用人工神经网络进行空间不完备数据的插补

何凯涛[1,4], 陈 明[2], 张治国[3], Jacques Yvon[5]
HE Kaitao[1,4], CHEN Ming[2], ZHANG Zhiguo[3]

1. 中国地质调查局, 北京 100011; 2. 国家地质实验测试中心, 北京 100037;
3. 吉林大学,吉林 长春 130026; 4. 国防科学技术大学, 湖南 长沙 410073
1. *China Geological Survey*, *Beijing*, 100011, *China*;
2. *National Research Center of Geoanalysis*, *Beijing*, 100037, *China*;
3. *Jilin University*, *Changchun* 130026, *Jilin*, *China*;
4. *University of National Defense Technology*, *Changsha* 410073, *Hunan*, *China*;
5. *Laboratoire Environnement et Mineralurgie*, *Nancy*, *France*

摘要:在地学研究中,特别是区域性资料处理过程中,常常遇到"不完备数据"的问题,即所谓的"数据不全"。在尽量减小估计误差的条件下对缺失数据进行预测或插补,对于充分利用历史资料和已知信息,提高预测质量具有重要意义。利用径向基人工神经网络(RBF)同时具有自组织神经网络和回归网络的优点,可以对缺失数据进行预测。实际区域地球化学数据处理的结果表明,RBF网络对空间不完备数据的建模和预测具有优异的效果。

关键词:空间不完备数据;人工神经网络;非线性;知识挖掘;数学地质;区域化探
中图分类号:P59 **文献标识码**:A **文章编号**:1671-2552(2005)05-0476-04

He K T, Chen M, Zhang Z G, Jacques Y. Complement of incomplete spatial information via RBF network. *Geological Bulletin of China*, 2005, 24 (5):476-479

Abstract: "Incomplete information" and its complement are encountered frequently in geo-information processing. It is of great significance to interpolate the lost data via the known historic datasets and improve the quality and accomplishment of information integration. The RBF network possesses the advantages of Kohonen and regression networks. A test was performed to prove the effectiveness of RBF to complement the incomplete spatial information.

Key words: incomplete spatial information; rule-based frame (RBF); artificial neural network; non-linearity; knowledge excavation; mathematical geology; geochemical exploration

地质、地球化学、地球物理、遥感等数据总是在一定的技术条件下,在一定的空间位置测量到的。在进行地学数据处理和信息整合处理时,常常要把研究采样的时间和空间域划分成若干个"单元"。如果在所有的时-空单元上,所有的变量都有足够的数据,则称为"数据是完备的";否则,如果在某些单元上缺失数据,或者出现若干个单元共用一个数据的情况,就称为"数据是不完备的"。对于大部分地学数据与信息处理方法而言,一般要求各种相关的资料具有系统性和完备性,也就是说,要求在每一个观测点上所有的观测变量都有观测值。然而,在处理历史资料时,"资料不全"或"资料缺失"是常见的问题。一般的处理方法是,暂时放弃或删除那些观测资料不全的点, 仅采用那些所有变量都有对应观测值的点。其缺陷有四:①大量的资料浪费;②可能遗漏重要的信息;③使得建模和预测结果产生一定的误差;④某些重要的靶区无法进行预测。当然,最好的方法是用一定的方法对缺失的数据插补,而且插补得到的数据与真实数据不会有很大

收稿日期:2004-03-16;**修订日期**:2005-02-25
基金项目:国家863计划项目(2001AA135120)、国家973项目(G1999045708)和国家自然科学基金项目(49973015)资助。
作者简介:何凯涛(1970-),男,高级工程师,在读博士,从事空间信息系统与技术、GIS、数学地质、非线性方法、网格技术等方面的研究工作。E-mail:hekt@vip.163.com

的误差。

在国内,胡旺等[1]研究了基于粗糙集理论不完备信息系统的数据挖掘;刘新立等[2]研究了空间不完备信息条件下的区域自然灾害风险评估:区域自然灾害风险评估中所用的数据不仅具有时间意义,而且具有空间意义。当数据的空间信息不完备时,需要对其进行优化处理,以减小风险评估的误差;将区域自然灾害风险评估中所遇到的空间不完备信息分为2类,分别用插补模型和校正模型进行处理;插补模型是针对空间数据缺失情况的,而校正模型是针对空间数据不符合精度需要情况的。唐建国[3]研究了"粗糙集理论处理不完备信息的可信度分析":用粗糙集理论得出的决策规则其依据不充分的问题,给出了可信度的概念和定义。在国外,"不完备数据"方面的研究也是比较新的课题。目前在Internet上用"incomplete information"和"geo"来查询只能得到大约1340条消息。纵观各种方法,不外乎概率论、多元统计分析、Kriging、粗糙集、空间插补、人工神经网络等方法。本文将利用径向基人工神经网络(RBF)进行缺失数据预测,并用实例说明其效果。

1　RBF网络的原理与算法

径向基函数(Radial Basis Function,简称RBF)[4,5]神经网络是在借鉴生物神经的局部调节功能和动物大脑中存在的交叠接受知识的区域的基础上提出的一种采用局部感知场来实现函数映射的人工神经网络,同时具有自组织神经网络和回归网络的优点,可用于任意多维实变量空间的插值问题,特别适用于地质曲面的插值重建、缺失数据的补齐和函数拟合。其优点在于,对原始数据的分布型式和边界条件没有特别的要求,收敛速度较快。

RBF网络结构如图1所示。网络由3层神经元组成,第1层是输入层,第2层是非线性径向基函数层(RBF层),第3层是线性输出层。径向基层的传输函数为径向基函数,而输出层的传输函数为纯线性函数。

所谓的径向基函数是一种高斯型函数,其输入是径向基层权值向量w_i与输入向量x_i之间的距离与偏差b的乘积:

$$z=\sqrt{\Sigma(w_i-x_i)^2}\times b \qquad (1)$$

其输出为:

$$y=e^{-z^2}=e^{-\left(\sqrt{\Sigma(w_i-x_i)^2}\times b\right)^2}=e^{-(\|W-X\|\times b)^2} \qquad (2)$$

径向基函数的输出值随着w_i与x_i之间距离的减小而增加;当w_i与x_i之间的距离为0时,输入为0,得到输出最大值1。在实际应用中,常取,$b=\sqrt{1n2}/C$,即

$$y=e^{-\left(\frac{\|W-X\|\times\sqrt{1n2}}{C}\right)^2}=e^{-1n2\left(\frac{\|W-X\|}{C}\right)^2} \qquad (3)$$

其中C被称为伸展常数,用于确定每个径向基层神经元对其输入向量x_i响应的邻域半径。当$\|W-X\|=C$时,有$y=e^{-1n2}=0.5$。C接近于0表示与权值向量w_i距离较近的输入样品,这时候神经元的输出也接近1,说明该神经元对这些样品比较敏感,而对其他输入样品不敏感。相反,当C取值较大时,RBF的响应范围可以扩大,RBF的平滑度也较好。所以,通过调整C值,可以达到调整RBF曲线的形状和曲面光滑程度的目的。

径向基层采用Kohonen网络的聚类方法确定训练输入样品聚类中心,从而确定径向基层的神经元。一般地,考虑N维空间的P个样品$x_1,x_2,\cdots,x_P$,首先将数据作归一化处理,然后将P个样品聚类。

设RBF网络结构为,输入层神经元节点数n,RBF层神经元节点数s,输出层神经元节点数m。设有P对已知训练样品模式(x_p,d_p),$p=1,2,\cdots,P$,则样品输入模式矩阵$X_{n\times P}=(x_1,x_2,\cdots,x_P)$,对应的目标输出模式矩阵$D_{m\times P}=(d_1,d_2,\cdots,d_P)$。

RBF网络的训练分为2个阶段。

第1阶段,采用无导师的方式训练RBF层的权值W^1和偏差b^1。RBF层的权值训练是不断地使权值向量w_s^1趋向于某个输入向量x_p,这样得到的w_s^1构成RBF层权值矩阵W^1,结果使输入样品向量x_p在等于或趋向于w_s^1处,使RBF的输出为1。当网络工作时,输入任一预测样品向量,RBF层中的每一个神经元都将按照输入向量接近每个神经元的权值向量的程度来计算其输出值。结果是,与权值向量的距离很远的输入向量,使RBF层的输出接近0,这些很小的输出对后面的线性输出层的影响很小,可以忽略;与权值向量的距离非常接近的输入向量,RBF层的输出值接近1。

第2阶段,采用有导师的方式训练线性输出层的权值W^2和偏差b^2。在确定了RBF层的权值W^1和偏差b^1之后,RBF层的输出矩阵$Y^1{}_{s\times P}$则可求出。由输出层的输入矩阵$Y^1{}_{s\times P}$和网络的目标输出模式矩阵$D_{s\times P}$,通过使网络的输出向量$y^2{}_m$与目标输出d_m的误差达到目标误差,来训练线性输出层的权值W^2和偏差b^2。

2　空间不完备数据的插补

为了方便,先定义"不完备度(faulty degree)"。如果总共有T个时空单元,其中有L个单元缺失数据,则称

$$R=L/T \qquad (4)$$

为这批数据的"不完备度",而$1-R$为这批数据的"完备度

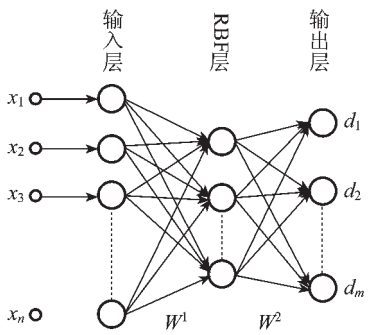

图1　RBF网络结构图

Fig.1　Structure of the RBF network

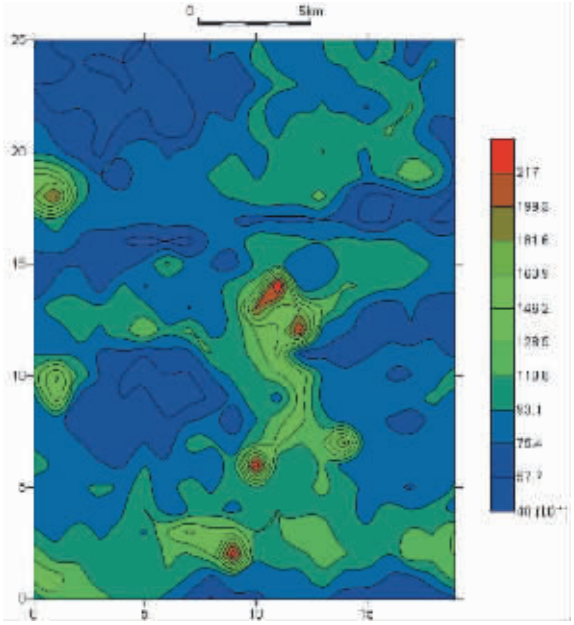

图2 Zn元素原始地球化学图

Fig.2 Original geochemical map of elemental Zn

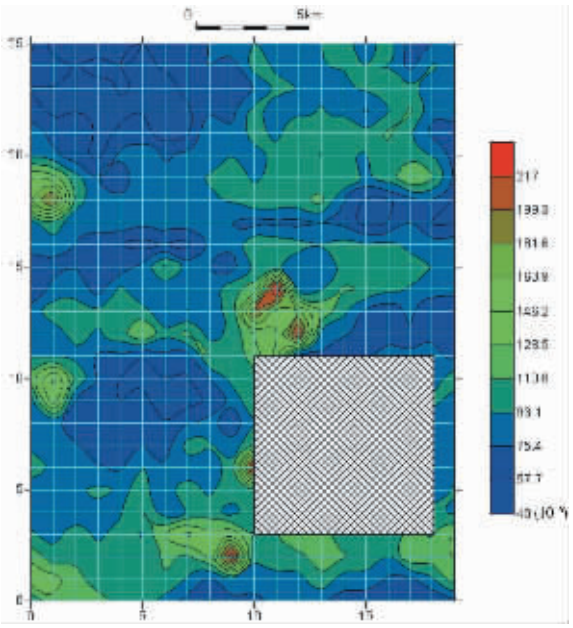

图3 Zn原始数据被去掉一部分后的地球化学图

Fig.3 Geochemical map constructed after a part of original Zn data are deleted

(degree of maturity)"。

下面以吉林某地化探数据为例说明RBF网络在不完备数据的插补上的应用情况。在吉林某地25 km×19 km的范围内按26 km×20 km的规则测网分析了520个化探组合样品,定量测试了Au、Ag、Co、Cr、Cu、Mo、Ni、Pb、Sn、W和Zn11个元素,陈明等[6]、Chen等[7]、陈明等[8]介绍了该地区的基本地质-地球化学特征。Zn元素的原始地球化学图如图2。由于原始数据本身就是网格数据,因此图2是用原始数据直接作出来的等值线图,未作任何插值处理。

笔者的研究思路是:在某任意区域中把Zn的原始数据去掉一部分,然后用RBF网络和剩余数据建立空间坐标和其他元素含量与Zn含量之间的定量对应关系(模型),再用这个模型来预测被挖去的数据,对比预测数据与实测数据之间的绝对误差和相对误差就可以评价RBF对空间不完备数据进行插补的效果了。

首先挖去4~12行和11~19列,此时缺失的数据点个数为81,数据的完备度为84.42%,不完备度等于15.58%,如图3所示。用剩余的439个点的X、Y坐标和另外10个元素的原始数据进行RBF网络建模,建立如下非线性映射关系

$$C_{Zn}=f(X,Y,C_{Au},C_{Ag},C_{Co},C_{Cr},C_{Mo},C_{Ni},C_{Pb}) \quad (5)$$

其中X和Y为X坐标和Y坐标,C_{Zn}表示建立Zn的含量,其余类推。建立模型后,把全区(包括Zn数据缺失区)其他10个元素的含量数据和空间坐标代入网络,重新得到全区Zn的网格化数据,其等值线图如图4。比较图2和图4可以发现,两者区别不大,说明RBF插补效果良好。

下面定量考察不完备度对RBF插补效果的影响。设Zn_{ij}为数据缺失区Zn元素的实际测量值,Zn_{ij}为用RBF估算得到

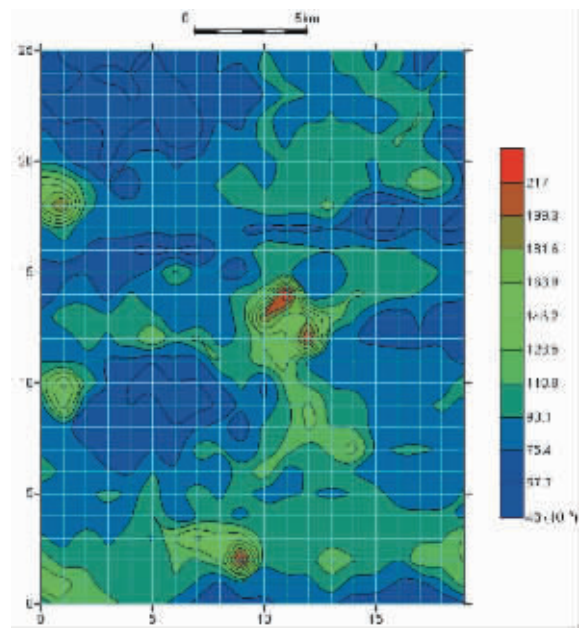

图4 Zn数据RBF插补后的地球化学图

Fig.4 Geochemical map constructed after original Zn data are completed using RBF

的Zn元素含量,则累计误差、累计相对误差和累计平均相对误差分别为

$$S=\Sigma|(Zn_i-Zn'_i)| \quad (6)$$

$$S\%=100\times\Sigma|(Zn_i-Zn'_i)|/\Sigma Zn_i \quad (7)$$

$$\bar{S}\%=\frac{100}{n}\times\Sigma|(Zn_i-Zn'_i)|/\Sigma Zn_i \quad (8)$$

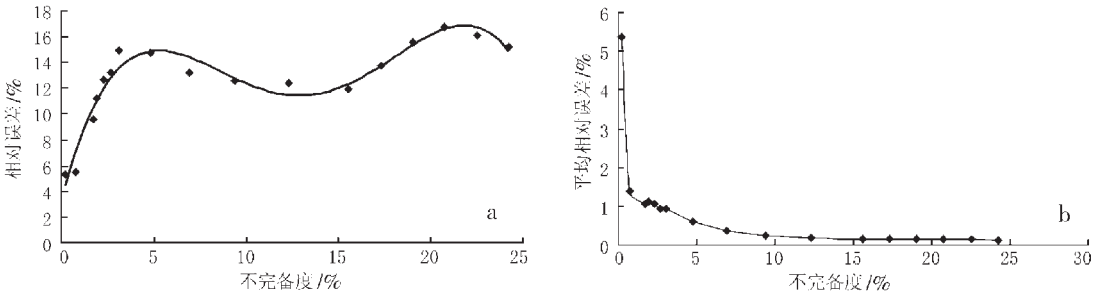

图5　不同不完备度条件下相对误差(a)和平均相对误差(b)的变化情况

Fig.5　Relative error (a) and average relative error (b) in different faulty degrees

表1　不同不完备度条件下RBF网络插补效果

Table 1　Effects of RBF to complement deficient data in different faulty degrees

序号	缺失单元格数	不完备度(R%)	相对误差(S%)	平均相对误差($\bar{S}$%)
1	1	0.19	5.36	5.36
2	4	0.77	5.51	1.38
3	9	1.73	9.55	1.06
4	10	1.92	11.24	1.12
5	12	2.31	12.66	1.06
6	14	2.69	13.24	0.95
7	16	3.08	14.90	0.93
8	25	4.81	14.73	0.59
9	36	6.92	13.23	0.37
10	49	9.42	12.54	0.26
11	64	12.31	12.41	0.19
12	81	15.56	11.96	0.15
13	90	17.31	13.71	0.15
14	99	19.04	15.55	0.16
15	108	20.77	16.77	0.15
16	117	22.50	16.08	0.13
17	126	24.23	15.19	0.12

它们可用于衡量插补误差的大小。

表1是不同的“不完备度”条件下相对误差和平均相对误差的变化情况,图5是相应的变化曲线。由图表可见,随着不完备度的加大,RBF网络的插补相对误差逐渐增大，这一点与“资料缺失程度越大,预测误差也越大”的传统理念是统一的。但有意思的是,随着不完备度的加大,平均相对误差却越来越小,说明RBF对“不完备度”具有稳健性。根据以上讨论可以认为，利用RBF网络进行不完备数据的插补是可行的,值得推荐使用。

3　小　结

针对地学信息处理和知识挖掘中常常遇见的数据缺失及其插补问题,本文扼要介绍了RBF网络的基本原理及其算法,并以吉林某地实测化探数据为例试验了不同数据缺失程度条件下RBF网络的插补效果。实验结果证明,RBF在空间不完备数据的插补方面具有很强的能力,并对数据“不完备度”具有一定的稳健性,值得推荐使用。

参考文献:

[1]胡旺,冯伟森,李志蜀,等.基于粗糙集理论不完备信息系统的数据挖掘[J].四川大学学报,2004,41(4):744-748.

[2]刘立新,史培军.空间不完备信息条件下的区域自然灾害风险评估[J].自然灾害学报,2000,9(1):26-32.

[3]唐建国.粗糙集理论处理不完备信息的可信度分析[J].控制与决策,2002,17(2).

[4]缪报通,陈发来.径向基函数神经网络在散乱数据插值中的应用[J].中国科技大学学报,2001,31(2):135-145.

[5]王明进,程乾生.基于径向基函数的非线性预测模型[J].管理科学学报,1999,2(4):28-33.

[6]陈明,吴锡生,马福生.法克立格法在吉林某地1:5万化探中的应用及其与其它方法的比较研究[J].吉林地质,1994,13(12):14-19.

[7]Chen Ming, Yan Guangsheng, Fan Jizhang, et al. Regional geochemical division—A tool for delineating geochemical block[J]. Journal of China University of Geosciences, 2000,11(2):150-153.

[8]陈明,何凯涛,王全明,等.地球化学场精细结构解析方案与应用[J].地质通报,2004,23(2):147-153.