مجله شیمی کوانتومی و اسپکتروسکوپی
۳۲:۹ (۱۳۹۸) ۶۹-۸۶
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مدل سازی و مطالعه ارتباط کمی ساختار-خاصیت (QSPR) جهت پیش بینی ثابت های اسیدی برخی از ترکیبات شیمیایی با استفاده از رگرسیون خطی چندگانه و ماشین بردار پشتیبان

سید عباس طاهری، مهدی نکوئی*، مجید محمدحسینی

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

تاریخ ثبت اولیه:۱۳۹۸/۱۰/۱۰، تاریخ دریافت نسخه اصلاح شده:۱۳۹۹/۰۲/۱۵ ، تاریخ پذیرش قطعی:۱۳۹۹/۰۳/۲۸

چکیده

مدل سازی و مطالعه ارتباط کمی ساختار-خاصیت(QSPR) جهت پیش بینی ثابت های اسیدی برخی از ترکیبات شیمیایی با استفاده از روش رگرسیون خطی چند گانه(MLR) و ماشین بردار پشتیبان (SVM) انجام شد. در ابتدا ساختار ترکیبات شیمیایی، ترسیم و گروه مناسبی از توصیف کننده‌ها محاسبه گردید. سپس با استفاده از روش انتخاب مرحله‌ای برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با خاصیت شیمیایی ترکیبات مورد نظر داشتند استفاده گردید. سپس از مدل خطی رگرسیون خطی چندگانه(MLR) و مدل غیرخطی ماشین بردارشتیبان (SVM) جهت پیش بینی ثابت های اسیدی ترکیبات استفاده گردید. داده های آماری، حاکی از برتری روش SVM نسبت به روش MLR بود.

واژه های کلیدی: ***ارتباط کمی ساختار- خاصیت، ثابت های اسیدی (pKa) ، رگرسیون خطی چندگانه، ماشین بردار پشتیبان.***

۱. مقدمه

مطالعه ارتباط کمی ساختار- خاصیت یکی از کارآمدترین روش های غیر آزمایشگاهی در پیش بینی و تعیین خواص ترکیبات شیمیایی محسوب می شود. این روش در سالیان اخیر توانسته بخوبی جهت تعیین خواص برخی از ترکیبات شیمیایی مانند زمان های بازداری ترکیبات در کروماتوگرافی، خواص شیمی فیزیکی ترکیبات، فعالیت برخی از ترکیبات دارویی و... مفید باشد[۹-۱]. در روش های آزمایشگاهی به دلایل عدیده ای از جمله گرانی و هزینه بالای مواد شیمیایی، وقت گیر بودن و زمان زیادی که صرف انجام آزمایش می شود، همچنین آسیب رسانی به محیط زیست و آلودگی های حاصل از مواد شیمیایی مخاطرات بسیاری را به دنبال

***عهده دار مکاتبات:** مهدی نکوئی

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

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

دارد در حالی که در روش های نوین مانند QSPR[1] انجام محاسبات شیمیایی برای پیش بینی خواص ترکیبات شیمیایی با هزینه ای بسیار کمتر و همراه با عدم آلودگی محیط زیست و نیز دقت بالا امکان پذیر می باشد. بنابراین می توان گفت بدلیل وقت گیر و هزینه بر بودن روش های آزمایشگاهی استفاده از شیوه هایی جهت پیش بینی خواص ترکیبات شیمیایی ضروری به نظر می رسد. مدل QSPR یک رابطه ریاضی بین ساختار و خاصیت گروهی از ترکیبات شیمیایی را توصیف می کند. از اینرو استفاده از اطلاعات مدل QSPR پیش از بررسی آزمایشگاهی ترکیبات شیمیایی در آزمایشگاه مهم به نظر می رسد. شیوه ها و روش های متنوعی از جمله رگرسیون خطی چندگانه(MLR[2])، حداقل مربعات جزیی(PLS[3])، ماشین بردار پشتیبان(SVM[4]) و شبکه های عصبی مصنوعی(ANN[5])در مدلسازی های QSPR می توانند مورد استفاده قرارگیرد[۱۵-۱۰].

این تحقیق با هدف پیش بینی ارتباط کمی ساختار-خاصیت (QSPR) به منظور پیش بینی ثابت های اسیدی برخی از ترکیبات شیمیایی با کمک روش های رگرسیون خطی چندگانه(MLR) و ماشین بردار پشتیبان(SVM) انجام شد.

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

در این تحقیق تعداد ۲۴۲ ترکیب از ترکیبات شیمیایی از طریق روشهای کمومتریکس مورد بررسی قرار گرفت[۱۶]. در این مقاله ثابت اسیدی این ترکیبات شیمیایی به صورت pK_a گزارش شده است.

در این تحقیق این ترکیبات به دو گروه سری آموزش و سری آزمون و به صورت تصادفی تقسیم شده است، سری آموزش شامل ۱۹۴ مولکول(۸۰٪ داده ها) و سری آزمون شامل ۴۸ مولکول (۲۰٪ داده ها) می‌باشد. مقادیر pK_a به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند. سری آموزش جهت ایجاد یک مدل مناسب و سری آزمون جهت ارزیابی مدل مورد استفاده قرار گرفت.

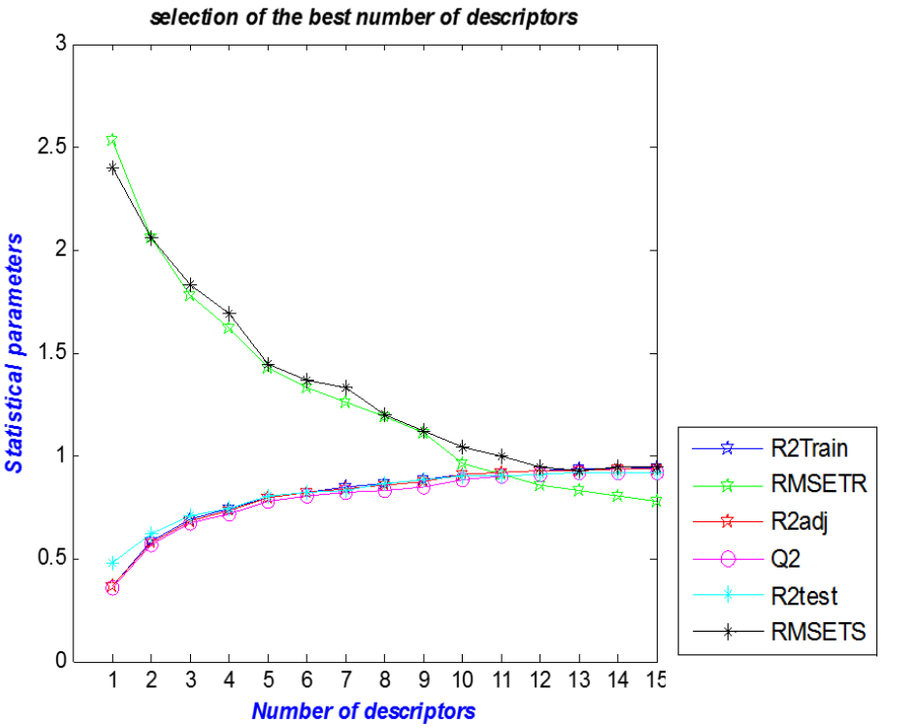
۲-۲. محاسبه توصیف‌کننده‌ها، غربالگری و گزینش بهترین آنها

توصیف کننده‌ها مقادیری هستند عددی که بیانگر خصوصیات متنوعی از مولکول می باشند. امروزه و با توجه به بروز بودن نرم افزارهای مربوط به انتخاب توصیف کننده، توصیف کننده های مولکولی بسیاری وجود دارد که بعد از غربالگری و حذف توصیف کننده های غیرضرور و به دنبال آن ارزیابی و یافتن مناسب‌ترین توصیف کننده از بین آنها می‌توانند به منظور پیش‌بینی خاصیت شیمیایی ترکیبات جدید در مطالعات QSPR استفاده نمود.

---

[1] Quantitative structure property relationship (QSPR)
[2] Multiple Linear Regression (MLR)
[3] Partial least squares regression (PLS regression)
[4] Support-vector machine (SVM)
[5] Artificial neural network (ANN)

به منظور انتخاب مناسب ترین توصیف کننده ها و مطلوب ترین آنها از روش رگرسیون مرحله ای[1] استفاده شد. بر طبق روش فوق الذکر، متغیرها بصورت پیوسته وارد مدل شده پس از آن، ابتدا متغیری که بالاترین میزان همبستگی را با متغیر وابسته داشت وارد مدل گردید. با ورود تک تک متغیرهای جدید، سایر متغیرهای موجود در معادله مورد بررسی قرار گرفته و در صورت از دست دادن سطح معناداری خود، با ورود متغیری جدید از مدل کنار گذاشته می‌شود. بدین سان ترتیب داده‌های pK_a یا همان متغیر وابسته و توصیفگرها، متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله‌ای اعمال شد. می‌دانیم که روش رگرسیون مرحله ای شمار بسیاری مدل ارائه می دهد. که اولین مدل دربرگیرنده یک توصیف کننده، مدل دوم شامل دو توصیف کننده و ... بوده و با افزایش شمار توصیف کننده ها بطورطبیعی مقدار R^2 افزایش و RMSE[2](میانگین خطای مجذور مربعات)کاهش می‌یابد. اما به سبب پیچیدگی مدل، نمی‌توانیم تعداد زیادی توصیف کننده را برای مدلسازی برگزینیم. لذا برای این منظور و جهت گزینش شمار توصیف کننده های مناسب، نمودار پارامتر های آماری مختلف از جمله $RMSE_{train}$, $RMSE_{test}$،R^2_{train}, R^2_{test} بر اساس شمار توصیف کننده ها ترسیم، که در شکل۱ آورده شده است. با توجه به نمودار، تعداد ۱۲ توصیف کننده به عنوان توصیف کننده هایی با بالاترین میزان ارتباط با خاصیت ترکیبات شیمیایی، انتخاب شدند. این ۱۲ توصیف کننده در جدول۱ ارائه شده است.



شکل ۱. نمودار پارامتر های آماری (R^2_{train}, R^2_{test}, $RMSE_{train}$, $RMSE_{test}$, Q_2 , R_{2adj}) برحسب تعداد توصیفگرها

[1] Stepwise
[2]Root-mean-square-error

جدول ۱. توصیفگرهای انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

نشانه توصیف کننده	معنی توصیف کننده	نوع توصیف کننده
IC1	Information content index (neighborhood symmetry of 1-order).	Topological descriptors.
SEigp	Eigenvalue sum from polarizability weighted distance matrix.	Topological descriptors.
MATS1m	Moran autocorrelation – lag 1 /weighted by atomic masses.	2D autocorrelations.
GATS3e	Moran autocorrelation – lag 3 /weighted by atomic Scanderson electronegativities.	2D autocorrelations.
GATS4e	Moran autocorrelation – lag 4 /weighted by atomic Scanderson electronegativities.	2D autocorrelations.
RPCG	Relative positive charge.	Charge descriptors.
MAXDN	Maximal electrotopological negative variation	Geometrical descriptors.
Mor30m	3D-MoRSE – signal 30 / weighted by atomic masses.	3D-MoRSE descriptors.
HATS2m	Leverage-weighted autocorrelation of lag 2/ weighted by atomic masses.	GETAWAY descriptors.
R2e	R autocorrelation of lag 2 / weighted by atomic Sanderson electronegativities.	GETAWAY descriptors.
n-OH	Number of total hydroxyl group.	Functional groups.
C-040	R-C(=X)-X / R-C#X / X-=C=X.	Atom-centred fragments.

۳. نتایج و بحث

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

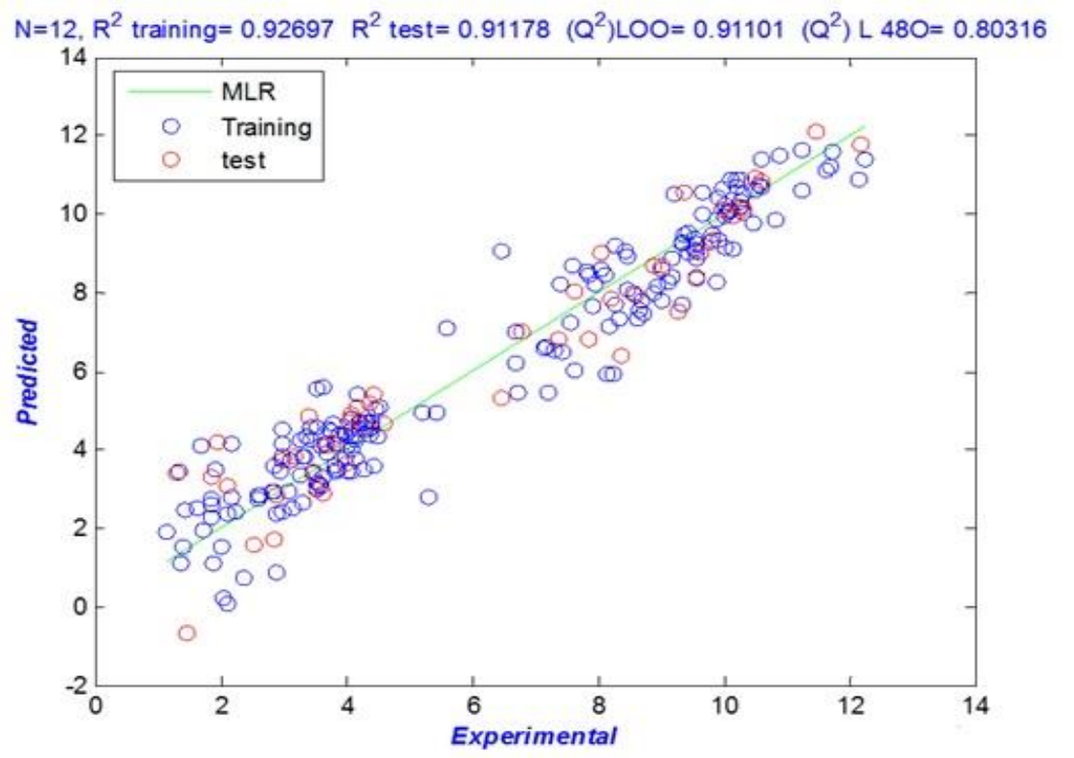
پس از گزینش مناسب‌ترین توصیف کننده ها با استفاده از روش مرحله‌ای، گام بعدی، تولید مدل بین توصیف کننده های برگزیده و pK_a می‌باشد. بین توصیف کننده‌ها و خاصیت ترکیبات شیمیایی برای سری آموزش با استفاده از روش MLR رابطه زیر تحت عنوان مدلی خطی بدست آمد:

$$\mathbf{pK_a}=46.725-1.951(\mathbf{IC1})-0.426(\mathbf{SEigp})-39.037(\mathbf{MATS1m})+1.015(\mathbf{GATS3e})+0.527(\mathbf{GATS4e})+7.647(\mathbf{RPCG})-2.903(\mathbf{MAXDN})+2.333(\mathbf{Mor30m})+5.358(\mathbf{HATS2m})+0.795(\mathbf{R2e})+0.661(\mathbf{nOH})-0.571(\mathbf{C\text{-}040})$$

[1] Multiple linear regression (MLR)

پس از آن از معادله ایجاد شده برای پیش‌بینی خاصیت ترکیبات شیمیایی سری آزمون استفاده گردید. داده های تجربی و پیش‌بینی شده pK_a برای تمامی ترکیبات سری آموزش و آزمون در جدول ۲ گردآوری شده است. شکل۲ نمودار مقادیر پیش بینی شده را نسبت به مقادیر تجربی (نمودار برگشتی) نشان می دهد.

در نمودار برگشتی مقادیر پیش‌بینی شده برحسب مقادیر تجربی رسم می‌گردد و با توجه به مقدار ضریب تعیین (R^2) به دست آمده از نمودار، پراکندگی نقاط در اطراف خط برگشت تعیین می‌شود. هر چه مقدار ضریب تعیین به یک نزدیک‌تر باشد، مدل ساخته شده، مدل بهتری است. همتنطور که مشاهده می گردد ضرایب تعیین ۰/۹۲۶ برای سری آموزش و ۰/۹۱۱ برای سری تست نشان از پیش بینی نسبتا خوب مدل پیشنهادی دارد.



شکل ۲. نمودار مقادیر پیش‌بینی شده pK_a بر حسب مقادیر تجربی برای سری آموزش و تست به روش MLR

جدول ۲. ساختار ترکیبات شیمیایی مورد استفاده و مقادیر «pK_a» تجربی و پیش بینی شده با استفاده از روشهای SW-MLR و SW-SVM

No.	Name	EXP. (pK_a)	Pred. (MLR)	Pred. (SVM)
m1	2,6-dinitrobenzoic acid	1.14	1.90	1.19
m2	2-chloro-6-nitronenzoic acid	1.34	1.45	1.39
m3	2-bromo-6-nitrobenzoic acid	1.37	1.09	1.42
m4	2,4,6-tribromobenzoic acid	1.41	1.55	1.46
m5	2,4-dinitrobenzoic acid	1.43	1.47	1.48
m6	2,5-dinitrobenzoic acid	1.62	2.51	1.67
m7	2,4,6-trihydroxybenzoic acid	1.68	4.08	1.73
m8	2-nitrobenzen-1,4-dicarboxylic acid	1.73	1.96	1.78
m9	2,3-dinitrobenzoic acid	1.85	2.29	1.80
m10	2-hydroxy-3-nitrobenzoic acid	1.87	2.75	1.92
m11	2-methyl-6-nitrobenzoic acid	1.87	2.63	1.92
m12	3-nitrobenzebe-1,2-dicarboxylic acid	1.88	1.12	1.83
m13	1,2,4,5-benzentetracarboxylic acid	1.92	3.49	1.97
m14	2-cholro-3-nitrobenzoic acid	2.02	1.52	1.97
m15	1,2,3,4-benzenetetracarboxylic acid	2.05	0.21	2.10
m16	4-nitrobenzene-1,2-dicaboxylic acid	2.11	0.10	2.16
m17	1,3,5- benzenetricaboxylic acid	2.12	2.38	2.17
m18	2-chloro-5-nitronenzoic acid	2.17	4.16	2.22
m19	2-nitrobenzoic acid	2.18	2.78	2.23
m20	2-hydroxy-6-nitrobenzoic acid	2.24	2.42	2.19
m21	1,2,3,5-benzenetetracarboxylic acid	2.38	0.72	2.33
m22	3-amino-1-naphtoic acid	2.61	2.77	2.66
m23	2-hydroxy-5-bromobenzoic acid	2.61	2.85	2.56
m24	2-hydroxy-5-chlorobenzoic acid	2.63	2.83	2.68
m25	3,4-dinitrobenzoic acid	2.82	2.93	2.77
m26	3,5-dinitrobenzoic acid	2.85	2.95	2.80
m27	2-Iodobenzoic acid	2.86	3.57	2.91
m28	2-chlorobenzoic acid	2.88	2.35	2.93
m29	1,2,3-benzenetricaboxylic acid	2.88	0.87	2.83
m30	orto-phthalic acid	2.95	3.44	3.00
m31	2,5-dihydroxybenzoic acid	2.97	4.53	3.02
m32	2-methyl-3,5-dinitrobenzoic acid	2.97	2.43	2.92
m33	4-chloro-2,6-dinitrophenol	2.97	3.83	3.02

m34	2-hydroxy-3-methylbenzoic acid	2.99	4.13	3.04
m35	Benzilic acid	3.09	2.92	3.14
m36	2-methyl-1-naphthoic acid	3.11	3.71	3.16
m37	2-cyanobenzoic acid	3.14	2.52	3.09
m38	2-fluorobenzoic acid	3.27	3.35	3.22
m39	2,4-dihydroxybenzoic acid	3.29	4.26	3.24
m40	2,6-di-iodo-4-nitrophenol	3.32	2.64	3.37
m41	2-hydroxy-6-methylbenzoic acid	3.32	3.82	3.37
m42	2,3-dimethylnaphthalene-1-carboxylic acid	3.33	3.82	3.29
m43	2,6-dimethylbenzoic acid	3.36	4.33	3.41
m44	2,6-dimethoxybenzoic acid	3.44	4.24	3.49
m45	2,4,6-trimethylbenzoic acid	3.45	4.55	3.50
m46	2-biphenylcarboxylic acid	3.46	3.46	3.51
m47	3-nitrobenzoic acid	3.46	3.44	3.51
m48	2-acetoxybenzoic acid	3.48	3.46	3.53
m49	3-methylsulfonylbenzoic acid	3.52	5.56	3.57
m50	2-phenoxybenzoic acid	3.53	4.57	3.58
m51	1,4-benzenedicarboxylic acid	3.54	3.02	3.49
m52	2-benzoylbenzoic acid	3.54	3.15	3.49
m53	4-cyanobenzoic acid	3.55	3.14	3.57
m54	Benzylamine-4-carboxylic acid	3.59	4.10	3.64
m55	3-cyanobenzoic acid	3.60	3.09	3.55
m56	2-acetamidobenzoic acid	3.63	3.33	3.58
m57	4-methylsulfonylbenzoic acid	3.64	5.59	3.69
m58	Anthracene-9-carboxylic acid	3.65	4.15	3.70
m59	1-naphthalenecarboxylic acid	3.70	3.92	3.75
m60	2,6-dinitrophenol	3.71	4.50	3.76
m61	2,3-dimethylbenzoic acid	3.77	4.49	3.82
m62	2-ethylbenzoic acid	3.79	4.68	3.84
m63	3-bromobenzoic acid	3.81	3.59	3.80
m64	3-chlorobenzoic acid	3.83	3.47	3.88
m65	3-iodobenzoic acid	3.86	3.50	3.84
m66	3-fluorobenzoic acid	3.87	4.16	3.88
m67	2-methylbenzoic acid	3.90	4.39	3.95
m68	2,5-dimethylbenzoic acid	3.99	4.38	3.94

m69	4-bromobenzoic acid	3.99	3.80	3.94
m70	4-iodobenzoic acid	4.00	3.45	3.95
m71	3-acetoxybenzoic acid	4.00	4.09	4.05
m72	3-acetamidobenzoic acid	4.07	3.43	4.12
m73	3-hydroxybenzoic acid	4.08	4.50	4.13
m74	2,4-dinitrophenol	4.08	4.80	4.13
m75	2-hydroxy-5-methylbenzoic acid	4.08	4.01	4.03
m76	2-methoxybenzoic acid	4.09	4.36	4.14
m77	2-acetylbenzoic acid	4.13	3.76	4.08
m78	4-fluorobenzoic acid	4.14	4.34	4.09
m79	2-naphthalenecarboxylic acid	4.16	4.23	4.11
m80	3,4,5-trihydroxybenzoic acid	4.19	5.42	4.24
m81	3-tert-butylbenzoic acid	4.20	4.74	4.25
m82	benzoic acid	4.20	4.65	4.26
m83	2,4-dimethylbenzoic acid	4.22	4.53	4.17
m84	4-acetamidobenzoic acid	4.28	3.51	4.23
m85	3,5-dimethylbenzoic acid	4.30	4.70	4.35
m86	Mesitylenic acid	4.32	4.71	4.36
m87	4-ethylbenzoic acid	4.35	4.72	4.30
m88	4-methylbenzoic acid	4.36	4.54	4.31
m89	4-acetoxybenzoic acid	4.38	4.39	4.33
m90	4-tert-butylbenzoic acid	4.39	4.72	4.34
m91	3,4-dimethybenzoic acid	4.41	4.58	4.36
m92	4-nitrobenzoic acid	4.44	3.60	4.39
m93	3,4-dihydroxybenzoic acid	4.48	5.04	4.43
m94	4-methoxybenzoic acid	4.49	4.36	4.14
m95	4-phenoxybenzoic acid	4.52	5.10	4.57
m96	2,5-dinitrophenol	5.22	4.95	5.16
m97	3,5-diaminobenzoic acid	5.30	2.79	5.25
m98	3,4-dinitrophenol	5.42	4.97	5.37
m99	2,2'methylenbis(4,6-dichlorophenol)	5.60	7.09	5.65
m100	4-nitrosophenol	6.48	9.06	6.53
m101	1,2-dihydroxy-3-nitrobenzene	6.68	6.20	6.63
m102	1,2-dihydroxy-4-nitrobenzene	6.70	6.99	6.75
m103	3,5-dinitrophenol	6.73	5.48	6.68

m104	4-nitrophenol	7.15	6.59	7.20
m105	2,6-dimethyl-4-nitrophenol	7.19	6.64	7.24
m106	2-nitrophenol	7.22	5.47	7.17
m107	1,3-dichloro-2,5-dihydroxybenzene	7.30	6.52	7.25
m108	4-hydroxy-3-methoxybenzaldehyde	7.40	8.22	7.69
m109	2,3-dichlorophenol	7.44	6.48	7.42
m110	3,4-dihydroxybenzaldehyde	7.55	7.25	7.60
m111	2,2'-methylenbis(4-chlorophenol)	7.60	8.71	7.65
m112	2-nitrohydroquinone	7.63	6.03	7.58
m113	4-methylsulfonylphenol	7.83	8.53	7.88
m114	2,4-dibromophenol	7.85	8.47	7.80
m115	2-hydroxy-3-methoxybenzaldehyde	7.91	7.65	7.96
m116	4-hydroxybenzonitrile	7.95	8.23	7.90
m117	4-acetylphenol	8.05	9.00	8.10
m118	3,5-dibromophenol	8.06	8.62	8.01
m119	3,5-diiodophenol	8.10	8.47	8.05
m120	4-methylfulfonyl-3,5-dimethylphenol	8.13	5.91	8.08
m121	3,5- dichlorophenol	8.18	7.16	8.13
m122	3,5- dimethyl-4-nitrophenol	8.25	5.92	8.19
m123	4-cyano-2,6-dimethylphenol	8.27	9.18	8.22
m124	4-(diethoxyphosphinyl)phenol	8.28	7.70	8.33
m125	2-hydroxybenzaldehyde	8.34	7.32	8.39
m126	1,3,5-trihydroxybenzene	8.45	9.08	8.40
m127	2-bromophenol	8.45	8.07	8.42
m128	2-Iodophenol	8.46	8.91	8.51
m129	2-chlorophenol	8.55	8.00	8.60
m130	3,4-dichlorophenol	8.63	7.36	8.58
m131	4-hydroxy-α,α,α-trifluorotoluene	8.68	7.57	8.62
m132	3-(diethoxyphosphinyl)phenol	8.68	7.80	8.63
m133	2-fluorophenol	8.73	7.48	8.68
m134	3-hydroxy-4-methoxybenzaldehyde	8.89	7.97	8.57
m135	3-trifluoromethylphenol	8.95	8.19	8.90
m136	3-hydroxybenzaldehyde	9.00	7.80	8.95
m137	1,2,3-trihydroxybenzene	9.03	8.62	8.98
m138	3-chlorophenol	9.10	8.25	9.05

m139	2-acetylphenol	9.19	8.39	9.24
m140	3'-hydroxyacetophenone	9.19	8.88	9.14
m141	4-Indophenol	9.20	10.53	10.25
m142	3-methylsulfonylphenol	9.33	7.71	9.28
m143	4-bromophenol	9.34	9.23	9.29
m144	3,5-dimethoxyphenol	9.35	9.28	9.29
m145	1,2-dihydroxybenzene	9.36	9.48	9.41
m146	4-chlorophenol	9.43	9.00	9.38
m147	1,3-dihydroxybenzene	9.44	9.52	9.41
m148	3-(s-methylthio)phenol	9.53	9.13	9.58
m149	4-(s-methylthio)phenol	9.53	9.40	9.58
m150	4-chloro-3-methylphenol	9.55	8.38	9.50
m151	2-phenylphenol	9.55	9.25	9.60
m152	4-phenylphenol	9.55	9.08	9.50
m153	1-hydroxy-2,4,6-trihydroxymethylbenzene	9.56	8.89	9.51
m154	3-methoxyphenol	9.65	10.01	9.70
m155	3-ethoxyphenol	9.66	10.57	9.60
m156	1-hydroxy-2,4-dihydroxymethylbenzene	9.79	9.32	9.74
m157	4-hydroxybenzyl alcohol	9.82	9.51	9.87
m158	4-fluorophenol	9.89	9.85	9.84
m159	2'-hydroxyacetophenone	9.90	8.29	9.85
m160	1,4-dihydroxybenzene	9.91	10.42	9.86
m161	2-hydroxybenzyl alcohol	9.92	9.36	9.89
m162	Phenol	9.99	10.65	9.94
m163	2-methoxy-4(2-propenyl)phenol	10.00	9.17	9.95
m164	m-cresol	10.00	9.97	10.05
m165	4-ethylphenol	10.00	10.23	10.05
m166	1,3-dihydroxy-2-methylbenzene	10.05	10.03	10.00
m167	3-ethylphenol	10.07	10.08	10.12
m168	3-tert-butylphenol	10.10	10.08	10.15
m169	2-ethoxyphenol	10.11	10.87	10.06
m170	(2-hydroxy-5-methylbenzene)-methanol	10.15	9.10	10.10
m171	2-ethylphenol	10.20	10.91	10.25
m172	4-methoxyphenol	10.20	10.71	10.15
m173	2,5-dimethylphenol	10.22	10.70	10.27

m174	o-cresol	10.26	10.88	10.31
m175	p-cresol	10.26	10.17	10.21
m176	2-allylphenol	10.28	10.19	10.23
m177	5,6,7,8-tetrahydro-1-naphthol	10.28	10.32	10.33
m178	3,4-dimethylphenol	10.32	10.03	10.27
m179	4-indanol	10.32	10.52	10.27
m180	5,6,7,8-tetrahydro-2-naphthol	10.48	9.77	10.43
m181	2,3-dimethylphenol	10.50	10.62	10.45
m182	2,4,5-trimethylphenol	10.57	10.76	10.58
m183	2,4-dimethylphenol	10.58	10.76	10.53
m184	2,6-dimethylphenol	10.59	11.41	10.54
m185	2-methyl-4-tert-butylphenol	10.59	10.70	10.64
m186	2,6-di-tert-butyl-4-bromophenol	10.83	9.86	10.78
m187	2,4,6-trimethylphenol	10.88	11.48	10.93
m188	2-tert-butylphenol	11.24	10.59	11.19
m189	1,4-dihydroxy-2,3,5,6-tetramethylbenzene	11.25	11.63	11.20
m190	2,4-di-tert-butylphenol	11.64	11.10	11.65
m191	2,6-di-tert-butylphenol	11.70	11.24	11.75
m192	6-methyl-2-butylphenol	11.72	11.59	11.67
m193	2,6-di-tert-butyl-4-methoxyphenol	12.15	10.89	12.10
m194	2,6-di-tert-butyl-4-methylphenol	12.23	11.43	12.18
TEST SET				
m1	2,6-dihydroxybenzoic acid	1.30	3.42	2.91
m2	3,6-dichlorophethalic acid	1.46	-0.69	1.39
m3	2-methyl-4-nitrobenzoic acid	1.86	3.33	2.72
m4	2-chloro-4-nitronenzoic acid	1.96	4.19	2.40
m5	2-hydroxy-5-nitrobenzoic acid	2.12	3.07	2.44
m6	1,2,4- benzenetricaboxylic acid	2.52	1.59	2.58
m7	2-bromobenzoic acid	2.85	1.71	2.84
m8	4-amino-2-naphtoic acid	2.89	2.83	3.00
m9	2-hydroxybenzoic acid	2.98	3.71	3.50
m10	2-hydroxy-4-methylbenzoic acid	3.17	3.81	3.75
m11	2,3,5,6-tetramethylbenzoic acid	3.42	4.84	4.46
m12	4-sulfamylbenzoic acid	3.47	3.42	3.29
m13	3-sulfamylbenzoic acid	3.54	3.00	3.19

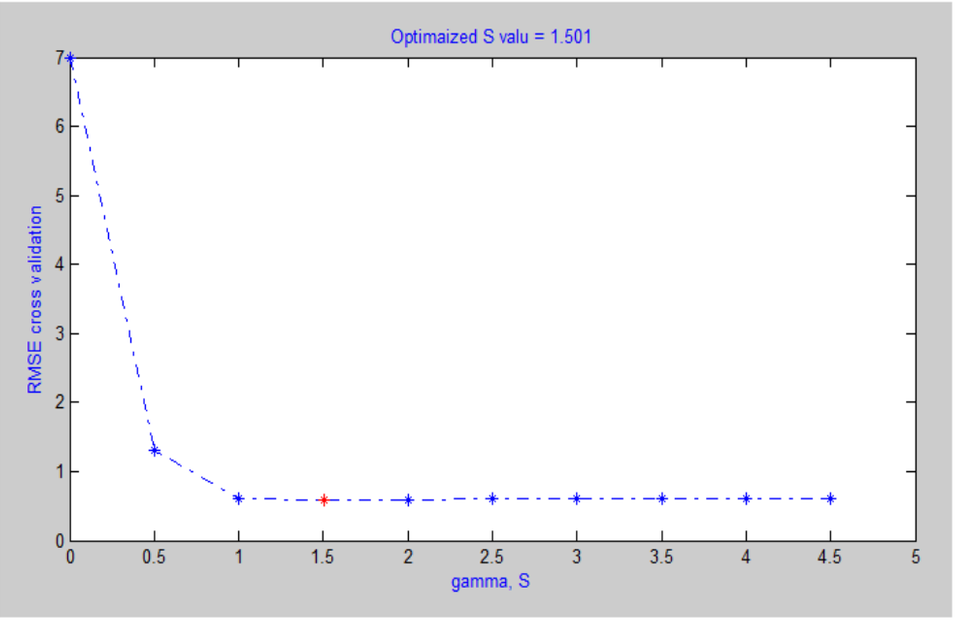
m14	1,3-benzenedicarboxylic acid	3.62	2.87	3.39
m15	4-acetylbenzoic acid	3.70	4.10	4.31
m16	3-acetylbenzoic acid	3.83	4.14	4.33
m17	4-chlorobenzoic acid	3.99	3.63	4.04
m18	3,5-dihydroxybenzoic acid	4.04	4.71	3.27
m19	3-methoxybenzoic acid	4.08	4.90	4.67
m20	Anthracene-2-carboxylic acid	4.18	5.09	4.48
m21	3-methylbenzoic acid	4.27	4.58	4.28
m22	4-hydroxy-3-methoxybenzoic acid	4.36	5.18	4.32
m23	1,4-dihydroxy-2,6-dinitrobenzene	4.42	5.42	5.12
m24	4-hydroxybenzoic acid	4.58	4.66	4.27
m25	4-chloro-2-nitrophenol	6.48	5.34	5.51
m26	2,6-dibromophenol	6.78	6.99	6.78
m27	2,4,5-trichlorophenol	7.37	6.83	7.76
m28	4-hydroxybenzaldehyde	7.62	8.05	8.91
m29	3,4,5-trichlorophenol	7.84	6.84	7.61
m30	4'-hydroxyacetophenone	8.05	9.01	8.24
m31	4-cyano-3,5-dimethylphenol	8.21	7.86	7.06
m32	3-nitrophenol	8.36	6.39	7.32
m33	3-cyanophenol	8.61	7.96	7.55
m34	3-iodophenol	8.88	8.67	8.62
m35	3-bromophenol	9.03	8.69	9.00
m36	3-fluorophenol	9.29	7.50	7.75
m37	3,5-diethoxyphenol	9.37	10.58	8.95
m38	1-chloro-2,6-dimethyl-4-hydroxybenzene	9.55	8.43	9.46
m39	3-phenylphenol	9.63	9.00	9.43
m40	3-hydroxybenzyl alcohol	9.83	9.32	9.88
m41	2-methoxyphenol	9.99	10.15	10.10
m42	2,4,6-tri-tert-butylphenol	12.19	11.76	12.02
m43	2,4,6-tripropylphenol	11.47	12.09	11.55
m44	2,3,4-trimethylphenol	10.59	10.83	10.56
m45	1-hydroxy-2-propylbenzene	10.50	10.92	10.66
m46	4-tert-butylphenol	10.31	10.14	10.29
m47	3,4,5-trimethylphenol	10.25	10.19	10.34
m48	3,5-dimethylphenol	10.15	9.96	10.17

۲-۳. ایجاد مدل با استفاده از ماشین بردار پشتیبان (SVM)

به منظور حصول نتایج مطلوب تر از روش SVM استفاده گردید. در این مرحله توصیف کننده های انتخاب شده به روش رگرسیون مرحله ای، به منظور مدل سازی و پیش‌بینی ثابت های اسیدی ترکیبات مذکور به روش غیر خطی ماشین بردار پشتیبان مورد بررسی قرار گرفت. در این روش قبل از شروع مدل سازی لازم بود تا پارامترهای تاثیرگذار بر روی قدرت مدلسازی SVM بهینه شود. این پارامترها عبارتند از:

Capacity parameter (C) - Epsilon (ε) – Gamma (γ) - Kernel function type (RBF)

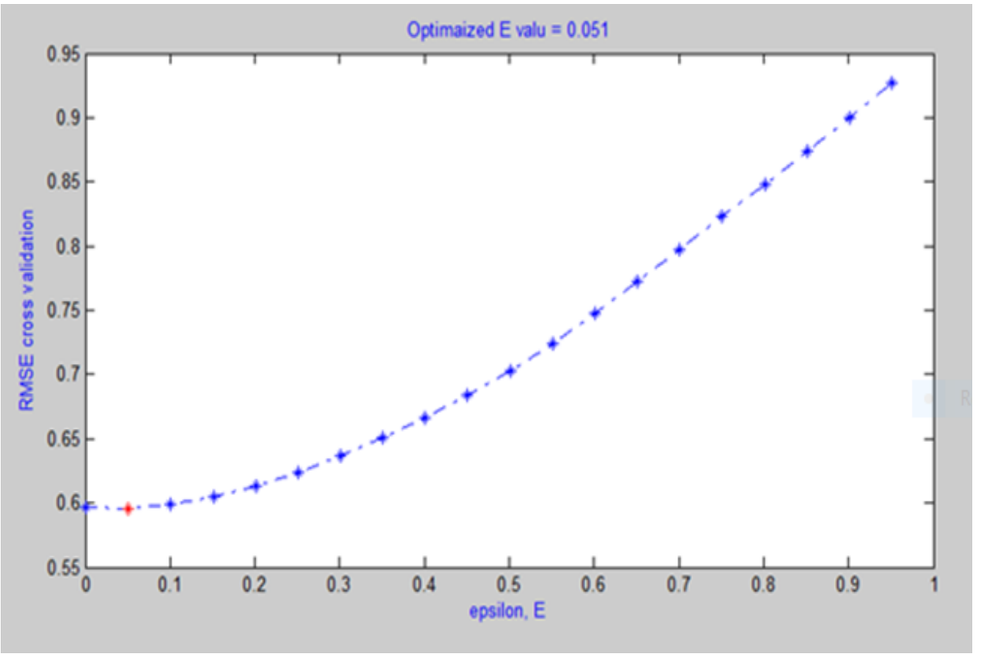
تعداد بردارهای پشتیبان بر زمان آموزش مدل تاثیر می گذارد بطوری که افزایش مقدار Gamma و در نتیجه تعداد بردار پشتیبان می تواند به افزایش زمان آموزش و همچنین Overfitting منجر شود. مقدار Gamma توانایی و قدرت SVM را در پیشگویی کنترل می کند. در شکل ۳ نمودار مقادیر متفاوت RMSE بر حسب Gamma نمایش داده شده است.



شکل ۳. نمودار تغییرات مقدار RMSE برحسب مقدار Gamma برای سری آموزش

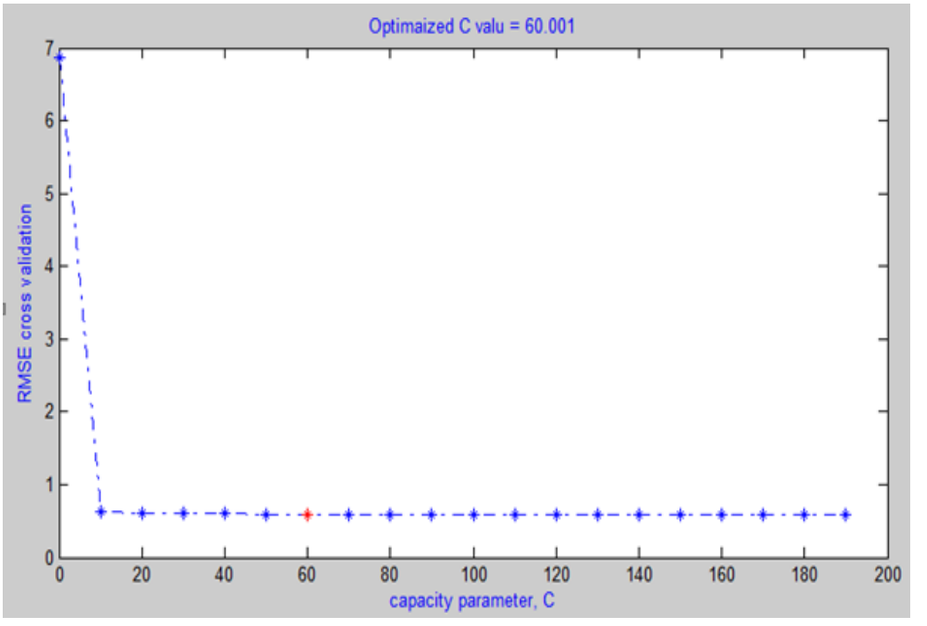
آنچه که از این شکل برآورد می شود این است که مقدار Gamma از ۰/۰۰۱ تا ۴/۵ با فواصل ۰/۵متغییر است و از نقطه ۱/۵ به بعد با افزایش Gamma اندک کاهشی در مقدار RMSE دیده نمی شود. بنابرین مقدار عددی ۱/۵تحت عنوان مقدار بهینه شده برای Gamma گزیده شد.

پارامتر (ε) epsilon یا همان فاکتور حساسیت دیگر پارامتری است که باید بهینه شود. این فاکتور به نوفه های موجود در داده ها که عمدتاً ناشناخته می باشند مربوط می شود. در شکل ۴ نمودار تغییرات RMSE بر حسب (ε) epsilon نمایش داده شده است.



شکل ٤. نمودار تغییرات مقدار RMSE برحسب مقدار Epsilon برای سری آموزش

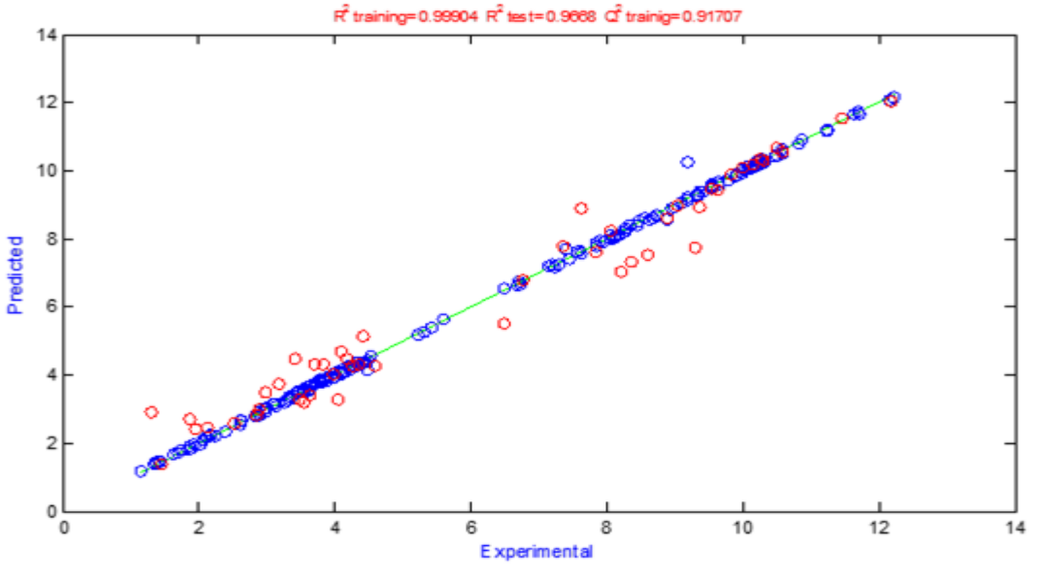
با تغییر مقدار epsilon از ۰/۰۰۱ تا ۱ با فواصل ۰/۰۵مشخص شد که کمترین مقدار RMSE در ۰/۰۵۱ مشاهده می شود بنابراین مقدار بهینه ۰/۰۵۱ برای epsilon انتخاب شد.



شکل ٥. نمودار تغییرات مقدار RMSE برحسب مقدار Capacity parameter (C) برای سری آموزش

با تغییر مقدار فاکتور ظرفیت در بازه ۰/۰۰۱ تا ۱۹۰ و با دامنه تغییرات ۱۰، مقدار بهینه ۶۰/۰۰۱ برای capacity factor برگزیده شد.

در مرحله آخر با استفاده از تمامی پارامترهای بهینه شده، مدل SVM ساخته شد و مقادیر ثابت های اسیدی ترکیبات شیمیایی(pK_a) پیش بینی شد. با استفاده از مدل SVM بهینه شده مقادیر pK_a ترکیبات مورد نظر در مجموعه آموزشی و تست مورد محاسبه قرار گرفت و در جدول۲ و شکل۶ نشان داده شده است. همانطور که در این شکل ها دیده می شود، میزان نزدیکی داده ها به خط راست قدرت پیش بینی مدل را نشان می‌دهد.



شکل٦. نمودار مقادیر پیش‌بینی شده pK_a بر حسب مقادیر تجربی برای سری آموزش و تست به روش SVM

٤. ارزیابی و مقایسه مدل های ایجاد شده

٤-١. ارزیابی مدل با استفاده از پارامترهای آماری

مطابق جدول ۳ تعداد پنج پارامتر آماری، برای ارزیابی توانایی پیش بینی مدل های ایجاد شده با روش‌های MLR, SVM مورد استفاده قرار گرفت.

جدول ۳. پارامترهای آماری برای مدل‌های انتخاب شده

		SW-MLR	SW-SVM
R^2	سری آموزش	۰/۹۲۶	۰/۹۹۰
	سری تست	۰/۹۱۱	۰/۹۶۶
RMSE	سری آموزش	۰/۸۵۷	۰/۰۹۸
	سری تست	۰/۹۴۷	۰/۵۹۳
REP	سری آموزش	۱۳/۷۱۹	۱/۵۷۰
	سری تست	۱۵/۲۰۵	۹/۵۳۰
AARD	سری آموزش	۱۶/۰۷۰	۱/۲۲۳
	سری تست	۲۲/۱۶۱	۱۰/۷۷۴

۲-٤. ارزیابی مدل‌ها توسط روش رد مرحله‌ای تک تک و گروهی

به منظور بررسی بیشتر قدرت پیش‌بینی مدل‌های خطی و غیر خطی، تکنیک رد مرحله‌ای تک تک و گروهی مورد استفاده قرار گرفت. در روش رد مرحله‌ای تک تک، هر بار یکی از ترکیبات به طور تصادفی از سری داده‌ها حذف شدند و در روش رد مرحله‌ای گروهی، هر بار یک گروه از ترکیبات(۵ ترکیب) به طور تصادفی ازسری داده‌ها حذف شدند. سپس با استفاده از مدل ساخته شده توسط بقیه ترکیبات، خاصیت شیمیایی ترکیب یا ترکیبات حذف شده، پیش‌بینی شدند. این فرایند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از رد مرحله‌ای و گروهی در جدول۳ ارائه شده است.

جدول۳. پارامترهای آماری برای مدل‌های انتخاب شده

		SW-MLR	SW-SVM
Q^2_{LOO}	کل داده ها	۰/۹۱۱	۰/۹۱۷
Q^2_{LGO}	کل داده ها	۰/۸۹۵	۰/۹۱۷

۵. نتیجه گیری

در این تحقیق از دو روش رگرسیون خطی چندگانه و ماشین بردار پشتیبان به منظور مدل سازی و پیش‌بینی ثابت های اسیدی دسته وسیعی از ترکیبات شیمیایی استفاده شد. توسط روش مرحله‌ای، دوازده توصیف‌کننده‌ای که بیشترین رابطه را با ثابت های اسیدی داشتند شامل IC1، SEigp، MATS1m ، GATS3e، GATS4e، RPCG ، MAXDN ، Mor30m ، R2e HATS2m ، n-OH ، C-040 انتخاب شدند. برای مدلسازی دو روش رگرسیون خطی چندمتغیره(MLR) و ماشین بردار پشتیبان (SVM) استفاده شدند. نتایج حاکی از رجحان مشخص روش SVM نسبت به MLR دارد لذا از مدل ایجاد شده توسط SVM می توان جهت پیش بینی ثابت های اسیدی دیگر ترکیبات شیمیایی استفاده کرد.

٦. مراجع

[1] Duchowicz, P.R., Talevi, A., Bruno-Blanch, L.E. and Castro, E.A., New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorganic & medicinal chemistry*, *16*(17) (**2008**) 7944-7955.

[2] Han, C., Yu, G., Wen, L., Zhao, D., Asumana, C. and Chen, X., Data and QSPR study for viscosity of imidazolium-based ionic liquids. *Fluid Phase Equilibria*, *300*(1-2) (**2011**) 95-104.

[3] Zhu, T., Yan, H., Singh, R.P., Wang, Y. and Cheng, H., QSPR study on the polyacrylate–water partition coefficients of hydrophobic organic compounds. *Environmental Science and Pollution Research*, *24* (**2019**) 1-11.

[4] Duchowicz, P.R., Aranda, J.F., Bacelo, D.E. and Fioressi, S.E., QSPR study of the Henry's law constant for heterogeneous compounds. *Chemical Engineering Research and Design*, *154* (**2020**) 115-121.
[5] Zhou, L., Wang, B., Jiang, J., Pan, Y. and Wang, Q., Quantitative structure-property relationship (QSPR) study for predicting gas-liquid critical temperatures of organic compounds. *Thermochimica Acta*, *655* (**2017**) 112-116.
[6] Belhassan, A., Chtita, S., Lakhlifi, T. and Bouachrine, M., QSPR study of the retention/release property of odorant molecules in water using statistical methods. *Orbital: The Electronic Journal of Chemistry*, *9*(4) (**2017**) 234-247.
[7] Chen, C.H., Tanaka, K. and Funatsu, K., Random forest approach to QSPR study of fluorescence properties combining quantum chemical descriptors and solvent conditions. *Journal of fluorescence*, *28*(2) (**2018**) 695-706.
[8] Safder, U., Nam, K., Kim, D., Shahlaei, M. and Yoo, C., Quantitative structure-property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. *Ecotoxicology and environmental safety*, *162* (**2018**) 17-28.
[9] Brusseau, M.L., The influence of molecular structure on the adsorption of PFAS to fluid-fluid interfaces: Using QSPR to predict interfacial adsorption coefficients. *Water research*, *152* (**2019**) 148-158.
[10] Petrosyan, L.S., Sizochenko, N., Leszczynski, J. and Rasulev, B., Modeling of Glass Transition Temperatures for Polymeric Coating Materials: Application of QSPR Mixture-based Approach. *Molecular informatics*, *189* (**2019**) 1154-1163.
[11] 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.
[12] 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.
[13] Zarei, K., Atabati, M. and Ebrahimi, M., Quantitative structure-property relationship study of the solvent polarity using wavelet neural networks. *Analytical Sciences*, *23*(8) (**2007**) 937-942.
[14] Todeschini, R. and Consonni, V., *Handbook of molecular descriptors* (Vol. 11). John Wiley & Sons, (**2008**).
[15] Van Aalten, D.M., Bywater, R., Findlay, J.B., Hendlich, M., Hooft, R.W. and Vriend, G., PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *Journal of computer-aided molecular design*, *10*(3) (**1996**) 255-262.
[16] Todeschini, R. and Consonni, V., *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons. (**2009**).
[17] Aziz Habibi-Yangjeh, Mohammad Danandeh-Jenagharad, and Mahdi Nooshyar, Prediction Acidity Constant of Various Benzoic Acids and Phenols in Water Using Linear and Nonlinear QSPR Models, Bull. Korean Chem. Soc., *26*(12) (**2005**) 2007-2016.

Modeling and quantitative structure-property relationship (QSPR) study to predict the acidic constants of some chemical compounds using multiple linear regression and support vector machine

S.Abbas Taheri [*], Mehdi Nekoei[*], Majid Mohammadhosseini

Department of Chemistry, Faculty of Science, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Submited: 31 December 2019, Revised: 04 May 2020, Accepted: 17 June 2020

Abstract

Modeling and studying the structure-property quantitative relationship (QSPR) to predict the acidic constants of some chemical compounds were performed using multiple linear regression (MLR) and support vector machine (SVM). First, the structure of chemical compounds was plotted and a suitable group of descriptors was calculated. Then, the step selection method was used to obtain the best descriptors that were most related to the chemical properties of the compounds. Then, linear multiple linear regression (MLR) model and nonlinear vector machine (SVM) model were used to predict the acid constants of the compounds. Statistical data showed that the SVM method was superior to the MLR method.

Keywords: *Acidic constant (pKa), multiple linear regression, Quantitative structure-property relationship, support vector machine.*

[*]**Corresponding author :** Mehdi Nekoei

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

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