人发微量元素与性别关系的模式识别分类研究*

章 元 朱尔一 庄峙厦 李 波 王小如
(厦门大学化学系, 材料和生命过程分析科学国家教委开放研究实验室, 厦门, 361005)

摘要 通过对人发样品中22种元素含量的数据进行变量扩维及压缩筛选处理, 选出了影响性别判断较显著的变量, 用PLS法处理这些变量组成的数据, 得到男性与女性分类清晰的二维判别图及预报模型, 并根据所建立的预报模型及人发微量元素的含量判别人的性别, 准确率为81%.

关键词 变量筛选, PLS回归, 微量元素
分类号 O657.31

人体中的化学元素特别是微量元素是维持人体生命活动的必要物质, 在体内具有重要的生理功能和营养作用. 人发作为人体组织的一部分, 其中元素含量能反映人体内微量元素运动变化的平衡水平及积累状况. 人发是最容易取样的部分, 并且化学元素与人发角质氨基酸具有亲合固化作用. 人发一经生长出来, 其中的微量元素含量就基本稳定, 具有较好的分析重现性. 对人发中微量元素进行测定分析, 并对所得大量数据进行分析研究, 对了解人体内微量元素与各种疾病之间的关系有积极的作用.

本研究对用原子光谱分析技术测定福州地区正常人发样品中22种微量元素所获得的数据进行处理, 考虑到变量间的非线性关系, 采用了变量扩维及压缩筛选方法(即引入原始变量的非线性项, 包括变量的平方项、所有二次交叉项及变量两两相除项, 再用正交递归选择法[1]对所有的变量进行压缩筛选, 该法可使二类样本分辨能力大大提高). 用PLS[2,3]方法对筛选的变量所组成的数据进行分类处理, 建立性别判断模型, 用于性别的判断. 通过人发微量元素含量的分析, 若能判别出人的性别,可以对公安侦破等方面起积极的辅助作用.

1 实验部分

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人发样品中22种微量元素[4], 表1列出了男性和女性头发样品中22种微量元素的平均含量、总平均值及标准偏差, 样本总数119, 其中男性74例, 女性45例.

Table 1 Spectroscopy analysis of trace elements in human hair/($\mu g \cdot g^{-1}$)

Element	Man average value	Woman average value	Total average value	Mean square deviation	Element	Man average value	Woman average value	Total average value	Mean square deviation
B	5 361.47	4 366.10	4 985.07	981.68	Fe	41.13	30.65	37.17	29.93
As	79.52	53.56	69.71	29.01	Mn	4.54	4.54	4.54	3.24
Mo	10.84	5.82	8.94	4.94	Cr	9.22	4.42	7.41	4.76
Zn	219.47	177.26	203.51	76.60	Mg	53.19	54.51	53.69	26.76
Se	56.39	31.21	46.87	25.97	V	11.53	5.69	9.32	5.29

收稿日期: 1997-11-25. 联系人: 朱尔一. 第一作者: 章 元, 男, 24岁, 硕士研究生.
* 国家教育委员会留学回国人员科研启动费及福建省自然科学基金资助课题.

Continued

Pb	36.89	22.02	31.26	13.96	Na	104.30	98.55	102.13	61.71
Ba	4.86	4.73	4.81	3.22	Al	85.99	51.72	73.03	39.85
Ni	12.31	6.25	10.02	5.72	Ca	655.84	742.21	688.50	278.19
Co	8.34	4.13	6.75	3.81	Cu	20.75	15.93	18.93	4.75
Cd	2.79	1.37	2.26	1.28	Ti	7.88	5.18	6.86	3.90
Si	23.13	15.43	20.22	11.26	K	331.40	233.91	294.53	160.84

2　数据处理结果与讨论

2.1　数据预处理

研究样本总数 119个, 为对判别模型预报能力进行检验, 先将样本分为两部分, 其中用于训练的样本 80个, 用于预报检验的样本 39个. 训练样本中第一类为 46个男性头发样本, 第二类为 34个女性头发样本. 测得的数据在进行计算机分析前, 先进行标准化, 使各变量均值为 0, 均方差为 1.

2.2　扩维与正交递归选择法筛选变量结果

为了能得到分类清晰的判别平面图及预报能力较强的模型, 对原始数据采用变量扩维及压缩筛选. 由于自变量与目标变量间存在非线性关系, 所以在筛选的变量中包含各变量线性和非线性因子, 非线性因子包括变量的平方项、所有二次交叉项及变量两两相除项, 将所有的变量全部参加筛选, 即对原有变量进行扩展. 在影响性别判定的因素中变量两两相除项比变量的平方项、所有二次交叉项重要得多, 所以将变量两两相除项保留, 平方项、所有二次交叉项删除, 所剩的变量重新用正交递归选择法筛选, 变量筛选用 PRESS判据[5], 根据 PRESS值为最低或接近最低, 从 2 000多个变量中筛选出 7个含信息量较多的变量 (表 2).

Table 2　Selected nonlinear factors in human hair

Parameter	K/Co	Mo	Cd/Mo	Co/Al	B/K	B/Al	V/Cu
Man average value	40.157	10.846	0.265	0.101	20.558	77.947	0.534
Woman average value	60.227	5.817	0.239	0.083	21.976	100.883	0.354
Total average value	47.746	8.944	0.255	0.094	21.095	86.620	0.466
Mean square deviation	21.051	4.945	0.056	0.021	10.039	47.007	0.154

2.3　PLS法处理结果

PLS方法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, 可用这两个分量构成判别平面. 采用 PLS1算法处理用正交递归选择法筛选出的 7个变量数据, 所得判别分类图见图 1.

由图 1可看出, 应用变量扩维及正交递归选择法筛选变量, 再用 PLS法处理, 能得到男女两类头发样本清晰分类的图, 两类样本点明显分布在两个不同的区域.

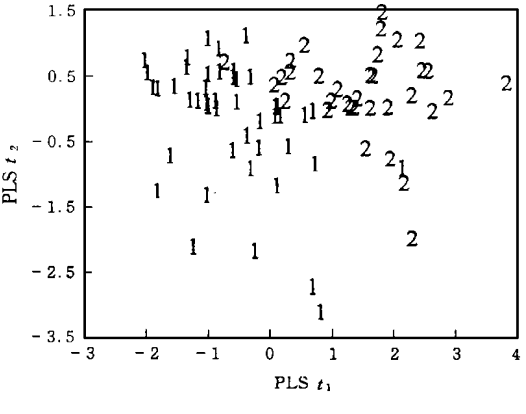

Fig. 1　Results for selected variables by PLS method

1. Man, 2. Woman.

2.4　预报模型

根据 PRESS判据用正交递归选择法筛选

出 7个变量后，再用 PLS法对筛选出的 7个变量进行数据处理，并根据 PRESS值最低，删除含噪音多的隐变量，保留 4个隐变量，建立预报模型，模型系数见表 3，其中系数 1为数据标准化后的模型系数，系数 2为数据未经过标准化的模型系数.

Table 3 Selected nonlinear factors and model coefficient

Factor	Model coefficient 1	Model coefficient 2	Contribution order	Factors	Model coefficient 1	Model coefficient 2	Contribution order
K/Co	0.076 711	0.006 427	6	Mo	-0.342 34	-0.184 1	1
Cd/Mo	-0.234 24	-6.990 63	2	Co/Al	-0.215 3	-11.367 9	3
B/K	-0.121 12	-0.016 48	4	B/Al	-0.015 15	-0.000 47	7
V/Cu	0.120 695	1.352 559	5	Const	0	5.177 216	

用先前预留下的 39个样本(其中男性 28例，女性 11例)代入预报模型，可得到模型的预测值，按照预测值与期望值的接近程度决定属于哪一类，预测准确率为 81%，比不经过变量扩维及筛选直接用 PLS回归所预报的准确率高 10%. 这一结果说明模型选择(变量选择)的准确性对模型判断准确性有重大的影响，因为模型中自变量与目标变量间可能存在非线性关系，所以只考虑线性关系，用 PLS等线性方法处理，得到模型的预报准确性较低. 以上分析还可得出，对于性别判断有较大影响的因素及排列顺序为 Mo Cd/Mo Co/Al, 从表 2可看出，这 3种因素男性均高于女性.

参 考 文 献

1 ZHU Er-Yi(朱尔一), YANG Peng-Yuan(杨 原), DENG Zhi-Wei(邓志威) *et al*.. Chem. J. Chinese Universities (高等学校化学学报), 1993, **14** 1518

2 Hoskuldsson A. J.. Chemoetrics, 1988, **2** 211

3 Haaland D. M., Thomas E. V.. Anal. Chem., 1988, **60** 1193

4 WANG Xiao-Ru(王小如), ZHU Er-Yi(朱尔一), YAN Xiao-Mei(颜晓梅) *et al*.. Acta Chimica Sinica(化学学报), 1993, **51** 1094

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

Classification Study by Pattern Recognition on the Relationship Between the Trace Elements in Human Hair and Sex

ZHANG Yuan, ZHU Er-Yi*, ZHUANG Zhi-Xia, LI Bo, WANG Xiao-Ru

(*Department of Chemistry, The SEDC Research Laboratory of Analytical Science for Material and Life Chemistry, Xiamen University, Xiamen*, 361005)

Abstract The data of 22 trace elements concentrations in human hair samples were obtained by ICP-AES and GFAAS. The variables which have significant influence on discriminating the sex are selected through the treatment of the concentration data by the variable dimension expansion and the variable selection methods. The discrimination plane figure with the good classification is obtained through the treatment of the data with selected variables by PLS method. The prediction models are built and used to distinguish the human sex according to the element concentrations data in human hair. The accuracy of the prediction is 81%.

Keywords Model selection, PLS regression, Trace element (Ed.: Z, G)