QSAR中ANN用于研究变量选择方法的回顾和比较

杨 蕾[1],王 鹏[2],王 虹[2,3],蒋益林[2]

(1. 哈尔滨工业大学 基础交叉与科学技术研究院,150001 哈尔滨,leiyang84@vip.sina.com;2. 哈尔滨工业大学市政环境工程学院,150090 哈尔滨;3. 黑龙江大学 化学化工与材料学院,150080 哈尔滨)

摘 要: 在定量构效关系(QSAR)建模中,人工神经网络(ANN)模型预测能力较好,但并不能提供化合物结构变量对活性影响的更多信息,因而被认为是个黑箱模型. 以35种硝基化合物对黑呆头鱼96 h的生物毒性为例,首次回顾和比较6种用于研究网络输入变量对输出相对贡献大小的方法. 结果表明:ANN中引入变量选择的方法,大大增强了QSAR模型的解释能力,其中偏微分方法能得出最为全面准确的结果,其次为轮廓图方法. 扰动法和权重法对输入参数能实现较好的分类,但过于简化且方法不稳定;而传统的逐步回归法结果最差.

关键词: 定量构效关系;人工神经网络;硝基芳烃;偏微分方法;权重法;扰动法;轮廓图法

中图分类号: U214.1 **文献标志码:** A **文章编号:** 0367-6234(2011)10-0060-07

Review and comparision of methods to study the contribution of variables in artificial neural network models for QSAR study

YANG Lei[1], WANG Peng[2], WANG Hong[2,3], JIANG Yi-lin[2]

(1. Academy of Fundamental and Interdisciplinary Science, Harbin Institute of Technology, 150001 Harbin, China, leiyang84@vip.sina.com; 2. School of Municipal and Environmental Engineering, Harbin Institute of Technology, 150090 Harbin, China; 3. School of Chemistry and Materials Science, Heilongjiang University, 150080 Harbin, China)

Abstract: Although Artificial Neural Network (ANN) shows superior predictive power in the study of quantitative structure-activity relationship (QSAR), it has been labeled as a "black box" because it provides little explanatory insight into the relative importance of the independent variables. In this paper, as an example of toxicity of 35 nitro-aromatics on fathead minnow, six methods which could give the relative contribution and/or the contribution profile of input factors were reviewed and compared. The Partial Derivative method was found to be the most useful as it gave the most complete results, followed by the Profile method that gave the contribution profile of input variables. The Perturb method allowed a good classification of input parameters as well as the Weights method that had been simplified, but these two methods lacked stability. Finally, the classical stepwise methods gave the poorest results.

Key words: QSAR; artificial neural network; nitro-aromatic; partial derivative method; weight method; perturb method; profile method

在环境化学领域,QSAR是进行有毒化学品生态风险评价的重要手段之一. 目前,QSAR研究中常用的方法有多元线性回归分析(MLR)、人工神经网络(ANN)等,后者在处理非线性关系方面有着非常强大的功能,而化合物的结构和毒性之间大多存在非线性的关系,使ANN成为QSAR研究的热点[1-3].

由输入变量ANN便能预测输出变量,但网络内部的作用机理往往被忽略,因而被认为是个黑箱模型. 近年来,研究者提出了许多方法来描述变量

收稿日期: 2010-05-12.

基金项目: 哈尔滨市科技创新人才(学科带头人)研究基金资助项目(2009RFXXN047);哈尔滨工业大学科研创新基金资助项目(HIT.NSRIF.2009081);哈尔滨工业大学创新团队计划(有机材料).

作者简介: 杨 蕾(1966—),女,教授,硕士生导师;
王 鹏(1957—),男,教授,博士生导师.

在神经网络 QSAR 模型中所起的作用,然而大多数方法被用来消除不相关变量,因而被称为修剪方法[4-6]. 简单地说,修剪算法就是由具有高度联结的网络(i. e. 神经元之间有许多连接)开始,逐步移除弱连接,或当连接移除时,网络误差无明显变化的连接. 事实上,在 QSAR 建模中,不仅需要好的预测能力,还要了解每个输入变量对输出的相对贡献大小. 本文回顾和比较了用 ANN 建模并解释 QSAR 模型的 6 种方法,这些方法被用来确定每个输入变量对输出的贡献,因而不是修剪算法.

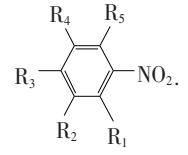
以 35 种硝基化合物对黑呆头鱼 96 h 的生物毒性为研究对象,探讨了 ANN 中引入变量选择方法后,QSAR 模型的解释能力. 结果表明,偏微分方法分析所建模型能得出最为全面准确的结果,模型具有良好的预测和解释能力;其次为分布图方法. 扰动法和权重法对输入参数能实现较好的分类,但过于简化且方法不稳定;传统的逐步回归法所得结果最差.

1　建模数据库及方法

以文献[7]报道的 35 种硝基芳烃化合物对黑呆头鱼的 96 h 半致死浓度 c_{150}(mmol/L)数据作为研究对象,来讨论和比较用于分析和解释人工神经网络 QSAR 模型的不同方法. 该硝基芳烃主要由具有不同硝基取代基的甲苯、苯胺和苯酚、卤代苯组成,具体结构和活性见表 1. 数据 $\lg 1/c_{150}$ 见文献[7].

表 1　硝基芳烃化合物结构及其毒性

序号	组成	R1	R2	R3	R4	R5	$\lg 1/c_{150}$(obs.)
1	2 - 硝基甲苯	CH_3	H	H	H	H	3.57
2	3 - 硝基甲苯	H	CH_3	H	H	H	3.63
3	4 - 硝基甲苯	H	H	CH_3	H	H	3.76
4	1,2 - 二硝基苯	NO_2	H	H	H	H	5.45
5	1,3 - 二硝基苯	H	NO_2	H	H	H	4.38
6	1,4 - - 二硝基苯	H	H	NO_2	H	H	5.22
7	2,3 - 二硝基甲苯	NO_2	CH_3	H	H	H	5.01
8	2,4 - 二硝基甲苯	CH_3	H	H	NO_2	H	3.75
9	2,5 - 二硝基甲苯	CH_3	H	NO_2	H	H	5.15
10	2,6 - 二硝基甲苯	CH_3	NO_2	H	H	H	3.99
11	3,4 - 二硝基甲苯	NO_2	H	CH_3	H	H	5.08
12	3,5 - 二硝基甲苯	H	CH_3	H	NO_2	H	3.91
13	1,3,5 - 三硝基苯	H	NO_2	H	NO_2	H	5.29
14	硝基苯	H	H	H	H	H	3.02
15	2 - 硝基苯胺	NH_2	H	H	H	H	3.70
16	2,4 - 二硝基苯胺	NH_2	H	H	NO_2	H	4.07
17	4 - 硝基苯酚	H	H	OH	H	H	3.36
18	4 - 氟硝基苯	H	H	F	H	H	3.70
19	2,4,6 - 三硝基甲苯	CH_3	NO_2	H	NO_2	H	4.88
20	2,3,6 - 三硝基甲苯	NO_2	CH_3	NO_2	H	H	6.37
21	2 - 甲基 - 3 - 硝基苯胺	CH_3	NH_2	H	H	H	3.48
22	2 - 甲基 - 4 - 硝基苯胺	H	CH_3	NH_2	H	H	3.24
23	2 - 甲基 - 5 - 硝基苯胺	H	NH_2	CH_3	H	H	3.35
24	2 - 甲基 - 6 - 硝基苯胺	NH_2	CH_3	H	H	H	3.80
25	3 - 甲基 - 6 - 硝基苯胺	NH_2	H	CH_3	H	H	3.80
26	4 - 甲基 - 2 - 硝基苯胺	NH_2	H	H	CH_3	H	3.79
27	3 - 硝基 - 4 - 羟基苯胺	OH	H	H	NH_2	H	3.65
28	4 - 甲基 - 3 - 硝基苯胺	CH_3	H	H	NH_2	H	3.77
29	2,4 - 二硝基苯酚	OH	H	H	NO_2	H	4.04
30	2 - 甲基 - 3,5 - 二硝基苯胺	CH_3	NH_2	H	NO_2	H	4.14
31	2 - 甲基 - 3,6 - 二硝基苯胺	CH_3	NH_2	NO_2	H	H	5.34
32	3 - 甲基 - 2,4 - 二硝基苯胺	NH_2	H	H	NO_2	CH_3	4.26
33	3 - 甲基 - 2,6 - 二硝基苯胺	NH_2	NO_2	CH_3	H	H	4.21
34	4 - 甲基 - 2,6 - 二硝基苯胺	NH_2	H	NO_2	CH_3	H	4.18
35	4 - 甲基 - 3,5 - 二硝基苯胺	CH_3	NO_2	H	NH_2	H	4.46



在 QSAR 研究中,用于描述化合物的结构参数有很多,包括拓扑的、量子的、实验值等[7-9]. 本文在 Hall[7] 和黄庆国等[10] 研究硝基芳烃类化合物的基础上,利用 HyperChem 6.03 软件和自编的

C + + 软件分别计算了 7 种量子化学参数和 5 种自相关拓扑指数来表征化合物的结构,具体见表 2(为表述方便,以下选择变量方法中都以表中的编号来代替变量).

表 2 硝基芳烃化合物的量子化学参数和拓扑指数

编号	变量	类型	意义
1[a]	Q_C	自变量	苯环上与硝基相连的 C 原子的净正电荷数
2[a]	Q_N	自变量	苯环上硝基中 N 原子的净正电荷数
3[a]	Q_{NO2}	自变量	苯环上所有硝基中 N 原子的最大净正电荷数
4[a]	F_H	自变量	分子的 mol 生成热/(J · mol^{-1})
5[a]	M	自变量	分子偶极矩/(C · m)
6[a]	$-E_{LUMO}$	自变量	分子最低未占用轨道能量/eV
7[a]	$-E_{HOMO}$	自变量	分子最高占用轨道能量/eV
8[b]	A_1	自变量	分子中邻接原子体积信息总和
9[b]	A_2	自变量	分子中间位原子体积信息总和
10[b]	B_1	自变量	分子中邻接原子电子信息总和
11[b]	B_2	自变量	分子中间位原子电子信息总和
12[b]	C_1	自变量	原子间共价连接信息总和
13	$\lg 1/c_{L50}$	因变量	黑呆头鱼 96 h 半致死浓度/(mmol · L^{-1})

注:a. 由 HyperChem 6.03 软件计算而得;b. 描述符定义见文献[11].

2 方 法

2.1 人工神经网络结构

目前,在 QSAR 建模中,多层前馈性人工神经网络结构得到了最为广泛的应用,其依据误差反向传播算法训练而得. 本文采用较为普遍的 3 层网络结构(其中输入层 12 个神经元,隐含层 5 个神经元和输出层 1 个神经元),具体结构见图 1.

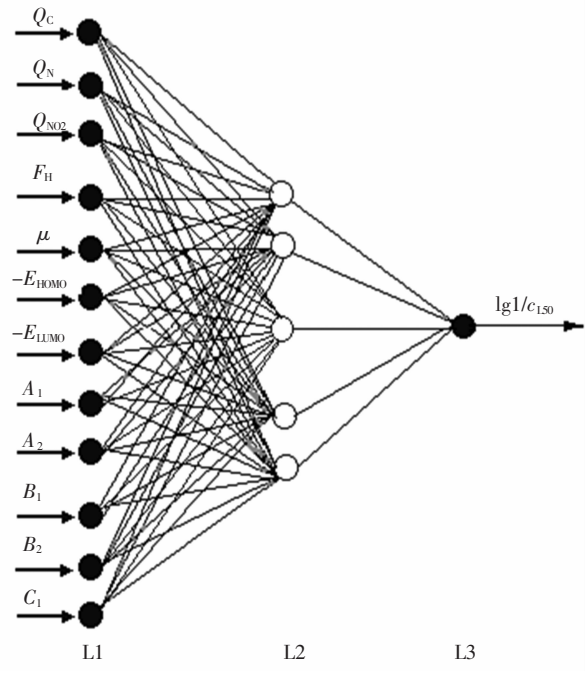


图 1 人工神经网络结构

建模过程主要分两步:

1)随机选择 75% 的化合物作为训练集,25% 的化合物作为测试集,利用训练数据集来训练模型,测试集来验证模型,反复多次来确定最佳的网络结构[12];

2)在整个数据集上,利用第一步所获得的网络最佳结构进行 QSAR 建模,采用不同方法研究输入变量对网络输出,即化合物的生物毒性的相对贡献大小,并分析解释 QSAR 模型.

2.2 偏微分法

由该法可以获得两种结果:一是每个输入变量的微小变化导致网络输出变化的偏微分图;二是每个输入变量相对输出的贡献大小排序.

为了获得输入变量的微小变化导致输出变化的偏微分图,计算输出对输入变量的偏微分. 以具有 n_i 个输入节点、n_h 个隐含节点和 1 个输出节点,第 j 个样本的输出 y_j 关于输入 x_j($j = 1 \cdots N$,N 为样本总数)的偏微分为

$$d_{ji} = S_j \sum_{h=1}^{n_h} w_{ho} I_{hj} (1 - I_{hj}) w_{ih}. \qquad (1)$$

其中:S_j 为输出神经元对其输入的偏微分;I_{hj} 为第 h 个神经元的响应;w_{ho} 和 w_{ih} 分别为输出与第 h 个隐含层神经元、第 i 个输入神经元与第 h 个隐含神经元之间的连接权重.

由式(1)可获得一系列输出相对输入变量的偏微分图,能直接评价每个输入对输出的影响. 例如,偏微分为负,对于研究变量,输出随输入的增大而减小.

另外,对于整个数据集,由偏微分方法可得到 ANN 输出对每个输入变量的相对贡献大小,具体

计算如下:

$$SSD_i = \sum_{j=1}^{N} (d_{ji})^2. \quad (2)$$

其中: SSD_i 为第 i 个变量对所有化合物毒性网络输出的偏微分平方之和, SSD 值越大,其对网络输出,即对化合物毒性的影响最大.

2.3　扰动法

该法旨在评价每个输入的微小变化对 ANN 输出的影响. 算法首先调整一个变量的值,而保持其他变量不变,同时记下每个输入对输出的响应. 输入变量变化对输出影响最大的变量被视为对网络输出影响最大[13],为最重要的变量.

基本思想如下:假定 x_i 为选定变量,δ 为变化量,则 x_i 的变化可表示为 $x_i = x_i + \delta$. 一般 δ 可选定变量值的 10% ~50% 不等,这样便可获得按重要性排序的输入变量分类.

2.4　权重方法

该法通过分割网络连接权重来确定输入变量的相对重要性,是由 Garson[14] 首先提出的. 方法主要涉及两部分:一是按隐含层节点分割隐含 - 输出层间连接权重;二是按输入层节点划分输入 - 隐含层间连接权重. 本文对此方法进行了简化,而所得结果一致,具体如下:

1)对于隐含神经元 h, 用其输入 - 隐含层连接权重的绝对值除以所有输入 - 隐含层间的连接权重绝对值之和,即

$$Q_{ih} = \frac{| W_{ih} |}{\sum_{i=1}^{n_i} | W_{ih} |}. \quad (3)$$

2)对于输入神经元 i, 用每个隐含神经元与其连接的输入所获得的 Q_{ih} 之和,除以所有隐含神经元与其连接的输入所获得的 Q_{ih} 之和,再乘以 100 便可获得每个输入变量对所有样本输出,权重分布贡献的相对重要性(Relative Importance),即

$$W_{\mathrm{RI}}(\%)_i = \frac{\sum_{h=1}^{N_h} Q_{ih}}{\sum_{h=1}^{N_h}\sum_{i=1}^{N_i} Q_{ih}} \times 100. \quad (4)$$

2.5　轮廓图法

Lek 等[15]首先提出了该方法,其主要思想是构建隶属于所有输入变量范围的假定矩阵,同一时刻固定其他变量的值,在假定矩阵范围内连续变化某个输入变量来观察网络输出的变化. 详细地说,就是每个输入变量在区间范围内被分成等间距的一系列值,该间距被称为标度. 其他变量被依次固定在不同倍数的标度上,一般取 5 个点,分别是最小值、1/4 区间、1/2 区间、3/4 区间和最大值上. 对于每个输入变量,根据不同的取值,便可得到输出变量的分布图. 由分布曲线图(见图 3)可以直观地看到随着输入变量的递增,网络输出变量的变化趋势和垂直波动范围,波动范围越大,表明该变量越重要.

本文在利用轮廓图法研究输入变量对输出贡献的过程中,分别将输入变量的最大值和最小值区间范围分成 12、24、48、96、144 和 192 标度. 图 3 代表了 24 标度的轮廓图. 事实上,不管采用什么标度,该方法相当稳定,不同标度下变量的轮廓图具有相似的形状,唯一不同的是标度越大,变量的轮廓图越精细.

2.6　传统逐步回归法

该法主要包括逐步地增加或删除一个输入变量来考察对输出结果的影响,根据网络输出均方差(MSE)的变化,输入变量便能按照重要性进行排序. 例如在逐步减少输入变量个数的过程中,引起均方差最大程度增大的变量,便是问题空间最重要的变量[16];反之,在逐步增加输入变量的过程中,引起均方差最大程度减小的变量,便是问题空间最重要的变量. 本文利用这两种逐步回归建模方法来评价 12 个输入变量的影响,分别可以获得变量之间的相对重要性排序.

1) 前进法:首先产生 12 个模型,每个模型仅包含一个输入变量,产生最小误差的变量 x 最为重要,并参与下一步建模;接着产生 11 个模型,每个模型由 x 和剩余变量中的任意一个组成,这个过程反复进行,直到所有的变量都进入模型. 网络模型中输入变量的出现排序即为它们对网络输出的相对重要性关系;

2) 后退法:首先产生 12 个模型,每个模型由 11 个变量组成,如果模型中不包含变量 x 引起网络输出的最大误差,则该变量 x 最为重要;接着产生 11 个模型,每个 ANN 模型由 10 个输入变量组成. 这个过程反复进行,直到模型中剩下一个变量为止. 网络中输入变量的消除顺序即为它们对网络输出的重要性排序.

3　结果与讨论

3.1　ANN - QSAR 模型的预测能力

由 2.1 给出的建模过程,最终确定最佳网络结构为 12 - 5 - 1(见图 1). 对于化合物学习样本集,步骤 1(见 2.1)的结果为 $R^2 = 0.923$ ($P < 0.01$);对于测试样本集,其结果为 $R^2 = 0.930$ ($P < 0.01$). 这表明该网络结构可以应用于

步骤 2(见 2.1),即分析结构参数对所有化合物毒性的相对重要性. 所有样本参与建模,结果为 $R^2 = 0.938$ ($P < 0.01$),验证了该网络模型的预测能力.

3.2 不同方法选择变量结果及其比较

3.2.1 偏微分方法

由偏微分方法可以获得一系列输入变量对输出的偏微分图. 图 2 给出了 Q_{NO2} 对硝基芳烃化合物毒性 $\lg 1/c_{L50}$ 的偏微分图. 可以看出,其偏微分值都为正,且随着 Q_{NO2} 的增大,偏微分接近于 0,表明随着 Q_{NO2} 的增大,$\lg 1/c_{L50}$ 也随之增大并最终达到一个稳定值,类似情况的还有变量 Q_C、Q_N、F_H、$-E_{LUMO}$、A_2,其中 Q_N 无明显的规律性;此外变量 μ、$-E_{HOMO}$、A_1、B_1、B_2、C_1 对硝基芳烃化合物毒性 $\lg 1/C_{L50}$ 的偏微分值大多为负,其中 μ、B_1 无明显的规律性.

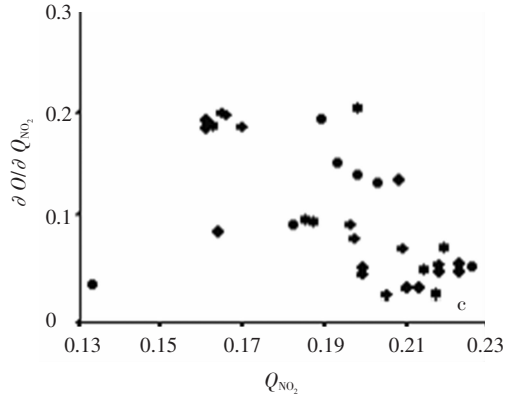


图 2 ANN 网络输出 $\lg 1/c_{L50}$ 对变量 Q_{NO2} 的偏微分图

3.2.2 轮廓图法

图 3 代表了 24 标度的轮廓图,可以看出,网络输出 $\lg 1/c_{L50}$ 随 Q_{NO2}、B_2 和 $-E_{HOMO}$ 的增大有明显变化,其中 Q_{NO2} 在整个取值范围内对网络输出影响最大,是最重要的变量. 另外,$\lg 1/c_{L50}$ 随 Q_{NO2} 的增大而增大,且当 Q_{NO2} 增大到一定程度时,$\lg 1/c_{L50}$ 保持恒定,而 B_2 和 $-E_{HOMO}$ 的增大将导致 $\lg 1/c_{L50}$ 减小,这与偏微分方法所得结果一致. 由轮廓图法获得的变量间的相对重要性关系见表 3.

3.2.3 权重法和扰动法

图 4(a)给出了由偏微分图得到的输入变量对输出的相对贡献图,可以看出,该方法非常的稳定且有较小的置信度区间,Q_{NO2} 是化合物结构变量中对生物毒性贡献最大的变量(45.2%),其次是 $-E_{LUMO}$(16.1%)和 B_2(11.3%). 由权重方法获得输入变量对输出的相对贡献率见图 4(b). 与偏微分方法比较,其置信度区间更大,因而稳定性较差. 由图可见,Q_{NO2} 变量对网络输出的贡献率最大,其次为 $-E_{LUMO}$、Q_C、F_H 和 B_2,而其他变量贡献率差异不大.

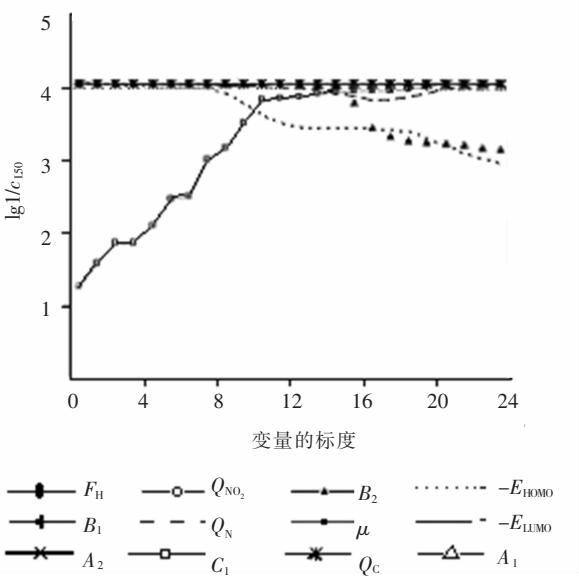


图 3 12 个参数变量对网络输出 $\lg 1/c_{L50}$ 的轮廓图

图 4(c)给出了利用扰动法($\delta = 50\%$)获得的输入对输出的相对贡献分布图. 由图可见,Q_{NO2} 变量对网络输出的贡献率最大,其次为 Q_C、$-E_{LUMO}$ 和 B_2. 该方法同样也不够稳定,因为有些变量如 Q_N 和 A_2、$-E_{HOMO}$ 和 B_1 等之间对输出的贡献差异并不明显.

3.2.4 逐步回归法

逐步回归方法分为前进法和后退法,获得的变量之间的相对重要性排序结果见表 3,可以看出,除了最重要的变量 Q_{NO2},两种方法对变量重要性分类结果不尽相同. 根据前进法,Q_{NO2} 之后依次为 F_H、Q_C、$-E_{HOMO}$,而后退法依次为 Q_C、$-E_{LUMO}$、Q_N.

3.3 ANN – QSAR 模型的解释性能力

本文采用 12 – 5 – 1 的 3 层神经网络结构,对硝基芳烃对黑呆头鱼生物毒性进行 QSAR 建模,并将各种选择变量方法作用于模型,来研究不同变量对网络输出,即生物毒性的相对重要性,进而来阐释硝基芳烃化合物的作用机理,提高网络模型的解释能力.

早期的研究表明[17],硝基芳烃化合物是一类重要的遗传毒性化合物,其致毒机理为:苯环上硝基 N 原子的亲电中心与生物组织中作为亲核中心的 DNA 分子相互反应引起的.

结构参数 Q_{NO2} 是用来表征苯环上所有硝基中 N 原子的最大净正电荷数,由表 3 可以看出,所有变量选择方法都得出 Q_{NO2} 是影响化合物毒性的最重要参数,这正好验证了文献[17]所报道的该类化合物的致毒机理.

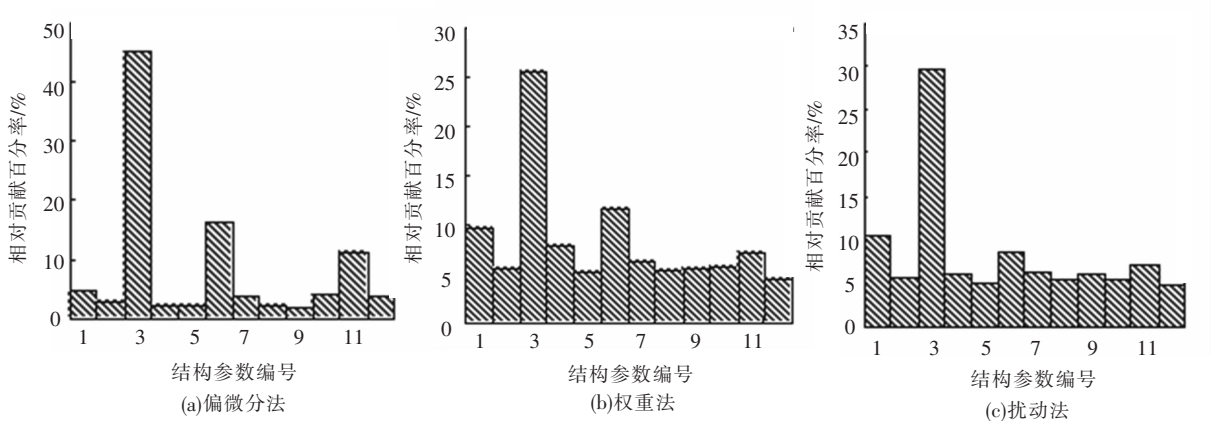


图 4　不同方法获得的 12 个结构参数对 ANN 输出的相对贡献分布图

表 3　采用不同方法对输入变量相对重要性分类结果

变量	前进法	后退法	偏微分法	扰动法	权重方法	轮廓图法
Q_C	3	2	4	2	3	5
Q_N	6	4	8	8	8	9
Q_{NO2}	1	1	1	1	1	1
F_H	2	8	10	7	4	6
M	10	12	11	11	11	12
$-E_{LUMO}$	5	3	2	3	2	4
$-E_{HOMO}$	4	7	7	5	6	3
A_1	9	11	9	9	10	11
A_2	11	10	12	6	9	8
B_1	12	5	5	10	7	7
B_2	8	9	3	4	5	2
C_1	7	6	6	12	12	10

比较表 3 和图 4 可以看出,$-E_{LUMO}$、Q_C、B_2 是另外 3 个影响硝基芳烃毒性的重要结构参数. 其中 $-E_{LUMO}$ 表示分子最低未占用轨道能量,其值愈大,接受电子的能力越强,化合物对黑呆头鱼毒性也越大. 这表明 $-E_{LUMO}$ 与生物毒性之间正相关,这与偏微分法和轮廓图法所得的结论一致. 可以认为,当化合物的亲电中心与 DNA 分子的亲核中心发生反应时,$-E_{LUMO}$ 越大则化合物越容易接受电子发生反应,因而化合物的生物毒性越强.

Q_C 代表苯环上与硝基相连的 C 原子的净正电荷数,其值越大,则与之相邻的硝基 N 原子亲电中心越强,越容易与 DNA 分子反应,因而 Q_C 与化合物毒性之间正相关,这与偏微分法的结论一致.

自相关拓扑指数 B_2 代表分子中间位原子电子信息总和,可以认为取代基电子相互作用同样影响了 DNA 分子的反应活性. 供电基团,如 NH_2、CH_3、OH (见表 1)可能离域了硝基上 N 原子的正电荷,提高了反应的活化能,因而化合物的毒性与 B_2 负相关,这与偏微分和轮廓图方法所得结论一致.

利用不同的变量选择方法,剩余变量的相对贡献大小排序有较大的出入,这主要是由方法的局限性引起的. 不管怎样,ANN 中引入选择方法有助于识别影响问题空间的主因子,提高模型的解释能力.

4　结　论

1)在 ANN 中引入不同的变量选择方法,可大大增强 QSAR 模型的解释能力,其中偏微分方法能得出最为全面准确的结果,其次为轮廓图方法. 扰动法和权重法对输入参数能实现较好的分类,但过于简化且方法不稳定;而传统的逐步回归法结果最差.

2)硝基芳烃对黑呆头鱼毒性的 QSAR 模型中,选择方法识别的重要变量包括 Q_{NO2}、$-E_{LUMO}$、Q_C 和 B_2,它们能准确揭示化合物的致毒机理,从而证明了变量选择方法的有效性.

参考文献:

[1]WU J H, MEI J, WEN S X, *et al*. A self – adaptive genetic algorithm – artificial neural network algorithm with leave – one – out cross validation for descriptor selection in QSAR study[J]. Journal of Computational Chemistry, 2010,31(10):1956 – 1968.

[2]JAGDISH C P, ONKAR S. Artificial neural networks – based approach to design ARIs using QSAR for diabetes mellitus[J]. Journal of Computational Chemistry,2009, 30(15):2494 – 2508.

[3]JAGDISH C P, BOON H C. Artificial neural network – based drug design for diabetes mellitus using flavonoids [J]. Journal of Computational Chemistry,2011,32(4): 555 – 567.

[4]APILAK W, CHANIN N, THANAKORN N, *et al*. Modeling the activity of furin inhibitors using artificial neural network[J]. European Journal of Medicinal Chemistry, 2008,44:1664 – 1673.

[5]陈国华, 陆瑶, 陈虹. 基于逐步回归所得变量集的遗传反向传播神经网络的 QSAR 研究[J]. 计算机与应用化学,2010,27(9):1257 – 1262.

[6]杜雨静, 范英芳. 人工神经网络用于三苯基丙烯腈衍生物的定量结构 – 活性关系模型[J]. 化工进展, 2010,29(1):25 – 28.

[7] KIER L B, HALL L H. Molecular connectivity in structure – activity analysis[M]. [S. l.]: Research Studies Press,1987: 232 – 256.

[8]李鸣建, 冯长君. 取代苯甲酸对植物生长调节活性的拓扑 QSAR 研究[J]. 哈尔滨工业大学学报,2009,41 (5):195 – 197.

[9]陈炫, 聂长明, 蒋司同,等. 量子拓扑方法对硫醇的定量构效关系研究[J]. 南华大学学报,2009,23(4): 84 – 87.

[10] HUANG Qingguo, LIU Yongbin. Genotoxicity of substituted nitro benzenes and the quantitative structure – activity relationship[J]. Journal of Environmental Sciences,1996, 8:103 – 109.

[11] 于秀娟, 王鹏, 龙明策,等. 有机化学品点价自相关拓扑指数与生物降解性的定量关系[J]. 环境科学学报,2000,20(增刊):93 – 96.

[12] GEMAN S, BIENENSTOCK E, DOURSAT R. Neural networks and the bias/valance dilemma[J]. Neural Computation,1992,4(3):51 – 58.

[13] SCARDI M, HARDING L W. Developing an empirical model of phytoplankton primary production: a neural network case study[J]. Ecological Modelling, 1999, 120(2/3):213 – 223.

[14]GARSON G D. Interpreting neural – network connection weight[J]. Artificial Intelligence, 2001,6(8):47 – 51.

[15] LEK S, DELACOSTE M, BARAN P, *et al*. Application of neural networks to modelling nonlinear relationships in ecology[J]. Ecological Modelling, 1996, 90 (32):39 – 52.

[16] SUNG A H. Ranking importance of input parameters of neural networks[J]. Expert Systems with Applications,1998,15(12):405 – 411.

[17]沈洪艳, 张国霞, 刘宝友,等. 地表水体中常见硝基芳烃对鲤鱼的联合毒性[J]. 环境科学与技术, 2011,34(2):17 – 21.

(编辑 刘 彤)