

امروزه استفاده از ابزارها و روش های محاسباتی به خصوص کامپیوتر و الگوریتم های هوشمند در تمامی جنبه های زیستی به امری عادی تبدیل شده است، همچنین این ابزارها و روش ها تمام جنبه های طراحی دارو را نیز تحت تاثیر خود قرار داده است. استفاده از تکنولوژی های اطلاعاتی و مدیریت آنها به عنوان بخش حیاتی در فرآیند طراحی دارو محسوب می شود، از این رو محققین این حوزه که به صورت آزمایشگاهی در طراحی دارو فعالیت دارند می دانند که احتمال کشف و طراحی یک داروی طبیعی که نیاز به تصحیح مجدد نداشته باشد بسیار نادر و کمیاب است. به همین دلیل پیشنهاد می شود که بواسطه به کارگیری علوم کامپیوتری و روش ها و مدل هالی محاسباتی شیمیایی، داروهای جدیدی طراحی شود که هم از نظر زمانی و هم از نظر مالی مقرون به صرفه تر است. در این میان استفاده از روش محاسباتی رابطه کمی ساختار-فعالیت¹ (QSAR) در طراحی دارو نقش مهمی را ایفا می کند. QSAR روشی است که سعی دارد تا ارتباط بین ویژگی های ساختاری ترکیبات و عملکرد آنها را یافته و با استفاده از ابزارهای آماری و تجزیه ای این رابطه را مدل کند. QSAR شامل سه بخش اصلی است که شامل داده های فعالیت (که از اثرات بیولوژیکی می آیند) ، توصیف گرها (که از ساختار شیمیایی می آیند) و روشی که می تواند مدل را فرموله کند، می باشد. بنابراین تغییر در ساختار شیمیایی باعث تغییر در فعالیت بیولوژیکی می شود. مدلی که از QSAR استخراج می شود برای پیش گویی فعالیت های جدید با ساختار جدید به کار می رود. این مدل وقتی کل داده های آزمایشگاهی جمع آوری شد، شروع به محاسبه پارامترهای هر کدام از ساختار های مولکولی موجود می کند و هزاران توصیف کننده مولکولی را مورد آنالیز قرار می دهد که از بین همه آنها فقط تعدادی هستند که با فعالیت بیولوژیکی همبستگی معناداری دارند که آنها از اولویت بالاتری نسبت به بقیه در ایجاد مدل دارند، پس انتخاب توصیف کننده های تاثیرگذار باعث بوجود آمدن مدل قوی برای QSAR خواهد شد. از این رو وجود تعداد زیادی روش انتخاب ویژگی (توصیف کننده) به ما در ساختن مدل با دقت بالا کمک می کند. انتخاب ویژگی به عنوان یک قسمت مهم از طراحی QSAR کمک می کند تا از تعدادی زیادی از ویژگی های موجود در داده های بیولوژیکی، ویژگی هایی که بهترین نتیجه در پیش گویی فعالیت را دارند، استخراج شود. در این مقاله از تکنیک رگرسیون خطی چند گانه² برای ساخت مدل QSAR برای چند پایگاه داده معروف دارویی به همراه الگوریتم ژنتیک، الگوریتم کرم شب تاب و نیز الگوریتم پیشنهادی جدیدی برای انتخاب ویژگی استفاده کرده ایم. هدف اصلی این مقاله ارائه یک ایده جدید به منظور بهبود نتایج حاصل از الگوریتم ژنتیک و کرم شب تاب و نیز ارائه یک الگوریتم جدید انتخاب ویژگی بر پایه مثلث آتش است.

¹ Quantitative structure–activity relationship

² Multiple linear regression

۱. دیتاست:

در این مطالعه یک مجموعه داده ای مشتمل بر ۵ مجموعه داده بیودارویی برای پیاده سازی و تست الگوریتم های انتخاب ویژگی استفاده شده است که همه مجموعه داده ها شامل لیست ساختارها و فعالیت می باشند. جدول شماره یک لیست مجموعه داده ها را نمایش می دهد.

Table 1. Overview of the Datasets Used in the Benchmarking Experiments^a

dataset	end point	n(class)	n(reg)	ref
COX2	cyclooxygenase-2	322	322	22
CPD	carcinogenic categorical activity in rats	1198	567	23
DHFR	dihydrofolate reductase	397	397	22
EPAFHM	fathead minnow acute toxicity	577	577	24
FDA	FDA maximum recommended daily dose	1216	1216	25

^an(class) and n(reg) indicate the number of compounds in each dataset with a classification (class) response and a regression (reg) response, respectively.

متغیرهای موجود و توضیحات مربوط به آنها در یکی از مجموعه داده در جدول شماره دو نشان داده شده است.

Table 2. Variables of the Selwood data set [4]

#	Property ^{a)}	#	Property ^{a)}	#	Property ^{a)}	#	Property ^{a)}
1	ATCH1	15	ESDL1	29	NSDL5	43	MOL_WT
2	ATCH2	16	ESDL2	30	NSDL6	44	S8_1DX
3	ATCH3	17	ESDL3	31	NSDL7	45	S8_1DY
4	ATCH4	18	ESDL4	32	NSDL8	46	S8_1DZ
5	ATCH5	19	ESDL5	33	NSDL9	47	S8_1CX
6	ATCH6	20	ESDL6	34	NSDL10	48	S8_1DY
7	ATCH7	21	ESDL7	35	VDWVOL	49	S8_1CZ
8	ATCH8	22	ESDL8	36	SURF_A	50	LOGP
9	ATCH9	23	ESDL9	37	MOFL_X	51	M_PNT
10	ATCH10	24	ESDL10	38	MOFL_Y	52	SUM_F
11	DIPV_X	25	NSDL1	39	MOFL_Z	53	SUM_R
12	DIPV_Y	26	NSDL2	40	PEAX_X		
13	DIPV_Z	27	NSDL3	41	PEAX_Y		
14	DIPMOM	28	NSDL4	42	PEAX_Z		

^{a)} ATCH1–ATCH10 = partial atomic charges for atoms 1–10; DIPV_X, DIPV_Y and DIPV_Z = dipole vectors; DIPMOM = dipole moment; ESDL1–ESDL10 = electrophilic superdelocalizability for atoms 1–10; NSDL1–NSDL10 = nucleophilic superdelocalizability for atoms 1–10; VDWVOL = van der Waals volume; SURF_A = surface area; MOFL_X, MOFL_Y and MOFL_Z = principal moments of inertia; PEAX_X, PEAX_Y and PEAX_Z = principal ellipsoid axes; MOL_WT = molecular weight; S8_1DX, S8_1DY and S8_1DZ = substituent dimensions; S8_1CX, S8_1CY and S8_1CZ = substituent centers; LOGP = partition coefficient; M_PNT = melting point; SUM_F and SUM_R = sums of the F and R substituent constants [4, 8].

برای ساخت مدل و نیز ارزیابی مدل ایجاد شده توسط رگرسیون چندگانه از داده های همین مجموعه ها به ترتیب ۷۰ و ۳۰ درصد است که به صورت تصادفی انتخاب شده است. که توسط آنها هم مدل ایجاد و هم قابلیت پیش گویی مدل استخراج شده بررسی خواهد شد.

۲. الگوریتم ژنتیک: توضیحات با دکتر مقصود لو

۳. الگوریتم کرم شب تاب : توضیحات با دکتر عرب فرد

۴. الگوریتم آنتی : توضیحات با دکتر مقصود لو

۵. طراحی آزمون:

جهت بررسی هر کدام از الگوریتم های انتخابی ۵ آزمون طراحی شده است که در هر کدام از آنها یک الگوریتم خاص انتخاب ویژگی با تکنیک MLR را برای تمام دیتا ست ها مورد بررسی قرار گرفته است. در جدول زیر اطلاعات هر کدام از آزمون ها را مشاهده می کنید.

توضیح	نام آزمون	شماره آزمون
Simple Genetic Algorithms With MLR	SGAMLR	۱
Advance Genetic Algorithms With MLR	AGAMLR	۲
Simple Firefly Algorithms With MLR	SFAMLR	۳
Advance Firefly Algorithms With MLR	AFAMLR	۴
Fire Algorithms With MLR	FMLR	۵

۶. مدل سازی:

پس از ارزیابی توصیف کننده ها و انتخاب مناسب ترین آنها جهت ساخت مدل با استفاده از روش های آماری مختلفی همچون رگرسیون جستجوی مدل مناسب آغاز می شود تا بتوان ارتباط بین توصیف کننده های انتخابی و پارامتر هدف مدل (فعالیت) را به درستی نشان داد. در این مطالعه مدل رگرسیون خطی چند گانه به عنوان روش آماری پرکاربرد در این صورت مسئله مورد استفاده قرار گرفت. این روش برای مدل سازی رابطه خطی بین یک متغیر وابسته و یک یا چند متغیر مستقل مورد استفاده قرار گرفت که توصیف کننده ها به عنوان متغیرهای مستقل و فعالیت به عنوان متغیر وابسته مدل سازی در نظر گرفته شد که رابطه کلی آن به صورت زیر است:

$$y_i = b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_kx_{i,k}$$

که در آن y_i کمییت فعالیت ، b_0 ثابت رگرسیون ، $x_{i,k}$ مقدار k امین توصیف گر مربوط به i امین ترکیب و b_k ضریب k امین توصیف گر است. در این روش هرچه درجه خطی بودن رابطه بین متغیرهای مستقل با وابسته بیشتر باشد ونیز برهم کنش متغیرهای مستقل با یکدیگر کمتر باشد مدل خروجی بهتری تولید خواهد شد.

۷. ارزیاب گر های مدل

جهت مقایسه مدل ها از آمارهای مختلفی استفاده می شود که در این جا به مهم ترین آن ها اشاره می شود:

۱. پارامتر ضریب همبستگی مدل (R -Squared³): ضریب همبستگی می تواند مقادیر بین صفر تا یک را بپذیرد. هر چه ضریب همبستگی به مقدار یک نزدیک تر باشد درجه خطی بودن رابطه بین متغیر وابسته و متغیرهای مستقل بیشتر می باشد.

$$R^2 = \frac{\sum (Y_{pred} - Y_{mean})^2}{\sum (Y_{actual} - Y_{mean})^2}$$

در این معادله Y_{actual} مقدار تجربی، Y_{pred} مقدار پیش بینی شده، Y_{mean} مقدار میانگین و n تعداد مشاهدات است.

³ Correlation Coefficient

۲. ریشه میانگین مجموع مربع خطا ($RMSE^4$): عاملی است که میزان خطای حاصل از مدل سازی را نشان می دهد. مدلی که RMSE کمتر داشته باشد دارای ارزش بالاتری خواهد بود.

$$RMSE = \sqrt{\frac{\sum (Y_{pred} - Y_{actual})^2}{n}}$$

۳. ارزیابی تقاطعی^۵ (Q^2): این پارامتر معیاری از عمومیت و قدرت پیش بینی مدل است.

$$Q^2 = 1 - \frac{\sum (Y_{pred} - Y_{actual})^2}{\sum (Y_{actual} - Y_{mean})^2}$$

۸. نرم افزار

تمامی آزمون های طراحی شده در نرم افزار متلب طراحی شده اند و کدهای آن در آدرس در دسترس می باشد و همچنین سیستم کامپیوتری مورد استفاده یک دستگاه PC با CPU اینتل چهار هسته ای با رم ۴ گیگابایت و سیستم عامل ویندوز ۷ بوده است.

نتایج:

ابتدا برای هر کدام از مجموعه داده ها سری داده آموزشی و تست به صورت تصادفی جدا شدند و بعد از آن فضای توصیف کننده ها بواسطه اجرای الگوریتم های مشخص شده در هر آزمون کاهش یافت و تعدادی از ویژگی ها جهت ساخت مدل استخراج گردید. در جدول ۴ بهترین توصیف کننده های انتخابی هر الگوریتم برای هر آزمون مشخص شده اند. بعد از انتخاب توصیف کننده ها مدل های QSAR برای هر مجموعه داده با استفاده از MLR ساخته شد که شامل معادلات هر مدل به همراه مقادیر ارزیاب گر ها می باشد را در زیر مشاهده می کنیم.

$$\begin{aligned} \text{Log } 1/EC_{50} = & 3.095 (\pm 1.94) \text{ ATCH4} + 0.766 (\pm 0.46) \text{ ESDL3} \\ & - 0.0137 (\pm 0.0060) \text{ VDWVOL} + 0.000433 \\ & (\pm 0.00022) \text{ MOFI_X} + 0.442 (\pm 0.16) \text{ LOGP} \\ & + 2.618 (\pm 1.76) \end{aligned} \quad (18)$$

$$\begin{aligned} (n = 31 \quad r = 0.905 \quad s = 0.385 \quad F = 22.709 \\ Q^2 = 0.676 \quad s_{PRESS} = 0.515) \end{aligned}$$

$$\begin{aligned} \text{Log } 1/EC_{50} = & 2.609 (\pm 1.69) \text{ ATCH4} - 0.125 (\pm 0.88) \text{ DIPV_X} \\ & - 0.0000734 (\pm 0.000025) \text{ MOFI_Z} + 0.490 \\ & (\pm 0.17) \text{ LOGP} + 1.972 (\pm 0.85) \text{ SUM_F} - 1.815 \\ & (\pm 0.83) \end{aligned} \quad (19)$$

$$\begin{aligned} (n = 31 \quad r = 0.904 \quad s = 0.386 \quad F = 22.476 \\ Q^2 = 0.699 \quad s_{PRESS} = 0.496) \end{aligned}$$

$$\begin{aligned} \text{Log } 1/EC_{50} = & 2.807 (\pm 1.78) \text{ ATCH4} + 0.842 (\pm 0.46) \text{ ESDL3} \\ & - 0.200 (\pm 0.069) \text{ PEAX_X} + 0.500 (\pm 0.16) \\ & \text{LOGP} + 1.791 (\pm 1.22) \end{aligned} \quad (20)$$

$$\begin{aligned} (n = 31 \quad r = 0.880 \quad s = 0.422 \quad F = 22.311 \\ Q^2 = 0.636 \quad s_{PRESS} = 0.535) \end{aligned}$$

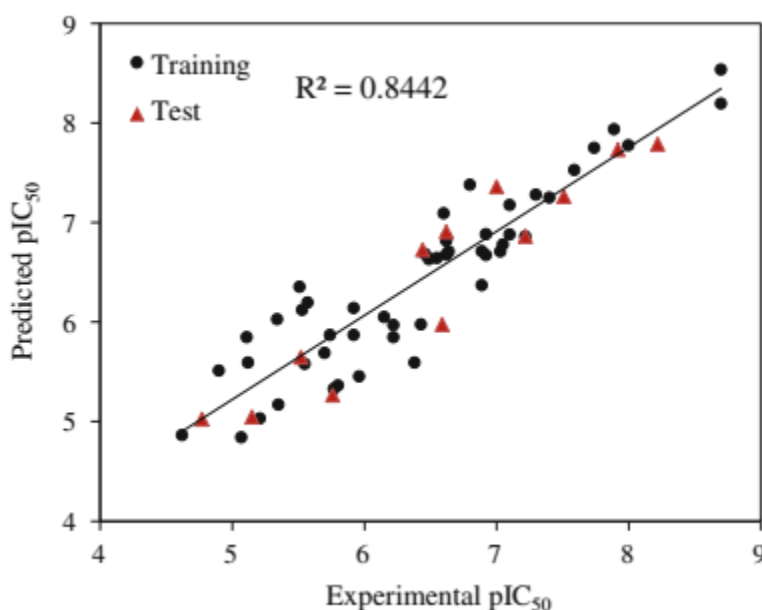
$$\begin{aligned} \text{Log } 1/EC_{50} = & 5.480 (\pm 2.09) \text{ ATCH4} + 21.03 (\pm 11.6) \text{ ATCH5} \\ & + 22.64 (\pm 26.6) \text{ ATCH6} - 0.153 (\pm 0.093) \\ & \text{DIPV_X} - 0.0000559 (\pm 0.000027) \text{ MOFI_Z} \\ & + 0.436 (\pm 0.18) \text{ LOGP} + 4.301 (\pm 6.61) \end{aligned} \quad (21)$$

$$\begin{aligned} (n = 31 \quad r = 0.920 \quad s = 0.362 \quad F = 22.015 \\ Q^2 = 0.721 \quad s_{PRESS} = 0.488) \end{aligned}$$

⁴ Root Mean Square error

⁵ Cross Validation

بعد از استخراج مدل های ساخته شده ، مدل ها با داده های تست مورد ارزیابی قرار گرفتند که در شکل زیر نمایی از رگرسیون استخراج شده برای داده های آموزش و تست را مشاهده می کنید



برای هر کدام از آزمون های موجود و نمایش نتایج حاصله از هر آزمون ارزیابگر های معرفی شده را مورد بررسی قرار دادیم که در جدول زیر مقادیر هر کدام مشخص شده است.

Table 4. Results of the MUSEUM program; stable best models (including up to 6 variables) for the Selwood data set, sorted by their F values

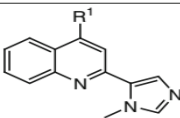
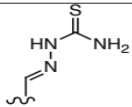
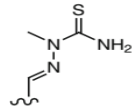
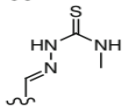
Variables	r	s	FIT	F	Q ²	S _{PRESS}
4, 5, 11, 39, 50	0.909	0.377	2.127	23.818	0.696	0.499
4, 5, 11, 38, 50	0.909	0.377	2.123	23.781	0.696	0.499
38, 50, 52	0.849	0.460	1.745	23.267	0.647	0.518
4, 11, 38, 48, 50, 52	0.924	0.354	2.081	23.240	0.754	0.458
4, 11, 39, 48, 50, 52	0.924	0.354	2.081	23.233	0.751	0.461
4, 11, 38, 47, 50, 52	0.924	0.354	2.077	23.191	0.749	0.463
4, 11, 39, 47, 50, 52	0.923	0.355	2.067	23.087	0.746	0.466
17, 36, 50	0.848	0.462	1.728	23.040	0.644	0.520
39, 50, 52	0.847	0.462	1.720	22.935	0.643	0.520
4, 17, 35, 37, 50	0.905	0.385	2.028	22.709	0.676	0.515
4, 17, 40, 50	0.880	0.422	1.899	22.311	0.636	0.535
4, 5, 35, 37, 42, 50	0.920	0.362	1.973	22.036	0.718	0.490
4, 32, 33, 34, 38, 50	0.919	0.365	1.939	21.647	0.398	0.717
4, 32, 33, 34, 39, 50	0.918	0.366	1.927	21.513	0.414	0.707
4, 5, 40, 50	0.873	0.433	1.774	20.847	0.606	0.557
40, 50	0.773	0.542	1.189	20.812	0.517	0.594
32, 33, 35, 50	0.873	0.433	1.769	20.781	0.369	0.705

نتایج: با مقصود لو و عرب فرد

تفسیر توصیف کننده ها: توضیحات با دکتر قرقانی

بحث: توضیحات با مقصود لو و دکتر قرقانی

نتیجه گیری: توضیحات با مقصود لو و عرب فرد و دکتر قرقانی

No.	General structure	R ₁	R ₂	X	Exp.	Pred.
1			-	-	6.60	7.09
2			-	-	6.92	6.67
3			-	-	5.77	5.33

مراجع: با مقصود لو و عرب فرد