به نام خدا

دانشکده: شیمی

مطالعه ارتباط کمی ساختار- خاصیت جهت پیش‌بینی اندیس بازداری بعضی از ترکیبات آلی فرار موجود در جو

دانشجو: **سارا مسگران کریمی**

استاد راهنما:

دکتر ناصر گودرزی

استاد مشاور:

دکتر منصور عرب چم‌جنگلی

پایان نامه کارشناسی ارشد جهت اخذ درجه کارشناسی ارشد

بهمن ماه ۱۳۹۱

سپاس بی‌کران پروردگار یکتا را که هستی‌مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه‌چینی از علم و معرفت را روزیمان ساخت.

تقدیم به :

پدرم به استواری کوه

مادرم به زلالی چشمه

همسرم به صمیمیت باران

دخترم به طراوت شبنم

و تقدیم به همه عزیزانی که صبورانه ادامه این راه را برایم هموار نمودند.

تشکر و قدردانی

اکنون که در پرتو الطاف الهی خداوند سبحان این مقطع تحصیلی را به پایان رساندم، وظیفه می‌دانم از زحمات خانواده عزیزم که با صبر و تحمل مشکلات مسیر را برایم تسهیل نمودند تشکر کنم.

از استاد گرامیم جناب آقای دکتر ناصر گودرزی که مساعدت و یاری‌های ایشان نقش مهمی در به ثمر رساندن این پروژه داشت بسیار سپاسگزارم. از جناب آقای دکتر منصور عرب به خاطر نظرات ارزنده به عنوان استاد مشاور نهایت قدردانی را دارم و از خداوند متعال مراتب موفقیت را برای این بزرگواران خواستارم.

چکیده

در قسمت اول این تحقیق، مطالعات ارتباط کمی ساختار- خاصیت (QSPR) بر روی شاخص بازداری ۶۰ ترکیب آلی فرار (VOCs) انجام گرفت. دو روش برازش مرحله‌ای (SR) و الگوریتم ژنتیک (GA) برای انتخاب توصیف‌کننده‌های مناسب استفاده شد. توصیف‌کننده‌های انتخاب شده از این دو روش برای مدل‌سازی و پیش‌بینی شاخص بازداری این ترکیبات وارد شبکه عصبی مصنوعی (ANN) شدند. به منظور بررسی اعتبار این مدل‌ها از روش‌های مختلفی مانند به کارگیری سری تست، رد مرحله‌ای تک‌تک داده‌ها و y- تصادفی استفاده گردید. نتایج به دست آمده بیانگر توانایی خوب هر دو روش SR-ANN و GA-ANN برای پیش‌بینی شاخص بازداری می‌باشد. ضرایب تعیین سری تست با روش‌های SR-ANN و GA-ANN به ترتیب ۰/۹۹۵ و ۰/۹۹۲ بود.

در قسمت دوم، مطالعات QSPR بر روی شاخص بازداری تعدای از ترکیبات استرول انجام شد. همانند قسمت اول، از روش‌های SR و GA برای انتخاب متغیرهای مناسب استفاده شد و توصیف-کننده‌های منتخب برای مدلسازی و پیش‌بینی شاخص بازداری این ترکیبات وارد شبکه عصبی مصنوعی (ANN) شدند. ضرایب تعیین سری تست با روش‌های SR-ANN و GA-ANN به ترتیب ۰/۹۴۴ و ۰/۹۲۰ بود. نتایج نشان می‌دهد در این قسمت نیز هر دو روش از قدرت پیش‌بینی مناسبی برخوردار بودند.

کلمات کلیدی: شاخص بازداری، SR-ANN، GA-ANN ، QSPR.

نتایج حاصل از این پایان‌نامه در پوستری با عنوان:

Prediction of retention index of atmospheric volatile organic compounds dataset using QSPR methods.

در پانزدهمین کنفرانس شیمی فیزیک ایران در پردیس علوم دانشگاه تهران ارائه گردید.

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل اول

مقدمه

۱-۱-پیشگفتار

توانایی حل مسائل برای درک مکانیسم فرایندهای مختلف شیمیایی، کشف و توسعه مواد جدید، حفظ محیط زیست و زمینه‌های دیگر شیمی، بطور کامل وجود ندارد. برای عملی کردن بعضی از مسائل نیاز به سیستم‌های بسیار پیچیده‌ای می‌باشد که انجام آن‌ها در گرو صرف هزینه‌های بسیار و مطالعات گسترده است. در جهت حل این مشکل، روش‌های محاسباتی کمومتریکس[1] می‌توانند مفید باشند. از تجزیه و تحلیل آماری و ریاضی داده‌های شیمیایی معمولاً تحت عنوان کمومتریکس یاد می-شود. بعبارت دیگر کمومتریکس یک روش کارآمد برای خلاصه کردن اطلاعات مفید از یکسری داده‌ی مشخص و پیش‌بینی سری دیگر داده‌هاست. درحقیقت هدف کمومتریکس، بهبود بخشیدن فرایندهای اندازه گیری و استخراج اطلاعات شیمیایی مفیدتر از داده‌های اندازه گیری شده فیزیکی و شیمیایی می‌باشد. چندین تعریف برای کمومتریکس بیان شده است که غالباً در متن‌های تجزیه‌ای بکار می‌روند. یکی از جامع‌ترین تعاریف بصورت زیر است:

کمومتریکس شاخه‌ای از شیمی است که از ریاضی، آمار و منطق استفاده می‌کند برای اینکه : الف) فرایندهای تجربی بهینه را طراحی و انتخاب کند. ب) حداکثر اطلاعات شیمیایی قابل حصول را از تحلیل اطلاعات فراهم کند. ج) بتوان اطلاعات بیشتری در مورد سیستم‌های شیمیایی بدست آورد. کمومتریکس بعنوان یک شاخه علمی جوان در دو دهه اخیر به سرعت توسعه پیدا کرده است. این رشد سریع مدیون پیشرفت سریع دستگاه‌های هوشمند و خودکار آزمایشگاهی و همچنین امکان استفاده از کامپیوترهای قدرتمند و نرم افزارهای ساده است. بنابراین کمومتریکس بعنوان یک وسیله در همه قسمت‌های شیمی و بیشتر در زمینه شیمی تجزیه مورد استفاده قرار گرفته است. یکی از زمینه‌های مهم کاربرد کمومتریکس در مطالعاتی است که خواص مولکول‌ها را به ویژگی‌های ساختاری آن‌ها نسبت می‌دهد. از نظر شیمی‌دانان فعالیت و خواص یک ترکیب ناشی از ویژگی‌های ساختاری آن

1-Chemometrics

است. این نوع از مطالعات به بررسی کمی ارتباط ساختار با فعالیت[1] یاQSAR و همچنین بررسی کمی ارتباط ساختار با ویژگی[2] یاQSPR معروف می‌باشد. نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختاری آن‌ها، به پژوهشگران در پیش بینی رفتار مولکول‌های جدید بر اساس رفتار مولکول‌های مشابه کمک می‌کند. به مجموعه ابزارها و روش‌هایی که به این منظور مورد استفاده قرار می‌گیرند، روش‌های پارامتری می‌گویند. در روش‌های پارامتری سعی می‌شود بین یک سری توصیف‌کننده‌های مولکولی[3] با فعالیت یا خاصیت موردنظر ارتباط منطقی برقرار شود. توصیف‌کننده‌های مولکولی که به این منظور استفاده می‌شوند، مقادیر عددی می‌باشند که جنبه‌های مختلف ساختاری مولکول را بطور کمی نشان می‌دهند.

در این پروژه از شاخه ارتباط کمی ساختار- ویژگی (QSPR) کمومتریکس برای پیش بینی شاخص‌های بازداری[4] تعدادی از ترکیبات آلی فرار[5] (VOCs) استفاده شده است. از جمله روش‌هایی که برای مطالعه ارتباط کمی ساختار- ویژگی به کار می‌روند، می‌توان به رگرسیون خطی چند-گانه[6](MLR)، رگرسیون اجزای اصلی[7](PCR)، روش حداقل مربعات جزئی[8](PLS) و شبکه عصبی مصنوعی[9] (ANN) اشاره نمود. در این پایان نامه، از دو نوع روش برازش مرحله ای[10] و الگوریتم ژنتیکی[11] برای انتخاب بهترین متغیرها استفاده شد، سپس متغیرهای انتخاب شده از این دو روش برای مطالعه ارتباط کمی ساختار- ویژگی ترکیبات موردنظر به کار گرفته شدند و مقایسه‌ای بین نتایج حاصل از دو روش انتخاب متغیر در مدل غیرخطی صورت پذیرفت. در بخش‌های سوم و چهارم تحقیق حاضر به ترتیب ارتباط بین ساختار و شاخص بازداری ۶۰ مولکول از ترکیبات آلی فرار و ۴۵

1- Quantitative Structure Activity Relationship
2- Quantitative Structure Property Relationship
3- Molecular descriptors
4-Retention Index
5- Volatile Organic Compounds
6- Multiple Linear Regression
7- Principal Components Regression
8- Partial Least Square
9- Artificial Neural Network
10- Stepwise Regression
11- Genetic Algorithm

ترکیب از استرول ها[1] مورد بررسی قرار گرفت. مدل‌های بدست آمده با روش‌های مختلف ارزیابی گردید و پارامترهای آماری مربوط به آن‌ها نیز محاسبه و گزارش شد.

۱-۱-۱- ترکیبات آلی فرار(VOCs)

این ترکیبات مواد شیمیایی آلی هستند که دارای فشار بخار زیاد در شرایط معمولی بوده و این فشار بخار زیاد ناشی از نقطه جوش پایین آن‌ها می‌باشد که باعث شده تعداد زیادی مولکول از سطح مایع یا جامد، تبخیر یا تصعید شده و وارد هوای اطراف شوند. به‌عنوان مثال فرمالدئید با نقطه جوش ۱۹ $^{\circ}c$- براحتی از سطح جدا شده و وارد محیط می‌شود. ترکیبات آلی فرار بسیار متنوع بوده و در همه جا حضور دارند و بیشتر عطرها و طعم‌ها شاملVOCها می باشند. این ترکیبات توسط بشر نیز ساخته می‌شوند و از هزاران محصول نیز منتشر می‌شوند، که می‌توان به رنگ‌ها و لاک‌ها، محصولات بهداشتی، حشره کش‌ها، مواد ساختمانی، لوازم و اثاثیه منازل، تجهیزات اداری مانند پرینترها و دستگاه‌های کپی، لاک‌های غلط‌گیر، کاغذهای کاربن، مواد صنعتی و گرافیکی شامل چسب‌ها، ماژیک‌های بادوام و محلول‌های عکاسی اشاره نمود. VOCها همچنین از سوختن بنزین، چوب، زغال یا گازهای طبیعی آزاد شده و وارد هوای اطراف می‌شوند. در طول سال‌های گذشته به اندازه گیری VOCها در هوای محیط توجه خاصی شده است، چرا که آن‌ها باعث تأثیرات نامطلوب برروی محیط و سلامتی انسان می‌شوند. گونه‌های کلردار این ترکیبات مانند تتراکلریدکربن مقاوم و پایدار بوده و کاهنده‌های ازون می‌باشند. برخی از آن‌ها نیز مانند نرمال هگزان سمی و بعضی نیز مانند بنزن سرطانزا هستند. بنابراین بدلیل این تأثیرات مضر، اندازه گیری این ترکیبات امری ضروری می‌باشد[۱].

1- Sterols

دی بلاس[1] و همکارانش در سال ۲۰۱۱ از کروماتوگرافی گازی-طیف سنجی جرمی و کروماتوگرافی گازی-یونیزاسیون شعله ای برای شناسایی و اندازه گیری ترکیبات آلی فرار موجود در اتمسفر استفاده و شاخص بازداری مربوط به آن‌ها را گزارش کردند[۱]. هدف آن‌ها بهینه سازی یک سیستم GC-MS برای اندازه گیری برخط، پیوسته، بدون ناظر و دراز مدت بیشتر از ۶۰ ترکیب VOC در یک محیط شهری بود. در تحقیق حاضر این نتایج بعنوان سری داده‌ها در فصل سوم مورد استفاده قرار گرفته و کارایی روش‌های QSPR برای پیش بینی شاخص بازداری این ترکیبات بررسی شده است.

۱-۱-۲-شاخص بازداری

شاخص بازداری معیاری کمی است که میزان بازداری نسبی هر یک از اجزای نمونه را نسبت به آلکان‌های نرمال (هیدروکربن‌ها) بر روی یک فاز ساکن و در یک دمای مشخص نشان می‌دهد. شاخص بازداری، متغیرهای مختلف سیستم کروماتوگرافی را نرمالیزه می‌کند، به گونه‌ای که امکان مقایسه نتایج سیستم‌های مختلف فراهم می‌شود. شاخص بازداری n- آلکان ها را می‌توان به صورت تعداد اتم های کربن × ۱۰۰ (به عنوان مثال، هگزان = ۶۰۰) تعریف نمود. شاخص بازداری (RI) ماده مورد تجزیه با استفاده از معادله (۱-۱) محاسبه می‌گردد که در این معادله t'_{Ra}، t'_{Rn}و t'_{Rn+1} به ترتیب عبارتند از : زمان های بازداری تنظیم شده ماده مورد تجزیه، آلکان های حاوی n و n+1 کربن که بلافاصله قبل و بعد از ماده مورد تجزیه از ستون خارج می‌شوند[۲].

$$RI = n \times 100 + 100 \frac{\log t'_{Ra} - \log t'_{Rn}}{\log t'_{Rn+1} - \log t'_{Rn}} \qquad (۱-۱)$$

شاخص بازداری زمانی مفید است که پیش بینی ترتیب خروج نسبی اجزاء نمونه از ستون مورد نظر باشد. اگر از شاخص بازداری با چنین منظوری استفاده شود، باید در نظر داشت که شاخص بازداری در

1- De Blas

یک شرایط معین و برای یک ستون خاص قابل کاربرد است. از شاخص بازداری، برای مقایسه بازداری و گزینش پذیری ستون های مشابه (ستون هایی که از لحاظ جنس و ابعاد لوله موئین، نوع و ضخامت فاز ساکن و سایر موارد کاملاً مشابه اند) نیز می‌توان استفاده کرد [۳].

۱-۲-هدف تحقیق

اندازه‌گیری شاخص‌های بازداری ترکیبات آلی به طور معمول با استفاده از تکنیک‌های مختلف کروماتوگرافی انجام می‌شود، وجود ستون‌های کروماتوگرافی مختلف برای اندازه‌گیری مواد گوناگون و پیشرفت‌هایی که از گذشته تا به حال در زمینه کروماتوگرافی پدید آمده‌اند اندازه‌گیری شاخص بازداری تعداد زیادی از ترکیبات را امکان‌پذیر کرده است و بسیاری از مسائل مربوط به جداسازی و شناسائی ترکیبات را حل نموده‌است. اما در کنار این مزایا، تکنیک‌های کروماتوگرافی، محدودیت‌هایی نیز دارند که از آن جمله می‌توان به عدم توانایی آن‌ها در اندازه‌گیری پاره‌ای از ترکیبات که تهیه و یا آماده‌سازی آن‌ها بسیار مشکل است یا ترکیباتی که در ستون‌های کروماتوگرافی ناپایدارند اشاره نمود.

هم‌چنین تکنیک‌های کروماتوگرافی نیاز به آماده‌سازی‌ها، انجام عملیات ویژه بر روی نمونه و ستون‌های کروماتوگرافی و در مواردی بهینه‌سازی پارامترها دارند که ممکن است با صرف زمان و هزینه‌های بسیار زیاد همراه باشند. بنابراین این محدودیت‌ها موجب می‌شوند که در کنار روش‌های آزمایشگاهی تجربی، روش‌های تئوری نیز برای استفاده از نتایج تجربی توسعه یابند. از جمله‌ی این روش‌ها مطالعات کمومتریکس به ویژه در زمینه‌ی QSPR می‌باشد. همانگونه که قبلا ذکر شد در این روش می‌توان از تکنیک‌های مختلفی برای ایجاد روابط خطی و غیرخطی میان خواص ساختاری و شاخص بازداری تجربی ترکیبات استفاده نمود. تدبیر اساسی QSPR به دست آوردن رابطه‌ی کمی بهینه‌ای از ساختار مولکولی برای پیش‌بینی خواص آن‌ها می‌باشد و هنگامی که رابطه‌ی معتبری به دست آمد، امکان پیش‌بینی شاخص‌های بازداری برای ساختارهای مشابه با ترکیبات اندازه‌گیری شده یا ساختارهای دیگر که هنوز اندازه‌گیری نشده وجود دارد[۴].

همچنین، با توجه به تأثیرات نامطلوب و مضر VOCها بر روی محیط و سلامتی انسان، ضرورت اندازه‌گیری این ترکیبات در هوای محیط شهری امری اجتناب ناپذیر می‌باشد. یکی از راه‌های کسب اطلاعات کیفی و کمی راجع به این ترکیبات اندازه گیری شاخص بازداری آن‌ها است که با روش‌های کروماتوگرافی قابل اندازه‌گیری می‌باشد. شاخص بازداری مستقل از متغیرهای محیطی نظیر دما، سرعت جریان فاز متحرک و ... می‌باشد. بنابراین پارامتر مناسبی برای کار کیفی می‌باشد که اطلاعات با ارزشی درخصوص چگونگی بازداری ترکیبات آلی با ساختارهای مختلف بدست می‌دهد. از طرفی دستیابی به تکنیک‌های اسپکتروسکوپی معتبر که دارای حد تشخیص مناسب و حساسیت بالا بوده براحتی امکانپذیر نمی‌باشد. چرا که اندازه گیری VOCها در هوای محیط بصورت برخط (برای مکش و جذب مستقیم هوا) و در دراز مدت می‌باشد[۱]. بنابراین بدیهی است که اندازه گیری‌ها وقف‌گیر و مستلزم صرف هزینه زیاد می‌باشد. در این تحقیق، سعی شده مدلی برای ایجاد رابطه کمی بین ساختارهای مولکولی و شاخص بازداری تعدادی از ترکیبات آلی فرار (VOCs) با به کارگیری روش های QSPR ارائه شود. سپس با استفاده از این مدل، امکان پیش بینی شاخص بازداری برای ترکیباتی با ساختار مشابه وجود خواهد داشت.

۱-۳-پیشینه تحقیق

ارتباطات کمی ساختار- خاصیت تاکنون در بسیاری موارد برای پیش‌بینی چگونگی رفتار بازداری ترکیبات در کروماتوگرافی و توضیح مکانیسم‌های جداسازی استفاده شده است. در ادامه به بررسی نمونه‌هایی از کارهای انجام شده می‌پردازیم.

در سال ۲۰۱۰، قوامی و همکارانش ۱۳۲ ترکیب از ترکیبات آلی فرار (VOCS) را با استفاده از روش ارتباط کمی ساختار- بازداری[1]، برای پیش‌بینی شاخص‌های بازداری بر روی ۱۲ فاز ساکن

1- Quantitative Structure-Retention Relationship(QSRR)

متفاوت درGC مورد بررسی قرار دادند .در واقع مدل QSRR پیشنهادی، RRIs[1] ترکیبات آلی را پیش‌بینی کرد. از رگرسیون مرحله‌ای (SR)[2] برای انتخاب بهترین متغیرها استفاده شد. در این تحقیق از روش MLR برای ساخت مدل استفاده شد که ضریب همبستگی[3] بر روی فازهای ساکن متفاوت در محدوده ۰/۹۶۷۳- ۰/۹۳۷۸ قرار گرفت و ضریب همبستگی برای رد مرحله‌ای تک‌تک داده‌ها در محدوده ۰/۹۳۲۵- ۰/۹۶۵۳بود و صحت مدل همچنین توسط روش Y–تصادفی تأیید شد[۵].

در سال ۲۰۱۲ سرخوش و همکارانش به بررسی مطالعه QSPR ۸۵ ترکیب آلی فرار بعنوان آلوده کننده‌های هوای محیط برای پیش بینی زمان های بازداری این ترکیبات پرداختند. از رگرسیون مرحله‌ای (SR) برای انتخاب بهترین متغیرها استفاده شد. بعد از انتخاب متغیر روش‌های ANN و MLR برای مدلسازی استفاده شدند. نتایج پیش بینی توافق خوبی با مقادیر تجربی داشتند[۶].

سری داده‌های مورد بررسی در تحقیق حاضر از اندازه‌گیری شاخص بازداری ترکیبات آلی فرار، در کروماتوگرافی گازی بر روی ستون‌های PLOT و BPI به دست آمده که بر اساس پیشینه تحقیق به دست آمده، مطالعه QSPR بر روی سری داده‌هایی که از کروماتوگرافی با این مشخصات ستون به دست آمده، انجام نشده است. بنابراین در این تحقیق سعی شده مطالعات کمی ارتباط ساختار-خاصیت را بر روی این سری داده‌ها انجام داده و مدل‌های SR-ANN و GA-ANN جهت پیش‌بینی شاخص بازداری این ترکیبات ارائه شود.

در بخش دوم این پایان‌نامه همانند قسمت اول مدل‌سازی با روش‌های SR-ANN و GA-ANN به منظور پیش‌بینی شاخص بازداری برخی از ترکیبات استرول ارائه شده است.

این ترکیبات دسته‌ی مهمی از ترکیبات آلی می‌باشند. که به طور طبیعی در گیاهان، جانوران و قارچ‌ها یافت می‌شوند. مهمترین نوع جانوری آن کلسترول بوده که یک عامل حیاتی در سلول‌ها می-

1-Relative Retention Indices
2-Stepwise Regression
3- Correlation coefficient

باشد و عاملی پیشرو برای ویتامین‌های محلول در چربی و هورمون‌های استروئیدی می‌باشد. بنابراین با توجه به اهمیت این ترکیبات در زندگی موجودات زنده اندازه‌گیری، شناسایی و جداسازی آن‌ها امری مهم بوده که با اندازه‌گیری شاخص بازداری آن‌ها در کروماتوگرافی این امر امکانپذیر می‌باشد. از طرفی اندازه‌گیری‌ها در کروماتوگرافی مستلزم صرف هزینه و وقت بسیار می‌باشد، بنابراین در این تحقیق، سعی شده مدلی برای ایجاد رابطه کمی بین ساختارهای مولکولی و شاخص بازداری تعدادی از ترکیبات استرول با به کارگیری روش های QSPR ارائه شود. سپس با استفاده از این مدل، امکان پیش بینی شاخص بازداری برای ترکیباتی با ساختار مشابه وجود خواهد داشت. تاکنون هیچ روش تئوری برای پیش‌بینی شاخص بازداری این دسته ترکیبات گزارش نشده است.

فصل دوم

کمومتریکس

۲-۱-کمومتریکس

کمومتریکس در حقیقت کاربرد علوم آمار، کامپیوتر و ریاضی در شیمی می‌باشد. از روش های ذکر شده برای درک بهتر اطلاعات شیمیایی که در آزمایشگاه بدست می‌آید استفاده می‌شود. در واقع با استفاده از آنالیز داده های شیمیایی بدست آمده اطلاعات مفید استخراج شده با توجه به این اطلاعات می‌توان آزمایش های موردنظر با بازدهی بهتر را طراحی کرد. کاربرد روش های ریاضی در شیمی سابقه دیرین دارد ولی با توجه به پیشرفت علوم کامپیوتر و کاربرد آن در علوم روش‌های کمومتریکس در دهه اخیر پیشرفت بسیار داشته است. در این دو دهه روش‌های کمومتریکس مختلفی توسط شیمیدان ها با کمک متخصصین علوم کامپیوتر، ریاضی و آمار ارائه شده است. عنوان کمومتریکس اولین بار توسط ولد[۱] در اوایل سال ۱۹۷۰ میلادی مطرح گردید و در سال ۱۹۷۴ با همکاری شیمیدان آمریکایی به نام کوالسکی[۲] انجمن بین المللی کمومتریکس[۳] تأسیس گشت[۷]. کارایی کمومتریکس بسیار زیاد است و شامل گستره ای از شیمی فیزیک (مطالعات سینتیک و تعادل) تا شیمی آلی (بهینه سازی واکنش ها و ارتباط ساختار با فعالیت) و شیمی نظری است. این علم همچنین دارای کاربردهای زیادی از جمله نظارت محیطی[۴]، باستان شناسی علمی[۵]، بیولوژی[۶]، علوم جنایی[۷]، نظارت فرایندهای صنعتی[۸]، شیمی خاک[۹] و غیره می‌باشد. اما از بین گروه‌های علاقمند به کمومتریکس، شیمیدانان تجزیه استفاده کنندگان اصلی آن می‌باشند[۸].

۱- S.Wold

۲ - Kowalski

۳- International Chemometrics Society (ICS)

4- Environmental monitoring

5- Scientific archaeology

6- Biology

7- Forensic science

8- Industrial process monitoring

9 - Geochemistry

۲-۲- ارتباط کمی ساختار- خاصیت(QSPR)

مطالعات مربوط به یافتن ارتباط کمی بین ساختار و فعالیت و یا ویژگی (QSPR/QSAR)، شاخه‌ای از علم کمومتریکس بنام کمومتریکس مولکولی است که اولین بار توسط هانش[۱]، فری[۲] و همکارانش شناخته شد. اساس روش های QSAR یا QSPR برقراری ارتباط بین فعالیت یا خاصیت مشاهده شده و ویژگی‌های ساختاری مولکول است که هدف آن پیش‌بینی فعالیت یا خواص بیولوژیکی، فیزیکی و شیمیایی یک ماده بر اساس ساختار شیمیایی آن می‌باشد. با این روش مکانیسم عمل ترکیبات در ابعاد مولکولی و پدیده های فیزیکوشیمیایی آن‌ها را نیز می‌توان تفسیر کرد[۹]. همانطور که گفته شد هرگاه مطالعات به صورت ارتباط بین ساختار مولکولی و خواص مشاهده شده مولکول انجام گیرد، به آن ارتباط کمی ساختار – خاصیت یا QSPR می‌گویند. در این روش خاصیت مورد نظر با پارامترهایی بر پایه خواص مولکولی که توصیف‌کننده‌های مولکولی نامیده می‌شود ارتباط داده می‌شود[۱۳-۱۰]. توصیف‌کننده‌های مولکولی پارامترهای نظری هستند که به ساختار مولکول ارتباط دارند و برای هر مولکول مقادیر خاصی را به خود اختصاص می‌دهند. این توصیف‌کننده‌ها از روی ساختارهای مولکولی و با استفاده از الگوریتم های شناخته شده ریاضی، نظریه گراف شیمیایی، روش مکانیک کوانتومی و غیره محاسبه می‌شوند[۱۴].

توسعه مدل در مطالعات QSPR شامل مراحل زیر می‌باشد:

۱- فراهم کردن سری داده‌ها

۲- رسم ساختار و بهینه سازی

۳- محاسبه توصیف‌کننده‌ها

۴- انتخاب بهترین توصیف‌کننده‌ها

۵- ساختن مدل

---

۱ - Hansh

2 - Free

۶- ارزیابی مدل

بطور کلی هدف QSPR یافتن روابط کمی بهینه میان ساختار و خواص مولکولی است که اگر نتایج این نوع مطالعات همبستگی قابل قبولی را ارائه کند علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختمانی آن‌ها، به پژوهشگران این امکان را می‌دهد که براساس رفتار مولکول‌های مشابه همان خواص را برای ساختارهایی را که تاکنون اندازه گیری نشده یا حتی هنوز سنتز نشده اند پیش بینی کنند[۱۶،۱۵].

۲-۲-۱-جمع‌آوری سری داده‌ها

در مرحله نخست، مجموعه ای از مقادیر تجربی از یک ویژگی خاص مانند شاخص بازداری برای تعدادی از مولکول‌ها از منابع علمی معتبر و در دسترس گردآوری می‌شود. بهتر است ترکیبات انتخابی دارای ساختارهای مشابه یا گروهی از مشتقات یک ترکیب بوده و کمیت هدف مدلسازی در شرایط عملی یکسان بدست آمده باشد تا نتایج قابل اعتمادتر و مناسبتری بدست آید. به این دسته مولکولی سری داده ها می‌گویند.

۲-۲-۲-بهینه سازی ساختار هندسی

در هر مطالعه QSPR اولین گام پس از انتخاب سری داده‌ها، رسم ساختارهای دوبعدی مولکول‌ها و تبدیل آن به ساختار سه بعدی بهینه[1] برای ایجاد توصیف‌کننده‌های مولکولی، به خصوص توصیف-کننده‌های هندسی به شکل صحیح می‌باشد. بدین منظور از نرم افزار HyperChem استفاده شده است. در این تحقیق، از روش AM1[2] برای رسم و محاسبه‌ی انرژی مولکول ها استفاده شده که از جمله روش‌های محاسباتی نیمه تجربی[3] می‌باشد[۱۸،۱۷].

1 - Optimize
2 - Austin methods
3 - Semi-empirical methods

۲-۲-۳- محاسبه توصیف‌کننده ها

روش های QSPR نیازمند توصیف کمی ساختار شیمیایی مولکول‌های موجود در سری داده ها می‌باشند. لذا برای یک مدلسازی موفق باید به روشی ساختارهای مولکولی را کدگذاری نمود. توصیف کننده‌ها در واقع ویژگی‌های ساختاری و الکترونی مولکول‌ها را بصورت کمی و عددی بیان می‌کنند. به عبارت دیگر می‌توان گفت که هر توصیف‌کننده، اطلاعات خاصی از مولکول را که بر کمیت مورد مدلسازی اثر می‌گذارد را در اختیار قرار می‌دهد و از اهمیت به سزایی برخوردار است. [۱۹]. در این تحقیق برای محاسبه توصیف‌کننده ها از نرم افزار Dragon استفاده شده است.

یک توصیف کننده مولکولی مناسب باید شامل ویژگی‌های زیر باشد[۲۰]:

- عدم پیچیدگی و ساده بودن
- عدم وابستگی به سایر توصیف‌کننده ها و مستقل بودن
- دارا بودن بیشترین وابستگی با خاصیت تجربی مورد نظر
- توانایی تفسیر ساختار ترکیبات
- قابلیت تمایز بین ایزومرهای مختلف مولکول

۲-۲-۴- انتخاب بهترین توصیف‌کننده‌ها

انتخاب متغیر (توصیف کننده) در روش های کمومتریکس، دارای اهمیت می‌باشد. قبل از مرحله انتخاب متغیر، فرایند کاهش متغیر[۱] صورت می‌گیرد. زیاد بودن تعداد توصیف‌کننده‌ها و همچنین همبستگی بالای برخی از آن‌ها به یکدیگر، باعث طولانی و دشوار شدن فرایند مدلسازی می‌شود. بنابراین با انتخاب زیرمجموعه‌ای از متغیرهای مستقل[۲] که اطلاعات اساسی مربوط به کل سری داده‌ها

1 - Variable reduction

۲ - Independent variable(s)

را دربردارند، متغیرهای با همبستگی بالا حذف می‌شوند. در حین انجام فرایند کاهش متغیر باید به نکات زیر توجه شود:

- توصیف‌کننده های با بیش از ۹۰٪ مقادیر یکسان حذف می‌شوند.
- از بین توصیف‌کننده هایی که با هم همبستگی بالای ۰/۹ دارند توصیف‌کننده‌ای که با متغیر وابسته همبستگی کمتری دارد، حذف می‌گردد.
- توصیف‌کننده هایی که محاسبه و تفسیر آن‌ها مشکل باشد، حذف می‌شوند.

سپس از فرایند انتخاب متغیر[1] استفاده می‌شود که طی آن، از بین توصیف‌کننده ها، آن‌هایی که با متغیر وابسته[2] همبستگی بیشتری دارند انتخاب می‌گردند. یکی از روش های انتخاب متغیر، روش برازش مرحله ای است که از طریق محاسبه ضریب همبستگی بین توصیف‌کننده و خاصیت تجربی انجام می‌شود (ضریب همبستگی، بیان کننده میزان وابسته بودن پارامتر تجربی به متغیر مستقل مورد نظر به صورت کمی است). اگر توصیف کننده‌ای دارای ضریب همبستگی برابر یک باشد، قادر است خواص مورد نظر را به درستی توصیف نموده و ضریب همبستگی صفر بدین معنی است که هیچ ارتباطی بین توصیف‌کننده و پارامتر تجربی وجود ندارد. پس توصیف‌کننده‌هایی که دارای ضرایب همبستگی هستند، در برازش مرحله‌ای برای ساخت مدل انتخاب می‌شوند[۲۱]. از دیگر روش های انتخاب متغیر می‌توان به الگوریتم ژنتیک و روش جایگزینی[3] اشاره کرد. که در این تحقیق از هر دو روش برازش مرحله‌ای و الگوریتم ژنتیک برای انتخاب متغیر استفاده شده است.

1 - Variable selection
۲ - Dependent variable
3 - Replacement method

۲-۳- روش برازش مرحله‌ای

در این روش متغیرهای مستقل به صورت مرحله‌ای انتخاب شده که در ابتدا، متغیری که بیشترین همبستگی را با متغیر وابسته داشته باشد انتخاب می‌شود. سپس در مرحله بعد، متغیر دوم با بیشترین همبستگی از بین متغیرهای باقیمانده انتخاب و به همین ترتیب ادامه می‌یابد.

۲-۴- الگوریتم ژنتیک

الگوریتم ژنتیک را می‌توان به طور ساده، یک روش جستجوگر نامید که بر پایه مشاهدات خصوصیات فرزندان نسل های متوالی، و انتخاب فرزندان بر اساس اصل بقای بهترین[۱] پایه ریزی شده است. الگوریتم ژنتیک برروی فرزندان یک نسل (از جواب های مسأله در یک مرحله)، از قوانین موجود در علم ژنتیک تقلید کرده و با به کاربردن آن‌ها، به تولید فرزندان با خصوصیات بهتر (جواب‌های نزدیک تر به هدف مسأله) می‌پردازد. در هر نسل به کمک فرآیند انتخابی متناسب با ارزش جواب‌ها و تولید مثل فرزندان (جواب های) انتخاب شده، تقریب‌های بهتری از جواب نهایی به دست می‌آید. این فرایند باعث می‌شود که نسل‌های جدید با شرایط مسأله سازگارتر باشند. این رقابت میان ژن‌ها و پیروز شدن ژن غالب (انتخاب شدن توسط الگوریتم برای تولید مثل بعدی) و کنار رفتن ژن‌های مغلوب (جواب‌های دور از هدف مسأله) روش کارآمدی برای حل مسائل پیچیده و دشوار می-باشد[۲۲]. بیشتر الگوریتم های ژنتیک تغیر یافته الگوریتم ژنتیک ماده هستند که توسط گولبرگ در سال ۱۹۸۹ پیشنهاد گردید. این الگوریتم دارای سه عملگر اصلی انتخاب، تکثیر و جهش می‌باشد. الگوریتم های ژنتیک به صورت کلی زیر مجموعه الگوریتم های تکاملی به حساب می‌آیند. امروزه الگوریتم ژنتیک شناخته شده‌ترین نوع الگوریتم تکاملی در این بین می‌باشد، چرا که الگوریتم‌های

۱- Principle of survival of the fittest

ژنتیکی، امروزه به دقت مورد نیاز دست یافته‌اند. همچنین، این الگوریتم‌ها با توانایی‌های فراوانی که دارا می‌باشند، توانسته‌اند در محدوده‌ی علوم مهندسی، به طور بسیار موفقی پاسخگو باشند[۲۳].

الگوریتم های ژنتیکی، تکنیک های تصادفی هستند که براساس مکانیزم انتخاب طبیعی پی ریزی شده‌اند. آن‌ها برای شروع کار، یک جمعیت از فضای جواب را انتخاب کرده و با توجه به تابع هدف شروع به جستجو می‌نمایند. جواب مسأله به شکل کروموزوم، که خود شامل یک رشته از اعداد و نمادها می‌باشند، نمایش داده می‌شود. دو عامل تقاطع و جهش در جهت رسیدن به جواب بهینه در الگوریتم مداخله می‌کنند. این دو عامل با توجه به ارزش‌گذاری کروموزوم‌ها سعی می‌کنند بهترین‌ها را نسل به نسل تبدیل دهند، تا جواب بهینه به دست آید. در بعضی مسائل، به دلیل ماهیت مسأله، فقط یکی از دو عامل تقاطع یا جهش را باید استفاده نمود. کدگذاری، از کلیدی ترین مفاهیم الگوریتم ژنتیک است که نسبت به دیگر عوامل اجرایی یا طراحی الگوریتم، باید دقت بیشتری صرف اجرای آن گردد[۲۴].

۲-۴-۱- تاریخچه الگوریتم ژنتیک

جان هلند[۱] در سال ۱۹۷۵ توانست ایده الگوبرداری از تکامل جانداران را در حل مسائل ریاضی عملی کند و کتابی در باره‌ی الگوریتم ژنتیک منتشر کند. سپس گلدبرگ[۲] در سال ۱۹۸۹ کتاب معروف خود درباره بهینه‌یابی با الگوریتم ژنتیک را به رشته تحریر در آورد. در سال ۱۹۹۲ جان کوزا[۳] الگوریتم ژنتیک را برای تولید برنامه‌هایی برای انجام کارهای خاص بکار برد و روش خود را برنامه نویسی ژنتیک[۴] نامید.

۱- John Holand

۲ - Goldberg

۳- John Koza

۴ - Genetic programming

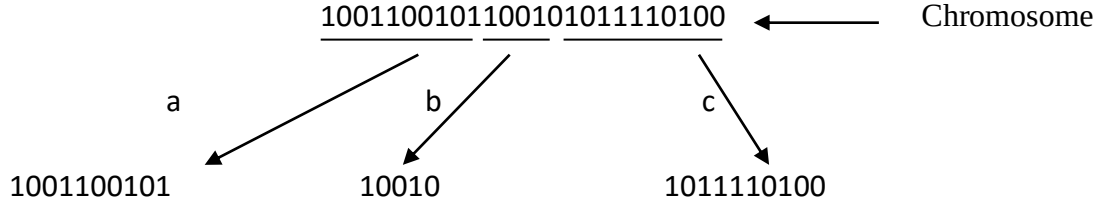
۲-۴-۲- ساختار الگوریتم ژنتیک

همانطور که گفته شد، سازوکار الگوریتم ژنتیک بر مبنای قوانین موجود در علم ژنتیک است..

در این بخش با برخی مفاهیم و تعاریف ژنتیک آشنا می شویم[۲۴].

۲-۴-۲-۱- افراد[۱] یا کروموزوم‌ها[۲]

هر کدام از افراد جمعیت، که تقریب‌هایی از جواب نهایی‌اند، به صورت رشته‌هایی از حروف یا ارقام، کدگذاری می شوند. این رشته ها را کروموزوم می نامند. متداولترین حالت، نمایش با ارقام صفر و یک است. حالت های دیگر مثل استفاده از اعداد حقیقی و اعداد صحیح هم مورد استفاده قرار می گیرند. برای مثال، یک کروموزوم با سه متغیر a, b و c با ساختار شکل (۲-۱) نمایش داده شود که هر کدام از این متغیرها، می توانند معرف خصوصیتی مستقل باشد. متغیر a با ۱۰ خانه اول سمت چپ و b با ۵ خانه بعدی و متغیر c با ۱۰ خانه باقیمانده، نمایش داده شده است. مقادیر موجود برروی کروموزوم‌ها به تنهایی معنی خاصی ندارند، بلکه باید ترجمه شوند، یعنی از حالت کد شده خارج شوند تا به عنوان متغیرهای تصمیم‌گیری دارای معنی و نتیجه باشند. باید توجه داشت که فرایند جستجو، برروی اطلاعات کد شده انجام می گیرد، مگر در صورتی که از ژن هایی با مقادیر حقیقی استفاده شود. بعد از این که کروموزوم‌ها از حالت کدگذاری شده خارج شدند، می توان به کمک اطلاعات اولیه ای که توسط مقادیر هدف مسأله به دست می‌آید، آن‌ها را مورد بررسی قرار داد.



۱ - Individuals

۲ - Chromosome

شکل (۲-۱)- نمایش یک کروموزوم با ارقام صفر و یک با سه متغیر(ژن)

۲-۴-۲-۲- جمعیت[۱]

الگوریتم های ژنتیک برروی مجموعه‌ای از جواب‌ها که جمعیت نامیده می شوند، کار خواهد کرد. افراد یک جمعیت، رشته‌هایی از اعداد هستند که کروموزوم نامیده می شوند و حاوی اطلاعات کد شده پارامترهای تصمیم‌گیری‌اند. به صورت معمول جمعیت شامل ۳۰ تا ۱۰۰ فرد است که در بعضی مسائل، این تعداد تا حدود ۱۰ فرد هم به کار رفته است..

۲-۴-۲-۳- نمایش و کدگذاری[۲] کروموزوم‌ها

تاکنون در حل مسائل بهینه‌سازی روش‌های مختلفی برای نمایش پارامترها و اطلاعات مسأله (کروموزوم‌ها) به کار برده شده است که انتخاب هر کدام از این روش ها باید با توجه به نوع مسأله و فضای جستجوی مورد نیاز برای حل و بهینه سازی آن صورت پذیرد. به طور مثال می توان سیستم کدگذاری به صورت رشته‌ای، آرایه، لیست و یا درخت صورت گیرد که در این میان، کدگذاری رشته‌ای، به دلیل قابلیت ایجاد تنوع کروموزوم‌های بیش تر در فضای کم‌تر، از کاربرد بیش‌تری نسبت به سایر روش‌ها برخوردار است.

۲-۴-۲-۴- تعیین جمعیت اولیه

بعد از تصمیم‌گیری در مورد شیوه کدگذاری کروموزوم‌ها، جمعیت اولیه باید اتخاذ گردد. این مرحله معمولا با انتخاب تصادفی مقادیر در محدوده مجاز صورت می گیرد.

---

۱ - Population

۲ - Encoding

۲-۴-۳- روش‌های انتخاب کروموزوم‌ها

روش‌های مختلفی برای انتخاب کروموزوم‌ها وجود دارند، که در زیر به معمولترین آن‌ها اشاره می‌شود:

۱- انتخاب نخبه‌گرا یا ممتازگرا[۱] : در این روش مناسب‌ترین عضو هر اجتماع انتخاب می‌شود.

۲- انتخاب چرخ رولت[۲] : یک روش انتخاب است که در آن هر عنصری که عدد برازش (تناسب) بیشتری داشته باشد، شانس انتخاب بیشتری دارد.

۳- انتخاب مقیاسی[۳] : در این روش به موازات افزایش متوسط عدد برازش جامعه، سنگینی انتخاب هم بیشتر می‌شود. این روش وقتی کاربرد دارد که مجموعه دارای عناصری باشد که عدد برازش بزرگی داشته باشد و فقط تفاوت‌های کوچکی آن‌ها را از هم تفکیک کند.

۴- انتخاب مسابقه ای[۴] : روشی است که در آن یک زیرمجموعه از صفات یک جامعه انتخاب می‌شوند و اعضای آن مجموعه با هم رقابت می‌کنند و سرانجام فقط یک صفت از هر زیر گروه برای ادغام، انتخاب می‌شوند[۱۹].

۲-۴-۴- روش های ادغام کروموزوم ها

وقتی با روش های انتخاب، کروموزوم‌ها انتخاب شدند، باید به طور تصادفی برای افزایش تناسبشان اصلاح شوند. دو راه حل اساسی برای این کار وجود داردکه اولین و ساده ترین آن‌ها جهش[۵] نام دارد

---

۱ - Elitist

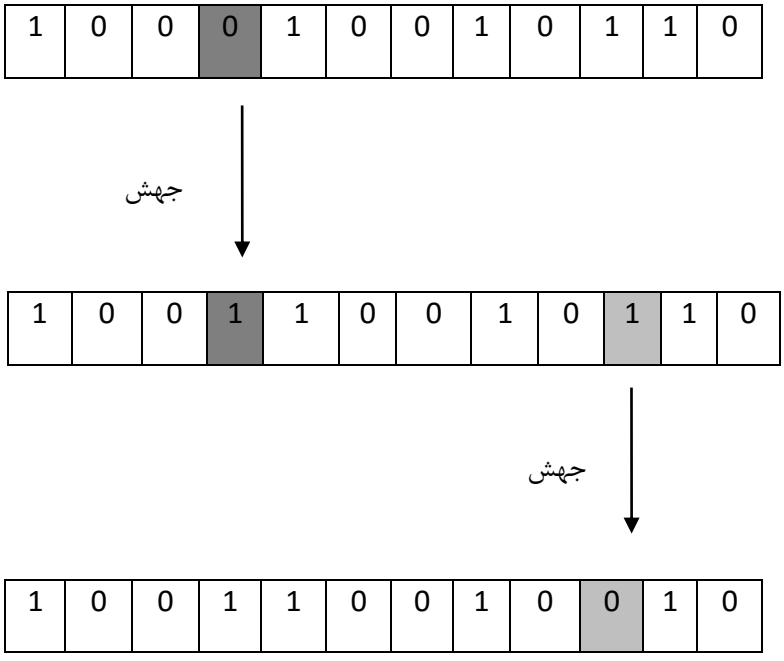
2-Roulette wheel

3- Scaling

4- Tournament

۵ - Mutation

که تغییر از یک ژن به دیگری می‌باشد. در الگوریتم های ژنتیک، جهش، تغیر کوچکی در یک نقطه از کد خصوصیاتی ایجاد می‌کند[۱۹]. در نمایش دودویی[۱] رشته‌ها، جهش، به معنای تغییر مقدار یکی از خانه‌های رشته، از صفر به یک و یا از یک به صفر می‌باشد. به عنوان مثال جهش در چهارمین خانه اولین فرزند و دهمین خانه دومین فرزند تولید شده، منجر به ایجاد شکل (۲-۲) می‌گردد. احتمال جهش در کروموزوم‌ها معمولا در حدود ۰/۰۱ تا ۰/۰۰۱ در نظر گرفته می‌شود. به کمک این عملگر می‌توان امید داشت که کروموزوم‌های خوبی که در مراحل انتخاب و یا تکثیر حذف شده اند، دوباره احیا شوند. این عملگر همچنین تضمین می‌کند که بدون توجه به پراکندگی جمعیت اولیه، احتمال جستجوی هر نقطه از فضای مسأله هیچ گاه صفر نشود[۲۵]. شکل (۲-۲) نحوه عملگر جهش در یک سیستم دودویی را نشان می‌دهد.



شکل (۲-۲) – نحوه عملکرد عملگر جهش در سیستم دودویی

6- Binary

دومین روش تغییر، تقاطع[1] نام دارد. در این روش دو کروموزوم برای معاوضه بخش‌های کد شده انتخاب می‌شوند. که اغلب شامل تقاطع تک‌نقطه ای هستند که نقطه تعویض در محلی تصادفی بین کروموزوم‌ها است[۱۹].

۲-۴-۵- دستور خاتمه الگوریتم ژنتیک

از آنجا که الگوریتم ژنتیک یک روش جستجوی تصادفی است، ارائه فرمول خاصی برای پایان آن مشکل است. ممکن است برازش جمعیت برای تعدادی از نسل ها، قبل از یافتن جواب نهایی، ثابت باقی مانده و باعث ایجاد این تصور شود که به جواب نهایی رسیده‌ایم. یک شیوه معمول پایان دادن به الگوریتم، متوقف کردن آن بعد از تولید تعداد نسل مشخص است. از آنجاییکه در بعضی عملگرها نیاز است که تعداد کل نسل‌ها را بدانیم، این شیوه مناسب به نظر می‌رسد. بعد از اتمام تکرار الگوریتم تا تعداد نسلی که به عنوان ورودی به الگوریتم داده‌ایم، کیفیت جواب‌های نهایی بررسی می‌شود که در صورت ارضاء نشدن جواب های مدنظر، الگوریتم تا تعداد نسل مشخص دیگری ادامه پیدا کرده و یا از ابتدا و با چینش جمعیت اولیه متفاوتی، اجرا می‌شود[۲۴].

۲-۴-۶- الگوریتم ژنتیک در یک نگاه

الگوریتم ژنتیک را می توان یک روش جستجوی کلی نامید که از قوانین تکامل بیولوژیک طبیعی تقلید می‌کند. الگوریتم ژنتیک برروی یک سری از جواب‌های مسأله به امید به دست آوردن جواب‌های بهتر، قانون بقای بهترین را اعمال می‌کند. در هر نسل به کمک فرایند انتخابی متناسب با ارزش جواب‌ها و تولید مثل جواب‌های انتخاب شده به کمک عملگرهایی که از ژنتیک طبیعی تقلید شده اند تقریب های بهتری از جواب نهایی به دست می‌آید. این فرایند باعث می شود که نسل های جدید با

۱ - Crossover

شرایط مسأله سازگارتر باشند. در هنگام تکثیر به کمک اطلاعات اولیه‌ای که از توابع هدف به دست می‌آید، برازش هر فرد مشخص می‌گردد و از این مقادیر در فرایند انتخاب استفاده شده تا آن را به سمت انتخاب مناسب سوق دهد. هر چه برازش فرد نسبت به جمعیت بالاتر باشد احتمال بیشتری دارد که انتخاب شود و هر چه برازش نسبی آن کم‌تر باشد، احتمال انتخاب آن برای تولید نسل بعدی کم‌تر می شود. وقتی که برازش تمام افراد جمعیت مشخص شد، هر کدام با احتمالی که متناسب با میزان برازش آن‌هاست، می توانند برای تولید نسل بعد انتخاب شوند. عمل تکثیر در الگوریتم برای تبادل اطلاعات ژنتیکی بین یک جفت و یا تعداد بیش‌تری از افراد و به کمک عملگرهای موجود در ژنتیک انجام می‌شود. معیار پایان الگوریتم ژنتیک می‌تواند رسیدن جواب‌ها به حد انتظار و یا تولید تعداد مشخصی از نسل‌ها تعریف شود[۲۴].

معمولا جهت ارزیابی یک کروموزوم از لحاظ درجه اهمیت یک تابعی به نام تابع هدف یا تابع ارزیابی تعریف شده و میزان سازگاری براساس این تابع محاسبه می‌شود. در بسیاری از موارد این تابع ریشه میانگین مربع خطا[۱](RMSE) محاسبه می‌شود. معمولا در نسل اول بعید است که کروموزوم-های انتخاب شده بهترین باشند زیرا کروموزوم‌ها به صورت تصادفی انتخاب شده‌اند. بنابراین لازم است که تغییراتی برروی کروموزوم‌ها صورت گیرد، بدین معنی که نسل‌های جدیدی از کروموزوم‌ها ایجاد شود که بتوانند پاسخ‌های مطلوب تری را ایجاد نموده و خطای کمتری داشته باشند.

۲-۴-۷ - الگوریتم ژنتیک و QSPR

توصیف‌کننده‌های انتخاب شده توسط الگوریتم ژنتیک، می‌توانند برای مدلسازی با روش‌هایی همچون رگرسیون خطی چندگانه، حداقل مربعات جزئی و شبکه عصبی استفاده شوند. در این تحقیق، متغیرهای انتخاب شده، برای مدلسازی بعنوان ورودی به شبکه عصبی داده شدند. که می‌توان آن را با نماد GA-ANN نشان داد.

۱- Root mean square errors

۲-۵- انتخاب مدل مناسب

یک مدل بر اساس روش‌های آماری گوناگون که از قواعد ریاضی بهره می‌برند، تشکیل شده و بیان‌گر ارتباط میان متغیر وابسته[۱] و متغیر (های) مستقل[۲] می‌باشد. متغیرهای مستقل، آن‌هایی هستند که سهمی در تابعی دارند که برای مدلسازی مناسب است. متغیرهای وابسته نیز متغیرهایی هستند که یافتن وابستگی آماری میان آن‌ها و تعدادی از متغیرهای مستقل مورد نظر بوده و معمولا از اندازه گیری‌های تجربی به دست می‌آیند. هدف از فرایند انتخاب متغیر که در بالا توضیح داده شد، دستیابی به یک مدل مناسب با تعدادی متغیر مستقل (توصیف‌کننده‌ها) است تا بتوان میزان متغیر وابسته را با کمک آن‌ها تخمین زد. چندین روش متفاوت برای ساختن مدل QSPR وجود دارد که از جمله می‌توان به رگرسیون خطی چندگانه (MLR) به عنوان روش خطی و شبکه عصبی مصنوعی (ANN) به عنوان یک روش غیرخطی اشاره کرد[۲۶]. در این تحقیق از روش غیرخطی شبکه عصبی برای ساخت مدل استفاده شد و متغیرهای به دست آمده ار دو روش برازش مرحله‌ای و الگوریتم ژنتیک، به طور جداگانه در ساخت مدل استفاده شدند. و در نهایت مقایسه‌ای بین دو مدل صورت گرفت.

۲-۶- شبکه عصبی مصنوعی

یک شبکه عصبی مصنوعی ایده ایست برای پردازش اطلاعات که از سیستم عصبی زیستی الهام گرفته شده و مانند مغز به پردازش اطلاعات می‌پردازد. شبکه های عصبی اطلاعات را به روشی مشابه با کاری که مغز انسان انجام می‌دهد پردازش می‌کنند. آنها از تعداد زیادی از عناصر پردازشی (سلول عصبی) که فوق العاده بهم پیوسته‌اند تشکیل شده‌اند که این عناصر به صورت موازی باهم برای حل

۱- Dependent variable

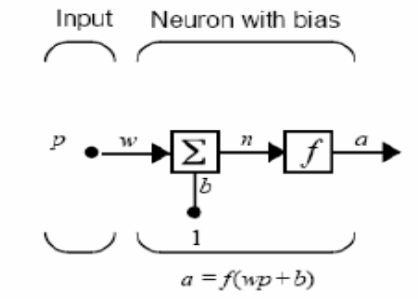
۲- Independent variable(s)

یک مسئله مشخص کار می‌کنند. استفاده از شبکه های عصبی مصنوعی به جای وش رگرسیون خطی چندگانه در مطالعات QSPR به خصوص هنگامی که ارتباط بین توصیف‌کننده و خاصیت مورد نظر خطی نبوده و یا اینکه بین آن‌ها برهمکنش‌هایی وجود داشته باشد باعث بهبود مدل حاصله خواهد شد[۲۷]. این شبکه‌ها قادر به یادگیری‌اند و قابلیت یادگیری یعنی تنظیم پارامترهای شبکه (وزن های سیناپتیکی) در مسیر زمان که محیط شبکه تغییر نموده و شبکه شرایط جدید را تجربه می‌کند، با این هدف که اگر شبکه برای یک وضعیت خاص آموزش دید و تغییر کوچکی در شرایط محیطی آن (وضعیت خاص) رخ داد، شبکه بتواند با آموزش مختصر برای شرایط جدید نیز کارآمد باشد. دیگر اینکه اطلاعات در شبکه‌های عصبی در سیناپس‌ها ذخیره می‌گردد و هر نرون در شبکه، به صورت بالقوه از کل فعالیت سایر نرون‌ها متأثر می‌شود. در نتیجه، اطلاعات از نوع مجزا از هم نبوده بلکه از کل شبکه می‌باشد. شبکه‌های عصبی مصنوعی جزء دسته ای از سیستم‌های دینامیکی قرار دارند، که با پردازش روی داده‌های تجربی، دانش یا قانون نهفته در ورای داده‌ها را به ساختار شبکه منتقل می-کنند. به همین خاطر به این سیستم‌ها هوشمند[1] گویند، چرا که براساس محاسبات روی داده‌های عددی یا مثال ها، قوانین کلی را فرا می‌گیرند[۲۸].

۲-۶-۱- سیستم عصبی مصنوعی

به تبعیت از شبکه های عصبی زیستی، می‌توان یک ساختار عصبی مصنوعی ساخت و با تنظیم مقادیر وزن اتصال، نحوه ارتباط میان اجزای شبکه، را تعیین نمود. نرون کوچکترین واحد پردازشگر اطلاعات می‌باشد، که اساس عملکرد شبکه های عصبی را تشکیل می‌دهد. عمل لایه‌ی ورودی، لایه‌ی میانی و لایه‌ی خروجی در یک شبکه عصبی به ترتیب شبیه عمل دندریت، جسم سلولی و آکسون در یک نرون زیستی می‌باشد. شکل زیر ساختار یک نرون تک ورودی را نشان می‌دهد.

1- Intelligent



شکل (۲-۳)- مدل نرون تک ورودی[۲۹]

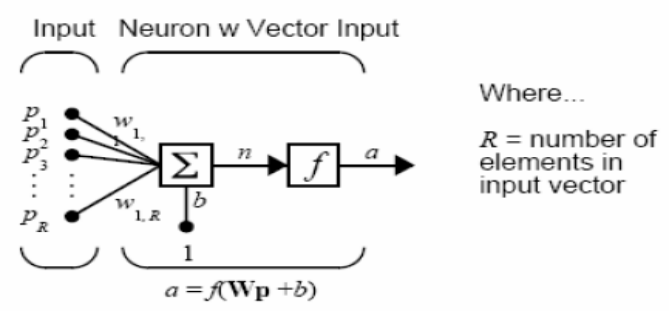
در این شکل از علامت p برای نشان دادن یک سیگنال ورودی استفاده شده و a نیز نشان دهنده‌ی خروجی نرون است. در واقع در این مدل یک سیگنال ورودی پس از تقویت (یا تضعیف) شدن با وزنw، wp را به وجود می‌آورد. میزان تأثیر ورودی p روی خروجی a، به وسیله‌ی w مشخص می-شود.ورودی دیگر که مقدار ثابت ۱ است، در جمله بایاس b ضرب شده و سپس با wp جمع می‌شود، سپس تابع تبدیل نرون f روی این مقدار یعنی ورودی خالص[۱] n اعمال می‌شود. بدین ترتیب خروجی نرون با معادله زیر تعریف می‌شود:

$$a = f\,(wp + b) \qquad (۱-۲)$$

b و w دو پارامتر تنظیم شونده در نرون‌ها می‌باشند و ایده اصلی شبکه عصبی این است که با تغییر مقادیر b و w، شبکه یک رفتار یا تصمیم را اتخاذ کند. در جعبه ابزار مورد استفاده در MATLAB، بایاس در نظر گرفته شده اما استفاده از آن اختیاری است[۲۸].

عموما یک نرون بیش از یک ورودی دارد. شکل (۲-۴) یک مدل نرون با R ورودی را ارائه می‌دهد. بردار ورودی با p نمایش داده می‌شود. اسکالرهای (p_i $(i=1,2,\ldots R$ عناصر بردار p هستند. مجموعه سیناپس‌های $w_{1,i}$، عناصر ماتریس وزن w را تشکیل می‌دهند. در این حالت w یک بردار سطری با عناصر $j=1,\ldots,R$ و $w_{1,j}$ است. هر عنصر از بردار ورودی p درعنصر متناظر w ضرب می‌شود. نرون یک جمله بایاس b داردکه با حاصلضرب ماتریس وزن w با بردار ورودی p جمع می‌شود[۲۸].

۱- Net input



شکل (۲-۴)- مدل یک نرون با چند ورودی [۲۹]

۲-۶-۲- توابع تبدیل

تابع تبدیل، وظیفه محدودسازی خروجی گره به یک بازه متناهی را بر عهده دارد. یک نرون مصنوعی را می‌توان به صورت شبکه‌ای از گره‌ها یا عناصر محاسباتی به هم پیوسته در نظر گرفت. هر گره، تبدیلی بر روی سیگنال‌های ورودی‌اش صورت داده و سیگنال خروجی تولید می‌کند که به نوبه‌ی خود به گره‌های دیگر فرستاده خواهد شد[۳۰]. انواع متفاوتی از توابع تبدیل از قبیل پله ای، تکه ای- خطی و سیگموئیدی وجود دارند که بنا به ماهیت مسأله کاربرد پیدا می‌کنند. این تابع توسط طراح مسأله انتخاب شده و بر اساس انتخاب الگوریتم یادگیری، پارامترهای مسأله (وزن ها) تنظیم می-گردند. از توابع تبدیل می‌توان به تابع تبدیل خطی[۱]، لگاریتم سیگموئید[۲] و تانژانت سیگموئید[۳] اشاره کرد. تابع تبدیل خطی همان ورودی را به عنوان خروجی برمی‌گرداند ولی تابع لگاریتم سیگموئید مقادیر ورودی را در محدوده ∞- تا ∞+ در یافت کرده و خروجی بین صفر و یک تولید می‌نماید. تابع تانژانت سیگموئید نیز با دریافت ورودی در همان محدوده، خروجی بین ۱ و ۱- تولید می‌کند[۳۱].

۲-۷- انواع شبکه‌های عصبی از نظر برگشت‌پذیری

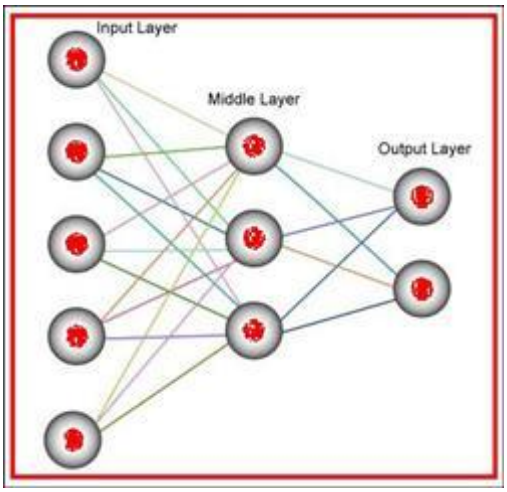
1-Linear
2- Log-sigmoid (logsig)
3- Tan-sigmoid (tansig)

- شبکه‌های پیشخور[1]
- شبکه‌های برگشتی[2]

در این تحقیق از شبکه پیشخور استفاده شده که توضیحات آن در ادامه آمده است.

۲-۸- شبکه‌های پیشخور

شبکه‌های پیشخور، شبکه‌هایی هستند که مسیر پاسخ در آن‌ها همواره رو به جلو پردازش شده و به نرون‌های همان لایه یا لایه قبل باز نمی‌گردد. در این نوع شبکه‌ها به سیگنال اجازه داده می‌شود که از مسیر یکطرفه، یعنی از ورودی تا خروجی عبور کند و در نتیجه بازخوردی[3] وجود نداشته و خروجی هر لایه تأثیری بر همان لایه و همچنین لایه‌های قبلی ندارد. شکل (۲-۴) نمونه‌ای از یک شبکه پیش خور را نشان می‌دهد.



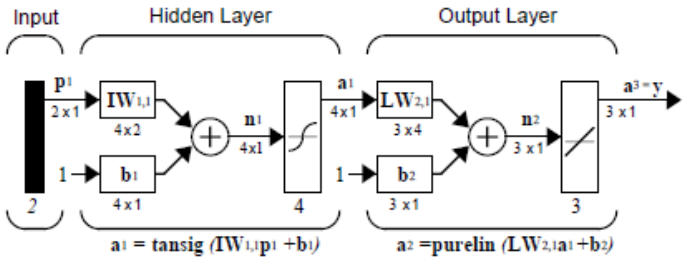
(۲-۵) – شبکه عصبی پیش خور[۳۲]

4- Feed forward
5 - Feedback
1- Recurrent

۲-۹- طراحی شبکه پیشخور

در شبکه های پیشخور، نرون‌ها در چند لایه سازمان دهی می‌شوند، بدین ترتیب که اولین و آخرین لایه به ترتیب لایه ورودی[۱] و لایه خروجی[۲] بوده، در حالی که لایه‌های بین این دو لایه را لایه‌های پنهان[۳] می‌نامند. در این نوع شبکه‌ها، ابتدا اطلاعات ورودی از محیط خارج به نرون‌های لایه ورودی ارسال می‌شود، اما در این لایه هیچ پردازشی بر روی ورودی‌ها صورت نمی‌گیرد. در واقع لایه ورودی صرفا نقش دریافت کننده و تبدیل دهنده را بر عهده دارد. خروجی این نرون‌ها به لایه بعدی که همان لایه پنهان است، منتقل شده و در نهایت خروجی شبکه از طریق نرون‌های لایه خروجی به محیط خارج ارسال می‌گردد.

طبق شکل (۲-۶)، شبکه‌های پیشخور اغلب از یک یا چند لایه مخفی از نرون‌های سیگموئیدی و از یک لایه پایانی خطی استفاده می‌کنند. شبکه چند لایه از نرون‌ها با یک تابع تبدیل غیرخطی به شبکه اجازه می‌دهد که توانایی یادگیری رابطه خطی و غیرخطی را بین ورودی‌ها و خروجی‌ها داشته باشد. لایه خروجی خطی به شبکه این امکان را می‌دهد که خروجی در هر محدوده دلخواهی از $-\infty$ تا $+\infty$ باشد. البته اگر بخواهیم خروجی در دامنه محدودی قرار گیرد، می‌توان از توابع محدودسازی مانند توابع سیگموئیدی در لایه خروجی استفاده کرد[۳۴].



شکل (۲-۶)- شبکه پیش خور[۲۹]

۱ - Input layer
2 - Output layer
3 - Hidden layer

۲-۱۰- الگوریتم‌های آموزشی شبکه عصبی

خاصیت یادگیری شبکه‌های عصبی از اهمیت ویژه‌ای برخوردار است. شبکه‌های عصبی به عنوان سیستم‌های یادگیری دارای این توانایی هستند که از گذشته، تجربه و محیط بیاموزند و رفتار خود را در حین هر یادگیری بهبود بخشند. بهبود در یادگیری در طول زمان باید بر اساس معیاری سنجیده شود. قانون یادگیری روندی است که توسط آن ماتریس وزن‌ها و بردارهای بایاس شبکه عصبی تنظیم می‌شوند. هدف قانون یادگیری آموزش شبکه عصبی جهت انجام کار مشخصی است و به عبارتی دیگر شبکه‌های عصبی در خلال آموزش پس از هر بار تکرار الگوریتم یادگیری، از محیط، شرایط و هدف کار خود بیشتر مطلع می‌گردند. نوع یادگیری هم توسط روندی که طبق آن پارامترهای شبکه تنظیم می‌گردند مشخص می‌شود[۲۸]. برای آموزش شبکه های عصبی از الگوریتم های یادگیری متفاوتی مانند الگوریتم نزول گرادیانی[۱]، الگوریتم پس انتشار گرادیان مزدوج[۲]و الگوریتم لونبرگ – مارکوارت[۳] استفاده می‌شود که همگی به عنوان زیرمجموعه‌ی یک تکنیک آموزشی به نام پس انتشار یا انتشار خطا به شمار می‌روند. انتخاب هر الگوریتم بر سرعت یادگیری و دقت شبکه مؤثر است.

پس انتشار خطا یک روش متداول آموزش با ناظر برای شبکه های پیشخور است؛ یعنی در یادگیری، برای به دست آوردن ارتباط بین متغیرهای ورودی و خروجی به الگوی آموزشی نیاز است. این روش، در شبکه‌های عصبی با بیش از یک لایه نرونی یا به عبارتی شبکه با لایه پنهان استفاده می‌شود. برای هر الگوی ورودی خاص، خروجی واقعی با خروجی مورد انتظار (هدف) مقایسه گردیده سپس اختلاف بین این دو خروجی به همه اتصال‌هایی که اولین بار برای به دست آوردن خروجی استفاده شده بودند، بازپراکنده (بازپخش) می‌شود. اگر خروجی شبکه با هدف همانندی مناسبی داشته باشد، اتصال واحدهایی که در خروجی مؤثر بوده اند تقویت می‌شود؛ اگر چنین نباشد، اتصال بین

۱- Gradient descent

2- Conjugate gradient

3- Levenberg- Marquart (LM)

واحدهای مورد نظر، تضعیف شده و بار دیگر که آن واحدهای خاص برانگیخته می‌شوند، نسبت به دفعه قبل تأثیر کمتری روی واحدهای پنهان و خروجی خواهند داشت.

۲-۱۱- بهبود تعمیم یا ارتقاء عمومیت[۱]

از جمله مشکلاتی که امکان دارد در هنگام آموزش شبکه عصبی رخ دهد این است که شبکه بیش از حد آموزش داده شود (بیش آموزی[۲] یا بیش برازش[۳]). بدین معنی که در هنگام آموزش شبکه، خطا به مقدار قابل قبولی می‌رسد ولی هنگام ارزیابی داده‌های جدید، خطای شبکه به مراتب از خطای داده های آموزشی بیشتر گردیده که گفته می‌شود شبکه تعمیم برای حالت های جدید را یاد نگرفته است. علت این امر آن است که شبکه کم کم داده‌ها را به خاطر سپرده و از پیدا کردن ارتباط موجود میان آن‌ها فاصله می‌گیرد. از جمله راه حل‌های ارائه شده در جهت ارتقاء عمومیت شبکه، به کارگیری شبکه‌های بزرگ برای ایجاد یک انطباق مناسب است و اگر شبکه‌ی مورد استفاده کوچک باشد، توانایی لازم برای انطباق داده‌ها را نخواهد داشت. متأسفانه از ابتدا نمی‌توان پیش بینی کرد که اندازه شبکه برای یک کاربرد خاص تا چه میزان باید بزرگ باشد. در جعبه ابزار شیکه عصبی برای بهبود تعمیم و جلوگیری از بیش آموزی شبکه دو راه حل تنظیم[۴] و توقف زودهنگام[۵] در نظر گرفته شده است[۳۶،۳۵].

۲-۱۱-۱- تنظیم

این عمل با اصلاح تابع کارایی متداول (MSE) از طریق افزایش یک عبارت که شامل میانگین مربعات وزن‌ها و بایاس‌های شبکه است، انجام می‌گیرد (رابطه ۲-۲).

۱- Improving generalization

۲ - Over training

3- Over fitting

4 - Regularization

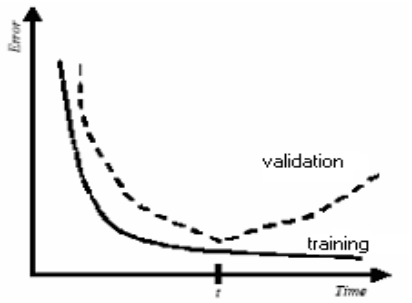
5- Early stoping

$$F=\beta E_D + \alpha E_W \qquad (۲-۲)$$

استفاده از این تابع کارایی اصلاح شده سبب می‌شود که شبکه، وزن‌ها و بایاس‌های کوچکتری داشته باشد و این امر، پاسخ شبکه را هموارتر می‌کند. در این رابطه α و β پارامترهای تابع کارایی، E_D و E_W به ترتیب میانگین مربعات خطاها و میانگین مربعات وزن‌ها می‌باشد[۳۷].

۲-۱۱-۲- توقف زود هنگام

در این تکنیک داده‌های در دسترس به سه زیر مجموعه تقسیم می‌شوند. اولین زیر مجموعه سری آموزش نام دارد و برای محاسبه گرادیان و به روز کردن وزن‌ها و بایاس‌های شبکه استفاده می‌شود. دومین زیر مجموعه به نام سری اعتبار یا ارزیابی است که خطای مربوط به آن، در طی فرایند آموزش نظارت می‌شود. خطای ارزیابی و همچنین خطای سری آموزشی،به طور طبیعی در طول فاز اولیه آموزش کاهش می‌یابند، ولی خطای سری ارزیابی همواره از خطای سری آموزش بیشتر است. زمانی که شبکه شروع به برازش اضافی داده‌ها بکند، خطای سری ارزیابی شروع به بالا رفتن می‌کند پس قبل از آن یعنی هنگامی که خطای سری ارزیابی به حداقل میزان خود رسید، آموزش متوقف شده و وزن‌ها و بایاس‌ها در مینیمم خطای اعتبار انطباق داده می‌شوند. مطابق شکل (۲-۷) روند آموزش می‌بایست در زمان t متوقف گردد.



(۲-۷)- تغییرات خطای سری آموزش و سری ارزیابی [۳۱]

زیر مجموعه سوم سری تست یا آزمایش نام دارد که در طول فرایند آموزش کاربرد نداشته و از آن برای مقایسه مدل‌های متفاوت استفاده می‌شود[۳۱].

۲-۱۲- ارزیابی مدل

برای اطمینان از اینکه مدل به دست آمده، مدل مناسبی است که توانایی پیش‌بینی نمونه‌های مختلفی از یک جمعیت را داراست، باید مدل را ارزیابی کرد. این ارزیابی از طریق شاخص‌های کمی است که به وسیله آن‌ها صحت نتایج ارائه شده توسط مدل مورد سنجش قرار می‌گیرند. برخی از این شاخص‌های کمی که پارامترهای آماری[۱] می‌باشند[۳۴]، عبارتند از:

ضریب همبستگی یا آماره R[۲]: ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر، محاسبه آماره‌ی ضریب همبستگی آنهاست. ضریب همبستگی دو متغیر x, y با رابطه (۲-۳) تعریف می‌شود.مقدار این آماره بین ۱ تا ۱- متغیر است. مقدار بزرگ‌تر آن نشان دهنده‌ی این است که ارتباط خطی بیشتری میان متغیر وابسته و متغیرهای مستقل وجود دارد.

$$R = \frac{\sum(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \qquad (۲-۳)$$

ضریب تعیین[۳] یا آماره R^2: به عنوان یک شاخص برای بیان دقت خط رگرسیون برآورد شده، به کار می‌رود و نشان‌دهنده‌ی نسبت تغییرات متغیر وابسته توضیح داده شده توسط متغیر مستقل است. رابطه ریاضی مربوط به ضریب تعیین به صورت زیر است:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{\mathrm{SST}} \qquad (۲-۴)$$

۱ - Statistical indices

2- Correlation coefficient

3- Determination coefficient

کهSSR[1] طبق رابطه (۲-۵) بیانگر مجموع مربعات انحراف مقادیر پیش‌بینی شده‌ی متغیر وابسته از میانگین مقادیر آن است.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (۵-۲)$$

SST[2] (مجموع مجذورات کل) طبق رابطه (۲-۶) نشانگر مجموع مربعات انحراف مقادیر واقعی (تجربی) متغیر وابسته از میانگین مقادیر آن است.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (۶-۲)$$

SSE[3](مجموع مجذورات خطاهای رگرسیون یا باقی مانده *ها) نیز* بیانگر مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش‌بینی شده برای آن است.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (۷-۲)$$

بنابراین با توجه به روابط فوق می‌توان نوشت:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (۸-۲)$$

طبق رابطه (۲-۸) اگر تمام مشاهدات بر روی خط برازش شده قرار گرفته باشند، یعنی به ازای تمام نقاط $y_i = \hat{y}_i$ باشد، مقدار R^2 برابر یک می‌شود و هرگونه انحرافی از این حالت باعث می‌شود که مقدار R^2 از یک کوچکتر شود.

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST} = 1 - (1 - R^2) \frac{n-1}{n-p-1} \qquad (۹-۲)$$

که در این رابطه p تعداد متغیرهای مستقل و n تعداد ترکیبات مورد بررسی می باشد.

1- Sum square regression

۲- Sum square total

3- Sum square error

مجموع مربع باقیمانده ها[1](PRESS): برابر مجموع مربعات تفاوت بین مقدار کمیت مشاهده شده(y_i) و مقدار تخمین زده شده ($\hat{y}_i$) است.

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (۱۰-۲)$$

خطای استاندارد پیش‌بینی[2](SEP)

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}} \qquad (۱۱-۲)$$

خطای مطلق میانگین[3](MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}|}{n} \qquad (۱۲-۲)$$

خطای نسبی پیش بینی[4](REP)

$$REP(\%) = \frac{100}{\overline{y}} \times \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}} \qquad (۱۳-۲)$$

میانگین مربع خطاها[5](MSE)

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n} \qquad (۱۴-۲)$$

میانگین خطای نسبی[1] (MRE)

۱-Predictive residual sum of squares
2-Standard error of prediction
3- Mean absolute error
4- Relative error of prediction
5 - Mean square error

$$MRE = \frac{\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}}{y_i}\right|}{n} \times 100 \qquad (۱۵-۲)$$

۲-۱۳- تکنیک های اعتبار سنجی

این تکنیک‌ها برای ارزیابی اعتبار مدل‌های به دست آمده و بررسی قدرت پیش بینی مدل‌ها به کار گرفته می‌شوند. به عبارت دیگر توسط این تکنیک‌ها، توانایی مدل‌ها برای انجام پیش‌بینی‌های مطمئن در موارد جدیدی که پاسخ ناشناخته دارند، سنجیده می‌شود[۳۸]. یک شرط لازم برای اعتبار یک مدل رگرسیون آن است که ضریب تعیین (R^2) تا حد ممکن به یک نزدیک بوده و ضریب همبستگی (R) حتی المقدور بزرگ باشد و همچنین خطای استاندارد (SE) کوچکی داشته باشد. به هر حال این توانایی برازش، حتی زمانی که مدل‌ها برازش نزدیک تری را برای تعداد زیادی از پارامترها و متغیرها در مدل به دست می‌دهند (SE کوچک‌تر و R^2 بزرگ‌تر)، در اعتبار سنجی مدل کافی نیست، زیرا این پارامترها ارتباطی به توانایی مدل برای ایجاد پیش بینی های مطمئن بر روی داده‌های بعدی ندارند [۳۹]. از طرف دیگر زمانی که مجموعه ای شامل تعداد زیادی توصیف کننده برای انتخاب از بین آن‌ها در دسترس باشد، مدل‌های به دست آمده برازش خوبی دارند ولی همراه با همبستگی شانسی[۲] هستند. اگر در مدلی متغیرهای مستقل به صورت تصادفی با متغیرهای پاسخ همبستگی داشته باشند، گفته می‌شود مدل دارای همبستگی شانسی است بدین معنی که منحنی مقادیر نظری محاسبه شده از مدل بر حسب مقادیر تجربی در یک خط مستقیم واقع شده بدون اینکه توانایی پیش بینی داشته باشد. برای جلوگیری از ایجاد مدل‌های با همبستگی شانسی، باید یک اعتبار سنجی متفاوت انجام گیرد. برای این منظور بهتر است یک سری ارزیابی مناسب از ترکیبات که در فرایند آموزش و بهینه

1- Mean relative error
۲ - Chance correlation

سازی دخالت نداشته باشند به نام سری ارزیابی بیرونی[1]، در دسترس باشد[۴]. تعدادی از تکنیک های اعتبار سنجی در زیر فهرست شده اند.

۲-۱۳-۱- ارزیابی متقاطع یا اعتبار سنجی تقاطعی

رایج ترین تکنیک اعتبار سنجی است که در آن در هر بار یکی یا یک گروه کوچک از داده‌ها کنار گذاشته شده و سپس برای داده‌های باقی مانده، مدلی به دست می‌آید. بعد از آن پاسخ برای داده‌های کنار گذاشته شده از روی این مدل پیش بینی می‌شود[۴۰].

این روش ها به ترتیب به نام‌های رد تک تک داده‌ها[2] و رد گروهی از داده‌ها[3] نامیده می‌شوند.

۲-۱۳-۲- تقسیم کردن مجموعه به آموزش و ارزیابی

در این تکنیک مجموعه داده‌ها به یک سری آموزش[4] و یک سری ارزیابی[5] تقسیم بندی می‌شوند. مدل از سری آموزش به دست آمده و قدرت تعمیم و پیش بینی توسط سری ارزیابی بررسی می‌گردد. تقسیم بندی داده ها به صورت کاملا تصادفی انجام گرفته و نتایج نیز تا حد زیادی به این تقسیم بندی وابسته اند[۴].

۲-۱۳-۳- اعتبار سنجی بیرونی[6]

تکنیکی است که در آن یک سری تست[7] نیز برای انجام یک سری بررسی اضافی بر روی قدرت پیش بینی مدل به دست آمده از سری آموزش و با قدرت پیش بینی بهینه شده توسط سری ارزیابی، در نظر گرفته می‌شود. سری تست نباید دخالتی در روند آموزش و انتخاب مدل داشته باشد[۴].

1 - External validation set
۲- Leave one out
2- Leave group out
4 - Train series
5 - Validation series
6 - External validation
7 - Test series

۲-۱۳-۴- آزمون Y- تصادفی[۱]

این تکنیک برای مطالعه همبستگی‌های شانسی مدل غیر خطی طراحی شده است. در این آزمون، مقادیر تجربی (که در اینجا بردار Y نامیده می شود) به صورت تصادفی در محدوده همان مقادیر، تغییر داده شده و سپس همبستگی متغیرهای مستقل با متغیرهای وابسته با استفاده از یکی از شاخص‌های آماری که معمولا R^2 است، مورد بررسی قرار می‌گیرد. اختلاف زیاد بین شاخص آماری به دست آمده از این روش با شاخص آماری به دست آمده از مدل اصلی، نشان دهنده عدم وجود همبستگی شانسی می‌باشد. به طور معمول این فرایند چندین بار انجام می‌شود[۴۱].

۲-۱۴- نرم‌افزارهای مورداستفاده در این تحقیق

دانش کمومتریکس بسته‌های نرم‌افزاری متنوعی را برای انجام تمام مراحل مدل‌سازی بکار گرفته است که در ادامه به اختصار بسته‌های نرم‌افزاری استفاده شده در این تحقیق معرفی می‌شوند.

۲-۱۴-۱- نرم‌افزار Hyperchem

از بسته نرم‌افزاری Hyperchem [۴۲] برای رسم شکل مولکول‌ها و بهینه‌سازی ساختار با استفاده از روش‌های کوانتومی و مکانیکی، استفاده می‌شود. به کمک این برنامه می‌توان طول پیوند، زاویه پیوندی وزوایای پیچشی را در مولکول تعیین کرد.داده‌های حاصل از این نرم‌افزار را می‌توان به عنوان ورودی به سایر نرم افزارها معرفی نمود. به عنوان مثال ساختار سه بعدی مولکول‌ها که با استفاده از روش مکانیک کوانتومی در Hyperchem بهینه شده است، را می‌توان جهت محاسبه توصیف‌کننده مختلف به وسیله نرم‌افزارDragon، به کار برد. همچنین به کمک این نرم‌افزار می‌توان تعدادی از توصیف‌کننده‌ها از جمله حجم مولی و قطبش‌پذیری را محاسبه کرد.

۱- Y- Randomization

۲-۱۴-۲- نرم‌افزار Dragon

نرم‌افزار Dragon که توسط گروه تحقیقاتی کمومتریکس میلانو ارائه شده[۴۳]، امکان محاسبه ۱۴۸۱ توصیف‌کننده مختلف را فراهم می‌کند. جهت محاسبه توصیف‌کننده به کمک این نرم‌افزار لازم است ساختار هندسی بهینه مولکول مورد استفاده قرار گیرد. برای این منظور می‌توان ساختار بهینه مولکول‌ها را به صورت فایل‌هایی با فرمت mol، solf، hin، ... به عنوان اطلاعات ورودی به کار برد. توصیف‌کننده ها برای ارزیابی روابط ساختار - فعالیت یا ساختار - ویژگی مولکولی به کار رفته و به ۱۸ دسته اصلی تقسیم می‌شوند.

۲-۱۴-۳- بسته نرم‌افزاری SPSS[1]

SPSS[۴۴] که نخستین نسخه آن در سال ۱۹۷۰ توسط جمعی از فارغ التحصیلان دانشگاه استانفورد آمریکا ارائه شده، امکان تجزیه و تحلیل آماری داده‌ها را فراهم می‌آورد. برخی از قابلیت‌های این بسته نرم‌افزاری عبارتند از:

- ✓ تعیین تعداد فراوانی‌های هر یک از گروهها در یک متغیر
- ✓ محاسبه میانگین ساده برای داده‌ها
- ✓ نمایش اطلاعات به صورت متنوع در قالب نمودار و جدول
- ✓ انجام رگرسیون تک متغیره و چند متغیره

۲-۱۴-۴- نرم‌افزار MATLAB

MATLAB[2] به معنای آزمایشگاه ماتریس است. ورودی‌ها اساساً به صورت ماتریس در نظر گرفته می‌شوند و هیچ نیازی به مشخص کردن ابعاد ماتریس نمی‌باشد. در MATLAB حتی اعداد اسکالر،

1- Statistical Package for the Social Science
2- Matrix Laboratory

ماتریس‌های (x) به حساب می‌آیند و بردارها، حالت خاصی از ماتریس‌های سطری یا ستونی در نظر گرفته می‌شوند.

در محیط MATLAB با استفاده از فرامین، آرایه ها و امکانات ویژه، مدلسازی غیر خطی فعالیت‌ها و خواص شیمیایی ترکیبات با ساختار آن‌ها امکانپذیر است. سایر روش های کمومتریکس مانند آنالیز رگرسیون اجزای اصلی، روش کمترین مربعات جزئی، الگوریتم ژنتیک و غیره نیز توسط این نرم افزار قابل اجرا هستند[۲۹].

فصل سوم

مدل سازی QSPR شاخص بازداری برخی از ترکیبات آلی فرار، با استفاده از روش غیر خطی شبکه عصبی (ANN) و استفاده از الگوریتم ژنتیک بعنوان یک روش انتخاب متغیر

۳-۱- سری داده‌ها

در این قسمت، شاخص بازداری ۶۰ ترکیب آلی فرار[1] مختلف موجود در هوا که با کروماتوگرافی گازی- اسپکترومتری جرمی[2] و کروماتوگرافی گازی- یونیزاسیون شعله‌ای[3]مورد آنالیز قرار گرفته‌اند به عنوان سری داده‌ها مورد استفاده قرار گرفت[۱]. نام و اندیس بازداری این ترکیبات در جدول (۳-۱) آمده است. داده های جدول بیانگر این هستند که اندازه مولکول، تعداد شاخه‌ها و نحوه‌ی آرایش اتم‌ها در مولکول، بر روی شاخص بازداری تأثیر می‌گذارند. لذا در این تحقیق سعی شده مدلی ارائه گردد که بتواند شاخص بازداری این ترکیبات را با استفاده از ویژگی‌های ساختاری آن‌ها پیش‌بینی کند.

جدول (۳-۱)- سری داده ها به همراه نام ترکیب و شاخص بازداری آن‌ها

شماره ترکیب	نام ترکیب	شاخص بازداری
1 (tr)*	2,3-dimethyl butane	562
2 (tr)	2-methyl pentane	565
3 (te)	3-methyl pentane	581
4 (tr)	n-hexane	600
5 (v)	2,2-dimethyl pentane	623
6 (tr)	Methyl cyclopentane	626
7 (tr)	2,4-dimethyl pentane	626
8 (v)	2,2,3-trimethyl butane	635
9 (te)	Benzene	650
10 (tr)	3,3-dimethyl pentane	655
11(tr)	Cyclohexane	659
12 (tr)	2-methyl hexane	667
13 (te)	2,3-dimethylpentane	669
14 (tr)	3-methyl hexane	677
15 (v)	3-ethyl pentane	687
16 (tr)	2,2,4-trimethyl pentane	691
17 (v)	n-heptane	700
18 (tr)	Methyl cyclohexane	728
19(te)	2,2-dimethyl hexane	729
20 (tr)	2,5-dimethyl hexane	741
21 (tr)	2,3,4-trimethyl pentane	760

1- Volatile Organic Compounds (VOCs)
2- Gas chromatography- mass spectrometry (GC-MS)
3- Gas chromatography- flame ionization detection (GD-FID)

ادامه جدول (۳-۱)

شماره ترکیب	نام ترکیب	شاخص بازداری
22(te)*	Toluene	764
23(tr)	2,3-dimethyl hexane	769
24(v)	2-methyl hexane	774
25(tr)	4-methyl heptanes	775
26(tr)	3-ethyl hexane	780
27(v)	3-methyl heptanes	781
28(tr)	cis-1,4-dimethyl cyclohexane	785
29(te)	n-octane	800
30(tr)	cis-1,2-dimethyl cyclohexane	834
31(tr)	Ethyl cyclohexane	839
32(te)	3,5-dimethyl heptanes	843
33(v)	Ethyl benzene	857
34(tr)	2,3-dimethyl heptanes	863
35(te)	m-xylene	865
36(tr)	p-xylene	865
37(tr)	3,4-dimethyl heptanes	867
38(tr)	2-methyl octane	870
39(te)	3-methyl octane	877
40 (tr)	Styrene	881
41(tr)	3,3-diethyl pentane	884
42 (v)	o-xylene	886
43(tr)	n-nonane	900
44(tr)	i-propyl benzene	919
45(tr)	2,2-dimethyl octane	922
46(v)	3,3-dimethyl octane	943
47(tr)	n-propyl benzene	950
48(v)	m-ethyl toluene	957
49(tr)	2,3-dimethyl octane	961
50(tr)	p-ethyl toluene	960
51(te)	1,3,5-trimethyl benzene	965
52(tr)	2-methyl nonane	969
53(tr)	3-ethyl octane	973
54(tr)	3-methyl nonane	975
55(tr)	o-ethylene toluene	976
56(v)	1,2,4-trimethyl benzene	990
57(tr)	n-decane	1000
58(tr)	1,2,3-trimethyl benzene	1019
59(te)	m-diethyl benzene	1046
60(tr)	p-diethyl benzene	1054

سری تست **te:**، سری ارزیابی **v:**، سری آموزش *tr:

۳-۲- رسم و بهینه سازی ساختمان هندسی و به دست آوردن توصیف کننده‌ها

در این مرحله، ابتدا ساختار مولکول‌ها به وسیله نرم افزار Hyperchem رسم و ساختار هندسی آن-ها با احتساب اتم‌های هیدروژن و با استفاده از روش AM1 بهینه شد. بهینه سازی تا زمانی ادامه یافت که گرادیان به ۰/۰۰۱ کیلوکالری بر مول برسد. سپس ساختارهای بهینه شده به عنوان ورودی به نرم افزاز Dragon وارد شدند. محاسبه توصیف کننده‌ها توسط این نرم افزار صورت گرفت و حدود ۱۴۸۱ توصیف کننده از هیجده گروه مختلف برای هر ترکیب محاسبه گردید.

۳-۳- انتخاب توصیف‌کننده‌های مهم

به منظور جلوگیری از طولانی شدن فرایند برازش و انتخاب توصیف‌کننده‌های مهم، روش‌های کاهش متغیر به کار گرفته شد. برای نیل به این هدف، توصیف‌کننده‌های به دست آمده از مرحله قبل، به عنوان متغیر مستقل، و اندیس بازداری (RI) نیز به عنوان متغیر وابسته وارد نرم افزارSPSS شدند. تعدادی از توصیف‌کننده‌ها که دارای بیش از ۹۰٪ مقادیر یکسان بودند، حذف شدند. همچنین از میان توصیف‌کننده‌هایی که هم، همبستگی زیادی با پارامتر وابسته داشتند و هم با پارامتر مستقل دیگری دارای همبستگی قابل توجه (ضریب همبستگی بیش از ۰/۹) بودند؛ پارامتری که همبستگی کمتری با پارامتر وابسته مورد نظر داشت حذف گردید. بدین ترتیب در پایان این مرحله تعداد ۱۴۶ توصیف-کننده باقی ماند. سپس با استفاده از دو روش برازش مرحله‌ای[1]و الگوریتم ژنتیک بهترین توصیف کننده‌ها که بیشترین همبستگی را با پارامتر وابسته داشتند، انتخاب گردیدند.

1- Stepwise

۳-۳-۱- انتخاب بهترین توصیف‌کننده‌ها به روش برازش مرحله‌ای

یکی از روش‌های انتخاب متغیر در این تحقیق روش برازش مرحله‌ای بوده که مراحل انجام آن به این صورت است که در منوی آنالیز در نرم‌افزار SPSS، گزینه رگرسیون خطی انتخاب شده و بعد از وارد کردن متغیرهای مستقل و وابسته، روش برازش مرحله‌ای (Stepwise) استفاده شد. در این روش متغیرها یکی پس از دیگری وارد مدل شده و در ابتدا متغیری که بیشترین همبستگی را با متغیر وابسته دارد، وارد می‌شود. با ورود هر متغیر جدید، کلیه متغیرهای موجود در معادله بررسی شده و اگر هر کدام از آن‌ها سطح معناداری خود را از دست داده باشند، قبل از ورود متغیر جدید از مدل خاج می‌شوند. به این ترتیب در روش برازش مرحله‌ای توصیف‌کننده‌های مناسب انتخاب می‌شوند. که در تحقیق حاضر، ۷ توصیف‌کننده انتخاب شدند. این ۷ توصیف‌کننده به همراه طبقه آن‌ها در جدول (۳-۲) آورده شده است. همچنین در جدول (۳-۳) ماتریس همبستگی بین این توصیف‌کننده‌ها ارائه شده است و این ماتریس عدم همبستگی بین این توصیف‌کننده‌ها را نشان می‌دهد.

جدول (۳-۲)- توصیف کننده های انتخابی توسط روش برازش مرحله ای به همراه معنی و طبقه آن‌ها

No	Symbol	Class	Meaning
1	G1	Geometrical descriptors	Gravitational index G1
2	Mor09m	3D MORSE	3D –MORSE-signal 09/weighted by atomic masses
3	RDF035v	RDF descriptores	Radial Distribution Function -3.5/weighted by atomic vander waals volumes
4	GATS3v	2D autocorrelations	Geary autocorrelation-lag 3/weighted by atomic vander waals volumes
5	RDF045u	RDF descriptores	Radial Distribution Function -4.5/unweighted
6	BEHm3	Burden eigenvalues	Highest eigenvalue n.3 of Burden matrix/ weighted by atomic masses
7	Mor06v	3D MORSE	3D –MORSE-signal 06/weighted by atomic vander waals volumes

جدول (۳-۳)- ماتریس همبستگی توصیف کننده های انتخاب شده توسط روش برازش مرحله ای

	G1	Mor09m	RDF035v	GATS3v	RDF045u	BEHm3	Mor06v
G1	۱						
Mor09m	-۰/۲۳۰	۱					
RDF035v	۰/۳۲۳	۰/۱۹۳	۱				
GATS3v	-۰/۰۴۲	۰/۲۱۴	۰/۳۶۶	۱			
RDF045u	۰/۵۱۴	۰/۱۸۲	۰/۵۱۳	۰/۰۷۶	۱		
BEHm3	۰/۶۷۸	-۰/۱۹۴	-۰/۱۳۱	-۰/۴۹۱	۰/۲۴۷	۱	
Mor06v	۰/۳۵۲	۰/۴۲۲	-۰/۱۳۱	۰/۱۲۳	۰/۰۲۲	۰/۴۲۹	۱

۳-۳-۲- انتخاب توصیف‌کننده‌ها به روش الگوریتم ژنتیک

همانطور که گفته شد، در این تحقیق از دو نوع روش انتخاب متغیر استفاده شده که روش دوم انتخاب متغیر الگوریتم ژنتیک (GA) می‌باشد. انتخاب متغیر با این روش با استفاده از جعبه‌ابزار GA- PLS در محیط نرم‌افزار MATLAB انجام شده است. الگوریتم ژنتیک هیچ چیزی در باره مسأله‌ای که حل می کند، نمی‌داند و برای اجرای جستجوی مؤثر فقط به مقادیر تابع هدف نیاز دارد، بنابراین شامل پیچیدگی‌های محاسباتی اضافی نمی‌شود از آنجایی GA که روشی غیرجبری است، پاسخ دقیق مسأله را نیافته و حتی ممکن است برای حل یک مسأله مشخص با هر بار بکارگیری، پاسخی متفاوت ارائه دهد، هرچند تمامی این پاسخ‌ها می‌توانند دقت مورد نظر را برآورده کنند. تاکنون هیچ روش استانداردی برای حل یک مسأله بخصوص در دسترس نبوده و با بکارگیری تجربیات موجود می‌توان الگوریتم ژنتیک را با احتیاجات خود منطبق کرد.

الگوریتم ژنتیک برای حل یک مسئله مجموعه بسیاربزرگی ازراه حل‌های ممکن راتولید می‌کند. هریک ازاین راه حل‌ها با استفاده از یک تابع تناسب مورد ارزیابی قرار می‌گیرند. فضای جستجو در جهتی تکامل پیدا می‌کند که به راه حل مطلوب برسد. در فرایند اجرای این الگوریتم، ابتدا به صورت اتفاقی یک جمعیت اولیه از توصیف کننده‌ها تولید شده و برازندگی یا شایستگی تک تک اعضای هر

نسل یعنی توصیف کننده‌ها، محاسبه و با توجه به شایستگی‌ها، نسل‌های بعدی با اعمال سه عملگر انتخاب، پیوند و جهش تولید یا باز ترکیب می‌شوند. این روند جستجو برای جمعیت‌های مختلف تا حصول ملاک خاتمه ادامه می‌یابد. مراحل ایجاد جمعیت جدید شامل موارد زیر است :

- انتخاب: انتخاب دو کروموزوم والد از جمعیت موجود بر طبق تناسب آن‌ها.
- پیوند: از ترکیب والدین یک فرزند جدید تولید می‌شود و اگر پیوندی صورت نگیرد فرزند دقیقا کپی یکی از والدین می‌باشد.
- جهش: ممکن است با یک احتمال در هر یک از کروموزوم‌های فرزند جهش ایجاد شود.
- پذیرش: فرزند جدید در یک جمعیت جدید پذیرفته می‌شود.
- جایگزینی: جمعیت جدید تولید شده جایگزین جمعیت قبلی می شود تا الگوریتم بار دیگر تکرار شود.

تست: اگر به شرط پایان الگوریتم رسیده باشیم اجرای برنامه متوقف می‌شود و بهترین راه حل از جمعیت جاری برگردانده می‌شود.

تا اتمام روند الگوریتم ژنتیک n بار برنامه اجرا می‌شود. در نیمه‌ی اول اجرای برنامه بردار Y یا تابع متغیرهای وابسته همان مقادیر اصلی در نظر گرفته می‌شود و در نیمه دوم یک بردار Y به روش اتفاقی تولید می‌شود. فرض می‌کنیم برنامه ۴۰ بار اجرا شود. در پایان ۲ بردار خواهیم داشت که اولی شامل میانگین ۲۰ اجرای اول الگوریتم بر روی مقادیر اصلی، و دومی شامل میانگین ۲۰ اجرای دوم الگوریتم بر روی مقادیر تصادفی خواهد بود. به طور ساده می‌توان گفت که اجرای الگوریتم ژنتیک با بردار Y اصلی نشان دهنده توانایی الگوریتم مورد نظر برای مدل سازی "اطلاعات همراه با مقادیر ناخواسته[1]" است و اجرای الگوریتم برای بردار Y تصادفی نیز نشان دهنده توانایی الگوریتم برای مدل سازی مقادیر" ناخواسته [2]"می‌باشد. حال تفاوت این دو می‌تواند توانایی الگوریتم ژنتیک را در ارتباط با

1-Information+ noise
2-Noise

مقادیر حقیقی نشان دهد. می‌توان گفت که بهترین لحظه توقف یک اجرای الگوریتم ژنتیک بر طبق ارزیابی زمانیست که ماکزیمم اختلاف[1] به دست آید[۴۶].

به طور متوسط هر بار برای انتخاب متغیر ۵ مرتبه برنامه اجرا می‌گردد. در جدول (۳-۴) توصیف-کننده‌های انتخاب شده در هر بار اجرای الگوریتم ژنتیک گزارش شده است.

جدول (۳-۴)- توصیف کننده‌های انتخاب شده توسط الگوریتم ژنتیک

توصیف کننده‌های انتخاب شده توسط الگوریتم ژنتیک	تعداد تکرار
G1-H2u -FDI- H2m -X2sol -Mor06v -X2v -Mor09m- RDF035v	۱
G1 - H2m - Mor06v- X3sol- H2u - GATS3m - RDF035e	۲
G1- H2m - H2u – X2v- BELe5- RDF035v- X3sol-FDI	۳
G1- H2u - X2sol - H2m - RDF035v- Mor06v-FDI	۴
G1- H2m - FDI - X2sol - Mor06v - Yindex – RDF035e	۵

سپس تمام متغیرهای انتخابی در ۵ بار، به عنوان ورودی جدید به GA-PLS داده شد و این بار متغیرهای انتخابی به عنوان متغیرهای برتر در نظرگرفته شده و برای مدلسازی به شبکه عصبی داده شد. در نهایت، از بین مجموع توصیف‌کننده‌های انتخاب شده در ۵ بار اجرای الگوریتم ژنتیک، ۸ توصیف‌کننده انتخاب شدند:

G1, H2u, X2sol, RDF035v, Mor06v, Mor09m, H2m, FDI

۳-۳-۲-۱- جمعیت

این برنامه کروموزوم ساخته شده به وسیله الگوریتم ژنتیکی را گرفته و بر اساس آن، توصیف کننده‌های منتخب را بطور خودکار از مجموعه‌ی داده‌ها انتخاب می‌کند. همچنین به صورت تصادفی

1-Maximum diffrence

ترکیبات را به دو سری آموزش و آزمون تقسیم می نماید. اکثر محققان برای اندازه جمعیت عددی بین ۳۰ تا ۱۰۰ پیشنهاد کرده‌اند. این مجموعه از جمعیت‌های انتخاب شده اولین نسل را ایجاد می-کند و از نسل اولیه نیز نسل‌های بعدی ایجاد می‌شود. سرانجام در نسل نهایی، یک جمعیت انتخاب شده که موجب حل بهینه‌ی مسأله است، به دست می‌آید. در این مطالعه، بنابرآنچه که در جعبه ابزار GA- PLS از پیش تعیین شده، اندازه جمعیت عدد ۳۰ می باشد[۴۷].

۳-۳-۲-۲- عملگر پیوند

عملگر پیوند می‌تواند دو نقطه‌ای باشد که در این صورت، دو نقطه به طور تصادفی انتخاب شده و پیوند بین آن دو اتفاق می‌افتد. این عملیات در چند نقطه انجام می‌شود که به آن بازترکیبی چندنقطه‌ای می‌گویند. معمولا احتمال پیوند برای هر زوج کروموزوم ۰/۱ تا ۰/۹ در نظر گرفته می‌شود که در این تحقیق این مقدار ۰/۵ فرض شده است[۴۷].

۳-۳-۲-۳- عملگر جهش

جهش در جمعیت تنوع ایجاد نموده و به پرهیز از همگرایی زودرس کمک می کند. احتمال وقوع این عملگر بین ۰/۰۰۵ تا ۰/۰۱ انتخاب می‌شود (۰/۵ تا ۱ درصد)، که در این مطالعه مقدار ۰/۰۱ در نظر گرفته شده که پیش فرض برنامه می‌باشد[۴۷].

۳-۳-۲-۴- توصیف‌کننده‌های انتخاب شده توسط روش الگوریتم ژنتیک

توصیف‌کننده های انتخاب شده نهایی توسط این روش در جدول (۳-۵) به همراه طبقه و معنی آن‌ها آورده شده است.

جدول (۳-۵)- توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک

No	Symbol	Class	Meaning
1	G1	Geometrical descriptors	Gravitational index G1
2	H2u	GETAWAY descriptors	H auto correlation of lag 2/ unweighted
3	X2sol	Connectivity indices	Salvation connectivity index chi-2
4	RDF035v	RDF descriptores	Radial Distribution Function - 3.5/weighted by atomic vander waals volumes
5	Mor06v	3D MORSE	3D –MORSE-signal 06/weighted by atomic vander waals volumes
6	Mor09m	3D MORSE	3D –MORSE-signal 09/weighted by atomic masses
7	H2m	GETAWAY descriptors	H auto correlation of lag 2/ weighted by atomic masses
8	FDI	Geometrical descriptors	Folding degree index

همچنین ماتریس همبستگی بین این توصیف کننده‌ها در جدول (۳-۶) ارائه شده که این ماتریس عدم همبستگی بین توصیف کننده‌ها را نشان می‌دهد.

جدول (۳-۶)- ماتریس همبستگی توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک

	G1	H2u	X2sol	RDF035v	Mor06v	Mor09m	H2m	FDI
G1	۱							
H2u	۰/۰۱۴	۱						
X2sol	۰/۸۱۲	۰/۰۸۰	۱					
RDF035v	۰/۳۲۲	۰/۱۳۴	۰/۴۲۸	۱				
Mor06v	۰/۳۵۲	-۰/۱۰۶	۰/۲۹۸	-۰/۱۳۱	۱			
Mor09m	-۰/۲۳۰	۰/۰۹۰	-۰/۱۶۹	۰/۱۹۳	۰/۴۲۲	۱		
H2m	۰/۷۷۶	۰/۰۱۶	۰/۶۳۲	-۰/۰۳۲	۰/۲۹۶	-۰/۳۰۶	۱	
FDI	۰/۲۲۰	-۰/۰۴۶	۰/۱۳۱	-۰/۵۴۰	-۰/۱۲۷	-۰/۷۶۹	۰/۴۷۰	۱

۳-۴- مدلسازی شبکه عصبی مصنوعی با توصیف‌کننده‌های انتخاب شده توسط روش برازش مرحله‌ای(SR-ANN)

با توجه به احتمال وجود روابط غیرخطی بین متغیرهای مستقل و متغیر وابسته، از شبکه عصبی مصنوعی برای مدل‌سازی استفاده گردید. شبکه مورد استفاده در این تحقیق، یک شبکه پیشخور با تکنیک پس‌انتشار است که برنامه نویسی و اجرای آن، در محیط MATLAB صورت گرفت. مقدار عددی توصیف کننده‌های انتخاب شده توسط روش برازش مرحله‌ای به عنوان ورودی و مقادیر تجربی اندیس بازداری (RI) نیز به عنوان هدف به شبکه عصبی داده شد تا پاسخ شبکه با آن‌ها سنجیده شود. داده ها به صورت تصادفی به سه سری آموزش (۳۶ ترکیب=۶۰٪ داده ها)، ارزیابی (۱۲ ترکیب=۲۰٪ داده ها) و تست (۱۲ ترکیب=۲۰٪ داده ها) تقسیم شدند. سری آموزش برای آموزش شبکه و سری ارزیابی برای بررسی قدرت تعمیم پذیری مدل به دست آمده از سری آموزش، به کار گرفته شد. سری تست نیز به عنوان داده های خارجی برای ارزیابی نهایی شبکه در نظر گرفته شد. برای آموزش موفق و در نتیجه بهتر شدن قدرت پیش بینی شبکه، می‌بایست عوامل مؤثر در آموزش شبکه عصبی مصنوعی مورد بررسی قرارگرفته و بهینه شوند. این پارامترها شامل تعداد متغیرهای ورودی شبکه (تعداد توصیف‌کننده ها=k)، نوع تابع آموزش، نوع تابع تبدیل، تعداد گره ها در لایه پنهان (n)، پارامتر μ(mu) و تعداد دورهای آموزش (epoch) می‌باشد. در این روش تابع کارایی شبکه نیز میانگین مربع خطای استاندارد (MSE) می‌باشد.

۳-۴-۱- بهینه سازی

هر شبکه دارای حداقل یک لایه ورودی، یک لایه خروجی و تعدادی لایه پنهان می‌باشد. تعداد ورودی‌ها (توصیف کننده ها) با تعداد نرون‌های لایه ورودی برابر است و لایه خروجی نیز دارای یک نرون است که نشان دهنده‌ی شاخص بازداری متناظر با هر ترکیب می‌باشد[۴۵]. لازم به ذکر است که هیچ راهنمای مناسبی برای انتخاب تعداد لایه‌های پنهان در دست نیست و ساختار شبکه عصبی،

اغلب به شکل آزمون و خطا ایجاد می‌شود. با وجود این در بیشتر موارد به نظرمی‌رسد که یک لایه پنهان مناسب باشد و بنابراین در این تحقیق نیز یک لایه پنهان مورد استفاده قرار گرفت[۳۴،۳۰]. برای بهینه سازی و شناسایی بهترین پارامترهای شبکه، شبکه‌هایی با ورودی های از ۲ تا ۷ توصیف کننده ایجاد شدند که این ورودی‌ها، همان توصیف کننده‌های به دست آمده از روش انتخاب متغیر برازش مرحله‌ای می‌باشند. هر شبکه با دو الگوریتم آموزشی لونبرگ- مارکوارت (trainlm) و تنظیم بایزین (trainbr) و تعداد گره‌های از ۲ تا ۱۰ در لایه پنهان و تعداد متفاوت دور آموزش ۲ تا ۳۰، آموزش داده شد. برای به دست آوردن بهترین تابع تبدیل در لایه پنهان مدل‌های شبکه عصبی، از توابع لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) استفاده گردید. همچنین از تابع تبدیل خطی (purelin) در لایه خروجی استفاده گردید.

در روند بهینه‌سازی شبکه از میان تمامی موارد ذکر شده، به حداقل رساندن مقدار میانگین مربعات خطای شبکه به عنوان معیار انتخاب گردید.

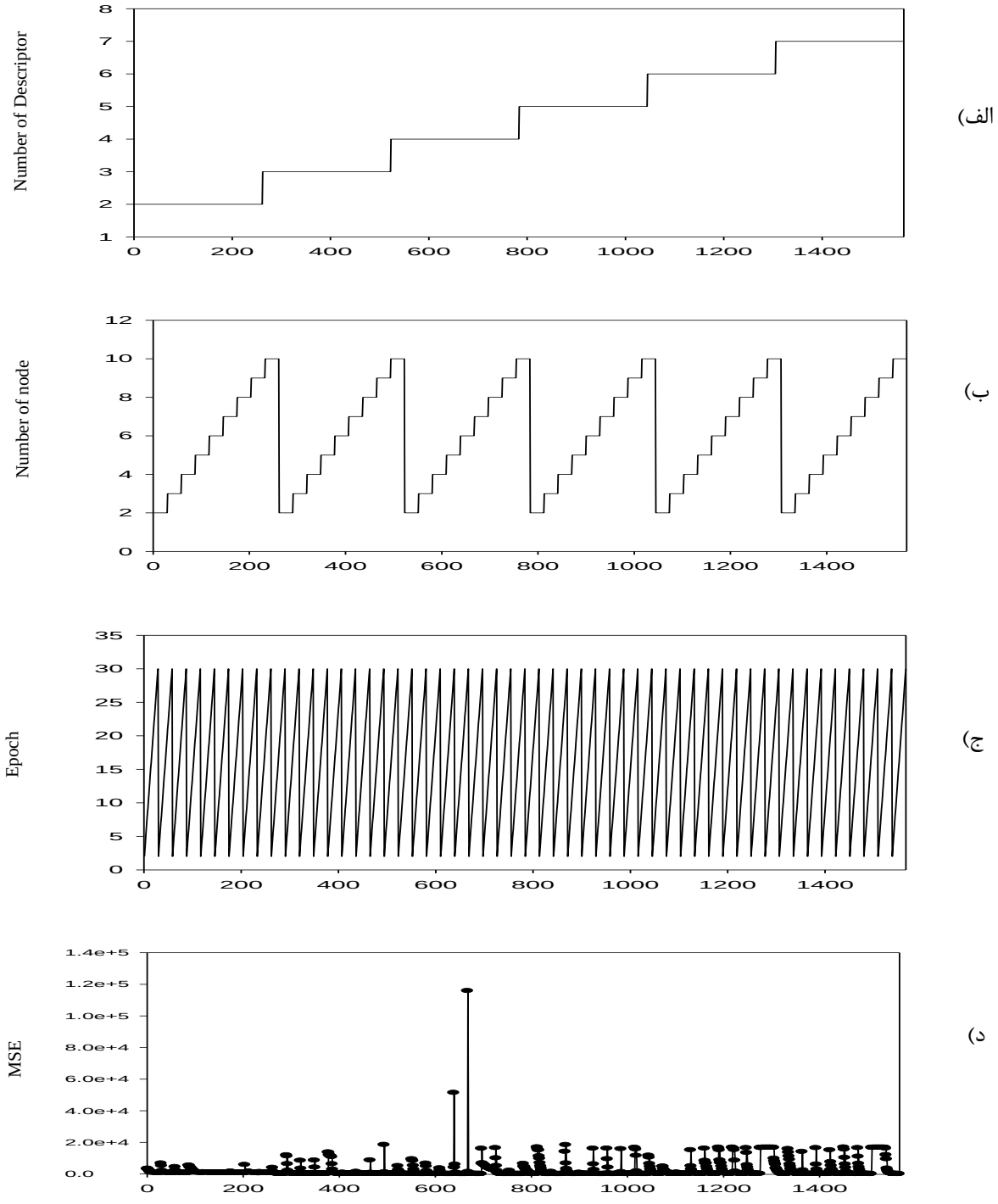
۳-۴-۱-۱- انتخاب تعداد متغیرهای ورودی شبکه، الگوریتم آموزش، نوع تابع تبدیل، تعداد گره‌های لایه‌های پنهان وتعداد دورهای آموزشی

در این مطالعه بهینه سازی وانتخاب بهترین تعداد متغیرهای ورودی شبکه، الگوریتم آموزشی، نوع تابع تبدیل، تعداد نرون‌های لایه پنهان وتعداد دورهای آموزش به طور همزمان انجام گرفت. این کار موجب می‌شود بتوانیم نقش پارامترها را به طور همزمان و یک جا بررسی نماییم (دراین مرحله، مقدار پارامترممنتوم، ۰/۰۰۵ درنظر گرفته شد). سپس در مقادیر بهینه‌ی این پارامترها، مقدار پارامترممنتوم از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد تا با توجه به مقدار خطای شبکه، مقدار بهینه انتخاب گردد.

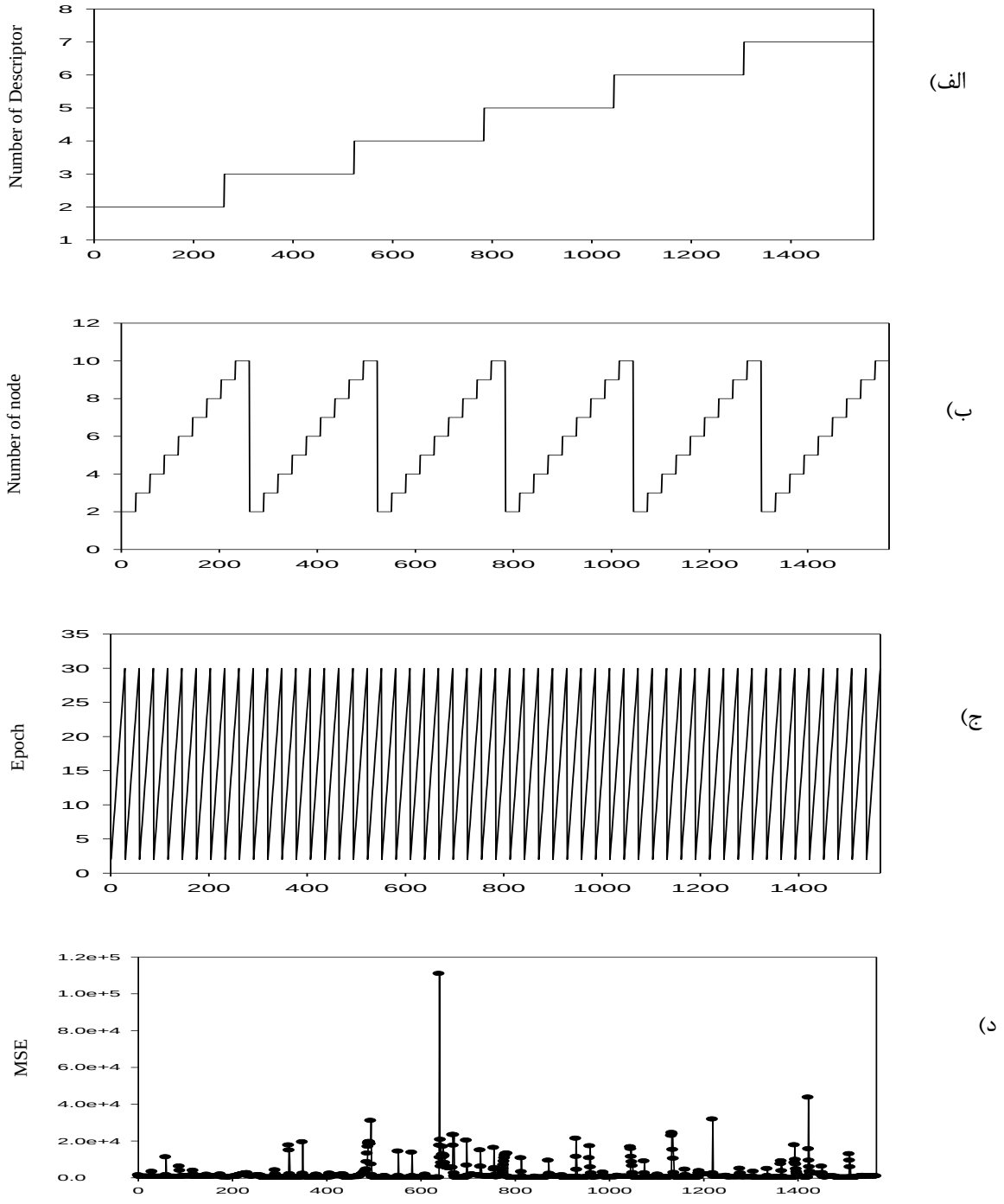
با توجه به پارامترهایی که باید برای شبکه بهینه می‌شد جهت بهینه سازی همزمان تمام پارامترها در مجموع ۶۲۶۴ شبکه مورد ارزیابی قرار گرفت که در زیر به تعداد تمام موارد مورد بررسی اشاره شده است.

۶۲۶۴ = دورهای آموزش (۲۹) × گره (۹) × تابع تبدیل (۲) × تابع آموزش (۲) × توصیف کننده (۶)

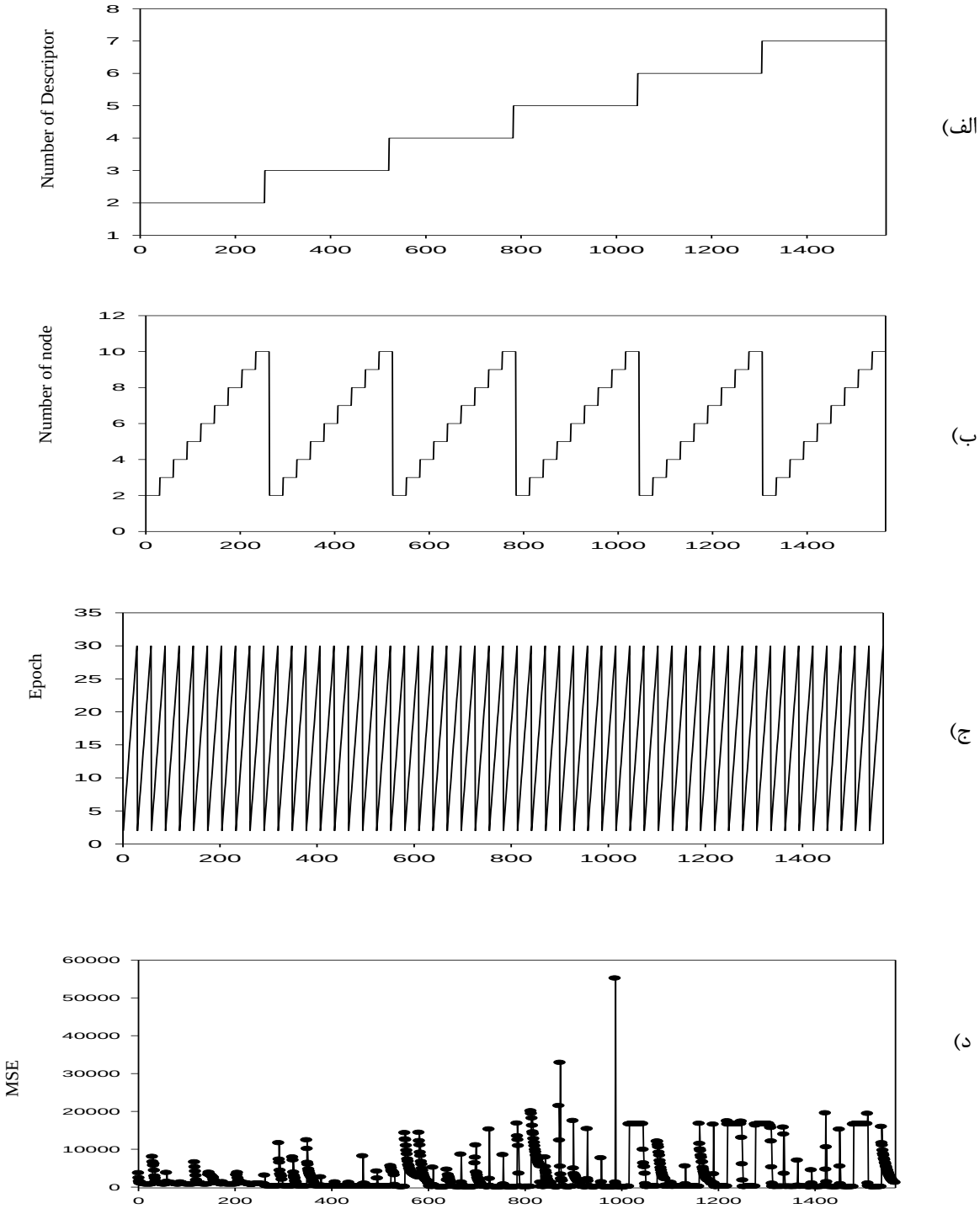
بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE به دست آمده به صورت نموداری بر حسب یک بردار مرجع فرضی در شکل‌های (۳-۱) تا (۳-۴) آمده است.



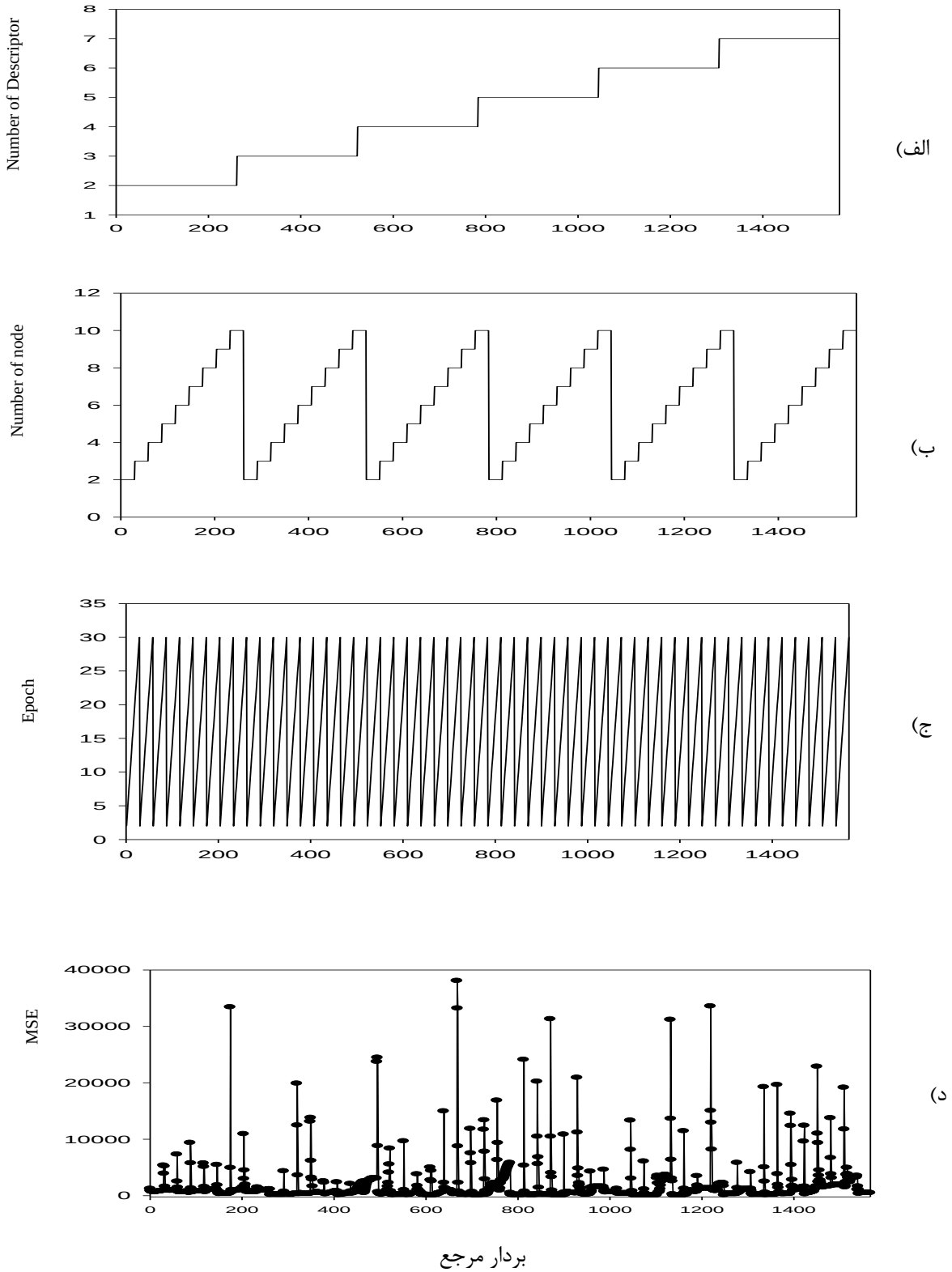
شکل (۳-۱)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



شکل (۳-۲)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی لونبرگ- مارکوارت



شکل (۳-۳)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



شکل (۳-۴)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی لونبرگ- مارکوارت

با توجه به نتایج حاصله برای تعداد مختلف توصیف کننده‌ها، گره‌ها، دورهای آموزش و همچنین توابع متفاوت آموزش و تبدیل، بهترین شبکه‌های به دست آمده براساس کمترین مقدار MSE در جدول (۳-۷) خلاصه شده‌اند. باتوجه به جدول (۳-۷) شبکه عصبی سه لایه با ۵ توصیف‌کننده درلایه ورودی، ۵ نرون درلایه مخفی و تعداد دور آموزشی ۲۳ با تابع آموزشی تنظیم بایزین و تابع تبدیل تانژانت سیگموئید کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین شبکه‌ای با این ساختار به عنوان بهترین مدل برای مدل‌سازی شاخص بازداری ترکیبات در نظر گرفته شد.

جدول(۳-۷)- توابع وپارامترهای شبکه های بهینه بدست آمده براساس مقادیر میانگین مربعات خطا

MSE	تعداد دورهای آموزش	تعدادگره ها	تابع تبدیل	تابع آموزش	تعدادتوصیف کننده ها
۵۸/۰۳	۱۳	۵	لگاریتم سیگموئید	لونبرگ - مارکوارت	۶
۳۴/۰۵	۲۷	۴	لگاریتم سیگموئید	تنظیم بایزین	۵
۴۹/۱۱	۱۳	۲	تانژانت سیگموئید	لونبرگ - مارکوارت	۶
۳۱/۳۸	۲۳	۵	تانژانت سیگموئید	تنظیم بایزین	۵

۳-۴-۱-۲- بهینه سازی پارامتر μ

برای بهینه سازی μ، مقدار این پارامتر در شبکه بهینه شده بین ۰/۰۰۰۵ تا ۰/۰۱ با گام‌های ۰/۰۰۰۵ تغییر داده شد و برای هر مورد مقدار MSE سری های ارزیابی و تست محاسبه گردید. سپس منحنی خطا بر حسب μ رسم و نقطه‌ای که کمترین خطا را برای سری ارزیابی داشت به عنوان مقدار بهینه انتخاب گردید که در این تحقیق، مقدار میانگین مربعات خطای شبکه با تغییرات μ تغییر نکرده و مقدار آن ثابت می‌باشد. بنابراین مقدار μ برابر با مقدار پیش فرض آن در تابع آموزشی بایزین یعنی ۰/۰۰۵ در نظر گرفته شد. (جدول ۳-۸).

جدول (۳-۸)- مقدار خطای شبکه با تغییر پارامتر μ

میانگین مربع خطا	ممنتوم	میانگین مربع خطا	ممنتوم
۳۱/۳۸	۰/۰۰۵۵	۳۱/۳۸	۰/۰۰۰۵
۳۱/۳۸	۰/۰۰۶	۳۱/۳۸	۰/۰۰۱
۳۱/۳۸	۰/۰۰۶۵	۳۱/۳۸	۰/۰۰۱۵
۳۱/۳۸	۰/۰۰۷	۳۱/۳۸	۰/۰۰۲
۳۱/۳۸	۰/۰۰۷۵	۳۱/۳۸	۰/۰۰۲۵
۳۱/۳۸	۰/۰۰۸	۳۱/۳۸	۰/۰۰۳
۳۱/۳۸	۰/۰۰۸۵	۳۱/۳۸	۰/۰۰۳۵
۳۱/۳۸	۰/۰۰۹	۳۱/۳۸	۰/۰۰۴
۳۱/۳۸	۰/۰۰۹۵	۳۱/۳۸	۰/۰۰۴۵
۳۱/۳۸	۰/۰۱	۳۱/۳۸	۰/۰۰۵

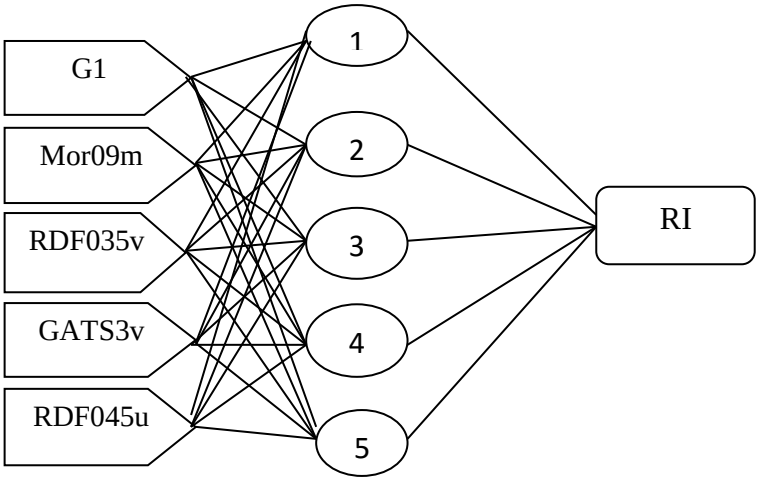
۳-۴-۱-۳-ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، مشخصات شبکه بهینه در جدول (۳-۹) ارائه شده است.

جدول (۳-۹)- مشخصات شبکه بهینه

تابع آموزش	trainbr
تابع تبدیل لایه پنهان	tansig
تابع تبدیل لایه خروجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۵
تعداد متغیرهای ورودی(توصیف‌کننده‌ها)	۵
تعداد چرخه های آموزشی	۲۳
پارامتر μ	۰/۰۰۵

ساختار شبکه عصبی به دست آمده پس از بهینه‌سازی فاکتورهای بحث شده در بالا به طور شماتیک در شکل (۳-۵) نشان داده شده است.



لایه خروجی=یک نرون لایه پنهان=۵ نرون لایه ورودی= ۵ نرون

شکل (۳-۵)- تصویر شماتیک ساختار شبکه عصبی به دست آمده پس از بهینه سازی

۳-۵- مدل‌شبکه عصبی مصنوعی با توصیف‌کننده‌های انتخاب شده توسط الگوریتم ژنتیک (GA-ANN)

در این بخش نیز از شبکه عصبی سه لایه متشکل از یک لایه ورودی، یک لایه پنهان و یک لایه خروجی استفاده شد. در این مرحله نیز شبکه با ورودی های از ۲ تا ۸ توصیف کننده، توسط دو الگوریتم آموزشی لونبرگ – مارکوارت و تنظیم بایزین، با تعداد متفاوت گره در لایه پنهان از ۲ تا ۱۰ و همچنین توابع لگاریتم سیگموئیدی (logsig) و تانژانت سیگموئیدی (tansig)، به عنوان توابع تبدیل لایه پنهان و دورهای آموزش ۲ تا ۳۰ ، آموزش داده شد. همچنین از تابع تبدیل خطی (purelin) در لایه خروجی استفاده شد. معیار نیز به حداقل رساندن میانگین مربع خطا (MSE) در نظر گرفته شد.

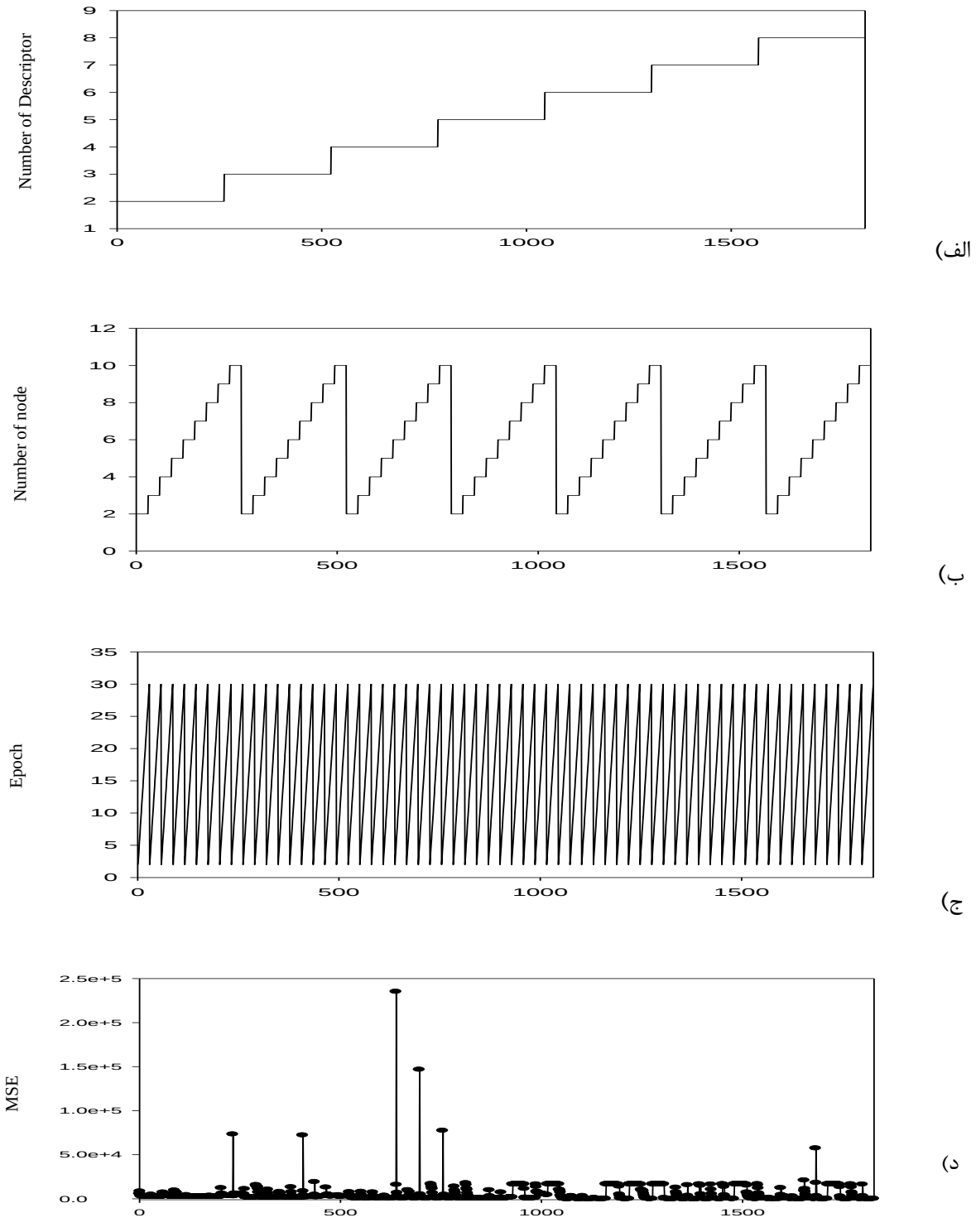
۳-۵-۱- انتخاب تعداد متغیرهای ورودی شبکه، الگوریتم آموزش، نوع تابع تبدیل، تعداد گره‌های لایه‌های پنهان وتعداد دورهای آموزشی

در این جا نیز بهینه سازی و انتخاب بهترین تعداد متغیرهای ورودی شبکه، الگوریتم آموزشی، نوع تابع تبدیل، تعداد نرون‌های لایه پنهان وتعداد دورهای آموزش به طور همزمان انجام گرفت (دراین مرحله، مقدار پارامتر ممنتوم، ۰/۰۰۵ درنظر گرفته شد). سپس در مقادیر بهینه‌ی این پارامترها، مقدار پارامتر ممنتوم از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد تا با توجه به مقدار خطای شبکه، مقدار بهینه انتخاب گردد.

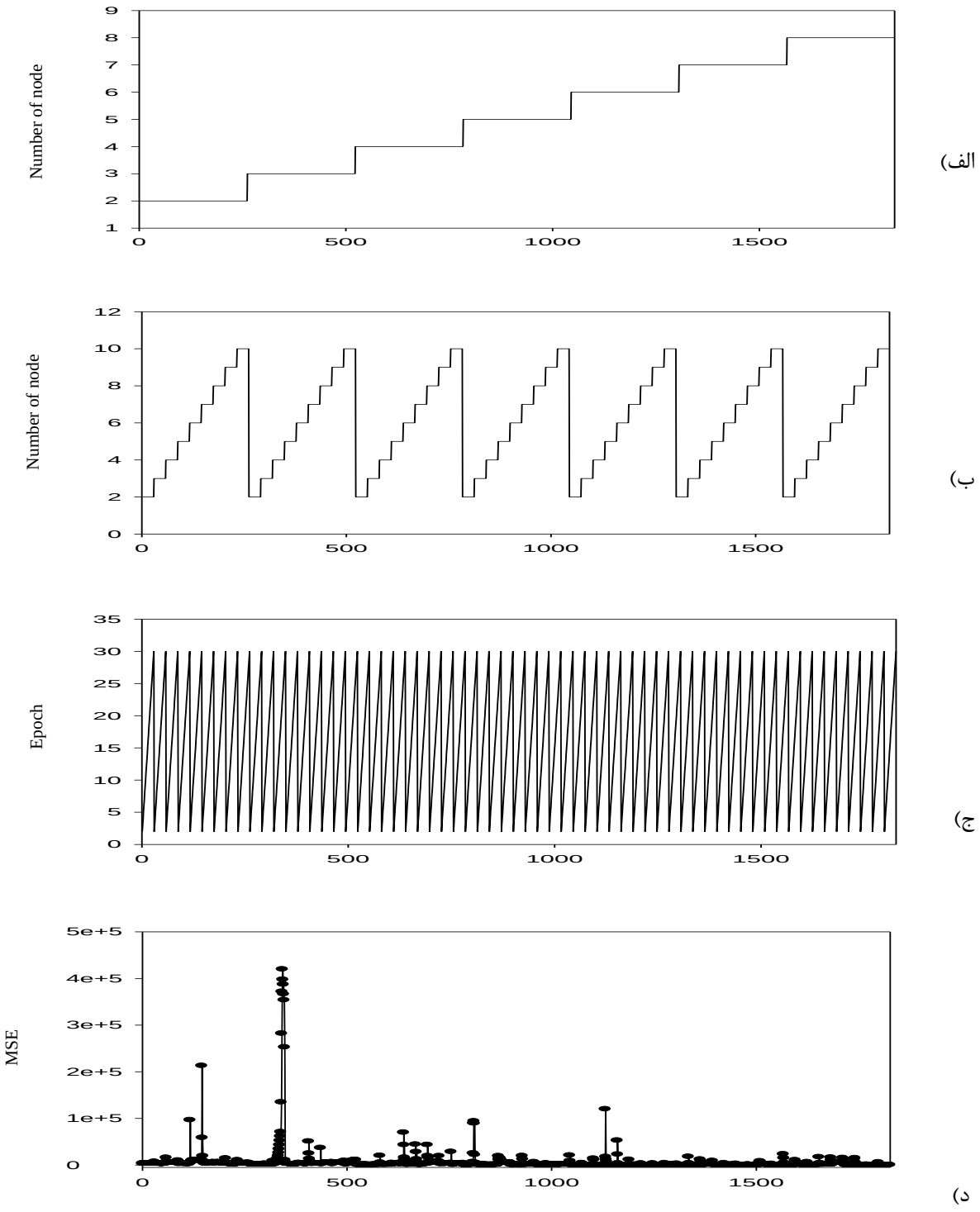
با توجه به پارامترهایی که باید برای شبکه بهینه می‌شد جهت بهینه سازی همزمان تمام پارامترها در مجموع ۷۳۰۸ شبکه مورد ارزیابی قرار گرفت که در زیر به تعداد تمام موارد مورد بررسی اشاره شده است.

۷۳۰۸ = دورهای آموزش (۲۹) × گره (۹) × تابع تبدیل (۲) × تابع آموزش(۲) × توصیف‌کننده (۷)

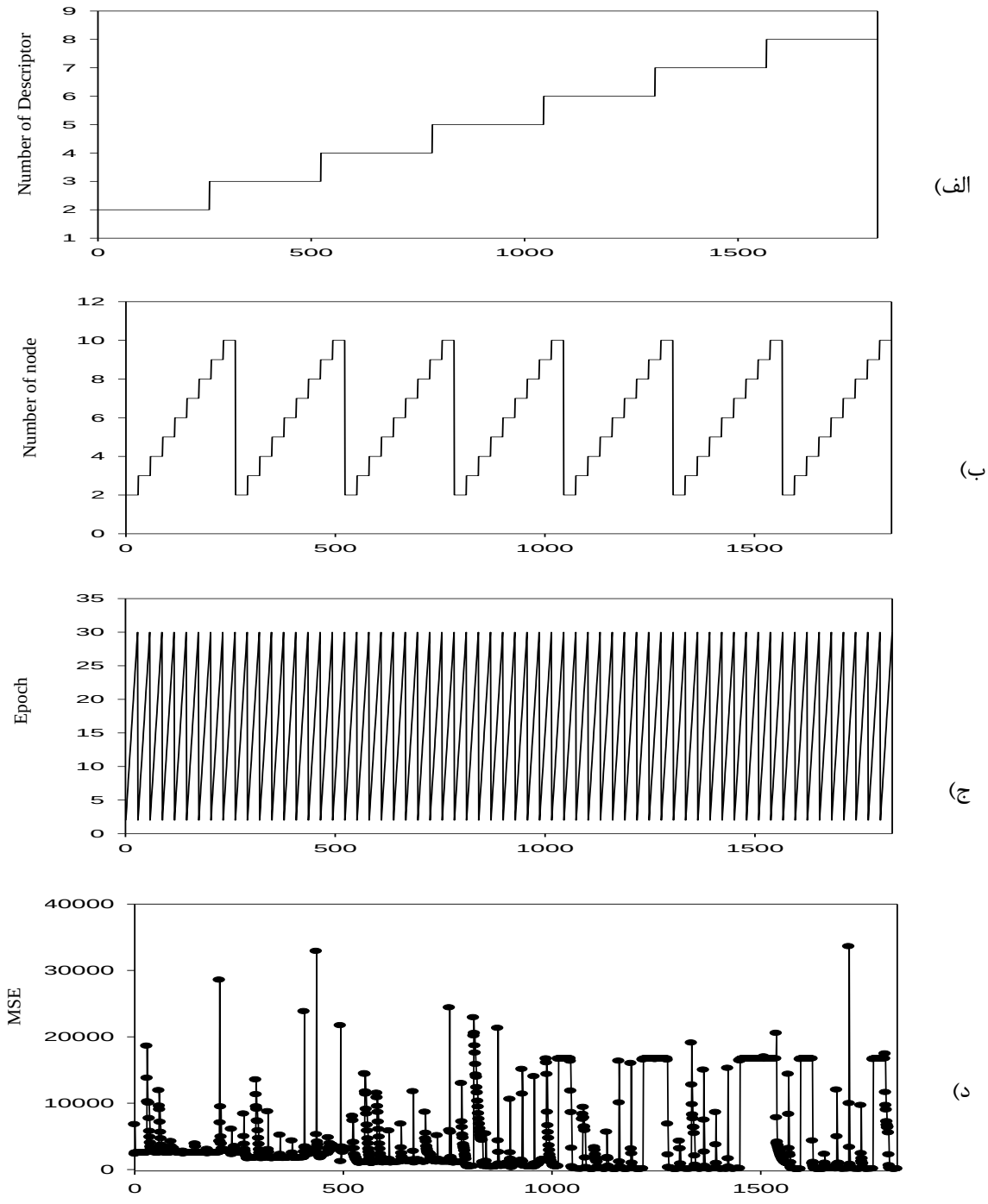
بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE به دست آمده به صورت نموداری بر حسب یک بردار مرجع فرضی در شکل‌های (۳-۶) تا (۳-۹) آمده است.



شکل (۳-۶)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دور آموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی تنظیم بایزین

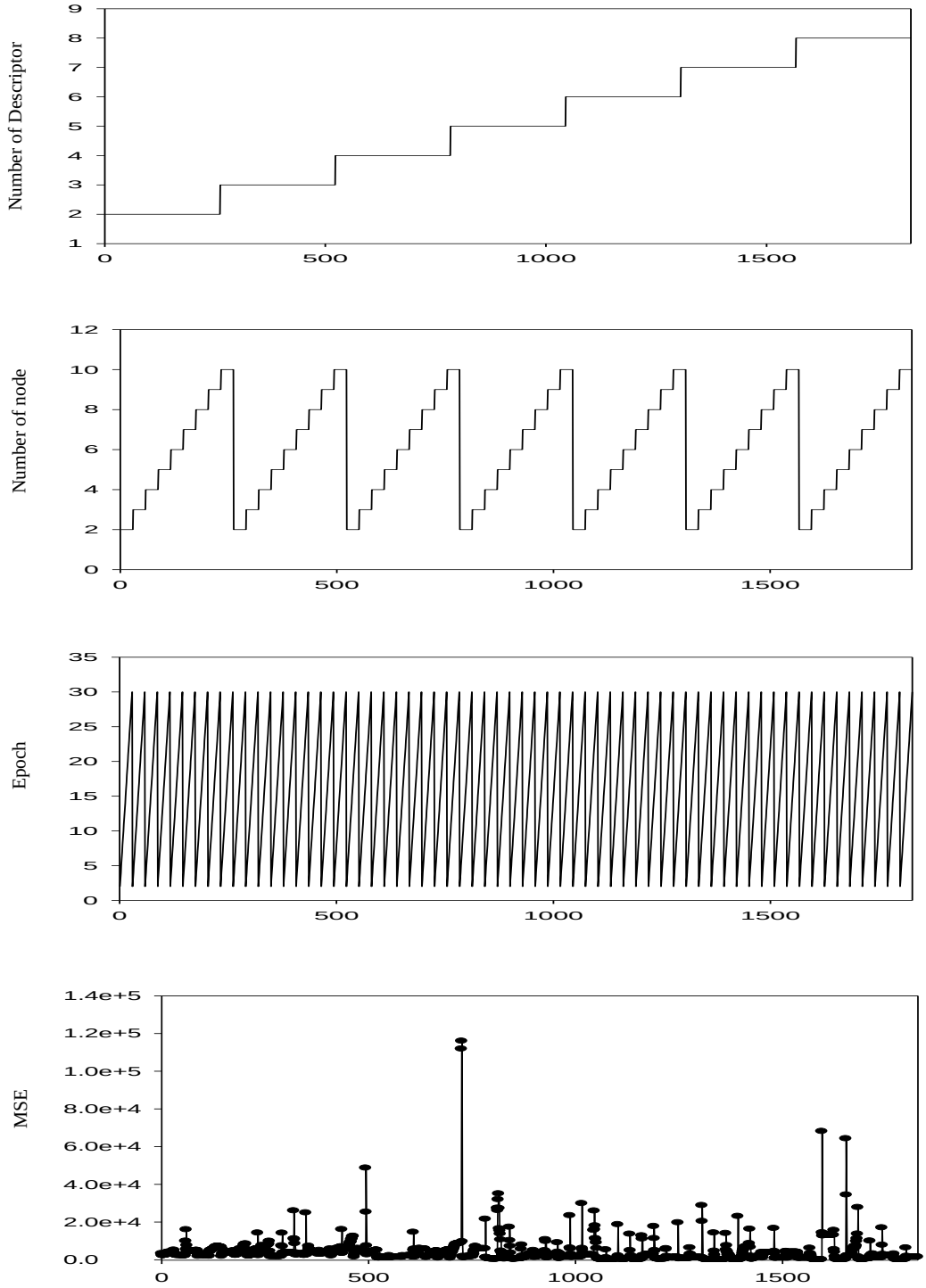


شکل (۳-۷)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دور آموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی لونبرگ- مارکوارت



بردار مرجع

شکل (۳-۸)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دور آموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



بردار مرجع

شکل (۳-۹)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی لونبرگ- مارکوارت

با توجه به نتایج حاصله برای تعداد مختلف توصیف‌کننده‌ها و گره‌ها و دورهای آموزش و همچنین توابع متفاوت آموزش و تبدیل، بهترین شبکه‌های به دست آمده براساس کمترین مقدار MSE در جدول (۳-۱۰) خلاصه شده‌اند. با توجه به جدول (۳-۱۰) شبکه عصبی سه لایه با ۶ توصیف‌کننده درلایه ورودی، ۳ نرون در لایه مخفی و تعداد دور آموزشی ۲۰ با تابع آموزشی تنظیم لونبرگ-مارکوارت و تابع تبدیل تانژانت سیگموئید کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین شبکه‌ای با این ساختار به عنوان بهترین مدل برای مدل‌سازی شاخص بازداری ترکیبات در نظر گرفته شد.

جدول(۳-۱۰)- توابع و پارامترهای شبکه های بهینه بدست آمده براساس مقادیر میانگین مربعات خطا

MSE	تعداد دورهای آموزش	تعدادگره ها	تابع تبدیل	تابع آموزش	تعدادتوصیف کننده ها
۱۱۰/۹۹	۷	۳	لگاریتم سیگموئید	لونبرگ - مارکوارت	۶
۷۰/۵۹	۲۷	۲	لگاریتم سیگموئید	بایزین	۸
۶۱/۲۵	۲۰	۳	تانژانت سیگموئید	لونبرگ - مارکوارت	۶
۶۵/۳۲	۲۴	۶	تانژانت سیگموئید	بایزین	۷

۳-۵-۲- بهینه سازی پارامتر μ

جهت بهینه کردن مقدار μ، ساختار شبکه با ۶ ورودی، ۳ گره در لایه پنهان، ۷ دور آموزش و الگوریتم آموزشی لونبرگ- مارکوارت در نظر گرفته شد. سپس مقدار μ از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. که در این مورد نیز، میانگین مربعات خطای شبکه با تغییرات μ تغییر نکرده و مقدار آن ثابت بود (۶۱/۲۵). بنابراین مقدار μ برابر با مقدار پیش فرض آن در تابع آموزشی لونبرگ- مارکوارت یعنی ۰/۰۰۱ می‌باشد.

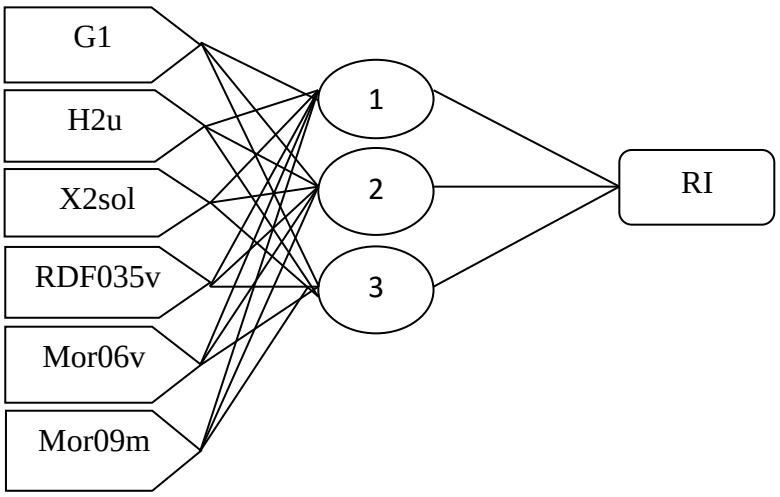
۳-۵-۳- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، مشخصات شبکه بهینه در جدول (۳-۱۱) ارائه شده است.

جدول (۳-۱۱)- مشخصات شبکه بهینه

تابع آموزش	trainlm
تابع تبدیل لایه پنهان	tansig
تابع تبدیل لایه خارجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۳
تعداد متغیرهای ورودی(توصیف‌کننده‌ها)	۶
تعداد چرخه های آموزشی	۲۰
پارامتر μ	۰/۰۰۱

ساختار شبکه عصبی به دست آمده پس از بهینه‌سازی فاکتورهای بحث شده در بالا به طور شماتیک در شکل (۳-۱۰) نشان داده شده است.



لایه خروجی=یک نرون لایه پنهان=۳ نرون لایه ورودی= ۶ نرون

شکل (۳-۱۰)- تصویر شماتیک ساختار شبکه عصبی به دست آمده پس از بهینه سازی

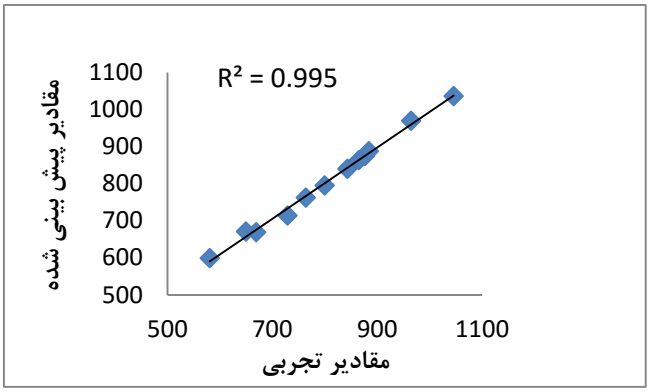
۳-۶- ارزیابی مدل‌های SR-ANN و GA-ANN

۳-۶-۱- ارزیابی مدل‌هایSR-ANN وGA-ANN با استفاده از سری تست

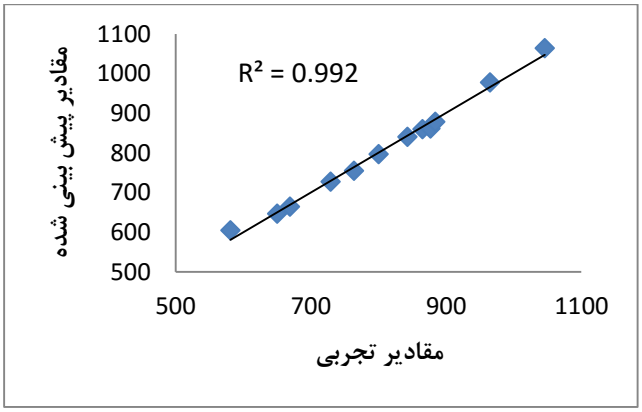
به منظور مطالعه‌ی قدرت پیش بینی مدل‌های ایجاد شده با استفاده از توصیف‌کننده‌های انتخاب شده از برازش مرحله‌ای و الگوریتم ژنتیک، سری تست مورد استفاده قرار گرفت. بدین ترتیب که کل داده ها به سه قسمت با نسبت های ۲۰٪، ۲۰٪ و۶۰٪ تحت عنوان سری های ارزیابی، تست و آموزش تقسیم شدند. در این راستا پارامترهای موثر بر شبکه جهت ایجاد مدل‌های GA-ANN و SR-ANN بهینه شدند و سپس مقادیر RI ترکیبات و سری داده‌ها پیش‌بینی شدند. مقادیر مجذور ضریب همبستگی در شکل‌های (۳-۱۱) و (۳-۱۲) مربوط به سری تست، قدرت مناسب هر دو روش SR-ANN و GA-ANN در پیش بینی شاخص بازداری این ترکیبات را نشان می‌دهد.

جدول (۳- ۱۲) – نتایج حاصل از ارزیابی مدل‌های SR-ANN وGA-ANN با استفاده از سری تست

شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۳	۵۸۱	۵۹۹	۶۰۵	۳/۰۹	۴/۱۳
۹	۶۵۰	۶۷۱	۶۴۷	۳/۲۳	-۰/۴۶
۱۳	۶۶۹	۶۷۰	۶۶۵	۰/۱۵	-۰/۶۰
۱۹	۷۲۹	۷۱۴	۷۲۸	-۲/۰۶	-۰/۱۴
۲۲	۷۶۴	۷۶۲	۷۵۵	-۰/۲۶	-۱/۱۸
۲۹	۸۰۰	۷۹۵	۷۹۷	-۰/۶۲	-۰/۳۷
۳۲	۸۴۳	۸۴۰	۸۴۱	-۰/۳۵	-۰/۲۴
۳۵	۸۶۵	۸۶۴	۸۶۰	-۰/۱۱	-۰/۵۸
۳۹	۸۷۷	۸۷۵	۸۶۲	-۰/۲۳	-۱/۷۱
۴۱	۸۸۴	۸۸۸	۸۷۹	۰/۴۵	-۰/۵۷
۵۱	۹۶۵	۹۷۰	۹۷۸	۰/۵۲	۱/۳۵
۵۹	۱۰۴۶	۱۰۳۶	۱۰۶۴	-۰/۹۶	۱/۷۲



شکل(۳-۱۱)- مقادیر پیش بینی شده RI بر حسب مقادیر تجربی برای سری تست با مدل SR-ANN



شکل(۳-۱۲)- مقادیر پیش بینی شده RI بر حسب مقادیر تجربی برای سری تست با مدل GA-ANN

۳-۶-۲- ارزیابی مدل‌های SR-ANN و GA-ANN به روش رد مرحله‌ای تک‌تک

از دیگر روش‌هایی که برای ارزیابی بیشتر مدل‌های برتر ارائه شده توسط SR-ANN و GA-ANN به کار گرفته شد، تکنیک رد مرحله‌ای تک‌تک (برای کل داده‌ها) می‌باشد. در روش حذف مرحله‌ای تک‌تک داده‌ها، یکی از ترکیبات از سری داده‌ها کنار گذاشته شده و مدل‌سازی با ۵۹ ترکیب باقی-مانده صورت گرفت. سپس مدل به دست آمده برای پیش‌بینی اندیس بازداری ترکیب کنار گذاشته شده به کار گرفته شد. این فرایند برای تمام ترکیبات در سری داده‌ها تکرار گردید.

نتایج حاصل از این روش در جدول (۳-۱۳) ارائه شده است. این نتایج تعمیم پذیری بالای مدل طراحی شده توسط هر دو روش را بیان می‌کند.

جدول(۳-۱۳)- نتایج حاصل از ارزیابی مدل برتر ارائه شده توسط SR-ANN و GA-ANN با استفاده از روش رد مرحله‌ای تک‌تک

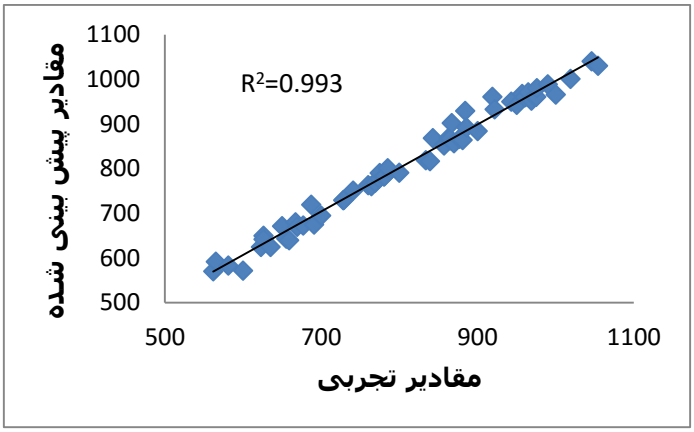
شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۱	۵۶۲	۵۸۱	۵۷۰	۳/۳۸	۱/۴۲
۲	۵۶۵	۵۷۵	۵۹۲	۱/۷۷	۴/۷۸
۳	۵۸۱	۵۹۱	۵۸۳	۱/۷۲	۰/۳۴
۴	۶۰۰	۶۱۸	۵۷۲	۳/۰۰	-۴/۶۷
۵	۶۲۳	۶۲۷	۶۲۵	۰/۶۴	۰/۳۲
۶	۶۲۶	۶۰۸	۶۵۰	-۲/۸۷	۳/۸۳
۷	۶۲۶	۶۳۱	۶۴۲	۰/۸۰	۲/۵۶
۸	۶۳۵	۶۳۳	۶۲۵	-۰/۳۱	-۱/۵۷
۹	۶۵۰	۶۷۱	۶۷۱	۳/۲۳	۳/۲۳
۱۰	۶۵۵	۶۵۱	۶۴۳	-۰/۶۱	-۱/۸۳
۱۱	۶۵۹	۶۲۵	۶۴۰	-۵/۱۶	-۲/۸۸
۱۲	۶۶۷	۶۶۸	۶۸۰	۰/۱۵	۱/۹۵
۱۳	۶۶۹	۶۶۴	۶۷۰	-۰/۷۵	۰/۱۵
۱۴	۶۷۷	۶۸۲	۶۷۳	۰/۷۴	-۰/۵۹
۱۵	۶۸۷	۶۸۸	۷۱۹	۰/۱۴	۴/۶۶
۱۶	۶۹۱	۷۰۲	۶۷۵	۱/۵۹	-۲/۳۱
۱۷	۷۰۰	۶۹۷	۶۹۵	-۰/۴۳	-۰/۷۱
۱۸	۷۲۸	۷۳۰	۷۲۹	۰/۲۷	۰/۱۴
۱۹	۷۲۹	۷۱۸	۷۳۰	-۱/۵۱	۰/۱۴
۲۰	۷۴۱	۷۴۲	۷۵۱	۰/۱۳	۱/۳۵
۲۱	۷۶۰	۷۴۷	۷۶۳	-۱/۷۱	۰/۳۹
۲۲	۷۶۴	۷۵۵	۷۶۱	-۱/۱۸	-۰/۳۹
۲۳	۷۶۹	۷۶۵	۷۶۹	-۰/۵۲	۰/۰۰۶
۲۴	۷۷۴	۷۷۱	۷۷۸	-۰/۳۹	۰/۵۲
۲۵	۷۷۵	۷۸۱	۷۹۰	۰/۷۷	۱/۹۳
۲۶	۷۸۰	۷۷۰	۷۸۲	-۱/۲۸	۰/۲۶
۲۷	۷۸۱	۷۷۷	۷۸۳	-۰/۵۱	۰/۲۶
۲۸	۷۸۵	۸۱۵	۸۰۱	۳/۸۲	۲/۰۴
۲۹	۸۰۰	۷۹۳	۷۹۱	-۰/۸۷	-۱/۱۲
۳۰	۸۳۴	۸۲۳	۸۱۹	-۱/۳۲	-۱/۸۰
۳۱	۸۳۹	۸۲۵	۸۱۷	-۱/۶۷	-۲/۶۲
۳۲	۸۴۳	۸۳۸	۸۶۸	-۰/۵۹	۲/۹۶

ادامه جدول(۳-۱۳)

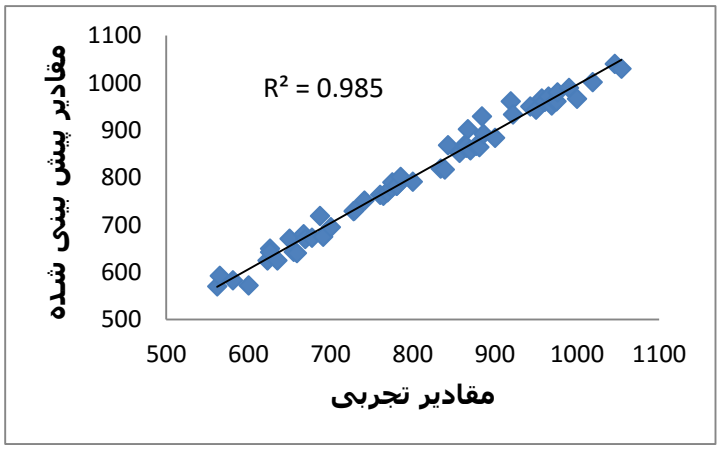
شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۳۳	۸۵۷	۸۵۰	۸۵۲	-۰/۸۲	-۰/۵۸
۳۴	۸۶۳	۸۶۲	۸۶۹	-۰/۱۲	۰/۶۹
۳۵	۸۶۵	۸۶۷	۸۶۳	۰/۲۳	-۰/۲۳
۳۶	۸۶۵	۸۶۶	۸۶۷	۰/۱۱	۰/۲۳
۳۷	۸۶۷	۸۸۲	۹۰۲	۱/۷۳	۴/۰۴
۳۸	۸۷۰	۸۷۱	۸۵۷	۰/۱۱	-۱/۴۹
۳۹	۸۷۷	۸۷۶	۸۶۷	-۰/۱۱	-۱/۱۴
۴۰	۸۸۱	۸۷۲	۸۶۴	-۱/۰۲	-۱/۹۳
۴۱	۸۸۴	۸۹۱	۹۲۹	۰/۷۹	۵/۰۹
۴۲	۸۸۶	۸۸۴	۸۹۲	-۰/۲۲	۰/۶۸
۴۳	۹۰۰	۸۹۸	۸۸۴	-۰/۲۲	-۱/۷۸
۴۴	۹۱۹	۹۳۷	۹۶۱	۱/۹۶	۴/۵۷
۴۵	۹۲۲	۹۲۹	۹۳۳	۰/۷۶	۱/۱۹
۴۶	۹۴۳	۹۴۳	۹۵۰	-۰/۰۲	۰/۷۴
۴۷	۹۵۰	۹۶۹	۹۴۳	۲/۰۰	-۰/۷۴
۴۸	۹۵۷	۹۶۷	۹۶۷	۱/۰۴	۱/۰۴
۴۹	۹۶۱	۹۵۴	۹۶۰	-۰/۷۳	-۰/۱۰
۵۰	۹۶۰	۹۵۸	۹۶۰	-۰/۲۱	۰/۰۲
۵۱	۹۶۵	۹۶۸	۹۷۱	۰/۳۱	۰/۶۲
۵۲	۹۶۹	۹۶۸	۹۵۲	-۰/۱۰	-۱/۷۵
۵۳	۹۷۳	۹۶۱	۹۶۱	-۱/۲۳	-۱/۲۳
۵۴	۹۷۵	۹۶۸	۹۶۵	-۰/۷۲	-۱/۰۲
۵۵	۹۷۶	۹۹۳	۹۸۰	۱/۷۴	۰/۴۱
۵۶	۹۹۰	۹۸۷	۹۸۹	-۰/۳۰	-۰/۱۰
۵۷	۱۰۰۰	۹۹۹	۹۶۶	-۰/۱۰	-۳/۴
۵۸	۱۰۱۹	۱۰۱۲	۱۰۰۱	-۰/۶۹	-۱/۷۷
۵۹	۱۰۴۶	۱۰۳۹	۱۰۴۰	-۰/۶۷	-۰/۵۷
۶۰	۱۰۵۴	۱۰۳۲	۱۰۳۰	-۲/۰۹	-۲/۲۸

ضرایب تعیین بدست آمده از رسم مقادیر پیش‌بینی شده شاخص بازداری و مقادیر تجربی در شکل‌های (۳-۱۳) و (۳-۱۴)، بر نزدیکی مقادیر پیش‌بینی شده توسط مدل‌ها نسبت به مقادیر تجربی

آن‌ها دلالت دارد. همچنین در شکل های (۳-۱۵) و (۳-۱۶) مقادیر باقیمانده‌ها بر حسب مقادیر تجربی RI ترکیبات مورد بحث، ترسیم شده است که توزیع متقارن داده‌ها حول محور افقی (خطای صفر) عدم وجود خطای سیستماتیک را نشان می‌دهد.

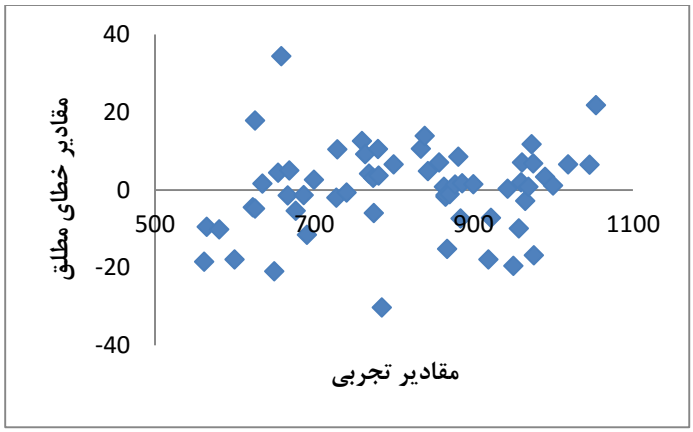


شکل (۳-۱۳) - نمودار پیش بینی شده بر حسب مقادیر تجربی RI به روش رد مرحله‌ای تک‌تک برای کل داده‌ها با مدل SR-ANN

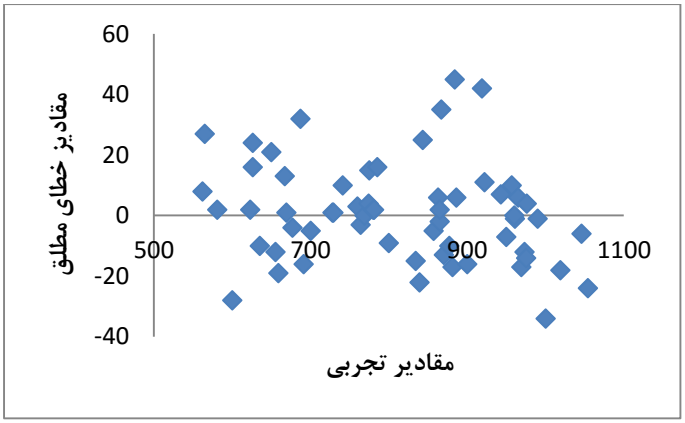


شکل (۳-۱۴)- نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی RI به روش رد مرحله‌ای تک‌تک برای کل داده‌ها با مدلGA-ANN

نمودار باقیمانده‌ها معیاری برای شایستگی مدل به دست آمده، می‌باشد و اگر باقیمانده‌ها به طور یکنواخت حول محور افقی پراکنده باشند، نشان می‌دهد که مدل مناسبی به دست آمده و هیچ خطای سیستماتیکی وجود ندارد.



شکل (۳ - ۱۵)- نمودار باقیمانده ها بر حسب مقدار تجربی RI برای کل داده ها توسط مدل SR-ANN



شکل (۳ -۱۶)- نمودار باقیمانده ها بر حسب مقدار تجربی RI برای کل داده ها توسط مدلGA-ANN

۳-۶-۳- ارزیابی مدل‌های برتر SR-ANN و GA-ANN با استفادهاز پارامترهای آماری

مطابق جدول (۳-۱۴)، هفت پارامتر آماری که توضیح آن‌ها در بخش (۲-۱۱) آمده‌است، جهت ارزیابی توانایی پیش‌بینی مدل‌های ارائه شده با روش SR-ANN وGA-ANN به کار گرفته شد. مقادیر این پارامترها نشان می‌دهد که هر دو مدل ارائه شده از قدرت پیش‌بینی مناسبی برخوردار

هستند ولی در عین حال نتایج حاصل از مدل SR–ANN به طور نسبی مقادیر بهتری دارد و این مدل توانایی پیش بینی نسبتا بالاتری نسبت به مدل GA-ANN دارد.

جدول (۳-۱۴)- پارامترهای آماری به دست آمده برای مدل های SR-ANN و GA-ANN

پارامتر	سری تست (N=12)		سری ارزیابی(N=12)		کل داده ها (N=60)	
	SR-ANN	GA-ANN	SR-ANN	GA-ANN	SR-ANN	GA-ANN
MAE	۷/۲۵	۸/۵۰	۴/۵۱	۵/۹۲	۸/۲۰	۱۲/۳۲
MSE	۹۷/۹۲	۱۲۲	۳۱/۳۸	۶۱/۲۵	۱۲۱/۴۰	۲۶۷/۰۲
PRESS	۱۱۷۵/۰۴	۱۴۶۴	۳۷۶/۵۶	۷۳۵	۷۲۸۴	۱۶۰۲۱/۲
SEP	۹/۸۹	۱۱/۰۴	۵/۶۰	۷/۸۳	۱۱/۰۲	۱۶/۳۴
REP(%)	۱/۲۲	۱/۳۶	۰/۶۹	۰/۹۶	۱/۳۶	۲/۰۱
MRE	۱/۰۰	۱/۰۹	۰/۵۵	۰/۷۵	۱/۰۷	۱/۵۶
R^2(Leave One Out)					۰/۹۹۳	۰/۹۸۵

نتایج به دست آمده از ارزیابی‌های فوق، همگی بیانگر این می‌باشند که هر دو مدل، از قدرت پیش-بینی خوبی برخوردار هستند، ولی در عین حال مدل SR-ANN با توجه نتایج آماری بهتر، می‌تواند مدل مناسب‌تری باشد.

۳-۶-۴- ارزیابی مدل ارائه شده توسط شبکه عصبی با استفاده از آزمون Y-تصادفی[1]

این تکنیک ارزیابی مدل با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام شد. در این آزمون مقادیر تجربی متغیر وابسته در محدوده‌ی همان مقادیر به طور تصادفی تغییر داده شدند و

1- Y- randomization test

مدل‌های جدیدی با استفاده از این مقادیر تصادفی ایجاد گردید. تفاوت قابل توجه بین مقادیر ضریب تعیین مدل اصلی و مدل QSPR که با پاسخ‌های تصادفی توسعه یافته، بیانگر عدم وجود ارتباط تصادفی در مدل اصلی است. نتایج حاصل از ۱۰ بار اجرای آزمون Y- تصادفی در جداول (۳-۱۵) و (۳-۱۶) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) در مدل اصلی و مدل‌های جدید، بیانگر عدم وجود ارتباط تصادفی در مدل اصلی است.

جدول (۳-۱۵)-مقادیرR2 برای سری ارزیابی و تست با استفاده از آزمونY- تصادفی روش SR-ANN

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2 (ارزیابی)	۰/۰۴۷	۰/۲۹۳	۰/۰۰۹	۰/۰۰۴	۰/۰۷۸	۰/۱۲۵	۰/۰۵۵	۰/۰۸۹	۰/۱۷۴	۰/۳۲۶
R^2 (تست)	۰/۱۱۹	۰/۱۴۶	۰/۰۰۶	۰/۰۳۴	۰/۰۱۵	۰/۰۸۲	۰/۰۲۸	۰/۰۱۷	۰/۰۷۵	۰/۱۰۴

جدول (۳-۱۶)-مقادیرR^2 برای سری ارزیابی و تست با استفاده از آزمونY- تصادفی روش GA-ANN

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2 (ارزیابی)	۰/۲۵۶	۰/۱۹۰	۰/۰۰۱	۰/۱۵۸	۰/۳۲۷	۰/۱۱۰	۰/۰۶۵	۰/۰۰۱	۰/۱۱۹	۰/۰۱۸
R^2 (تست)	۰/۰۸۷	۰/۰۹۰	۰/۰۷۱	۰/۰۰۸	۰/۰۵	۰/۰۲۳	۰/۰۰۲	۰/۰۳۹	۰/۰۰۵	۰/۰۸۴

۳-۷- بررسی ارتباط توصیف‌کننده‌های منتخب با اندیس بازداری ترکیبات

در این بخش به طور خلاصه ارتباط بین توصیف کننده‌های وارد شده در مدل و شاخص بازداری ترکیبات، مورد بررسی قرار خواهد گرفت. با توجه به نتایج به دست آمده در مدل SR-ANN بعنوان مدل برتر، ۵ توصیف‌کننده انتخاب شده که هر کدام بیانگر خصوصیات متفاوتی از مولکول‌های مورد بررسی است. این توصیف کننده‌ها شامل G1, Mor09m , RDF035v, GATS3v, RDF045u

می‌باشد. جدول زیر مقادیر اثر متوسط[1] هر توصیف‌کننده را نشان می‌دهد. اثر متوسط یک متغیر مستقل با استفاده از فرمول (۳-۱) به دست می‌آید:

$$t = \frac{\beta_x \times \sum x_n}{\sum RI} \qquad (۱-۳)$$

که در آن، β_x ضریب متغیر مستقل x در مدل، x_n مقدار متغیر مستقل x مورد نظر برای ترکیب nام و RI نیز شاخص بازداری تجربی می‌باشد.

جدول (۳-۱۷)–اثر متوسط توصیف‌کننده‌ها

توصیف‌کننده	G1	Mor09m	RDF035v	GATS3v	RDF045u
اثر متوسط	۶۶/۵۸۴	-۱۴/۷۶۹	-۹/۵۹۴	-۹/۰۱۵	-۴/۹۰۴

۳-۷-۱- توصیف کننده های هندسی(Geometrical)

G1 یکی از این دسته توصیف‌کننده‌ها می‌باشد که در مدل ظاهر شده این توصیف‌کننده از رابطه زیر محاسبه می‌شود، که m_i و m_j وزن اتمی اتم‌های i و j بوده و r_{ij} فاصله بین دو اتم می‌باشد[۴۸].

$$G1 = \sum \frac{m_i m_j}{{r_{ij}}^2} \qquad (۲-۳)$$

با توجه به این رابطه، توصیف‌کننده G1 به وزن اتمی وابسته بوده که با افزایش وزن اتمی مقدار آن افزایش می‌یابد. از طرفی اثر این توصیف کننده بر روی شاخص بازداری مثبت (۶۶/۵۸۴) می‌باشد که با افزایش آن میزان شاخص بازداری افزایش می‌یابد. بنابراین می‌توان گفت، یکی از عوامل موثر بر روی شاخص بازداری ترکیبات مورد نظر، وزن اتمی می‌باشد یعنی ترکیبات با تعداد کربن بیشتر، شاخص بازداری بزرگتری دارند.

1-Mean effect

۳-۷-۲- توصیف‌کننده‌های 3D MORSE[1]

توصیفگرهای 3D - Morse (نمایش سه بعدی ساختار مولکول براساس تفرق الکترون) از طریق معادله تبدیلی که در پراش الکترون استفاده می‌شود، محاسبه می‌گردند:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (۳-۳)$$

که در این رابطه I شدت الکترون پراکنده شده، A_i و A_j خاصیت اتمی اتم i و j، s زاویه پراکندگی، r_{ij} فاصله بین اتم‌های i و j و N نیز تعداد کل اتم‌هارا نشان می‌دهد. این روش باعث می‌شود که ساختار سه بعدی مولکول به یک کد ثابت تبدیل شود[۴۹].

توصیف‌کننده Mor09m از این گروه توصیف‌کننده‌ها است که وارد مدل برتر شده است. این توصیف‌کننده آرایش سه بعدی اتم‌ها را که با جرم اتمی وزن دار شده بیان می‌کند. با توجه به سری داده‌ها تغییر آرایش مولکول از حالت نرمال به شاخه‌دار شدن، باعث کاهش شاخص بازداری مولکول شده که توصیف‌کننده Mor09m بیانگر چگونگی این حالت می‌باشد. بنابراین می‌توان تأثیر منفی این توصیف‌کننده را بر روی شاخص بازداری توجیه کرد (۱۴/۷۶۹-) که با بزرگ شدن این توصیف‌کننده شاخص بازداری کاهش می‌یابد.

جدول (۳-۱۸) چگونگی تغییرات مقدار توصیف‌کننده Mor09m و شاخص بازداری را به عنوان مثالبرای سه مولکول از سری داده‌ها نمایش می‌دهد. همانطور که مشاهده می‌شود، برای سه مولکول با تعداد اتم‌های یکسان، تغییر آرایش مولکول از حالت نرمال به شاخه‌دار شدن، باعث افزایش این توصیف‌کننده و کاهش شاخص بازداری می‌شود.

1-3D- Molecular representation of structure based on electron diffractio

جدول (۳-۱۸)- مثال‌هایی از اثر توصیف‌کننده Mor09m بر شاخص بازداری

نام ترکیب	مقدارMor09m	شاخص بازداری
n-octane	-۰/۰۰۹	۸۰۰
2,2-dimethyl hexane	۰/۰۶۵	۷۲۹
2,2,4-trimethyl pentane	۰/۲۸۴	۶۹۱

۳-۷-۳- توصیف‌کننده‌های RDF[1]

می‌توان تابع توزیع شعاعی یک ترکیب شاملN اتم را به احتمال یافتن یک اتم در یک حجم کروی با شعاعr تعریف کرد و فرم کلی این تابع نیز به صورت زیر است:

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{\beta (r - r_{ij})^2} \qquad (۴-۳)$$

که در این رابطه N تعداد اتم‌ها، A_i و A_j خاصیت اتمی (الکترونگاتیویته، جرم اتمی، قطبش پذیری اتمی و...)، r_{ij} شعاع بین اتم i و j، f فاکتور مقیاس و β نیز فاکتور دما می‌باشد که β باعث حرکت و جابجایی اتم‌ها می‌شود[۵۰].

RDF می‌تواند در موارد متفاوتی به اقتضای نیازمندی‌های اطلاعات ارائه شده به کار برده شود. این ویژگی‌های اتمی قادرند میان اتم‌های یک مولکول، برای تقریبا هر خصلتی که می‌تواند به یک اتم مربوط باشد تمایز قائل شوند. تابع توزیع شعاعی به این شکل، با تمامی نیازها برای یک توصیف‌کننده ساختار سه‌بعدی روبرو می‌شود، این تابع مستقل از تعداد اتم‌ها یا به عبارت دیگر اندازه مولکول است، و با توجه به آرایش سه‌بعدی اتم‌ها تابعی منحصر به فرد است، ولی در برابر حرکات تبدیلی و چرخش کل مولکول نامتغیر است. کد RDF می‌تواند در انواع خاصی از اتم‌ها یا بردهایی از فاصله محدود شود تا اطلاعات مشخصی را درباره یک ساختار فضایی سه‌بعدی معین به دست می‌دهد. توصیف‌کننده‌های RDF با به کارگیری مجموعه قوانین ساده قابل تفسیرند و از این رو امکان تبدیل کد را به ساختار سه

1 - Radial Distribution Function

بعدی متناظر فراهم می‌کنند. علاوه بر فراهم‌سازی اطلاعات درباره فواصل بین اتمی در کل مولکول، کد RDF اطلاعات ارزشمند دیگری را نیز مانند فاصله‌ی پیوند، نوع حلقه‌ها، سیستم‌های مسطح یا غیر مسطح و انواع اتم‌ها به دست می‌دهد[۵۱]. RDF035v و RDF045u از جمله توصیف‌کننده‌های این دسته هستند که در مدل برتر ظاهر شده‌اند. توصیف‌کننده RDF035v با حجم‌های واندوالس اتمی وزن دار شده که با توجه به رابطه‌ی فوق هر چه حجم اتمی زیاد شود مقدار این توصیف‌کننده نیز افزایش می‌یابد و از طرفی علامت منفی این توصیف‌کننده (۹/۵۹۴-) در مدل نشان‌دهنده اثر منفی این توصیف‌کننده برروی شاخص بازداری می‌باشد. توصیف‌کننده RDF045u وزن دار نشده و با توجه به اثر منفی آن بر روی شاخص بازداری (۴/۹۰۴-)، افزایش آن باعث کاهش شاخص بازداری می‌شود. این رابطه عکس بین این توصیف‌کننده‌ها و شاخص بازداری را، با توجه به توضیحات فوق، می‌توان اینگونه توجیه کرد که چگونگی آرایش اتم‌ها و وضعیت قرارگیری شاخه‌ها بر روی اتم‌ها در مولکول بر روی مقدار شاخص بازداری تأثیر می‌گذارد. در جدول (۳-۱۹) نمونه‌هایی از این اثرات نشان داده شده است.

جدول(۳-۱۹)-مثال‌هایی از اثر توصیف‌کننده RDF035v و RDF045u بر شاخص بازداری

نام ترکیب	مقدارRDF035v	مقدارRDF045u	شاخص بازداری
n-octane	۰/۵۴۷	۱/۳۲۴	۸۰۰
2,5-dimethyl hexane	۳/۰۹۰	۷/۹۹۳	۷۴۱
2,2-dimethyl hexane	۳/۵۶۸	۸/۳۱۹	۷۲۹

۳-۷-۴- توصیف‌کننده‌های 2D-autocorrelation

توصیفگرهای دوبعدی خودارتباطی از گراف مولکولی و از طریق محاسبه مجموع اوزان اتم‌های انتهایی کل مسیرها با طول مسیر مورد نظر (lag)، به دست می‌آیند. نوعی از این توصیفگرها گروه Autocorrelation Geary است که ضریب گری نام دارد و بدین صورت محاسبه می‌شود:

$$C(d) = \frac{\frac{1}{2\Delta}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}.(w_i - w_j)^2}{\frac{1}{A-1}.\sum_{i=1}^{A}(w_i - \overline{w})^2} \qquad (۵-۳)$$

که w یک ویژگی اتم، $\overline{w}$ میانگین مقدار آن ویژگی در مولکول، A تعداد اتم‌ها و d فاصله‌ی توپولوژیکی است و δ_{ij} نیز که به تابع کرونکر[1] معروف است در حالتی که $d = d_{ij}$ باشد برابر با یک است و در غیر اینصورت صفر بوده و همچنین Δ هم مجموعه‌ی δ هاست. مقدار این توصیف‌کننده که از جنس فاصله است، از صفر تا بی‌نهایت متغیر است[۵۲]. از میان این توصیفگرها GATS3v که با حجم واندروالس وزن‌دار شده در مدل انتخاب شده و اثر آن روی شاخص بازداری منفی و برابر (۹/۰۱۵-) می‌باشد. بدین معنی که افزایش مقادیر این توصیف‌کننده باعث کاهش شاخص بازداری می‌شود.

۳-۸- بررسی میزان مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی

میزان مشارکت توصیفگرهای منتخب به صورت زیر تعیین شد:

۱- توصیفگر مورد نظر به همراه اوزان مربوطه اش از شبکه بهینه شده حذف گردید.

۲- با استفاده از بقیه توصیفگرها مقدار متغیر وابسته (RI) برای هر ترکیب سری آموزش به روش ارزیابی تقاطعی پیش‌بینی شد.

1 - Kronecker

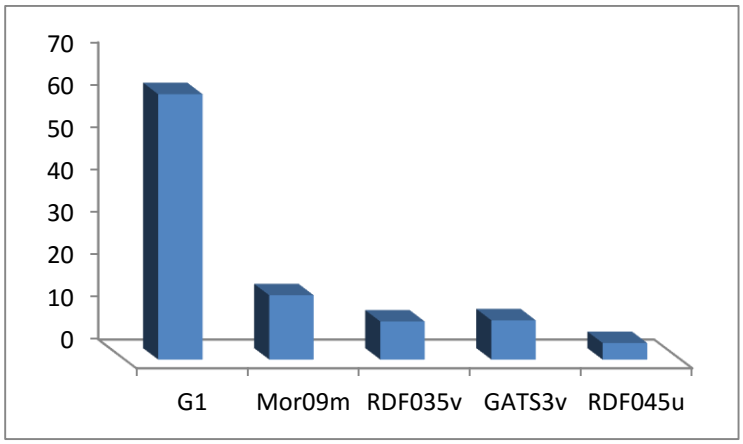
۳- میانگین خطای مطلق حاصل از ارزیابی تقاطعی ترکیبات سری آموزش محاسبه گردید.

۴- مراحل ۱تا ۳ برای دیگر توصیفگرهای منتخب نیز تکرار شد.

۵- سرانجام درصد مشارکت هر توصیفگر توسط رابطه (۳-۴) برآورد شد[۵۵-۵۳].

$$c_i = 100 \frac{\Delta m_i}{\sum_{i=1}^{N} \Delta m_i} \qquad (۶-۳)$$

که در این رابطه c_i درصد مشارکت توصیفگر حذف شده iام، N تعداد توصیفگرهای مدل و Δm_i میانگین خطای مطلق حاصل از ارزیابی تقاطعی سری آموزش در غیاب توصیفگر iام را نشان می‌دهد. که بر این اساس درصد مشارکت توصیفگرهای منتخب در ترکیبات مورد بررسی به شکل زیر می‌باشد.



شکل(۳-۱۷)- مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینه

طبق شکل فوق توصیف‌کننده G1 دارای بیشترین اثر مشارکت می‌باشد. این توصیف‌کننده جزء دسته توصیفگرهای هندسی می‌باشد که با وزن اتمی رابطه مستقیم دارد بدین معنی که با بزرگ شدن مولکول شاخص بازداری افزایش می‌یابد.

۳-۹ - نتیجه‌گیری

در این تحقیق از مدل‌های SR-ANN و GA-ANN برای پیش بینی اندیس بازداری یک سری ترکیبات آلی فرار استفاده شده است، بر اساس این مدل‌ها می‌توان گروه‌های مختلف از توصیف کننده‌ها و در نتیجه خواص مختلف ساختارهای مولکول‌ها را به طور ویژه روی مکانیسم بازداری مورد بررسی قرار داد و از آنجا که اندازه گیری این ترکیبات به صورت تجربی و آزمایشگاهی وقت‌گیر و هزینه بر است، بنابراین با صرف زمان و هزینه کمتر می‌توان با استفاده از مدل معتبر شاخص بازداری را برای ترکیبات مشابه از این دسته و یا ترکیباتی که هنوز سنتز نشده‌اند، پیش بینی نمود.

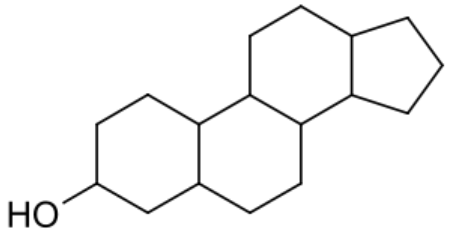
فصل چهارم

مدل‌سازی QSPR شاخص بازداری (RI) تعدادی از استرول‌ها با استفاده از شبکه عصبی و انتخاب متغیر با کمک الگوریتم ژنتیک

۴-۱- استرول‌ها[1]

استرول‌ها به عنوان الکل‌های استروئیدی شناخته شده‌اند که زیر گروه استروئیدها می‌باشند. این ترکیبات دسته‌ی مهمی از ترکیبات آلی می‌باشند. که به طور طبیعی در گیاهان، جانوران و قارچ‌ها یافت می‌شوند. مهمترین نوع جانوری آن کلسترول بوده که یک عامل حیاتی در سلول‌ها می‌باشد و عاملی پیشرو برای ویتامین‌های محلول در چربی و هورمون‌های استروئیدی می‌باشد.

استرول‌های گیاهی فیتواسترول[2] نامیده می‌شوند و استرول‌های حیوانی, زئواسترول[3] نام دارند. از فیتواسترول‌های مهم کمپسترول[4]، سیتواسترول[5] و استیگماسترول[6] می‌باشند. ارجسترول[7] نوعی استرول است که در غشای سلولی قارچ وجود دارد که نقشی مشابه کلسترول در سلول‌های حیوانی دارد. استرول‌ها با گروه هیدروکسیل در موقعیت ۳ در حلقه A، لیپیدهای آمفی‌پاتیک[8]بوده که برروی مولکول چربی‌دوست، گروه هیدروکسیل قطبی در حلقه A قرار دارد. این ترکیبات از استیل- کوآنزیم A[9] از طریق HMG-CoA reductase سنتز می‌شوند[۵۶]. شکل (۴-۱) ساختار یک استرول را نمایش می‌دهد.



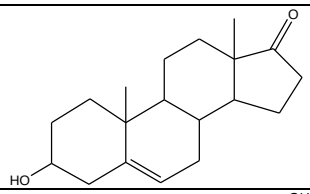
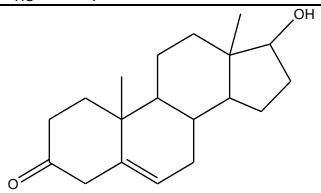
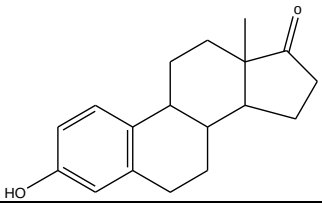
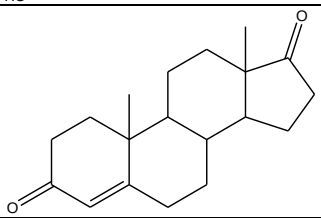
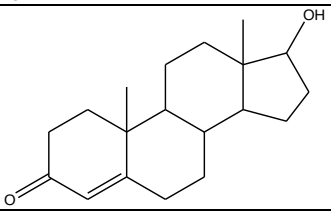
شکل (۴-۱)- ساختار استرول

1- Sterols
2- Phytosterols
3-Zoosterols
4-Campesterol
5-Sitosterol
6- Stigmasterol
7- Ergesterol
8- Amphipathic Lipids
9- Acetyl-Coenzyme A

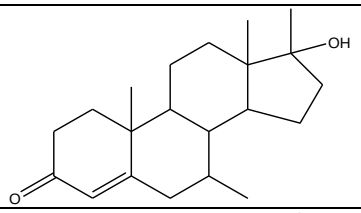
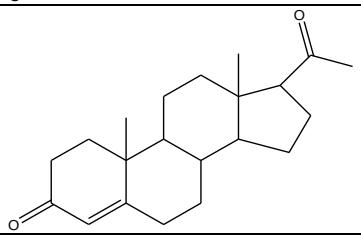
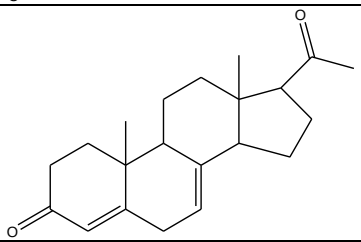
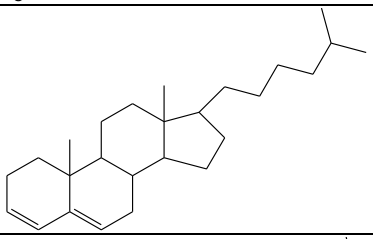
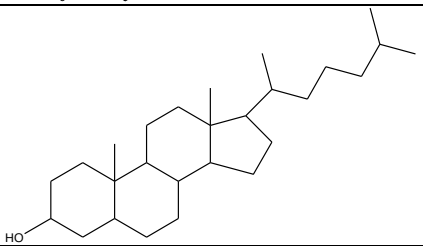
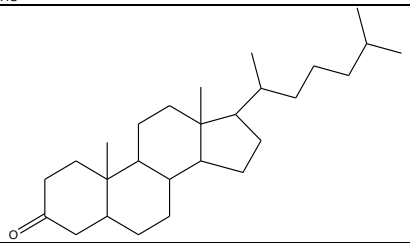
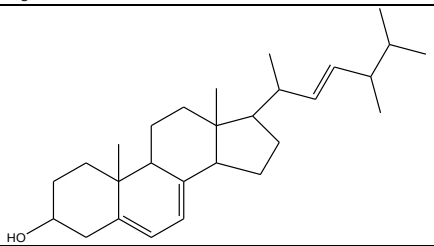
۴-۲- سری داده‌ها

در ابتدا شاخص بازداری (RI) دسته‌ای از استرول‌ها، به عنوان سری داده‌ها مورد استفاده قرار گرفت. آنالیز این ترکیبات با استفاده از GC-MS و GC-FID انجام گرفت[۵۷]. نام، ساختمان و شاخص بازداری این ترکیبات در جدول (۴-۱) آمده است.

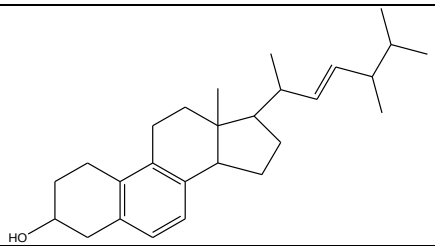
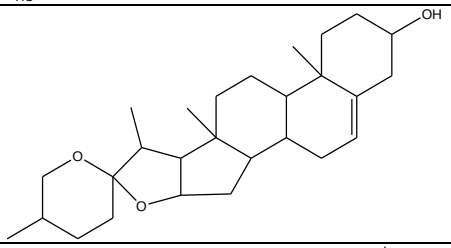
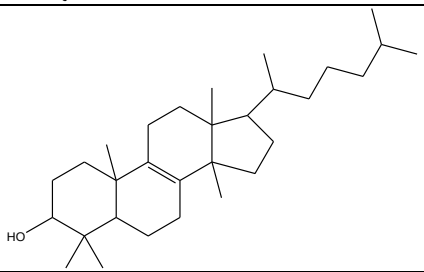
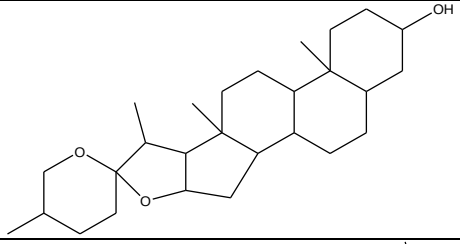
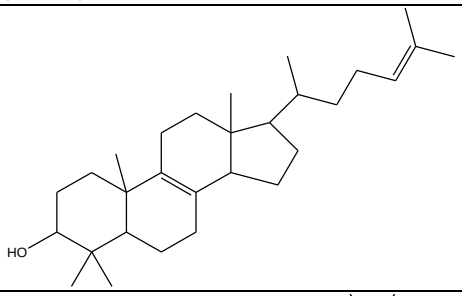
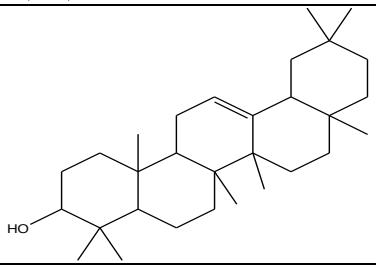
جدول ۴-۱- مولکول‌های سری داده‌ها همراه با نام، ساختار و شاخص بازداری آن‌ها

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۱	Prasterone		۲۵۱۳	tr
۲	Androst-5-en-17 β -ol-3-one		۲۵۵۵	v
۳	Estrone		۲۶۱۴	tr
۴	Androst-4-ene-3,17-dione		۲۶۳۱	te
۵	Testosterone		۲۶۷۸	tr

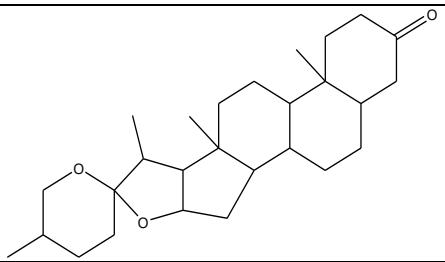
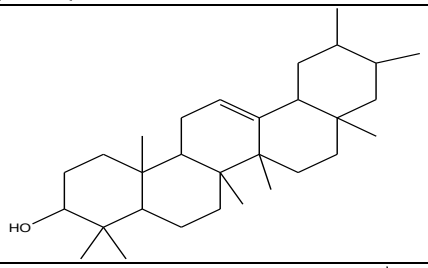
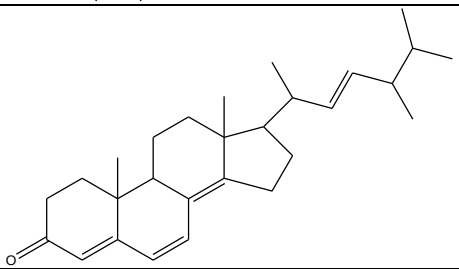
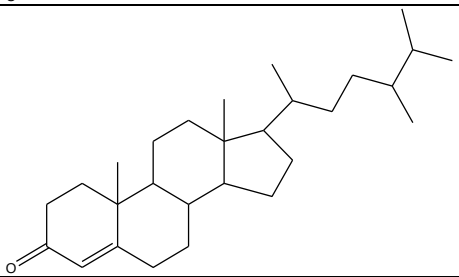
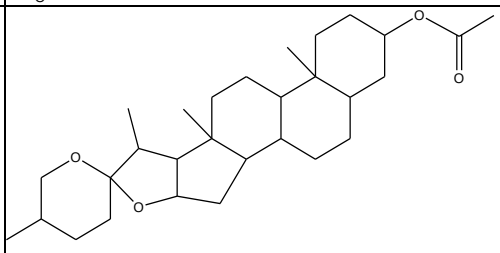
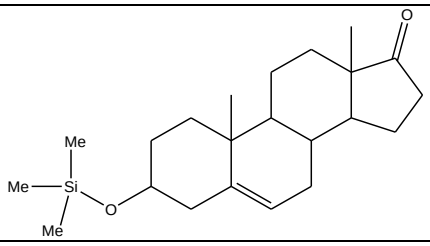
ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۶	17-epi-Bolasterone		۲۷۲۲	tr
۷	Progesterone		۲۸۴۳	v
۸	Pregna-4,7-diene-3,20-dione		۲۸۴۹	te
۹	Cholesta-3,5-diene		۲۸۷۴	tr
۱۰	Cholestan-3 α -ol		۳۰۵۸	tr
۱۱	Coprostan-3-one		۳۰۷۸	v
۱۲	Ergosterol, (3 β)-ergosta-5,7,22-trien-3-ol		۳۱۵۲	tr

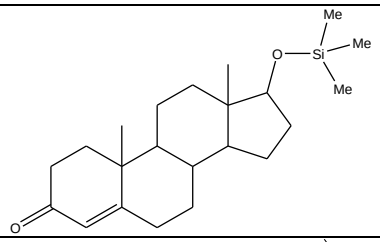
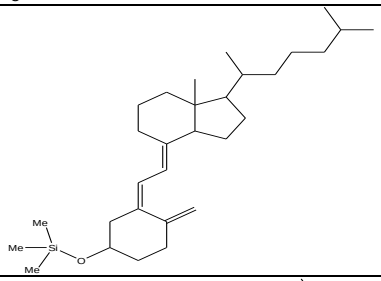
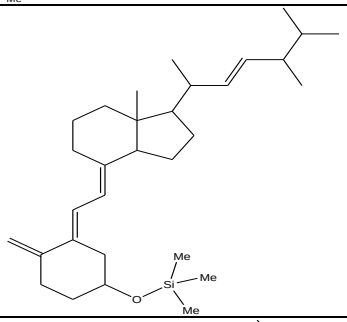
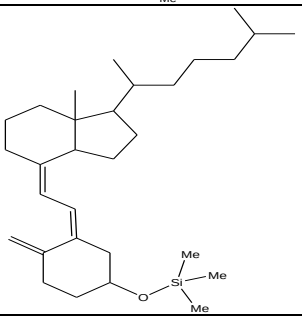
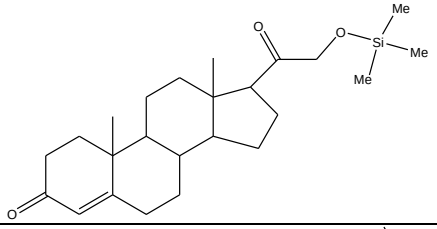
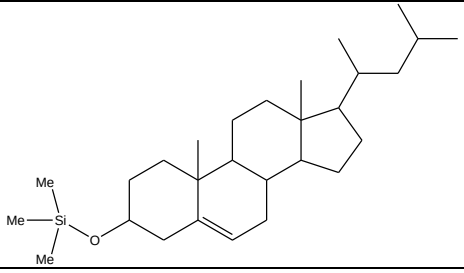
ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۱۳	Neoergosterol		۳۲۱۵	tr
۱۴	Diosgenin		۳۲۶۷	te
۱۵	Lanost-8-en-3-ol (dihydrolanosterol)		۳۲۷۵	tr
۱۶	Tigogenone		۳۲۷۷	v
۱۷	Lanosterol		۳۳۱۰	tr
۱۸	β-Amyrine		۳۳۱۲	tr

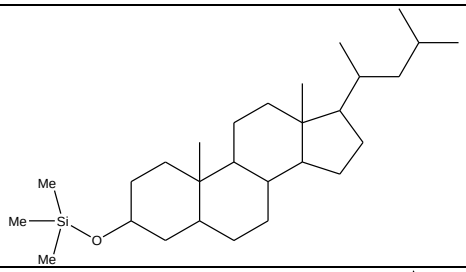
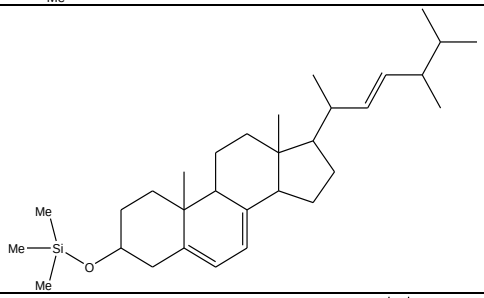
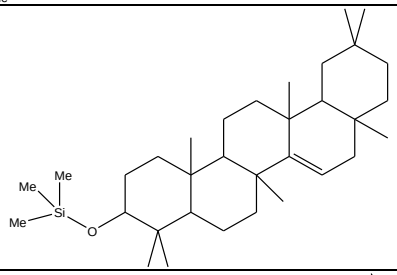
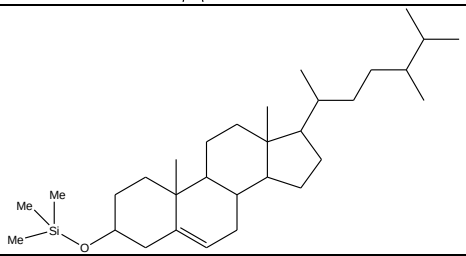
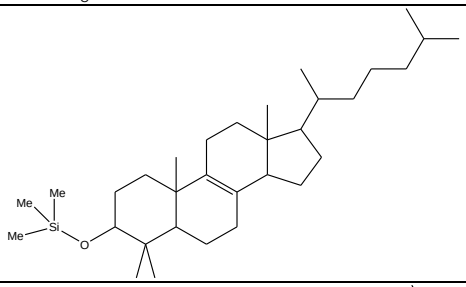
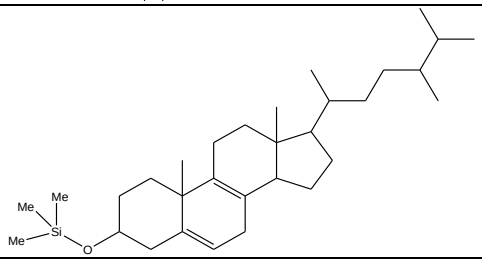
ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۱۹	Sarsasagenone		۳۳۲۱	v
۲۰	α-Amyrine		۳۳۵۴	te
۲۱	Ergosta-4,6,8(14),22-tetraen-3-one		۳۳۶۳	tr
۲۲	Sitostenone		۳۴۳۵	te
۲۳	Sarsasagenin acetate		۳۴۹۵	tr
۲۴	Dehydroepiandrosterone		۲۵۸۴	tr

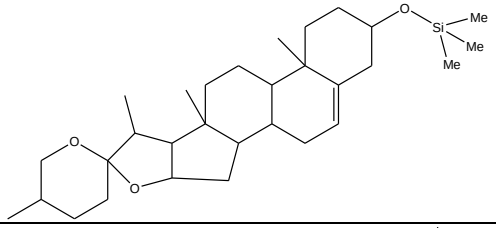
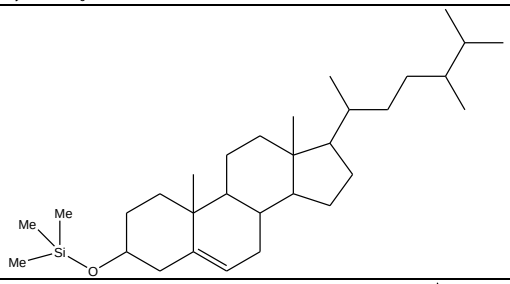
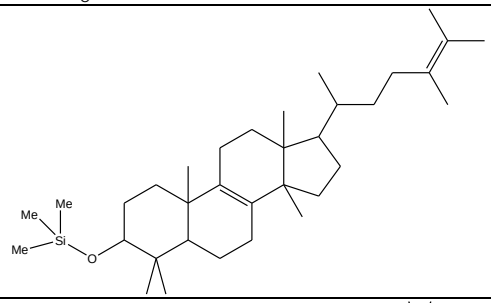
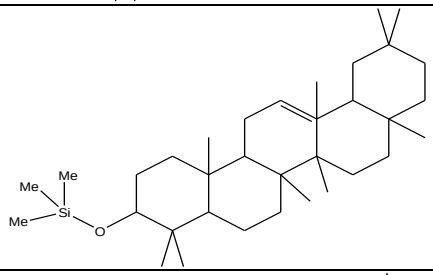
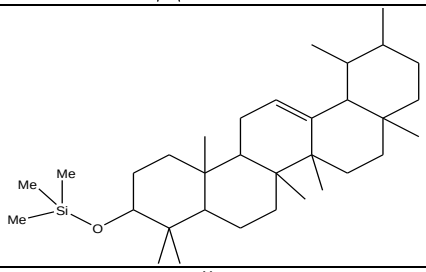
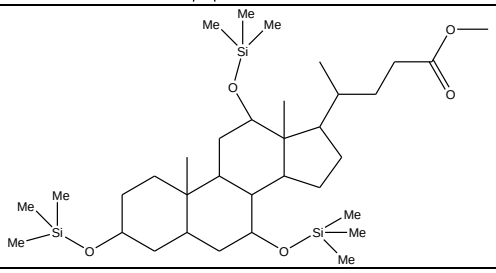
ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۲۵	Testosterone		۲۶۷۸	v
۲۶	Cholecalciferol 1		۲۹۸۲	tr
۲۷	Ergocalciferol 1		۳۰۲۴	tr
۲۸	Cholecalciferol 2		۳۱۲۹	tr
۲۹	Deoxycorticosterone		۳۱۴۷	v
۳۰	Cholesterol		۳۱۴۷	te

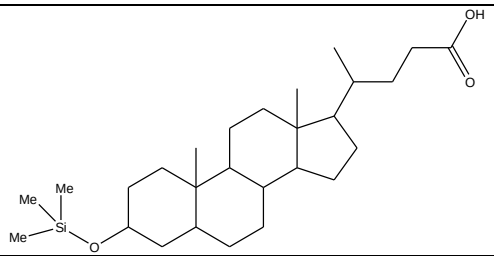
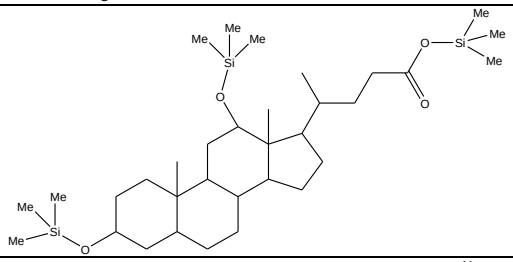
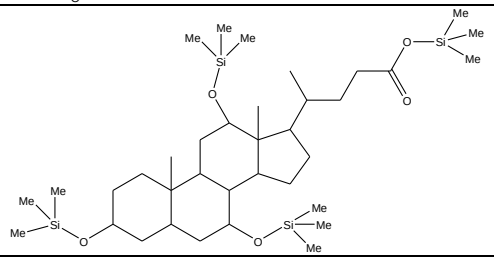
ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۳۱	Cholestanol (3-hydroxycholestane)		۳۱۵۴	tr
۳۲	Ergosterol		۳۲۳۲	v
۳۳	Taraxerol		۳۲۴۷	tr
۳۴	Campesterol		۳۲۵۱	te
۳۵	Dihydrolanosterol, mono-TMS		۳۲۹۱	v
۳۶	Ergosta-5,8-dien-3 β -ol		۳۲۹۳	tr

ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۳۷	Diosgenin		۳۳۰۵	tr
۳۸	γ-Ergostenol		۳۳۰۹	tr
۳۹	Lanosterol		۳۳۳۱	te
۴۰	β-Amyrine		۳۳۵۰	tr
۴۱	α-Amyrine		۳۳۷۸	v
۴۲	5 β -Cholan-24-oic acid, 3 α, 7 α,12 α -tri-hydroxy-, methyl ester		۳۲۴۱	tr

ادامه جدول (۴-۱)

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری*
۴۳	Litocholic acid		۳۲۶۹	te
۴۴	Desoxycholic acid		۳۳۲۳	tr
۴۵	Cholic acid		۳۳۴۳	tr

*v:سری ارزیابی, te:سری تست, tr:سری آموزش.

۴-۳- رسم و بهینه‌سازی ساختار هندسی ترکیبات

در مرحله دوم، ساختمان مولکولی ترکیبات در نرم‌افزار HyperChem8.0 رسم و ساختار هندسی آن‌ها با محاسبه‌ی اتم‌های هیدروژن و به کمک روش نیمه تجربی AM1 بهینه شد. بهینه‌سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی[1] (بر حسب کیلوکالری بر مول)، به ۰/۰۰۱ برسد.

1- Root Mean Square Gradient of Energy

۴-۴- محاسبه توصیف‌کننده‌های مولکولی

ساختارهای بهینه شده، به عنوان ورودی به نرم‌افزار Dragon وارد شدند. محاسبه توصیف‌کننده‌ها توسط این نرم‌افزار صورت پذیرفت و حدود ۱۴۸۱ توصیف‌کننده از هیجده گروه مختلف برای هر ترکیب محاسبه گردید.

۴-۵- انتخاب توصیف‌کننده‌های مهم

توصیف‌کننده‌های به دست آمده از مرحله قبل، به عنوان متغیر مستقل، و اندیس بازداری(RI) به عنوان متغیر وابسته وارد نرم افزار SPSS شدند. سپس همانطور که در فصل قبل (بخش ۳-۳) توضیح داده شد، و با به کارگیری تکنیک‌های کاهش متغیر توصیف‌کننده‌های مناسب انتخاب شدند. بدین ترتیب در پایان این مرحله تعداد ۱۷۲ توصیف کننده باقی ماند.

سپس با دو روش برازش مرحله‌ای و الگوریتم ژنتیک، بهترین توصیف‌کننده‌ها انتخاب گردیدند.

۴-۵-۱- انتخاب بهترین توصیف‌کننده‌ها به روش برازش مرحله‌ای

همانطور که در فصل قبل (بخش ۳-۳-۱) توضیح داده شد، انتخاب توصیف‌کننده‌ها با این روش به این ترتیب است که با استفاده از منوی آنالیز در نرم‌افزار، گزینه رگرسیون خطی انتخاب شده و بعد از وارد کردن متغیرهای مستقل و وابسته، روش برازش مرحله‌ای (Stepwise) استفاده شد. در تحقیق حاضر، ۱۲ توصیف‌کننده انتخاب شدند که این ۱۲ توصیف‌کننده به همراه طبقه و معنی آن‌ها در جدول (۴-۲) آورده شده است. و همچنین در جدول (۴-۳) ماتریس همبستگی بین این توصیف‌کننده ها ارائه شده و این ماتریس عدم همبستگی بین این توصیف‌کننده‌ها را نشان می‌دهد.

جدول (۴-۲)- توصیف کننده های انتخابی توسط روش برازش مرحله ای

No	Symbol	Class	Meaning
1	MR	Property descriptors	Ghose-crippen molar refractivity
2	SEige	Eigenvalue-based indices	Eigenvalue sum from electronegativity weighted distance matrixe
3	E1p	WHIM descriptors	1st component shape directional WHIM index weighted by atomic polarizabilities
4	E1u	WHIM descriptors	1st component accessibility directional WHIM index unweighted
5	GATS2e	2D autocorrelations	Geary autocorrelation lag2 weighted by atomic sanderson electronegativities
6	FDI	Geometrical descriptors	Folding degree index
7	Mor09m	3D MoRSE descriptors	3D MoRSE – signal 09 weighted by atomic masses
8	R8e	GETAWAY descriptors	R autocorrelation of lag 8 weighted by atomic sanderson electronegativities
9	R7m	GETAWAY descriptors	R autocorrelation of lag 7 weighted by atomic masses
10	R2u	GETAWAY descriptors	R autocorrelation of lag 2 unweighted
11	MATS4m	2D autocorrelations	Moran autocorrelation lag 4 weighted by atomic masses
12	Dp	WHIM descriptors	D total accessibility index weighted by atomic polarizabilities

جدول (۴-۳)- ماتریس همبستگی کل توصیف کننده های انتخاب شده توسط روش برازش مرحله ای

	MR	SEige	E1p	E1u	GATS2e	FDI	Mor09m	R8e	R7m	R2u	MATS4m	Dp
MR	۱											
SEige	۰/۵۱۵	۱										
E1p	-۰/۳۰۶	۰/۳۸۴	۱									
E1u	-۰/۰۱۷	۰/۲۳۴	۰/۷۳۱	۱								
GATS2e	۰/۶۹۰	-۰/۶۸۹	۰/۳۰۸	-۰/۰۴۲	۱							
FDI	۰/۲۳۶	-۰/۶۲۶	-۰/۱۲۴	-۰/۰۸۷	-۰/۳۱۳	۱						
Mor09m	۰/۱۱۰	-۰/۳۷۱	۰/۰۶۶	-۰/۴۸۱	۰/۳۳۴	-۰/۱۲۹	۱					
R8e	-۰/۰۶۶	-۰/۱۰۵	-۰/۲۲۳	-۰/۳۰۹	-۰/۰۸۷	۰/۰۵۸	۰/۳۲۲	۱				
R7m	-۰/۰۶۱	۰/۰۱۳	-۰/۵۳۳	-۰/۵۵۵	-۰/۰۰۵	۰/۱۳۳	۰/۲۳۷	۰/۵۲۲	۱			
R2u	۰/۰۰۵	-۰/۰۰۳	۰/۰۹۹	۰/۰۷۹	۰/۰۳۹	۰/۱۴۲	-۰/۰۲۸	-۰/۲۱۳	-۰/۲۱۷	۱		
MATS4m	-۰/۴۵۴	۰/۱۲۰	-۰/۲۶۵	-۰/۵۳۸	-۰/۰۳۳	۰/۶۳۷	۰/۳۲۸	۰/۳۲۴	۰/۴۹۰	-۰/۰۱۶	۱	
Dp	۰/۶۰۸	-۰/۵۲۵	۰/۶۳۰	۰/۴۹۴	۰/۴۲۵	-۰/۳۸۱	۰/۱۴۵	-۰/۰۴۱	-۰/۲۱۵	۰/۱۶۴	-۰/۳۶۸	۱

۴-۵-۲- انتخاب توصیف‌کننده‌ها به روش الگوریتم ژنتیک

در این قسمت، برای انتخاب متغیر با روش الگوریتم ژنتیک، از جعبه ابزار GA-PLS در نرم‌افزار MATLAB استفاده شد. پارامترهای مهم در اجرای این برنامه جمعیت، عملگر پیوند و عملگر جهش می‌باشند که بنابرآنچه در جعبه ابزار GA-PLS از پیش تعیین شده بود انتخاب شدند. یعنی اندازه جمعیت عدد ۳۰ و برای عملگرهای پیوند و جهش به ترتیب مقادیر ۰/۵ و ۰/۰۱ درنظر گرفته شد. پس از مشخص شدن این مقادیر برنامه اجرا شد. به طور متوسط هر بار برای انتخاب متغیر ۵ مرتبه برنامه اجرا می‌شود. در جدول (۴-۴) توصیف‌کننده‌های انتخاب شده در هر بار اجرای الگوریتم ژنتیک گزارش شده است.

جدول (۴-۴)- توصیف کننده‌های انتخاب شده توسط الگوریتم ژنتیک

توصیف کننده‌های انتخاب شده توسط الگوریتم ژنتیک	تعداد تکرار
MR, IC5, Mor05m, Gs, RDF015u, RTV-A, R_1v-A	۱
RDF015u, MR, Gs, IC5, R_1v-A, DP20, Mor05m, R3u, R_2v-A, R_2e	۲
IC5, RDF015u, MR, Gs, R_2m	۳
MR, RDF015u, Gs, IC5, Mor05m, RTV-A, DP20, R_2m	۴
RDF015u, MR, Gs, IC5,HATSu, R_1v-A, DP20, R_2m	۵

سپس تمام متغیرهای انتخابی در ۵ بار، به عنوان ورودی جدید به GA-PLS داده شد و این بار متغیرهای انتخابی به عنوان متغیرهای برتر در نظرگرفته شده و برای مدلسازی به شبکه عصبی داده شد.در نهایت، از بین مجموع توصیف‌کننده‌های انتخاب شده در ۵ بار اجرای الگوریتم ژنتیک، ۸ توصیف‌کننده انتخاب شدند، عبارتند از:

R2m, Mor05m, IC5, Gs, R2e, R3u, DP20, R1v-A

۴-۵-۲-۱- توصیف‌کننده‌های انتخاب شده توسط روش الگوریتم ژنتیک

توصیف‌کننده‌های نهایی انتخاب شده با این روش در جدول (۴-۵) همراه با طبقه و معنی آن‌ها آورده شده است.

جدول (۴-۵)- توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک

No	Symbol	Class	Meaning
1	R2m	GETAWAY descriptors	R autocorrelation of lag 2 weighted by atomic masses
2	Mor05m	3D MORSE	3D –MORSE-signal 05/weighted by atomic masses
3	IC5	Information indices	Information content index neighborhooid symmetry of 5-order
4	Gs	WHIM descriptors	G total symmetry index weighted by atomic electrotopological states
5	R2e	GETAWAY descriptors	R autocorrelation of lag 2 weighted by atomic sanderson electronegativities
6	R3u	GETAWAY descriptors	R autocorrelation of lag3 unweighted
7	DP20	Randic molecular profiles	molecular profiles no.20
8	R1v-A	GETAWAY descriptors	R maximal autocorrelation of lag 1 weighted by atomic van der waals volumes

همچنین ماتریس همبستگی بین این توصیف کننده‌ها در جدول (۴-۶) ارائه شده که این ماتریس عدم همبستگی بین توصیف کننده‌ها را نشان می‌دهد.

جدول (۴-۶)- ماتریس همبستگی توصیف کننده های انتخاب شده توسط الگوریتم ژنتیک

	R2m	Mor05m	IC5	Gs	R2e	R3u	DP20	R_1v-A
R2m	۱							
Mor05m	۰/۴۴۳	۱						
IC5	-۰/۳۰۲	-۰/۵۸۸	۱					
Gs	۰/۵۶۱	۰/۴۰۴	۰/۵۱۵	۱				
R2e	-۰/۰۰۱	-۰/۱۸۱	۰/۱۵۵	۰/۲۰۴	۱			
R3u	۰/۴۷۳	۰/۲۶۵	۰/۰۰۳	۰/۶۲۳	۰/۵۶۹	۱		
DP20	-۰/۳۲۷	-۰/۱۰۰	۰/۵۰۳	-۰/۷۲۵	-۰/۱۸۶	-۰/۳۶۷	۱	
R_1v-A	-۰/۰۰۲	-۰/۰۹۴	۰/۳۰۵	-۰/۵۹۴	۰/۶۵۱	۰/۱۸۹	-۰/۰۵۰	۱

۴-۶- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌کننده‌های انتخاب شده توسط روش برازش مرحله‌ای (SR-ANN)

در این بخش نیز مانند قسمت ۳-۴ فصل ۳ در محیط MATLAB شبکه عصبی اجرا می‌شود. به این ترتیب که، مقدار عددی توصیف کننده‌های انتخاب شده توسط روش برازش مرحله‌ای به عنوان ورودی و مقادیر تجربی اندیس بازداری (RI) نیز به عنوان هدف به شبکه عصبی داده شد تا پاسخ شبکه با آن‌ها سنجیده شود. داده‌ها به طور تصادفی به سه سری آموزش (شامل ۲۵ ترکیب)، سری ارزیابی (شامل ۱۰ ترکیب) و سری تست (شامل ۱۰ ترکیب) تقسیم‌بندی شدند. برای به دست آوردن بهترین معادله و کمترین خطا، پارامترهای شبکه (تعداد متغیرهای ورودی شبکه، نوع توابع آموزش و تبدیل، تعداد گره‌ها در لایه پنهان، μ و تعداد دورهای آموزش بهینه‌سازی شدند.

۴-۶-۱- بهینه سازی

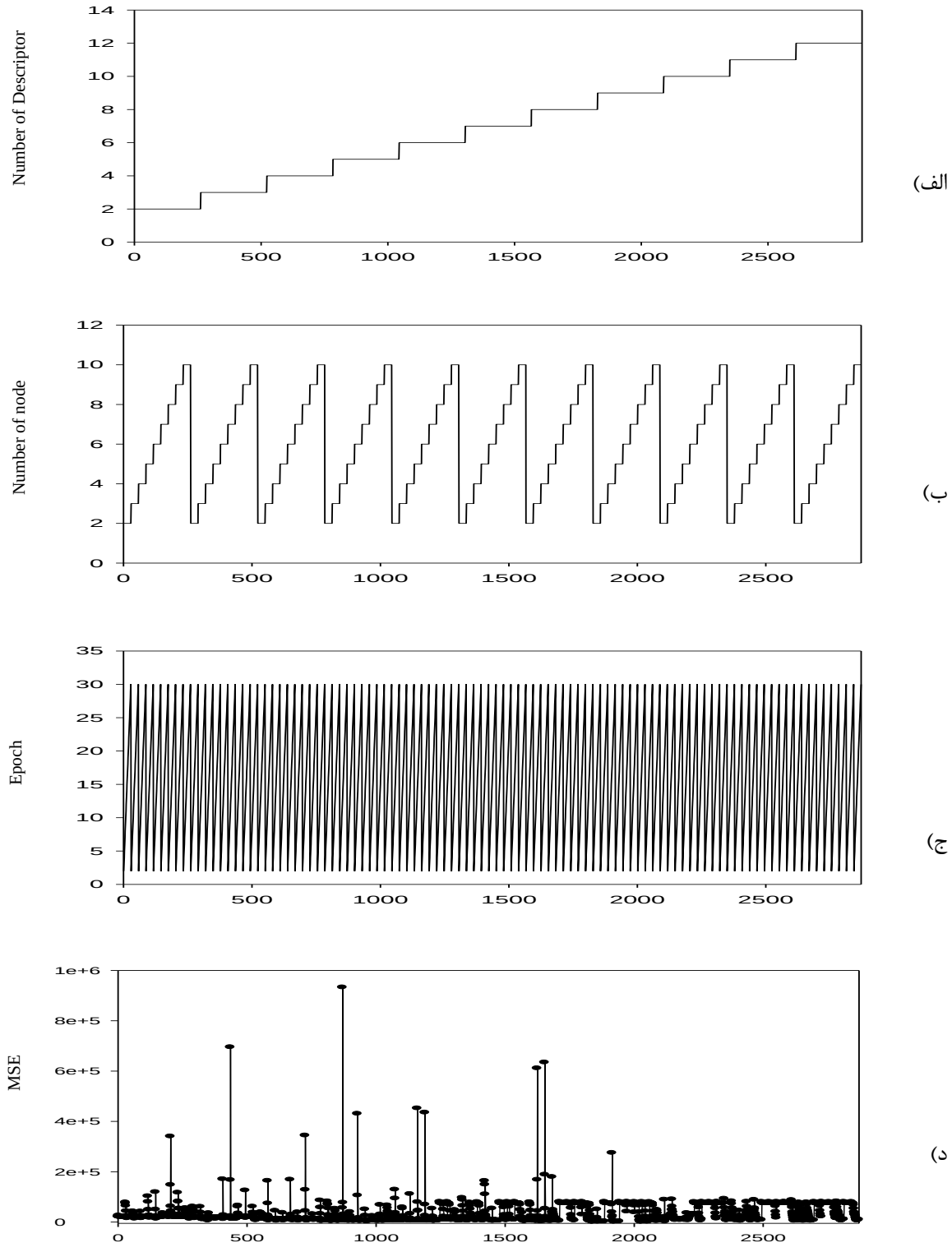
در این بخش نیز بهینه سازی وانتخاب بهترین تعداد متغیرهای ورودی شبکه، الگوریتم آموزشی، نوع تابع تبدیل، تعداد نرون‌های لایه پنهان وتعداد دورهای آموزش به طور همزمان انجام گرفت. برای بهینه سازی و شناسایی بهترین پارامترهای شبکه، شبکه‌هایی با ورودی های از ۲ تا ۱۲ توصیف کننده

ایجاد شدند که این ورودی‌ها، همان توصیف کننده‌های به دست آمده از روش انتخاب متغیر برازش مرحله‌ای می‌باشند. هر شبکه با دو الگوریتم آموزشی لونبرگ- مارکوارت (trainlm) و تنظیم بایزین (trainbr) و تعداد گره‌های از ۲ تا ۱۰ در لایه پنهان و تعداد متفاوت دور آموزش ۲ تا ۳۰، آموزش داده شد. برای به دست آوردن بهترین تابع تبدیل در لایه پنهان مدل‌های شبکه عصبی، از توابع لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) استفاده گردید. همچنین از تابع تبدیل خطی (purelin) در لایه خروجی استفاده گردید. نقطه‌ای که کمترین خطا را برای سری ارزیابی داشت به عنوان نقطه بهینه انتخاب گردید.

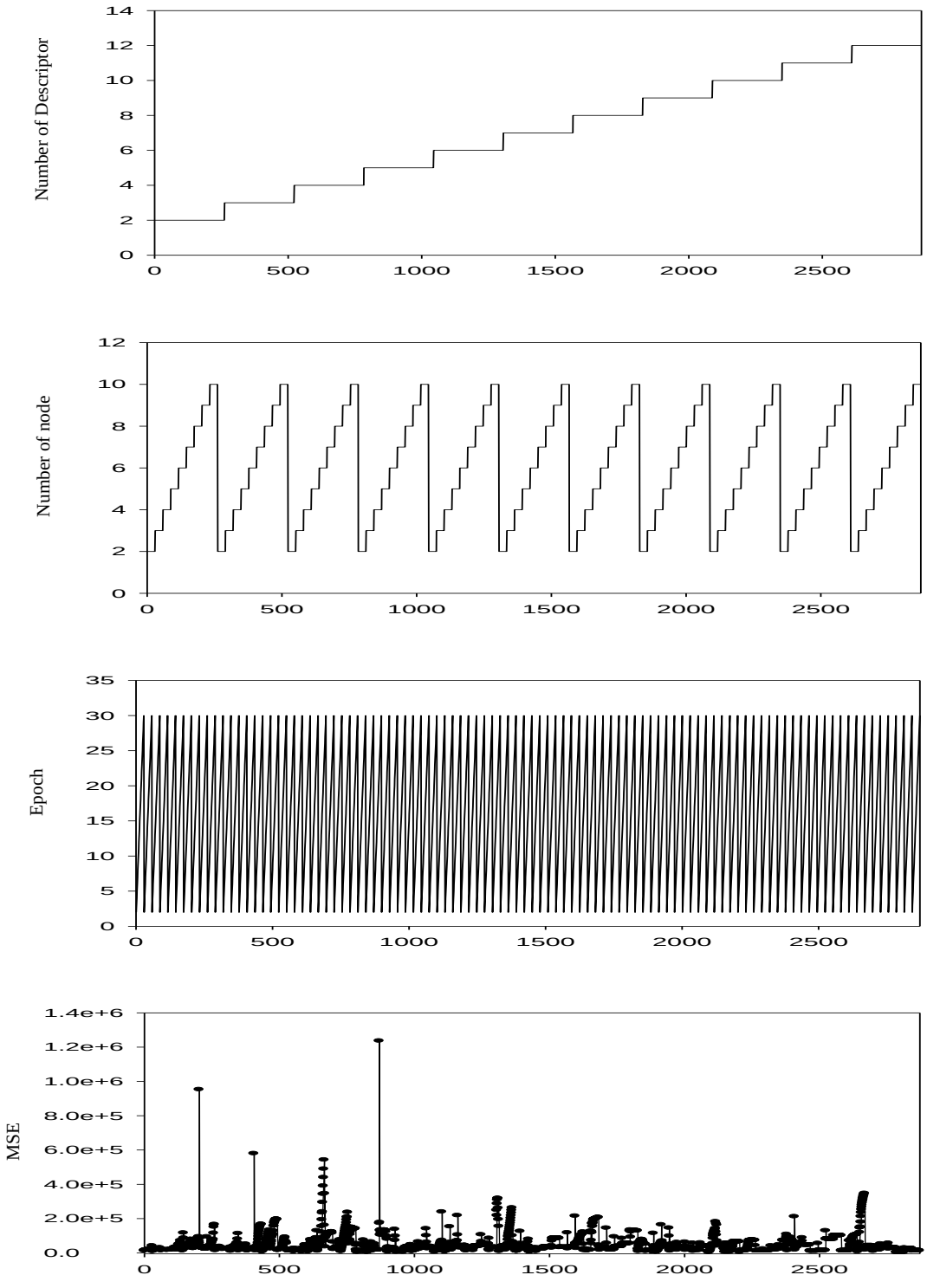
باتوجه به پارامترهایی که باید برای شبکه بهینه می‌شد جهت بهینه سازی همزمان تمام پارامترها در مجموع۱۱۴۸۴ شبکه مورد ارزیابی قرارگرفته است که در زیر به تعداد تمام موارد مورد بررسی اشاره شده است.

۱۱۴۸۴= دورهای آموزش(۲۹) × گره(۹) × تابع تبدیل(۲) × تابع آموزش(۲) × توصیف‌کننده (۱۱)

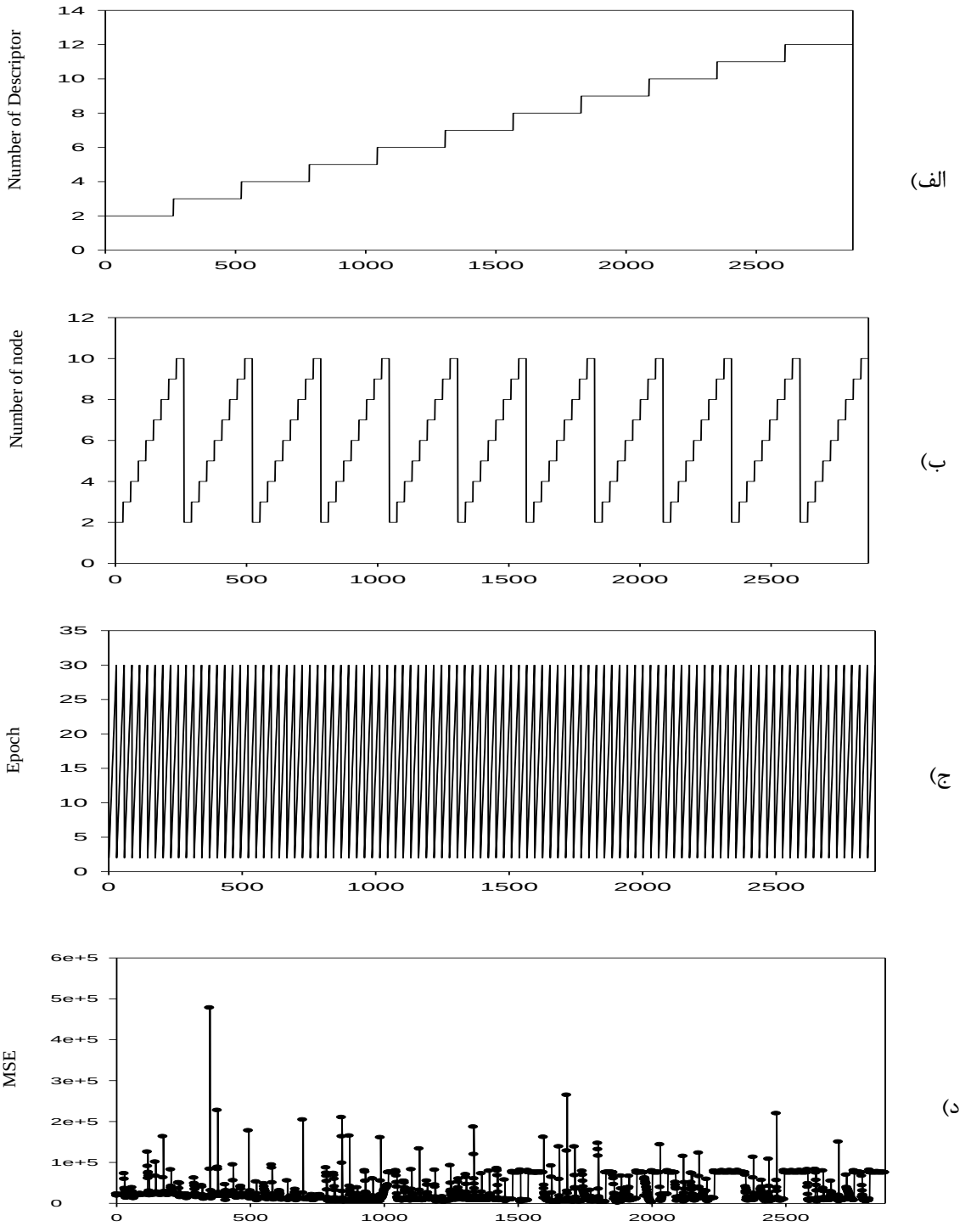
بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE به دست آمده به صورت نموداری بر حسب یک بردار مرجع فرضی در شکل‌های (۴-۲) تا (۴-۵) آمده است.



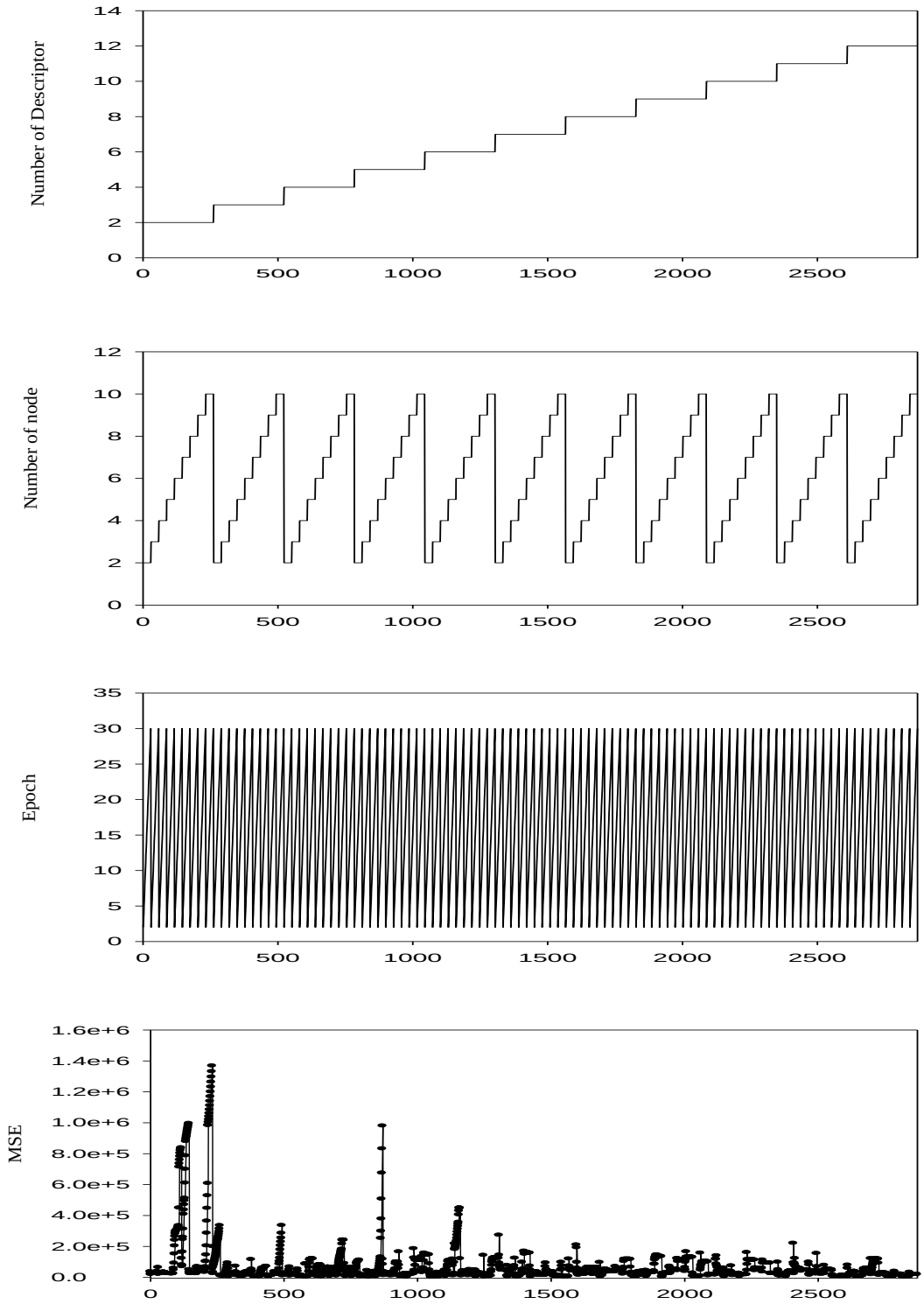
شکل (۴-۲)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



شکل (۴-۳)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی تنظیم لونبرگ- مارکوارت



شکل (۴-۴)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



شکل (۴-۵)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم لونبرگ- مارکوارت

با توجه به نتایج حاصله برای تعداد مختلف توصیف‌کننده‌ها و گره‌ها و دورهای آموزش و همچنین توابع متفاوت آموزش و تبدیل، بهترین شبکه‌های به دست آمده براساس کمترین مقدار MSE درجدول (۴-۷) خلاصه شده‌اند. با توجه به جدول (۴-۷) شبکه عصبی سه لایه با ۹ توصیف‌کننده درلایه ورودی، ۳ نرون در لایه مخفی و تعداد دور آموزشی ۲۰ با تابع آموزشی تنظیم بایزین و تابع تبدیل لگاریتم سیگموئید کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین شبکه‌ای بااین ساختار به عنوان شبکه بهینه انتخاب شد.

جدول (۴-۷)- توابع و پارامترهای شبکه های بهینه بدست آمده براساس مقادیر میانگین مربعات خطا

MSE	تعداد دورهای آموزش	تعدادگره ها	تابع تبدیل	تابع آموزش	تعدادتوصیف کننده ها
۲۴۹۳/۹	۸	۲	لگاریتم سیگموئید	لونبرگ - مارکوارت	۶
۱۵۸۷/۷	۲۰	۳	لگاریتم سیگموئید	بایزین	۹
۱۷۴۱/۱	۶	۴	تانژانت سیگموئید	لونبرگ - مارکوارت	۱۰
۱۶۲۲/۲	۱۳	۲	تانژانت سیگموئید	بایزین	۹

۴-۶-۱-۱- بهینه سازی پارامتر μ

جهت بهینه کردن مقدار μ، ساختار شبکه با ۹ ورودی، ۳ گره در لایه پنهان و الگوریتم آموزشی بایزین در نظر گرفته شد. سپس مقدار μ از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. که در این مورد نیز، میانگین مربعات خطای شبکه با تغییرات μ تغییر نکرده، لذا مقدارμ برابر با مقدار پیش فرض آن یعنی ۰/۰۰۵ می‌باشد.

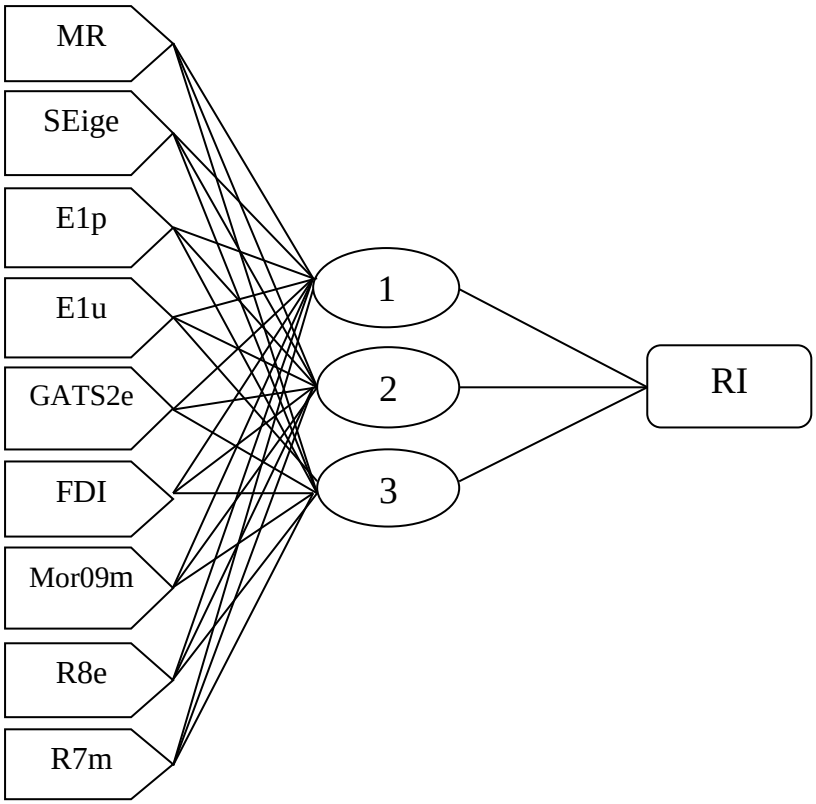
۴-۶-۱-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، مشخصات شبکه بهینه در جدول (۴-۸) ارائه شده است.

جدول (۴-۸)- مشخصات شبکه بهینه

تابع آموزش	trainbr
تابع تبدیل لایه پنهان	logsig
تابع تبدیل لایه خارجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۳
تعداد متغیرهای ورودی(توصیف‌کننده)	۹
تعداد چرخه های آموزشی	۲۰
پارامتر μ	۰/۰۰۵

ساختار شبکه عصبی به دست آمده پس از بهینه‌سازی فاکتورهای بحث شده در بالا به طور شماتیک در شکل (۴-۶) نشان داده شده است.



شکل (۴-۶)- تصویر شماتیک ساختار شبکه عصبی به دست آمده پس از بهینه سازی

۴-۷- مدلسازی شبکه عصبی مصنوعی باتوصیف‌کننده‌های انتخاب شده توسط روش الگوریتم ژنتیک (GA-ANN)

در این بخش نیز از شبکه عصبی سه لایه متشکل از یک لایه ورودی، یک لایه پنهان و یک لایه خروجی استفاده شد. در این مرحله نیز شبکه با ورودی های از ۲ تا ۸ توصیف کننده، توسط دو الگوریتم آموزشی لونبرگ – مارکوارت و تنظیم بایزین، با تعداد متفاوت گره در لایه پنهان از ۲ تا ۱۰ و همچنین توابع لگاریتم سیگموئیدی (logsig) و تانژانت سیگموئیدی (tansig)،به عنوان توابع تبدیل لایه پنهان و تعداد متفاوت دور آموزش ۲ تا ۳۰، آموزش داده شد. همچنین از تابع تبدیل خطی (purelin) در لایه خروجی استفاده شد. معیار نیز به حداقل رساندن میانگین مربع خطا (MSE) در نظر گرفته شد.

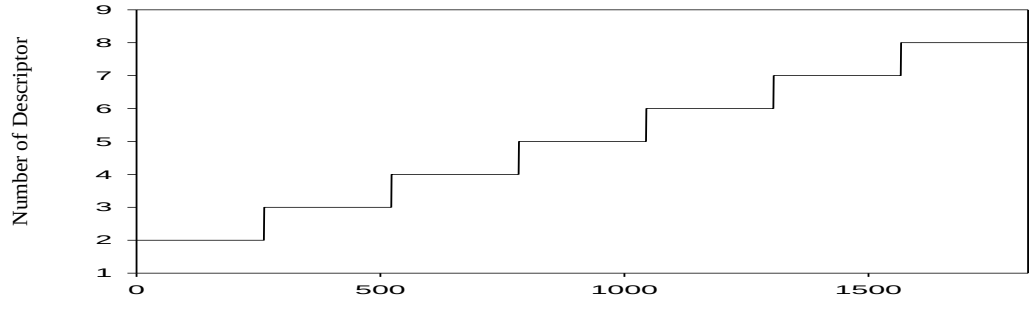
۴-۷-۱- انتخاب تعداد متغیرهای ورودی شبکه، الگوریتم آموزش، نوع تابع تبدیل، تعداد گره‌های لایه‌های پنهان وتعداد دورهای آموزشی

در این جا نیز بهینه سازی و انتخاب بهترین تعداد متغیرهای ورودی شبکه، الگوریتم آموزشی، نوع تابع تبدیل، تعداد نرون‌های لایه پنهان وتعداد دورهای آموزش به طور همزمان انجام گرفت (دراین مرحله، مقدار پارامتر ممنتوم، ۰/۰۰۵ درنظر گرفته شد). سپس در مقادیر بهینه‌ی این پارامترها، مقدار پارامتر ممنتوم از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد تا با توجه به مقدار خطای شبکه، مقدار بهینه انتخاب گردد.

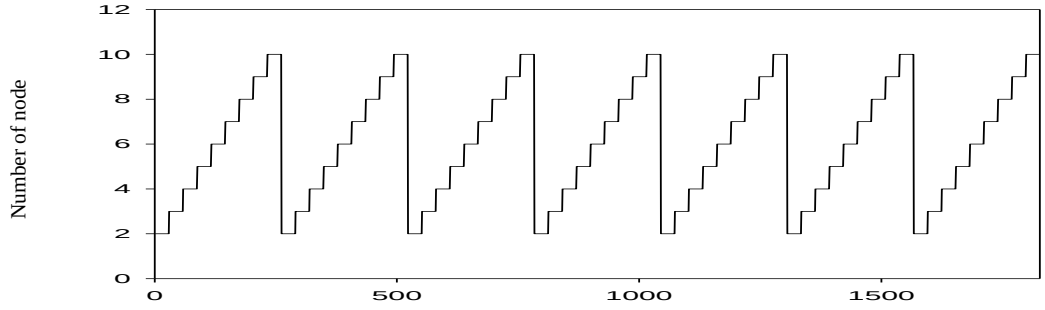
باتوجه به پارامترهایی که باید برای شبکه بهینه می‌شد جهت بهینه سازی همزمان تمام پارامترها در مجموع ۷۳۰۸ شبکه مورد ارزیابی قرارگرفته است که در زیر به تعداد تمام موارد مورد بررسی اشاره شده است.

۷۳۰۸= دورهای آموزش(۲۹) × گره(۹) × تابع تبدیل(۲) × تابع آموزش(۲) × توصیف‌کننده (۷)

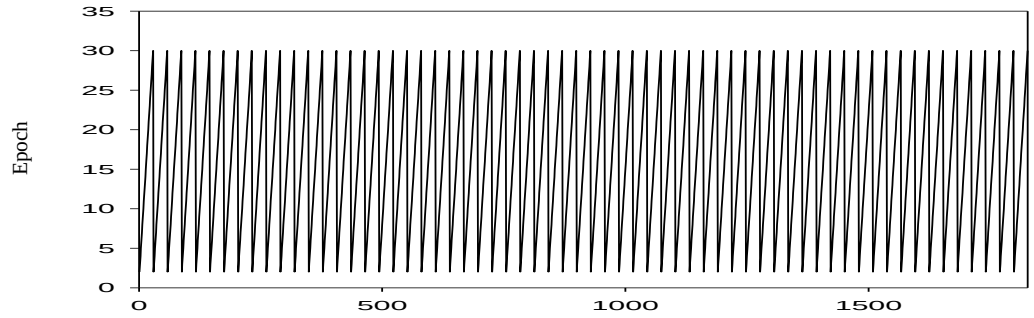
بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیرMSE به دست آمده به صورت نموداری بر حسب یک بردار مرجع فرضی در شکل‌های (۴-۷) تا (۴-۱۰) آمده است.



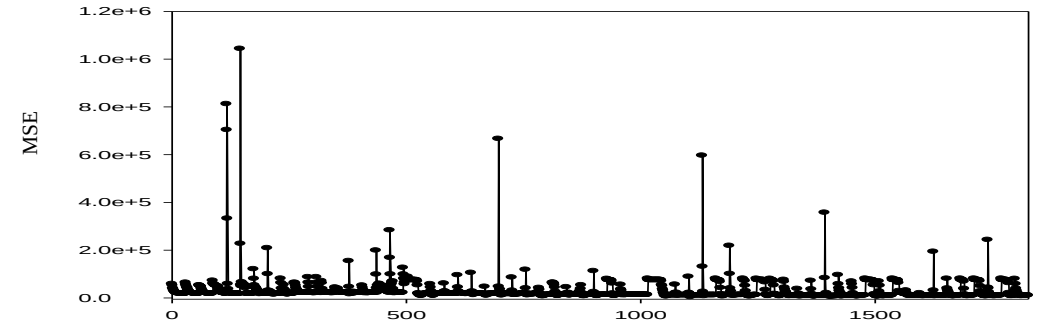
الف)



ب)



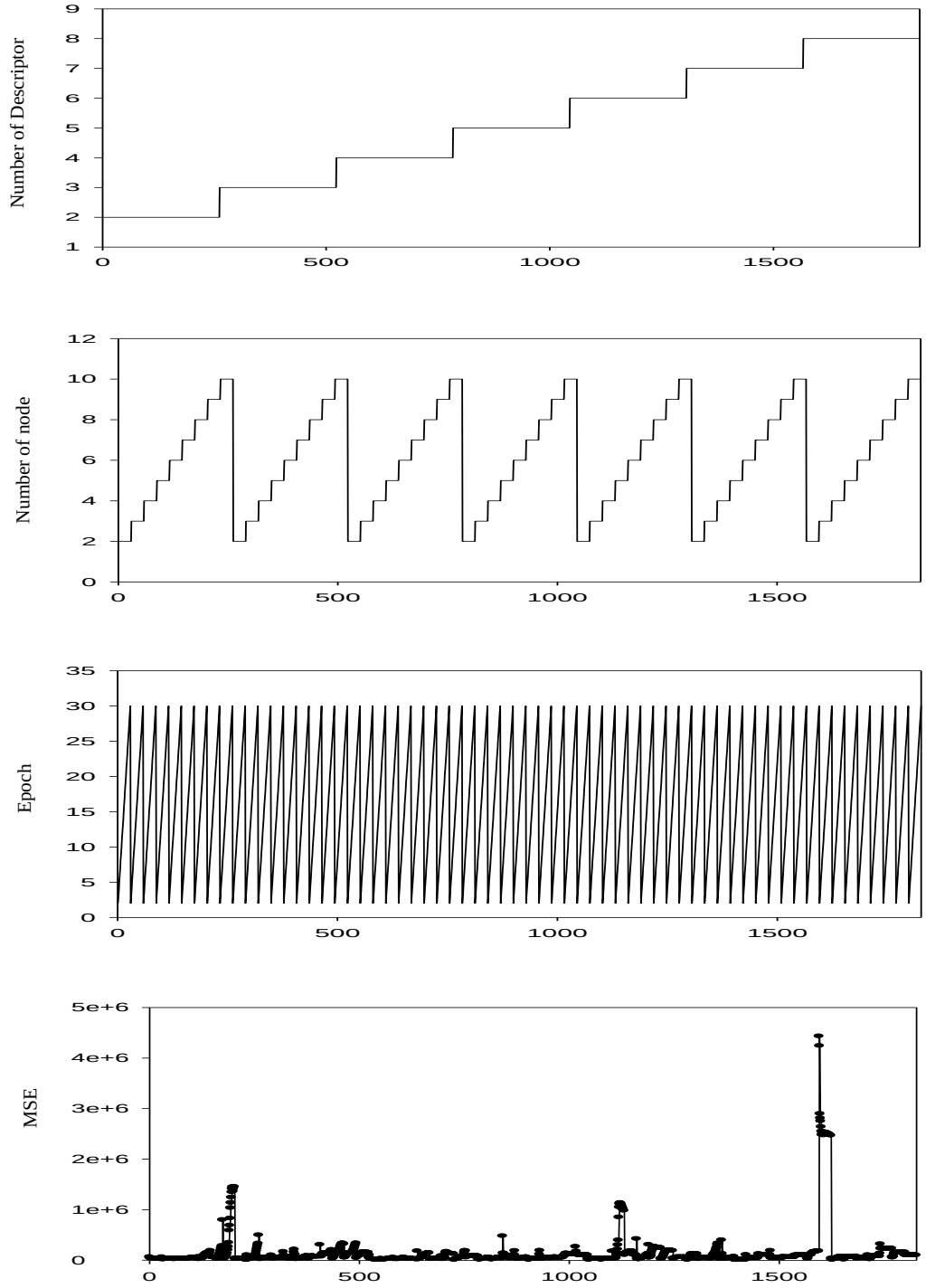
ج)



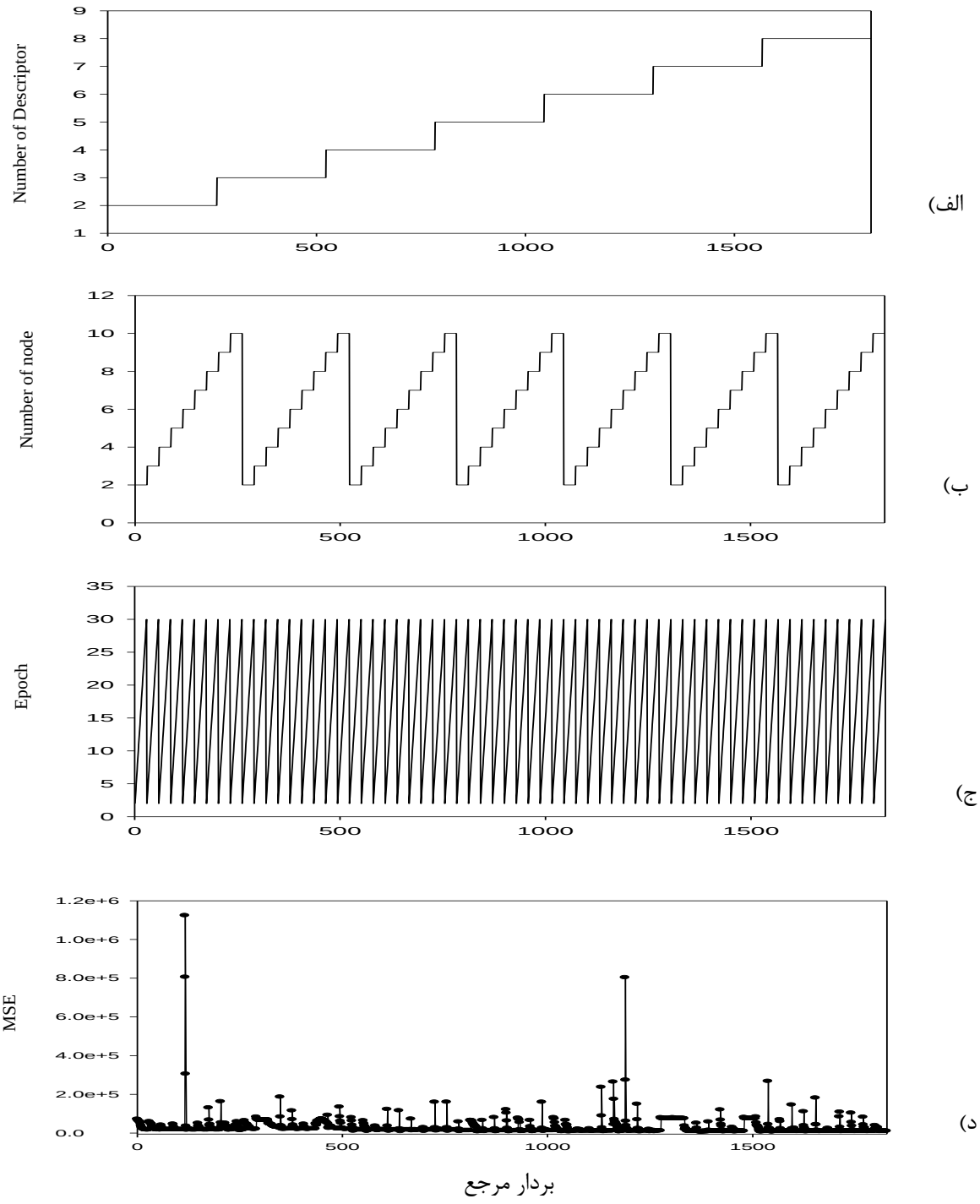
د)

بردار مرجع

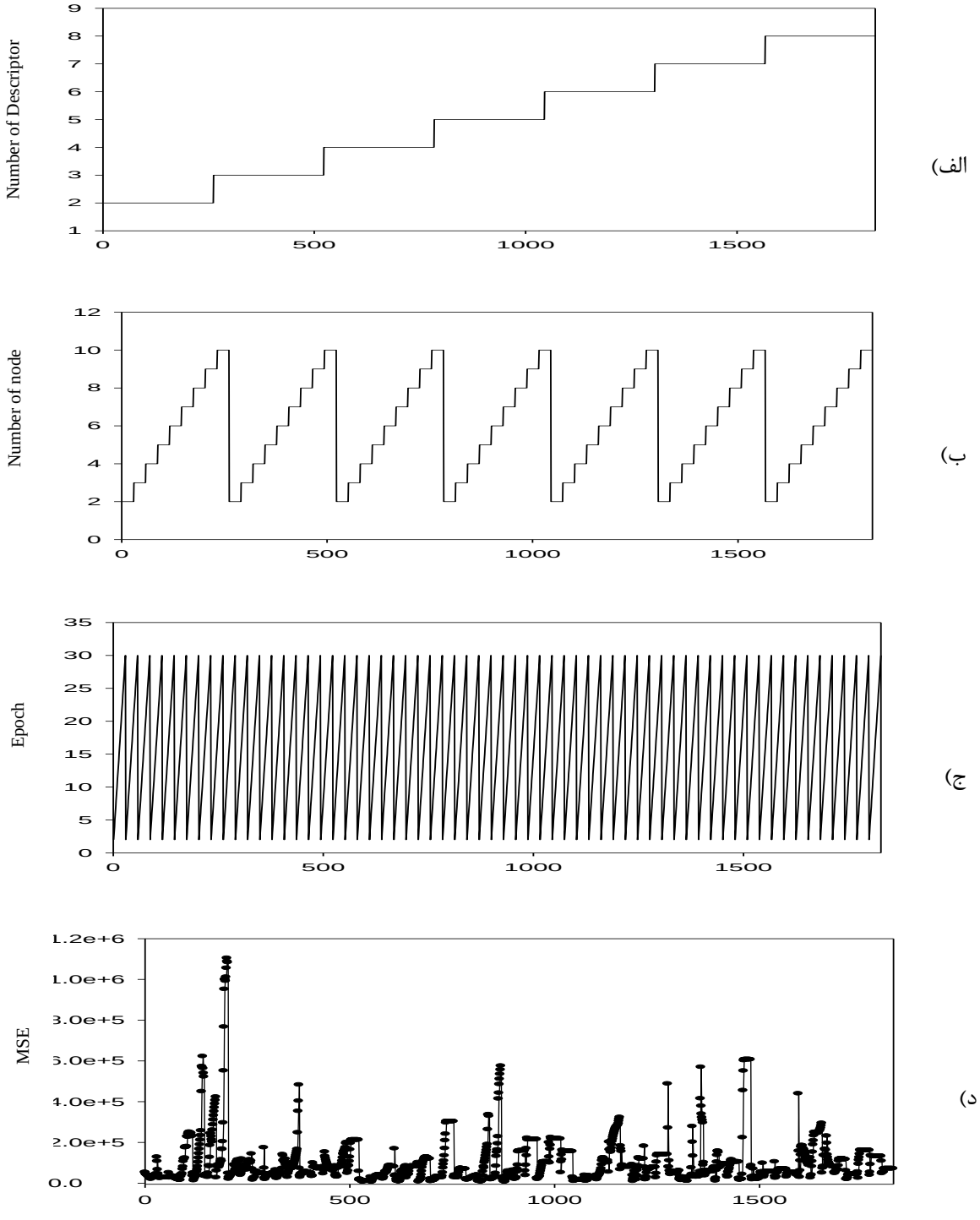
شکل (۴-۷)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم آموزشی بایزین



شکل (۴-۸)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل تانژانت سیگموئیدی و الگوریتم لونبرگ- مارکوارت



شکل (۴-۹)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرون لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم آموزشی تنظیم بایزین



شکل (۴-۱۰)- نمودارهای (الف) تعداد توصیف‌کننده، (ب) تعداد نرو ن لایه مخفی، (ج) دورآموزش، (د) MSE سری ارزیابی برای تابع تبدیل لگاریتم سیگموئیدی و الگوریتم لونبرگ- مارکوارت

با توجه به نتایج حاصله برای تعداد مختلف توصیف‌کننده‌ها و گره‌ها و دورهای آموزش و همچنین توابع متفاوت آموزش و تبدیل، بهترین شبکه‌های به دست آمده براساس کمترین مقدار MSE درجدول (۴-۹) خلاصه شده‌اند. با توجه به این جدول مشاهده می‌کنیم که دو شبکه عصبی با تابع آموزشی تنظیم بایزین و تابع تبدیل تانژانت سیگموئید و شبکه عصبی با تابع آموزشی تنظیم بایزین و تابع تبدیل لگاریتم سیگموئید دارای مقدار MSE با تفاوت بسیار کم می‌باشند در حالیکه مقادیر MRE شبکه عصبی با تابع آموزشی تنظیم بایزین و تابع تبدیل تانژانت سیگموئید به ترتیب برای دو سری ارزیابی و تست ۱/۷۸۴ و ۱/۸۱۹ بوده و مقدار این پارامتر آماری برای شبکه عصبی با تابع آموزشی تنظیم بایزین و تابع تبدیل لگاریتم سیگموئید برای دو سری ارزیابی و تست به ترتیب ۱/۷۸۸ و ۲/۱۱۶ می‌باشد بنابراین در مقایسه‌ی این دو شبکه، با توجه به بهتر بودن مقادیر MRE برای شبکه‌ی با تابع آموزشی تنظیم بایزین و تابع تبدیل تانژانت سیگموئید نسبت به شبکه دیگر و داشتن کمترین مقدار MSE نسبت به شبکه‌های بهینه، این شبکه به عنوان بهترین شبکه انتخاب شد. در نتیجه شبکه‌ای با ساختار ۷ توصیف‌کننده درلایه ورودی، ۵ نرون درلایه مخفی و تعداد دور آموزشی ۱۳ و تابع آموزشی تنظیم بایزین و تابع تبدیل تانژانت سیگموئید بهترین شبکه می‌باشد.

جدول(۴-۹)- توابع و پارامترهای شبکه های بهینه بدست آمده براساس مقادیر میانگین مربعات خط

MSE	تعداد دورهای آموزش	تعدادگره ها	تابع تبدیل	تابع آموزش	تعدادتوصیف کننده ها
۶۵۲۴/۶	۵	۳	لگاریتم سیگموئید	لونبرگ - مارکوارت	۴
۶۱۸۳/۸	۱۲	۴	لگاریتم سیگموئید	بایزین	۷
۶۹۱۳/۴	۱۰	۵	تانژانت سیگموئید	لونبرگ - مارکوارت	۷
۶۱۸۳/۶	۱۳	۵	تانژانت سیگموئید	بایزین	۷

۴-۷-۲- بهینه سازی پارامتر μ

جهت بهینه کردن مقدار μ، ساختار شبکه با ۷ ورودی، ۵ گره در لایه پنهان و الگوریتم آموزشی بایزین و تابع تبدیل تانژانت سیگموئید در نظر گرفته شد. سپس مقدار μ از ۰/۰۰۰۵ تا ۰/۰۱ با گام های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری ارزیابی محاسبه گردید. که در این مورد نیز، میانگین مربعات خطای شبکه با تغییرات μ تغییر نکرده و مقدار آن ثابت بود (۶۱۸۳/۶). بنابراین مقدار μ برابر با مقدار پیش فرض آن در تابع آموزشی بایزین یعنی ۰/۰۰۵می-باشد.

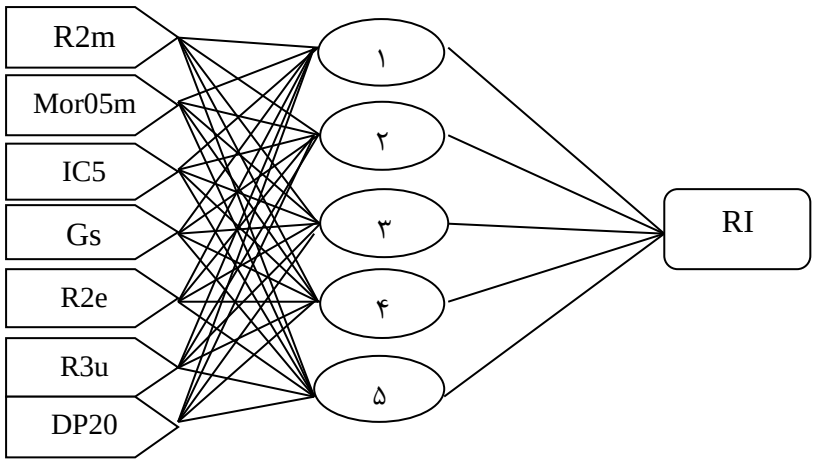
۴-۷-۳- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، مشخصات شبکه بهینه در جدول (۴-۱۰) ارائه شده است.

جدول (۴-۱۰)- مشخصات شبکه بهینه

تابع آموزش	trainbr
تابع تبدیل لایه پنهان	tansig
تابع تبدیل لایه خارجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۵
تعداد متغیرهای ورودی(توصیف‌کننده)	۷
تعداد چرخه های آموزشی	۱۳
پارامتر μ	۰/۰۰۵

ساختار شبکه عصبی به دست آمده پس از بهینه‌سازی فاکتورهای بحث شده در بالا به طور شماتیک در شکل (۴-۱۱) نشان داده شده است.



شکل (۴-۱۱)- تصویر شماتیک ساختار شبکه عصبی به دست آمده پس از بهینه سازی

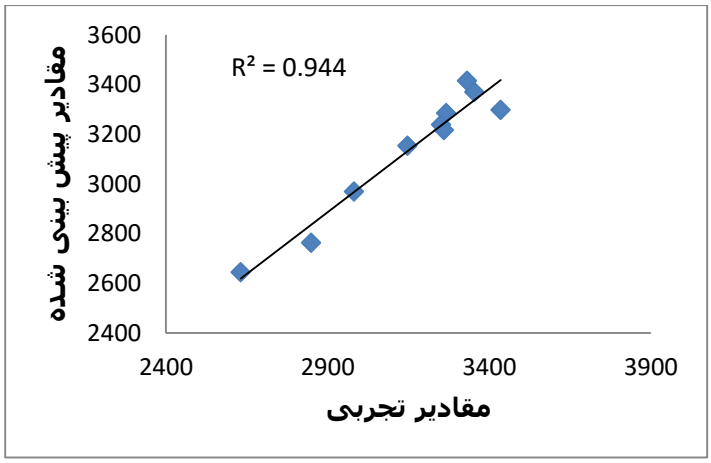
۴-۸- ارزیابی مدل‌های SR-ANN و GA-ANN

۴-۸-۱- ارزیابی مدل‌هایSR-ANN وGA-ANN با استفاده از سری تست

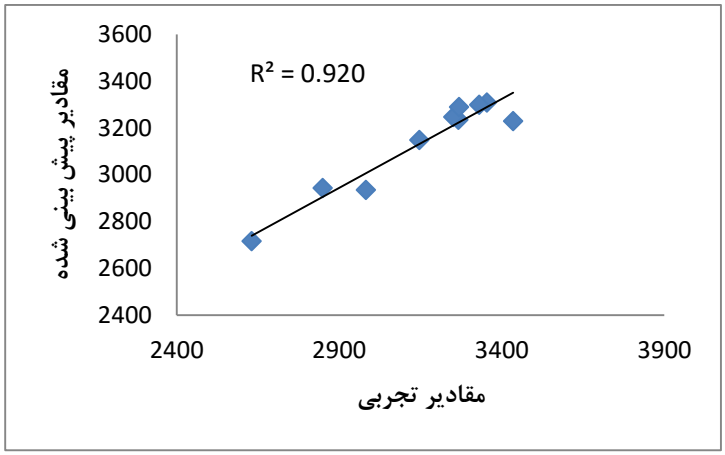
با استفاده از روش‌های SR-ANN و GA-ANN مقادیر اندیس بازداری برای سری‌های ارزیابی و تست پیش‌بینی شدند. در جدول (۴-۱۱) مقادیر محاسبه شده و درصد خطای نسبی برای داده‌های سری تست با دو روش SR-ANN و GA-ANN آورده شده است. نمودارهای مقادیر پیش‌بینی شده در مقابل مقادیر تجربی شاخص بازداری برای سری تست مربوط به هر دو روش به ترتیب در شکل‌های (۴-۱۲) و (۴-۱۳) رسم شده است. ضرایب تعیین برای سری تست روش‌های SR-ANN و GA-ANN به ترتیب ۰/۹۴۴ و ۰/۹۲۰ می‌باشد.

جدول (۴- ۱۱) – نتایج حاصل از ارزیابی مدل‌های SR-ANN و GA-ANN با استفاده از سری تست

شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۴	۲۶۳۱	۲۶۴۵	۲۷۱۶	۰/۵۳	۳/۲۳
۸	۲۸۴۹	۲۷۶۳	۲۹۴۴	-۳/۰۲	۳/۳۳
۱۴	۳۲۶۷	۳۲۸۵	۳۲۳۴	۰/۵۵	-۱/۰۱
۲۰	۳۳۵۴	۳۳۶۹	۳۳۰۹	۰/۴۵	-۱/۳۴
۲۲	۳۴۳۵	۳۲۹۸	۳۲۲۹	-۳/۹۹	-۵/۹۹
۲۶	۲۹۸۲	۲۹۷۰	۲۹۳۵	-۰/۴۰	-۱/۵۸
۳۰	۳۱۴۷	۳۱۵۳	۳۱۴۸	۰/۱۹	۰/۰۳
۳۴	۳۲۵۱	۳۲۳۹	۳۲۴۸	-۰/۳۹	-۰/۰۹
۳۹	۳۳۳۱	۳۴۱۵	۳۲۹۹	۲/۵۲	-۰/۹۶
۴۳	۳۲۶۹	۳۲۱۷	۳۲۹۰	-۱/۵۹	۰/۶۴



شکل(۴-۱۲)- مقادیر پیش بینی شده RI بر حسب مقادیر تجربی برای سری تست با مدل SR-ANN



شکل(۴-۱۳)- مقادیر پیش بینی شده RI بر حسب مقادیر تجربی برای سری تست با مدلGA-ANN

۴-۸-۲- ارزیابی مدل‌های SR-ANN و GA-ANN به روش رد مرحله‌ای تک‌تک داده‌ها

برای ارزیابی مدل، روش رد مرحله‌ای تک‌تک داده‌ها مورد استفاده قرار گرفت و نتایج حاصل از آن با مقادیر تجربی برای سری داده‌ها توسط مدل‌های SR-ANN و GA-ANN مقایسه گردید. مقدار واقعی، پیش‌بینی شده، درصد خطای نسبی برای سری داده‌ها در جدول (۴-۱۲) مشاهده می‌شود. نزدیک بودن نتایج حاصل از ارزیابی به روش رد مرحله‌ای تک‌تک داده‌ها با نتایج به دست آمده از مدل منتخب برای هر دو روش نشان‌دهنده صحت مدل می‌باشد.

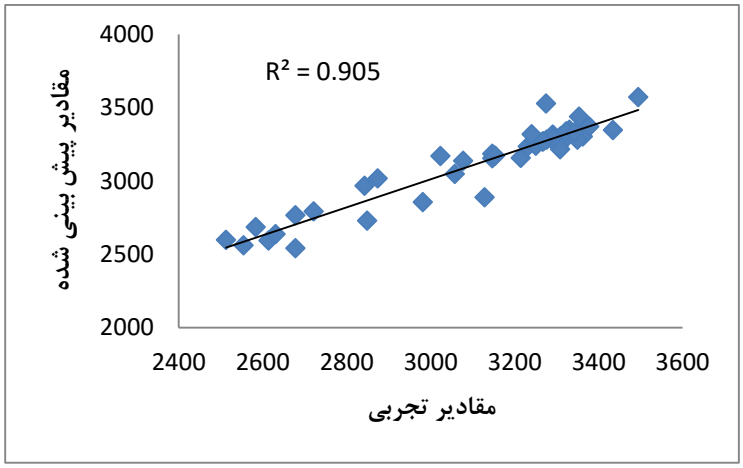
جدول(۴-۱۲)- نتایج حاصل از ارزیابی مدل برتر ارائه شده توسط SR-ANN و GA-ANN با استفاده از روش رد مرحله‌ای تک‌تک

شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۱	۲۵۱۳	۲۵۹۷	۲۶۲۴	۳/۳۴	۴/۴۲
۲	۲۵۵۵	۲۵۶۱	۲۶۴۶	۰/۲۳	۳/۵۶
۳	۲۶۱۴	۲۵۹۵	۲۵۵۹	-۰/۷۳	-۲/۱۰
۴	۲۶۳۱	۲۶۳۸	۲۶۱۳	۰/۲۷	-۰/۶۸
۵	۲۶۷۸	۲۵۴۱	۲۶۵۸	- ۵/۱۲	-۰/۷۵
۶	۲۷۲۲	۲۷۹۳	۲۶۴۲	۲/۶۱	-۲/۹۴
۷	۲۸۴۳	۲۹۶۷	۲۸۴۷	۴/۳۶	۰/۱۴
۸	۲۸۴۹	۲۷۲۹	۲۹۴۰	-۴/۲۱	۳/۱۹
۹	۲۸۷۴	۳۰۱۷	۲۸۸۵	۴/۹۸	۰/۳۸
۱۰	۳۰۵۸	۳۰۴۷	۳۰۶۲	-۰/۳۶	۰/۱۳
۱۱	۳۰۷۸	۳۱۳۸	۳۰۹۷	۱/۹۵	۰/۶۲
۱۲	۳۱۵۲	۳۱۷۱	۳۲۷۸	۰/۶۰	۳/۹۹
۱۳	۳۲۱۵	۳۱۵۷	۳۲۳۲	-۱/۸۰	۰/۵۳
۱۴	۳۲۶۷	۳۲۶۹	۳۲۷۲	۰/۰۶	۰/۱۵
۱۵	۳۲۷۵	۳۵۲۸	۳۳۳۴	۷/۷۲	۱/۸۰
۱۶	۳۲۷۷	۳۲۸۱	۳۲۷۰	۰/۱۲	-۰/۲۱
۱۷	۳۳۱۰	۳۳۰۵	۳۲۶۷	-۰/۱۵	-۱/۳۰
۱۸	۳۳۱۲	۳۲۶۳	۳۲۹۶	-۱/۴۸	-۰/۴۸
۱۹	۳۳۲۱	۳۲۹۵	۳۲۸۰	-۰/۷۸	-۱/۲۳
۲۰	۳۳۵۴	۳۴۳۸	۳۳۸۹	۲/۵۰	۱/۰۴
۲۱	۳۳۶۳	۳۳۰۴	۳۲۵۵	-۱/۷۵	-۳/۲۱
۲۲	۳۴۳۵	۳۳۴۶	۳۲۶۳	-۲/۵۹	-۵/۰۱
۲۳	۳۴۹۵	۳۵۷۲	۳۵۰۵	۲/۲۰	۰/۲۹
۲۴	۲۵۸۴	۲۶۸۶	۲۷۲۱	۳/۹۵	۵/۳۰
۲۵	۲۶۷۸	۲۷۶۷	۲۶۸۸	۳/۳۲	۰/۳۷
۲۶	۲۹۸۲	۲۸۵۶	۳۱۸۱	-۴/۲۲	۶/۶۳
۲۷	۳۰۲۴	۳۱۶۹	۳۲۷۶	۴/۷۹	۸/۳۳
۲۸	۳۱۲۹	۲۸۸۹	۳۲۲۵	-۷/۶۷	۳/۰۷
۲۹	۳۱۴۷	۳۱۵۴	۲۹۳۴	۰/۲۲	-۶/۷۷
۳۰	۳۱۴۷	۳۱۸۵	۳۱۷۴	۱/۲۱	۰/۸۶
۳۱	۳۱۵۴	۳۱۷۲	۳۱۹۰	۰/۵۷	۱/۱۴
۳۲	۳۲۳۲	۳۲۳۶	۳۲۳۷	۰/۱۲	۰/۱۵

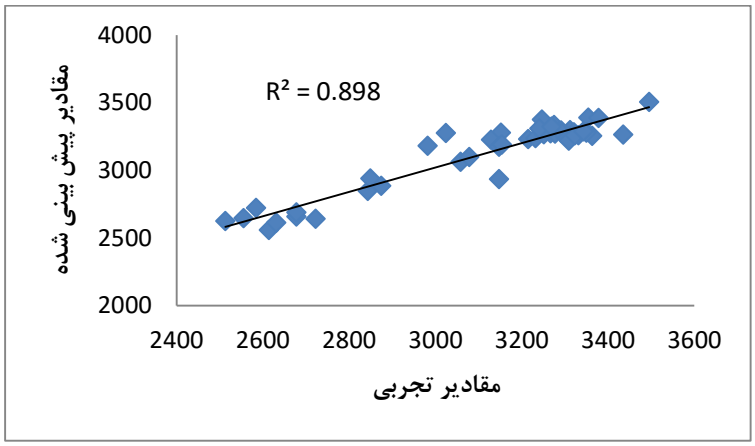
ادامه جدول(۴-۱۲)

شماره ترکیب	RI				
	مقدار تجربی	مقدار پیش بینی شده		درصد خطا	
		SR-ANN	GA-ANN	SR-ANN	GA-ANN
۳۳	۳۲۴۷	۳۲۶۹	۳۳۷۶	۰/۶۸	۳/۹۷
۳۴	۳۲۵۱	۳۲۴۰	۳۲۶۶	-۰/۳۴	۰/۴۶
۳۵	۳۲۹۱	۳۳۱۷	۳۲۹۴	۰/۷۹	۰/۰۹
۳۶	۳۲۹۳	۳۲۹۰	۳۲۷۶	-۰/۰۹	-۰/۵۲
۳۷	۳۳۰۵	۳۲۳۸	۳۲۳۹	-۲/۰۳	-۱/۹۹
۳۸	۳۳۰۹	۳۲۱۵	۳۲۲۰	-۲/۸۴	-۲/۶۹
۳۹	۳۳۳۱	۳۳۵۰	۳۲۶۰	۰/۵۷	-۲/۱۳
۴۰	۳۳۵۰	۳۲۸۱	۳۲۷۹	-۲/۰۶	-۲/۱۲
۴۱	۳۳۷۸	۳۳۷۱	۳۳۸۸	-۰/۲۱	۰/۲۹
۴۲	۳۲۴۱	۳۳۲۰	۳۳۱۰	۲/۴۴	۲/۱۳
۴۳	۳۲۶۹	۳۲۷۱	۳۳۲۲	۰/۰۶	۱/۶۲
۴۴	۳۳۲۳	۳۳۳۹	۳۲۵۹	۰/۴۸	۱/۹۳
۴۵	۳۳۴۳	۳۳۳۲	۳۲۸۷	-۰/۳۳	-۱/۶۷

نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی داده‌ها نیز در ارزیابی دو مدل SR-ANN وGA-ANN به روش رد مرحله‌ای تک‌تک داده‌ها، رسم گردید که نتایج به دست آمده در شکل‌های (۴-۱۴) و (۴-۱۵) آورده شده است. نتایج این ارزیابی توانایی خوب هر دو روش را در پیش‌بینی شاخص بازداری نشان می‌دهد.

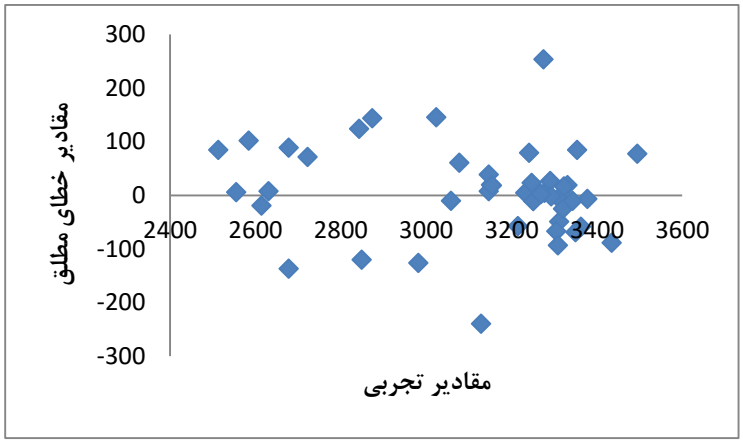


شکل (۴-۱۴) - نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی RI به روش رد مرحله‌ای تک‌تک برای کل داده‌ها با مدل SR-ANN

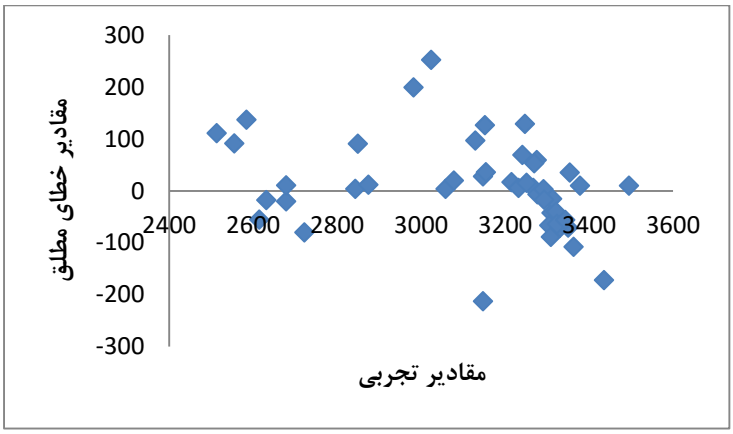


شکل (۴-۱۵)- نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی RI به روش رد مرحله‌ای تک‌تک برای کل داده‌ها با مدل GA-ANN

همچنین نمودار مقادیر باقی‌مانده‌ی محاسبه شده بر حسب مقادیر تجربی شاخص بازداری، در ارزیابی دو مدل در شکل‌های (۴-۱۶) و (۴-۱۷) رسم شد. توزیع متقارن داده‌ها حول محور افقی نشان‌دهنده عدم وجود خطای معین می‌باشد.



شکل(۴-۱۶)- نمودار باقیمانده ها بر حسب مقدار تجربی RI برای کل داده ها توسط مدل SR-ANN



شکل(۴- ۱۷)- نمودار باقیمانده ها بر حسب مقدار تجربی RI برای کل داده ها توسط مدل GA-ANN

۴-۸-۳- ارزیابی مدل‌های برتر SR-ANN و GA-ANN با استفاده از پارامترهای آماری

مطابق جدول (۴-۱۳) پارامترهای آماری موجود، جهت ارزیابی توانایی پیش‌بینی مدل‌های ساخته شده به روش SR-ANN و GA-ANN آورده شده‌اند. مقادیر این پارامترها نشان می‌دهد که مدل SR-ANN ارائه شده تا حدودی از قدرت پیش‌بینی بهتری نسبت به مدل GA-ANN برخوردار است. بنابراین مدل SR-ANN به عنوان مدل برتر انتخاب گردید.

جدول (۴-۱۳)- پارامترهای آماری برای مدل های برتر طراحی شده توسط SR-ANN و GA-ANN

پارامتر	سری تست (N=10)		سری ارزیابی(N=10)		کل داده ها(N=45)	
	SR-ANN	GA-ANN	SR-ANN	GA-ANN	SR-ANN	GA-ANN
MAE	۴۳/۸	۵۶/۸	۲۹/۳	۵۳/۹	۶۰/۰۴	۶۲/۹۱
MSE	۳۷۰۴/۰	۶۵۳۲/۷	۱۵۸۷/۷	۶۱۸۳/۶	۷۱۴۶/۰	۷۵۲۵/۸
PRESS	۳۷۰۴۰/۴	۶۵۳۲۷/۲	۱۵۸۷۷/۴	۶۱۸۳۵/۸	۳۲۱۵۷۰	۳۳۸۶۶۱
SEP	۶۰/۸۶	۸۰/۸۲	۳۹/۸۵	۷۸/۶۴	۸۴/۵۳	۸۶/۷۵
REP (%)	۱/۹۵	۲/۵۹	۱/۲۸	۲/۵۲	۲/۷۱	۲/۷۸
MRE	۱/۳۷	۱/۸۲	۰/۹۸	۱/۸۱	۱/۹۸	۲/۰۵
R^2(Leave One Out)					۰/۹۰۵	۰/۸۹۸

۴-۸-۴- ارزیابی مدل ارائه شده توسط شبکه عصبی با استفاده از آزمون Y- تصادفی

در این روش، مقادیر متغیر وابسته به صورت تصادفی تغییر داده شد. سپس همبستگی متغیرهای مستقل با متغیرهای وابسته توسط شاخص R^2 مورد بررسی قرار گرفتند. در جداول (۴-۱۴) و (۴-۱۵) مقادیر ضریب تعیین برای دو مدل، برای سری ارزیابی و تست آورده شده است. مقادیر پایین R^2 نشان‌دهنده عدم وجود همبستگی شانسی در مدل اصلی هستند.

جدول (۴-۱۴)– مقادیرR^2 برای سری ارزیابی و تست با استفاده از آزمون Y- تصادفی روش SR-ANN

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2 (ارزیابی)	۰/۲۶۰	۰/۰۰۱۵	۰/۰۱۰	۰/۰۰۲	۰/۰۳۶	۰/۰۰۳	۰/۰۰۸	۰/۰۳۹	۰/۰۰۲	۰/۳۶
R^2 (تست)	۰/۰۷۲	۰/۰۰۷	۰/۰۰۹	۰/۰۳۵	۰/۰۲۳	۰/۰۱۹	۰/۲۵	۰/۰۰۱	۰/۰۱۹	۰/۰۱۱

جدول (۴-۱۵)– مقادیر R^2 برای سری ارزیابی و تست با استفاده از آزمون Y- تصادفی روش GA-ANN

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2 (ارزیابی)	۰/۱۷۴	۰/۰۰۱	۰/۱۰۸	۰/۱۵۰	۰/۲۵۰	۰/۱۳۳	۰/۲۴۵	۰/۰۷۰	۰/۱۶۸	۰/۳۲۱
R^2 (تست)	۰/۱۹۷	۰/۰۵۸	۰/۰۸۰	۰/۰۱۴	۰/۰۹۸	۰/۰۹۲	۰/۱۱۳	۰/۰۶۴	۰/۰۰۹	۰/۲۵۴

۴-۹- بررسی ارتباط توصیف‌کننده‌های منتخب با شاخص بازداری ترکیبات

مدل برتر انتخاب شده شامل ۹ توصیف‌کننده است که هرکدام بیان‌گر خصوصیات متفاوتی از مولکول-های مورد بررسی می‌باشد. این توصیف‌کننده‌ها عبارتند از:

MR, SEige, E1p, E1u, GATS2e, FDI, Mor09m, R8e, R7m

جدول زیر اثر متوسط توصیف‌کننده‌ها را نشان می‌دهد. اثر متوسط یک متغیر مستقل با استفاده از معادله (۳-۱) که در بخش (۳-۷) آمده است، به دست می‌آید.

جدول (۴-۱۶)- اثر متوسط توصیف‌کننده‌ها

توصیف-کننده	MR	SEige	E1p	E1u	GATS2e	FDI	Mor09m	R8e	R7m
اثر متوسط	۱۶/۲۳۲	۶/۷۷۹	۸/۰۳۸	-۶/۹۲۰	-۳/۸۵۶	-۳/۶۵۷	-۴/۰۱۴	۴/۳۷۱	-۲/۹۲۰

۴-۹-۱- توصیف‌کننده‌های Property

MR جزء این دسته توصیف‌کننده‌ها می‌باشد این توصیف‌کننده به حجم مولی و قطبش پذیری وابسته است بطوریکه افزایش حجم مولکول باعث افزایش این توصیف‌کننده می‌شود. این توصیف‌کننده از رابطه لورنز-لورنزز[1] محاسبه می‌شود.

$$MR = \frac{(n^2-1)(n^2+1).MW}{\rho} = 1.333\pi Na \qquad (۱-۴)$$

MW وزن مولکولی، ρ دانسیته، n ضریب شکست، N عدد آواگادرو و a قطبش‌پذیری است. بنابراین MR به اندازه و قطبش‌پذیری گونه‌های شیمیایی وابسته است[۵۸]. همچنین با توجه به رابطه فوق با افزایش وزن مولکول نیز مقدار این توصیف‌کننده افزایش می‌یابد و از آنجا که MR رابطه مستقیم با شاخص بازداری دارد (اثر متوسط۱۶/۲۳۲) بنابراین می‌توان نتیجه گرفت که با افزایش وزن مولکول شاخص بازداری افزایش می‌یابد.

۴-۹-۲- توصیف‌کننده SEige

این توصیف‌کننده، جزء گروه توصیف‌کننده‌های توپولوژیکی هستند که از گراف مولکولی تهی از هیدروژن[2] محاسبه می‌شود. گراف مولکولی را با علامت

$$G=G(V,E) \qquad (۲-۴)$$

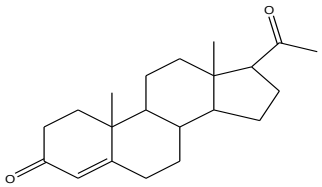
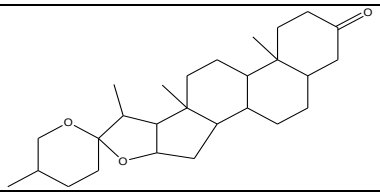
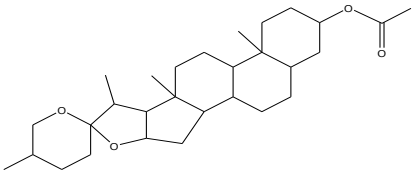
تعریف می‌کنند. در این رابطه V نشان‌دهنده‌ی رئوس که بیان‌گر تعداد اتم‌های موجود در ساختار مولکول بوده و E به عنوان یال‌ها که نشان‌دهنده‌ی پیوندهای متصل‌کننده این رئوس (اتم‌ها) به یکدیگر هستند، معرفی می‌شوند. تعداد یال‌های متصل به یک رأس را درجه آن رأس می‌نامند. یک گراف مولکولی که بدون درنظرگرفتن اتم‌های هیدروژن به دست بیاید، گراف مولکولی تهی شده از

1- Lorentz- Lorentz Equation
2- H-depleted Molecular Graf or Hydrogen- Suppressed Molecular Graf

هیدروژن نامیده می‌شود. با استفاده از نظریه گراف و با در نظرگرفتن موقعیت رئوس مولکول نسبت به یکدیگر می‌توان ساختار یک مولکول را به صورت یک ماتریس نشان داد. این ماتریس با افزایش تعداد اتم‌ها در مولکول (رئوس) و نیز با بزرگتر شدن مولکول و ورود حلقه‌ها به ساختار آن و همچنین پرشاخه شدن مولکول پیچیده‌تر می‌شود[۵۹].

توصیف‌کننده SEige از جمله توصیف‌کننده‌های اندیس VEA است که توسط ضریب 1_{iA} از یک بردار مشخصه تعریف می‌شود که این بردار با بزرگترین مقدار مشخصه منفی در ارتباط است. به عبارت دیگر، 1_{iA} مقدار مشخصه‌ی Aام از توالی رو به کاهش این مقادیر است[۶۰]. اثر توصیف‌کننده SEige بر روی شاخص بازداری مثبت می‌باشد (جدول۴-۱۶) به این معنی که با بزرگتر شدن و پرشاخه شدن مولکول مقدار این توصیف‌کننده افزایش یافته و در نتیجه شاخص بازداری افزایش می‌یابد. جدول (۴-۱۷) چگونگی ارتباط این توصیف‌کننده را با شاخص بازداری برای سه مولکول از سری داده‌ها نشان می‌دهد.

جدول (۴-۱۷)- مثال‌هایی از اثر توصیف‌کننده SEige بر شاخص بازداری

نام ترکیب	ساختار	مقدارSEige	شاخص بازداری
Progesterone		۰/۴۹۷	۲۸۴۳
Sarsasagenone		۰/۷۴۵	۳۳۲۱
Sarsasagenin acetate		۰/۹۹۴	۳۴۹۵

۴-۹-۳- توصیف‌کننده‌های WHIM

این شاخص از مختصات کارتزین ساختار سه‌بعدی مولکول، با استفاده از صورتبندی با حداقل انرژی محاسبه می‌شود و شامل اطلاعاتی درباره‌ی اندازه، شکل، تقارن و توزیع اتمی ساختار سه‌بعدی مولکول می‌باشد. این توصیف‌کننده از رابطه (۴-۳) به دست می‌آید:

$$S_{jk} = \frac{\sum_{i=1}^{A} w_i (q_{ij} - \overline{q_j})\,(q_{ik} - \overline{q_k})}{\sum_{i=1}^{A} w_i} \qquad (۳-۴)$$

که s_{jk} کوواریانس وزن‌دار شده بین کئوردینه j ام و k ام، A تعداد اتم‌ها، w_i وزن i امین اتم، q_{ij} وq_{ik} به ترتیب j امین و k امین کئوردینه‌های اتم i ام، و $\overline{q}$ مقدار میانگین متناظر است. شش طرح وزن‌دار شدن پیشنهاد شده است که عبارتند از:

۱ - حالت بدون وزن (u) ۲- وزن‌دار شده با جرم اتمی ۳- وزن‌دار شده با حجم واندروالس ۴- وزن‌دار شده با الکترونگاتیویته ساندرسون ۵- وزن‌دار شده با حالت الکتروتوپولوژیکی کییر و هال[1] ۶- وزن‌دار شده با قطبش‌پذیری[۶۱ و۵۲].

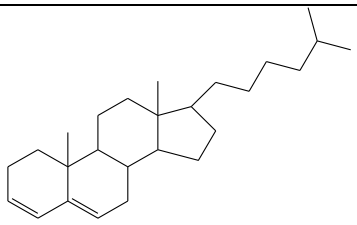
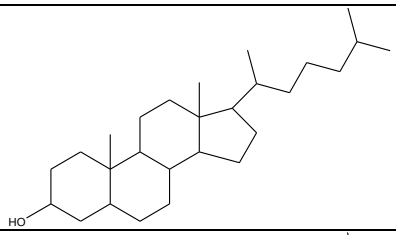
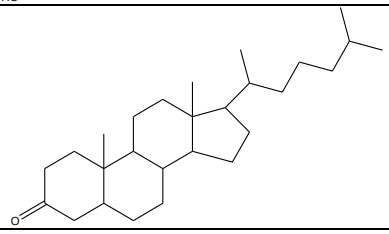
از بین این توصیف‌کننده‌ها E1p و E1u توسط مدل انتخاب شده که نشان‌دهنده تأثیر ساختار سه بعدی، شکل و قطبیت مولکول‌ها برروی شاخص بازداری می‌باشد. E1p که با قطبش‌پذیری وزن‌دار شده، طبق جدول (۴-۱۶) رابطه مستقیم با شاخص بازداری دارد یعنی با افزایش میزان قطبیت مولکول این توصیف‌کننده افزایش می‌یابد و به همین ترتیب شاخص بازداری افزایش می‌یابد و این رابطه را می‌توان با کمی قطبی بودن فاز ساکن در کروماتوگرافی توجیه کرد[۵۷]. بطوریکه مولکول‌های قطبی‌تر زمان بازداری بیشتر و در نتیجه شاخص بازداری بزرگتری دارند.

1 - Kier-hall

۴-۹-۴- توصیف‌کننده‌های2D-autocorrelation

همانطور که در بخش (۳-۷-۴) توضیح داده شد، این توصیف‌کننده‌ها جزء توصیفگرهای دوبعدی، گروه Autocorrelation Geary می‌باشند. توصیف‌کننده GATS2e یکی از این توصیف‌کننده‌ها بوده که با الکترونگاتیویته ساندرسون وزن‌دار شده و به عبارتی نقش فیزیکولوژی الکترونگاتیویته‌ی ساندسون را بیان می‌کند. و این توصیف‌کننده دارای اثر متوسط منفی بوده (جدول ۴-۱۶)، بدین معنی که افزایش آن باعث کاهش شاخص بازداری می‌شود. جدول (۴-۱۸) چگونگی تغییرات مقدار توصیف‌کننده GATS2e و شاخص بازداری را به عنوان مثال برای سه مولکول از سری داده‌ها نمایش می‌دهد.

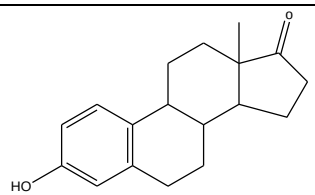
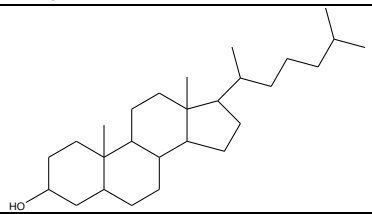
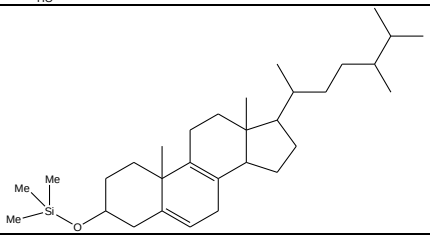
جدول (۴-۱۸)- مثال‌هایی از اثر توصیف‌کننده GATS2e بر شاخص بازداری

نام ترکیب	ساختار	مقدارGATS2e	شاخص بازداری
Cholesta-3,5-diene		۲/۲۰۳	۲۸۷۴
Cholestan-3 α -ol		۱/۸۸۰	۳۰۵۸
Coprostan-3-one		۱/۱۹۴	۳۰۷۸

۴-۹-۵- توصیف‌کننده‌های هندسی(Geometrical)

توصیف کننده‌های هندسی با ساختار سه بعدی مولکول‌ها در ارتباط می‌باشند. FDI یکی از این دسته توصیف‌کننده‌ها می‌باشد که در مدل ظاهر شده، مقدار این توصیف‌کننده بین ۰ و ۱ تعریف می‌شود (۰<FDI<۱) که این مقدار برای مولکول‌های خطی به یک همگرا می‌شود و با میزان بهم آمیختگی و پیچیدگی مولکول‌ها کاهش می‌یابد. بنابراین توصیف‌کننده FDI می‌تواند به عنوان شاخصی از درجه انحراف یک مولکول از حالت خطی بودن محض استفاده شود[۶۲]. با بزرگتر و پیچیده‌تر شدن مولکول مقدار توصیف‌کننده FDI کاهش یافته ولی شاخص بازداری افزایش می‌یابد. منفی بودن اثر متوسط این توصیف کننده بر روی شاخص بازداری نیز نشان می‌دهد که با افزایش آن (نزدیک شدن به مقدار یک) میزان شاخص بازداری کاهش می‌یابد. جدول (۴-۱۹) چگونگی تغییرات مقدار توصیف‌کننده FDI و شاخص بازداری را به عنوان مثال برای سه مولکول از سری داده‌ها نمایش می‌دهد.

جدول (۴-۱۹)- مثال‌هایی از اثر توصیف‌کننده FDI بر شاخص بازداری

نام ترکیب	ساختار	مقدارFDI	شاخص بازداری
Estrone		۰/۹۹۲	۲۶۱۴
Cholestan-3α-ol		۰/۹۵۲	۳۰۵۸
Ergosta-5,8-dien-3β-ol		۰/۹۲۱	۳۲۹۳

۴-۹-۶- توصیف‌کننده‌های 3D MORSE[1]

توصیف‌کننده‌های 3D - Morse (نمایش سه بعدی ساختار مولکول براساس تفرق الکترون) از طریق معادله تبدیلی که در پراش الکترون استفاده می‌شود، محاسبه می‌گردند:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (۴-۴)$$

I شدت الکترون پراکنده شده، A_iوA_j خاصیت اتمی اتم i و j، s زاویه پراکندگی، r_{ij} فاصله بین اتم‌های i و j و N نیز تعداد کل اتم‌ها را نشان‌می‌دهد. این روش باعث می‌شود که ساختار سه بعدی مولکول به یک کد ثابت تبدیل شود[۴۸].

توصیف‌کننده Mor09m از این گروه توصیف‌کننده‌ها است که وارد مدل برتر شده است. این توصیف‌کننده آرایش سه بعدی اتم‌ها را که با جرم اتمی وزن دار شده بیان می‌کند. میزان تأثیر آن بر شاخص بازداری منفی است. بنابراین با افزایش این توصیف‌کننده شاخص بازداری کاهش می‌یابد.

۴-۹-۷- توصیف‌کننده‌هاGETAWAY

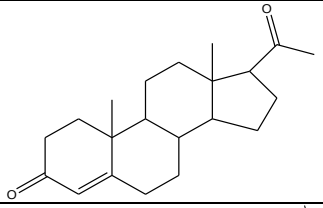
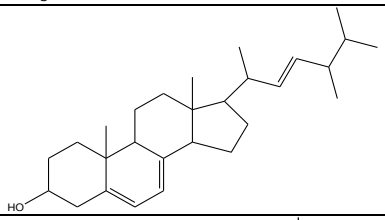
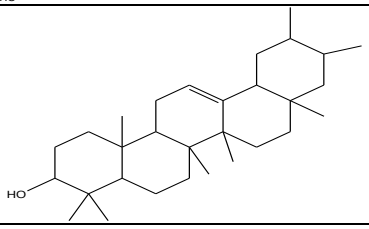
دو نوع توصیف‌کننده‌یGETAWAY طراحی شده‌اند، یک گروه از این توصیف‌کننده‌ها به نام H- GETAWAY از یک نمایش جدید از ساختار مولکولی به نام ماتریس تأثیر مولکولی[2] که با نماد H نمایش داده می‌شود به دست می‌آیند. این توصیف‌کننده‌ها به الکترونگاتیویته، اندازه اتمی و محل قرار گرفتن اتم‌ها در مولکول بستگی دارند. هر چه الکترونگاتیویته، اندازه اتمی و فاصله بین اتم و مرکز مولکول بیشتر باشد، مقدار این توصیف‌کننده بیشتر است.

گروه دیگر، توصیف‌کننده‌های R-GETAWAY هستند که R8e و R7m از این نوع هستند و از ماتریس اثر- فاصله (R) گرفته شده‌اند و در آن عناصر ماتریس تأثیر مولکولی با عناصر ماتریس

1-3D- Molecular Representation of Structure based on Electron diffraction
2 - Molecular Influence Matrix

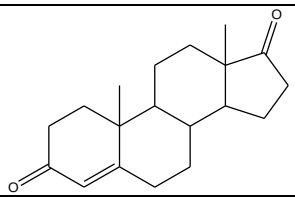
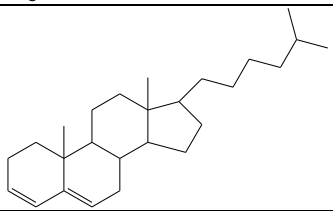
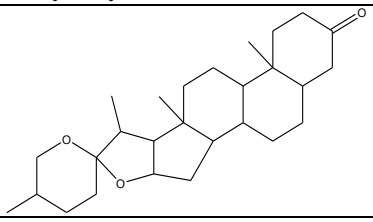
هندسی[1] ترکیب شده‌اند[۵۹]. توصیف‌کننده R8e با الکترونگاتیویته ساندرسون وزن‌دار شده و اثر آن روی شاخص بازداری مثبت بوده یعنی افزایش الکترونگاتیویته مولکول باعث افزایش این توصیف‌کننده و به تبع آن شاخص بازداری افزایش می‌یابد. ولی توصیف‌کننده R7m با جرم اتمی وزن‌دار شده و اثر آن روی شاخص بازداری منفی می‌باشد(جدول ۴-۱۶). جداول (۴-۲۰) و (۴-۲۱) چگونگی ارتباط این توصیف‌کننده‌ها را با شاخص بازداری برای سه مولکول از سری داده‌ها نشان می‌دهد.

جدول (۴-۲۰)- مثال‌هایی از اثر توصیف‌کننده R8e بر شاخص بازداری

نام ترکیب	ساختار	مقدارR8e	شاخص بازداری
Progesterone		۰/۰۸۲	۲۸۴۳
Ergosterol, (3 β)-ergosta-5,7,22-trien-3-ol		۰/۰۹۲	۳۱۵۲
α-Amyrine		۰/۱۲۲	۳۳۵۴

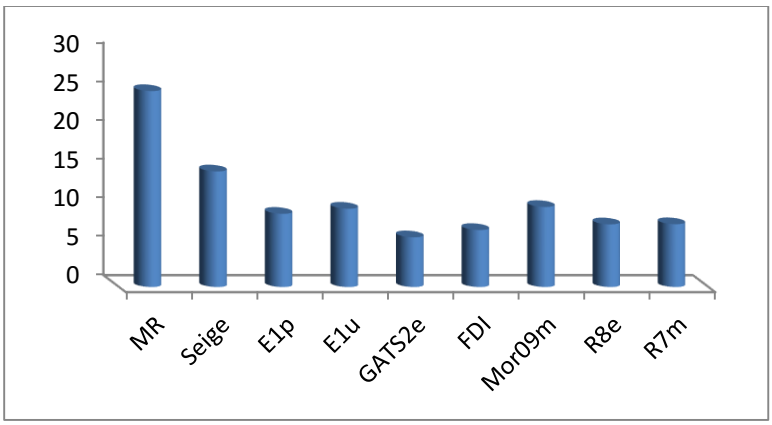
1-Geometry Matrix

جدول (۴-۲۱)- مثال‌هایی از اثر توصیف‌کننده R7m بر شاخص بازداری

نام ترکیب	ساختار	مقدارR8e	شاخص بازداری
Androst-4-ene-3,17-dione		۰/۰۳۵	۲۶۳۱
Cholesta-3,5-diene		۰/۰۲۶	۲۸۷۴
Sarsasagenone		۰/۰۱۹	۳۳۲۱

۴-۱۰- بررسی میزان مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی

بر اساس روشی که در بخش (۳-۸) بیان شد، میزان مشارکت توصیف‌کننده‌های انتخاب شده در پیش‌بینی اندیس بازداری موردنظر برآورد گردیده که نتایج آن در شکل ۴-۱۸ ارائه شده است.



شکل(۴-۱۸)- مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینه

بر اساس شکل فوق، توصیف‌کننده MR بیشترین اثر مشارکت را دارا می‌باشد. توصیف‌کننده‌ی دیگر که درصد مشارکت بالایی دارد، توصیف‌کننده SEige می‌باشد. با توجه به اینکه این توصیف‌کننده‌ها به بزرگتر شدن و پرشاخه شدن مولکول وابسته‌اند و از طرفی این دو عامل یعنی شاخه‌دار شدن و اندازه مولکول روی شاخص بازداری تأثیر مستقیم دارند می‌توان درصد مشارکت بالای این دو توصیف‌کننده را توجیه کرد.

۴-۱۱- نتیجه‌گیری

با توجه به تنوع گسترده استرول‌ها و کاربرد این ترکیبات ، مطالعه بر روی آن‌ها در زمینه‌های مختلف امری اجتناب ناپذیر به نظر می‌رسد. بنابراین با صرف زمان و هزینه کمتر، می‌توان با استفاده از مدل-های معتبر خطی و غیرخطی شاخص بازداری را برای دسته‌ای از این ترکیبات با درصد اطمینان مشخص تخمین زد و سپس آن را برای سری مشابهی از ترکیبات که تاکنون سنتز نشده‌اند پیش‌بینی نمود و از آنجا که اندازه‌گیری این پارامتر در شرایط خاصی از ستون کروماتوگرافی با برنامه دمایی متفاوت ستون صورت می‌گیرد، لذا برای مدل‌سازی شاخص بازداری و تعریف دسته‌ای از توصیف-کننده‌ها برای آن‌ها، بررسی ستون اندازه‌گیری‌کننده لازم است.

آینده‌نگری

- در این تحقیق از روش‌های SR-ANN و GA-ANN برای پیش‌بینی شاخص بازداری ترکیبات VOC و استرول‌ها استفاده شده که نتایج ارزیابی مدل منتخب در هر دو بخش، تا حدی قابلیت تعمیم آن را برای پیش‌بینی شاخص بازداری ترکیباتی که در این مدل‌سازی‌ها به کار نرفته یا حتی هنوز سنتز نشده، نشان می‌دهد.

- در این تحقیق برای انتخاب توصیف‌کننده‌های برتر از روش رگرسیون خطی و الگوریتم ژنتیک استفاده شده و سپس توصیف‌کننده‌های حاصل از این دو روش به عنوان ورودی به شبکه عصبی مصنوعی خورانده شدند. علاوه بر آن می‌توان از روش‌هایی مثل بهینه‌سازی اجتماع مورچگان[1] و طرح‌ریزی پیاپی[2] برای انتخاب توصیف‌کننده‌های مهم بهره برد و از توصیف‌کننده‌های انتخابی این روش‌ها برای ساخت مدل استفاده کرد و آنگاه مقایسه‌ای بین مدل‌ها انجام داد.

- روش‌هایی مانند KPLS، poly-PLS، SVM و LS-SVM[3] را می‌توان به جای شبکه عصبی مصنوعی یا در کنار آن برای مدلسازی غیرخطی استفاده کرده و نتایج را با هم مقایسه نمود.

1- Ant Colony Optimization
2- Successive Projection Algorithm (SPA)
3- Least Square Supported Vector Machine

فهرست منابع

[1] De Blas M., Navazo M., Alonso L.,Durana N., Iza J., (2012), “ Automatic on-line monitoring of atmospheric volatile organic compounds: Gas chromatography-mass spectrometry and gas chromatography-flame ionization detection as complementary systems”, *Science of the Total Environment*, **409**, pp 5459- 5469.

[۲] راجر.ام.اسمیت، (۱۳۸۳)، "کروماتوگرافی گاز و مایع در شیمی تجزیه" ارباب زوار م. سرافراز یزدی ع.، انتشارات دانشگاه فردوسی مشهد، ۴۸۶، ص ۱۸۵-۱۸۴ .

[۳] رووד، دین، (۱۳۸۳)، "روش‌های کروماتوگرافی: راهنمای عملی استفاده، نگهداری و رفع عیب از دستگاه‌های کروماتوگرافی" امامی س. رحیمی ح.، چاپ اول، انتشارات نورپردازان تهران، ص ۵۲ .

[۴] غفاری راد ح.، (۱۳۸۸)، پایان نامه کارشناسی ارشد، "مدل‌سازی QSPR فاکتور بازداری ترکیبات آلی بر روی فازهای ثابت تری فلوئوروپروپیل"، دانشکده شیمی، دانشگاه صنعتی شاهرود، ص ۲.

[5] Ghavami R., Faham S., (2010), “QSPR Models for kov′ats’Retention Indicesof avariety of Volatile Organic Compounds on Polar and A polar GC stationary phasesusing Molecular Connectivity indexes”,*Chromatographia*, **72**, pp 893-903.

[6]Sarkhosh M., GhasemiJ. B. , and Ayati M. , (2012), “A quantitative structure-property relationship of gas chromatographic/ mass spectrometric retention data of85 volatile organic compounds as pollutant materials by multivariate methods”, *Chemistry Centraljournal*, 6(suppl 2):S4.doi:10.1186/1752-153X-6-S2-S4.

[7] Enix G. W., Zwanziger H. W., Geiss S.,(1997), “*Chemometrics in Environmental Analysis*”, Winheim, VCH.

[8] Wold S., Sjostrom M., (1998), *Chemmometrics & Intelligent Laboratory Systems*. 44,33.

[9] http:// www.Home. Neo. Rr.com/Catbar/Chem- Txt. Htm.

[10] Mendeleev D.I., (1891), Principle of Chemistry. London: Longmans. 5 th ed.

[11] Borman s., (1990), “New QSAR techniques eyed for environmental assess-ments”,*Chemical & Engineering News*, **68**, pp 20-23.

[12] Crum B.A. and Fraser T. R., (1869), “ On the connection between chemical constitution physical action. Patr I. On the physiological action of the ammonioum based, derived from strychia, Brucia, Thebabi, codeia, Morphia and Nicotia.”*Transaction of the Royal Society*, **25**, pp 151-203.

[13] Crum B. A. and Fraser T. R., (1869), “ On the connection between chemical constitution physical action. Patr II. On the physiological action of the ammonioum based, derived from strychia, Brucia, Thebabi, codeia, Morphia and Nicotia.”*Transaction of the Royal Society*, **25**, pp 693-739.

[14] Lipnick R. L. and Overton E., (1986), “Narcosis studies and a contribution to general pharmacology”, *Trends in Pharmacological Sciences..*, **7**, pp 161-164.

[15] Katritzky A. R., Dobchev D. A., Karelson M. (2006), “Physical, Chemical and Thecnological Property Correlation with Chemical Structure: Potential of QSPR”, *Zertschrift furNaturforschung- B.*,**61**, PP 373-384.

[16] Gharagheizi F., Sattari M., (2008), “Prediction of Some Important Physical Properties of Sulfur Compounds Using QSPR Models”, *MolecularDiversity*,**12**, pp 143-155.

[17] Todeschini R.and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley-VCH, New York, pp 98.

[18] HyperChem8.0 for Windows, (2007), Toronto, Canada:Hypercube Inc, http:// www.hyper.com.

[۱۹] باهر ا.،(۱۳۸۸)، پایان نامه دکتری،"پیش‌بینی و مطالعه QSPRبازداری ترکیبات آلی، دارویی و آلاینده‌ها با استفاده از رگرسیون بردارهای پشتیبان"، دانشکده شیمی، دانشگاه مازندران.

[20] Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley-VCH, New York, pp 307.

[۲۱] یزدان دوست ح. ،(۱۳۹۰)، پایان نامه کارشناسی ارشد، " پیش بینی شاخص بازداری تعدادی روغن های ضروری با استفاده از روش‌های QSPR خطی و غیر خطی"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[22] Goldberg D. E., (1989), “Genetic Algorithms in Search, Optimization and Machine Learning. Addison – Wesley. Reading, MA.

[۲۳] صادقیه ا.، (۱۳۸۵)، "تصمیم‌گیری بر اساس الگوریتم ژنتیک در بهینه‌سازی"،انتشارات دانشگاه یزد.

[۲۴] باوی ا. ، صالحی م. ، (۱۳۸۹)، " الگوریتم‌های ژنتیک و بهینه‌سازی سازه‌های مرکب"، چاپ دوم، انتشارات عابد، ص ۲۵ .

[25] Gaurdal Z, Haftka RT., “Optimization of composite laminates”.Presented at the NATO Advanced Study Institute on Optimization of Large Structural Systems, Berchtesgaden, Germany, September 23 october 4,1991.

[26] Toplis J. G. and Edward R. P., (1979), “Chance Fraction in Studies of Quantitative Structure- Activity Relationships”, *Journal of Medicinal Chemistry*,**22**, pp 1224-1238.

[27] http://pro-programming.com/p=134

[۲۸]- منهاج م.، (۱۳۸۹)، "مبانی شبکه‌های عصبی"، چاپ هفتم، انتشارات دانشگاه صنعتی امیرکبیر، ص ۲۰.

[29] MATLAB, (The Language of Technical Computing), Version 7.6.0.324, (R2008a), The Math Works Inc.,February 10, 2008.

[30] Zupan J.and Gasteiger J., (1993), “*Neural Networks for Chemist*”, VCH Publishers, New York, Chapter 2.

[۳۱]- کیا م.، (۱۳۸۷)، "*شبکه‌های عصبی در MATLAB* "، چاپ دوم، انتشارات کیان رایانه سبز.

[32]http://foram. Takdownload.ir/threads/25029.

[33]http://openai.sourceforge.net/docs/nn_algorithms/networksarticle/index.html.

[۳۴]- اشرفی م.، (۱۳۸۹)، پایان نامه کاشناسی ارشد، "مطالعه ارتباط کمی ساختار- فعالیت مشتقات تیوکربامات‌ها به عنوان دسته‌ی جدیدی از بازدارنده‌های غیر نوکلئوزیدی HIV."، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[35] Jalali-Heravi M., Mani-Varnosfaderani A., (2009), “QSAR Modeling of 1-(3,3-diphenilpropyl)-Piperidinylamides as CCR_5 Moodulators Using Multivariate Adaptive Regression Spline and Bayesian Regularization Genetic Neural Networks”, *QSAR Combinatorial Science*,**9**, pp 946-958.

[36] Caballero J., Fernandes M., (2006), “Linear and Nonlinear Modeling of Antifungal Activity of Some Heterocyclic Ring Derivatives Using Multiple Linear Regression and Bayesian- Regularized Neural Networks”, *Journal of Molecular Modeling*, **12**, pp 168-181.

[37]Caballero J., Fernandes M., (2006), “Linear and Nonlinear Modeling of Antifungal Activity of Some Heterocyclic Ring Derivatives Using Multiple Linear Regression and Bayesian- Regularized Neural Networks”,*Journal of Molecular Modeling*, **12**, pp 1255-1268.

[38] Rawling J. O., (1988), “*Applied Regression Analysis*”, Wadsworth&Brooks/Cole, Pacific Grove (CA).

[39] Allen D. M., (1971), “Mean Square Error of Prediction as a Criterion for Selecting Variables”, *Technometrics*, **13**, pp 469-475.
[40] Cruciani G., Baroni M., Clementi S., Constantino G., Riganelli D., and Skagerberg B., (1992), “Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDPE)”, *journal of chemometrics*., **6**, pp 335-346.
[41] Todeschini R., Consonni V., (2000), *“Handbook of molecular descriptors”*, Wiley, New York, pp 461.

[42] HyperChem7.0 Toronto, Canada: HyperCube Inc, http:// www.hyper.com.

[43]Todeschini R. Milano Chemometrics and QSPR Group, http://www. disat.unimib.it/vhml.

[44] SPSS for windows Statistical package for IBM PC, SPSS Inc, http: //www. spss.com.

[45] Zahouily M., Rakik J., Lazar M., Banlaoui M. A., Rayadh A., Komiha N., (2007), “Exploring QSAR of Non-Nucleoside Reverse Transcriptase Inhibitors by Artificial Neural Networks: HEPT Derivatives”, *ARKIVOC*, PP 245-256.
[46] Leardi R., Lupianez Gonzalez A., (1998), “ Genetic algorithms applied to feature selection in PLS regression: how and when to use them ”, *Chemometrics and Intelligent Laboratory System* , **41**, pp 195-207.
[47] Ghasemi J., Niazi A., Leardi R., (2003), “Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture”, *Talanta*, **59**, pp 311-317.
[48] CODESSA 2.13, Semichem, 7204 Mullen, Shawnee, KS 66216, U.S.A., E-mail andy@semichem.com, www http://www.semichem.com.
[49] Todeschini R., Consonni V., (2000), *“Handbook of molecular descriptors”*, Wiley, New York, pp 513-514.
[50] Hemmer M.C., Steinhauer V., Gasteiger J., (1999),”The prediction of the 3D structure of organic molecules from their infrared spectra”, *Vibrational Spectroscopy Journal*. ,**19**, pp 151-164.
[51] Todeschini R., Consonni V., (2000), *“Handbook of molecular descriptors”*, Wiley, New York, pp 366-367.

[52] Todeschini R., Consonni V., (2000), *"Handbook of molecular descriptors"*, Wiley-VCH, Weinheim, Germany.

[53] Douali L., Villemin D., Cherqaoui D.; (2003), "Neural networks: Accurate nonlinear QSAR model for HEPT derivatives", *Journal of Chemical Information and Computer Sciences*, **43**, pp 1200-1207.

[54] Cabarello J., Garriga M., Fernandez M., (2006), "2D autocorrelation modeling of negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks",*Bioorganic & Medicinal Chemistry Journal*, **14**, pp 3330-3340.

[55] chen Z., Zhang Y., Fu W., (2010),"QSAR study of carboxylic acid derivatives as HIV-1integrase inhibitors", *European Journal of Medicinal Chemistry*, **45**, pp 3970-3980.

[56] http://en. Wikipedia. Org/wiki/sterol.

[57] Valery A. Isidorov, Lech Szczepaniak, (2009), "Gas chromatographic retention indices of biologically and environmentally important organic compounds on capillary columns with low- polar stationary phases", *Journal of Chromatography A*, **1216**, pp 8998- 9007.

[58] Alan R. Katritzky, Liliana Pacureanu, Dimitar Dobchev, Mati Karelson, (2007), "QSPR modeling of hyperpolarizabilities",*Molecular Modeling*, 13(9), 951-63.

[۵۹]- سلیمی ه.، (۱۳۸۹)، پایان نامه کارشناسی ارشد، "پیش بینی پارامتر حلالیت تعدادی حلال آلی با استفاده از روش‌هایQSPR. دانشکده شیمی، دانشگاه صنعتی شاهرود"، خطی و غیر خطی

[60] http://www.Strandls.com/sarchitect/documents/manual-html/desctheory.html.

[61] Todeschini R., Gromatica P., (1997), "3D- modeling and prediction by WHIM descriptors. Part5. Theory development and chemical meaning of WHIM descriptors", *Quantitative Structure -Activity Relationship*., **16**, pp 113-119.

[62] www.igi.global.com/chapter/nanoparticles-towards-predicting.../56450.

Abstract

In the first section of this project, quantitative structure - property relationship (QSPR) has been developed for modeling and prediction of retention indices of 60 volatile organic compounds (VOCs). The stepwise regression (SR) and genetic algorithm (GA) were used as variable selection methods. The selected variable by these methods was used as input of artificial neural network (ANN) to construct of model to predict of retention indices of these compounds. The obtained models were validated by different techniques such as using the external test set, leave one out (LOO) and y-randomization. The obtained results Shaw that both models of SR-ANN and GA-ANN have a good prediction ability. The determination coefficient of the test set for SR-ANN and GA-ANN were obtained 0.995, 0.992 respectively.

In the second part of this project, a QSPR model has been developed to predict of retention indices of some sterol compounds, same the first part, SR and GA were used as variable selection methods. Also, ANN technique was applied for modeling and prediction of retention indices of these compounds. The obtained results of determination coefficient of the test set for SR-ANN and GA-ANN were obtained 0.944, 0.920 respectively.

Keyword: Retention index, QSPR, SR-ANN, GA-ANN.

Shahrood University of Technology

Faculty of Chemistry

Prediction of retention index of atmospheric volatile organic compounds data set using QSPR methods

Sara Mesgaran karimi

Supervisor:

Dr. N. Goudarzi

Advisor:

Dr. M. Arab Chamjangali

Date:**Feb** – 2013