

模式识别-人工神经网络方法 评价化探异常

徐驰 李明

(中国科学院上海冶金研究所)

提 要 人工神经网络和模式识别(包括主成分分析和非线性映照方法)应用于化探异常评价。样本取自湖南南部不同矿点,特征变量包括化学成分 Pb、Mo、As、Sb、Ni、Sn 以及物理和地质变量。二种方法对训练样本的分类正确率达 100%。据此模型预报了若干个矿点为锡矿区。

关键词 模式识别 人工神经网络 化探

计算机模式识别技术是近年发展起来,用于信息处理、分类、决策等方面的一门新技术^[1]。1980年起,我们将数学地质和模式识别方法相结合,开展了地质化学模式识别的系列研究,如对我国数十个主要矿床铁帽进行评价,利用地表浅层信息预报深部矿产类型^[2];在广西大厂锡多金属矿田作成矿预测,均取得了令人满意的结果。由于计算机技术的发展,近年来,人工神经网络方法无论在算法上,还是应用上都有了明显的进展。本文将模式识别和人工神经网络方法相结合,以湘南地区化探锡异常评价为对象,将人工神经网络方法引入地质勘探领域,并讨论其应用的若干特点。

1 地质概况

研究区位于湘南宝山—黄沙坪地区,处于东亚新华夏系第二、第三隆起带和第二沉降带的南部,南岭纬向构造带北缘,地层出露较全,构造、岩浆活动十分强烈,多金属矿产丰富。本区西部有来阳-临武南北构造带通过,由强烈挤压的南北向线型褶皱及压性断裂层组成,该带产出大量铅锌多金属,锡多金属和稀有矿产。东部有千里山-杨梅山南北构造带通过,该带控制了东坡金属矿田和五盖山铅锌锑汞矿带。研究区内各类矿产较多,本文仅对锡异常作出评价。成型锡矿床有两种类型,即与斑岩有关的云英岩型锡矿床(如果牌岭)和与花岗岩接触带有关的硫化物矿床(如安源)。测区面积约 3200km²。

2 样本和特征变量

研究区内共有各类锡异常(样本)50个,由1:25000化探次生晕数据平面图上获得。其中14个异常列为已知有矿异常。选出13个已知有矿异常进入训练样本集,另一个已知有矿异常作为待查样本,以检验模式识别-人工神经网络方法预报的正确性。在有矿异常中,计有大型有望隐伏锡矿床1个,小型锡矿床3个,与钨或铅矿伴生的中小型锡矿床6个,矿点3个。选定7个无矿异常进入训练样本集。这些无矿异常经钻探检验查明为无矿的属少数,大部分系根据成矿地质条件差,化探异常规模小,找到矿的可能性甚小确定的。余下的列为待查异常样本。这类异常地表未见矿化,深部未进行详细地质工作,将从中预测成矿远景区。

所用特征变量包括化探的元素含量,物探磁异常和地质定性变量。分别根据1:25000化探数据图、物探磁异常图和1:50000地质图上各自对应的异常范围内统计获得。对化探元素定量变量采用R型因子分析、逐步回归分析及权分析等多种方法进行筛选,对地质变量则根据矿床地质特征和成矿地质条件的分析加以挑选,据此选定Pb、Mo、As、Sb、Ni、Sn等6种化探元素的平均值(以 $X_1 \sim X_6$ 表示)以及磁异常的有无、有利地层、断裂构造、褶皱构造、岩浆岩蚀变是否出现等五个二态变量(以 $X_7 \sim X_{11}$ 表示)对异常进行描述。在二态变量中,有利于成矿的取1,不利于成矿的取0。这11个特征变量构成区分矿与非矿锡异常以及矿床规模的特征变量集。其特征变量平均值和方差如表1所示。

表1 特征变量的均值和方差(ppm)

Table. 1 Mean deviation and variance of charateristic variables (ppm)

特征	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
变量	Pb	Mo	As	Sb	Ni	Sn	磁异常	有利地层	断裂构造	褶皱构造	岩浆岩蚀变
均值	176.19	5.92	34.19	8.93	103.81	28.10	0.34	0.70	0.62	0.54	0.52
方差	333.75	3.63	31.88	7.34	63.69	18.02	0.48	0.46	0.49	0.50	0.50

3 模式识别评价化探异常

模式识别方法是将多维空间信息降到低维空间信息的一种信息处理技术。统计模式识别一般分训练阶段和实现阶段。在训练阶段,对已知样本集(包括已知有矿异常、已知无矿异常)寻找一个有效的分类规律,包括聚类特点,数学模型、判决函数。在实现阶段,则根据已建立的分类规律,对输入性质不明的模式(本文中的待查异常样本)进行分类、判别、预报和决策。从信息加工过程来说,一个模式识别系统实际上包括三个空间:由原始信息构成的高维数据空间(如本文中11维数据空间);经模式识别后得到的高维性质空间;降维处理后成为人们能用目

测的低维映射空间(一般为二维平面)。

用于模式识别成矿预测的方法很多,须根据具体研究对象的特点和数据结构灵活运用。本工作中应用的主成分分析法(PCA)和非线性映射性(NLM),其原理在文[2]中已作介绍,在此不作赘述。

以有矿(包括矿点)、无矿作为评价准则,对 20 个训练样本进行判别。图 1 是 NLM 结果,数据经标准化处理。图 1 表明:在 NLM 图中,有矿和无矿异常分别占据不同区间,分类正确率达 100%。图 2 是 PCA 结果,数据经正规化处理,亦按有矿、无矿作为评价准则。由图 2 可知:已知有矿异常和已知无矿异常分别聚集在不同空间。直线 Z_1 (实际上为一超平面)上方为有矿区($Z_1 < 0$)。直线 Z_2 下方为无矿区($Z_2 > 0$)。对 30 个待查样本,可根据其落入位置或 Z_1 、 Z_2 值判别其为有矿异常或无矿异常。对于 Z_1 、 Z_2 超平面之间的区间,由于该区间缺少训练样本,因而较难定性。

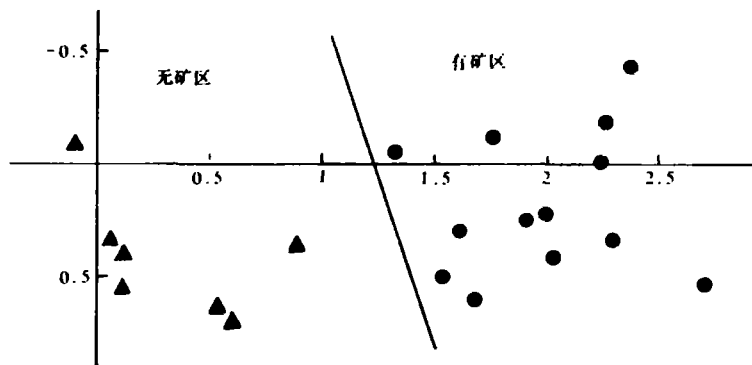


图 1 NLM 结果(●已知有矿样本;▲已知无矿样本)

Fig. 1 NLM results(●ore samples ▲nonore samples)

X 、 Y 及 Z_1 的方程分别是:

$$X = 0.056X_1 + 0.146X_2 + 0.113X_3 + 0.097X_4 + 0.142X_5 + 0.139X_6 + 0.207X_7 + 0.528X_8 + 0.477X_9 + 0.442X_{10} + 0.411X_{11}$$

$$Y = 0.047X_1 - 0.032X_2 + 0.149X_3 + 0.023X_4 - 0.114X_5 + 0.084X_6 + 0.757X_7 - 0.310X_8 + 0.0008X_9 - 0.374X_{10} + 0.381X_{11}$$

$$Z_1 = -0.131X_1 - 0.495X_2 - 0.210X_3 - 0.285X_4 - 0.564X_5 - 0.357X_6 + 0.098X_7 - 1.987X_8 - 1.504X_9 - 1.778X_{10} - 0.922X_{11} + 4.6$$

由图 2 判别,落入有矿区的样本为 No. 21、22、23、26、32、33、34、41、43、44、45、49、50 等 13 个样本。落入无矿区的样本为 No. 36、37、38、39、40、46、47、48 等 8 个样本。其它样本落入 Z_1 和 Z_2 之间,性质难以认定。

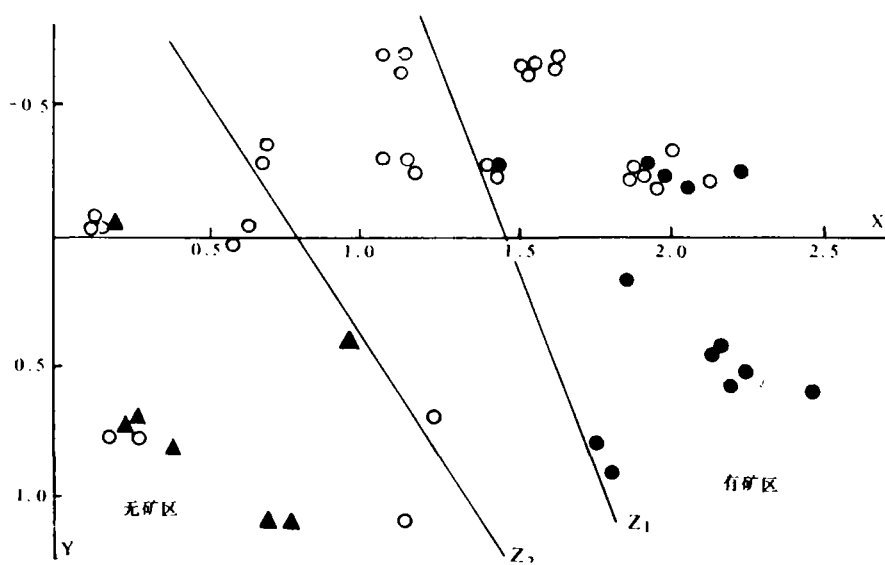


图2 PCA 结果(第1、第2主成分)(●已知有样本;▲已知无矿样本;○待查样本)

Fig. 2 PCA results, (1, 2 principles)

4 人工神经网络评价化探异常

4.1 原理和方法

人工神经网络(ANN)是由众多神经元连成的动态运行非线性动力学系统,具有高维性,不可预测性,不可逆性和多个吸引子(即当时间趋于无穷大时,系统所能达到的状态)。它可以象人类那样根据环境的变化不断改变自己,具有自适应、自组织和自学习的能力。它对信息的处理有这样的特点:一是信息的存贮是分布在整个网络,而不是存放在一个地方;二是每个神经元都是独立的信息处理单元,可以根据接收到的信息独立进行运算。

本工作中采用三层 B-P(Back-propagation)模型^[3]。模型的网络拓扑结构如图3所示。B-P 网络的学习过程是将输出误差反向传播,以逐层修正权值的过程。通过对已知性质样本的训练,调节各层节点间连接权值,就可完成从输入层到输出层的“非线性映照”或建立起输入信息与输出信息之间的非参数化联系。

对于输入模式 P,如输出层 M 的第 j 个节点的实际输出的 O_j^M ,正确输出的 Y_j ,则误差定义为:

$$E_P = \frac{1}{2} \sum_j (Y_j - O_j^M)^2 \quad (1)$$

B-P 模型的建立就是通过调整各连接权值以使 E_P 达到最小如用最速下降法解 E_P 的极小问题,则第 K-1 层 i 节点到 K 层 J 节点的连接权修正量为:

$$\Delta W_{ij}^{k-1,k} = -\varepsilon \frac{\partial E_p}{\partial W_{ij}^{k-1,k}} \quad (2)$$

式中 ε 为修正步长,或学习率。由(1)、(2)式不难导出具体的修正权值公式:

$$\Delta W_{ij}^{k-1,k} = -\varepsilon \delta_j^k O_i^{k-1} \quad (3)$$

其中

$$\delta_i^k = \begin{cases} (O_j^k - Y_j) \cdot f'(i_j^M) & K=M \\ f'(i_j^k) \cdot \sum_l W_{jl}^{k+1} \cdot \delta_l^{k+1} & K \neq M \end{cases}$$

i_j^k 为第 K 层 j 节点的输入; $f(X)$ 称为活动函数(activation function),表示节点的输出对输入的函数关系, f 的适当形式可使网络性态良好。

为了加速收敛,减少振动,可以在权值修正公式中加入一动量项,即:

$$\Delta W_{ij}^{k-1,k}(t) = -\varepsilon \delta_j^k O_i^{k-1} + \alpha \Delta W_{ij}^{k-1,k}(t-1) \quad (3a)$$

X 为一小正数,可通过试验确定, t 表示学习次数。

对所有的训练样本输入模式重复运用(3a)式,直到网络的系统总误差小于某一给定精度 σ ,网络的学习过程结束。

将预报样本的输入模式输入到训练好的网络,网络就可以给出分类信息,作出预报。

活动函数 f 的常用形式有线性、正弦、双曲正切、阈值型和 Sigmoid 型。其中 Sigmoid 函数由于对线性和非线性的问题都能很好地适应,应用量广泛,其函数形式如(4)所示。

$$f_s(x) = \frac{1}{1 + e^{-(x-\theta)/\theta_0}} \quad (4)$$

式中 θ 称偏置,确定函数的区域; θ_0 称增益,确定函数幅值变化斜率。本文的网络中各层节点的活动函数均采用 Sigmoid 函数。

本工作中输入层节点数为 11,隐含层节点数为 8 或 10,输出层节点数为 1,学习率 $\varepsilon = 1.0$ 。动量项系数 $\alpha = 0.02$,增益 $\theta_0 = 1.0$ 。

4.2 计算结果

将 20 个已知样本(No. 1~20)分成有矿异常(包括矿点)和无矿异常两类,其输出值分别以 1 和 2 表示。对该网络进行训练后,输入 30 个待查样本,其输出结果如表 2 所示。如输出值 < 1.5 的预报为有矿异常,反之则预报为无矿异常。据此预报 No. 21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、41、42、43、44、45、49、50 等 22 个样本为有矿异常,而 No. 36、37、38、39、40、46、47、48 等 8 个样本为无矿异常。

为进一步预报含矿规模,将 13 个已知有矿样本(包括矿点)分为有矿异常和矿点异常二类,其输出值分别以 1 和 2 标志。对该网络进行训练后,将上述 22 个预报为有矿异常的样本作为待查样本,输入网络。其输出结果如表 3 所示。如输出值 < 1.5 ,预报为有矿异常,反之则预报为矿点异常。据此,预报 No. 22、34、35、49、50 等 5 个待查样本为具有相当含矿规模的有矿异常,23 号样本输出值为 1.5,属于有矿异常—矿点之间,其余待查样本预报为矿点异常。

表 2 人工神经网络和模式识别预报结果比较

Table. 2 Predicting results of artificial neural network model recognition

样本号	输出值	ANN 预报	模式识别预报	样本号	输出值	ANN 预报	模式识别预报
21	1.00	有矿	有矿	36	1.97	无矿	无矿
22	1.01	有矿	有矿	37	2.00	无矿	无矿
23	1.02	有矿	有矿	38	2.03	无矿	无矿
24	1.03	有矿	未定	39	2.03	无矿	无矿
25	1.01	有矿	未定	40	2.06	无矿	无矿
26	1.01	有矿	有矿	41	1.00	有矿	有矿
27	1.02	有矿	未定	42	1.03	有矿	未定
28	1.03	有矿	未定	43	1.01	有矿	有矿
29	1.02	有矿	未定	44	0.99	有矿	有矿
30	1.03	有矿	未定	45	0.97	有矿	有矿
31	1.05	有矿	未定	46	1.99	无矿	无矿
32	1.00	有矿	有矿	47	2.00	无矿	无矿
33	1.02	有矿	有矿	48	2.01	无矿	无矿
34	0.97	有矿	有矿	49	1.02	有矿	有矿
35	0.03	有矿	未定	50	1.01	有矿	有矿

输出变量 0

下面对预测结果作若干地质分析。22 号样本，在 ANN 中预报为具有相当含矿规模的有矿异常。该异常位于桂阳宝山以北约 7km 处，异常区处于下石炭统含矿地层和中上石炭统壶天群组成的次级向斜翼部（主要落在含矿地层中）。距异常东面 200~300m 处，有岩体出露，成矿地质条件有利。化探异常约 1km²，Sn 具浓度分带，最高值 70ppm 该异常与界牌岭锡异常（已知大型有望隐伏矿异常）极相

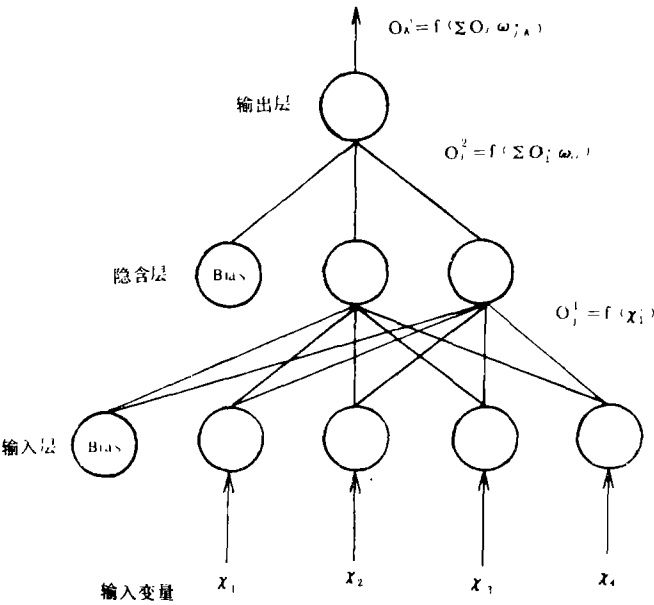


图 3 三层 B-P 网络结构示意图

Fig. 3 Structural sketch of 3-layered B-P networks

似,因而 ANN 预报与地分析相符,又如 50 号样本,在 ANN 中亦预报为具有相当含矿规模的有矿异常。该异常处于下石炭统含矿地层及中上石炭统壶天群组成的次级背斜轴部及南北向与东西断裂交叉处,地表出露花岗斑岩体,成矿地质条件有利。异常面积约 0.3km^2 , Sn 具浓度分带,中心最高值达 30ppm。该异常的地理位置与界牌岭大型有望隐伏矿异常紧邻。在 50 号异常南缘 143 号勘探线剖面上,已完工的 ZK2, ZK3 孔分别在 587.64m 和 403.9m 处见到工业锡矿体。

表 3 含矿规模的预报结果

Table. 3 Predicting results of scales of ore-bearing

样本号	输出值	样本号	输出值	样本号	输出值	样本号	输出值
21	1.82	27	2.90	33	1.71	44	3.65
22	0.93	28	2.03	34	1.10	45	3.54
23	1.50	29	3.00	35	0.82	49	1.37
24	2.38	30	2.20	41	3.91	50	1.01
25	2.25	31	1.93	42	3.16		
26	1.78	32	3.80	43	3.75		

5 结果比较和讨论

(1)对照模式识别(PR)和人工神经网络(ANN)二种方法预报结果,PR 预报为有矿异常的 13 个待查样本,在 ANN 中全部预报为有矿异常;PR 预报为无矿异常的 8 个待查样本在 ANN 中亦全部预报为无矿异常。用 PR 方法性质难以认定的 9 个样本(在图 2 中处于 Z_1 和 Z_2 之间),在 ANN 中全部预报为有矿异常。由此可见:ANN 除了能验证 PR 的预报结果外,对 PR 所难以认定的“中间”区域,ANN 亦能提供分类信息以供参考。

(2)如进一步将训练样本中有矿异常按含矿规模分为二类,ANN 也能对待查样本的含矿规模作出预报。从表 3 所预报的 5 个具有相当含矿规模的待查样本中,除 35 号样本在图 2 中落入“中间”区域外,其余 4 个样本在图 2 中都靠近已知的大型隐伏锡多金属矿或已知小型锡矿异常附近。可见,二种方法的计算结果互为印证。

(3)No. 50 样本是已证实的大型隐伏锡多金属矿,为验证 PR 和 ANN 方法预报的可信性,随机抽样作为待查样本预测属性。在 PR 和 ANN 中, No. 50 样本均被预报为有矿异常,且均被判别为具有相当含量规模的有矿异常。

(4)在计算方案中,我们分二步走。第一步先判别有无矿;第二步,对预报为有矿异常的样本再判别含矿规模大小,即把复杂问题逐步分解为二类判别问题。实践证明,此计算方案较有效。

(5)数学地质中接触到的特征变量值,既有定量变量,又有定性变量(如本文中 0 或 1)。本

工作中,为使化探数据(定量变量)和地质变量(定性变量)更好匹配,对数据作正规化处理是一种行之有效的办法,经这样处理后,所有数据都分布在 0~1 之间。

将人工神经网络方法应用于地球化学的信息处理才刚刚起步。笔者认为,将人工智能和传统的数学地质方法结合起来,能够较全面地揭示事物的本质,片面强调单一的信息处理技术是不可取的。

参考文献

- 1 McCabe G P. principal variables, Technometrics. 1984, 26; 137
- 2 徐驰,陈念贻,龚美菱. 模式识别技术在矿床铁帽评价中的应用. 地球化学, 1989; 167
- 3 Lippmann R P. An introduction to computing with neural networks. IEEE ASSP Magazine, 1987; 4

PATTERN RECOGNITION AND ARTIFICIAL NEURAL NETWORK APPLIED TO EVALUATION OF GEOCHEMICAL ANOMALIES IN THE SOUTH OF HUNAN

Xu Chi Li Ming

(Shanghai Institute of Metallurgy, Academia Sinica)

Abstract

Artificial neural network and pattern recognition (including principal component analysis and nonlinear mapping) methods have been applied to the evaluation of geochemical anomalies. The samples are taken from different prospects in the south of Hunan. The distinct variations are their chemical compositions such as Pb, Mo, As, Sb, Ni, Sn etc., as well as their physical and geological characteristics. The correct classification rate is 100% for training samples by the two methods, and some targets are predicted as potential Sn ore-field.