2000年第58卷
第11期，1409～1412

化学学报
ACTA CHIMICA SINICA

Vol.58，2000
No.11，1409～1412

一种改进BP网络结构用于化学中的非线性校正

罗明亮　　李梦龙*
（四川大学化学学院　成都　610064）

摘要　针对化学领域中的非线性关系特点，在常规BP网络基础上，提出了一种“杂交”型BP网络，包含两个隐层，并有输入层到输出层的直连接.它可很好地解释数据中同时存在的线性及非线性关系，效果优于多元回归法及普通BP算法.

关键词　BP网络，非线性校正，定量构效关系

Using Modified BP Neural Network for Non－Linear Modeling in Chemistry

LUO Ming－Liang　　LI Meng－Long*
（*Department of Chemistry*，*Sichuan University*，*Chengdu*，610064）

Abstract　BP neural network's excellence in non－linear modeling is shown with two examples. The first is the non－linear calibration for the relationship between the infrared reflectance rates and contents of protein. The second is QSAR, predicting the physical property of some compounds by using the structural parameters. By employing a new mode of BP neural network, with two hidden layers and direct connections from input to output layer, better results than those of multivariate linear regression and normal BP neural network are achieved.

Keywords　BP neural network, non－linear calibration, QSAR

非线性校正在化学领域中应用较多，因为真正符合线性关系的化学体系并不多见，典型的非线性体系如QSAR问题，一般可视为线性与非线性复合.对这类非线性体系的处理，常规化学计量学方法是采用MLR加上一些非线性校正项，如高次项，交互项等.这类算法带有较大的主观性，究竟该对哪一项进行校正，哪两项间应采用交互项等，都是实际应用中难以判断的问题.回归的结果常带有较大的误差.采用神经网络进行非线性校正就不必考虑这些问题.它具有自适应学习能力，能主动地学会存在于输入/输出数据之间的非线性关系.对于这类非线性关系，当我们对其中的因果关系不甚明了时，利用神经网络进行预测通常能获得较好的结果[1-4].本文针对化学中常见的线性与非线性关系交错的复杂局面，提出一种改进BP网络结构来解决这一问题.

1　一种通用的改进BP网络结构

常规BP网络[1]采用三层结构，隐层和输出层均采用Sigmoid函数，实践表明，这种结构在描述非线

* E－mail：liml@mail.sc.cninfo.net
收稿日期：1999－12－23，修回日期：2000－05－08，定稿日期：2000－07－16，国家自然科学基金（29877016）资助课题
（Received December 23，1999. Revised May 8，2000. Accepted July 16，2000）

性关系时比较成功,但面对线性关系时,其效果并不比多元线性回归方法更好[5].化学中的定量校正问题通常是线性和非线性关系同时存在,如何才能使网络模型将数据中的线性和非线性关系同时描述出来呢? 本文提出了如图 1 所示的一种网络结构来达到这一目的.

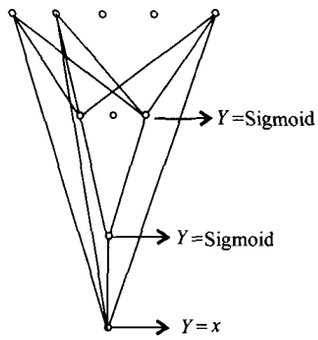

图 1 "杂交"BP 网络

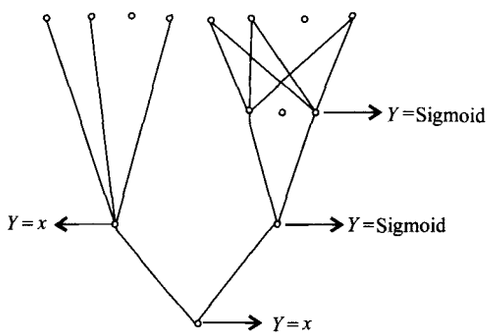

图 2 常规 BP 网络结合线性网络

图中使用了输入到输出层的直连接和两个隐层的结构.隐层各节点使用 Sigmoid 函数(或其他非线性函数),输出层节点使用线性变换函数: $Y = x$.

该网络实际相当于一常规 BP 网络和一个线性网络"杂交"(hybrid)(如图 2).右边是一个常规 BP 网络,它可表示为如下数学式:

$$y = Sig[\sum_j B_j Sig \sum_i (A_{ij}X_i)]$$

其中 X_i 为输入层第 i 个节点输入值,A_{ij} 为输入层第 i 个节点与第 1 隐层第 j 个节点的权连接,B_j 为第 1 隐层到第 2 隐层权连接. Sig 表示 Sigmoid 函数.很明显,它将输入层数据经过线性组合后进行两次 Sigmoid 变换得到输出值,能拟合数据中存在的非线性关系.

左边相当于一个线性网络,可用下述数学公式表示(C_i 为相应权连接),它可描述数据中的线性关系.

$$y = \sum_i C_i X_i$$

在"杂交"网络的输出层,上述两部分线性组合而成网络的输出:

$$y = \sum_i C_i X_i + Sig[\sum_j B_j Sig \sum_i (A_{ij}X_i)]$$

输入/输出数据中的线性关系可以通过左边的线性网络获得描述,而用线性关系不能解释的那一部分,则可通过右边的网络获得解释.

2 定量校正的数据预处理方法

本文以下定量校正关系实例均采用此网络结构,并用尝试法确定隐层节点数据,以交互校验法来防止过拟合.所有原始数据均做标准化和归一化处理,并采用标准预测误差(SEP)来衡量预测效果,其中 N 代表样本个数,T_i 代表第 i 个样本的目标输出值,O_i 为实际输出值.

$$SEP = [\frac{1}{N}\sum_i^N (T_i - O_i)^2]^{1/2}$$

2.1 红外光谱反射率与蛋白质含量之间的非线性校正

表 1 反射率和蛋白质含量

No.	X_1	X_2	X_3	%	%*
1	246	374	386	9.23	9.04
2	240	359	353	10.95	10.82
3	236	352	340	11.67	11.35
4	242	370	377	8.67	9.16
5	243	367	378	9.95	9.43
6	324	448	467	11.87	12.13
7	271	407	451	8.09	7.39
8	274	406	407	8.38	10.30
9	260	385	374	9.64	11.18
10	242	366	353	9.70	10.71
11	255	376	383	10.75	10.45
12	276	396	404	11.47	11.34
13	258	393	377	8.05	10.25
14	288	415	443	10.57	9.94
15	236	386	383	8.01	6.99
16	243	366	371	10.41	9.87
17	273	404	433	9.51	8.85
18	238	370	353	7.75	9.88
19	264	384	398	11.39	10.54
20	233	365	365	8.25	8.75
21	293	421	450	10.23	9.97
22	360	484	524	12.55	12.21
23	269	389	391	11.35	11.39
24	285	410	445	10.75	9.67

本部分采用了一组 3 个波长处的红外光谱反射

率以及对应的蛋白质含量,以此建立反射率与蛋白质含量间的非线性关系.所用数据见表 1,其中 X_1, X_2, X_3 分别代表 3 个波长处的反射率,%栏为样本的真实蛋白质百分含量,%*栏为预测出的蛋白质百分含量(最后二栏的数据均已乘以 100).文献[6]指反射率和百分含量之间存在着较强的非线性关系,并以 1~15 号样本为训练集,以剩余样本为预测集,分别采用了 PLS, PCR, 常规 BP 网络, PCA - NN 以及 PLS 联合 RES - NN 的方法来处理数据[三种 NN 均采用相同结构:单隐层,三个输入节点,一个隐节点(Sigmoid 变换),一个输出节点(Sigmoid 变换)].各方法所得预测集 *SEP* 值:

PLS: *SEP* = 1.34; PCR: *SEP* = 1.34; 常规 NN: *SEP* = 0.51; PCA - NN: *SEP* = 0.39; RES - NN: *SEP* = 0.34.

本文采用的 BP 网络结构为:输入层 3 个节点,对应 3 个反射率,第一隐层含 2 个 Sigmoid 节点,第二隐层含一个 Sigmoid 节点,输出层 1 个节点.使用数据与文献中的其他方法相同,并对输入数据做了预处理.

结果为:训练集 *SEP* = 0.09895,预测集所得 *SEP* = 0.09182;可以看出,本文的 BP 网络所得结果不仅优于 PLS, PCR 等方法,而且优于常规的 BP 网络方法.

2.2　利用分子连接性指数 *X* 和分子拓扑指数 *Am*[7] 预测稀土显色剂的显色反应灵敏度 *ε*

文献[8]利用 20 个稀土显色剂的结构参数: Am_2, Am_3, ${}^4X_{PC}$ 和 ${}^2X_P$ 来研究了和镱显色反应灵敏度 ε 的相关性,根据表 2 数据得到了一个多元线性回归公式:

$$\log\varepsilon = 28.8123 + 1.1369\,{}^2X_p - 2.0313\,{}^4X_{pc} + 6.8844\,Am_2^{1/2} - 7.9967\,Am_3^{1/2}$$

$R = 0.9325$, $F = 24.9908$, $S = 0.1112$, $n = 20$.

对训练集的 *SEP* = 0.100728.

本文所用 BP 网络结构为:输入层 4 个节点,对应 4 个拓扑指数,第一隐层采用了 2 个节点,第二隐层采用了 1 个节点,输出层 1 个节点.输入数据作了预处理,所得训练集 *SEP* = 0.089653.

由二者的 *SEP* 可以看出,采用 BP 网络对训练集结果较好.

表 2　20 个稀土显色剂的结构参数和显色反应灵敏度

No.	Am_2	Am_3	${}^4X_{pc}$	${}^2X_p$	ε	ε*	ε**
1	108.58974	162.29741	15.84230	5.90110	4.60206	4.67141	4.69816
2	108.51160	162.49953	15.83040	5.93512	4.59106	4.50461	4.52629
3	104.89896	159.15538	16.44350	6.08149	4.65321	4.78160	4.77653
4	104.84162	159.14806	16.43160	6.17111	4.62325	4.55415	4.56400
5	104.93790	159.89880	16.27620	6.36491	3.81291	3.76976	3.78835
6	110.51097	163.79880	15.67170	5.82335	4.77815	4.80641	4.82384
7	110.45758	164.14297	16.95020	6.60354	4.59106	4.59547	4.56761
8	108.44515	162.19144	15.56990	5.72918	4.56820	4.69556	4.72318
9	110.31574	164.61942	16.10860	6.20703	4.36173	4.38759	4.22118
10	105.25385	160.16751	15.69750	5.95462	3.76343	3.92953	3.98500
11	105.21507	159.37346	15.80640	5.79215	4.63347	4.68348	4.67701
12	105.15026	159.37068	15.79450	5.85254	4.59106	4.50361	4.51994
13	108.61433	163.41353	15.51380	5.75306	4.30103	4.26040	4.28382
14	106.71552	160.71091	15.39240	5.61500	4.75588	4.65409	4.64520
15	106.64507	160.80193	15.38050	5.64558	4.56820	4.51538	4.51737
16	105.06721	159.27171	15.18860	5.51159	4.47712	4.57499	4.52713
17	105.00593	159.26692	15.17670	5.54364	4.34242	4.46450	4.42943
18	105.21507	159.37346	15.63990	5.71656	4.70757	4.65731	4.64126
19	105.15026	159.37068	15.62800	5.76931	4.66276	4.49526	4.49971
20	103.30121	157.84682	15.00610	5.36106	4.59106	4.59724	4.48231

注:ε 为文献查出的真实值,ε* 为 BP 网络计算值,ε** 为 MLR 计算值

表3 预测结果

No.	Am_2	Am_3	$^4X_{pc}$	^2X_p	ε^*	ε^{**}
21	104.86271	159.56618	16.43163	6.17628	4.40791	4.43
22	104.52242	158.91022	17.06877	6.46045	4.71373	4.66
23	104.54143	159.29367	17.06877	6.46562	4.59599	4.56
24	106.37199	160.95795	16.01772	5.99396	4.34803	4.40
25	104.71265	159.43860	15.78202	5.88128	4.24357	4.28

注:以上五个样本没有标准值,只把 BP 网络和 MLR 的计算结果作对比

3 结论

由前面几个例子可看出,对于这类大量存在于化学中的非线性定量关系,采用本文推荐的 BP 网络结构,第一隐层一般只需用 2 ~ 3 个 Sigmoid 变换节点,第二隐层一般只需 1 个 Sigmoid 变换节点,大多能建立较好的计算模型,且比常规 BP 网络更优,与多元线性回归方法比较,BP 网络的结果更令人满意,而且当对数据中的因果关系不甚明了时,使用 BP 网络免除了模型选择的麻烦,也不用进行逐步回归筛选变量.也正是 BP 网络的这一优点,蕴含着其缺点,即:不象多元回归法的参数那样,我们不能从训练所得的权矢量中确切地获得其物理意义,对于这一不足,需要在以后进一步研究中找到解决办法.

References

1 H. Yoshida, Y. Miyashita, S. Sasaki, *Chemom. Intell. Lab. Syst.*, **1996**, *32*(2), 193.

2 LIU Ping, LIANG Yi - Zeng, WANG Su - Guo, SONG Xin - Hua, YU Ru - Qin, *Acta Chimica Sinica*, **1997**, *55*(4), 386 (in Chinese).

3 F. R. Burden, R. G. Brereton, P. T. Walsh, *Analyst (Cambridge, U. K.)*, **1997**, *122*(10), 1015.

4 J. Tetteh, S. Howells, E. Metcalfe, T. Suzuki, *Chemom. Intell. Lab. Syst.*, **1998**, *41*(1), 17.

5 T. B. Blank, S. D. Brown, *Anal. Chim. Acta*, **1993**, *277*, 273.

6 B. Walczak, W. Wegscheider, *Anal. Chim. Acta*, **1993**, *283*, 508.

7 YAO Yu - Yuan, XU Lu, YUAN Xiu - Shun, *Acta Chimica Sinica*, **1993**, *51*(11), 1041 (in Chinese).

8 LI Hua, XU Lu, SU Qiang, *Chem. J. Chin. Univ.*, **1995**, *16*(5), 688 (in Chinese).

(Ed. CHENG Biao)
(DONG Hua - Zhen)