# 结构多样性化合物沸点QSPR模型研究

周新奇[1]，何勤[1]，赵晨曦[1*]，曾映旭[2]，梁逸曾[2*]
（1. 长沙学院生物工程与环境科学系，湖南，长沙，410003；2. 中南大学化学化工学院中药现代化研究中心，湖南，长沙，410083）

**摘要**：在定量结构-性质/活性关系（QSPR/QSAR）研究中，分子结构的数值化和建立良好预测的数学模型是2个重要的问题。建立具有良好适应性有实际应用价值的模型是进行QSPR/QSAR的最终目标。本文针对148种来自不同植物挥发油的具有结构多样性化合物，分别采用主成分回归（PCR）、偏最小二乘（PLS）和遗传算法（GA）对其沸点与结构之间的定量结构性能关系进行了研究，用拓扑指数建立了沸点预测模型。结果表明，PLS模型的预测能力最佳，模型训练集的平均相关系数为0.996，平均训练偏差为7.05，检验集的平均相关系数为0.986，平均检验偏差为12.91。

**关键词**：结构多样性化合物；沸点预测；QSPR/QSAR；拓扑指数
**中图分类号**：R 284.1
**文献标识码**：A
**文章编号**：1001-4160(2008)06-717-720

# QSPR study for compounds with diversity structures

Zhou Xinqi[1], He Qin[1], Zhao Chenxi[1*], Zeng Yingxu[2] and Liang Yizeng[2*]
(1. Department of Biology Engineering and Environmental Science, Changsha, 410003, Hunan, China; 2. Research Center of Modernization of Chinese Herbal Medicine, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, Hunan, China)

**Abstract**: IIt is key that how to convert numerical value from molecular structure and to build good model in quantitative structure-property/activity relationship (QSPR/QSAR) study. The constructing of QSAR/QSPR prediction model of good compatibility and practical application is our ultimate objective. In the present research, we focus on a set of 148 compounds (mainly collected from three different volatile oils) to construct predicting models between their structures and normal boiling points by PCR, PLS or GA through topological indices. It showed that the PLS model is the best one in training and prediction. Its training regulations average correlation coefficient is 0.996, the average training deviation is 7.05. The average testing correlation coefficient is 0.986 and the average testing deviation is 12.91.

**Key words**: compounds with diversity structures, prediction of boiling point, QSPR/QSAR, topological index

Zhou XQ, He Q, Zhao CX, Zeng YX and Liang YZ. QSPR study for compounds with diversity structures. Computers and Applied Chemistry, 2008, 25(6):717-720.

## 1 引言

拓扑指数是定量构效活性关系（quantitative structure-activity relationship, QSAR）与定量结构性能关系（quantitative structure-property relationship, QSPR）研究中常用而有效的一类描述子[1]，其中最著名的是Randic和Kier提出的分子连接性指数，它们被用来预测许多化合物的物理化学性质[2~4]。

沸点是化工设计和化工计算中常用的物性数据之一，也常常作为预测其他性质的基础。大部分化合物的沸点可以通过实验测定，但是对于测量上有困难或尚未合成的化合物，就有必要对其沸点进行估算或预测，以确定更好的实验条件或者工艺。多年来，许多研究者提出了计算和预测沸点的方法[5,6]，但这些研究比较单一而且往往只针对某一类化合物（卤代烃、苯类等），虽然预测计算值比较准确，但适用范围受到一定限制。

本论文拟对一组来自不同植物挥发油且具有多种结构特征的化合物进行沸点与结构之间的相关关系研究，采用不同的计算方法如主成分回归分析[7]（PCR）、偏最小二乘[8]（PLS）以及遗传算法[9]（GA）等建立沸点预测模型，为复杂体系的气相色谱-质谱（GC-MS）分析中某些单独用质谱难以

**收稿日期**：2007-11-07；**修回日期**：2008-04-18
**基金资助**：长沙市科技计划重点基金（No. K069054-12）；湖南省教育厅研究基金（No. 06C164）和湖南省科技计划基金（No. 2007FJ3094）支助.
**作者简介**：周新奇（1981—），硕士研究生，E-mail：zhouxingqi2000@163.com.

准确定性的化合物的鉴定起到辅助定性的作用。

## 2 理论部分

### 2.1 主成分回归法(PCR)

主成分回归法(principal component regression,PCR)[7],是采用多元统计中的主成分分析方法,先对混合物量测矩阵 **Y** 矩阵直接进行分解,然后只取其中的主成分来进行回归分析。对 **Y** 矩阵直接进行分解在化学计量学中一般采用的方法是非线性迭代偏最小二乘算法(nonlinear iterative partial least squares,NIPALS),另一种方法是线性代数中常用的奇异值分解法(single value decomposition,SVD)。奇异值分解法可将任意阶实数分解成为 3 个矩阵的积,即,$Y = USVt$。用奇异值分解法求出量测矩阵的广义逆 $Y^{0+}$,然后用它求出回归系数矩阵 **P**

$$\mathbf{P} = C\ Y^{0+} = CV^{*}(S^{*})^{-1}U^{t*}$$

用求得的 **P** 直接就可计算未知混合体系的浓度矢量 $c_{未知}$ 或浓度矩阵 $C_{未知}$:

$$c_{未知} = Py_{未知} \text{ 或 } C_{未知} = PY_{未知}$$

### 2.2 偏最小二乘法(PLS)

在多元校正中偏最小二乘方法(partial least squares,PLS)[8]是一种基于高维投影思想的新的非参数回归方法。该法的优点在于它不但对量测矩阵 **X** 进行正交分解,而且在分解 **X** 的同时对相应矩阵 **Y** 也进行正交分解,是一种同时进行分解的特征变量回归法。PLS 对 **X** 和 **Y** 同时进行分解的方法如下:

$$\mathbf{X} = USVt = U^{*}S^{*}Vt^{*} + EX = T^{*}Vt^{*} + EX$$

$$\mathbf{Y} = PGQt = P^{*}G^{*}Qt^{*} + EY = R^{*}Qt^{*} + EY$$

由样本量测矩阵 **X** 分解得到的矩阵 $\mathbf{T}^{*}$ 以及由响应矩阵 **Y** 分解得到的矩阵 $\mathbf{R}^{*}$,代表了除去大部分噪声后的变量和响应的信息。而且得到的 2 个矩阵 $\mathbf{T}^{*}$ 和 $\mathbf{R}^{*}$,经回归建立起内部联系,即线性关系,这样的关系在 PLS 中称为内相关。

### 2.3 遗传算法(GA)

遗传算法(genetic algorithm,GA)[9]是一类借鉴生物界自然选择和遗传机制的高度并行、随机、自适应搜索算法,它是由复制、杂交和变异 3 个算子组成。它通过模拟生物的遗传、变异和自然选择过程,将一代群体变换到新一代的群体。在这里每代群体由一组染色体组成,每条染色体代表搜索空间内的一个解。染色体由待求参数排列在一起构成。通过对上一代(父代)群体中的染色体进行有选择的复制、交叉和变异,可产生新一代(子代)群体,此过程一直重复,直至达到最优解。

## 3 数据及处理

### 3.1 化合物结构及沸点数据采集

本文采用的数据集由 2 个部分组成,一部分数据来自本实验室对杜鹃挥发油[10]、紫丁香挥发油[11]以及莪术挥发油的 GC-MS 分析鉴定的化合物的正常沸点数据,另一部分数据是从有关物性手册[12]上找到的与挥发油化合物结构相近的化合物的正常沸点数据,其分子结构和沸点主要是从 http://webbook.nist.gov/chemistry/cas-ser.htmlh 和 http://chemdbs.com/source/sdict.php 两个网站获得。共收集到 148 种化合物及其沸点数据(表略)。

### 3.2 化合物结构描述

采用 ISIS/BASE 软件保存分子结构及其沸点数据,然后将其结构和沸点数据分别导出并保存为 SDF 文件。分别采用 DROGEN 软件和中南大学化学化工学院中药现代化研究中心编制的软件,计算得到了 4 组描述子,分别是分子连结性指数,分子形状指数,分子电负性距离指数和电荷指数。

### 3.3 数据预处理

将数据中心化(中心化公式:$x_{ij}$, new = $x_{ij}$, old/mean($x_{ij}$, old),式中 mean 表示第 $j$ 列变量的平均值)并自标度化($x_{ij}$, new = ($x_{ij}$, old − $m_j$)/$V_j$,式中 $V_j$ 为变量 $j$ 的方差,$m_j$ 为变量 $j$ 的均值)。并采用相关系数计算方法将 0 或常数变量预先剔除。相关系数计算方法如下:

$$\cos(\alpha_{ij}) = [\sum (x_{ik} - m_i) \times (x_{jk} - m_j)] / (\sqrt{\sum (x_{ik} - m_i)^2} \times \sqrt{\sum (x_{jk} - m_j)^2})$$

其中 $m_i$ 和 $m_j$ 分别表示第 $i$ 和第 $j$ 个样本的均值,即 $m_i = (\sum x_{ik})/n, k = 1, 2, \cdots, n$。当 2 个变量的相关性超过 0.95 则剔除掉其中一个变量。

## 4 建模分析

### 4.1 样本选择

本文中采用 PCR 回归方法尝试剔除奇异点(坏样本)。当选用的变量数为 10 时,剔除奇异点前后 PCR 模型的相关系数分别为 0.91 和 0.98,方差分别为 35.5 和 18.3。由图 1 可以看出,剔除奇异点后回归结果显著改善。

### 4.2 变量选择

#### 4.2.1 PLS 交互检验变量选择

PLS 蒙特卡罗交互检验实际上是随机检验,即我们随机抽取样本分子作为检验集进行检验。此时所采用的参数为:随机抽取 20 个样本分子作为检验集,其余的作为训练集;所有变量采用自标度预处理方法,程序进行一百次检验。在取 10 个变量时其检验集的平均检验偏差最小(见图 2),因此我们可以确定建模的最优变量个数为 10,此时训练集的平均相关系数 $R$ 为:0.996,平均训练集计算方差 $S$ 为 7.05,检验集的平均相关系数 $R$ 为 0.986,平均检验偏差 $S$ 为 12.91。此结果比 PCR 建模的结果有了很大的提高。

由此可见,采用 PLS 方法对沸点进行建模是可取的,但是 PLS 方法只能得到对沸点影响较大的变量有多少个,而不能提供具体的变量的内容,所以我们下面采用遗传算法结果 PLS 模型确定具体的变量的内容。

#### 4.2.2 GA 算法选择变量

遗传算法中各个参数设置如下:最大变量个数取 30;30

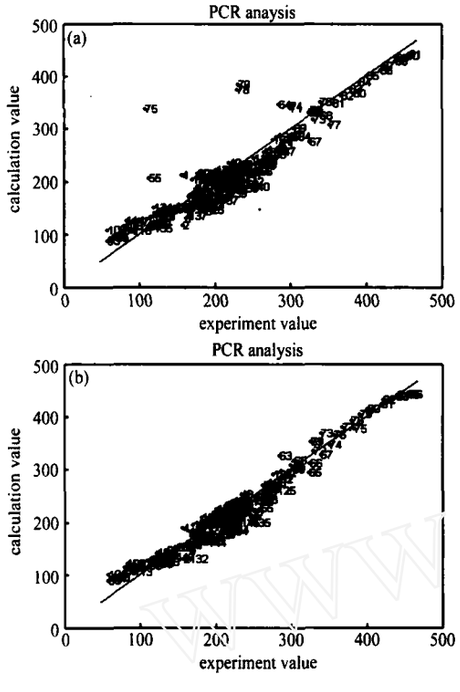


Fig. 1　PCR analysis results obtained by eliminating the outlier (a) or not (b).

图 1　剔除奇异点前(a)后(b)PCR 回归结果

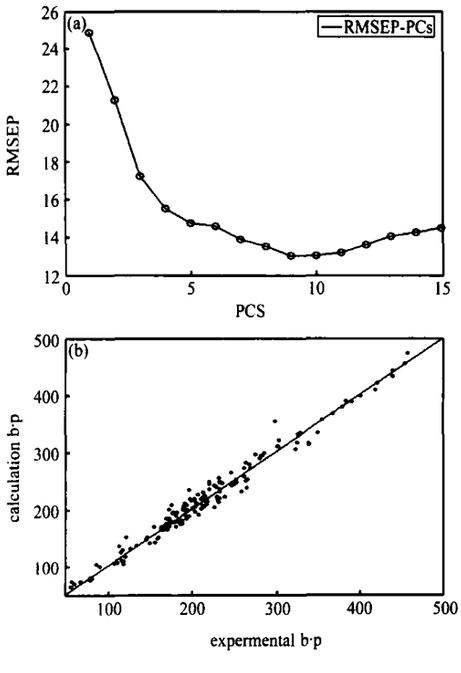


Fig. 2　Results of variable selection (a) and regression (b) by PLS cross validation.

图 2　PLS 交互检验变量选择(a)及回归结果(b)

条染色体;繁殖一百代;基因变异率为 0.01;基因交叉率为 0.5,采用 PLS 交互检验方法,以预测偏差达到最小作为目标函数。

从图 3(b)中可以看出当取 10 个变量时,交互检验的相关系数已经接近 0.97,9 个分子作为检验集,此时的沸点的检验平均偏差为:14.91。从图 3(a)中可知,第 43、136、10、17、62、84、102、14、114、33 列变量的适应性最强,因此他们是显著性变量。

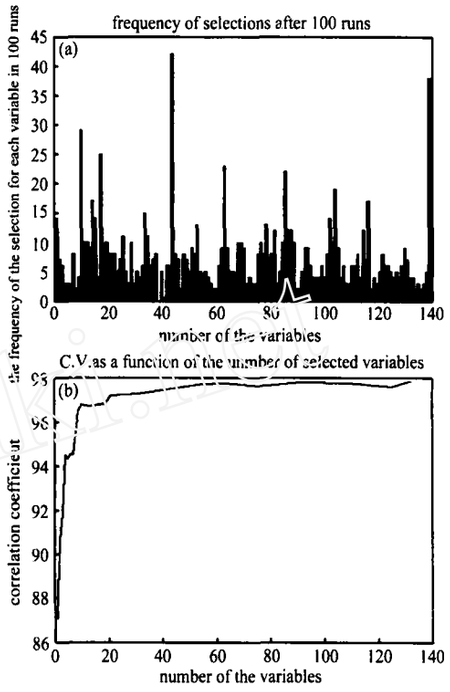


Fig. 3　Plot of variable selection (a) and result obtained by cross validation based on the selected variables (b).

图 3　变量选择图(a)及所取变量个数交互检验结果(b)

## 5　结论

本论文先后采用了 PCR、PLS 和 GA 3 种方法进行沸点模型研究,通过样本选择和变量选择剔除奇异点,分别建立了相应的沸点预测模型。结果表明,偏最小二乘方法(PLS)所建立的模型具有最佳的沸点预测能力,训练集相关系数 $R$ 为 0.996,此相关程度明显优于拓扑指数$^1X^F$[13] ($R$ = 0.981),也优于 MDE 矢量[14] ($R$ = 0.995)。检验集的相关系数 $R$ 为 0.986,预测偏差 $S$ 为 12.9。它虽然比文献[6,14]报道的某一类物质的沸点预测偏差稍高,但是本文首次对具有多结构特征的数据集进行研究,模型所得预测偏差与实验误差接近,说明所建立的模型具有较强的预测能力,具有实际应用价值,特别是对于 GC-MS 分析中单独用质谱难以准确定性的化合物的结构确定具有很好的辅助作用。

**References**

1　Ren BY, Xu Y and Chen GB. QSPR/QSAR research on organic compounds using a new topological index. Acta Chimica Sinica, 1997,57:563 - 571.

2　Xin Houwen and Zhang Hongguang. Topological theory of relationship between molecular structyre and its properties. Chinese Journal of Atomic and Molecular Physics, 1990, 7(4):16 - 28.

3　Xin HW, Zhang HG. Bond parameter topological index and its application in the properties of related molecules with alike atomic form of XY1, XY2, XY3, XY4. Acta Physico-Chmimica Sinica, 1989, (3):26 - 28.

4 Wang ZD, Huang YP and Yang F, et al. Signification and application of a novel molecular topological index. Wuhan University Journal (Natural Science Edition), 2003, 49(4):441 - 444.

5 Wang KQ and Sun XZ. Correlation of boiling points with molecular structure for halogenated propanes. J Capital Normal University (Natural Science Edition), 2000, 21(4):52 - 55.

6 Zhang Feng. Correlation studies between the molecular topological index and boiling point for haloalkanes. J Yunnan Nationalities University (Natural Sciences Edition), 2006, 15(4):320 - 323.

7 Liang YZ. White, gray and black multicomponent systems and their chemometrics algorithms. Hunan Publishing House of Science and Technology, Changsha, Hunan, 1996:29 - 32.

8 Xu Lu. Chemometric Methods. Beijing: Publishing House of Science. 1995:80 - 128.

9 Liu Y and Kang LS. Nonnumerical Parallel Algorithm-genetic Algorithm. Beijing: Publishing House of Science, 1995.

10 Zhao CX, Li XN and Liang YZ, et al. Comparative analysis of chemical components of essential oils from different samples of rhododendron with the help of chemometrics methods. Chemometr Intell Lab Syst, 2006, 82:218 - 228.

11 Zhao CX, Liang YZ and Fang HZ, et al. Temperature - programmed retention indices for gas chromatography mass spectroscopy analysis of plant essential oils. Chromatogr A, 2005, 1096:76 - 85.

12 Ma PS. A Handbook of Experimental Physical Property Data of Organical Compounds. Beijing: Publishing House of Chemical Industry. 2006:1 - 297.

13 He XY and Chen YD. Correlation studies between a new topological index $^{1}X^{F}$ and boiling point for alcoholic compounds. J Changde Normal College, 2003, 14(1):27 - 29.

14 Lin ZH, Liu SS and Xu H. On structural descriptors for fatty alcohols and quantitative structure property relationship (QSPR). J Congqing University (Natural Science Edition), 2000, 23(1):105 - 108.

## 中文参考文献

1 任碧野，许友，陈国斌. 一个新的拓扑指数用于有机化合物的QSPR/QSAR研究. 化学学报，1997，57:563 - 571.

3 辛厚文，张宏光. 键参数拓扑指数及其在相关XY1、XY2、XY3、XY4同构型原子分子性质中的应用. 化学物理学报，1989，(3):26 - 28.

4 王振东. 一种新的拓扑指数的意义及应用. 武汉大学学报，2003，49(4):441 - 444.

5 王克强，孙献忠. 卤代丙烷的沸点与分子结构的关联及预测. 首都师范大学学报，2000，21(4):52 - 55.

6 张凤. 分子拓扑指数与饱和卤代烃沸点的相关性研究. 云南民族大学学报，2006，15(4):320 - 323.

7 梁逸曾. 白灰黑复杂多组份分析体系及其化学计量学算法. 长沙:湖南科学技术出版社，1996:29 - 32.

8 许禄. 化学计量学方法. 北京:科学出版社，1995:80 - 128.

9 刘勇，康力山. 非数值并行算法——遗传算法. 北京:科学出版社，1995.

12 马沛生. 有机化合物实验物性数据手册—含碳、氢、氧、卤部分. 北京:化学工业出版社，2006:1 - 297.

13 何旭元，陈远道. 一种新的拓扑指数$^{1}X^{F}$及其与醇类化合物沸点的相关性研究. 常德师范学院学报，2003，14(1):27 - 29.

14 林治华，刘树深，徐红，等. 脂肪族饱和一元醇的结构描述及沸点定量计算. 重庆大学学报，2000，23(1):105 - 108.