مجله شیمی کوانتومی و اسپکتروسکوپی
۹:۲۹ (۱۳۹۸) ۳۷-۵۲
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مطالعه ارتباط کمی ساختار- خاصیت جهت پیش‌بینی قطبیت حلال با استفاده از توصیف‌کننده‌های مکانیک‌کوانتومی و ماشین بردار پشتیبان

مهدی نکوئی*، بهزاد چهکندی

دانشگاه آزاد اسلامی، واحد شاهرود، دانشکده علوم پایه، گروه شیمی، شاهرود، ایران

تاریخ ثبت اولیه:۱۳۹۷/۱۱/۰۶، تاریخ دریافت نسخه اصلاح شده:۱۳۹۷/۱۲/۲۳، تاریخ پذیرش قطعی:۱۳۹۸/۰۱/۲۵

چکیده

مطالعه ارتباط کمی ساختار-خاصیت(QSPR) جهت پیش بینی قطبیت برخی از حلال ها با استفاده از توصیف کننده های مکانیک کوانتومی و تکنیک ماشین بردار پشتیبان(SVM) انجام شد. مقادیر تجربی قطبیت برای ۶۹ حلال گردآوری شد. این حلال ها شامل هیدروکربن های اشباع و غیراشباع، حلال های شامل هالوژن، سیانید، نیترو، آمید، سولفید، مرکاپتو، فسفات، استر، اتر و غیره بود. در ابتدا ساختار حلال ها، رسم و گروه مناسبی از توصیف کننده‌ها محاسبه شدند. سپس از روش انتخاب مرحله‌ای برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با قطبیت حلال مورد نظر داشتند استفاده گردید. در ابتدا مدل خطی رگرسیون خطی چندگانه(MLR) ساخته شد. سپس برای به دست آوردن نتایج بهتر از SVM استفاده گردید. داده های آماری، برتری روش SVM را نسبت به روش MLR نشان می دهد.

واژه های کلیدی : ارتباط کمی ساختار- خاصیت، قطبیت حلال، رگرسیون خطی چند گانه، ماشین بردار پشتیبان ، توصیف کنند های مکانیک کوانتمی

۱. مقدمه

حلال یکی از اجزاء مهم و درگیر در واکنش است که علی‌رغم حضور دائمی‌اش در واکنشها، به ندرت در معادلات شیمیایی وارد می‌شود و البته بدون آن، انجام بیشتر واکنش‌ها غیرممکن است. حضور یک حلال می‌تواند سرعت یک واکنش را با ضریب

***عهده دار مکاتبات:** مهدی نکوئی

نشانی: گروه شیمی، دانشکده علوم پایه، دانشگاه آزاد اسلامی، شاهرود، ایران

تلفن:۰۲۳۳۲۳۹۴۲۸۹ **پست الکترونیک:**E-mail: m_nekoei1356@yahoo.com

10^{20} تسریع کند و یا آن را آهسته نماید[۱]. تغییر یک حلال به حلال دیگر می‌تواند میلیونها بار سرعت واکنش را تغییر دهد. اثرات حلال به مراتب قوی‌تر از اثرات سایر عوامل می‌باشد. انتخاب یک حلال خاص می‌تواند عامل تعیین کننده اصلی سرعت واکنش و حتی چگونگی انجام و یا عدم انجام واکنش باشد. حلال می تواند از بین چندین مسیر برای واکنش مسیری را که در واقع واکنش دنبال می کند، تعیین نماید[۳-۲]. حلال در هر واکنشی که در آن رخ می دهد ذاتاً دخالت دارد و یافتن میزان این دخالت و نحوه آن برای ما اهمیت دارد. مولکول‌های حل شونده و یونها به صورت ذراتی لخت در محلول وجود ندارد بلکه حلال پوشیده می‌باشند. به هر ذره حل شده، یک توده از مولکول‌های حلال توسط پیوندهایی متصل است. حلال ها به طور کلی به دو دسته حلال های قطبی و حلال‌های غیرقطبی تقسیم می‌شوند. در حلال قطبی، ذرات تشکیل دهنده حلال قطبی بوده و یکدیگر را با نیروی جاذبه‌ی الکتروستاتیکی جذب می‌نمایند. در حلال‌های غیرقطبی، ذرات حلال، غیرقطبی بوده و بنابراین تنها نیروی جاذبه‌ی ضعیف واندروالسی بین ذرات وجود دارد، به همین دلیل این حلال‌ها اغلب، دارای نقطه‌ی جوش بسیار پایین بوده و فرار هستند[۴]. میزان قطبیت حلال یک پارامتر مهم در انجام واکنش‌های شیمیایی است. بنابراین پیش بینی میزان قطبیت حلال دارای اهمیت زیادی است[۵]. روش ارتباط کمی ساختار- خاصیت[1] (QSPR) یک روش مناسب برای تخمین میزان قطبیت حلال براساس توصیف کننده‌های بدست آمده از ساختار مولکولی آنها می‌باشد. 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 رابطه‌ای است که بین رفتار فیزیکوشیمیایی یک مولکول با پارامترهای ساختاری آن وجود دارد[۱۲-۶]. نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختاری آنها به پژوهشگران در پیش بینی رفتار مولکول‌های جدید براساس رفتار مولکول‌های مشابه کمک می کند.

هدف از مطالعه حاضر، پیش بینی قطبیت حلال ترکیبات شامل هیدروکربن های اشباع و غیراشباع، حلال های شامل هالوژن، سیانید، نیترو، آمید، سولفید، مرکاپتو، فسفات، استر، اتر و غیره با استفاده از توصیف کننده‌های شیمیایی و مکانیک کوانتمی و ماشین بردار پشتیبان است.

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

سری داده‌ها مربوط به قطبیت ۶۹ حلال است که توسط زارعی و همکارانش گزارش شده است[۱۳]. نام و فرمول این ترکیبات در جدول ۴ آورده شده است. در ابتدا ترکیبات به صورت تصادفی به دو سری شامل سری آموزش و سری پیش‌بینی تقسیم شدند(جدول۴). سری آموزش شامل ۴۸ ترکیب و سری پیش‌بینی یا تست نیز شامل ۲۱ ترکیب است. مقادیر قطبیت حلال به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند.

[1]-Quantitative Structure-Property Relationship

۲-۲. محاسبه توصیف‌کننده‌ها

توصیف کننده‌ها 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[۱۶-۱۴]. بعد از ارزیابی و پیدا کردن مناسب‌ترین آنها می‌توانند جهت پیش‌بینی خاصیت ترکیبات جدید بکار روند. محاسبه توصیف‌کننده‌های الکترونی بوسیله بسته نرم افزاری HyperChem انجام شد. مشخصات این توصیف کننده ها در جدول ۱ آورده شده است. برای محاسبه بقیه توصیف کننده‌ها از نرم‌افزار دراگون استفاده گردید.

جدول۱. مشخصات برخی از توصیف کننده‌های محاسبه شده

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

۲-۳. ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از روش‌های یادگیری تحت نظارت است که هم برای دسته‌بندی و هم رگرسیون قابل استفاده است. این روش توسط وپنیک[1] بر پایه تئوری یادگیری آماری بنا نهاده شده است. ماشین بردار پشتیبان روشی برای طبقه‌بندی دوتائی در فضای ویژگی‌های دلخواه است و از این رو روشی مناسب برای مسائل پیش بینی به شمار می رود[۱۷]. ماشین بردار پشتیبان در اصل یک دسته بندی کننده دو کلاسه است که کلاس‌ها را توسط یک مرز خطی از هم جدا می‌کند. در این روش نزدیکترین نمونه‌ها به مرز تصمیم گیری را بردارهای پشتیبان می‌نامند. این بردارها معادله مرز تصمیم گیری را مشخص می‌کنند[۱۸].

در سال ۱۹۹۶، واپنیک و همکارانش، نسخه‌ای از SVM را پیشنهاد دادند که به جای طبقه‌بندی، عمل رگرسیون را انجام می‌دهد. این مورد به Support Vector Regression یا SVR معروف است. همانند SVM در این مدل نیز از تابع کرنل و ابرپارامتر C استفاده می‌شود[۱۹].

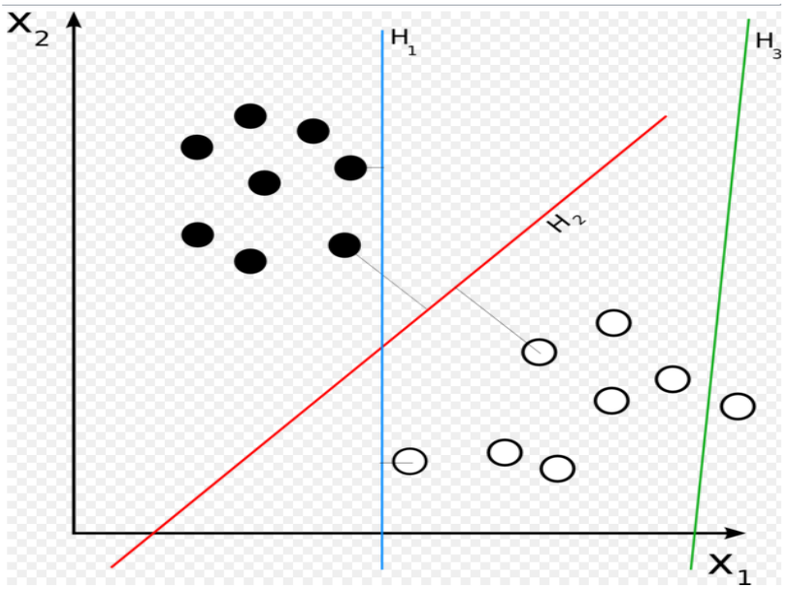
[1] Vapnic

بردارهای پشتیبان به زبان ساده، مجموعه‌ای از نقاط در فضای n بعدی داده‌ها هستند که مرز دسته‌ها را مشخص می‌کنند و مرزبندی و دسته بندی داده‌ها براساس آنها انجام می‌شود و با جابجایی یکی از آنها، خروجی دسته بندی ممکن است تغییر کند. در فضای دوبعدی، بردارهای پشتیبان، یک خط، در فضای سه بعدی یک صفحه و در فضای n بعدی یک ابر صفحه را شکل خواهند داد. ماشین بردار پشتیبان، یک دسته بند یا مرزی است (شکل۱) که با معیار قرار دادن بردارهای پشتیبان، بهترین دسته بندی و تفکیک بین داده ها را برای ما مشخص می کند[۲۱-۲۰].

برای مجموعه داده‌های آزمایشی شامل n عضو(نقطه) رابطه زیر برقرار می‌باشد:

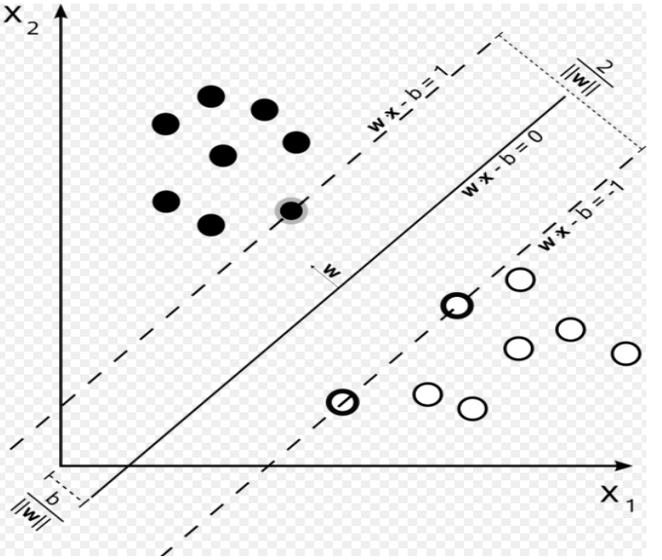
$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n} \qquad (۱)$$

به ازای داده‌های موجود، تعداد زیادی مرزبندی می‌توانیم داشته باشیم که سه تا از این مرزبندی‌ها در زیر نمایش داده شده است.



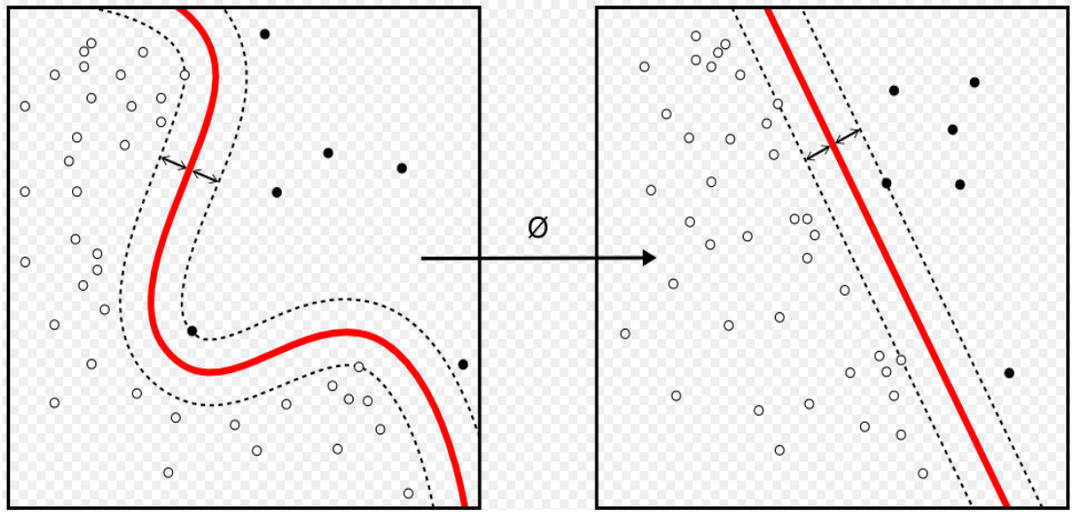
شکل۱. انواع مرز در تفکیک و دسته بندی داده ها

یک راه ساده برای بدست آوردن بهترین مرزبندی و ساخت یک دسته بند بهینه، محاسبه فاصله‌ی مرزهای به دست آمده با بردارهای پشتیبان هر دسته (مرزی ترین نقاط هر دسته یا کلاس) و در نهایت انتخاب مرزیست که از دسته‌های موجود، مجموعاً بیشترین فاصله را داشته باشد(شکل۲). این عمل تعیین مرز و انتخاب خط بهینه (در حالت کلی، ابر صفحه مرزی) به راحتی با انجام محاسبات ریاضی نه چندان پیچیده قابل پیاده سازی است. به عبارت دیگر، اگر داده‌های آموزشی جدایی پذیر خطی باشند، ما می‌توانیم دو ابر صفحه در حاشیه نقاط به‌طوری‌که هیچ نقطه مشترکی نداشته باشند، در نظر بگیریم و سپس سعی کنیم، فاصله آن‌ها را ماکسیمم کنیم[۲۲].



شکل۲. نمایش دو ابرصفحه در حاشیه نقاط دو مجموعه از داده ها و بدست آوردن بهترین بردار پشتیبان

اگر داده‌ها به صورت خطی قابل تفکیک باشند، الگوریتم فوق می‌تواند بهترین بردار یا ابرصفحه را برای تفکیک داده‌ها ایجاد کند اما اگر داده‌ها به صورت خطی توزیع نشده باشند نیاز داریم داده‌ها را به کمک یک تابع ریاضی (Kernel functions) به یک فضای دیگر ببریم (نگاشت کنیم) که در آن فضا، داده‌ها تفکیک پذیر باشند و بتوان SVM آنها را به راحتی تعیین کرد(شکل۳). ماشین های بردار پشتیبان برای حل مسائل غیرخطی، ابعاد مسئله را از طریق توابع کرنل تغییر می‌دهند. انتخاب کرنل برای SVM به حجم داده‌های 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 ورودی‌های مساله را داشته باشد. در عمل چهار نوع کرنل خطی، کرنل چند جمله ای، کرنل تانژانت هیپربولینک و کرنل RBF به کار گرفته می‌شوند. در جدول ۲ معادلات برخی از کرنل‌های رایج ارائه شده اند[۲۲].



شکل۳. استفاده از تابع کرنل برای تغییر فضای داده ها

جدول ۲. 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	$K(x_i, x_j) = exp(-\gamma \lvert x_i - x_j \rvert 2)$

۳. بررسی نتایج

۳-۱. مدل‌سازی به روش رگرسیون خطی چندگانه[1](MLR)

در رگرسیون خطی چندگانه، پارامترهای یک مدل خطی به کمک یک تابع هدف و مقدارهای متغیرها، برآورد می‌شوند. در رگرسیون خطی، مدل در نظر گرفته شده، یک رابطه خطی برحسب پارامترهای مدل است. به این ترتیب اگر n مشاهده از متغیر مستقل p بعدی X داشته باشیم و بخواهیم یک رابطه خطی با متغیر پاسخ y برقرار کنیم، می‌توانیم از مدل رگرسیون خطی زیر استفاده کنیم.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \qquad (۲)$$

از آنجایی که متغیر مستقل X دارای p بعد است، مقدار آن را در هر بعد با یک متغیر مستقل یک بعدی جایگزین کرده‌ایم. مشخص است که اندیس i نیز شماره مشاهده را نشان می‌دهد. در انتها نیز ε جمله خطای مدل رگرسیونی محسوب می‌شود.

جهت مدل سازی به روشMLR در ابتدا توصیف‌کننده‌های محاسبه شده، به عنوان متغیرهای مستقل و مقادیر قطبیت حلال 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، 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 قطبیت ترکیبات با ساختار آن‌ها انتخاب شد. با این روش ۱۰ مدل بررسی شد که مدل سوم با تعداد ۳ 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 به همراه توصیف مختصری از آنها در جدول۳ آورده شده است. در جدول ۴ نیز مقادیر عددی توصیف کننده‌ها آورده شده است.

[1] Multiple linear regression (MLR)

جدول ۳. توصیف کننده‌های انتخاب شده با SPSS وتوصیف آنها

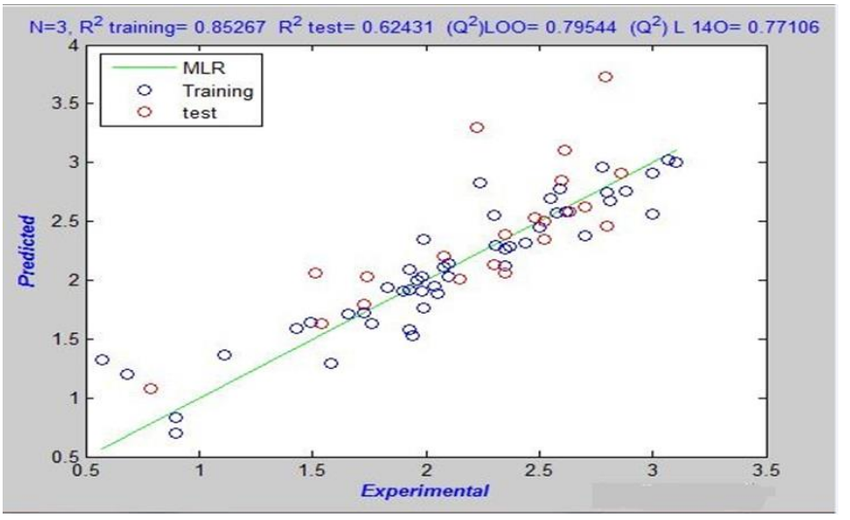
نام توصیف کننده	نوع توصیف کننده	ضرایب	اثر متوسط توصیف کننده	T-value
Dipole moment	Quantum mechanics descriptors	0.282	-0.505	13.247
Surface area	Geometrical descriptors	-0.004	0.923	-7.450
$E_{HOMO-LUMO}(E_{GAP})$	Quantum mechanics descriptors	-0.059	0.581	-2.624
Constant	-	3.308	-	11.477

پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها توسط روش مرحله‌ای با استفاده از SPSS، مرحله بعد، ایجاد مدل بین توصیف‌کننده‌های انتخاب شده و قطبیت حلال ترکیبات است. از نرم افزار SPSS برای این منظور استفاده گردید و بین توصیف‌کننده‌ها و قطبیت حلال ترکیبات سری آموزش با استفاده از روش MLR رابطه زیر به دست آمد:

$$S' = 3.308 + 0.282\ (\text{Dipole moment}) - 0.004\ (\text{surface area}) - 0.059\ (E_{HOMO-LUMO}) \qquad (۳)$$

$R^2_{train}=0.852$, $F_{train}=84.88$, $R^2_{test}=0.624$, $F_{test}=13.404$, $R^2_{adj}=0.842$, $Q^2_{LOO}=0.795$, $Q^2_{LGO}=0.780$

سپس از معادله به دست آمده برای پیش‌بینی قطبیت حلال (S') سری پیش‌بینی (تست) استفاده گردید. شکل ۴ نمودار مقادیر S' محاسبه شده با کمک مدلSW-MLR برای مجموعه‌های آموزش و پیش‌بینی را برحسب مقادیر تجربی نشان می‌دهد.



شکل ٤. نمودار مقادیر S' محاسبه شده با کمک مدلSW-MLR برای مجموعه‌های آموزش و پیش‌بینی برحسب مقادیر تجربی

همانطور که پارامترهای آماری بالای شکل نشان می‌دهد مقدار R^2_{test} نسبتاً پایین بوده و روش رگرسیون خطی چندگانه نتوانسته است پیش‌بینی‌های مناسبی را نشان دهد. بنابراین جهت حصول نتایج بهتر، از ماشین بردار پشتیبان(SVM) برای پیش‌بینی S' ترکیبات استفاده شد.

جدول٤. مقادیر تجربی و محاسبه‌شده قطبیت حلال(S') برای مجموعه‌های آموزش و پیش‌بینی در مدل SW-SVM همراه با مقادیر عددی توصیف‌کننده‌ها

NO.	Solvent	Formula	$S'_{exp.}$	SVM	Dipole moment	Surface area	$E_{HOMO.LUMO}$
1[a]	1,1,1-Trichloroethane	$Cl_3C\text{-}CH_3$	1.93	2.09	1.75	248.90	11.72
2[a]	Cis-1,1,2- Trichloroethane	$ClH_2C\text{-}CHCl_2$	2.35	2.31	2.45	251.09	11.83
3[b]	Trans-1,1,2-Trichloroethane	$ClH_2C\text{-}CHCl_2$	2.35	2.07	1.71	251.32	11.75
4[a]	1,1,2- Trichloroethane	$Cl_2C{=}CHCl$	1.90	1.93	1.16	251.09	11.73
5[a]	1,2-Dichlorobenzene	$1,2\text{—}Cl_2C_6H_4$	2.10	2.05	1.97	283.20	9.46
6[a]	boat 1,4-Dioxane	$O(CH_2CH_2)_2O$	1.93	2.00	1.91	238.00	13.03
7[a]	chair 1,4-Dioxane	$O(CH_2CH_2)_2O$	1.93	1.77	0.00	235.33	12.94
8[a]	Butan-2-one	$CH_3C(O)C_2H_5$	2.50	2.49	2.81	237.27	11.42
9[a]	3-Methylsulfolane	$C_5H_{10}SO_2$	2.55	2.63	4.36	287.51	11.23
10[a]	Pentan-3-one	$C_2H_5COC_2H_5$	2.37	2.34	2.69	269.98	11.33
11[b]	4-Butyrolactone	$C_4H_6O_2$	2.86	2.56	4.56	232.70	12.34
12[a]	4-Methylpyridine	$4\text{-}CH_3C_2H_4N$	2.31	2.27	2.32	261.01	10.08
13[a]	Acetone	$(CH_3)_2CO$	2.58	2.58	2.85	208.57	11.51
14[a]	Acetonitrile	CH_3CN	3.00	2.84	2.89	177.06	14.12
15[b]	Acetophenone	$C_6H_5C(O)CH_3$	2.52	2.23	2.85	292.08	9.55
16[a]	Anisole	$C_6H_6OCH_3$	2.04	1.89	1.25	281.80	9.48
17[a]	Benzene	C_6H_6	1.73	1.70	0.00	239.98	10.20
18[b]	Benzonitrile	C_6H_5CN	2.63	2.41	3.36	270.55	9.62
19[a]	Bromobenzene	C_6H_5Br	2.10	1.97	1.45	272.95	9.66
20[a]	Butyl acetate	$CH_3C(O)OBu$	1.99	1.87	1.84	323.56	12.40
21[b]	Carbon disulfide	CS_2	1.51	1.66	0.00	181.48	8.52
22[a]	Carbon tetrachloride	CCl_4	1.49	1.65	0.00	244.77	11.25
23[a]	Chlorobenzene	C_6H_5Cl	1.98	1.96	1.31	262.83	9.71
24[a]	Cyclohexane	C_6H_{12}	1.11	0.95	0.00	263.25	14.58
25[b]	Cyclohexanone	$(CH_2)_5CO$	2.35	2.44	1.31	266.34	11.22
26[a]	Di-isopropyl ether	$Pr^i{}_2O$	1.76	1.69	0.00	303.91	13.23
27[a]	Di-n- butyl ether	$(n\text{-}C_4H_9)_2O$	1.58	1.48	2.97	382.06	13.27
28[b]	Dichloromethane	CCl_2H_2	2.08	2.22	1.26	199.60	11.98
29[b]	Diethyl ether	$(C_2H_5)_2O$	1.73	1.85	1.18	262.05	13.37
30[a]	Dimethylaniline	$C_6H_5N(CH_3)_2$	1.96	1.86	1.50	310.06	8.81
31[a]	Dimethyl sulfoxide	$(CH_3)_2SO$	3.00	2.84	1.25	220.71	10.34
32[b]	Ethyl acetate	$CH_3C(O)OC_2H_5$	2.15	2.07	1.70	263.62	12.39
33[a]	Ethyl formamide	$HC(O)NEt_2$	2.59	2.69	3.95	238.30	11.54

34[a]	Ethyi formate	$HCOOC_2H_5$	2.24	2.40	1.83	229.53	12.30
35[b]	Hexamethylphosphoramide	$[(CH_3)_2N]_3PO$	2.52	2.39	3.99	365.61	11.09
36[a]	Hexyl acetate	$CH_3C(O)Ohex$	1.94	1.78	4.22	383.47	12.28
37[a]	Methyl acetate	$CH_3C(O)OCH_3$	2.35	2.20	4.78	229.12	12.50
38[b]	N,N-Dimethylacetamide	$CH_3CON(CH_{3)2}$	2.70	2.65	3.63	257.79	10.98
39[a]	N,N-Dimethylcyanamide	$(CH_3)_2NCN$	2.81	2.65	3.55	232.66	11.59
40[a]	N,N-Dimethylformamide	$HCON(CH_3)_2$	2.80	2.74	3.69	231.02	11.12
41[b]	N-Methylimidazole	$C_4H_6N_2$	2.60	2.69	3.95	240.61	10.02
42[a]	N-Methylpyrrolidinone	$CH_2CH_2CH_2CONCH_3$	2.62	2.61	3.59	265.86	10.99
43[a]	n-Butyronitrile	$CH_3CH_2CH_2CN$	2.70	2.54	3.00	239.53	13.45
44[a]	n-Decane	$C_{10}H_{22}$	0.90	1.01	0.00	424.23	14.63
45[b]	n-Heptane	C_7H_{16}	0.79	0.92	0.01	331.45	14.75
46[a]	n-Hexane	C_6H_{14}	0.68	0.84	0.00	300.80	14.82
47[a]	n-Nonane	C_9H_{20}	0.90	0.99	0.01	393.57	14.66
48[a]	n-Pentane	C_5H_{12}	0.57	0.73	0.01	270.61	14.88
49[b]	Nitrobenzene	$C_6H_5NO_2$	2.61	2.51	5.24	272.67	9.49
50[a]	Nitroethane	$C_2H_5NO_2$	2.78	2.73	4.40	217.33	11.76
51[a]	Nitromethane	CH_3NO_2	3.07	2.91	4.17	185.01	11.84
52[a]	Tetrahydropyran	$(CH_2)_5O$	1.98	1.97	1.51	251.47	13.52
53[b]	Propionitrile	C_2H_5CN	2.80	2.66	2.94	210.94	13.69
54[a]	Propyl acetate	$CH_3C(O)OPr$	2.05	1.95	1.83	293.60	12.34
55[a]	Propylene carbonate	$(CH_2)_3(O-)_2CO$	3.10	2.94	5.27	246.96	13.21
56[a]	Pyridine	C_5H_5N	2.44	2.28	1.97	231.39	10.07
57[b]	Quinoline	C_9H_7N	2.30	1.83	1.88	298.96	8.71
58[a]	Tetrahydrofuran	$(CH_2)_4O$	2.08	2.24	1.90	230.89	13.31
59[a]	Tetrahydrothiophene	$(CH_2)_4S$	1.99	2.15	2.13	244.23	9.44
60[a]	Tetramethylene sulfone	$(CH_2)_4SO_2$	2.88	2.72	4.25	262.90	11.32
61[b]	Tetramethylurea	$[(CH_3)_2N)]_2CO$	2.48	2.51	3.79	298.22	10.61
62[a]	Thiophene	$(CH)_4S$	1.83	1.79	0.35	222.55	9.45
63[a]	Toluene	$C_6H_5CH_3$	1.66	1.71	0.26	265.98	9.85
64[a]	Tributyl phosphate	$(n\text{-}C_4H_9O)_3PO$	2.30	2.14	7.99	563.50	11.80
65[b]	Trichloromethane	CCl_3H	1.74	2.03	1.16	225.08	11.47
66[a]	Triethylamine	$(C_2H_5)_3N$	1.43	1.59	0.87	308.04	11.92
67[a]	Triethyl phosphate	$(C_2H_5O)_3PO$	2.22	2.36	8.02	381.94	11.87
68[a]	Trimethyl benzene	$C_6H_3(CH_3)_3$	1.54	1.69	0.54	306.20	9.70
69[b]	Trimethyl phosphate	$(CH_3O)_3PO$	2.79	2.50	8.05	279.30	11.90

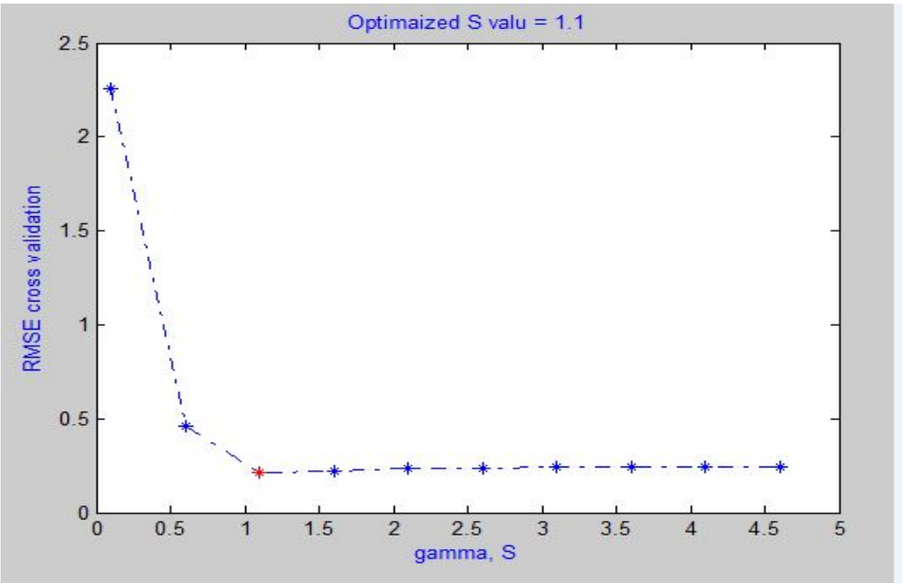
a: سری آموزش. b: سری پیش بینی

۳-۲. مدل‌سازی با استفاده از ماشین بردار پشتیبان

در مرحله دوم، جهت حصول نتایج بهتر از ماشین بردار پشتیبان برای ایجاد مدل و پیش‌بینی قطبیت حلال استفاده شد. در ابتدا پارامترهای مربوط به SVM بهینه گردید:

۱- نوع تابع کرنل[1] ۲- پارامتر ظرفیت[2] ۳- فاکتور حساسیت (سیگما[3]) و ۴- اپسیلون یا گاما[4]

نوع تابع کرنل، نحوه توزیع نمونه‌ها را در فضا مشخص می‌کند. RBF یکی از توابعی است که به طور معمول استفاده می‌شود و نتایج خوبی نیز می‌دهد 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 می‌گذارد نیز باید بهینه شود. تعداد بردارهای پشتیبان بر زمان آموزش مدل تاثیر می‌گذارد طوری که افزایش مقدار Gamma و در نتیجه تعداد بردار پشتیبان می‌تواند به افزایش زمان آموزش و همچنین Overfitting منجر شود. مقدار Gamma توانایی و قدرت SVM را در پیشگویی کنترل می‌کند. در شکل ۵ نمودار مقادیر متفاوت Gamma بر حسب RMSE نمایش داده شده است.

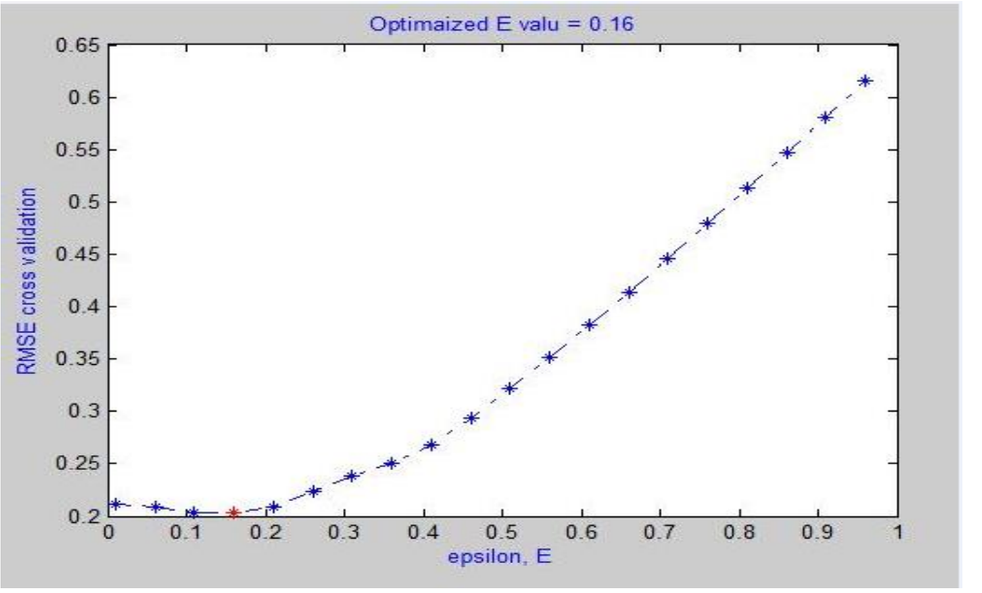


شکل ۵. نمودار تغییرات مقدار Gamma برحسب مقدار RMSE برای سری آموزش

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

[1] kernel function type
[2] Capacity parameter (C)
[3] Sensitive factor (ε)
[4] Gamma (eps)

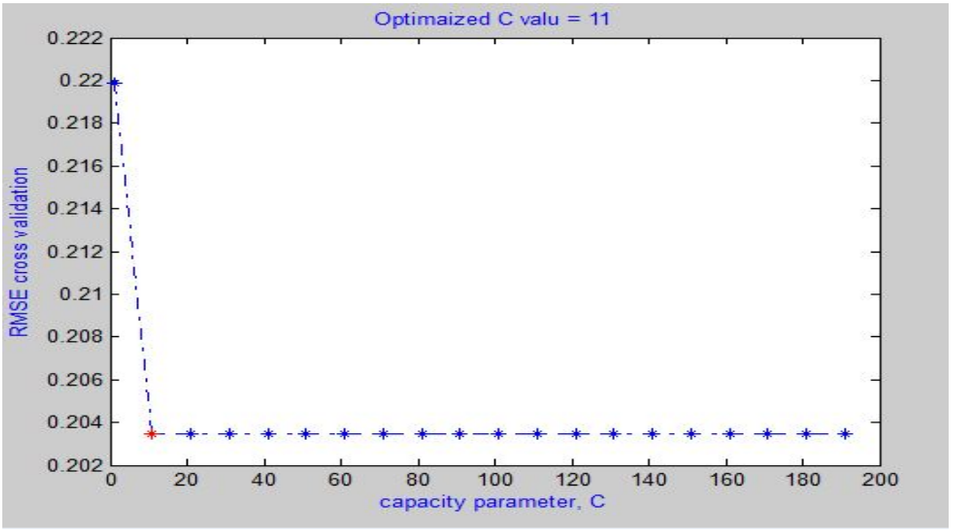
فاکتور حساسیت یکی دیگر از پارامترهایی است که باید بهینه شود. فاکتور حساسیت به نویزهای موجود در داده‌ها مربوط می‌شود که معمولاً ناشناس هستند. در شکل ۶ نمودار تغییرات RMSE برحسب epsilon (ε) نمایش داده شده است.



شکل ٦. نمودار تغییرات مقدار epsilon (ε) برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار epsilon از ۰/۰۱ تا ۰/۹۶ مقدار بهینه ۰/۱۶ برای epsilon انتخاب شد.

و در نهایت پارامتر ظرفیت(C) باید بهینه شود اگر مقدار C پایین باشد یک پراکندگی در پیش‌بینی دیده خواهد شد و در بعضی اوقات با افزایش بیش از حد C، Overfitting رخ می‌دهد هرچند که مقدار زیادی C تاثیر چندانی روی پیش‌بینی ندارد ولی با این حال مقدار این پارامتر نیز باید بهینه شود. در شکل ۷ نمودار تغییرات(Capacity parameter (C بر حسب RMSE نمایش داده شده است.

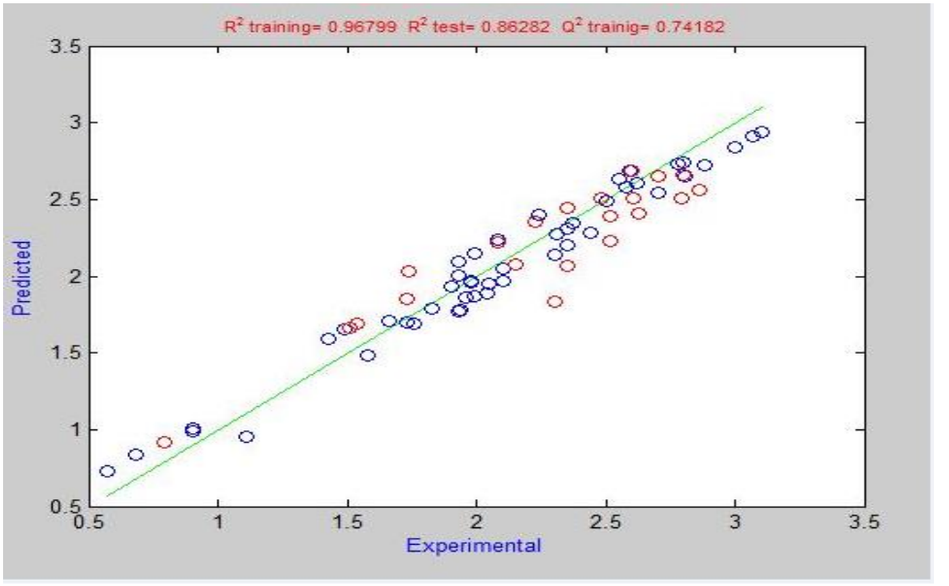


شکل ۷. نمودار تغییرات مقدار Capacity parameter (C) برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار Capacity parameter از ۱ تا ۱۹۱ مقدار بهینه ۱۱ برای Capacity parameter انتخاب شد.

در مرحله آخر با استفاده از تمامی پارامترهای بهینه شده، مدل SVM ساخته شده و مقادیر قطبیت حلال پیش بینی شد. با استفاده از مدل SVM بهینه شده مقادیر قطبیت حلال ترکیبات مورد نظر در مجموعه آموزشی و پیش‌بینی مورد محاسبه قرار گرفت و در جدول۴ نشان داده شده است. همانطور که در این جدول مشاهده می‌شود SVM توانسته است پیش‌بینی‌های بسیار خوبی را برای مقادیر قطبیت حلال‌ها نشان دهد.

مقادیر قطبیت‌های محاسبه‌شده و تجربی برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست در شکل ۸ آورده شده است در این شکل میزان نزدیکی داده‌ها به خط راست قدرت پیش‌بینی مدل را نشان می‌دهد.



شکل ۸. نمودار مقادیر قطبیت حلال محاسبه شده برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست برحسب مقادیر تجربی

۳-۳. نتایج پارامترهای آماری جهت مقایسه مدل‌های انتخاب شده

مطابق جدول ۵، سه پارامتر آماری، جهت ارزیابی توانایی پیش‌بینی مدل‌های ساخته شده به روش‌های SW-MLR و SW-SVM به کار گرفته شد. نتایج جدول نشان دهنده برتری روش SVM به روش MLR می‌باشد.

جدول ۵. پارامترهای آماری برای مدل های انتخاب شده

	RMSE		R^2		F	
	آموزش	تست	آموزش	تست	آموزش	تست
SW-MLR	۰/۲۳۲	۰/۳۸۶	۰/۸۵۲	۰/۶۲۴	۸۴/۸۸۶	۱۳/۴۰۴
SW-SVM	۰/۱۱۹	۰/۲۰۳	۰/۹۶۷	۰/۸۶۲	۳۱۷/۵۰۷	۲۴/۷۹۱

۳-٤. مروری بر توصیف کننده های انتخاب شده

اولین توصیف‌کننده انتخاب شده در مدل، گشتاور دوقطبی می‌باشد که یکی از توصیف‌کننده‌های الکترونی است که از ساختار سه بعدی مولکول به دست می‌آید و رفتار کششی و چرخشی یک مولکول را در یک میدان الکترواستاتیک توجیه می‌کند. این توصیف‌کننده با استفاده از بارهای جزئی اتمی و کئوردینانسیون اتمی تخمین زده می‌شود و دارای واحد دبای می‌باشد. فرمول محاسبه این توصیف‌کننده به صورت زیر است:

$$\mu = -\sum_{t=1(v)}^{0cc} \phi_1 \hat{r} \phi_1 dv + \sum_{a=1}^{M} z_a \vec{R}_a \qquad (۴)$$

که در این معادله μ گشتاور دوقطبی، $\emptyset$ اوربیتال مولکولی، $\hat{r}$ اپراتور موقعیت الکترون، Z_a بار هسته aام و R_a بردار مکانی هسته aام می‌باشد. علامت این توصیف‌کننده در مدل مثبت است که نشان می‌دهد با افزایش یا کاهش گشتاور دوقطبی مولکول، قطبیت نیز افزایش یا کاهش می‌یابد[۲۳].

دومین توصیف‌کننده، سطح مولکول می‌باشد که از نوع توصیف‌کننده‌های هندسی است. این توصیف‌کننده‌ها جزء توصیف‌کننده های سه‌بعدی هستند، یعنی برای محاسبه آن‌ها به ساختار سه‌بعدی مولکول نیاز است. در نمایش سه‌بعدی مولکول، اتم‌ها به صورت کرات صلب با شعاع واندروالسی درنظر گرفته می‌شوند که در محل پیوندها با هم همپوشانی دارند. علامت این توصیف‌کننده در مدل، منفی بوده که نشان‌دهنده تاثیر معکوس اندازه سطح در قطبیت می‌باشد.

آخرین توصیف کننده (اختلاف بین انرژی بالاترین تراز اشغال شده و انرژی پایین ترین تراز اشغال نشده) یک توصیف‌کننده الکترونی بوده که نشان دهنده خواص الکترونی مولکول‌ها می‌باشد و اطلاعات مفیدی در مورد برهمکنش‌های الکترونی بین مولکولی و درون مولکولی در اختیار قرار می‌دهد. این نوع برهمکنش‌ها خواص فیزیکی و شیمیایی مولکول‌ها را شدیداً تحت تاثیر قرار می‌دهند. علامت این توصیف‌کننده نیز در مدل منفی بوده که نشان دهنده تاثیر معکوس E_{Gap} برمقدار قطبیت می‌باشد[۲۳].

٤. نتیجه گیری

در این تحقیق از دو روش رگرسیون خطی چندگانه و ماشین بردار پشتیبان جهت مدل سازی و پیش‌بینی قطبیت برخی از حلالهای آلی استفاده شد. توسط روش مرحله‌ای، سه توصیف‌کننده‌ایی که بیشترین ارتباط را با قطبیت حلال داشتند شامل گشتاور دوقطبی، سطح مولکولی و E_{Gap} انتخاب شدند. سپس رابطه‌ای بین توصیف‌کننده‌های انتخاب شده و قطبیت حلال با استفاده از روش‌های MLR و SVM بدست آمد. نتایج نشان از برتری روش SVM نسبت به MLR دارد. بنابراین می‌توان نتیجه گرفت که توصیف کننده های الکترونی و مکانیک کوانتومی می‌توانند بخوبی جهت پیش بینی قطبیت حلال، بکار گرفته شوند.

٥. مراجع

[1] Reichardt, C. and Welton, T., 2011. *Solvents and solvent effects in organic chemistry*. John Wiley & Sons.
[2] Wong, M.W., Frisch, M.J. and Wiberg, K.B., 1991. Solvent effects. 1. The mediation of electrostatic effects by solvents. *Journal of the American Chemical Society*, *113*(13), pp.4776-4782.

[3] Katritzky, A.R., Fara, D.C., Yang, H., Tämm, K., Tamm, T. and Karelson, M., 2004. Quantitative measures of solvent polarity. *Chemical reviews*, *104*(1), pp.175-198.
[4] Haidekker, M.A., Brady, T.P., Lichlyter, D. and Theodorakis, E.A., 2005. Effects of solvent polarity and solvent viscosity on the fluorescent properties of molecular rotors and related probes. *Bioorganic chemistry*, *33*(6), pp.415-425.
[5] Snyder, L.R., 1974. Classification of the solvent properties of common liquids. *Journal of Chromatography A*, *92*(2), pp.223-230.
[6] Ojha, P.K. and Roy, K., 2018. Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. *Food and Chemical Toxicology*, *112*, pp.551-562.
[7] Banchero, M. and Manna, L., 2018. Comparison between Multi-Linear-and Radial-Basis-Function-Neural-Network-Based QSPR Models for The Prediction of The Critical Temperature, Critical Pressure and Acentric Factor of Organic Compounds. *Molecules*, *23*(6), p.1379.
[8] Safder, U., Nam, K., Kim, D., Shahlaei, M. and Yoo, C., 2018. Quantitative structure-property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. *Ecotoxicology and environmental safety*, *162*, pp.17-28.
[9] Brusseau, M.L., 2019. The influence of molecular structure on the adsorption of PFAS to fluid-fluid interfaces: Using QSPR to predict interfacial adsorption coefficients. *Water research*, *152*, pp.148-158.
[10] Petrosyan, L.S., Sizochenko, N., Leszczynski, J. and Rasulev, B., 2019. Modeling of Glass Transition Temperatures for Polymeric Coating Materials: Application of QSPR Mixture-based Approach. *Molecular informatics*.
[11] Rahimi, M. and Nekoei, M., 2013. Quantitative Structure-Property Relationship Study for Prediction of Flash Point of Some Organic Compounds Based On SW-MLR Method. *Analytical Chemistry Letters*, *3*(4), pp.278-286.
[12] Pourbasheer, E., Beheshti, A., Vahdani, S., Nekoei, M., Danandeh, M., Abbasghorbani, M. and Ganjali, M.R., 2015. Simple QSPR modeling for prediction of the GC retention indices of essential oil compounds. *Journal of Essential Oil Bearing Plants*, *18*(6), pp.1298-1309.
[13] Zarei, K., Atabati, M. and Ebrahimi, M., 2007. Quantitative structure-property relationship study of the solvent polarity using wavelet neural networks. *Analytical Sciences*, *23*(8), pp.937-942.
[14] Todeschini, R. and Consonni, V., 2008. *Handbook of molecular descriptors* (Vol. 11). John Wiley & Sons.
[15] Van Aalten, D.M., Bywater, R., Findlay, J.B., Hendlich, M., Hooft, R.W. and Vriend, G., 1996. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *Journal of computer-aided molecular design*, *10*(3), pp.255-262.
[16] Todeschini, R. and Consonni, V., 2009. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons.
[17] Doucet, J.P., Barbault, F., Xia, H., Panaye, A. and Fan, B., 2007. Nonlinear SVM approaches to QSPR/QSAR studies and drug design. *Current Computer-Aided Drug Design*, *3*(4), pp.263-289.
[18] Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Igglessi-Markopoulou, O. and Kollias, G., 2010. A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs. *Molecular diversity*, *14*(2), pp.225-235.
[19] Nekoei, M., Mohammadhosseini, M. and Pourbasheer, E., 2015. 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), pp.3037-3046.
[20] Wang, Y., Chen, J., Tang, W., Xia, D., Liang, Y. and Li, X., 2019. Modeling adsorption of organic pollutants onto single-walled carbon nanotubes with theoretical molecular descriptors using MLR and SVM algorithms. *Chemosphere*, *214*, pp.79-84.

[21] Padierna, L.C., Carpio, M., Rojas-Domínguez, A., Puga, H. and Fraire, H., 2018. A novel formulation of orthogonal polynomial kernel functions for SVM classifiers: The Gegenbauer family. *Pattern Recognition*, *84*, pp.211-225.

[22] Roxas, E.A., Vicerra, R.R.P., Lim, L.A.G., Dadios, E.P. and Bandala, A.A., 2018. SVM Compound Kernel Functions for Vehicle Target Classification. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *22*(5), pp.654-659.

[23] Zohari, N., Sheibani, N. and Chavoshi, H.Z., 2018. Investigation of the most effective molecular descriptors on the thermal behaviour of energetic azido-ester plasticizers through QSPR approach. *Journal of Thermal Analysis and Calorimetry*, *131*(3), pp.3157-3167.

Quantitative Structure-Property Relationship Study for Prediction of the Solvent Polarity Using Quantum Mechanics Descriptors and Support Vector Machine

Mehdi Nekoei*, Behzad Chahkandi

Department of Chemistry, Faculty of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Submited: 26 January 2019, Revised: 14 March 2019, Accepted: 14 April 2019

Abstract

Quantitative structure-property relationship (QSPR) study for prediction of the polarity some of solvents using quantum mechanics descriptors and support vector machine. Experimental S′ values for 69 solvents were assembled. This set included saturated and unsaturated hydrocarbons, solvents containing halogen, cyano, nitro, amide, sulfide, mercapto, sulfone, phosphate, ester, ether, etc. After drawing the structure of the molecules, the suitable molecular descriptors were calculated. Then, the stepwise multiple linear regressions (SW-MLR) variable selection method was subsequently employed to select and implement the prominent descriptors having the most significant contributions to the polarity of the molecules. At first, multiple linear regressions (MLR) model was constructed. Then, support vector machine (SVM) model was used for to obtain better results. A comparison of results by the two methodologies indicated the superiority of SW-SVM over the SW-MLR method.

Keywords: *Quantitative structure- property relationship (QSPR), Solvent polarity, Quantum mechanics descriptors, Support vector machine (SVM).*

***Corresponding author :** Mehdi Nekoei

Adress: Department of Chemistry, Faculty of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Tel: 02332394289 **E-mail:** m_nekoei1356@yahoo.com