مجله شیمی کوانتومی و اسپکتروسکوپی
۳۳:۱۰ (۱۳۹۹) ٦٤-٧٦
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مدل سازی و مطالعه ارتباط کمی ساختار- بازداری (QSRR) ترکیبات تشکیل دهنده‌ی پوست میوه *Citrus. sinensis* CV. *Thamson* استحصال شده با کروماتوگرافی گازی-طیف سنج جرمی با استفاده از الگوریتم ژنتیک-رگرسیون خطی چند گانه

سعید نکوئی

دانشکده شیمی، دانشگاه صنعتی شاهرود، شاهرود، ایران

تاریخ ثبت اولیه:۱۳۹۹/۰۴/۲۵، تاریخ دریافت نسخه اصلاح شده:۱۳۹۹/۰۶/۱۱ ، تاریخ پذیرش قطعی:۱۳۹۹/۰۶/۲۱

چکیده

مطالعه ارتباط کمی ساختار-بازداری(QSRR) جهت پیش بینی شاخص کواتس ترکیبات تشکیل دهنده پوست میوه *Citrus. sinensis* CV. *Thamson* با استفاده از روش رگرسیون خطی چند گانه(MLR) انجام شد. بعد از استحصال اسانس و تزریق آن به دستگاه GC-MS ترکیبات مختلف آن شناسایی گردید. سپس برای انجام مدل سازی و پیش بینی مقادیر اندیس کواتس(KI) ترکیبات، در ابتدا ساختار ترکیبات، رسم و گروه مناسبی از توصیف کننده‌ها محاسبه شد. سپس از روش انتخاب مرحله‌ای(SW) و الگوریتم ژنتیک(GA) برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با KI ترکیبات مورد نظر داشتند استفاده گردید. برای مدل سازی از روش خطی رگرسیون خطی چند-گانه ساخته شد. داده های آماری نشان می دهد که هر دو روش SW-MLR و GA-MLR پیش بینی های قابل قبولی را ارائه نموده است.

واژه های کلیدی: ***ارتباط کمی ساختار- بازداری، الگوریتم ژنتیک، رگرسیون خطی چند گانه.***

۱. مقدمه

اندیس بازداری پارامتری است که برای جداسازی و شناسایی کمی و کیفی بسیاری از ترکیبات شیمیایی استفاده می‌شود. این خاصیت با روشهایی چون کروماتوگرافی مایع با عملکرد بالا و کروماتوگرافی گازی با آشکارساز طیف سنج جرمی قابل اندازه گیری است[۱]. اندیس بازداری بازتابی از برهمکنش مولکول با فاز ساکن ستون کروماتوگرافی است این برهمکنش وابسته به ویژگی‌های ساختاری مولکول است[۲]. با توجه به اهمیت ترکیبات موجود در روغن های اسانسی باید یک سری اطلاعات کیفی و کمی راجع به این ترکیبات کسب کرد یکی از راه‌های آنالیز این ترکیبات اندازه گیری اندیس بازداری است. با توجه به مشکلات و

*عهده دار مکاتبات: سعید نکوئی
نشانی: دانشکده شیمی، دانشگاه صنعتی شاهرود، شاهرود، ایران
تلفن:۰۲۳۳۲۳۹۴۲۸۹ پست الکترونیک:E-mail:S_nekoei68@yahoo.com

محدودیت های کارهای آزمایشگاهی استفاده از روش های تئوری برای محاسبه خواص و ویژگی های ترکیبات بسیار حائز اهمیت می باشد. امروزه مطالعات کمومتریکس راهی برای استخراج حداکثر اطلاعات از نتایج تجربی با استفاده از انجام یک سری محاسبات آماری و ریاضی می باشد. مطالعات ارتباط کمی ساختار- خاصیت(QSPR) یکی از روش‌های کمومتریکس است که با استفاده از آن می توان روابط خطی و یا غیرخطی بین ویژگی های ساختاری و خواص ترکیبات را پیدا نمود واز روی آن خاصیت موردنظر را برای ترکیبات دیگر پیش‌بینی نمود[۱۰-۳].

هدف از انجام این تحقیق پیش‌بینی اندیس بازداری دسته ایی از ترکیبات موجود در روغن های اسانسی پوست میوه *Citrus. sinensis* CV. *Thamson* با استفاده از روش‌های رگرسیون مرحله ایی و الگوریتم ژنتیک به عنوان روشهای انتخاب متغیر (توصیف کننده) و رگرسیون خطی چندگانه(MLR) به عنوان روش مدل سازی می باشد.

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

سری داده‌ها که برای مطالعه ارتباط کمی ساختار-بازداری انتخاب گردید، مربوط به اندیس کواتس ۲۷ ترکیب از روغن های استحصال شده از پوست میوه *Citrus. sinensis* CV. *Thamson* می‌باشد که درآزمایشگاه تخصصی شیمی تجزیه واحد علوم و تحقیقات استخراج شده است.

در این کار این ترکیبات به صورت تصادفی به دو گروه سری آموزش و سری تست تقسیم شدند(جدول ۱)، سری آموزش شامل ۲۲ مولکول و سری تست شامل ۵ مولکول می‌باشد. مقادیر شاخص بازداری (اندیس کواتس) به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شد. سری آموزش جهت ایجاد یک مدل مناسب و سری تست جهت ارزیابی مدل مورد استفاده قرار گرفت.

۲-۲. محاسبه توصیف‌کننده‌ها

برای محاسبه توصیف گرهای نظری ابتدا ساختارهای مولکولی به کمک نرم افزار Hyper Chem 7 رسم شدند[۱۱]. شکل و پیکربندی مولکولی نقش بسیار مهمی در پیش بینی و توصیف خاصیت یا فعالیت های بیولوژیکی بازی می کنند. بنابراین ابتدا ساختارهای مولکولی به وسیله الگوریتم AM1 بهینه شدند. سپس خروجی نرم افزار Hyper Chem برای هر ترکیب به برنامه Dragon منتقل شده و توصیف گرها محاسبه شدند[۱۵-۱۲]. به این ترتیب تعداد ۱٤۹۷ توصیف گر مولکولی برای هر ترکیب محاسبه شدند.

۲-۳. کاهش تعداد توصیف‌کننده‌ها

با توجه به این که بعضی از متغیرهای مستقل (توصیف کننده ها) ثابت بوده و همچنین برخی دیگر با یکدیگر همبستگی نشان می دهند، لذا به روش زیر بعضی از متغیرها حذف شدند.

۱) توصیف کننده هایی که مقادیر ثابت و یا تقریباً ثابت دارند (بیش از ۹۰٪ داده های ثابت دارند)، حذف می شوند. در این مرحله تعداد ۳۰۵ توصیف کننده کنار گذاشته شدند. بدین ترتیب ۱۱۷۶ توصیف کننده باقی ماند.

۲) با توجه به اینکه در برخی از موارد بعضی از متغیرهای مستقل با یکدیگر همبستگی بالایی دارند، و وجود تنها یکی از این متغیر ها در مدل کردن کافی بوده و نیازی به حضور بقیه متغیر ها نمی باشد. لذا داده ها از این نظر مورد بررسی قرار گرفتند. به منظور حذف متغیرهایی که همبستگی خطی بزرگتر از ۰/۹ با یکدیگر دارند، ماتریس ۱۱۷۷×۱۱۷۷ بین همه متغیرهای مستقل و نیز متغیر وابسته تشکیل شد. متغیرهای مستقلی که همبستگی بیشتر از ۰/۹ با متغیرهای دیگر دارند، بایستی حذف شوند. اما بایستی دقت داشت که هنگام حذف متغیر اضافی، متغیری کنار گذاشته شود که همبستگی کمتری با متغیر وابسته داشته باشد. بدین ترتیب تعداد ۷۷۶ توصیف کننده اضافی کنار گذاشته شدند و تعداد ۴۰۰ توصیف کننده باقی ماندند.

جدول ۱. مقادیر تجربی و محاسبه شده اندیس کواتس برای ترکیبات مختلف برای مجموعه های آموزش و پیش بینی در مدلهای GA-MLR و SW-MLR همراه با مقادیر درصد خطا

No.	Compound	RI(Exp)	SW-MLR	E (%)[d]	GA-MLR	E (%)[d]
1[a]	Octane	800	853	6.74	828	3.50
2[a]	4-methyl-thiazole	819	811	-2.37	848	3.54
3[a]	n-nonane	900	918	2.04	922	2.44
4[a]	α-thujene	930	907	2.69	920	-1.07
5[a]	α-pinene	939	946	2.77	943	0.42
6[b]	Sabinene	975	943	-3.18	945	-3.07
7[a]	Myrcene	991	1017	-3.22	1003	1.21
8[a]	n-octane	999	970	-0.62	934	-6.50
9[a]	α-phellandrene	1003	1021	1.88	1034	3.09
10[b]	n-hexyl acetate	1009	1037	-0.39	1026	1.68
11[a]	Limonene	1029	1018	4.94	1024	-0.48
12[a]	γ-terpinene	1060	1039	-1.95	1049	-1.03
13[a]	Terpinolene	1089	1053	-2.27	1044	-4.13
14[a]	Linalool	1097	1097	-0.89	1076	-1.91
15[b]	undecane	1100	1076	-2.17	1110	0.90
16[a]	Hexyl isobutanoate	1152	1144	0.81	1202	4.34
17[a]	Hexyl butanoate	1193	1164	-2.80	1225	2.68
18[a]	Decanal	1202	1125	-6.32	1123	-6.57
19[b]	Hexyl isovalerate	1244	1239	-1.05	1289	3.61
20[a]	α-copaene	1377	1389	0.009	1412	2.54
21[a]	Tetradecane	1391	1365	-1.84	1397	0.43
22[a]	β-elemene	1400	1469	-2.42	1439	2.78
23[a]	β-caryophyllene	1419	1407	0.87	1407	-0.84
24[b]	Valencene	1496	1511	1.06	1482	-0.93
25[a]	δ-cadinene	1514	1479	-0.81	1487	-1.78
26[a]	Hexadecane	1600	1592	-0.45	1587	-0.81
27[a]	Octadecane	1800	1846	2.60	1791	-0.50

a)سری آموزش b) سری پیش بینی c) درصد خطا

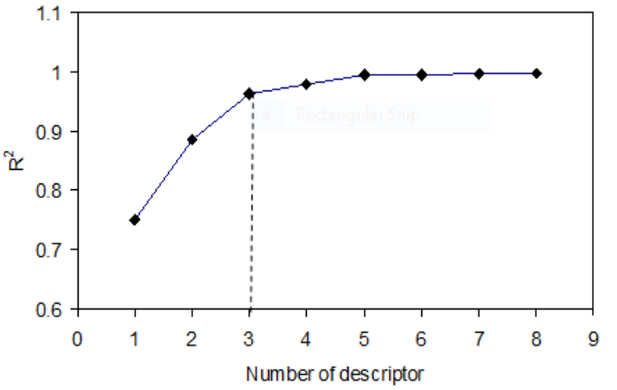
۲-٤. انتخاب توصیف‌کننده‌های مناسب

مهم ترین بخش در ایجاد یک مدل کار آمد، انتخاب توصیف کننده های مناسب است. پس از محاسبه توصیف کننده های مختلف، تعدادی از آنها به عنوان توصیف کننده های مناسب برای ساخت مدل انتخاب می شوند. این مرحله شامل یافتن توصیف کننده های حاوی اطلاعات مفید است به طوری که قدرت پیش بینی مدل در سطح قابل قبولی باشد. در این کار، از دو روش افزایش مرحله ای[1] و الگوریتم ژنتیک[2] برای انتخاب توصیف کننده های مناسب استفاده گردید.

۳. نتایج و بحث

۳-۱. انتخاب مناسب ترین توصیف کننده ها به روش رگرسیون مرحله ایی

مقادیر توصیف‌کننده‌های محاسبه شده توسط نرم‌افزار Dragon بعنوان متغیرهای مستقل و مقادیر انـدیس کـواتس مولکـول‌هـای مورد نظر بعنوان متغیرهای وابسته بعنوان ورودی به نرم افزارSPSS وارد شد. سـپس بـا اسـتفاده از منـوی آنـالیز، گزینـه‌ی رگرسـیون خطی و روش مرحله ای انتخاب و نهایتاً چندین مدل مختلف به‌طور جداگانه به دست آمد، که با توجه به خصوصـیات آمـاری آن‌هـا از جمله ضریب تعیین (R^2)، آماره F و خطای استاندارد و پس از رسم مقادیر R^2 بر حسب تعداد توصیف‌کننـده‌هـا بهتـرین مـدل کـه دارای بیش‌ترین مقدار R^2 و F و کم ترین مقدار خطای استاندارد و شامل توصیف‌کننده‌های تا حد امکان قابل توجیه باشد، بـه‌عنـوان مدل نهایی برای ارتباط اندیس کواتس مولکول‌ها با ساختار آن‌ها انتخاب شد. با این روش مدل سـوم بـا تعـداد ۳ توصـیف کننـده بـه عنوان مناسب ترین آنها انتخاب شد و توسط روش MLR مدل سازی و مورد ارزیابی قرار گرفت . شکل۱تاثیر تعداد توصیف کننـده ها را بر مقدارR^2 نشان می دهد. همانطور که ملاحظه می شود، تغییراتR^2بعد از ۳ توصـیف کننـده خیلـی کـم مـی باشـد. بنـابراین ۳ توصیف کننده جهت مدل سازی انتخاب گردید.



شکل۱. تاثیر تعداد توصیف کننده ها بر مقدارR^2

فهرست توصیف کننده های انتخاب شده توسط نرم افزار SPSS به همراه توصیف مختصری از آنها در جدول۲آورده شده است.

[1] Step-wise
[2] Genetic Algorithem

جدول۲. توصیف کننده های انتخاب شده با SPSS و توصیف آنها

توصیف کننده	نوع توصیف کننده	علامت اختصاری	ضریب
3D-MoRSE-signal 01 / weighted by atomic masses	3D-MoRSE	Mor01m	5.39
3D-MoRSE-signal 18 / weighted by atomic masses	3D-MoRSE	Mor18m	-200.08
R maximal autocorrelation of lag1/unweighted	GETAWAY	$R1u^+$	-284.19
Constant			758.73

۳-۲. ارزیابی توصیف کننده های انتخاب شده

به منظور ارزیابی توصیف کننده های انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول۳ ضرایب همبستگی توصیف کننده های انتخاب شده نسبت به یکدیگر آورده شده است. همانطور که از جدول مشاهده می شود بیشترین ضریب همبستگی بین توصیف کننده Mor01m و توصیف کننده $R1u^+$ با مقدار ضریب همبستگی ۰/۶۰ می باشد این نتایج نشان می دهد که بین توصیف کننده های انتخاب شده همبستگی بالایی وجود نداشته و توصیف کننده ها تقریبا مستقل از هم هستند و نتایج بدست آمده از مدلسازی دال بر وابستگی توصیف کننده ها نمی باشد.

جدول۳. ماتریس ضرایب همبستگی توصیف کننده های انتخاب شده

	Mor01m	Mor18m	$R1u^+$
Mor01m	1.00		
Mor18m	0.12	1.00	
$R1u^+$	0.60	0.35	1.00

۳-۳. ایجاد مدل با استفاده از SW-MLR

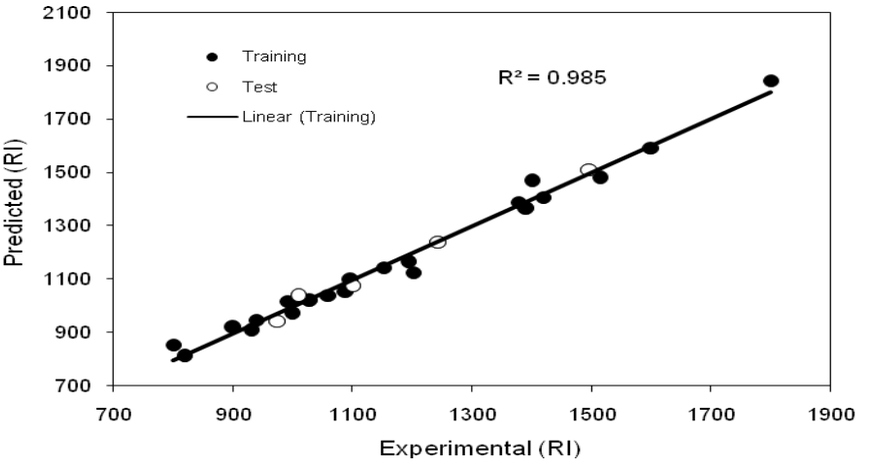
پس از انتخاب مناسب ترین توصیف کننده ها توسط روش مرحله ای با استفاده از SPSS، مرحله بعدی ایجاد مدل، میان توصیف کننده های انتخاب شده و اندیس کواتس ترکیبات می باشد. از نرم افزار SPSS برای این منظور استفاده گردید. بین توصیف کننده ها و شاخص بازداری (اندیس کواتس) ترکیبات سری آموزش با استفاده از روش MLR را بطه زیر بدست آمد:

$KI = 758.73 + 5.39(\mathbf{Mor01m}) - 200.08\ (\mathbf{Mor18m}) - 284.19\ (\mathbf{R1u^+})$

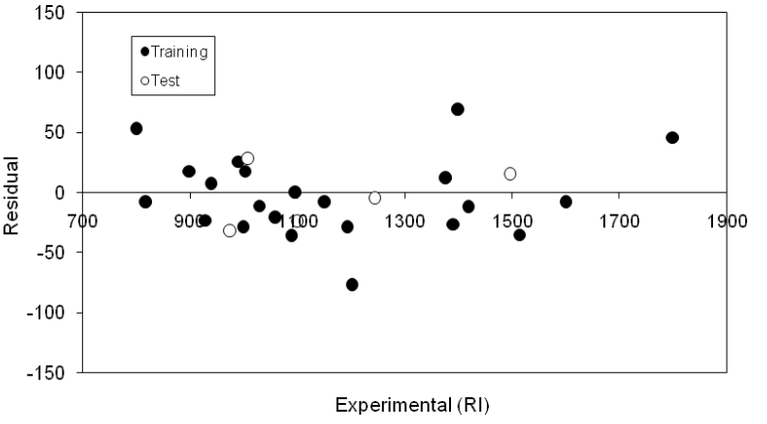
$N = 22;\ R^2 = 0.9835;\ F = 354.82;\ \%REP = 3.85;\ Q^2_{LOO} = 0.9605;\ Q^2_{LGO} = 0.9613$

سپس از معادله بدست آمده برای محاسبه اندیس کواتس سری تست استفاده گردید. مقادیر واقعی و پیش بینی شده اندیس کواتس و همچنین خطاهای پیش بینی برای کلیه ترکیبات مجموعه آموزش و پیش بینی در جدول۱آورده شده است. شکل۲ نمودار مقادیر اندیس کواتس محاسبه شده برای سری آموزش و تست بر حسب مقادیر تجربی را نشان می دهد. در این شکل، میزان نزدیکی داده

ها به خط راست قدرت پیش بینی مدل را نشان می دهد. شکل۳ مقادیر باقیمانده خطاها را نسبت به مقادیر تجربی نشان می دهد. میزان پراکندگی خطاها در اطراف محور نشان دهنده این است که خطای سیستماتیک در مدل وجود ندارد.



شکل۲. نمودار مقادیر اندیس کواتس محاسبه شده با کمک مدلSW-MLR برای مجموعه‌های آموزشی و پیش‌بینی برحسب مقادیر تجربی



شکل۳. نمودار تغییرات خطا برای مقادیر اندیس کواتس محاسبه شده با کمک مدلSW-MLR برای مجموعه‌های آموزشی و پیش‌بینی

۳-۴. انتخاب مناسب ترین توصیف کننده ها توسط الگوریتم ژنتیک

برای حصول نتایج بهتر، از الگوریتم ژنتیک برای انتخاب توصیف کننده های مناسب تر استفاده شد. نرم افزار مورد استفاده برای این کار تحت مطلب نوشته شده است. ابتدا یک سری از توصیف کننده ها به صورت شانسی به عنوان جمعیت یا کروموزوم اولیه در نظر گرفته می شود و به صورت یک رشته با کد گذاری دوتایی (۰ و ۱) مشخص می گردد. تعداد کروموزوم های اولیه توسط کاربر تعیین می شود. با افزایش تعداد اعضاء جمعیت، زمان محاسبه نیز افزایش می یابد. برای یافتن مناسب ترین تعداد جمعیت اولیه عمل

بهینه سازی با ۱۰ بار تکرار برای هر جمعیت اولیه صورت گرفته و کمترین ارزش محاسبه شده برای میانگین ۱۰ بار تکرار، بدست آمد. پس از ایجاد جمعیت اولیه کلیه کروموزوم ها (راه حل ها) ارزیابی می گردد. برای این کار هر راه حل اولیه برای ساخت مدل به کار گرفته می شود. راه حل های بهتر، راه حل هایی هستند که مدل بهتری ساخته و خطای کمتری در پیش بینی سری حذف شده ایجاد کنند. جدول۴ پارامترهای بهینه شده الگوریتم ژنتیک را نشان می دهد. با این روش از بین ۴۰۰ توصیف کننده باقی مانده تعداد ۲ توصیف کننده به عنوان مناسب ترین آنها انتخاب و توسط روش MLR مدل سازی شد و مورد ارزیابی قرار گرفت. لیست توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک به همراه توصیف مختصری از آنها در جدول۵ آورده شده است.

جدول٤. پارامترهای بهینه شده برای الگوریتم ژنتیک

مشخصات الگوریتم ژنتیک	
اندازه جمعیت[1]	۱۵۰
(٪)جمعیت اولیه[2]	۱۰
حداکثر تعداد نسل[3]	۱۰۰
نرخ جهش[4]	۰/۰۵
همگذری	دوتایی

جدول۵. توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک و توصیف آنها

Descriptor	Type of descriptor	Notation	Coefficient	Mean effect
Modified Randic chi-1 index	Topological	XMOD	31.28	0.968
3D-MoRSE-signal 30 / weighted by atomic van der Waals volumes	3D-MoRSE	Mor18p	-169.70	0.031
Constant			78.02	

۳-۵. ارزیابی توصیف کننده های انتخاب شده

به منظور ارزیابی توصیف کننده های انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول ۵ ضرایب همبستگی توصیف کننده های انتخاب شده نسبت به یکدیگر آورده شده است. همانطور که از جدول مشاهده می شود بین توصیف کننده های انتخاب شده همبستگی چندانی وجود نداشته و توصیف کننده ها تقریبا مستقل از هم هستند و نتایج بدست آمده از مدلسازی دال بر وابستگی توصیف کننده ها نمی باشد.

1- Population size
2- Initial terms
3- Max Generation
4- Mutation rate

جدول۵. ماتریس ضرایب همبستگی توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک

	XMOD	Mor18p
XMOD	1	0
Mor18p	0.039	1

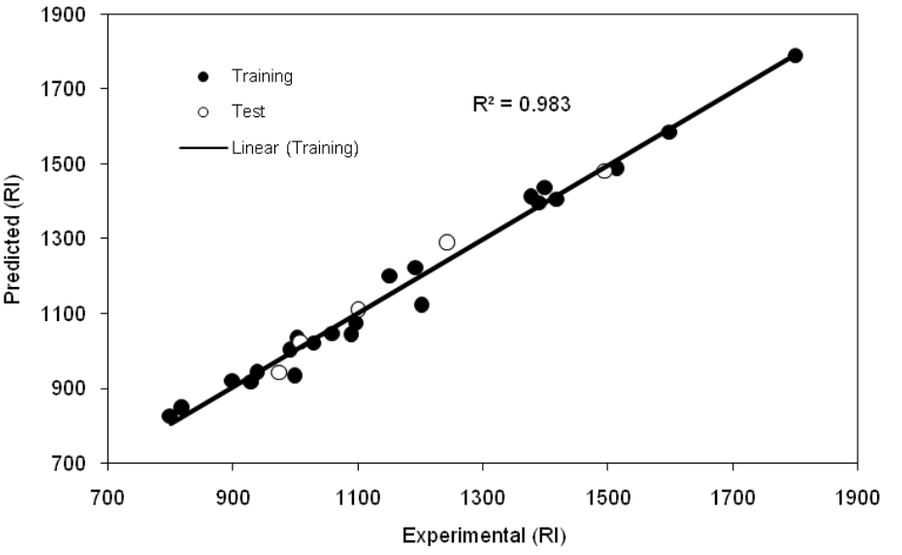
۳-٦. ایجاد مدل با استفاده از GA-MLR

پس از انتخاب مناسب ترین توصیف کننده ها توسط الگوریتم ژنتیک مرحله بعدی ایجاد مدل، میان توصیف کننده های انتخاب شده و اندیس کواتس ترکیبات می باشد. که با استفاده از روش MLR رابطه زیر بدست آمد:

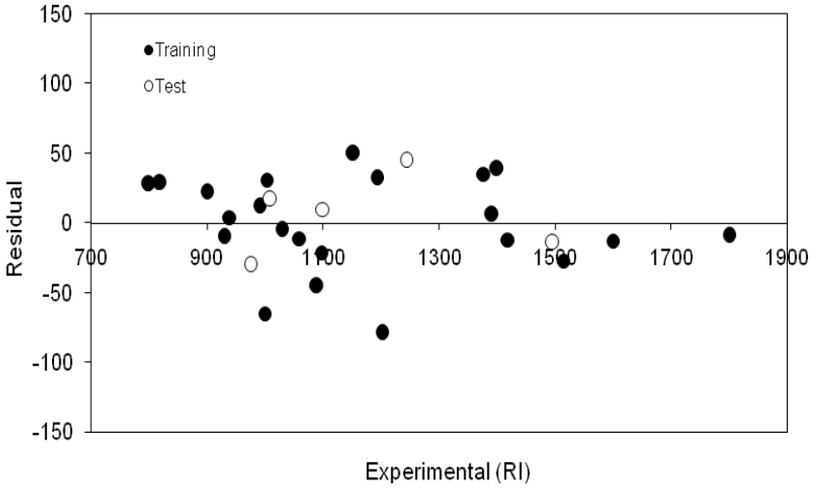
KI=78.02 + 31.28 (**XMOD**) – 169.70 (**Mor18p**)

سپس از معادله بدست آمده برای پیش‌بینی اندیس کواتس سری تست استفاده گردید.

مقادیر واقعی و پیش‌بینی شده اندیس کواتس و همچنین خطاهای پیش‌بینی برای کلیه ترکیبات مجموعه آموزش و تست در جدول۱ آورده شده است. همچنین مقادیر اندیس کواتس محاسبه شده و تجربی برای ترکیبات براساس مدل GA-MLR در دو مجموعه آموزشی و تست در شکل۴ آورده شده است. در این شکل، میزان نزدیکی داده ها به خط راست قدرت پیش‌بینی مدل را نشان می دهد. شکل۵ مقادیر باقیمانده خطاها را نسبت به مقادیر تجربی نشان می دهد. میزان پراکندگی خطاها در اطراف خط صفر نشان دهنده این است که خطای سیستماتیک در مدل وجود ندارد.



شکل٤. نمودار مقادیر اندیس کواتس محاسبه شده با کمک مدل GA-MLR برای مجموعه‌های آموزشی و پیش‌بینی برحسب مقادیر تجربی



شکل۵. نمودار تغییرات خطا برای مقادیر اندیس کواتس محاسبه شده توسط مدل GA-MLR برای مجموعه‌های آموزشی و پیش‌بینی

۳-۷. ارزیابی اعتبار مدل‌های انتخاب‌شده

یکی از مواردی که می‌توان با استناد به آن از معتبر بودن مدل انتخابی اطمینان پیدا کرد، به‌دست‌آوردن ضریب تعیین (R^2) می‌باشد که هرچه این مقدار به یک نزدیک‌تر باشد مدل ما معتبرتر خواهد بود. برای نیل به این امر مقادیر پیش‌بینی شده‌ی اندیس‌های بازداری را برحسب مقادیر تجربی آن رسم می‌کنیم که شکل های۲و۴ مؤید این مطلب می‌باشند. در نمودارهای فوق مقدار ضریب همبستگی، برای مدل SW-MLR برابر۰/۹۸۵ و برای مدل GA-MLR برابر ۰/۹۸۳ است که مقادیر بالای این پارامتر حاکی از اعتبار مدل‌ها می‌باشد.

مورد دیگری که می‌توان برای اعتبار مدل به آن استناد کرد رسم مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی شده‌ی اندیس‌های بازداری بر حسب مقادیر تجربی آن می‌باشد که شکل های۳ و۵ آن را به‌خوبی نشان می‌دهد. در نمودارهای فوق می‌توان پراکندگی نسبتاً یکسان نقاط را حول مقادیر صفر، ناشی از عدم وجود خطای معین در روش و اعتبار مدل دانست.

درصد خطای نسبی برای سری پیش بینی (REP) نیز جهت بررسی اعتبار مدل مورد استفاده قرار می گیرد. هر چه مقدار REP کوچکتر باشد مدل، مدل مناسبتری خواهد بود. مقدار این آماره برای مدلSW-MLR برابر ۱/۹۷٪ و برای مدل GA-MLR برابر۲/۲۷ می باشد که مقادیر کوچک آن اعتبار مدل را نشان می دهد.

یکی دیگر از روش‌ها ارزیابی مدل، روش اعتبارسنجی تقاطعی است. در این روش هر بار یک (LOO) یا تعداد معینی (LGO) از مولکولها کنار گذاشته می شوند و مدل با مولکول های باقی مانده ساخته می شود، سپس کمیت مدلسازی برای مولکول های کنار گذاشته شده، توسط مدل حاصله پیش بینی می شود. این عمل به صورت چرخه ای برای تمام مولکول ها تکرار می شود به طوری که همه مولکولها یک بار در سری پیش بینی قرار می گیرند. پارامتر آماری Q^2 برای ارزیابی مدل حاصل از ارزیابی تقاطعی بکار می

رود. مقدار این پارامتر برای هر دو مدل در جدول۶ آورده شده است که مقادیر بالای آن نشان دهنده اعتبار مدل می باشد. و در نهایت برای اینکه نشان دهیم مدل ارایه شده یک مدل شانسی نیست از روشY-randomization استفاده گردید. در این روش تمام اندیس های کواتس، بطور تصادفی بهم ریخته و سپس مدلسازی و پیش بینی انجام می‌شود. مسلما باید ضریب همبستگی بسیار پایین بدست آید که اگر اینطور شد نشان دهنده این است که مدل اولیه شانسی نبوده و دارای اعتبار می باشد. نتایج ۱۰ بار تکرار در جدول ۶ آورده شده است. همانطور که مشاهده می شود در همه تکرارها ضریب همبستگی بسیار پایین می باشد که نشان دهنده اعتبار مدل ارایه شده و شانسی نبودن مدل می باشد.

جدول٦. مقادیر R^2 و Q^2 برای ۱۰ بار Y-randomization

Iteration	R^2	Q^2
1	0.168	0.052
2	0.068	0.208
3	0.253	0.003
4	0.290	0.001
5	0.203	0.071
6	0.044	0.242
7	0.145	0.093
8	0.200	0.057
9	0.228	0.024
10	0.182	0.009

۳-۷. پارامترهای آماری

جدول۷ پارامترهای آماری مختلف را برای دو مدل SW-MLR و GA-MLR با یکدیگر مقایسه می‌کند. به‌عنوان مثال مقادیر R^2 برای این دو مدل برای مجموعه تست به ترتیب عبارتند از: ۰/۹۸۸ و ۰/۹۸۲ و همچنین درصد خطای پیش بینی برای اندیس کواتس محاسبه شده برای مجموعه تست با کمک این دو مدل به ترتیب ۱/۹۷و۲/۲۷ می‌باشند. همانگونه که مشاهده می‌شود هر دو مدل نتایج بسیار خوبی را ارایه می دهند. هر چند که تفاوت بین دو روش بسیار کم است ولی در روش GA-MLR از ۲ توصیف کننده برای مدلسازی استفاده گردید ولی در روش SW-MLR از ۳ توصیف کننده برای مدلسازی استفاده شد. در هر روش مدلسازی، هرچه تعداد توصیف کننده های استفاده شده کمتر باشد مدل از اعتبار بیشتری برخوردار است و پیچیدگی مدلسازی کمتر خواهد بود. بنابراین توصیه می شود از این مدل‌ها جهت پیش بینی اندیس کواتس سایر ترکیبات نیز استفاده گردد.

جدول۷. مقایسه پارامترهای آماری مختلف برای دو مدل SW-MLR و GA-MLR

Model	R^2_{train}	R^2_{pred}	REP(%)	Q^2_{LOO}	Q^2_{LGO}	RMSEP
SW-MLR	0.985	0.988	1.97	0.974	0.948	22.95
GA-MLR	0.983	0.982	2.27	0.970	0.941	26.49

٤. نتیجه گیری

در این تحقیق برای انتخاب مناسب ترین توصیف کننده ها از دو روش رگرسیون مرحله ایی و الگوریتم ژنتیک استفاده گردید. توسط روش مرحله‌ای، ۳ توصیف‌کننده و توسط الگوریتم ژنتیک ۲ توصیف کننده که بیشترین ارتباط را با اندیس کواتس ترکیبات داشتند انتخاب شدند. جهت مدلسازی از روش رگرسیون خطی چندگانه(MLR) استفاده گردید. نتایج نشان داد که هم روش رگرسیون مرحله ایی و هم الگوریتم ژنتیک به خوبی قادر به انتخاب مناسب ترین توصیف کننده ها بوده است. هر دو روش نتایج خوب و قابل قبولی ارائه داده است و بنابراین از هر دو مدل می‌توان برای پیش بینی اندیس کواتس ترکیبات مشابه استفاده نمود.

٦. مراجع

[1] Gonzalez, F. R., & Nardillo, A. M. (1999). Retention index in temperature-programmed gas chromatography. *Journal of Chromatography A*, *842*(1-2), 29-49.

[2] Stein, S. E., Babushok, V. I., Brown, R. L., & Linstrom, P. J. (2007). Estimation of Kovats retention indices using group contributions. *Journal of chemical information and modeling*, *47*(3), 975-980.

[3] Villaverde, J. J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J. L., & Sandín-España, P. (2018). Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Science of the Total Environment*, *634*, 1530-1539.

[4] Raffetti, E., Donato, F., Speziani, F., Scarcella, C., Gaia, A., & Magoni, M. (2018). Polychlorinated biphenyls (PCBs) exposure and cardiovascular, endocrine and metabolic diseases: a population-based cohort study in a North Italian highly polluted area. *Environment international*, *120*, 215-222.

[5] Li, M., Yu, H., Wang, Y., Li, J., Ma, G., & Wei, X. (2020). QSPR models for predicting the adsorption capacity for microplastics of polyethylene, polypropylene and polystyrene. *Scientific reports*, *10*(1), 1-11.

[6] Sepehri, B. (2020). A review on created QSPR models for predicting ionic liquids properties and their reliability from chemometric point of view. *Journal of Molecular Liquids*, *297*, 112013.

[7] Sepehri, B. (2020). A review on created QSPR models for predicting ionic liquids properties and their reliability from chemometric point of view. *Journal of Molecular Liquids*, *297*, 112013.

[8] Borhani, T. N., García-Muñoz, S., Luciani, C. V., Galindo, A., & Adjiman, C. S. (2019). Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics*, *21*(25), 13706-13720.

[9] Zhu, T., Gu, L., Chen, M., & Sun, F. (2021). Exploring QSPR models for predicting PUF-air partition coefficients of organic compounds with linear and nonlinear approaches. *Chemosphere*, *266*, 128962.

[10] Yan, F., Shi, Y., Wang, Y., Jia, Q., Wang, Q., & Xia, S. (2020). QSPR models for the properties of ionic liquids at variable temperatures based on norm descriptors. *Chemical Engineering Science*, *217*, 115540.

[11] Laxmi, D., & Priyadarshy, S. (2002). HyperChem 6.03. *Biotech Software & Internet Report: The Computer Software Journal for Scientists*, *3*(1), 5-9.

[12] Marleau, G., Hébert, A., & Roy, R. (2011). A user guide for DRAGON Version 4. *Institute of Genius Nuclear, Department of Genius Mechanical, School Polytechnic of Montreal*.
[13] Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, *56*(2), 237-248.
[14] Wang, Y., Chen, J., Tang, W., Xia, D., Liang, Y., & Li, X. (2019). Modeling adsorption of organic pollutants onto single-walled carbon nanotubes with theoretical molecular descriptors using MLR and SVM algorithms. *Chemosphere*, *214*, 79-84.
[15] Réti, T., Sharafdini, R., Dregelyi-Kiss, A., & Haghbin, H. (2018). Graph irregularity indices used as molecular descriptors in QSPR studies. *MATCH Commun. Math. Comput. Chem*, *79*, 509-524.

Modeling and quantitative structure-retention relationship (QSRR) studying of the constituents of *Citrus. sinensis* CV. *Thamson* extracted by gas chromatography-mass spectrometry using genetic algorithm-multiple linear regression

Saeed Nekoei*

Faculty of Chemistry, Shahrood University of Technology, Shahrood, Iran

Submited: 15 July 2020, Revised: 01 September 2020, Accepted: 11 September 2020

Abstract

Quantitative structure-retention relationship (QSRR) study to predict the kovats index (KI) of *Citrus. sinensis* CV. *Thamson* was performed using multiple linear regression (MLR). After extracting the essential oil and injecting it into the GC-MS device, its various compounds were identified. Then, in order to model and predict the quantum index (KI) values of the compounds, first the structure of the compounds, the drawing and the appropriate group of descriptors were calculated. Then the step-wise selection method (SW) and genetic algorithm (GA) were used to obtain the best descriptors that were most related to the KI of the desired compounds. Multiple linear regression was constructed for linear modeling. Statistical data show that both SW-MLR and GA-MLR methods provide acceptable predictions.

Keywords: *Quantitative structure-retention relationship, genetic algorithm, multiple linear regression.*

***Corresponding author :** Saeed Nekoei

Address: Faculty of Chemistry, Shahrood University of Technology, Shahrood, Iran

Tel: 02332394289 **E-mail:** S_nekoei68@yahoo.com