基于遗传算法的 PLS 分析在 QSAR 研究中的应用

张大仁, 赵立新(中国科学院生态环境研究中心, 北京 100085, E-mail: darenzh@mail.rcees.ac.cn)

摘要: 将遗传算法和偏最小二乘法结合应用于定量结构活性关系研究中, 进行变量选择和建立最终模型. 以较少的变量个数包含较多的变量信息, 且变量间没有线性相关问题, 因而得到较好的 QSAR 模型. 将这种方法应用于氯代酚和单取代苯 2 种系列化合物, 可以得到几种常规多元回归分析方法不能得到的质量较高的 QSAR 模型.

关键词: 定量结构活性关系; 偏最小二乘法; 遗传算法; 变量选择
中图分类号: X17 文献标识码: A 文章编号: 0250-3301(2000)06-05-0011

Application of GA-Based PLS Analysis in QSAR Studies

Zhang Daren, Zhao Lixin(Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China E-mail: darenzh@mail.rcees.ac.cn)

Abstract: Genetic algorithms(GA) are an effective optimization method, and partial least squares(PLS) are a well-known multivariate statistical method. The combination of two methods is proposed to construct GA-based PLS analysis, which is applied to QSAR studies for variable selection and model construction. By the method, a few latent variables can include the information of more variables, thus better QSAR models can be obtained. The method is applied to chlorophenols and mono-substituted benzene derivatives, and some models are better than the best multivariate regression results.

Keywords: QSAR; PLS; genetic algorithms; variable selection

在定量结构活性关系(quantitative structure-activity relationship, QSAR)研究中, 其结构参数包括实验测定的物理化学参数和理论计算的参数, 数量越来越多, 而化合物的生物活性实验需耗费较多人力物力. 为找到最好的模型, 变量选择的工作量通常很大. 近年来广泛应用于化学各个领域的遗传算法[1], 也被应用于变量选择中[2], 这种基于达尔文(Darwin)的自然选择和进化理论而发展起来的计算方法, 是一种重要的优化工具. 然而, 尽管用交互验证作为适应函数, 但用遗传算法选择的变量并不能保证都对回归模型有显著贡献, 用其他方法对数据进行一定的处理就变得必要.

偏最小二乘法(partial least squares, PLS)是一种新的回归分析方法[3], 它是从自变量和因变量中分别提取潜变量, 并使之尽可能相关, 这种潜变量变为互相正交, 就克服了共线性问题, 同时保留了自变量中的有用信息. 但是, PLS 方法对于数据中存在的偏离点(outlier)是非常灵敏的, 这能导致整个模型缺乏预言能力. 因此, 这样的变量必须排除在 PLS 回归模型之外, 这就需要进行变量选择, 如上所述, 遗传算法能较好地达到这个目的. 将这 2 种方法结合起来应用于 QSAR 研究中, 近年来已引起人们的注意[4,5]. 本文的目的就是将这 2 种方法结合起来, 用于氯代酚和单取代苯系列化合物的 QSAR 研究中, 以期得到质量较好的模型, 并与多元回归分析进行比较, 以评价 PLS 所得到的改进.

1 方法

1.1 PLS

PLS[3,6]主要用于多因变量与多自变量之间的回归分析, 当然, 它也能用于单个因变量的

基金项目: 国家自然科学基金资助项目(29577287)
作者简介: 张大仁(1940~), 男, 主要从事化学和环境中的模拟计算.
收稿日期: 2000-04-12

回归分析, 这是 QSAR 中常遇到的情况. 偏最小二乘法的核心是从自变量和因变量各自提取潜变量, 设为 T 和 U(T 为 t_1, t_2… 组成的矩阵, U 为 u_1, u_2… 组成的矩阵), 设原始数据已经经过标准化处理, 并设自变量矩阵为 X, 因变量矩阵为 Y, 则有:

$$q_i = \frac{Y_i^T t_i}{t_i^T t_i} \quad (1)$$

$$p_i = \frac{X_i^T t_i}{t_i^T t_i} \quad (2)$$

这里 Y_i 和 X_i 是提取 $i-1$ 个潜变量后, 剩余部分. 上标 T 表示转置矩阵. 设提取 a 个分量, 则:

$$X = \sum_{i=1}^{a} t_i p_i^T + E \quad (3)$$

$$Y = \sum_{i=1}^{a} u_i q_i + F \quad (4)$$

这里 E 和 F 是残差.

在 PLS 分析中, 先是建立 U 和 T 的回归关系, 然后可还原到 Y 与 X 的回归关系. 设 W 为 X 的权重矩阵[6], 则:

$$T = XR \quad (5)$$

$$R = W(P^T W)^{-1} \quad (6)$$

$$Y = RQ^T X \quad (7)$$

式(7)即为得到的原始变量间的回归方程.

在 PLS 中提取潜变量 t_i 和 u_i 时, 有 2 点被满足: ① t_i 和 u_i 应尽可能大地获得本变量系统中的信息; ② t_i 与 u_i 相关程度达到最大. 这样保证潜变量之间是独立的, 又使自变量的潜变量 t_i 能更好地解释因变量的潜变量 u_i, 从而最终达到建立比多元回归分析更有预言能力的模型. 同时, 由于在遗传算法中要进行大量的 PLS 计算, 采用节省时间的 PLS 算法是必要的, 笔者采用了 Dayal 等[6]改进的 PLS 核心(kernel)算法, 与通常的算法相比, 明显地节省了时间.

1.2 遗传算法

遗传算法可参考文献[1, 4], 方法是随机地设置染色体(即解的个体), 然后通过复制, 交叉, 突变等操作, 优胜劣汰, 得到一些最好的染色体, 即解. 在本文中, 先随机选取初始的 30 个染色体, 进化 100 代, 并进行 100 次这样的运行. 所用适应函数为交互验证(cross-validation)的决定系数(CVR^2):

$$CVR^2 = 1 - \sum_i (Y_{i,\mathrm{obs}} - Y_{i,\mathrm{pred}})^2 \Big/ \sum (Y_{i,\mathrm{obs}} - \bar{Y}) \quad (8)$$

这里 $Y_{i,\mathrm{obs}}$和 $Y_{i,\mathrm{pred}}$分别为样品 i 的测定和预言值, $\bar{Y}$ 为 Y 变量的平均值.

经过遗传算法进行变量选择, 用 CVR^2 作判断, 就能排除一些对建立 PLS 模型不适合的变量, 而得到优化组合.

2 结果和讨论

遗传算法中的交叉几率为 0.5, 突变几率为 0.01, 并且以全部变量进行计算, 选取各种较好的变量组合, 用这种方法分别计算了氯代酚和单取代苯 2 种系列化合物.

2.1 氯代酚[7]

该系列有 20 个化合物, 生物毒性为对发光菌的毒性(用- $\log EC_{50}$表示), 所用参数基本上为量子化学计算所得, 包括(括号中数字为变量的代号): 分子总表面积的对数(1), 生成热(2), 总能量(3), 离子化势(4), OH 基表面积的对数(5), O 原子的净电荷(6), 偶极矩(7), C1 原子的净电荷(8~12), OH 基上 H 原子的净电荷(13), 氯原子数(14)等共 14 种参数. 对这些参数进行了基于遗传算法的 PLS 变量选择, 所得结果被列于表 1.

在第 1 和第 2 个模型中, 不仅 R^2、CVR^2 与回归分析结果相差很小, 在最后还原为原始数据的回归方程中, 截止到 4 位有效数字所有系数也是相同的. 如第 1 个变量组合, 经过 3 轮计算后, 得到了构成 X 的 3 个潜变量的主轴:

W_1: - 0.7709 0.4141 0.4840

W_2: - 0.0232 - 0.7776 0.6283

W_3: - 0.6365 - 0.4732 - 0.6091

由(5)式就可求得 3 个潜变量 t_1, t_2, t_3, 由于 y 是单因变量, 对这些潜变量的回归方程为:

$$y = 0.6752t_1 + 0.2864t_2 + 1.2421t_3 + f \quad (9)$$

这里 f 是残差. 将这个方程还原为原始数据之间的方程, 并忽略残差 f, 则有

$$y = -1.5113(3) - 0.8415(6) + 0.1939(7) \quad (10)$$

这个方程与回归分析得到的完全一致. 第 2 个变量组合也是这样. 这说明每组变量间相互独立性好, 潜变量数等于变量数, 则 PLS 回归方程收敛到常规的多元回归方程. 但对于含 4 个变量及其以上的回归方程, 常规多元回归分析均不能得到高质量的模型, 这可能是变量之间存在共线性问题. 而在 PLS 分析中, 从表 1 可以看出, 采用 5 个变量和 6 个变量时, 可以得到更好一些的结果. 在这里最有意义的结果是模型 5 和模型 8, 这里虽然分别用了 4 个变量和 5 个变量, 但潜变量均只用 2 个, 其对于各自的 t_1, t_2 有如下方程:

$$y = 0.3834t_1 + 1.5569t_2 + f \quad (11)$$

$$y = 0.3517t_1 + 1.3818t_2 + f \quad (12)$$

还原为原始数据方程则为:

$$y = -0.9324(3) - 1.0693(6) - 0.1652(13) + 0.9322(14) \quad (13)$$

$$y = 0.6744(1) - 0.6093(3) - 1.0044(6) - 0.2521(13) + 0.6092(14) \quad (14)$$

由表 1 可知, R^2 分别为 0.9333 和 0.9388, CVR^2 分别为 0.9060 和 0.9028, 均优于回归分析中 2 个变量的模型, 普通回归分析 2 变量的最好模型其 R^2 为 0.8930~ 0.9060, 而 CVR^2 仅为 0.8589~ 0.8595. 这说明即使对于比较简单的体系, PLS 分析也可以得到维数低的较好的模型.

2.2　单取代苯[8]

这一系列包含 30 个化合物, 生物活性也为对发光菌的毒性. 所用参数为量子化学计算参数和分子形状指数等参数, 包括(括号中数字为参数的代号): 诱导指数(1), 代基当量(2), 生成热(3), 取代基所连 C 原子的净电荷(4), 偶极矩(5), HOMO(6), LUMO(7), 分子总表面积的对数(8), 总能(9), 分子中最负的原子净电荷(10), 最正的原子净电荷(11), ${}^0K_\alpha$(12), ${}^1K_\alpha$(13), ${}^2K_\alpha$(14), ${}^3K_\alpha$(15), 活性基团表面积的对数(16), $({}^1K_\alpha)^2$(17), $({}^2K_\alpha)^2$(18), $({}^3K_\alpha)^2$(19)等共 19 种参数. 对这些参数进行的变量选择结果列于表 2.

对于变量组合1和2, 基于前面关于潜变

表 1　氯代酚 PLS 变量选择的结果

Table 1　PLS variable selection of chlorophenols

模型编号	变量	变量数	潜变量数	模型的决定系数(R^2)	交互验证的决定系数(CVR^2)
1	3, 6, 7	3	3	0.9424	0.9216
2	6, 7, 14	3	3	0.9424	0.9216
3	3, 6, 13, 14	4	4	0.9475	0.9236
4	3, 6, 13, 14	4	3	0.9375	0.9106
5	3, 6, 13, 14	4	2	0.9333	0.9060
6	1, 3, 6, 13, 14	5	5	0.9739	0.9523
7	1, 3, 6, 13, 14	5	3	0.9399	0.9183
8	1, 3, 6, 13, 14	5	2	0.9388	0.9028
9	1, 2, 6, 13, 14	5	5	0.9738	0.9521
10	1, 2, 6, 13, 14	5	4	0.9402	0.9114
11	1, 2, 6, 13, 14	5	3	0.9246	0.8881
12	1, 3, 6, 11, 13, 14	6	6	0.9767	0.9532
13	1, 3, 6, 11, 13, 14	6	4	0.9425	0.9132
14	1, 3, 6, 11, 13, 14	6	3	0.9401	0.8931

表 2 单取代苯 PLS 变量选择的结果

Table 2 PLS variable selection of mono-substituted benzene

模型编号	变量	变量数	潜变量数	模型的决定系数(R^2)	交互验证的决定系数(CVR^2)
1	8, 16, 19	3	3	0.8660	0.8368
2	8, 15, 16	3	3	0.8425	0.8075
3	3, 8, 16, 19	4	4	0.8839	0.8529
4	8, 11, 16, 19	4	4	0.8811	0.8481
5	8, 11, 16, 19	4	3	0.8652	0.8295
6	5, 8, 11, 16, 19	5	4	0.9184	0.8897
7	5, 8, 11, 16, 19	5	3	0.9153	0.8846
8	8, 11, 13, 16, 19	5	5	0.8879	0.8579
9	5, 8, 11, 15, 16	5	5	0.9167	0.8780
10	5, 8, 11, 15, 16	5	4	0.9039	0.8700
11	7, 8, 11, 15, 16	5	5	0.9116	0.8823
12	7, 8, 11, 15, 16	5	4	0.9095	0.8618
13	7, 8, 11, 16, 19	5	5	0.9082	0.8693
14	7, 8, 11, 16, 19	5	4	0.9017	0.8573
15	5, 8, 11, 16, 18, 19	6	6	0.9286	0.9024
16	5, 8, 11, 16, 18, 19	6	4	0.9272	0.9060
17	5, 8, 11, 16, 18, 19	6	3	0.9111	0.8919
18	5, 8, 11, 14, 18, 19	6	4	0.9297	0.9097
19	5, 8, 11, 14, 18, 19	6	3	0.9098	0.8936
20	7, 8, 11, 14, 15, 16	6	5	0.9117	0.8758
21	5, 8, 11, 14, 15, 16	6	5	0.9278	0.8983
22	5, 8, 11, 14, 15, 16	6	4	0.9150	0.8875
23	5, 8, 11, 15, 16, 19	6	5	0.9250	0.8884
24	5, 8, 9, 11, 15, 16, 18	7	6	0.9397	0.9131
25	5, 8, 9, 11, 15, 16, 18	7	5	0.9339	0.9020
26	5, 8, 9, 11, 14, 16, 19	7	6	0.9351	0.9126
27	5, 8, 9, 11, 14, 16, 19	7	5	0.9298	0.9036
28	5, 8, 9, 11, 14, 16, 19	7	4	0.9117	0.8785
29	5, 8, 9, 11, 15, 16, 19	7	5	0.9238	0.8818
30	5, 8, 9, 11, 14, 15, 16	7	5	0.9342	0.9096
31	5, 8, 11, 15, 16, 18, 19	7	4	0.9285	0.9061
32	5, 8, 11, 15, 16, 18, 19	7	3	0.9090	0.8859

量数与变量数相等情况的说明, 也与回归分析的结果是一致的. 重要的改进在含有 5 个变量以上的模型, 这种改进比氯代酚更明显. 不但用较多潜变量得到较好的模型, 特别是用较少潜变量时, 模型仍明显优于同样变量数下的回归模型. 比较典型的模型有 6, 7, 16, 17, 18, 19,

31, 32 等, 如在第 19 个模型中, 潜变量数为 3, 其 R^2 和 CVR^2 分别为 0.9111 和 0.8919, 而在常规回归分析中, 最好的 3 变量模型其 R^2 和 CVR^2 分别为 0.8649 和 0.8300, 前者明显优于后者. 为节省篇幅, 这里不再列出这些模型的方程, 其构成方式同 2.1 节的描述.

基于遗传算法的 PLS 模型能得到维数较低的高质量模型, 一是得益于遗传算法的变量选择, 另一是如前面所叙述的 PLS 的特点, 潜变量的提取尽可能多地包含了原始变量的信息, 且自变量的潜变量与因变量的潜变量之间的相关程度达到最大, 这样就使得较少的潜变量含有较多变量的信息, 且与因变量相关程度大, 因而有利于建立低维的更具有预言能力的 QSAR 模型. 从表 1 和表 2 也可知, 这种方法, 同时得到了很多模型, 可供挑选.

3　结论

基于遗传算法的 PLS 分析将 2 种方法的特点结合起来, 因而具有明显的优点.

遗传算法进行的变量选择排除特别不适合的变量, 使 PLS 使用的变量得到优化组合, 同时, 很多的变量组合模型被同时建立.

PLS 的特点使得能用较少的潜变量含有较多变量的信息, 并且潜变量是相互正交的, 因而能用较低维数建立具有预言能力的模型, 也排除了得到没有统计意义的模型的可能.

应用于氯代酚和单取代苯系列, 均得到一些优于常规回归分析的模型, 对于体系较复杂的后者效果更明显.

致谢　Leardi 教授热情地提供他的程序和建议, 谨表谢意.

参考文献:

1　Lucasius C B, Kateman G. Understanding and Using Genetic Algorithms. Part 1. Concepts, Properties, and Context. Chemomom. Intell. Lab. Syst., 1993, **19**: 1~ 33.

2　Leardi R. Application of Genetic Algorithms to Feature Selection under Full Validation Conditions and to Outlier Detection. J. Chemom., 1994, **8**: 65~ 79.

3　Geladi P, Kowalski B R. Partial Least-Squares Regression: A Tutorial. Anal. Chim. Acta, 1986, **185**: 1~ 17.

4　Leardi R, Gonzalez A L. Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. Chemom. Intell. Lab. Syst., 1998, **41**: 195~ 207.

5　Dunn W J, Rogers D. Genetic Partial Least Squares in QSAR. Genetic Algorithm in Molecular Modeling(ed. by Devilliers). London: Academic Press, 1996, 109~ 130.

6　Dayal B S, MacGregor J F. Improved PLS Algorithms. J. Chemom., 1997, **11**: 73~ 85.

7　张大仁等. 分子表面积的精确和经验计算及其在 QSAR 中的应用. 环境化学, 1994, **13**(3): 234~ 238.

8　张大仁. 单取代苯衍生物的 QSAR 研究. 环境化学, 1996, **15**(6): 536~ 540.