1995 年 9 月 重庆大学学报 (自然科学版) Vol. 18, №. 5
第 18 卷第 5 期 Journal of Chongqing University (Natural Science Edition) Sep. 1995

(14)
83-89

神经元网络技术和偏最小二乘法在定量分析和构效关系方面的比较*

O655.4

Structure-Activity Relation Study of Fentanyl Derivatives and Multicompontens Quantitative Analysis Using BP and PLS Algorithm

刘信安 Liu Xinan　　陶长元 Tao Changyouang
(重庆大学化学化工学院,重庆,630044)
唐宏志 Tang Hongzhi　　郭力** Guo Li　　李北坡 Li Beipo
(北京总参防化研究院)

A **摘 要** 在光度计/计算机联机系统上利用人工神经网络反向传递算法和偏最小二乘法对五组分有机物混合体系的紫外光谱数据和芬太尼构效关系数据同时进行了定性定量分析、对比和研究。结果表明,反向传递模型具有良好的定性分类能力和定量预测能力;偏最小二乘算法的定性分类效果稍差,但定量预测能力优于反向传递算法。

关键词 神经元网络;偏最小二乘法;模式识别;定量分析
中国图书资料分类法分类号 O655.4; O657.32

神经网络,构效关系,

ABSTRACT Using the back-propagation alogrithm and Partial Least-Squares Method the Structure-Activity Relation of Fentanyl derivatives and quantitative analysis of five organic compounds mixtures were studied on the UV Spectrophotometer-PC/AT On-Line system. Results show that the BP Model has good *qualitative pattern recognition* and *quantitative estimation ability*, and the PLS algorithm has more accurate quantitative ability and less accurate qualitative pattern recognition ability than BP model.

KEYWORDS ANN; PLS; pattern recognition; quantitative analysis

0 引 言

人工神经元网络(Artificial Neural Networks, ANN)技术在化学中的应用大致起始于1989年,Thomsen 等人首先利用神经元网络中的反向传递模型(Back-Propagation Model, BP)

* 收文日期 1994-08-24
** 中国科学院化工冶金研究所工作(北京)

对六个糖类化合物的氢核磁共振谱进行模式识别，获得初步结果[1]. Aoyama 等人将 BP 模型应用于丝裂霉素类药物和芳基丙酰哌嗪衍生物的构效关系(Structure-Activity Relation，SAR)研究，分类和预测结果明显优于自适应最小二乘法[2]. Long 等人利用 BP 模型对再生糖浆中假麻黄碱盐酸盐、吡啶盐酸盐、灭菌苯甲酸钠和羟苯甲酸甲酯进行了紫外光谱定量分析，以及对小麦中蛋白质的近红外光谱进行了定量分析，研究结果表明 BP 模型对非线性响应的复杂体系具有较好的预测能力[3]. 国内从九零年以来，也开始了这方面的研究。石乐明等人利用 BP 模型对四十七种化学杂交剂和九十六种磺酰脲类除草剂进行了 SAR 研究，结果表明 BP 模型具有良好的预测和分类能力[4]. 钟雷鸣利用我们提供的六个氨基酸紫外光谱测试数据，训练了包含六个氨基酸在内的 BP 模型，并对其中的酪氨酸进行了定量的分析，就这一个组分而言，分析效果略好于我们原来用卡尔曼滤波的计算结果[5].

偏最小二乘法(Partial Least-Squares Method，PLS)是一种将函数关系中自变量和因变量各自分解成两个或多个独立变量，并对这些变量进行外相关、内相关和混合相关处理，并从所得到的这些本征量来预测新变量的计量学算法。这种算法最初由 H. Wold 在 1966 年提出并应用于经济学中，1979 年才开始应用于化学[6]. 国内学者对多种计量学算法进行了研究和对比，由于 PLS 在计算中不用求逆，自变量和因变量相互利用了对方的信息，而且在模型训练期间本身包含了各组分间的非线性影响，因此一般都认为这是一种在多组分混合体系定量分析中最好的算法[7~9].

1 理论和算法

1.1 BP 模型理论和算法

目前在化学中应用最为广泛的 ANN 是如图 1 所示的 BP 模型，这种模型采用广义 δ 规则进行学习. BP 模型的第一层为输入层(Input Layer)，经过某种标准化预处理后的特征数据从输入层进入网络中，网络的第二层为隐含层(Hidden Layer)，用来抽取输入模式的特征，第三层为输出层(Output Layer)，网络的分类、预测或者目标函数估计值就从输出层输出。BP 实际是真实神经元网络的一种数学抽象，可以看出，这种三层结构的数学模型对于信号的接受、知识的积累和信息的反馈类似于细胞体、轴突、树突和突触组成的生物神经网络，不同的是，生物神经元网络神经末梢传出冲动信号，通过具有结构可塑性的细胞相互关联特征来学习和遗忘事物。在 ANN 以数据化的方式来接受外来信息、用调节权重的方式来学习，通过这些保留在网络中的权重值来记忆或者预测新的事实。

为了便于阅读和参考，下面给出 BP 模型的具体算法[10]. 根据图 1，其算法可以描述为：设输入层、隐含层和输出层的单元(节点)数分别为 I、J 和 K，其中 $a_1, a_2, \cdots, a_I$ 为输入矢量；$h_1, h_2, \cdots, h_J$ 为隐含层输出矢量；$y_1, y_2, \cdots, y_K$ 为输出矢量，并用 $d_1, d_2, \cdots, d_K$ 表示训练组中各模式的目标输出矢量。输入单元 i 到隐含单元 j 的权重是 V_{ij}，而隐含单元到输出单元 k 的权重是 W_{jk}. 另外，用 θ_k 和 φ_j 分别表示输出单元和隐含单元的阈值。于是，隐含层各单元的输出为：

$$h_j = f(\beta_j) = f\left(\sum_{i=1}^{I} V_{ij} a_i - \varphi_j\right) \tag{1}$$

而输出层各单元的输出是：

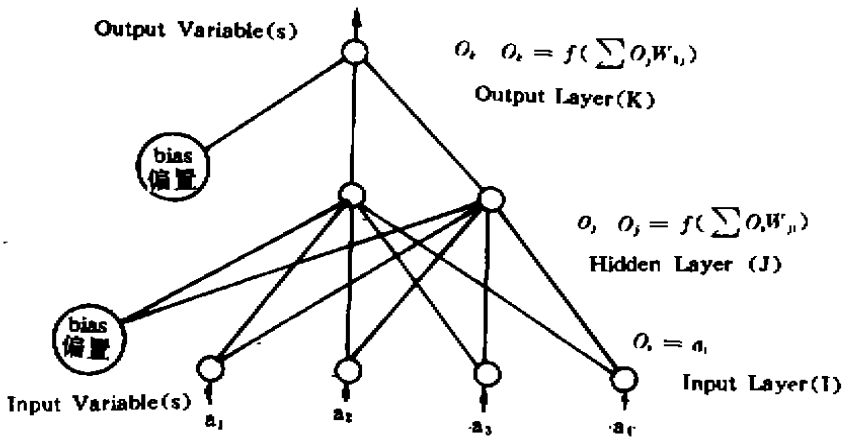

图 1　BP 模型示意图

$$y_k = f(a_k) = f\Big(\sum_{j=1}^{J} W_{jk} h_j - \theta_k\Big) \tag{2}$$

上述表达式中的 $f(x)$ 通常采用形如(3)式的 S 型函数，然后确定作为校正用的训练集数据组，并将各权重 V_{ij}、W_{jk} 以及阈值 φ_j、θ_k 设置成一个小的随机值。

$$f(x) = 1 / (1 + e^{-x}) \tag{3}$$

1) 从训练集数据组中取一个模式 $a_1, a_2, \cdots, a_I$(例如，混合体系在不同波长处测得的吸光度所构成的特征矢量、构效关系中化合物的特征值，如量化参数、官能团和拓扑参数等)加到网络中，而该模式的目标输出矢量(如混合体系中各个组分已知的浓度值矢量，或者构效关系中化合物的宏观指标如镇痛效果等)为 $d_1, d_2, \cdots, d_K$。

2) 利用(1)式计算出一个隐含输出矢量 $h_1, h_2, \cdots, h_J$，再根据(2)式计算出网络的实际输出矢量 $y_1, y_2, \cdots, y_K$。

3) 将输出矢量中的元素 y_k 与目标矢量中的元素 d_k 进行比较，按(4)式计算出 K 个误差项 δ_k，按(5)式计算出隐含层单元的 J 个误差项 e_j，按(6)式和(7)式分别计算出各自的权重调整量，其中 η 就是控制学习速率的步长。

$$\delta_k = (d_k - y_k) y_k (1 - y_k) \tag{4}$$

$$e_j = h_j (1 - h_j) \sum_{k=1}^{K} \delta_k W_{jk} \tag{5}$$

$$\triangle W_{jk}(n) = \eta \delta_k h_j \tag{6}$$

$$\triangle V_{ij}(n) = \eta e_j a_i \tag{7}$$

4) 根据(8)式和(9)式调整权重，式中 $\triangle W_{jk}(n-1)$ 和 $\triangle V_{ij}(n-1)$ 是上次训练时计算出的调整量，而 μ 是一个小于 1 的正常数：

$$W_{jk}(n+1) = W_{jk}(n) + \triangle W_{jk}(n) + \mu \triangle W_{jk}(n-1) \tag{8}$$

$$V_{ij}(n+1) = V_{ij}(n) + \triangle V_{ij}(n) + \mu \triangle V_{ij}(n-1) \tag{9}$$

5) 返回第 1) 步，继续迭代，直到权重 V_{ij} 和 W_{jk} 达到稳定，即所有调整量接近 0。

至此，便完成了 BP 模型的训练，该网络以其计算得到的权重矢量形成一个模式分类器。当我们加入需要进行预测的其它模式时，可根据(1)、(2)和(3)式计算出相应的输出矢量，进而判断(预测)该模式所属的类别。在计算多组分数据和构效关系数之前还必须利用(10)

式进行预处理，使输入的数据在尽可能拉开距离的同时，取值范围又控制在0和1之间[2]，式中 $\tilde{x}_i$ 为第 i 个经预处理后的值：

$$\tilde{x}_i = (x_i - x_{min} + 0.1) / (x_{max} - x_{min} + 0.1) \quad (10)$$

1.2 PLS 理论和算法

如前所述，PLS 算法在对特征函数(如吸光度矩阵、构效关系中的特征参数矢量等进行分解的同时，还对目标函数(如浓度矩阵、活性参数矢量等) 进行分解，并相互交换特征变量(t 和 u 的相互位置)。用 Y 表示 n 个校正集样品中 l 个组分的浓度(或者活性参数) 矩阵，用 X 表示校正集样品在 m 个波长处的吸光度(或者特征参数) 矩阵。PLS 同时将 Y 和 X 分解成两个小矩阵的乘积：

$$X = T_{nxa}P'_{axm} + E \quad (11)$$

$$Y = U_{nxa}Q'_{axm} + F \quad (12)$$

T 和 U 分别为 X 和 Y 的特征变量矩阵(列正交矩阵)，P' 和 Q' 分别为 X 和 Y 的载荷矩阵(行正交矩阵)，E 和 F 分别为 X 和 Y 的残余矩阵，a 为主成分数。在 PLS 中利用 T 对 U 进行回归，并获取用于预测的回归系数 b. 有关 PLS 的详细算法和实现可参阅文献[7].

2 实验和计算

2.1 仪器和试剂

DMS-200 型紫外/可见分光光度计(美国 Varian 公司)，1.0 cm 光程的石英比色皿(日本岛津公司)IBM 286 微机(联机用)，IBM386 DX-33 微机(供计算时用，装有数学协处理器 387 DX-33)Turbo C++3.0(美国 Borland 公司)，Neural Works-2 软件包(NeuralWare 公司).

药品：α 萘酚(α NAP)、α 萘胺(α NAA)、2,7—二羟甲基萘(DHN)、2,4—二甲氧基苯甲醛(DMO)[英国进口分装]和水杨酸甲酯(MSA)，溶剂为 95%的乙醇，除注明外，其余均为北京化工厂产品，分析纯。准确称取适量上述样品并用乙醇稀释至所需浓度的工作贮备液待用。

2.2 实验和计算结果

表 1 PLS 方法分析五组分混合物的结果 (μg/mL)

样品号	加入量					测试量				
	α NAP	α NAA	DHN	DMO	MSA	α NAP	α NAA	DHN	DMO	MSA
1	0.960	0.845	0.844	1.527	2.309	0.971	0.846	0.841	1.508	2.290
2	0.576	0.845	1.182	0.916	3.233	0.603	0.860	1.175	0.882	3.210
3	1.344	0.507	0.506	1.527	3.233	1.346	0.523	0.501	1.503	3.217
4	0.576	1.183	1.182	0.916	2.309	0.593	1.215	1.173	0.899	2.253
5	0.768	1.352	0.675	1.832	2.771	0.782	1.396	0.659	1.846	2.722
6	1.152	1.014	1.013	1.221	1.847	1.163	1.035	1.006	1.208	1.802

按正交方式 $L_{16}(4^5)$配制上述五个有机物混合而成的校正集共十六个标准样品。这样的标准混合试样在经过处理后，所得到的主因子数 a 正好等于该混合试样中的组分数。具体样正样的配制和数据参见文献[7]. 用 PLS 算法和 BP 算法估算得到的未知混合样的数据如表 1 和表 2 所示。用 PLS 计算芬太尼的构效关系与 BP 相似，但是，在计算多组分混合体系时，

PLS 标准化原始数据的方法为平均中心化法[7]，而不是(10)式。

表 2　BP 模型分析五组分混合物的结果　　(μg/mL)

样品号	加入量					测试量				
	α NAP	α NAA	DHN	DMO	MSA	α NAP	α NAA	DHN	DMO	MSA
1	0.960	0.845	0.844	1.527	2.309	0.885	0.802	0.838	1.486	2.279
2	0.576	0.845	1.182	0.916	3.233	0.597	0.790	1.198	0.959	3.518
3	1.344	0.507	0.506	1.527	3.233	1.372	0.541	0.538	1.538	3.398
4	0.576	1.183	1.182	0.916	2.309	0.602	1.190	1.216	0.971	2.345
5	0.768	1.352	0.675	1.832	2.771	0.827	1.436	0.595	1.873	2.669
6	1.152	1.1014	1.013	1.221	1.847	1.106	1.008	1.067	1.178	1.949

表 3　十种芬太尼衍生物的结构和活性　(小白鼠试验)

芬太尼衍生物		取代基				镇痛用量
编号	代码名	R_1	R_3	R_4	R_5	
1	1379	$phCH_2CH_2-$	$-COOCH_3$	ph	△	8.06
2	1505	$CH_3=CHCH_2CH$	$-COCH_2$	ph	Et	301
3	1410	$phCH_2CH_2-$	$-CH_2OCH_3$	ph	Et	21.3
4	1501	$phCH_2CH_2-$	$-COCH_4$	ph	Et	8.03
5	1502	$CH_3C(CH_3)=CHCH(CH_3)CH_3$	$-COCH_4$	ph	Et	10.9
6	1503	$CH_3C(CH_3)=CHCH_2CH_2-$	$-COCH_2$	ph	Et	17.9
7	1504	$-CH_2CH_2-$	$-CHOCH_3$	ph	Et	21.3
8	1506	$-CH_2CH_2-$	$-COCH_3$	ph	Et	11.4
9	F6420	$phCH_2CH_2-$	$-H$	ph	Et	166
10	R31833	$phCH_2CH_2-$	$-COOCH_3$	ph	Et	13.0

构效关系计算用化合物：合成十种芬太尼衍生物，并用小白鼠做动物实验。这十种芬太尼的名称、取代基和生物效应如表 3 所示，芬太尼母体结构式如图 2 所示。用 PLS 算法和 BP 算法计算得到的构效关系数据如表 4 所示，进行标准化处理的芬太尼化合物构效数据列在表 5 中，对小白鼠镇痛半量低于 30 nmol/kg 以下的定为有效芬太尼，其期望模式，即 PB 算法中的 d 矢量为(1　0)，其余的定为无效芬太尼，其期望模式为(0　1)。BP 模型输入节点/隐含层节点/输出节点＝6/12/2，偏值 Bias 为 1.000，模型训练次数为 1500 次时已经没有显著误差。我们采用 Borland C＋＋3.0 编制的 BP 程序所计算出的结果和 NeuralWare 公司的 Neural Works-2 软件包计算出的结果完全一致，学习速率 η 为 0.300，阈值 φ、θ 由系统自动设置为一个小于 0.20 的正值随机数。BP 定量预测五组分混合体系所采用的参数为：输入节点/隐含层节点/输出节点＝82/20/5，偏值 Bias 为 0.30，迭代次数为 8000。BP 程序中采用了指

针类型变量，因此源程序编译时必须设置为大模式或者巨模式。

表 4 BP 模型和 PLS 模型计算 10 种芬太尼的构效关系结果

组号	测试样号	期待模式	预测结果	组号	测试样号	期待模式	预测结果
1	1—2	1 0	0.9980 0.021	1	1—2	1 0	0.7305 0.2670
		0 1	0.0955 0.9070			0 1	0.0740 0.9408
2	3—5	1 0	0.9975 0.0025	2	3—5	1 0	0.9077 0.1055
		1 0	0.9986 0.0012			1 0	0.8463 0.2077
		1 0	0.9980 0.0012			1 0	0.4408 0.5701#
3	6—8	1 0	0.8702 0.1164	3	6—8	1 0	0.9706 0.0997
		1 0	0.9975 0.0026			1 0	0.7633 0.3011
		1 0	0.9938 0.0064			1 0	0.6018 0.4276
4	9—10	0 1	0.0827 0.9074	4	9—10	0 1	0.7509 0.3018#
		1 0	0.9986 0.0014			1 0	0.9633 0.0874
	BP 模型的预测正确率 100%				PLS 模型的预测正确率 80%		

*(1 0) 为有效化合物模式，(0 1)为无效化合物模式，# 为预测错误的结果

表 5 十种芬太尼衍生物的标准化特征参数和目标参数

样号	量化指数		分子连接指数				目标参数
	DAN	DPN	X(0)	X(1)	X(2)	X(3)	
1	0.4753	0.5206	1.0000	0.9809	1.0000	0.9342	0.0005
2	0.3854	0.4009	0.8321	0.7317	0.5863	0.5770	0.0456
3	0.0099	0.4022	0.7606	0.6885	0.5562	0..6374	0.0456
4	0.1951	0.6734	0.9724	1.0000	0.7125	0.5725	0.0003
5	0.1553	0.6711	0.6539	0.4363	0.5014	0.2830	0.0099
6	0.0820	0.4465	0.9399	0.9733	0.8629	1.0000	0.0339
7	0.4775	1.0000	0.0345	0.0445	0.0459	0.0631	1.0000
8	0.2592	0.1083	0.7174	0.7570	0.7511	0.8839	0.0116
9	1.0000	0.8684	0.1282	0.3283	0.1762	0.1681	0.5427
10	0.2498	0.4098	0.9016	0.7388	0.6030	0.7415	0.0171

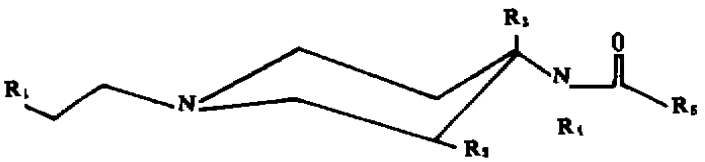

图 2 芬太物衍生物的结构 （对全部芬太尼都有 $R_3=H$）

3 讨 论

从上述计算结果可以看出：BP 和 PLS 均能够同时定量分析多组分混合体系和进行构效

关系的模式识别，在这一定程度上说明这两种算法的适用性。但是，仔细分析这两种方法的原理和计算结果，可以看出，BP 在构造预测模型时，利用数据训练时建立在隐含层中的某种相对简单的权重关系来预测未知的模式，这就非常适合于那种因果关系不明确、知识背景不清楚和推理规则不确定的问题。在芬太尼的构效关系识别中，特征参数与目标参数之间不存在明显的对应关系，很难建立一种非线性映射方程，所以 BP 的这种独特的模型构造方式就可以有效地解决这类问题。但是，BP 预测多组分混合体系的结果又不如 PLS 算法，BP 模型的这种局限性还可以从文献[3,5]的结论中可以看出。当 BP 模型利用单一信号源（如光度分析中不同波长处的吸光度矢量）作为其特征参数矢量，并对多个组分（目标参数）进行定量分析时，其结果难以和 PLS 这类以某种非线性映射方程所预测的结果相比。我们推测，这可能与 BP 算法容易陷入局部最小值、在最优解中不容易稳定以及其计算结果本身就具有某种非精确性有关。同理，PLS 算法在多组分定量分析中由于充分和相互利用了自变量和因变量之间的信息，并借此克服了基体效应、溶剂效应和组分间影响等非线性的干扰，这样预测结果应该是相当好的。但是，在进行构效关系识别预测时，由于此时特征参数和目标参数之间难以建立准确的描述方程、或者其对应关系是模糊的、甚至不存在对应关系，而在这种情况下用基于相关性原理的方法，如 PLS、因子分析和主成分分析等算法所建立的模型则可能效果不理想。

综上所述，尽管以后还会不断出现新的理论和新算法，这些技术可以更好地解决某些特定领域的问题，但一般不会成为某种普遍适应的化学计量学理论。

参 考 文 献

1　Thomsen J U, Meyer B. Pattern Recogntion of the 1H NMR Spectra of Surgar Alditols Using a Neural Network, J. Magn. Reson. 1989(84): 212～217

2　Ayoama T, Suzuki Y, Ichikawa H. Nerual Networks Applied to Pharmaceutical Problems, Med. Chem. 1990(33): 22～32

3　Long R J, Gregoriou G V, Gempline J P. Neural Networks Applied to Multi-component Analysis. Anal. Chem. 1990(62): 2122～2134

4　石乐明，刘信安．神经元网络 BP 算法研究磺酰类除草剂的构效关系，见：复旦大学化学系，第三届全国化学类研究生学术报告会论文集．上海：复旦大学出版社，1991

5　钟雷鸣．六种氨基酸混合物溶液的紫外光谱的人工神经元网络定量分析研究．生物物理学报．1992, 8(4): 706～710

6　Kowalski B R. Partial Least-Squares Path Modelling With Latent Variables, Anal. Chim. Acta, 1979, (112): 417～430

7　刘信安，石乐明．五种有机化合物混合体系的偏最小二乘法定量同时分析．防化研究．1990, (1): 25～30

8　李梦龙．偏最小二乘法定量分析研究．分析实验室．1991, 9(6): 66～68

9　倪永年．偏最小二乘法在分析化学中的应用．分析化学．1990, 18(4): 35～40

10　中国科学技术大学生物医学工程跨系委员会编．神经网络及其应用．北京：中国科学技术大学出版社，1992, 98～108