مجله شیمی کوانتومی و اسپکتروسکوپی
۶:۱۹ (۱۳۹۵) ۱۷-۳۳
Journal of Quantum Chemistry and Spectroscopy (JQCS)

روش‌های محاسباتی و مطالعه ارتباط کمی ساختار- خاصیت جهت پیش‌بینی نقاط ذوب ترکیبات نیتروآروماتیک حلقوی با استفاده از توصیف‌کننده‌های شیمیایی و مکانیک‌کوانتومی: ترکیب محاسبات DFT و QSPR

مهدی نکوئی[1*]، مهدی مهام[2]، بهزاد چهکندی[1]، سید محسن بابازاده[1]
[1]دانشگاه آزاد اسلامی، واحد شاهرود، دانشکده علوم پایه، گروه شیمی، شاهرود، ایران
[2]دانشگاه آزاد اسلامی، واحد علی آباد کتول، گروه شیمی، علی آباد کتول، ایران

تاریخ ثبت اولیه: ۱۳۹۵/۰۵/۱۲، تاریخ دریافت نسخه اصلاح شده: ۱۳۹۵/۰۷/۱۸، تاریخ پذیرش قطعی: ۱۳۹۵/۰۸/۲۵

چکیده

روش DFT-B3LYP با سری پایه 6-31G(d) جهت محاسبه چندین توصیف کننده مکانیک کوانتومی در ۶۰ ترکیب نیتروآروماتیک حلقوی بکار گرفته شد. مناسب‌ترین توصیف کننده ها جهت مطالعه ارتباط کمی ساختار- خاصیت در پیش بینی نقاط ذوب ترکیبات نیتروآروماتیک حلقوی با استفاده از توصیف‌کننده های مکانیک‌کوانتومی و شیمیایی و ماشین بردار پشتیبان(SVM) انجام شد. در ابتدا ساختار ترکیبات رسم و گروه مناسبی از توصیف‌کننده‌های شیمیایی و مکانیک‌کوانتومی محاسبه شدند. سپس از روش انتخاب مرحله‌ای برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با نقاط ذوب ترکیبات مورد نظر داشتند استفاده گردید. در ابتدا مدل خطی رگرسیون خطی چند-گانه(MLR) ساخته شد. سپس برای به دست آوردن مدل بهتر، از SVM استفاده گردید. داده های آماری، برتری روش SVM را نسبت به روش خطی MLR نشان می دهد.

واژه‌های کلیدی: ارتباط کمی ساختار- خاصیت، نقطه ذوب، رگرسیون خطی چند گانه، ماشین بردار پشتیبان ، توصیف کنند های مکانیک کوانتمی

۱. مقدمه

جستجو برای ترکیبات پرانرژی جدید با خواص فیزیکی و شیمیایی مطلوب، کارایی و عملکرد بالا و حساسیت کاهش یافته به محرک های خارجی از اهمیت بسیار بالایی در صنایع شیمیایی برخوردار است[۱]. در سالیان اخیر تلاش های مداوم و تحقیقات گسترده ایی در جهت توسعه ترکیبات پرانرژی جدید به علت کاربردهای زیادشان به عنوان پیشران ها در صنایع نظامی و مواد منفجره انجام گرفته است. اما سنتز ترکیبات

***عهده دار مکاتبات:** مهدی نکوئی
نشانی: شاهرود، دانشگاه آزاد اسلامی، گروه شیمی
تلفن: ۳۲۳۹۴۲۸۹- ۰۲۳ **پست الکترونیک:** E-Mail: m_nekoei1356@yahoo.com

پرانرژی همراه با مخاطرات و مضرات فراوانی است بعلاوه بررسی و اندازه گیری اطلاعات و خواص ترکیبات پرانرژی، گران و وقت گیر و گاهی اوقات حتی غیرممکن است. بنابراین توانایی تخمین خواص مطلوب مواد پرانرژی بطور مستقیم از روی ساختار مولکولیشان بسیار جذاب و با اهمیت است. این امر می تواند در تصمیم گیری برای اینکه آیا تلاش برای سنتز این ترکیبات ارزش ویژه ایی دارد یا خیر کمک شایانی خواهد کرد[۱]. اخیرا روشهای مختلفی برای پیش بینی خواص ترموشیمی از قبیل گرمای تصعید، گرمای تشکیل و غیره برای دسته های مختلف ترکیبات پرانرژی معرفی شده است[۳-۲].

پیش بینی نقطه ذوب(Mp) ترکیبات پرانرژی دارای اهمیت زیادی است زیرا نقطه ذوب از خواص فیزیکوشیمیایی پایه ایی است که در شناسایی و تعیین خلوص ترکیبات وهمچنین در محاسبه برخی دیگر از خواص فیزیکوشیمیایی از قبیل فشار بخار و حلالیت کاربرد دارد. دینامیک مولکولی جهت شبیه سازی انتقال فاز جامد به مایع در مواد پر انرژی به منظور پیش بینی نقاط ذوبشان بکار رفته است اما این کار به علت سد انرژی آزاد جهت تشکیل فاز جامد-مایع، نسبتا مشکل و وقت گیر است.

روش ارتباط کمی ساختار- خاصیت[1] (QSPR) یک روش جایگزین برای تخمین نقاط ذوب ترکیبات پرانرژی براساس توصیف کننده های بدست آمده از ساختار مولکولی آنها می باشد. مزیت این روش این است که فقط به فهم و دانش ساختار شیمیایی احتیاج است و به هیچ خاصیت تجربی دیگری وابسته نیست. هدف از مطالعات QSPR پیدا کردن رابطه ای است که بین رفتار فیزیکوشیمیایی یک مولکول با پارامترهای ساختاری آن وجود دارد[۱۰-۴]. نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکولها و ویژگی های ساختاری آنها به پژوهشگران در پیش بینی رفتار مولکولهای جدید براساس رفتار مولکول‌های مشابه کمک می‌کند. کار اصلی در مطالعات ارتباط کمی ساختار-ویژگی، برقراری رابطه و مدلی بین خواص مولکولی مشخص و توصیف‌گرهای مولکولی توسط آمار و یا روش‌های دیگر است. باید توجه شود که مدل ممکن است کاملاً پیچیده باشد، و بنابراین اغلب انواعی از تقریب‌ها به کار می‌رود. تعدادی از روش‌هایی که به طور وسیعی در مطالعات مدل‌سازی ارتباط کمی ساختار-ویژگی استفاده می‌شوند، روش‌های خطی هستند که مدل‌های خطی ساده و قابل تفسیری را می‌سازند، از قبیل روش حداقل مربعات متداول[2] (OLS)، رگرسیون اجزاء اصلی[3] (PCR) و حداقل مربعات جزئی[4] (PLS). برخی دیگر از روش‌ها مدل‌های غیرخطی‌اند، مانند شبکه های عصبی مصنوعی[5] و ماشین بردار پشتیبان[6] که بطور گسترده از آنها استفاده می گردد[۱۷-۱۱].

سهم اصلی برای گسترش استفاده از مدل‌های ارتباط کمی ساختار-ویژگی مربوط است به توسعه توصیف‌گرهای ساختاری و معادلات ریاضی که خواص فیزیکی و شیمیایی را به ساختار شیمیایی مرتبط می‌کنند. موفقیت ارتباط کمی ساختار-ویژگی را می‌توان توسط بینشی که به تخمین ساختاری خواص شیمیایی و امکان تخمین خواص ترکیبات شیمیایی جدید می‌انجامد توضیح داد. یک فرض اساسی در روش ارتباط کمی ساختار-ویژگی این است که همه خواص (فیزیکی- شیمیایی) از یک ماده شیمیایی به ساختار مولکولی مرتبط است. بر اساس این فرض، تحقیقات ارتباط کمی ساختار-ویژگی در دهه‌های اخیر توسعه بسیاری پیدا کرده‌اند.

1. Quantitative Structure-Property Relationship
2. Ordinary Least Squares
3. Principal Component Regression
4. Partial Least Squares
5. Arteficial Neural Networks(ANN)
6. Support Vector Machine (SVM)

هدف از مطالعه حاضر، پیش بینی نقاط ذوب ترکیبات حلقوی نیتروآروماتیک با استفاده از توصیف کننده های شیمیایی و مکانیک کوانتمی و ماشین بردار پشتیبان است.

۲. بخش محاسباتی

۲-۱. سری داده ها

سری داده‌ها مربوط به نقاط ذوب ۶۰ ترکیب از نیتروآروماتیک های حلقوی است که توسط وانگ وهمکارانش گزارش شده است[۱۸]. ساختار این ترکیبات در جدول ۱ آورده شده است. در ابتدا ترکیبات به صورت تصادفی به دو سری شامل سری آموزش و سری پیش‌بینی تقسیم شدند(جدول۱). سری آموزش شامل ۳۵ ترکیب (۶۰٪ داده ها) و سری پیش‌بینی یا تست نیز شامل ۱۱ ترکیب(۲۰٪ داده ها) است. مقادیر نقاط ذوب به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند.

۲-۳. محاسبه توصیف کننده‌ها

توصیف کننده ها مقادیر عددی هستند که خصوصیات مختلفی از مولکول را بیان می‌کنند. در حال حاضر تعداد زیادی توصیف کننده مولکولی وجود دارد که در مطالعات QSPR مورد استفاده قرار می‌گیرند. بعد از ارزیابی و پیدا کردن مناسب‌ترین آنها می‌توانند جهت پیش بینی خاصیت ترکیبات جدید بکار روند.

محاسبه توصیف کننده های الکترونی بوسیله بسته نرم افزاری Gaussian03W انجام شد. ساختار هندسی ۶۰ ترکیب حلقوی نیتروآروماتیک به روش DFT با تابع B3LYP و سری پایه 6-31G(d) بهینه شد. سپس چندین پارامتر ساختاری مرتبط (توصیف کننده) از نتایج محاسبات کوانتومی انتخاب شدند. که مشخصات این توصیف کننده ها در جدول ۲ آورده شده است. برای محاسبه بقیه توصیف کننده‌ها از نرم‌افزار دراگون استفاده گردید.

جدول۲. مشخصات برخی از توصیف کننده های محاسبه شده.

توصیف کننده‌ها	نماد	اختصار	توصیف کننده‌ها	نماد	اختصار
توصیف کننده‌های مکانیک کوانتومی	Highest Occupied Molecular Orbital	HOMO	توصیف کننده‌های مکانیک کوانتومی	Hardness [η=1/2 (HOMO+LUMO)]	H
	Lowest Unoccupied Molecular Orbital	LUMO		Softness (S=1/ η)	S
	difference between LUMO and HOMO	E GAP		Electro negativity [χ= -1/2 (HOMO–LUMO)]	χ
	Molecular Polarizability	MP		Electrophilicity (ω=χ2 /2 η)	Ω
ویژگی‌های شیمیایی	Molecule surface area (Approx)	SA (A)	ویژگی‌های شیمیایی	Partition Coefficient	Log P
	Molecule surface area (Grid)	SA (G)		Hydration Energy	HE
	Mass	M		Refractivity	REF
	Molecule volume	V			

۲-٤. ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از روش های یادگیری تحت نظارت است که هم برای دسته بندی و هم رگرسیون قابل استفاده است. این روش توسط وپنیک [۱۹] بر پایه تئوری یادگیری آماری بنا نهاده شده است. ماشین بردار پشتیبان روشی برای طبقه بندی دوتائی در فضای ویژگی های دلخواه است و از این رو روشی مناسب برای مسائل پیش بینی به شمار می رود[۲۰]. ماشین بردار پشتیبان در اصل یک دسته بندی کننده دو کلاسه است که کلاس ها را توسط یک مرز خطی از هم جدا می کند. در این روش نزدیکترین نمونه ها به مرز تصمیم گیری را بردارهای پشتیبان می نامند. این بردارها معادله مرز تصمیم گیری را مشخص می کنند. الگوریتم های شبیه سازی معمولا هوشمند کلاسیک مانند شبکه های عصبی مصنوعی ، معمولا قدر مطلق خطا یا مجموع مربعات خطای داده های آموزشی را حداقل می کنند، ولی مدل های SVM اصل حداقل سازی خطای ساختاری را به کار می گیرند [۲۱].

در یک مدل رگرسیونی SVM لازم است وابستگی تابعی متغیر وابسته y به مجموعه ای از متغیرهای مستقل X تخمین زده شود. فرض بر این است که مانند دیگر مسائل رگرسیونی ، رابطه بین متغیرهای وابسته و مستقل توسط یک تابع معین f به علاوه یک مقدار اضافی نویز مشخص می شود.

$$y = f(x) + Noise \qquad (۱)$$

بنابراین موضوع اصلی ، پیدا کردن فرم تابع f است که بتواند به صورت صحیح ، موارد جدیدی را که SVM تاکنون تجربه نکرده است پیش بینی کند. این تابع به وسیله آموزش مدل SVM بر روی یک مجموعه داده به عنوان مجموعه آموزش که شامل فرآیندی به منظور بهینه سازی دائمی تابع خطا است. قابل دسترسی است. بر مبنای تعریف این تابع خطا، دو نمونه از مدل های SVM شناخته شده است که عبارتند از الف) مدل های رگرسیونی SVM نوع اول که مدل های SVM-v نیز نامیده می شوند و ب) مدل های رگرسیونی SVM نوع دوم که با نام SVM– ε شناخته شده هستند. در این مطالعه SVM– ε به دلیل کاربرد گسترده آن در مسائل رگرسیونی مورد استفاده قرار گرفت. برای این مدل ، تابع خطا به صورت زیر تعریف می شود :

$$\frac{1}{2}\mathrm{W^T W} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}}\delta_{\mathrm{i}} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}}\delta_{\mathrm{i}}^{*} \qquad (۲)$$

تابع خطای فوق لازم به ذکر است که با توجه به محدودیت های زیر کمینه گردد (۱۳):

$$W^T\emptyset(X_i) + b - y_i \le \varepsilon + \delta_i^* \qquad (۳)$$

$$y_{\mathrm{i}} - \mathrm{W^T}\emptyset(\mathrm{x_i}) - b \le \varepsilon + \delta_i^*$$
$$\delta_{\mathrm{i}}\,,\delta_{\mathrm{i}}^{*} \ge 0$$

که در این رابطه C ثابت گنجایش ، W بردار ضرایب ، W^{T} ترا نهاده بردار ضرایب ، $\delta_{\mathrm{i}}\,,\delta_{\mathrm{i}}^{*}$ ضرایب کمبود ، b ضریب ثابت ، N الگوهای آموزش مدل و Ø تابع کرنل است. اطلاعات کمی در مورد انتخاب تابع غیر خطی مناسب Ø در دسترس می‌باشد. ماشین های بردار پشتیبان برای حل مسائل غیر خطی ، ابعاد مسئله را از طریق توابع کرنل تغییر می‌دهند. انتخاب کرنل برای SVM که حجم داده های آموزشی و ابعاد بردار ویژگی بستگی دارد. به عبارت دیگر ، بایستی با توجه به این پارامترها تابع کرنلی را انتخاب نمود که توانایی آموزش برای ورودی های مساله را داشته باشد. در عمل چهار نوع کرنل خطی ، کرنل چهار جمله ای ، کرنل تانژانت هیپربولینک و کرنل گوشی (RBF) به کار گرفته می‌شوند. در جدول ۳ معادلات برخی از کرنل های رایج ارائه شده اند.

جدول ۳. توابع کرنل رایج در ماشین های بردار پشتیبان

نوع تابع	تابع کرنل
خطی	$K(x_i, x_j) = x_i^T, x_j$
چند جمله ای	$K(x_i, x_j) = (\gamma x_i^T . x_j + C)d$
تانژانت هیپربولیک	$K(x_i, x_j) = tanh(\gamma x_i^T . x_j + C)$
RBF	$K(x_i, x_j) = exp(-\gamma \lvert x_i - x_j \rvert 2)$

در نهایت ، تابع تصمیم رگرسیون بردار پشتیبان غیر خطی به صورت معادله زیر خواهد بود که کنترل کننده میزان نوسان تابع گوسی و همچنین کنترل کننده نتایج پیش بینی و تعمیم دهنده مدلSVM است [۲۲].

$$f(x_i) = \sum_{i=1}^{I} (-\partial_i - \partial_i^*) K(x_i, x_j) + b \qquad (۴)$$

۳. نتایج و بحث

۳-۱. مدل‌سازی با روش رگرسیون خطی چندگانه

برای ساختن مدلی که بیانگر ارتباط ساختاری ترکیبات مورد بررسی با نقطه ذوب (Mp) آنها باشد، از روش رگرسیون خطی چندگانه(MLR) استفاده شد. در ابتدا توصیف‌کننده‌های محاسبه شده، به عنوان متغیرهای مستقل و مقادیر Mp ترکیبات مورد نظر به عنوان متغیرهای وابسته، به عنوان ورودی به نرم افزارSPSS وارد شدند. در نهایت با استفاده از منوی آنالیز، گزینه‌ی رگرسیون خطی و روش مرحله‌ای انتخاب و نهایتاً چندین مدل مختلف بطور جداگانه به دست آمد، که با توجه به خصوصیات آماری آنها از جمله ضریب رگرسیون(R)، آمارهF و خطای استاندارد و پس از رسم مقادیر R^2_{train} ، R^2_{test} ، $RMSE_{train}$ و $RMSE_{test}$ بر حسب تعداد توصیف‌کننده‌ها، بهترین مدل که دارای بیش‌ترین مقدار R^2 و F و کم ترین مقدار خطای استاندارد و شامل توصیف‌کننده‌های تا حد امکان قابل توجیه باشد، به‌عنوان مدل نهایی برای ارتباط Mp ترکیبات با ساختار آن‌ها انتخاب شد. با این روش ۱۰ مدل بررسی شد که مدل هفتم با تعداد ۷ توصیف کننده به عنوان مناسب ترین آنها انتخاب و توسط روش MLR مدل سازی و مورد ارزیابی قرار گرفت. فهرست توصیف کننده های انتخاب شده توسط نرم افزار SPSS به همراه توصیف مختصری از آنها در جدول۴ آورده شده است. جدول ۵ نیز مقادیر عددی توصیف کننده ها را نشان می دهد.

جدول ٤. توصیف کننده های انتخاب شده با SPSS وتوصیف آنها.

توصیف کننده	نوع توصیف کننده	ضرایب	MF	VIF
MATS3e	2D autocorrelations	-298.37	0.076	1.099
Mor02m	3D-MoRSE descriptors	14.09	0.746	1.142

Mor32e	3D-MoRSE descriptors	96.74	-0.057	1.240
G2v	WHIM descriptors	92.80	0.090	1.079
H5m	GETAWAY descriptors	51.97	0.025	1.028
nHDon	Functional groups	28.10	0.085	1.032
E_{Gap}	Mechanic quantum descriptors	80.72	0.033	1.216
Constant	-	133.98	-	-

پس از انتخاب مناسب ترین توصیف کننده‌ها توسط روش مرحله ای با استفاده از SPSS، مرحله بعد، ایجاد مدل بین توصیف کننده های انتخاب شده و Mp ترکیبات است. از نرم افزار SPSS برای این منظور استفاده گردید و بین توصیف کننده‌ها و Mp ترکیبات سری آموزش با استفاده از روش MLR رابطه زیر به دست آمد:

$$Mp = 133.98 - 298.37\,(MATS3e) + 14.09\,(Mor02m) + 96.74\,(Mor32e) + 92.80\,(G2v) + 51.97\,(H5m) + 28.10\,(nHDon) + 80.72\,(E_{Gap})$$

$R^2_{train}=0.804$, $F_{train}=23.54$, $R^2_{test}=0.865$, $F_{test}=4.022$, $R^2_{adj}=0.770$, $Q^2_{LOO}=0.721$, $Q^2_{LGO}=0.594$

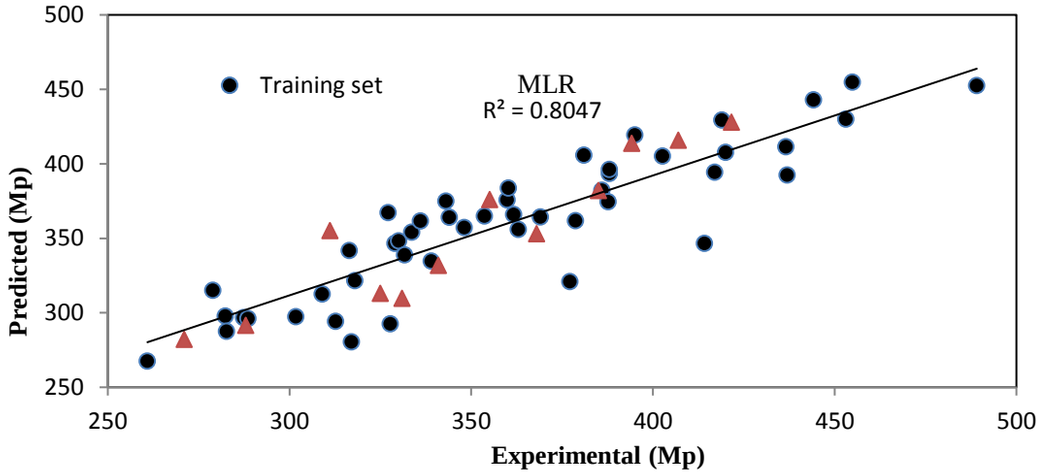
جدول۵. مقادیر عددی توصیف کننده ها.

MATS3e	Mor02m	Mor32e	G2v	H5m	nHDon	E_{Gap}
0.054	9.902	-0.100	0.188	0.029	0	0.0097
-0.003	9.787	-0.144	0.416	0.000	0	0.2177
-0.009	10.866	-0.123	0.193	0.016	0	0.0150
-0.013	10.847	-0.233	0.188	0.026	0	0.0083
-0.007	10.596	-0.268	0.188	0.039	0	0.2193
-0.023	9.930	-0.169	0.197	0.000	0	0.1671
-0.032	10.648	-0.176	0.188	0.007	0	0.0366
0.165	11.283	0.051	0.204	0.000	1	0.2097
0.024	10.079	-0.238	0.193	0.024	1	0.0138
0.058	13.795	-0.171	0.171	0.040	1	0.0152
-0.002	10.906	-0.300	0.181	0.060	0	0.0149
-0.010	10.901	-0.178	0.204	0.000	1	0.0147
-0.157	11.760	-0.147	0.193	0.011	0	0.2027
0.032	11.382	-0.143	0.188	0.025	0	0.0356
-0.032	11.819	-0.163	0.191	0.489	0	0.1117
-0.016	11.671	-0.181	0.191	0.695	0	0.1077
0.005	10.740	-0.087	0.193	0.000	1	0.2165
-0.064	11.073	0.126	0.183	0.007	0	0.1919
-0.047	13.052	-0.219	0.217	0.029	1	0.0146
-0.062	12.241	-0.143	0.191	0.105	0	0.0043
-0.040	11.605	-0.120	0.191	0.018	2	0.0278
0.009	10.491	-0.064	0.200	0.013	2	0.1958

0.016	14.571	-0.256	0.175	0.046	1	0.0088
-0.168	12.427	-0.171	0.191	0.019	0	0.0429
-0.150	13.032	-0.080	0.188	0.057	0	0.0083
-0.093	12.475	-0.161	0.188	0.010	1	0.1948
0.026	11.770	-0.036	0.191	0.057	2	0.0053
-0.106	11.676	-0.101	0.200	0.000	0	0.2100
-0.109	11.886	-0.133	0.172	0.439	0	0.0552
0.005	11.456	-0.234	0.193	0.010	1	0.0387
0.107	16.522	-0.280	0.165	0.051	1	0.0982
-0.155	13.688	-0.211	0.197	0.094	1	0.0226
-0.091	10.046	-0.086	0.200	0.000	2	0.1627
-0.035	11.611	-0.089	0.242	0.834	0	0.1149
-0.155	12.509	-0.127	0.197	0.030	1	0.0162
-0.122	14.205	-0.016	0.236	0.000	0	0.0657
-0.203	13.742	-0.176	0.191	0.027	1	0.0114
-0.174	13.075	-0.120	0.177	0.487	0	0.0602
-0.084	11.265	-0.186	0.197	0.000	1	0.0042
-0.094	12.342	-0.176	0.197	0.388	1	0.1088
-0.067	10.847	-0.193	0.197	0.631	3	0.0715
-0.091	10.608	-0.040	0.310	0.005	2	0.1943
-0.107	12.290	-0.084	0.197	0.264	0	0.6009
-0.064	12.682	0.150	0.322	0.023	0	0.1861
-0.106	13.451	-0.059	1.000	0.000	0	0.0070
-0.113	12.275	-0.006	0.193	0.016	2	0.1848
-0.199	14.760	-0.249	0.188	0.063	2	0.0089
-0.294	13.793	-0.143	0.185	0.080	1	0.0113
0.003	10.351	-0.179	0.197	0.024	0	0.0107
0.011	10.442	-0.083	0.188	0.055	0	0.0163
-0.157	10.980	-0.142	0.193	0.035	0	0.1686
-0.023	10.491	-0.175	0.255	0.008	0	0.2129
-0.028	11.464	-0.166	0.193	0.024	0	0.0335
-0.078	12.261	-0.192	0.191	0.035	0	0.0087
-0.153	12.946	-0.145	0.185	0.113	0	0.0594
-0.158	10.123	-0.198	0.204	0.000	1	0.0166
-0.158	10.716	-0.140	0.204	0.001	1	0.2027
-0.226	12.559	0.005	0.193	0.000	0	0.2100
-0.202	11.999	-0.151	0.197	0.226	1	0.1107
-0.063	13.369	-0.012	0.191	0.017	2	0.1628

سپس از معادله به دست آمده برای پیش بینی Mp سری پیش بینی (تست) استفاده گردید. شکل ۱ نمودار مقادیر Mp محاسبه شده با کمک مدل SW-MLR برای مجموعه‌های آموزش و پیش‌بینی را برحسب مقادیر تجربی نشان می دهد.

با توجه به پارامترهای آماری، هر چند روش رگرسیون خطی چندگانه توانسته پیش بینی نسبتا مناسبی را نشان دهد اما برای حصول نتایج بهتر، از ماشین بردار پشتیبان(SVM) برای پیش بینی Mp ترکیبات استفاده شد.



شکل ۱. نمودار مقادیر Mp محاسبه شده با کمک مدلSW-MLR برای مجموعه‌های آموزش و پیش‌بینی را برحسب مقادیر تجربی.

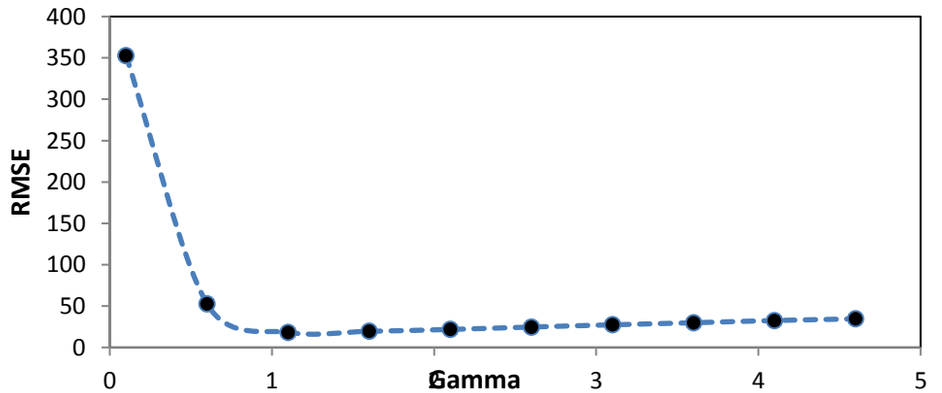
۳-۲. مدل‌سازی با استفاده از ماشین بردار پشتیبان

در این مرحله توصیف کننده‌های انتخاب شده برای ایجاد مدل و پیش‌بینی فعالیتهای بازدارندگی داروها توسط روش غیر خطی ماشین بردار پشتیبان مورد بررسی قرار گرفت . در این روش قبل از شروع مدل سازی لازم است یک سری از پارامترها بهینه شود. پارامترهایی که روی قدرت مدلسازی SVM تاثیر می گذارند عبارتند از:

kernel function type, capacity parameter (C), ε -insensitive (ε), Gamma (eps)

در ابتدا باید نوع تابع مورد استفاده بهینه شود که نوع تابع، نحوه توزیع نمونه ها را در فضا مشخص می کند. RBF یکی از توابعی است که به طور معمول استفاده می شود و نتایج خوبی نیز می دهد در این پروژه از RBF به عنوان تابع برای SVM استفاده گردید.

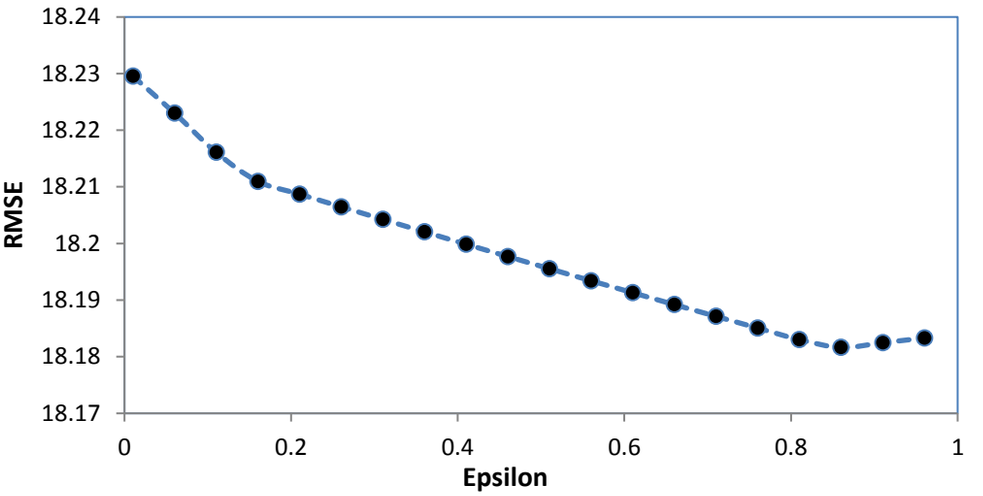
بعلاوه پارامتر متناظر با نوع تابع یعنی Gamma که روی تعداد بردارهای پشتیبان تاثیر می گذارد نیز باید بهینه شود. تعداد بردارهای پشتیبان بر زمان آموزش مدل تاثیر می گذارد طوری که افزایش مقدار Gamma و در نتیجه تعداد بردار پشتیبان می تواند به افزایش زمان آموزش و همچنین overfitting منجر شود. مقدار Gamma توانایی و قدرت SVM را در پیشگویی کنترل می کند. در شکل ۲ نمودار مقادیر متفاوت Gamma بر حسب RMSE نمایش داده شده است.



شکل ۲. نمودار تغییرات مقدار Gamma برحسب مقدار RMSE برای سری آموزش

همانطوری که از این شکل ملاحظه می شود مقدار Gamma از ۰/۱ تا ۴/۶ تغییر می یابد و از نقطه ۱/۱ به بعد با افزایش Gamma مقدار RMSE افزایش می یابد. بنابرین مقدار ۱/۱ به عنوان نقطه بهینه برای Gamma انتخاب شد.

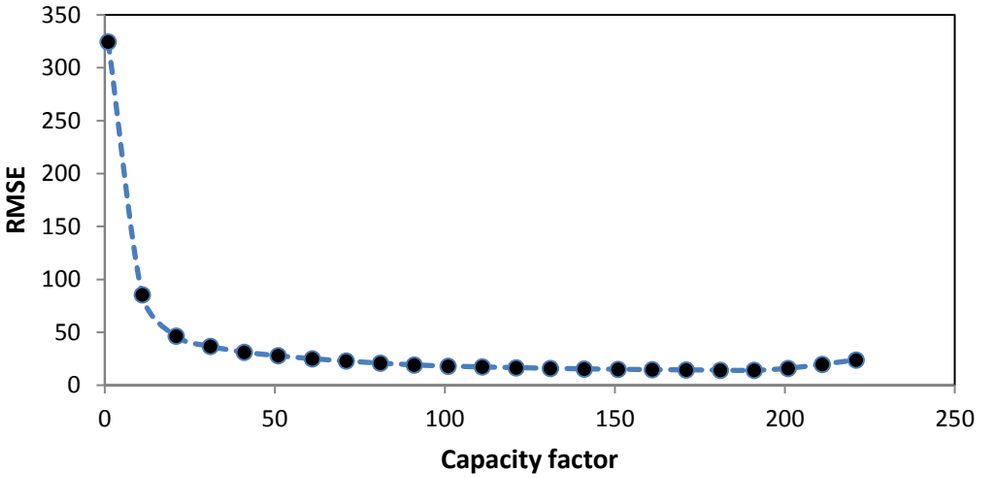
پارامتر ε-insensitive parameter یا فاکتور حساسیت یکی دیگر از پارامترهایی است که باید بهینه شود. فاکتور حساسیت به نویز های موجود در داده ها مربوط می شود که معمولا ناشناس هستند. در شکل ۳ نمودار تغییرات(ε) epsilon بر حسب RMSE نمایش داده شده است.



شکل ۳. نمودار تغییرات مقدار(ε) epsilon برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار epsilon از ۰/۰۰۱ تا ۰/۶۹ مقدار بهینه ۰/۸۶ برای epsilon انتخاب شد.

و در نهایت پارامتر ظرفیت C باید بهینه شود اگر مقدار C پایین باشد یک پراکندگی در پیشگویی دیده خواهد شد و در بعضی اوقات با افزایش بیش از حد C، Overfitting رخ می دهد هرچند که مقدار زیادی C تاثیر چندانی روی پیشگویی ندارد ولی با این حال مقدار این پارامتر نیز باید بهینه شود. در شکل ۴ نمودار تغییرات(C) Capacity parameter بر حسب RMSE نمایش داده شده است.



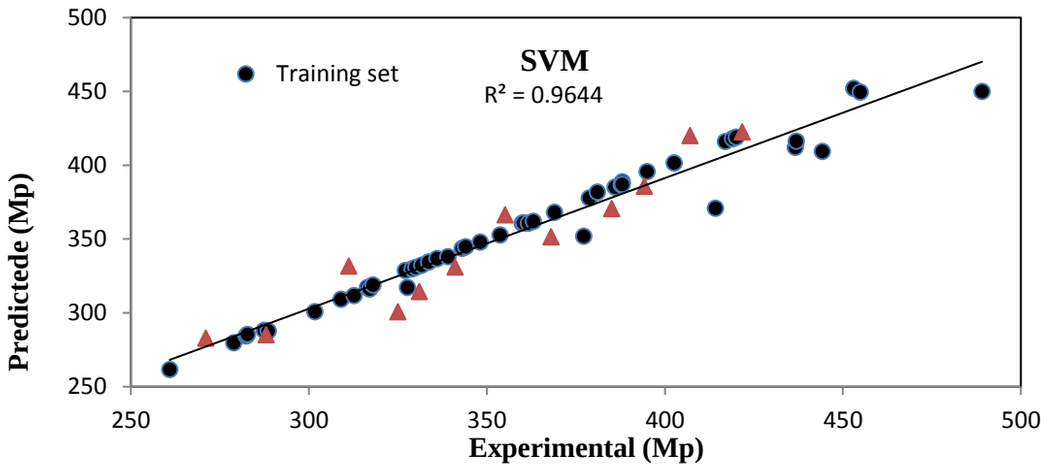
شکل ٤. نمودار تغییرات مقدار(C) Capacity parameter برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار Capacity parameter از ۱ تا ۲۲۱ مقدار بهینه ۱۹۱ برای Capacity parameter انتخاب شد.

در مرحله آخر با استفاده از تمامی پارامترهای بهینه شده، مدل SVM ساخته شده و مقادیر فعالیتهای بازدارندگی ترکیبات دارویی پیش بینی شد.

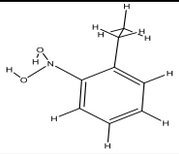
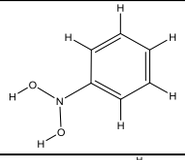
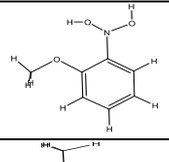
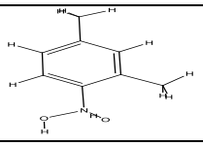
با استفاده از مدل SVM بهینه شده مقادیر فعالیتهای بازدارندگی ترکیبات مورد نظر در مجموعه آموزشی و پیش‌بینی مورد محاسبه قرار گرفته و در جدول۱ نشان داده شده است

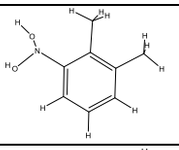
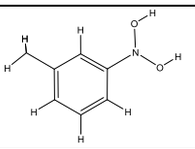
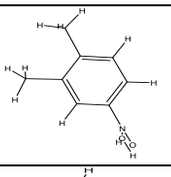
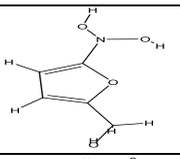
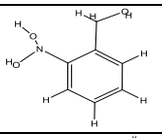
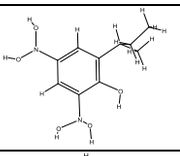
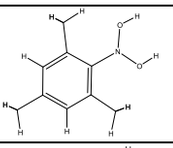
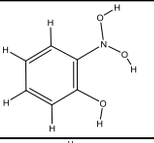
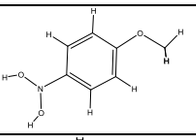
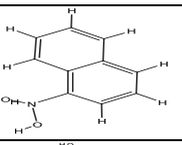
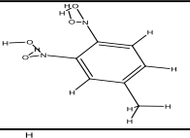
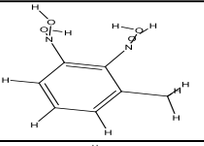
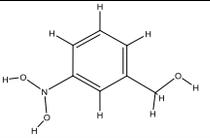
مقادیر فعالیتهای محاسبه شده و تجربی برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست در شکل ۵ آورده شده است در این شکل میزان نزدیکی داده ها به خط راست قدرت پیشگویی مدل را نشان می دهد.

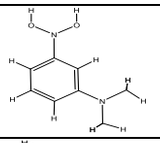
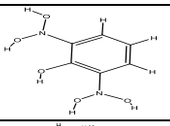
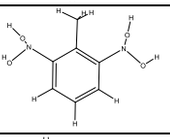
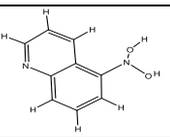
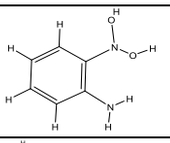
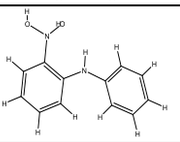
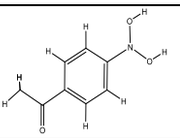
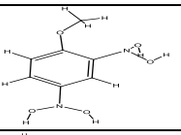
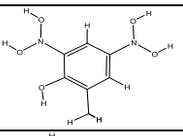
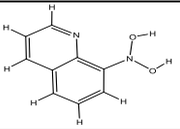
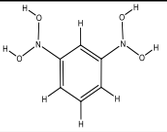
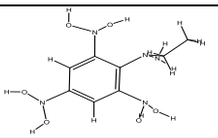
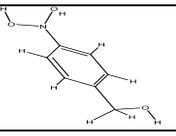


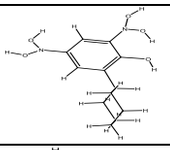
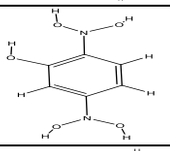
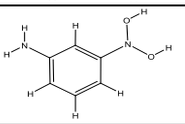
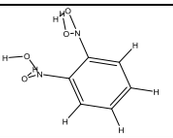
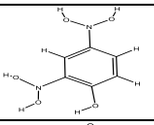
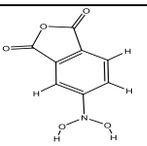
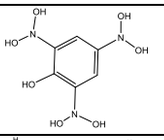
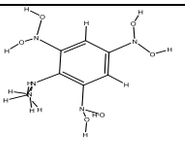
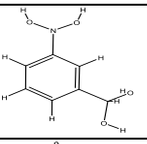
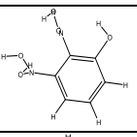
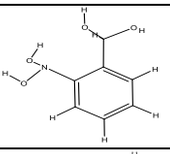
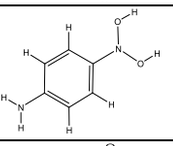
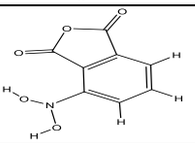
شکل ۵. مقادیر Mp محاسبه شده برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست برحسب مقادیر تجربی

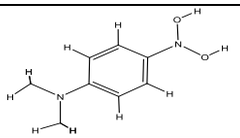
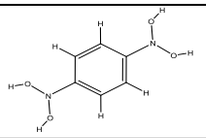
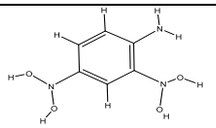
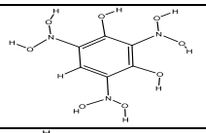
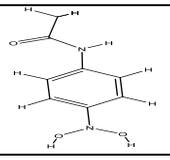
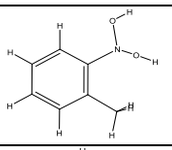
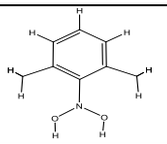
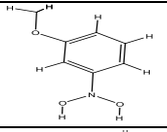
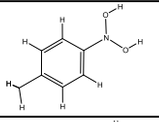
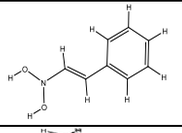
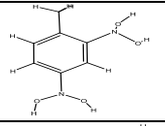
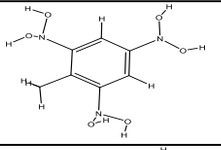
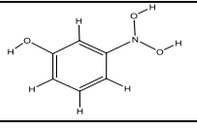
جدول۱. مقادیر تجربی و محاسبه شده سمیت(pEC_{50}) برای ترکیبات مختلف برای مجموعه های آموزش، ارزیابی و پیش بینی در مدل SW-ANN همراه با مقادیر خطای نسبی

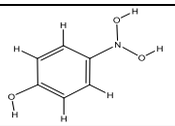
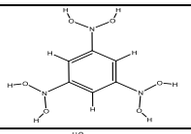
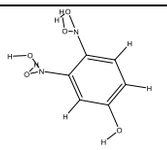
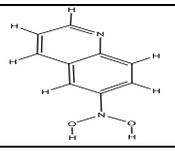
No.	Compound	Exp [a] (Mp)	MLR[b] (Mp)	SVM [c](Mp)	Residual[d] (for SVM)
Training set					
1		260.90	267.48	267.48	6.58
2		278.90	315.05	315.05	36.15
3		282.35	297.85	297.85	15.50
4		282.68	287.65	287.65	4.97

5		287.40	296.65	296.65	9.25
6		288.59	296.21	296.21	7.62
7		301.70	297.33	297.33	-4.37
8		309.00	312.65	312.65	3.65
9		312.65	294.21	294.21	-18.44
10		316.42	341.82	341.82	25.40
11		317.00	280.37	280.37	-36.63
12		318.00	321.59	321.59	3.59
13		327.10	367.18	367.18	40.08
14		327.70	292.62	292.62	-35.08
15		329.00	346.48	346.48	17.48
16		330.00	348.26	348.26	18.26
17		331.65	338.92	338.92	7.27

18		333.65	354.15	354.15	20.50
19		336.00	361.68	361.68	25.68
20		339.00	334.68	334.68	-4.32
21		343.00	374.96	374.96	31.96
22		344.00	364.20	364.20	20.20
23		348.10	357.23	357.23	9.13
24		353.65	364.87	364.87	11.22
25		359.90	375.73	375.73	15.83
26		360.25	383.75	383.75	23.50
27		361.65	365.93	365.93	4.28
28		363.00	355.90	355.90	-7.10
29		369.00	364.38	364.38	-4.62
30		377.15	320.95	320.95	-56.20

31		378.70	361.79	361.79	-16.91
32		381.00	405.81	405.81	24.81
33		386.00	382.29	382.29	-3.71
34		387.70	374.52	374.52	-13.18
35		388.00	393.48	393.48	5.48
36		388.00	396.22	396.22	8.22
37		395.00	419.33	419.33	24.33
38		402.60	405.15	405.15	2.55
39		414.15	346.53	346.53	-67.62
40		417.00	394.26	394.26	-22.74
41		419.00	429.32	429.32	10.32
42		420.00	407.67	407.67	-12.33
43		436.60	411.49	411.49	-25.11

44		436.90	392.41	392.41	-44.49
45		444.20	442.83	442.83	-1.37
46		453.05	429.97	429.97	-23.08
47		454.90	454.91	454.91	0.01
48		489.10	452.59	452.59	-36.51
Test set					
1		271.00	282.04	282.04	11.04
2		288.00	291.45	291.45	3.45
3		311.15	355.17	355.17	44.02
4		325.00	313.03	313.03	-11.97
5		331.00	309.70	309.70	-21.30
6		341.00	331.72	331.72	-9.28
7		355.10	375.88	375.88	20.78
8		368.00	353.00	353.00	-15.00

9		385.00	382.05	382.05	-2.95
10		394.20	413.75	413.75	19.55
11		407.00	415.81	415.81	8.81
12		421.65	427.97	427.97	6.32

a) مقادیر تجربی سمیت(pEC_{50})

b) مقادیر پیش بینی شده توسط رگرسیون خطی چندگانه

c) مقادیر پیش بینی شده توسط ماشین بردار پشتیبان

d) باقیمانده(اختلاف بین مقدار پیش بینی شده و مقدار واقعی برای مدل SVM)

۳-۳. ارزیابی اعتبار مدل‌های انتخاب شده

جهت اطمینان از اعتبار مدل بدســـت آمده، باید مدل را ارزیابی کرد. برای ارزیابی مدل از آماره‌های مختلفی نظیر F، R^2،Q^2، RMSE و آنالیز باقی‌مانده‌ها ا ستفاده می شود. نتایج پارامترهای آماری بد ست آمده در جدول۶ آورده شده ا ست. مقدار R^2 در واقع ن شانگر میزان تطابق نتایج تجربی با مقادیر محاسباتی است یا به عبارتی دیگر معیاری از قدرت تشریح مدل رگرسیون می‌باشد.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (۵)$$

در معادله بالا y ، $\hat{y}$ و $\bar{y}$ به ترتیب نشـــان دهنده مقادیر تجربی، مقادیر محاســـباتی و میانگین مقادیر تجربی هســـتند. برای مدل های MLR و SVM مقادیر R^2 به ترتیب ۰/۸۰۴ و ۰/۹۶۴ بدســـت آمد که نشـــان دهنده قدرت پیش بینی خوب مدل ها خصـــوصـــا SVM می باشـــد. مقدار F تخمینی از اهمیت آماری معادله رگرسیون می‌باشد مزیت مقدار F برR^2 این است که این پارامتر وابسته به درجه آزادی است و به طور واضح‌تر تحت تأثیر تعداد نمونه‌ها و تعداد توصیف کننده‌هایی است که در مدل دخیل هستند بنابراین ممکن است مدلی ساخته باشیم کهR^2 خیلی بالا و مقدار F آن کم باشد این بدان معنی است که در مدل تعداد نمونه‌ها کم و در عوض تعداد توصیف کننده‌ها بالا بوده است

$$F = \frac{\sum_i (\hat{y} - \bar{y})^2 / p}{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)} \qquad (۶)$$

در فرمول بالا n تعداد مولکولها و p نشـانگر تعداد توصـیف کننده‌های موجود در مدل به اضـافه یک (عرض از مبدأ به عنوان یک درجه آزادی درنظر گرفته می‌شود) می‌باشد. در مدل بدست آمده برای MLR و SVM مقادیر F به ترتیب ۲۳/۵۴ و ۹۷/۲۰ بدست آمد.

یکی دیگر از راه‌های ارزیای مدل، پارامتر RMSE می‌باشد که از رابطه زیر محاسبه خواهد شد:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{n} \quad (۷)$$

پارامتر RMSE مقدار پراکندگی مقادیر تجربی حول خط رگرسیون را مشخص می‌کند. پایین بودن مقدار RMSE نشان دهنده اعتبار بالای مدل می‌باشد. مقادیر RMSE برای سری های آموزش و پیش بینی در هر دو مدل در جدول ۶ آورده شده است.

جدول٦. پارامترهای آماری بدست آمده برای مدل های MLR و SVM

	Training set			Test set		
	R^2	RMSE	F	R^2	RMSE	F
MLR	0.804	23.299	23.549	0.865	18.073	4.022
SVM	0.964	11.566	97.207	0.910	14.040	6.093

٤. نتیجه گیری

در این مطالعه، رگرسیون QSPR برای پیش بینی نقاط ذوب ترکیبات پرانرژی نیتروآروماتیک های حلقوی بررسی شده است. در ابتدا توصیف کننده های الکترونی و مکانیک کوانتومی و شیمیایی برای هر ترکیب بدست آمد. سپس رابطه ای بین مناسب‌ترین توصیف کننده ها و نقاط ذوب ترکیبات با استفاده از روش های MLR و SVM بدست آمد. نتایج نشان داد که هر دومدل MLR و SVM نتایج قابل قبولی ارائه داده است اگر چه مقایسه این دو روش نشان از برتری روش SVM نسبت به روش MLR دارد. بنابراین می‌توان نتیجه گرفت که از توصیف کننده های الکترونی و مکانیک کوانتومی نیز می‌توان جهت پیش بینی نقاط ذوب ترکیبات فوق استفاده کرد. از طرف دیگر از این روش‌ها می‌توان جهت پیش بینی سایر خواص فیزیکوشیمیایی ترکیبات نیز استفاده نمود.

٥. مراجع

[1] Hamadanian, M., Keshavarz, M.H., Nazari, B. and Mohebbi, M., Reliable method for safety assessment of melting points of energetic compounds. *Process Safety and Environmental Protection*, *103* (**2016**)10-22.

[2] Pagoria, P.F., Lee, G.S., Mitchell, A.R. and Schmidt, R.D., A review of energetic materials synthesis. *Thermochimica Acta*, *384*(1) (**2002**) 187-204.

[3] Rice, B.M., Pai, S.V. and Hare, J., Predicting heats of formation of energetic materials using quantum mechanical calculations. *Combustion and flame*, *118*(3) (**1999**) 445-458.

[4] Shahlaie, M., Fassihi, A., Pourhossein, A. and Arkan, E., Statistically validated QSAR study of some antagonists of the human CCR5 receptor using least square support vector machine based on the genetic algorithm and factor analysis. *Medicinal Chemistry Research*, 22(3) (**2013**) 1399-1414.

[5] Ma, S., Lv, M., Deng, F., Zhang, X., Zhai, H. and Lv, W., Predicting the ecotoxicity of ionic liquids towards Vibrio fischeri using genetic function approximation and least squares support vector machine. *Journal of hazardous materials*, 283 (**2015**) 591-598.

[6] Prasoona, R.K., Jyoti, A., Mukesh, Y., Nishant, S., Anuraj, N.S. and Shobha, J., Optimization of Gaussian Kernel Function in Support Vector Machine aided QSAR studies of C-aryl glucoside SGLT2 inhibitors. *Interdisciplinary Sciences: Computational Life Sciences*, 5(1) (**2013**) 45-52.

[7] Nekoei, M., Mohammadhosseini, M. and Pourbasheer, E., QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach. *Medicinal Chemistry Research*, 24(7) (**2015**) 3037-3046.
[8] Pourbasheer, E., Aalizadeh, R., Ganjali, M.R. and Norouzi, P., Prediction of Superoxide Quenching Activity of Fullerene (C60) Derivatives by Genetic Algorithm-Support Vector Machine. *Fullerenes, Nanotubes and Carbon Nanostructures*, 23(4) (**2015**) 290-299.
[9] Zhao, H., Zhang, X., Ji, L., Hu, H. and Li, Q., Quantitative structure–activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design. *Corrosion Science*, 83 (**2014**) 261-271.
[10] Pourbasheer, E., Beheshti, A., Vahdani, S., Nekoei, M., Danandeh, M., Abbasghorbani, M. and Ganjali, M.R., Simple QSPR Modeling for Prediction of the GC Retention Indices of Essential Oil Compounds. *Journal of Essential Oil Bearing Plants*, 18(6) (**2015**) 1298-1309.
[11] Jalali, A., Nekoei, M. and Mohammadhosseini, M., Novel QSPR Study on the Melting Points of a Broad Set of Drug-Like Compounds Using the Genetic Algorithm Feature Selection Approach Combined With Multiple Linear Regression and Support Vector Machine. *Journal of Chemical Health Risks*, 6(1) (**2016**).
[12] Nekoei, M., A Robust Quantitative Structure-Property Relationship-Based Model for Estimation of Refractivity Indices of 101 Common Paraffin Derivatives Based on Their Molecular Structures. *Journal of Chemical Health Risks*, (**2016**).
[13] Mohammadhosseini, M., Deeb, O., Alavi-Gharabagh, A. and Nekoei, M., Exploring novel QSRRs for simulation of gas chromatographic retention indices of diverse sets of terpenoids in Pistacia lentiscus L. essential oil using stepwise and genetic algorithm multiple linear regressions. *Analytical Chemistry Letters*, 2(2) (**2012**) 80-102.
[14] Adimi, M., Salimi, M., Nekoei, M., Pourbasheer, E. and Beheshti, A., A quantitative structure-activity relationship study on histamine receptor antagonists using the genetic algorithm-multi-parameter linear regression method. *J Serb Chem Soc*, 77(5) (**2012**) 639-650.
[15] Rahimi, M. and Nekoei, M., Quantitative Structure-Property Relationship Study for Prediction of Flash Point of Some Organic Compounds Based On SW-MLR Method. *Analytical Chemistry Letters*, 3(4) (**2013**) 278-286.
[16] Nekoei, M., Goudarzi, N., Nekoei, S. and Mohammadhosseini, M., QSAR Study of Arylsulfonylpiperazine Inhibitors of 11β-HSD1 by GA-MLR, GA-PLS and GA-ANN. *Analytical Chemistry Letters*, 4(1) (**2014**) 14-28.
[17] Beheshti, A., Pourbasheer, E., Nekoei, M. and Banaei, A., Quantitative Structure-Activity Relationship Study of Amino Acid Derivatives as Histone Deacetylase Inhibitors using the Genetic Algorithm–Multiple Linear Regression. *Analytical Chemistry Letters*, 2(1) (2012) 33-43.
[18] Wang, D., Yuan, Y., Duan, S., Liu, R., Gu, S., Zhao, S., Liu, L. and Xu, J., QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network. *Chemometrics and Intelligent Laboratory Systems*, 143 (**2015**) 7-15.
[19] Ma, S., Lv, M., Deng, F., Zhang, X., Zhai, H. and Lv, W., Predicting the ecotoxicity of ionic liquids towards Vibrio fischeri using genetic function approximation and least squares support vector machine. *Journal of hazardous materials*, *283* (**2015**) 591-598.
[20] Golmohammadi, H., Dashtbozorgi, Z. and Vander Heyden, Y., Support vector regression based QSPR for the prediction of retention time of peptides in reversed-phase liquid chromatography. *Chromatographia*, *78*(1-2) (**2015**) 7-19.
[21] Ghanbari, A., Application of support vector machine in QSAR study of triazolyl thiophenes as cyclin dependent kinase-5 inhibitors for their anti-alzheimer activity. *Indian Journal of Chemical Technology (IJCT)*, *23*(1) (**2016**) 9-21.
[22] Golmohammadi, H. and Dashtbozorgi, Z., QSPR studies for predicting polarity parameter of organic compounds in methanol using support vector machine and enhanced replacement method. *SAR and QSAR in Environmental Research*, *27*(12) (**2016**) 977-997.