

天然山楂香料挥发性组分的结构表征及其保留时间的估计和预测

廖立敏^{1,2}, 梅 虎¹, 李建凤¹, 舒 茂¹, 朱万平¹, 杨 娟¹, 李志良^{1,*}

(1. 重庆大学化学化工学院/生物工程学院, 重庆 400044

2. 内江师范学院化学与生命科学系, 四川 内江 641112)

摘 要: 用分子电距矢量(MEDV)对天然山楂香料挥发油的51种组分在Ultra 2柱上气相色谱保留时间(t_R)进行了定量构效关系的研究。通过多元线性回归, 得到的模型相关系数 $R=0.977$ 。用逐步回归的方法筛选5个变量进行建模, 相关系数 $R=0.950$, 再用留一法(leave-one-out, LOO)交互检验对模型进行评价, 得到的相关系数 $R_{cv}=0.924$ 。所以此模型具有较好的稳定性和预测能力。

关键词: 天然山楂; 分子电性距离矢量; 描述子; 定量结构-保留相关

Structural Characterization of Some Components from Essential Oil of Natural Hawthorn Perfume and Estimation and Prediction of Their Retention Times

LIAO Li-min^{1,2}, MEI Hu¹, LI Jian-feng¹, SHU Mao¹, ZHU Wan-ping¹, YANG Juan¹, LI Zhi-liang^{1,*}

(1. College of Chemistry and Chemical Engineering/College of Life Science and Biological Engineering, Chongqing University, Chongqing 400044, China 2. Department of Chemistry and Life Science, Neijiang Teachers College, Neijiang 641112, China)

Abstract: Quantitative structure-retention relationship (QSRR) studies of the retention time of 51 components of natural hawthorn perfume were performed by molecular electronegativity-distance vector (MEDV). Through multiple linear regression (MLR), the correlation coefficient 0.977 was obtained. 5 variables were selected to build the model through the method of stepwise multiple regression (SMR), the correlation coefficient of which 0.950. Then the model was evaluated by performing the cross validation with the leave-one-out (LOO) procedure, and R_{cv} obtained was 0.924. So it can be concluded that the stability and predictability of the model is good.

Key words natural hawthorn perfume; molecular electronegativity-distance vector (MEDV); descriptor; quantitative structure-retention relationship

中图分类号: R284

文献标识码: A

文章编号: 1002-6630(2008)03-0074-010

山楂又名山里红、红果, 具有消食化滞, 行气散瘀的功效, 山楂果实含有丰富的营养物质, 已被公认为是加工保健食品的优质原料之一, 我国种植面积非常广阔。山楂含有多种香味成分和丰富的微量元素, 在食品、保健品、日化、烟草行业得到广泛应用。张峻松等^[1]以山楂果为原料经生物技术处理得到了天然山楂香料, 它除了具有山楂基本香韵外, 还兼有优美的果香韵和酿甜香韵; 利用同时蒸馏萃取装置提取此香料中

的挥发油, 用GC-MS法分离并分析鉴定其成分及质量分数, 最后鉴定出51个挥发性化学成分。

定量结构-色谱保留值关系(quantitative structure-retention relationship, QSRR)的研究对预测保留值、选择分离条件及探索色谱保留机理都具有重要意义^[2]。实验采用这51种物质进行建模研究(表1)。用分子电距矢量(MEDV)^[3]对山楂果中分离出的51种化合物进行结构表征, 借助多元线性回归、逐步回归和留一法(leave-one-

收稿日期: 2007-01-16

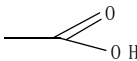
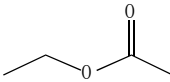
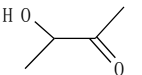
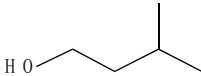
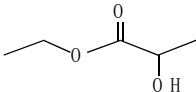
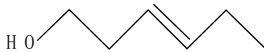
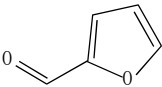
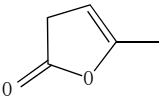
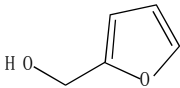
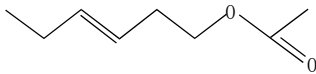
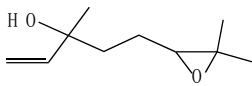
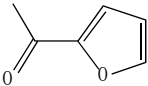
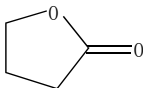
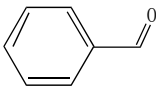
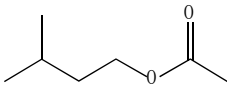
基金项目: 重庆市应用基础-研究基金项目(01-3-6); 重庆大学自主创新基金科技项目(030506; 040909)

作者简介: 廖立敏(1981-), 男, 硕士研究生, 主要从事天然药物的分离分析及药物分子结构与活性研究。

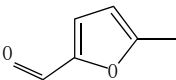
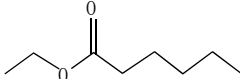
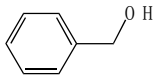
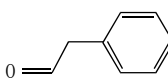
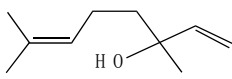
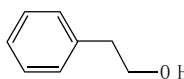
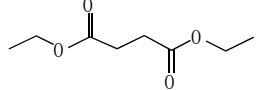
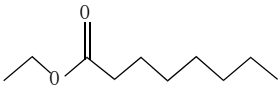
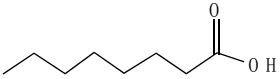
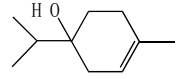
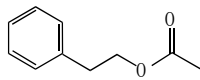
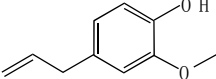
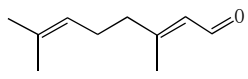
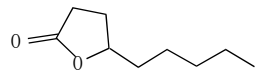
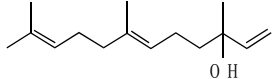
E-mail: liaolimin523@126.com

*通讯作者: 李志良(1962-), 男, 教授, 博士, 主要从事化学生物信息学及定量构效优化分子设计研究。E-mail: zlli-cqu@163.com

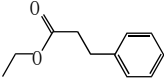
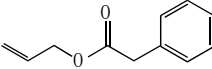
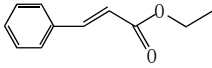
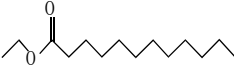
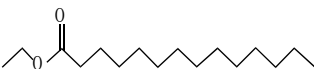
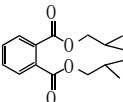
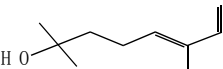
表1 天然山楂挥发油中51种组分及其在气相色谱-质谱中的保留时间
Table 1 51 Components of natural hawthorn perfume and their retention times

序号	化合物名称	英文名称	分子结构	分子式	t_R (min)
1	乙酸	Acetic acid		$C_2H_4O_2$	3.573
2	乙酸乙酯	Ethylacetate		$C_4H_8O_2$	3.638
3	3-羟基-2-丁酮	3-Hydroxy-2-butanone		$C_4H_8O_2$	3.895
4	异戊醇	Isoamyl alcohol		$C_5H_{12}O$	4.190
5	乳酸乙酯	Ethyllactate		$C_5H_{10}O_3$	4.926
6	顺-3-己烯	<i>cis</i> -3-Hexenol		$C_6H_{12}O$	5.140
7	糠醛	Furfural		$C_5H_4O_2$	5.501
8	5-甲基-2(3H)-呋喃酮	5-Methyl-2(3H)-furanone		$C_5H_6O_2$	5.542
9	糠醇	2-Furanmethanol		$C_5H_6O_2$	5.816
10	顺-3-己烯醇乙酸酯	<i>cis</i> -3-Hexenylacetate		$C_{10}H_{18}O_2$	6.458
11	氧化芳樟醇	Linalool oxide		$C_{10}H_{18}O_2$	6.911
12	2-乙酰基呋喃	2-Acetylfuran		$C_6H_6O_2$	6.998
13	γ -丁内酯	γ -Butyrolactone		$C_4H_6O_2$	7.109
14	苯甲醛	Benzaldehyde		C_7H_6O	7.445
15	乙酸异戊酯	Isoamylacetate		$C_7H_{14}O_2$	8.019

(续表1)

序号	化合物名称	英文名称	分子结构	分子式	t _R (min)
16	5 甲基糠醛	5-Methy-2-furancarboxaldehyde		C ₆ H ₆ O ₂	8.414
17	己酸乙酯	Ethyl hexanoate		C ₈ H ₁₆ O ₂	9.446
18	苯甲醇	Benzyl alcohol		C ₇ H ₈ O	10.842
19	苯乙醛	Benzeneacetaldehyde		C ₈ H ₈ O	11.077
20	芳樟醇	Linalool		C ₁₀ H ₁₈ O	11.524
21	苯乙醇	Phenylethyl alcohol		C ₈ H ₁₀ O	14.420
22	丁二酸二乙酯	Diethylsuccinate		C ₈ H ₁₄ O ₄	16.609
23	辛酸乙酯	Ethyl octanoate		C ₁₀ H ₂₀ O ₂	17.209
24	辛酸	Octanoic acid		C ₈ H ₁₆ O ₂	17.987
25	4-萜品醇	4-Terpinol		C ₁₀ H ₁₈ O	19.400
26	乙酸苯乙酯	Phenethylacetate		C ₁₀ H ₁₂ O ₂	19.792
27	丁香酚	Eugenol		C ₉ H ₁₀ O ₂	20.391
28	柠檬醛	Citral		C ₁₀ H ₁₆ O	23.622
29	γ-壬内酯	γ-Nnonalactone		C ₁₀ H ₁₈ O ₂	24.238
30	橙花叔醇	Neroidol		C ₁₅ H ₂₆ O	25.549

(续表1)

序号	化合物名称	英文名称	分子结构	分子式	t_R (min)
31	3-苯丙酸乙酯	Ethyl 3-phenylpropionate		$C_{11}H_{14}O_2$	27.434
32	苯乙酸烯丙酯	Allylphenylacetate		$C_{11}H_{12}O_2$	28.067
33	肉桂酸乙酯	Ethyl cinnamate		$C_{11}H_{12}O_2$	28.403
34	十二酸乙酯	Ethyl dodecanoate		$C_{14}H_{28}O_2$	33.444
35	十五烷	Pentadecane		$C_{15}H_{32}$	36.910
36	十七烷	Heptadecane		$C_{17}H_{36}$	37.196
37	肉豆蔻酸乙酯	Ethylmyristate		$C_{16}H_{32}O_2$	40.021
38	月桂酸异戊酯	Isoamyl laurate		$C_{17}H_{34}O_2$	40.375
39	十八烷	Octadecane		$C_{18}H_{38}$	40.570
40	邻苯甲酸二异丁酯	Disisobutylphthalate		$C_{16}H_{22}O_4$	43.114
41	十九烷	Nonadecane		$C_{19}H_{40}$	44.054
42	2,6-二甲基-5,7-辛二烯-2-醇	2,6-dimethyl-5,7-octadien-2-ol		$C_{10}H_{18}O$	44.618
43	十六酸	Hexadecanoic acid		$C_{16}H_{32}O_2$	46.458
44	11-十六烯酸乙酯	Ethyl 11-hexadecenoate		$C_{18}H_{34}O_2$	46.773
45	十六酸乙酯	Ethyl hexadecanoate		$C_{18}H_{36}O_2$	47.158

(续表1)

序号	化合物名称	英文名称	分子结构	分子式	$t_R(\text{min})$
46	二十烷	Eicosane		$\text{C}_{20}\text{H}_{42}$	49.501
47	二十一烷	Heneicosane		$\text{C}_{21}\text{H}_{44}$	50.304
48	亚油酸乙酯	Ethyllinoleate		$\text{C}_{20}\text{H}_{36}\text{O}_2$	52.147
49	油酸乙酯	Ethylolate		$\text{C}_{20}\text{H}_{38}\text{O}_2$	52.304
50	硬脂酸乙酯	Ethylstearate		$\text{C}_{20}\text{H}_{40}\text{O}_2$	53.146
51	邻苯二甲酸辛酯	Dioctylphthalate		$\text{C}_{24}\text{H}_{38}\text{O}_4$	62.566

out, L00) 交互检验建立分子电距矢量和 51 种化合物的气相色谱保留值之间的相关模型(QSRR), 以期对天然药物及精细化工产品的色谱行为提供有益参考。

1 QSRR 研究方法和对化合物分子结构表征

1.1 MEDV 方法原理

本实验中 MEDV 矢量是由化学分子图理论方法衍生得到结构参数的一种矢量。分子由各种不同属性的原子按不同连接方式(化学键)构成。而原子之间的相互作用与原子电性及相隔距离密切相关。一般而言, 依据库仑定律, 电性越大, 距离越小, 相互作用就越大。另一方面, 不同原子类型之间的相互作用方式也不同。原子间相互作用取决于各原子的电负性和它们之间的距离。本实验中 MEDV 矢量的提出^[4]即基于这一原理。

1.2 相对电负性和相对键长

MEDV 方法提出了原子的相对电负性 q (relative electronegativity) 和化学键(A-B)的相对键长 d (relative bond length)。 q 定义为以 A 原子的鲍林电负性 X_A 比上 C 原子的鲍林电负性 X_C , 即 $q_A = X_A / X_C$; d 定义为以 A-B 键的键长 BL_{A-B} 比 C-C 单键的键长 (BL_{C-C}), 即 $d_{A-B} = BL_{A-B} / BL_{C-C}$ 。本研究所使用的相对电性与相对键长的计算参考文献^[5]得到。

1.3 原子类型和原子属性

“原子类型”用于间接地描绘不同分子的拓扑结构特征。绝大多数有机或药物分子中包含的非氢原子通常

为 C、O、N、P、S 和卤素原子(F、Cl、Br 和 I)。分子内的碳原子依据其所连接的非氢原子数可分为 C1、C2、C3、C4 四类, 分别表示与 1、2、3、4 个非氢原子相连。而其他非氢原子(包括 O、N、P、S、F、Cl、Br 与 I)则视为拟碳原子, 依据其所连接的非氢原子数, 它们同样也分为以上四种类型。为表征分子结构, 也需要对分子中不同原子之间相连的化学键类型进行区分, 这里将某一个原子的“原子属性”用于表达该原子键连的化学键类型。原子属性用两个大写的英文字母来表示, 第一个字母代表该原子所属元素类型, 第二个代表与其他原子或基团连接的化学键类型。例如, 原子属性“CC”指一个以单键或双键与其他原子连接形成共轭体系的 C 原子, 而原子属性“N1”指仅以单键与其他原子相连的 N 原子。

1.4 MEDV 矢量的计算方法

依据以上原子类型与原子属性的定义, 化合物分子中四类非氢原子发生相互作用可以组合出以下几种方式: m_{11} 、 m_{12} 、 m_{13} 、 m_{14} 、 m_{22} 、 m_{23} 、 m_{24} 、 m_{33} 、 m_{34} 、 m_{44} , 简写成 x_1 、 x_2 、 x_3 、 x_4 、 x_5 、 x_6 、 x_7 、 x_8 、 x_9 、 x_{10} 10 个结构描述子。分子中 MEDV 各元素的计算公式为:

$$x_i = m_{ij} = \sum_{i=n, j=1} \frac{q_i \cdot q_j}{d_{ij}^2} \quad (n=1, 2, 3, 4; n \leq i \leq 4)$$

式中, n 、 i 表示原子 i 、 j 的原子类型; q_i 、 q_j 表示原子 i 、 j 的相对电负性; d_{ij} 表示原子 i 、 j 之间的

相对距离。

这里以 3-己烯醇为例说明 MEDV 矢量的计算过程。MEDV 矢量的计算可用图 1 的分子图来说明, 首先对化合物分子中的各个非氢原子进行编号(见图 1), 然后判断各个非氢原子的原子类型和原子属性。对于 3-己烯醇, 分子内有 7 个非氢原子, 其原子类型分别为 2、2、2、2、2、1 和 1; 其相应的原子属性分别为 C1、C1、C2、C2、C1、C1 和 O1。计算 $x_1: x_1$ 为 6 与 7 号两个第一类原子之间的相互作用。这两个原子间的最短路径为: $d(C^6-O^7)=0.9286+1.0000+1.0000+0.8701+1.0000+1.0000=5.7987$ 。另外, 根据相对电负性的定义^[6], $q_{O1}=1.3490$,

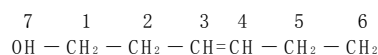


图 1 3-己烯醇的分子结构

Fig.1 Molecular matrix skeleton of cis-3-hexenol

$q_{C1}=1.0000$, 因此 x_1 数值计算如下:

$$x_1=m_{11}=1.0000 \times 1.3490/5.7987^2=0.0401。$$

同样, 可计算出其他几个 MEDV 矢量(x_2 、 x_5) 分别为 3.7168 和 5.4523 (x_3 、 x_4 、 x_6 、 x_7 、 x_8 、 x_9 、 x_{10} 均为 0)。以此类推, 可得到 51 种化合物的每个 MEDV 元素值(详见表 2, 表中数据为 True Basic 程序 Imp Vmed 在 PC 机上运行而得)。

表 2 51 种天然山楂挥发油组分所得的 MEDV 和 t_R

Table 2 MEDV and retention times of 51 natural hawthorn perfumes components

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$t_{R(exp)}(\text{min})$
1	1.3972	0	4.7139	0	0	0	0	0	0	0	3.573
2	0.5885	2.6546	3.2720	0	1.5644	1.8543	0	0	0	0	3.638
3	1.4699	0	6.9966	0	0	0	0	1.0000	0	0	3.895
4	0.4248	2.6493	2.1573	0	1.0000	1.2500	0	0	0	0	4.190
5	0.9841	2.7110	5.9363	0	1.5644	2.3395	0	1.0000	0	0	4.926
6	0.0401	3.7168	0	0	5.4523	0	0	0	0	0	5.140
7	0	2.8310	0.4697	0	6.1085	4.6168	0	0	0	0	5.501
8	0.1013	1.9776	3.4642	0	1.7796	5.9857	0	0.2899	0	0	5.542
9	0	2.1465	0.3627	0	6.0201	4.3890	0	0	0	0	5.816
10	0.4606	3.2143	3.1716	0	7.6888	2.1193	0	0	0	0	6.458
11	1.0810	4.3037	0.6921	5.0549	1.9366	2.8769	4.3029	0	1.1111	0.0625	6.911
12	0.4200	1.1204	3.8955	0	5.1315	4.3660	0	1.2277	0	0	6.998
13	0	1.3998	2.1495	0	4.5398	3.1043	0	0	0	0	7.109
14	0	2.8547	0.4882	0	7.4056	4.5266	0	0	0	0	7.445
15	0.8128	2.3575	5.3391	0	2.9271	3.3841	0	0.0672	0	0	8.019
16	0.0651	3.5609	1.6991	0	2.8781	7.5572	0	0.2899	0	0	8.414
17	0.0162	3.1685	0	0	9.3776	0	0	0	0	0	9.446
18	0	2.1573	0.3627	0	7.2910	4.2058	0	0	0	0	10.842
19	0	2.9394	0.1730	0	8.7117	4.4558	0	0	0	0	11.077
20	1.0886	4.2011	2.1311	2.9347	2.6736	1.7704	2.3611	0	0.0668	0	11.524
21	0	2.2738	0.1573	0	8.7117	4.4558	0	0	0	0	14.420
22	0.3634	6.0841	4.7899	0	5.7281	6.4683	0	0.1111	0	0	16.609
23	0.1338	4.6990	2.2924	0	9.0909	3.3457	0	0	0	0	17.209
24	0.6582	3.0358	3.7344	0	6.4983	1.4914	0	0	0	0	17.987
25	0.6757	2.4566	3.5701	2.1269	2.7222	3.2222	2.5000	0.0625	1.1111	0	19.400
26	0.4200	1.7605	3.2543	0	10.9855	6.7638	0	0.0672	0	0	19.792
27	0.1545	4.8391	2.7639	0	4.4991	9.9864	0	1.6711	0	0	20.391
28	0.4607	4.7105	3.3587	0	4.1779	4.6989	0	0.0668	0	0	23.622
29	0.0231	2.9542	2.4030	0	6.7044	7.3137	0	0.2899	0	0	24.238
30	1.1232	5.4386	3.2994	2.9187	5.8323	4.9696	2.4535	0.0668	0.0835	0	25.549
31	0.1013	3.2040	2.3950	0	11.3369	7.9385	0	0.1111	0	0	27.434
32	0.0660	3.6165	2.4250	0	11.4139	7.8702	0	0.2500	0	0	28.067
33	0.1013	3.3252	2.4145	0	11.8220	8.5653	0	0.1364	0	0	28.403
34	0.1162	4.8360	2.2803	0	15.2632	3.4041	0	0	0	0	33.444
35	0.0051	3.1418	0	0	17.2415	0	0	0	0	0	36.910
36	0.0039	3.1609	0	0	20.3884	0	0	0	0	0	37.196
37	0.1124	4.8748	2.2779	0	18.3998	3.4193	0	0	0	0	40.021
38	0.3522	4.4980	4.3095	0	16.8468	5.0928	0	0.0672	0	0	40.375
39	0.0035	3.1687	0	0	21.9688	0	0	0	0	0	40.570
40	0.9849	5.0366	10.8361	0	9.0715	13.9499	0	4.9907	0	0	43.114
41	0.0031	3.1756	0	0	23.5532	0	0	0	0	0	44.054

(续表2)

序号	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₀	t _{R(exp)} (min)
42	1.2746	3.7739	1.4479	3.6367	2.7531	2.8505	1.4044	0	0.0657	0	44.618
43	0.6253	3.3108	3.7184	0	18.8124	1.5760	0	0	0	0	46.458
44	0.1100	4.9122	2.2765	0	22.0283	3.4362	0	0	0	0	46.773
45	0.1098	4.9039	2.2765	0	21.5572	3.4303	0	0	0	0	47.158
46	0.0028	3.1818	0	0	25.1410	0	0	0	0	0	49.501
47	0.0025	3.1873	0	0	26.7319	0	0	0	0	0	50.304
48	0.1083	4.9351	2.2756	0	25.6796	3.4406	0	0	0	0	52.147
49	0.1082	4.9302	2.2755	0	25.2075	3.4400	0	0	0	0	52.304
50	0.1081	4.9266	2.2755	0	24.7303	3.4387	0	0	0	0	53.146
51	0.1511	6.8918	5.9937	0	28.0107	12.4486	0	4.6492	0	0	62.566

表3 逐步回归变量分析(n=51)
Table 3 Stepwise regression analysis of variables(n=51)

序号	a ₀	a ₁	a ₂	a ₃	a ₄	a ₅	a ₆	a ₇	a ₈	a ₉	a ₁₀	R	S D	F	R _{CV}	S D _{CV}	F _{CV}
1	-94.496	—	—	—	—	1.846	—	—	—	—	—	0.883	9.270	14.147	0.410	24.332	15.632
2	-139.437	12.120	—	—	—	2.159	—	—	—	—	—	0.918	7.839	21.379	0.782	18.995	20.321
3	-140.786	13.581	—	—	—	2.224	1.180	—	—	—	—	0.914	6.658	31.180	0.855	10.562	25.654
4	-153.295	12.754	1.639	—	—	2.071	0.866	—	—	—	—	0.947	5.930	99.382	0.908	9.651	80.652
5	-11.579	13.410	1.910	—	—	2.040	0.803	—	—	—	-167.760	0.950	5.813	83.298	0.924	7.139	78.564
6	-9.782	9.814	0.964	—	3.825	2.127	1.003	—	—	—	-390.494	0.960	5.291	85.513	0.717	6.993	78.965
7	-9.348	9.407	0.954	—	9.992	2.111	1.004	-8.726	—	—	-287.230	0.966	4.909	86.292	0.717	6.120	80.230
8	-10.957	10.295	1.551	—	14.686	2.083	0.912	-19.149	—	24.882	-417.308	0.976	4.180	106.335	0.809	11.203	92.321
9	-11.881	11.219	1.565	—	14.504	2.108	1.064	-19.258	-0.712	25.249	-411.523	0.977	4.203	93.548	0.825	23.540	83.321
10	-11.223	7.670	1.373	0.696	15.755	2.101	0.894	-19.520	-0.913	23.721	-403.089	0.977	4.230	82.170	—	20.413	70.562

2 线性回归模型的建立

多元线性回归(multiple linear regression,MLR)是一种经典的建模方法,它对自变量和因变量加以线性拟合以得到最小二乘(least square,LS)意义下的最佳结果。借助多元线性回归(MLR)方法对MEDV 矢量和 t_{R(exp)} 进行相关性研究,得到10 变量的预测模型(M₁): t_R=-11.223+7.670X₁+1.373X₂+0.696X₃+15.755X₄+2.101X₅+0.894X₆-19.520X₇-0.913X₈+ 23.721X₉-403.089X₁₀。

该模型的复相关系数R 较高,达到0.977,标准偏差SD 为4.230。以多元线性回归建立的模型对各系列化合物气相色谱保留时间的估计值和误差分别见表4 的Cal₁和Err₁。栏。为了说明模型的稳定性和对外部样本的预测能力,对模型进行了留一法(leave-one-out,LOO)交互检验。留一法交互检验是评价模型稳定性和对外部样本预测能力最重要也最有效的一种方法^[4]。其过程是从样本集n 个样本中每次抽取一个样本作为外部预测集,以其余所有的(n-1)个样本作为训练集建立线性回归模型,并以这个模型预测所抽出的那个外部预测集样本,连续作n 次不重复抽样直到每一个样本均被抽出检验一次也仅一次为止。最后是计算样本集交互检验的预测值与所观测的实验值的复相关系数等统计量,以对模型稳定性和预测能力进行评价。留一法交互检验结果见表4 的Pre_{LOO1}。和Err_{LOO1}。栏。交互检验的结果并不理想,其R²_{cv}

值为负值(-3.189),说明该10 参数模型并不稳定。原因可能是由于只有51 个样本,而自变量数有10 个,样本数太少而自变量数过多造成多元线性回归模型有着显著的过拟合现象。由于自变量之间关系复杂,各个自变量对因变量的作用是不相同的。有的自变量对因变量影响显著,有的则不显著,或者有的自变量单独看来与因变量关系密切,但与其他几个变量合起来求回归方程,该变量在回归方程中的作用又变得无足轻重了,如果还把这样的变量留在回归方程中将增加方程的不稳定性。为了进一步考察MEDV 各矢量对QSRR 建模过程的影响和对模型的贡献大小,对变量进行了逐步回归(SMR)分析,逐步回归的结果见表3、图2 和图3。

从表3、图2 和图3 可以看出,随变量个数m 增加,复相关系数R 逐渐增加,标准偏差SD 逐渐降低;但在交互检验预测过程中,随变量个数m 增加,R_{cv} 先增加,SD 减小,当变量数m 为5 时R_{cv} 达到最大值(0.924),而后变量个数m 继续增加后,R_{cv} 开始减小,而SD_{cv} 增加。综合考虑R、SD、R_{cv} 和SD_{cv},5 变量模型的估计能力比较强,R_{cv} 变化较小,模型的稳定性和预测能力较好。

当m=2 时,R 达到了0.9 以上,但SD 和SD_{cv} 仍然较大。当m=3 时,R_{cv} 增大,SD 和SD_{cv} 都明显减小。用3 个变量对此模型进行建模,所得模型如下:

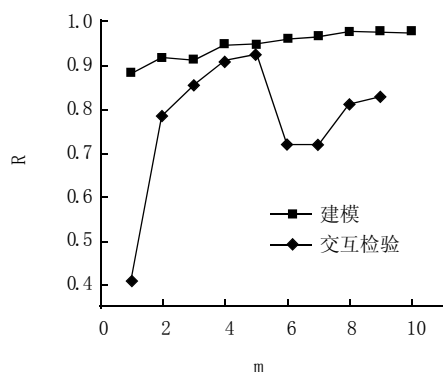


图2 逐步回归结果R随变量个数m的变化曲线
Fig.2 Plot of R versus number parameters with SMR

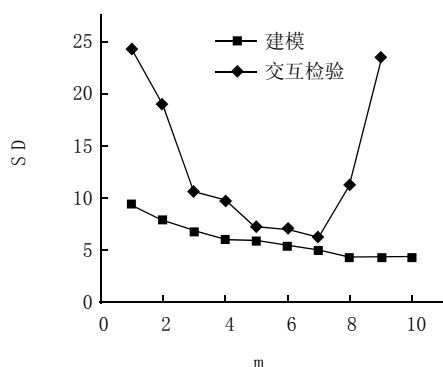


图3 逐步回归结果SD随变量个数m的变化曲线
Fig.3 Plot of SD versus number parameters with SMR

$$t_R = -5.000 + 10.905X_1 + 2.072X_5 + 1.089X_6 \quad (M2)$$

分子建模估计所得结果为: $N=51$, $m=3$, $R=0.914$, $SD=6.658$, $F=31.180$ 。交互检验预测结果为: $N=51$, $m=3$, $R_{CV}=0.855$, $SD_{CV}=10.562$, $F_{CV}=25.654$ 。

当 $m=4$ 时, R 和 R_{CV} 增大, SD 和 SD_{CV} 都减小。用 4 个变量对此模型进行建模, 所得模型如下:

$$t_R = -11.204 + 12.754X_1 + 1.639X_2 + 2.071X_5 + 0.866X_6 \quad (M3)$$

分子建模估计所得结果为: $N=51$, $m=4$, $R=0.947$, $SD=5.930$, $F=99.382$ 。交互检验预测结果为: $N=51$, $m=4$, $R=0.908$, $SD_{CV}=9.651$, $F_{CV}=80.652$ 。

当 $m=6$ 时, R_{CV} 较 5 变量时有所减小, 但 R 继续增大, SD 和 SD_{CV} 都减小。用 6 个变量对此模型进行建模, 所得模型如下:

$$t_R = -9.782 + 9.814X_1 + 0.964X_2 + 3.825X_4 + 2.127X_5 + 1.003X_6 - 390.494X_{10} \quad (M4)$$

分子建模估计所得结果为: $N=51$, $m=6$, $R=0.960$, $SD=5.291$, $F=85.513$ 。交互检验预测结果为: $N=51$, $m=6$, $R=0.717$, $SD_{CV}=6.993$, $F_{CV}=78.965$ 。

当 $m=7$ 时, R_{CV} 不变, 但 R 继续增大, SD 和 SD_{CV} 继续减小。用 7 个变量对此模型进行建模, 所得模型如下:

$$t_R = -9.348 + 9.407X_1 + 0.954X_2 + 9.992X_4 + 2.111X_5 + 1.004X_6 - 8.726X_7 - 287.230X_{10} \quad (M5)$$

分子建模估计所得结果为: $N=51$, $m=7$, $R=0.966$, $SD=4.909$, $F=86.292$ 。交互检验预测结果为: $N=51$, $m=7$, $R_{CV}=0.717$, $SD_{CV}=6.120$, $F_{CV}=80.230$ 。

$m=5$ 时, R_{CV} 达到最大, 建模和预测中的 R 、 SD 、 SD_{CV} 都基本达到了较好值。所以选 5 个变量为最佳变量个数, 用 5 个变量进行建模, 所得模型为:

$$t_R = -11.579 + 13.410X_1 + 1.910X_2 + 2.040X_5 + 0.803X_6 - 167.760X_{10} \quad (M6)$$

分子建模估计所得结果为: $N=51$, $m=5$, $R=0.950$, $SD=5.813$, $F=83.218$ 。交互检验预测结果为: $N=51$, $m=5$, $R_{CV}=0.924$, $SD_{CV}=7.139$, $F_{CV}=78.564$ 。

上述各模型对保留时间的估计和预测值见表 4。这里只给出 10 变量和 5 变量估计和预测误差, 分别见表 4 中的 Err_1 、 Err_{CV1} 、 Err_6 和 Err_{CV6} 栏。10 变量和 5 变量模型对化合物保留时间的估计和预测对实验值作图, 得到图 4。

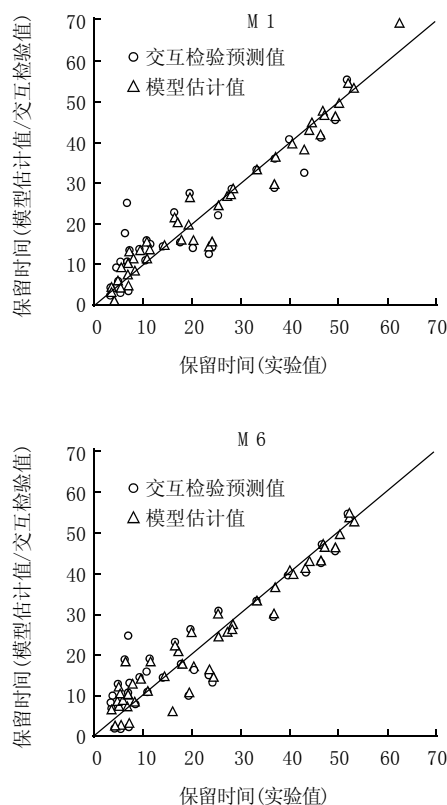


图4 天然山楂香料挥发性成分的色谱保留时间实验值与模型估计值相关图(M1和M6)

Fig.4 Plot of estimated and observed retention time(M1 and M6)

从以上结果可以看出, 虽然用 10 个变量建模得到的复相关系数最高(0.977), 但模型的交互检验预测的复

表4 各模型的估计、预测及误差
Table 4 Estimations and predictions and errors of models

序号	Exp	Cal ₁	Err ₁	Pre _{cv1}	Err _{cv1}	Cal ₂	Pre _{cv2}	Cal ₃	Pre _{cv3}	Cal ₄	Pre _{cv4}	Cal ₅	Pre _{cv5}	Cal ₆	Err ₆	Pre _{cv6}	Err _{cv6}
01	3.573	2.773	-0.800	2.183	-1.390	10.237	11.533	6.616	7.623	3.930	4.074	3.795	3.885	7.158	3.585	8.363	4.790
02	3.638	4.156	0.518	4.199	0.561	6.680	6.845	5.497	5.605	3.740	3.749	3.884	3.905	6.062	2.424	6.212	2.574
03	3.895	4.006	0.111	4.059	0.164	11.030	12.590	7.543	8.880	4.644	4.986	4.479	4.746	8.133	4.238	9.714	5.819
04	4.190	0.391	-3.799	-0.027	-4.217	3.066	2.986	1.708	1.508	0.322	-0.059	0.542	0.181	2.221	-1.969	2.055	-2.135
05	4.926	8.642	3.716	9.241	4.315	11.522	12.025	11.055	11.524	8.164	8.627	8.146	8.607	11.865	6.939	12.450	7.524
06	5.140	5.643	0.503	5.779	0.639	6.736	6.901	6.691	6.954	5.794	5.913	6.085	6.257	7.181	2.041	7.534	2.394
07	5.501	9.953	4.452	10.328	4.827	12.688	13.104	10.085	10.366	10.572	10.890	0.880	11.221	9.996	4.495	10.273	4.772
08	5.542	3.505	-2.037	2.709	-2.833	6.312	6.376	2.197	1.853	2.907	2.625	3.256	3.009	1.992	-3.550	1.624	-3.918
09	5.816	8.549	2.733	8.811	2.995	12.256	12.633	8.583	8.786	9.496	9.789	9.812	10.133	8.326	2.510	8.513	2.697
10	6.458	16.979	10.521	17.521	11.063	18.265	18.606	17.698	18.052	16.319	16.770	6.408	16.863	18.125	11.667	18.515	12.057
11	6.911	6.911	0.000	24.667	17.756	13.936	14.557	16.137	17.396	6.911	24.667	6.911	24.667	6.911	0.000	24.667	17.756
12	6.998	9.812	2.814	10.451	3.453	14.970	15.202	10.398	10.749	10.715	11.106	10.885	11.294	10.167	3.169	10.497	3.499
13	7.109	4.508	-2.601	3.256	-3.853	7.789	7.840	3.181	2.768	4.338	4.017	4.685	4.402	2.849	-4.260	2.395	-4.714
14	7.445	12.643	5.198	13.034	5.589	15.278	15.670	12.732	13.016	13.264	13.587	3.550	13.892	12.616	5.171	12.894	5.449
15	8.019	11.076	3.057	11.465	3.446	13.616	13.895	12.018	12.238	10.088	10.256	0.122	10.292	12.511	4.492	12.771	4.752
16	8.414	7.885	-0.529	7.815	-0.599	9.906	10.047	7.966	7.922	7.991	7.949	8.322	8.313	8.032	-0.382	7.994	-0.420
17	9.446	12.955	3.509	13.483	4.037	14.610	15.029	13.618	14.017	13.381	13.763	13.622	14.028	13.823	4.377	14.244	4.798
18	10.842	11.071	0.229	11.091	0.249	14.691	14.886	11.074	11.090	12.026	12.114	12.321	12.432	10.793	-0.049	10.789	-0.053
19	11.077	15.221	4.144	15.634	4.557	17.907	18.207	15.515	15.725	16.053	16.297	16.317	16.576	15.386	4.309	15.590	4.513
20	11.524	13.309	1.785	15.028	3.504	14.341	14.608	16.634	17.380	23.641	27.846	21.041	24.971	17.918	6.394	18.996	7.472
21	14.420	14.296	-0.124	14.283	-0.137	17.907	18.060	14.424	14.425	15.411	15.481	15.681	15.770	14.115	-0.305	14.095	-0.325
22	16.609	20.965	4.356	22.514	5.905	17.879	17.931	20.865	21.654	18.322	18.769	18.459	18.942	21.791	5.182	22.817	6.208
23	17.209	19.941	2.732	20.236	3.027	18.943	18.996	19.929	20.153	18.756	18.906	18.941	19.109	20.423	3.214	20.697	3.488
24	17.987	15.578	-2.409	15.394	-2.593	17.269	17.239	16.916	16.869	14.924	14.698	14.953	14.729	17.501	-0.486	17.477	-0.510
25	19.400	19.426	0.026	264.702	245.302	11.520	11.189	9.868	9.447	16.376	15.750	7.769	-6.487	10.314	-9.086	9.893	-9.507
26	19.792	25.748	5.956	27.233	7.441	29.714	30.127	25.648	26.528	26.190	27.190	26.258	27.268	25.258	5.466	26.091	6.299
27	20.391	15.383	-5.008	13.753	-6.638	16.886	16.451	16.662	16.125	15.985	15.330	16.242	15.623	16.929	-3.462	16.428	-3.963
28	23.622	14.032	-9.590	12.323	-11.299	13.800	13.477	15.112	14.368	12.881	11.444	13.015	11.595	15.890	-7.732	15.156	-8.466
29	24.238	15.042	-9.196	14.051	-10.187	17.112	16.612	14.151	13.346	14.890	14.103	15.179	14.412	13.922	-10.316	13.093	-11.145
30	25.549	23.863	-1.686	21.836	-3.713	24.749	24.662	28.416	28.948	35.040	38.580	31.459	34.467	29.757	4.208	30.650	5.101
31	27.434	26.435	-0.999	26.298	-1.136	28.246	28.295	25.694	25.557	26.379	26.291	26.558	26.485	25.401	-2.033	25.239	-2.195
32	28.067	26.726	-1.341	26.591	-1.476	27.946	27.939	26.020	25.883	26.526	26.419	26.714	26.620	25.818	-2.249	25.666	-2.401
33	28.403	28.172	-0.231	28.131	-0.272	29.934	30.045	27.440	27.351	28.157	28.133	28.327	28.319	27.125	-1.278	27.005	-1.398
34	33.444	33.008	-0.436	32.977	-0.467	31.606	31.552	32.764	32.727	31.905	31.811	31.993	31.904	33.089	-0.355	33.069	-0.375
35	36.910	29.357	-7.553	28.624	-8.286	30.786	30.341	29.721	29.208	29.976	29.478	30.090	29.600	29.668	-7.242	29.151	-7.759
36	37.196	35.987	-1.209	35.862	-1.334	37.295	37.303	36.255	36.178	36.677	36.633	36.739	36.700	36.109	-1.087	36.019	-1.177
37	40.021	39.635	-0.386	39.607	-0.414	38.081	38.003	39.289	39.250	38.593	38.510	38.629	38.549	39.524	-0.497	39.497	-0.524
38	40.375	40.541	0.166	40.564	0.189	39.301	39.254	39.963	39.945	38.957	38.888	38.927	38.856	40.194	-0.181	40.187	-0.188
39	0.570	39.315	-1.255	39.175	-1.395	40.566	40.565	39.537	39.440	40.043	39.992	40.078	40.031	39.343	-1.227	39.228	-1.342
40	43.114	37.760	-5.354	32.417	-10.697	39.735	38.204	40.479	39.258	38.027	35.304	37.870	35.062	40.952	-2.162	39.939	-3.175
41	44.054	42.651	-1.403	42.477	-1.577	43.845	43.823	42.825	42.693	43.416	43.346	43.425	43.356	42.584	-1.470	42.425	-1.629
42	44.618	44.516	-0.102	-16.045	-60.663	17.710	14.148	19.407	15.705	28.992	19.183	38.997	-10.522	20.625	-23.993	16.609	-28.009
43	46.458	41.643	-4.815	41.032	-5.426	42.522	42.169	42.527	42.163	41.148	40.593	40.982	40.410	42.777	-3.681	42.434	-4.024
44	46.773	47.306	0.533	47.352	0.579	45.593	45.517	46.850	46.855	46.341	46.312	46.318	46.287	46.980	0.207	46.994	0.221
45	47.158	46.298	-0.860	46.227	-0.931	10.237	44.454	45.853	45.770	45.323	45.202	45.308	45.186	45.995	-1.163	45.921	-1.237
46	49.501	45.993	-3.508	45.497	-4.004	6.680	46.843	46.120	45.699	46.797	46.446	46.780	46.427	45.831	-3.670	45.370	-4.131
47	50.304	49.342	-0.962	49.184	-1.120	11.030	50.443	49.420	49.291	50.184	50.165	50.140	50.115	49.084	-1.220	48.903	-1.401
48	52.147	55.000	2.853	55.347	3.200	3.066	53.249	54.433	54.652	54.119	54.308	54.035	54.217	54.454	2.307	54.675	2.528
49	52.304	54.000	1.696	54.196	1.892	11.522	52.153	53.445	53.549	53.108	53.182	53.032	53.099	53.480	1.176	53.587	1.283
50	53.146	52.990	-0.156	52.973	-0.173	6.736	7.360	52.449	52.388	52.087	51.995	52.019	51.921	52.497	-0.649	52.441	-0.705
51	62.566	69.308	6.742	76.297	13.731	12.688	70.668	70.813	74.207	70.417	73.655	70.265	73.442	70.751	8.185	74.120	11.554

相关系数却为负,说明该模型的变量数过多(10个),而样本数相对太少,从而使得模型存在显著的过拟合现象。进行的逐步回归表明,去掉第3个、第4个、第7个、第8个、第9个变量后,选用5个变量时,模型的复相关系数、标准偏差和交互检验的 R_{cv} 值均达到较好的值(分别为 $R=0.950$ 、 $SD=5.813$ 和 $R_{cv}=0.942$),说明模型的稳定性和预测能力更高。由于MEDV矢量是一个二维拓扑描述子,暂时还不能区分顺反异构和旋光异构的化合物结构;并且由于化合物分子的实际结构是立体的,因而用化学键相对键长加和的最小值作为两个原子之间的距离尚有一定局限性。这些问题还有待进一步的探讨。另外,作为输出变量的实验值,必定也有一定的误差,从而可能会影响计算的结果。

3 结 论

MEDV矢量由于其具有直观、简单易行的优点,将其应用于烃类的某些物化性质,如沸点、密度、热容量焓和色谱保留指数等的QSAR/QSPR研究^[6-8]等都得到了满意的结果。本研究对51种天然山楂香料挥发性有机物所建立的QSRR模型,充分考虑了各非氢原子对该类化合物气相色谱保留时间的贡献,且模型所采用的参数均为计算值,简单易得。采用LOO交互检验,对模型稳定性和预测能力进行了评价,结果显示,所得模型具有良好的稳定性和预测能力。通过构建分子

电性距离矢量(MEDV),对天然山楂香料挥发性有机物的气相色谱保留时间进行了定量结构色谱保留相关(QSRR)研究。研究结果表明,MEDV能较好地描述天然山楂香料挥发性有机物分子结构特征,采用MEDV建立的QSRR模型具有预测该类化合物气相色谱保留时间的能力。这对于天然产物中挥发性有机化合物的定量结构色谱保留相关(QSRR)研究具有一定参考价值。

参考文献:

- [1] 张俊松,张文叶,刘程毅,等. 生物技术制备的天然山楂香料的化学成分分析[J]. 精细化工, 2006, 23(10): 977-978.
- [2] 杨学谨,李延东,王善伟,等. 二茂铁衍生物疏水参数的测定及其保留值与结构定量关系的研究[J]. 色谱, 1996, 14(2): 86-90.
- [3] LIU S S, CAO C Z, LI Z L. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector[J]. J Chem Inf Comput Sci, 1998, 38(3): 387-394.
- [4] 刘树深,刘堰,李志良,等. 一个新的分子电性距离矢量(MEDV)[J]. 化学学报, 2000, 58(11): 1353-1357.
- [5] 周润琦,陈石根. 生物化学教程[M]. 北京: 化学工业出版社, 1992: 196-197.
- [6] LIU S S, LIU H L, XIA Z L, et al. Molecular distance-edge vector: an extension from alkanes to alcohols[J]. J Chem Inf Comput Sci, 1999, 39(6): 951-953.
- [7] LIU S S, YIN C S, CAI S X, et al. QSAR study of steroid benchmark and dipeptides based on MEDV-13[J]. J Chem Inf Comput Sci, 2001, 41(2): 321-329.
- [8] 周丽平,夏之宁,李伯玉,等. 多环芳烃分子结构的距边矢量表征及其气相色谱保留指数预测[J]. 色谱, 2001, 19(1): 25-31.