2X_7X_{15}$	−0.000499		const.	0.1933
10	$X_3^2X_{15}$	−9.8E−05			

由图 2 可看出，应用上述方法，头发样本中的两类人即癌病患者及正常人可更清晰分类，两类样本点之间可明显划出界限. 分类愈清晰表明模型辨别癌症的能力愈强，对未知样本的预报准确性也愈高.

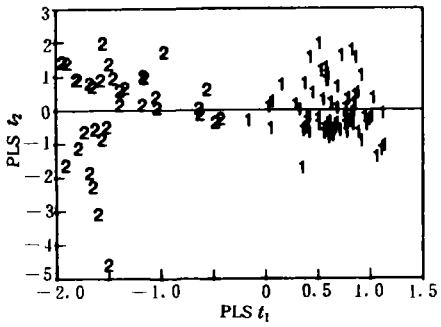

Fig. 2 Results for selected variables by PLS method

1. Normal people, 2. Cancer patient.

4 预报模型

根据 PRESS 判据值为最低，用逐步回归正向选择法选出 18 个变量后，再用 PLS 法对变量筛选出的 18 维数据处理，再根据 PRESS 判据值为最低，删除含噪音多的隐变量，建立了预报模型，模型的系数见表 2 中的第三列. 为了考验所建立的模型，采集了 5 个检验样本，原子光谱分析结果见表 3，再用所得模型对 5 个样本进行预测，按其预测值与期望值的接近程度，决定其属于哪一类，预测结果表明，有 3 个是第 2 类样本(癌症病人)，另 2 个是第 1 类样本(正常人)，这些预报结果与实际结果完全一致(表 3).

Table 3 Predicted results for test samples(μg/g)

Samples No.	1	2	3	4	5	Samples No.	1	2	3	4	5
Zn	141.8	168.1	146.7	198.6	326.9	V	1.19	3.12	0.08	0.01	4.78
Pb	7.46	11.85	4.93	6.29	4.02	Al	7.304	11.852	6.074	0.010	3.357
Ba	2.09	3.49	3.03	1.75	6.02	Ca	457.0	620.0	616.8	405.5	1492.1
Ni	1.02	2.04	0.27	0.49	0.23	Cu	10.92	9.74	7.18	10.07	14.29
Co	0.38	1.03	0.11	0.20	0.10	Ti	0.707	2.911	2.940	0.237	0.001
Cd	0.38	0.94	0.48	0.28	0.29	Se	0.273	0.643	0.543	0.031	0.055
Mn	3.87	1.03	0.33	0.18	1.06	Predicted value	0.842	0.743	1.202	0.178	0.085
Cr	1.03	2.48	1.18	0.38	0.01	Predicted class	2	2	2	1	1
Mg	45.8	49.4	66.1	70.4	1.5	Real class	2	2	2	1	1

5 头发中的微量元素与癌症的关系

从判别分析图 1 和图 2 可知，正常人和癌症病人头发中的微量元素是有明显区别的. 根据各变量与目标变量的相关系数分析，所得预报模型的因子分析以及表 1 中两类样本均值之差分析可知，正常人与癌症病人头发中有显著区别的微量元素有 Se、Ti、Cu、Cr、Zn、Ca，其中癌症病人头发中 Se、Ti、Cr 含量明显高于正常人，而 Cu、Zn、Ca 的含量明显低于正常人.

6 所用软件

本工作所用软件是本实验室在多变量分析方法软件的基础上自编的. 由于本研究中数据扩展增维后，单个数组远超过 64KB，因此所用软件用 Borland C^{++}语言编写、用 huge 指针命令可使单个数组大大超过 64KB，该软件可在 286，386 机器上运行，内存为 2 兆.

进一步的研究工作将集中在扩大样品量及不同癌症的分类.

参 考 文 献

1 Hoskuldsson A.; J. Chemoetrics, 1988, 2: 211

2 Haaland D. M., Thomas E. V.; Anal. Chem., 1988, 60: 1193

3 LUAN Jie-Gu(李介谷), CAI Guo-Lian(蔡国廉); Jishuanji Moshi Shibie Jishu(计算机模式识别技术). Shanghai: Shanghai Jiaotong University Press, 1986: 286

4 ZHANG Yao-Ting(张尧庭), FANG Kai-Tai(方开泰): Duoyuan Tongji Fenxi Yinlun(多元统计分析引论), Beijing: Science Press, 1983: 298

5 Myers R. H.; Classical and modern Regression with Application, Boston, Massachusetts: Duxbury Press, 1986: 105

Research on an Application of Multivariate Discriminant Analysis to Cancer Diagnosis

ZHU Er-Yi*, WANG Xiao-Ru, DENG Zhi-Wei, YANG Peng-Yuan, HUANG Ben-Li

(*Department of Chemistry, Xiamen University, Xiamen*, 361005)

Abstract The concentrations of 15 trace elements in human hair samples including both normal people and cancer patient samples are determined by ICP-AES and GFAAS. These concentration data are then treated by use of a technique which involves polynomial expanding, model selection and PLS method. The good classification results between two classes of samples are obtained and showed in discrimination plane figure. It is believed that the hair sample can be used as a test sample instead of the blood sample in the diagnosis for the cancer.

Keywords ICP-AES, PLS, Polynomial expanding, Stepwise regression, Cancer diagnosis

(Ed.: Z, S)