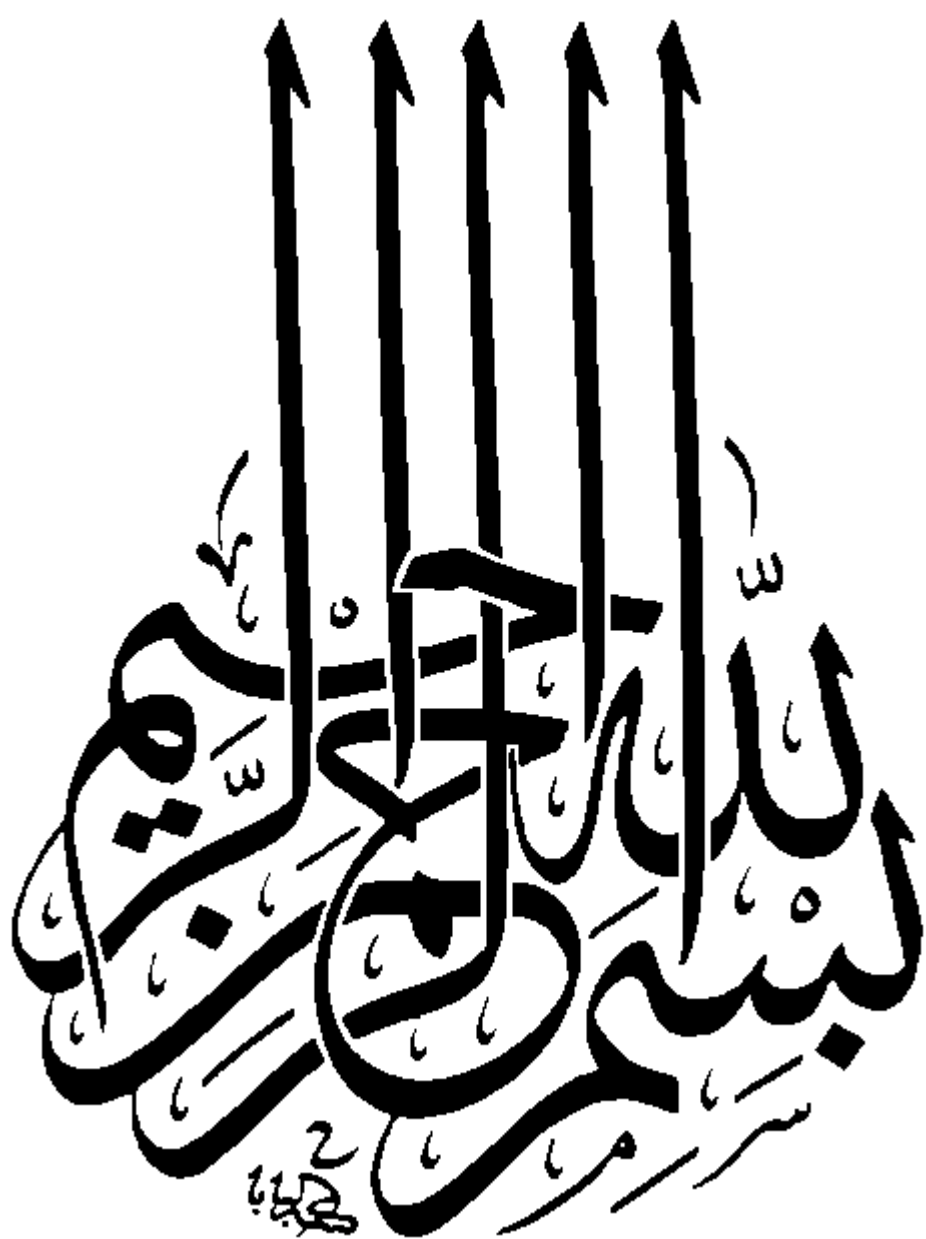






دانشکده: علوم پایه

گروه: شیمی تجزیه

پیش‌بینی شاخص بازداري تعدادی روغن‌های ضروری با استفاده

از روش‌های QSPR خطی و غیرخطی

دانشجو: حانیه یزدان دوست

استاد راهنما:

دکتر ناصر گودرزی

استاد مشاور:

دکتر زهرا کلانتر

پایان نامه جهت اخذ درجه کارشناسی ارشد

تیر ماه ۱۳۹۰

خدای متعال را سپاس که نعمتش بی شمار، قدرتش بی انتها، رحمتش فراگیر و الطافش لایتناهی است؛
او که در خطه زندگی ما وروپشتیانم بوده و در سختی ما و ناملایمات تکیه گاهم...

تقدیم به:

پدر و مادر بزرگوار، خواهر عزیز و همسر مهربانم
به سگزانہ زحمات بی دریغ، حمایت ما و دلسوزی ما نشان

ابر از شکر و قدردانی

حال که به یاری پروردگار، این دوره از تحصیلات خود را به پایان رساندم، بر خود لازم می‌دانم از زحمات بی‌شائبه‌ی پدر و مادر نازنینم و خواهر عزیزم که همواره مشوق و پشتیبان من بوده‌اند و نیز همسر فداکارم که صبورانه ادامه این راه را برایم هموار نمود تشکر کنم.

از اساتید گرامی جناب آقای دکتر ناصر کوردزی که زحمت راهنمایی این پایان نامه را بر عهده داشتند و سرکار خانم دکتر زهرا کلاتر به عنوان استاد مشاور سپاسگزاری می‌نمایم. همچنین از تمامی دوستانم و کسانی که مراد انجام این پروژه یاری کردند صمیمانه قدردانی می‌کنم.

چکیده:

روغن‌های ضروری گیاهی و ترکیبات مرتبط، کاربرد گسترده‌ای در طب سنتی، صنایع غذایی، صنایع عطرسازی و داروسازی دارند. در این تحقیق، مطالعات ارتباط کمی ساختار - ویژگی بر روی شاخص بازداري ۷۷ و ۱۱۹ روغن ضروري به ترتيب در بخش‌های سوم و چهارم انجام پذیرفت. رگرسیون مرحله‌ای خطی برای انتخاب توصیف‌کننده‌ها به کار گرفته شد. برای مدل‌سازی اندیس بازداري این ترکیبات از روش رگرسیون خطی چندگانه (MLR) به عنوان روش خطی و همچنین شبکه عصبی مصنوعی (ANN) به عنوان روش غیر خطی استفاده شد. برای بررسی اعتبار این مدل‌ها علاوه بر به کارگیری سری تست، از روش‌های دیگری همانند حذف مرحله‌ای تک‌تک داده‌ها و Y -تصادفی استفاده گردید. هر دو روش خطی و غیرخطی توانایی پیش‌بینی خوبی دارند، اما با توجه به نتایج حاصله، مدل ANN دارای قدرت پیش‌بینی بیشتری در هر دو مورد بوده است. در قسمت اول این تحقیق، ضرایب همبستگی سری تست توسط مدل‌های MLR و ANN به ترتیب ۰/۹۷۵ و ۰/۹۸۲ و در قسمت دوم آن، ضرایب همبستگی سری تست مدل‌های MLR و ANN به ترتیب ۰/۹۴۱ و ۰/۹۵۶ بودند.

کلیدواژه‌ها: ارتباط کمی ساختار - ویژگی، رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، شاخص بازداري

نتایج حاصل از این پایان نامه در دو پوستر با عناوین:

“Prediction of gas chromatographic retention indices of essential oils with Linear and nonlinear methods”

9

“Quantitative structure-property relationship study of chromatographic retention indices for essential oils using linear and non-linear methods”

در هفدهمین سمینار شیمی تجزیه ایران در دانشگاه کاشان ارائه گردید.

فهرست مطالب

فصل اول: مقدمه

۱-۱- پیشگفتار..... ۲

۱-۱-۱- روغن‌های ضروری..... ۳

۱-۲- شاخص بازداري..... ۴

۲-۱- هدف تحقيق..... ۵

۳-۱- پيشينه تحقيق..... ۶

فصل دوم: کمومتریکس

۱-۲- کمومتریکس..... ۱۰

۲-۲- ارتباط کمی ساختار - ویژگی..... ۱۲

۲-۲-۱- جمع‌آوری سری داده‌ها..... ۱۳

۲-۲-۲- بهینه‌سازی ساختار هندسی..... ۱۴

۲-۲-۳- محاسبه توصیف‌کننده‌ها..... ۱۵

۲-۲-۴- انتخاب بهترین توصیف‌کننده‌ها..... ۱۶

۲-۲-۵- انتخاب مدل مناسب..... ۱۷

۲-۲-۵-۱- ورود اجباری..... ۱۸

۲-۲-۵-۲- روش برازش مرحله‌ای..... ۱۸

۲-۲-۵-۳- حذف پس‌رو..... ۱۸

۲-۲-۵-۴- انتخاب پیش‌رو..... ۱۹

۲-۳- مدل‌سازی با استفاده از روش MLR..... ۱۹

۲-۴- شبکه عصبی مصنوعی..... ۲۰

۲۱	۲-۴-۱- سیستم عصبی زیستی
۲۲	۲-۴-۲- سیستم عصبی مصنوعی
۲۵	۲-۵- معماری شبکه‌های عصبی
۲۶	۲-۶- آموزش شبکه عصبی
۲۸	۲-۷- انواع شبکه‌های عصبی
۲۸	۲-۷-۱- شبکه‌های پیش‌خور
۲۹	۲-۷-۲- شبکه‌های پس‌خور یا برگشتی
۲۹	۲-۸- طراحی شبکه پیش‌خور
۳۰	۲-۹- آموزش شبکه‌های پیش‌خور
۳۰	۲-۹-۱- ایجاد یک شبکه پیش‌خور
۳۰	۲-۹-۲- مقداردهی آغازین به وزن‌ها
۳۱	۲-۹-۳- آموزش شبکه
۳۱	۲-۹-۴- شبیه‌سازی
۳۱	۲-۱۰- الگوریتم‌های آموزشی شبکه عصبی
۳۳	۲-۱۰-۱- الگوریتم آموزشی نزول گرادیانی
۳۴	۲-۱۰-۲- الگوریتم‌های آموزشی لونبرگ-مارکوارت
۳۷	۲-۱۱- بهبود تعمیم یا ارتقاء عمومیت
۳۸	۲-۱۱-۱- تنظیم
۳۹	۲-۱۱-۲- توقف زودهنگام
۴۰	۲-۱۲- ارزیابی مدل
۴۳	۲-۱۳- تکنیک‌های اعتبارسنجی
۴۴	۲-۱۳-۱- ارزیابی متقاطع یا اعتبارسنجی تقاطعی
۴۴	۲-۱۳-۲- تقسیم کردن مجموعه به دو سری آموزش و ارزیابی

۴۵	۲-۱۳-۳- اعتبارسنجی بیرونی
۴۵	۲-۱۳-۴- آزمون Y-تصادفی
۴۵	۲-۱۴- نرم افزارهای مورد استفاده در این تحقیق
۴۶	۲-۱۴-۱- نرم افزار Hyperchem
۴۶	۲-۱۴-۲- نرم افزار Dragon
۴۸	۲-۱۴-۳- نرم افزار SPSS
۴۸	۲-۱۴-۴- نرم افزار MATLAB

فصل سوم: مدل سازی QSPR شاخص بازداری برخی از اجزای فرآر روغن های ضروری، با استفاده از

روش رگرسیون خطی چندگانه (MLR) و شبکه عصبی مصنوعی (ANN)

۵۰	۳-۱- سری داده ها
۶۰	۳-۲- رسم و بهینه سازی ساختمان هندسی و به دست آوردن توصیف کننده ها
۶۱	۳-۳- انتخاب بهترین توصیف کننده ها و ایجاد مدل نهایی با روش رگرسیون خطی چندگانه
۶۵	۳-۴- مدل سازی با استفاده از شبکه عصبی مصنوعی (ANN)
۶۶	۳-۴-۱- بهینه سازی
۷۳	۳-۴-۱-۱- بهینه سازی تعداد چرخه های آموزشی
۷۴	۳-۴-۱-۲- بهینه سازی پارامتر μ
۷۶	۳-۴-۱-۳- مشخصات شبکه عصبی مصنوعی بهینه شده
۷۷	۳-۵- ارزیابی مدل های خطی و غیر خطی
۷۷	۳-۵-۱- ارزیابی مدل های MLR و ANN به کمک سری های ارزیابی و تست
۸۴	۳-۵-۲- ارزیابی مدل های خطی و غیر خطی به روش حذف مرحله ای تک تک داده ها
۹۲	۳-۵-۳- ارزیابی شبکه به روش Y-تصادفی
۹۲	۳-۵-۴- ارزیابی مدل های برتر خطی و غیرخطی با استفاده از پارامترهای آماری

- ۹۳-۶-۳ بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اندیس بازداري ترکیبات.....
- ۹۴-۶-۳-۱ توصیف‌کننده‌های Dz و Whetp و VEA2.....
- ۹۵-۶-۳-۲ توصیف‌کننده‌های Mor29p و Mor28e.....
- ۹۶-۶-۳-۳ توصیف‌کننده RDF095e.....
- ۹۷-۳-۷ بررسی میزان مشارکت توصیف‌کننده‌های منتخب توسط شبکه عصبی.....
- ۹۹-۳-۸ نتیجه‌گیری.....
- فصل چهارم: مدل‌سازی QSPR اندیس بازداري (RI) روغن‌های ضروري با استفاده از روش رگرسیون خطی چندگانه (MLR) و شبکه عصبی مصنوعی (ANN)
- ۱۰۱-۴-۱ سری داده‌ها.....
- ۱۱۳-۴-۲ رسم و بهینه‌سازی ساختار هندسی ترکیبات.....
- ۱۱۳-۴-۳ محاسبه‌ی توصیف‌کننده‌های مولکولی.....
- ۱۱۴-۴-۴ انتخاب بهترین توصیف‌کننده‌ها به روش رگرسیون خطی چندگانه.....
- ۱۱۸-۴-۵ مدل‌سازی با استفاده از شبکه عصبی مصنوعی (ANN).....
- ۱۱۹-۴-۵-۱ بهینه‌سازی.....
- ۱۲۶-۴-۵-۱-۱ بهینه‌سازی تعداد چرخه‌های آموزشی.....
- ۱۲۷-۴-۵-۱-۲ بهینه‌سازی پارامتر μ
- ۲۹۱-۴-۵-۱-۳ ساختار شبکه عصبی مصنوعی بهینه شده.....
- ۱۳۰-۴-۶ ارزیابی مدل‌ها.....
- ۱۳۰-۴-۶-۱ ارزیابی مدل‌های MLR و ANN به کمک سری‌های ارزیابی و تست.....
- ۱۳۹-۴-۶-۲ ارزیابی مدل‌های خطی و غیر خطی به روش حذف مرحله‌ای تک‌تک داده‌ها.....
- ۱۵۰-۴-۶-۳ ارزیابی شبکه به روش Y-تصادفی.....
- ۱۵۰-۴-۶-۴ ارزیابی مدل‌های برتر خطی و غیر خطی با استفاده از پارامترهای آماری.....

۵۱۱.....	۷-۴- بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اندیس بازداری ترکیبات.....
۱۵۱.....	۴-۷-۱- توصیف‌کننده Ss.....
۱۵۳.....	۴-۷-۲- توصیف‌کننده‌های MATS1e و MATS1p.....
۵۳۱.....	۴-۷-۳- توصیف‌کننده BAC.....
۱۵۴.....	۴-۷-۴- توصیف‌کننده RDF020m.....
۱۵۴.....	۴-۷-۵- توصیف‌کننده R3v_A.....
۵۵۱.....	۴-۷-۶- توصیف‌کننده MLOGP.....
۱۵۶.....	۴-۷-۷- توصیف‌کننده Mor01v.....
۱۵۷.....	۴-۸- بررسی میزان مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی.....
۵۸۱.....	۴-۹- نتیجه‌گیری.....
۱۵۹.....	آینده‌نگری.....
۱۶۱.....	فهرست منابع.....

فهرست اشکال

صفحه	عنوان
۱۳	شکل ۱-۲- نمایش خلاصه مراحل به کار گرفته شده در یک فرایند QSPR
۲۲	شکل ۲-۲- سلول عصبی زیستی
۲۳	شکل ۳-۲- مدل ریاضی نورون
۲۵	شکل ۴-۲- برخی از توابع انتقال در شبکه‌های عصبی
۲۵	شکل ۵-۲- نمایش نرون تک‌لایه با یک بردار ورودی
۲۷	شکل ۶-۲- شیوه یادگیری با ناظر
۲۸	شکل ۷-۲- شماییک شبکه عصبی مصنوعی از نوع پیش‌خور
۲۹	شکل ۸-۲- شماییک شبکه عصبی مصنوعی از نوع پس‌خور
۳۰	شکل ۹-۲- شبکه پیش‌خور tansig/purelin
۳۹	شکل ۱۰-۲- تغییرات خطای سری آموزش و سری ارزیابی
۶۳	شکل ۱-۳- تأثیر تعداد توصیف‌کننده‌ها بر مقادیر R^2 (الف)، F (ب) و SE (ج) در مدل‌های MLR
۶۷	شکل ۲-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل لگاریتم سیگموئید
۶۸	شکل ۳-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل تانژانت سیگموئید
۶۹	شکل ۴-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل خطی
۷۰	شکل ۵-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونیبرگ-مارکوارت و تابع تبدیل لگاریتم سیگموئید
۷۱	شکل ۶-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونیبرگ-مارکوارت و تابع تبدیل تانژانت سیگموئید
۷۲	شکل ۷-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونیبرگ-مارکوارت و تابع تبدیل خطی
۷۴	شکل ۸-۳- منحنی مقدار مربع میانگین خطاها در برابر تغییر تعداد چرخه‌های آموزشی
۷۵	شکل ۹-۳- منحنی مقادیر خطای شبکه در مقابل پارامتر μ
۷۶	شکل ۱۰-۳- تصویر شماییک ساختار شبکه عصبی مصنوعی به دست آمده پس از بهینه‌سازی
۸۰	شکل ۱۱-۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری ارزیابی
۸۰	شکل ۱۲-۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش ANN برای سری ارزیابی
۸۳	شکل ۱۳-۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای سری تست توسط روش MLR
۸۳	شکل ۱۴-۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای سری تست توسط روش ANN

- شکل ۳-۱۵- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل خطی به روش حذف مرحله‌ای
۹۰ تک‌تک داده‌ها
- شکل ۳-۱۶- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل غیرخطی به روش حذف مرحله‌ای
۹۰ تک‌تک داده‌ها
- شکل ۳-۱۷- منحنی مقادیر باقیمانده بر حسب مقادیر تجربی اندیس بازداري در ارزیابی مدل MLR
۹۱
- شکل ۳-۱۸- رسم منحنی مقادیر باقیمانده در مقابل مقادیر تجربی شاخص بازداري در ارزیابی مدل ANN
۹۱
- شکل ۳-۱۹- نمایش میزان مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینه
۹۸
- شکل ۴-۱- تأثیر تعداد توصیف‌کننده‌ها بر مقادیر R^2 (الف)، F (ب) و SE (ج) در مدل‌های MLR
۱۱۶
- شکل ۴-۲- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل لگاریتم سیگموئید
۱۲۰
- شکل ۴-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل تانژانت سیگموئید
۱۲۱
- شکل ۴-۴- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل خطی
۱۲۲
- شکل ۴-۵- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونبرگ - مارکوارت و تابع تبدیل لگاریتم سیگموئید
۱۲۳
- شکل ۴-۶- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونبرگ - مارکوارت و تابع تبدیل تانژانت سیگموئید
۱۲۴
- شکل ۴-۷- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونبرگ - مارکوارت و تابع تبدیل خطی
۱۲۵
- شکل ۴-۸- منحنی مقادیر خطای شبکه در مقابل تعداد چرخه‌های آموزشی
۱۲۷
- شکل ۴-۹- منحنی مقادیر خطای شبکه در مقابل پارامتر μ
۱۲۹
- شکل ۴-۱۰- تصویر شماتیک ساختار شبکه عصبی مصنوعی به دست آمده پس از بهینه‌سازی
۱۳۰
- شکل ۴-۱۱- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری ارزیابی
۱۳۴
- شکل ۴-۱۲- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ANN برای سری ارزیابی
۱۳۴
- شکل ۴-۱۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری تست
۱۳۸
- شکل ۴-۱۴- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش ANN برای سری تست
۱۳۸
- شکل ۴-۱۵- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل خطی به روش حذف مرحله‌ای
۱۴۸ تک‌تک داده‌ها
- شکل ۴-۱۶- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل ANN به روش حذف مرحله‌ای
۱۴۸ تک‌تک داده‌ها
- شکل ۴-۱۷- رسم مقادیر خطای مطلق در ارزیابی مدل MLR در مقابل مقادیر تجربی شاخص بازداري
۱۴۹
- شکل ۴-۱۸- منحنی مقادیر خطای مطلق محاسبه شده بر حسب مقادیر اندیس بازداري در روش ANN
۱۴۹

فهرست جداول

صفحه	عنوان
۴۷	جدول ۱-۲- انواع توصیف‌کننده‌های محاسبه شده توسط نرم افزار Dragon
۵۰	جدول ۱-۳- سری داده‌های به کار رفته در مدل سازی
۶۲	جدول ۲-۳- مقایسه آماره‌های چند مدل به دست آمده از روش MLR
۶۴	جدول ۳-۳- توصیف‌کننده‌های وارد شده در مدل MLR منتخب
۶۴	جدول ۴-۳- ماتریس همبستگی توصیف‌کننده‌های به کار رفته در مدل نهایی
۶۷	جدول ۵-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۶۸	جدول ۶-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۶۹	جدول ۷-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۷۰	جدول ۸-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لونبرگ- مارکوارت (trainlm) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۷۱	جدول ۹-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لونبرگ - مارکوارت (trainlm) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۷۲	جدول ۱۰-۳- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لونبرگ - مارکوارت (trainlm) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان
۷۳	جدول ۱۱-۳- مقادیر خطای شبکه با تغییر تعداد چرخه‌های آموزشی
۷۵	جدول ۱۲-۳- مقادیر خطای شبکه با تغییر پارامتر μ
۷۶	جدول ۱۳-۳- مشخصات شبکه بهینه
۷۸	جدول ۱۴-۳- مقایسه مقادیر تجربی و پیش‌بینی شده‌ی شاخص بازدارى برای سری ارزیابی در روش MLR
۷۹	جدول ۱۵-۳- مقایسه مقادیر تجربی و پیش‌بینی شده‌ی اندیس بازدارى برای سری ارزیابی در روش ANN

- جدول ۳-۱۶- مقایسه مقادیر مورد انتظار و پیش‌بینی شده سری تست با روش MLR ۸۱
- جدول ۳-۱۷- مقادیر مورد انتظار و پیش‌بینی شده سری تست با روش ANN ۸۲
- جدول ۳-۱۸- ارزیابی مدل خطی به روش حذف مرحله‌ای تک‌تک داده‌ها ۸۴
- جدول ۳-۱۹- ارزیابی مدل غیر خطی به روش حذف مرحله‌ای تک‌تک داده‌ها ۸۷
- جدول ۳-۲۰- مقادیر R^2 برای سری ارزیابی و تست در روش Y-Randomization ۹۲
- جدول ۳-۲۱- پارامترهای آماری برای دو روش MLR و ANN ۹۳
- جدول ۳-۲۲- مثال‌هایی از اثر توصیف‌کننده Dz بر شاخص بازداری ۹۹
- جدول ۴-۱- مجموعه داده‌های به کار گرفته شده در مدل‌سازی ۱۰۱
- جدول ۴-۲- مقایسه آماره‌ی مدل‌های به دست آمده از روش MLR ۱۱۴
- جدول ۴-۳- توصیف‌کننده‌های وارد شده در مدل MLR ۱۱۷
- جدول ۴-۴- ماتریس همبستگی توصیف‌کننده‌های به کار رفته در مدل نهایی ۱۱۸
- جدول ۴-۵- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۰
- جدول ۴-۶- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۱
- جدول ۴-۷- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۲
- جدول ۴-۸- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت (trainlm) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۳
- جدول ۴-۹- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت (trainlm) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۴
- جدول ۴-۱۰- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت (trainlm) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان ۱۲۵
- جدول ۴-۱۱- مقادیر خطای شبکه با تغییر تعداد چرخه‌های آموزشی ۱۲۶
- جدول ۴-۱۲- مقادیر خطای شبکه با تغییر میزان μ ۱۲۸
- جدول ۴-۱۳- مشخصات شبکه بهینه ۱۲۹
- جدول ۴-۱۴- مقایسه مقادیر تجربی و پیش‌بینی شده در روش MLR برای سری ارزیابی ۱۳۱
- جدول ۴-۱۵- مقایسه مقادیر تجربی و پیش‌بینی شده در روش ANN برای سری ارزیابی ۱۳۲

۱۳۵	جدول ۴-۱۶ مقایسه مقادیر تجربی و پیش‌بینی شده‌ی سری تست در MLR
۱۳۶	جدول ۴-۱۷ مقایسه مقادیر تجربی و پیش‌بینی شده‌ی سری تست در ANN
۱۳۹	جدول ۴-۱۸ ارزیابی مدل خطی به روش حذف مرحله‌ای تک تک داده‌ها
۱۴۳	جدول ۴-۱۹ ارزیابی مدل غیر خطی به روش حذف مرحله‌ای تک تک داده‌ها
۱۵۰	جدول ۴-۲۰ مقادیر R^2 برای سری ارزیابی و تست در روش Y-Randomization
۱۵۱	جدول ۴-۲۱ مقایسه برخی پارامترهای آماری دو روش MLR و ANN

فصل اول

مقدمه

امروزه به دلیل رشد سریع تجهیزات مورد استفاده در شیمی (و نیز سایر زمینه های علمی) و حجم بسیار زیاد داده ها جهت جمع آوری، پردازش، تفسیر اطلاعات و استخراج نتایج مفید از آنها، استفاده از کامپیوتر اجتناب ناپذیر است. در سال های اخیر پیشرفت های زیادی در جهت تطابق بیش از پیش روش های آزمایشگاهی با محیط زیست صورت گرفته است، آنچه که امروز اصطلاحاً شیمی سبز نامیده می شود. در عین حال، گاهی شیمیدانان، با روش هایی سرو کار دارند که دارای محدودیت هایی نظیر عدم توانایی در اندازه گیری همه ترکیبات، هزینه بالا یا عدم دسترسی به امکانات مورد نظر می باشند همچنین با موادی که بسیار گران قیمت هستند یا سمی و خطرناک بوده و یا به آسانی در دسترس نیستند. در این موارد، می توان از روش های ریاضی و آماری کمک گرفت و به شکل تئوری نتایج آزمایش های مختلف را توصیف و توجیه نمود [۱]. با به کارگیری این روش ها، علاوه بر پیش بینی خواص مورد نظر، پارامترهای مؤثر بر نتایج آزمایش های تجربی و مکانیسم های درگیر در آنها نیز قابل توضیح خواهند بود. از جمله روش های نظری که در زمینه های مختلف شیمی و به خصوص شیمی تجزیه مورد استفاده قرار گرفته، کمومتریکس^۱ است که در حال حاضر به دلایل ذکر شده جایگاه برجسته ای را در نزد شیمیدانان به خود اختصاص داده است.

یکی از شاخه های مهم کمومتریکس که در این پروژه مورد استفاده قرار گرفته، ارتباط کمی ساختار - ویژگی^۲ می باشد که در این تحقیق برای پیش بینی شاخص های بازداری^۳ تعدادی از ترکیبات تشکیل دهنده روغن های ضروری^۴ در بخش های سوم و چهارم به کار گرفته شده است. از جمله روش هایی که برای مطالعه ارتباط خطی ساختار - ویژگی به کار می روند، می توان به رگرسیون خطی

۱- Chemometrics

۲- Quantitative Structure-Property Relationship (QSPR)

۳- Retention Indices (RI)

۴- Essential Oils (Eos)

چندگانه^۱، رگرسیون اجزای اصلی^۲، روش حداقل مربعات جزئی^۳ و شبکه‌ی عصبی مصنوعی^۴ اشاره نمود. در این پایان‌نامه، مطالعه ارتباط کمی ساختار - ویژگی ترکیبات موردنظر با استفاده از روش‌های خطی و غیرخطی انجام گرفت و مقایسه‌ای بین نتایج حاصل از مدل‌های خطی و غیرخطی نیز صورت پذیرفت. در بخش‌های اول و دوم تحقیق حاضر به ترتیب، ارتباط بین ساختار و شاخص بازداري ۷۷ و ۱۱۹ ترکیب از روغن‌های ضروری مورد بررسی قرار گرفت. مدل‌های به دست آمده با روش‌های مختلفی ارزیابی^۵ گردید و پارامترهای آماری مربوط به آنها نیز گزارش شد.

۱-۱-۱- روغن‌های ضروری

روغن‌های ضروری که روغن‌های فرّار یا اتری^۶ نیز نامیده می‌شوند، مایعات روغنی معطری هستند که به روش‌های تخمیر، استخراج و همینطور تقطیر با بخار آب (که در تولید تجاری از همه رایج‌تر است) از مواد خام گیاهی (گل‌ها، جوانه‌ها، دانه‌ها، برگ‌ها، چوب، میوه و ریشه‌ها) به دست می‌آیند. مشخصه مشترک اکثر این ترکیبات این است که دارای بوی تند و زننده هستند. یک روغن از این جهت اتری نام می‌گیرد که حاوی یک رایحه مشخص یا "اسانس" از یک گیاه معین است. خواص آنتی اکسیدانی، ضد میکروبی، ضد ویروسی و همینطور ضد انگلی این ترکیبات شناخته شده است. این روغن‌ها به عنوان ادویه در خوردنی‌ها و آشامیدنی‌ها، در مصارف بهداشتی نظیر صابون‌ها و شوینده‌ها، در صنعت داروسازی و همچنین مواد آرایشی و عطر و اسپری کاربرد دارند. امروزه با پیشرفت بهداشت عمومی و تلاش صنایع غذایی برای سالم‌سازی هرچه بیشتر فرآورده‌های غذایی، استفاده از این ترکیبات به عنوان جایگزینی برای مواد افزودنی سنتزی و همینطور نمک‌ها (که می‌توانند احتمال بیماری‌های قلبی را افزایش دهند) پیشنهاد شده است [۲،۳].

۱- Multiple Linear Regression (MLR)

۲- Principal Components Regression (PCR)

۳- Partial Least Square (PLS)

۴- Artificial Neural Network (ANN)

۵- Validation Techniques

۶- Ethereal Oils (Aetherolea)

سیموناتو^۱ و همکارانش در سال ۲۰۰۷ از کروماتوگرافی گازی- طیف سنجی جرمی برای تعیین کمی روغن‌های ضروری موجود در پوسته‌ی تنه‌ی درختی موسوم به کرچک هندی استفاده و شاخص بازداری مربوط به آنها را گزارش کردند [۴]. هدف آنها بررسی فعالیت‌های آنتی میکروبی و آنتی اکسیدانی این ترکیبات بود. در تحقیق حاضر این نتایج به عنوان سری داده‌ها در فصل سوم مورد استفاده قرار گرفته و کارآمدی روش‌های QSPR برای پیش‌بینی شاخص بازداری این ترکیبات بررسی شده است.

در سال ۲۰۰۰ نیز چویی^۲ و ساوامورا^۳ با استفاده از GC-MS به تعیین کیفی و کمی اجزای فرار روغن‌های گیاه هیوگاناتسو به صورت cold-pressed، در زمان‌های مختلفی از برداشت محصول، درجات متفاوتی از تازگی آن و اندازه‌های گوناگون میوه روی آوردند. همچنین مطالعه‌ای به منظور بررسی ساختار شیمیایی روغن‌های ضروری میوه هیوگاناتسو برای به‌کارگیری در مواد غذایی، آشامیدنی‌ها و محصولات دیگر انجام گرفت [۵]. نتایج حاصل از این مطالعات به عنوان سری داده‌ها در فصل چهارم مورد استفاده قرار گرفت.

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 معین 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. مقدار شاخص بازداری از معادله ۱-۱ محاسبه می‌شود:

۱- Simionatto

۲- Choi

۳- Sawamura

$$I = 100Y(Z - Y) \left[\frac{\log t'_X - \log t'_Y}{\log t'_Z - \log t'_Y} \right] \quad \text{معادله (۱-۱)}$$

در معادله ی ۱-۱، t'_Y زمان بازداري اصلاح شده برای آلکان نرمال که قبل از ترکیب X از ستون خارج می‌شود، t'_X زمان بازداري اصلاح شده برای ترکیب مورد نظر و t'_Z نیز زمان بازداري اصلاح شده برای آلکان نرمال که بعد از ترکیب X از ستون خارج می‌شود، می‌باشد. همچنین Y بیان‌گر تعداد کربن‌های آلکان نرمالی است که قبل از ترکیب مورد نظر از ستون خارج شده و Z نشان‌دهنده تعداد کربن‌های آلکان نرمالی است که بعد از ترکیب مورد نظر از ستون خارج می‌شود. به عنوان مثال، پیک ترکیبی با شاخص بازداري ۱۲۵۰ در کروماتوگرام دقیقاً در میان پیک‌های $C_{12}H_{26}$ (با شاخص بازداري ۱۲۰۰) و $C_{13}H_{28}$ (با شاخص بازداري ۱۳۰۰) قرار می‌گیرد. شاخص بازداري زمانی مفید است که پیش‌بینی ترتیب خروج نسبی اجزاء نمونه از ستون مورد نظر باشد. اگر از شاخص بازداري با چنین منظوری استفاده شود، باید در نظر داشت که شاخص بازداري در یک شرایط دمایی معین و برای یک ستون خاص قابل کاربرد است. از شاخص بازداري، برای مقایسه بازداري و گزینش‌پذیری ستون‌های مشابه (ستون‌هایی که از لحاظ جنس و ابعاد لوله موئین، نوع و ضخامت فاز ساکن و سایر موارد کاملاً مشابه اند) نیز می‌توان استفاده کرد [۶].

۱-۱- هدف تحقیق

با توجه به کاربرد گسترده‌ی روغن‌های ضروری، اندازه‌گیری آنها امری اجتناب ناپذیر است؛ ولی در عین حال، به دلیل فرآر بودن و وجود مقادیر کمی از هر یک از آنها در نمونه‌های گیاهی این اندازه‌گیری‌ها با مشکلاتی نیز همراه می‌باشد. یکی از راه‌های کسب اطلاعات کیفی و کمی راجع به این ترکیبات که اکثراً ایزومر بوده و دارای خواص فیزیکی و شیمیایی مشابهی هستند، اندازه‌گیری اندیس بازداري آنها است که با روش‌های کروماتوگرافی قابل اندازه‌گیری می‌باشد. این تکنیک‌ها به دلیل وجود تنوع زیاد ستون‌های کروماتوگرافی و پیشرفت‌های پدید آمده در این زمینه، توانایی بسیاری در حل

مسائل مربوط به جداسازی و شناسایی ترکیبات دارند. اما با وجود این مزایا، در مورد اندازه‌گیری ترکیباتی که تهیه یا آماده‌سازی آنها مشکل بوده و یا شامل مواد ناپایدار هستند دارای محدودیت می‌باشند. وجود این محدودیت‌ها در روش‌های آزمایشگاهی باعث شد که در کنار روش‌های عملی، شیوه‌های تئوری نیز برای استفاده از نتایج تجربی توسعه یابند. مطالعات کمومتریکس با استفاده از انجام محاسبات آماری و ریاضی، راهی برای استخراج حداکثر اطلاعات از نتایج تجربی می‌باشد. مطالعات کمی روابط ساختار-خاصیت (QSPR)، یکی از روش‌های کمومتریکس است که با استفاده از آن می‌توان ارتباط خطی یا غیرخطی بین ویژگی‌های ساختاری و خواص ترکیبات (مانند اندیس بازداری) برقرار نموده و خاصیت مورد نظر را برای آن ترکیبات پیش‌بینی کرد. در این تحقیق، سعی شده مدلی برای ایجاد رابطه‌ی کمی بین ساختارهای مولکولی و شاخص بازداری ترکیب روغن‌های ضروری با به‌کارگیری روش‌های QSPR ارائه شود. سپس با استفاده از این مدل، امکان پیش‌بینی شاخص بازداری برای ترکیباتی با ساختار مشابه وجود خواهد داشت.

۱-۲- پیشینه تحقیق

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

در سال ۲۰۰۷، همتی‌نژاد و همکارانش ۵۷۳ ترپنوئید^۱ را که دسته‌ای از روغن‌های ضروری می‌باشند با استفاده از روش CDFS^۲ و چند روش دیگر در QSPR مورد بررسی قرار دادند. در CDFS، آنها به کمک روش رتبه‌بندی y^3 داده‌ها را به ۵ زیرمجموعه مختلف تقسیم نمودند.

۱- Terpenoid

۲- Combined Data Splitting Feature Selection (CDFS)

۳- y-Ranking

مقادیر آماری مشابهی برای مدل‌های خطی QSRR^۱ ایجاد شده از همه زیرمجموعه داده‌ها به دست آمد. تمامی این مدل‌ها R^2 در حدود ۰/۹۰۰ داشتند. این گروه تحقیقاتی روش‌های متفاوت دیگری را نیز برای سری داده‌های مذکور مورد استفاده قرار دادند. از جمله این که، نتایج R^2 حاصل از مطالعات آنها برای پیش‌بینی فاکتور بازداري پنجاه روغن ضروري خارج از سری ترکیبات (سری تست) برای دو مدل خطی و غیر خطی به ترتیب ۰/۹۰۶۹ و ۰/۹۹۶۲ بود [۷].

ریاحی و همکارانش در سال ۲۰۰۸، روش GA^۲-MLR را برای مدل‌سازی شاخص بازداري ۴۳ روغن ضروري به کار بردند. R^2 گزارش شده برای سری آموزش متشکل از ۳۵ ترکیب، ۰/۹۷۷ و برای بقیه ترکیبات که به عنوان سری تست در نظر گرفته شدند، ۰/۹۷۸ بود [۸].

همین گروه تحقیقاتی در سال ۲۰۰۹، PLS و MLR را به عنوان روش‌های خطی و SVM^۳ و poly-PLS را به عنوان روش‌های غیر خطی QSRR، برای پیش‌بینی شاخص بازداري صد ترکیب موجود در روغن‌های ضروري به کار بردند. انتخاب توصیف‌کننده‌ها توسط روش الگوریتم ژنتیک انجام پذیرفت. R^2 به دست آمده برای ۲۰ مولکول که به عنوان سری تست انتخاب شده بودند در روش‌های MLR، PLS و poly-PLS به ترتیب ۰/۹۴۹، ۰/۹۴۴ و ۰/۹۵۳ بود. نتایج حاصله نشان از برتری SVM ($R^2=0/962$ تست) نسبت به سایر روش‌ها داشت [۹].

در سال ۲۰۱۰، نوری‌زاده و فرمانی با استفاده از روش‌های GA-MLR، GA-PLS، و شبکه عصبی مصنوعی ارتباط شاخص بازداري و ۴۰ ترکیب حلقوی روغن ضروري را در رزماری و گیاه ساج بررسی نمودند. در تکنیک GA-MLR، R^2 برای سری تست شامل ۸ ترکیب برابر با ۰/۹۰۲ گزارش شد. توصیف‌کننده‌های حاصل از این روش به عنوان ورودی به ANN داده شدند و R^2 روش غیرخطی برای سری ارزیابی ۰/۹۸۰ بود. همچنین در GA-PLS، R^2 برای تست ۰/۹۷۵ گزارش شد [۱۰].

۱- Quantitative Structure-Retention Relationships

۲- Genetic Algorithm

۳- Supported Vector Machine (SVM)

همین گروه در سال ۲۰۱۰، از چهار روش GA-MLR، GA-PLS، GA-KPLS^۱ و ANN برای یافتن همبستگی میان اندیس بازداری ۱۱۳ ترکیب مختلف روغن ضروری با توصیف‌کننده‌ها بهره گرفتند. توصیف‌کننده‌های به دست آمده از GA-KPLS به عنوان ورودی به شبکه عصبی داده شدند. ضرایب همبستگی در این روش‌ها برای سری تست شامل ۲۵ مولکول، به ترتیب ۰/۸۸، ۰/۹۱، ۰/۹۴ و ۰/۹۷ بود [۱۱].

همچنین نوری‌زاده و همکارانش در همین سال، چهار تکنیک گفته شده را برای انجام مطالعات روی سری دیگری از روغن‌های ضروری شامل ۱۱۶ ترکیب به کار گرفتند. ضرایب همبستگی به دست آمده برای سری تست با ۲۴ ترکیب، به ترتیب ۰/۹۱۷، ۰/۸۹۰، ۰/۹۱۵ و ۰/۹۵۴ بود [۱۲].

نوری‌زاده و همکارانش ایضاً در سال ۲۰۱۱، دو روش GA-PLS و GA-KPLS را در QSRR برای آنالیز اندیس بازداری ۱۱۷ ترکیب مختلف از روغن‌های ضروری به کار بردند. R^2 برای ۲۵ ترکیب سری تست به ترتیب ۰/۹۱۴ و ۰/۹۳۸ بود [۱۳].

۱- Kernel Partial Least Square (KPLS)

فصل دوم

کمو متریکس

۲-۱- کمومتریکس

اصطلاح کمومتریکس برای اولین بار توسط اسوانت ولد^۱ دانشمند جوان سوئدی مطرح شد و همکاری او با بروس آر. کووالسکی^۲ در مطالعه بر روی الگوشناسی در شیمی تجزیه در دانشگاه واشنگتن، منجر به تأسیس انجمن بین المللی کمومتریکس (ICS)^۳ در سال ۱۹۷۴ گردید [۱۴]. بنا به تعریف ICS، کمومتریکس کاربرد روش‌های ریاضی و آماری برای برقراری ارتباط بین سنجش‌های انجام شده روی یک سیستم یا فرایند شیمیایی به منظور درک بهتر اطلاعات شیمیایی می‌باشد [۱۵]. تعریفی که اکنون از کمومتریکس ارائه می‌شود عبارت است از: شاخه‌ای از شیمی که روش‌های ریاضی و آماری را برای انتخاب و طراحی مطلوب‌ترین روش‌ها و آزمایش‌ها؛ تجزیه و تحلیل داده‌های شیمیایی به منظور فراهم آوردن بیشترین اطلاعات شیمیایی و شناخت سیستم‌های شیمیایی به کار می‌برد [۱]. از جمله قابلیت‌های کمومتریکس می‌توان به کنترل فرآیندها با نگرش چند متغیره‌ای به مسائل، ساخت مدل‌ها با قابلیت پیش‌بینی و ارزیابی آنها، مقایسه نتایج به دست آمده از روش‌های مختلف و همچنین تعریف شاخص‌هایی که کیفیت اطلاعات استخراج شده و مدل‌های به دست آمده را مورد سنجش قرار دهند، اشاره نمود [۱۶، ۱۷]. کاربردهای عمومی روش‌های کمومتریکس توسعه ارتباط کمی ساختار- فعالیت^۴ و یا ارتباط کمی ساختار- خاصیت یا ارزیابی داده‌های شیمیایی هستند. مهمترین دلیل برای افزایش کاربرد روش‌های کمومتریکس نزد شیمیدانان، داده‌های فراوانی بودند که به وسیله‌ی ظهور تجهیزات مدرن تجزیه‌ای (کامپیوترها و منابع نرم افزاری) تولید شدند. از اوایل دهه‌ی ۸۰ میلادی با در دسترس قرار گرفتن رایانه‌های شخصی، دوره‌ی نوینی برای کشف، پردازش و تفسیر داده‌های شیمیایی آغاز گردید.

۱- Svante wold

۲- Bruce R. Kowaski

۳- International Chemometrics Society

۴- Quantitative Structure - Activity Relationship (QSAR)

هاوری^۱ و هیرش^۲ در اوایل سال ۱۹۸۰ پیشرفت‌های کمومتریکس را در مراحل مختلف طبقه‌بندی کردند؛ بدین ترتیب که اولین مرحله قبل از ۱۹۷۰ رخ داد که در آن تعدادی از روش‌های ریاضی توسعه یافتند. در این زمان، عملکرد شیمیدانان محدود به روش‌های آنالیز داده‌ها، شامل محاسبه پارامترهای آماری مانند میانگین، انحراف استاندارد و حدود اطمینان می‌شد. دومین مرحله تکامل کمومتریکس در سال ۱۹۷۰ بود که کمومتریکس توجه شیمیدانان، به خصوص متخصصین شیمی تجزیه را که با این روش، راه‌های بیشتری برای آنالیز داده‌های در دسترس و همچنین روش‌های نوینی برای به دست آوردن اطلاعات پیش روی خود می‌دیدند، به خود جلب کرد. مرحله بعدی تکامل کمومتریکس که ابتدا توسط هاوری و هیرش و سپس توسط براون^۳ پیش‌بینی شده بود از اوایل ۱۹۸۰ آغاز گردید. در ابتدا تمرکز تحقیقات هاوری و هیرش به‌طور ویژه روی همبستگی تعداد زیادی از داده‌های شیمیایی و ویژگی‌های مولکولی مربوطه، قرار داشت که این اقدامات اولیه اساس یکی از شاخه‌های مهم کمومتریکس یعنی ارتباط کمی ساختار- فعالیت (QSPR) را فراهم نمود. هم‌اکنون کمومتریکس ابزاری قوی و کارآمد برای شیمیدانان به شمار رفته و به عنوان یک زیرشاخه رشته شیمی در بسیاری از دانشگاه‌های آمریکا و اروپا و بعضی دانشگاه‌های چین و سایر کشورها تدریس می‌شود. پیشرفت‌ها در علوم کامپیوتر، فناوری اطلاعات، آمار و ریاضی کاربردی عناصر جدیدی را در کمومتریکس معرفی کرده است [۱۴، ۱۸، ۱۹].

هم‌زمان با رشد سریع کمومتریکس در جهان، در ایران نیز افزایش استفاده از این روش در تحقیقات متخصصین شیمی تجزیه مشاهده گردید. انجمن کمومتریکس ایران به عنوان یکی از شاخه‌های فعال انجمن شیمی ایران در سال ۲۰۰۰ تأسیس شد. زمینه‌هایی که تاکنون بیشتر در ایران

۱- Howery

۲- Hirsch

۳- Brown

مورد توجه قرار گرفته‌اند، مطالعات (QSPR/QSAR)، کالیراسیون چندمتغیره^۱ و هوش مصنوعی^۲ هستند [۲۰]. مقاله‌ای که در ارتباط با کار جلالی هروی و گرکانی نژاد از دانشگاه شهید باهنر کرمان می‌باشد و با عنوان پیش‌بینی اندیس بازداري کروماتوگرافي گازی تعدادی از مشتقات بنزن در سال ۱۹۹۳ به چاپ رسیده است، به عنوان اولین مقاله چاپ شده در موضوعیت کمومتریکس در ایران مطرح است [۲۱].

۲-۲- ارتباط کمی ساختار- ویژگی

با توسعه‌ی نظریه‌ی ساختاری^۳ در نیمه‌ی دوم قرن نوزدهم، این نتیجه به دست آمد که خواص شیمیایی، فیزیکی و زیستی تابعی از تغییرات ساختاری هستند. یکی از مهم‌ترین روش‌هایی که به صورت کمی از این قاعده برای مطالعات ارتباط بین ساختار مولکولی و خواص مشاهده شده‌ی مولکول بهره می‌برد، ارتباط کمی ساختار- خاصیت (QSPR)، نام دارد. علاوه بر شیمی تجزیه، رشته‌های گوناگونی مانند شیمی فیزیک آلی، شیمی معدنی، اسپکتروسکوپی و شیمی محیط‌زیست نیز از این روش جهت پیش‌بینی خواص مختلف استفاده می‌کنند [۲۲]. با وجود کاربرد وسیع، قواعد اصلی و طرزکار این روش در همه‌ی آنها یکسان یا مشابه است. در این روش خاصیت مدّ نظر با پارامترهایی که اساس آنها ویژگی‌های مولکولی است و توصیف‌کننده‌های مولکولی^۴ نام دارند، ارتباط داده می‌شوند [۷،۲۳].

از جمله این ویژگی‌ها خصوصیات الکترونیکی، هندسی، فضایی، آبگریزی و غیره بوده و برخی از این خواص شامل ضریب شکست پلیمر، نقطه‌ی ذوب یا جوش ترکیبات، کشش سطحی مایعات، زمان بازداري و اندیس بازداري ترکیبات آلی می‌باشند. توصیف‌کننده‌های مولکولی پارامترهای کمی هستند

۱- Multivariate Calibration (MVC)

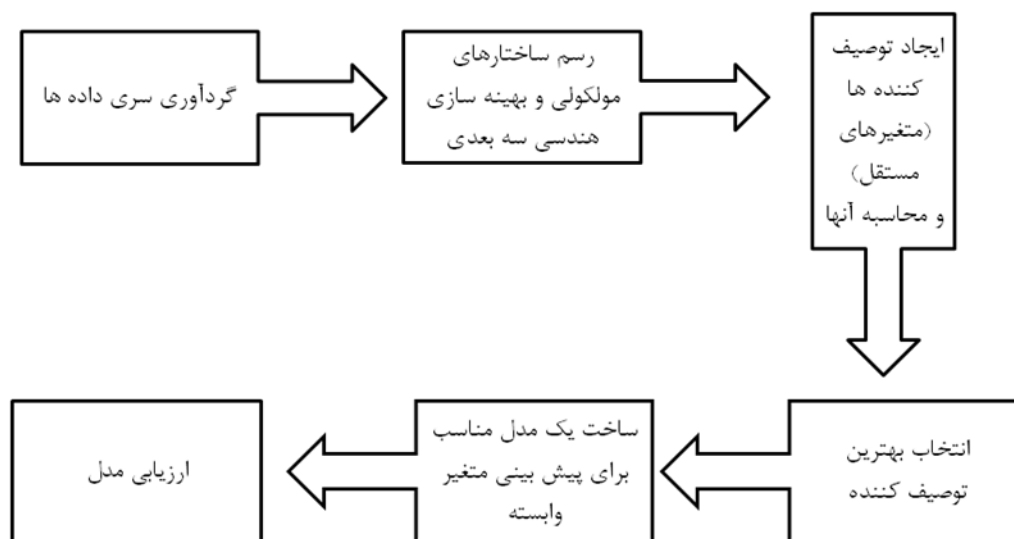
۲- Artificial Intelligence (AI)

۳- Structural Theory

۴- Molecular Descriptors

که با استفاده از الگوریتم‌های شناخته شده ریاضی یا توسط فرمول‌های به دست آمده از روش‌هایی مثل مکانیک کوانتومی از ساختارهای مولکولی محاسبه می‌شوند. به کمک یک سری اصول خاص، تعداد معین و محدودی از توصیف‌کننده‌های مولکولی به عنوان ورودی QSPR انتخاب شده که از لحاظ نظری نیز مفهوم فیزیکی مشخصی را برای ترکیبات یک سری بیان می‌کنند. به‌طور 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 شفاف‌سازی نحوه‌ی ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختمانی آنها، به پژوهشگران این امکان را می‌دهد که بر اساس رفتار مولکول‌های مشابه همان خواص را برای ساختارهایی که تاکنون اندازه‌گیری نشده یا حتی هنوز تهیه نشده‌اند پیش‌بینی کنند [۲۴، ۲۵].

مراحل اساسی QSPR در شکل ۱-۲ نشان داده شده است:



شکل ۱-۲- نمایش خلاصه مراحل به کار رفته در یک فرایند QSPR

۱-۲-۲- جمع‌آوری سری داده‌ها

در ابتدا، گردآوری گروهی از مولکول‌ها که از نظر ساختاری به یکدیگر شباهت دارند و دارای مقادیر اندازه‌گیری شده (تجربی) از یک ویژگی خاص مولکولی نظیر اندیس بازداري در شرایط یکسان

باشند، ضرورت دارد. به این دسته مولکولی سری داده^۱ها می‌گویند که هرچه تعداد گونه‌ها در این سری بیشتر و همچنین مولکول‌های مورد استفاده تقریباً مشابه بوده یا گروهی از مشتقات یک ترکیب باشند، مدل‌های ارائه شده بهتر خواهند بود.

۲-۲-۲- بهینه‌سازی ساختار هندسی

برای ایجاد توصیف‌کننده‌های مولکولی به خصوص توصیف‌کننده‌های هندسی به شکل صحیح، باید ساختار هندسی مولکول‌ها بهینه^۲ شود. بدین منظور از نرم‌افزار HyperChem استفاده شده است که در آن ساختمان مولکول‌ها رسم شده و سپس برای دستیابی به ساختار بهینه، صورت‌بندی مولکول تا رسیدن به میزان حداقل انرژی تغییر می‌کند. این اعمال، در زیر مجموعه‌ی روش‌هایی به نام شیمی محاسباتی^۳ قرار می‌گیرند که هدف آنها استخراج اطلاعات از داده‌های شیمیایی است. به طور کلی شیمی محاسباتی ساختارهای مولکولی را به صورت پارامترهای عددی معرفی و رفتار آنها را با معادلات کوانتومی یا فیزیک کلاسیک شبیه‌سازی می‌نماید که از این طریق، اطلاعاتی راجع به مولکول‌ها نظیر ساختار هندسی، خواص الکترونیکی و انرژی آنها به دست می‌آید. روش‌های محاسباتی در شیمی به چندین دسته تقسیم می‌شوند. در این تحقیق، از روش AM1^۴ برای رسم و محاسبه‌ی انرژی مولکول‌ها استفاده شده که از جمله روش‌های محاسباتی نیمه‌تجربی^۵ می‌باشد [۲۶، ۲۷].

۱- Data Set

۲- Optimize

۳- Computational Chemistry

۴- Austin Methods

۵- Semi-empirical Methods

توصیف‌کننده‌ها مقادیر عددی هستند و به طور کمی ویژگی‌های ساختاری و الکترونی مولکول‌ها را بیان می‌کنند؛ یا به عبارت دیگر توصیف‌کننده‌های مولکولی نتیجه نهایی یک فرآیند ریاضی و منطقی می‌باشند که اطلاعات شیمیایی مربوط به ساختار یک مولکول را به اعداد تبدیل می‌کنند. این اعداد می‌توانند خواص مولکولی را تفسیر نموده یا برای پیش‌بینی تعدادی از ویژگی‌های مولکولی در یک مدل شرکت کنند. هر توصیف‌کننده نشان‌دهنده خصوصیت ویژه‌ای از مولکول است که می‌تواند بر خاصیت تجربی مورد نظر تأثیر داشته باشد [۲۸].

یک توصیف‌کننده مولکولی مناسب باید شامل ویژگی‌های زیر باشد [۲۹]:

- عدم پیچیدگی و ساده بودن
 - عدم وابستگی به دیگر توصیف‌کننده‌ها و استقلال از یکدیگر
 - دارا بودن بیشترین وابستگی با خاصیت تجربی مدنظر
 - توانایی تفسیر ساختار ترکیبات
 - توانایی تمیز دادن ایزومرهای مختلف مولکول از یکدیگر
- توصیف‌کننده‌های مولکولی از دیدگاه‌های متفاوت به دسته‌های مختلفی تقسیم می‌شوند. ساده‌ترین طبقه‌بندی بر اساس توصیف‌کننده‌های تجربی و تئوری است. توصیف‌کننده‌ی تجربی بر اساس اندازه‌گیری‌های تجربی به دست آمده و برای تعداد محدودی از ترکیبات شیمیایی قابل استفاده است؛ در حالی که توصیف‌کننده‌ی تئوری به کمک ساختار مولکول و بدون نیاز به داده‌های تجربی محاسبه می‌گردد و برای هر مولکول واقعی و یا فرضی که هنوز سنتز نشده، در دسترس است بنابراین در وقت، هزینه، مواد و تجهیزات صرفه‌جویی به عمل می‌آید [۳۰، ۳۱]؛ به همین دلیل در این تحقیق نیز از توصیف‌کننده‌های تئوری بهره برده‌ایم. توصیف‌کننده‌های تئوری خود بر اساس ابعاد توصیف مولکول به انواع صفر بعدی، یک بعدی، دوبعدی و سه بعدی تقسیم می‌شوند.

۲-۲-۴- انتخاب بهترین توصیف‌کننده‌ها

با وجود تعداد زیاد توصیف‌کننده‌ها و همچنین همبستگی بالای برخی از آنها به یکدیگر، فرایند مدل‌سازی طولانی و دشوار خواهد شد؛ بنابراین با انتخاب زیر مجموعه‌ای از متغیرهای مستقل (توصیف‌کننده‌ها) که اطلاعات اساسی مربوط به کل سری داده را دربردارند، متغیرهای با همبستگی بالا حذف می‌شوند. این روش، فرایند کاهش متغیر^۱ نام دارد که در حین انجام آن باید به نکات زیر توجه شود:

- توصیف‌کننده‌های با بیش از ۹۰٪ مقادیر یکسان حذف می‌شوند.
 - از بین توصیف‌کننده‌هایی که با هم همبستگی بالای ۰/۹ دارند توصیف‌کننده‌ای که با متغیر وابسته همبستگی کمتری دارد، حذف می‌گردد.
 - توصیف‌کننده‌هایی که محاسبه و تفسیر آنها مشکل باشد، حذف می‌شوند.
- سپس از فرایند انتخاب متغیر^۲ استفاده می‌شود که طی آن، از بین توصیف‌کننده‌ها، آنهایی که با متغیر وابسته همبستگی بیشتری دارند انتخاب می‌گردند. یکی از روش‌های انتخاب متغیر، روش برازش مرحله‌ای^۳ است که از طریق محاسبه ضریب همبستگی^۴ بین توصیف‌کننده و خاصیت تجربی انجام می‌شود (ضریب همبستگی، بیان‌کننده میزان وابسته بودن پارامتر تجربی به متغیر مستقل موردنظر به صورت کمی است). اگر توصیف‌کننده‌ای دارای ضریب همبستگی برابر یک باشد، قادر است خواص مورد نظر را به درستی توصیف کند و ضریب همبستگی صفر بدین معنی است که هیچ ارتباطی بین توصیف‌کننده و پارامتر تجربی وجود ندارد. پس توصیف‌کننده‌هایی که دارای ضرایب همبستگی بزرگ هستند، در برازش منحنی برای ساخت مدل انتخاب می‌شوند.

۱- Variable Reduction

۲- Variable Selection

۳- Stepwise

۴- Correlation Coefficient

۲-۲-۵- انتخاب مدل مناسب

یک مدل بر اساس روش‌های آماری گوناگون که از قواعد ریاضی بهره می‌برد، 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 موردنظر بوده و معمولاً از اندازه‌گیری‌های 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 به کمک آن 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. مدل‌های رگرسیون که شامل تعداد کمی از متغیرهای مستقل هستند بر سایر مدل‌ها برتری دارند، زیرا تفسیر این مدل‌ها به سهولت انجام می‌گیرد و قدرت پیش‌بینی بالایی دارند[۳۲]. رگرسیون به معنای رابطه ایجاد شده میان یک متغیر از طرفی و یک یا چند متغیر از طرف دیگر می‌باشد. در صورتی که تنها یک متغیر مستقل داشته باشیم، رگرسیون را ساده و در غیر این صورت، رگرسیون را چندگانه می‌گویند. در بسیاری موارد برای پیش‌بینی مقادیر یک متغیر، دانستن مقادیر دو یا چند متغیر دیگر لازم بوده و در این حالت از روش‌های رگرسیون چندگانه استفاده می‌شود[۳۳]. در رگرسیون خطی چندگانه چندین روش رگرسیون مختلف برای ایجاد مدل وجود دارد که به برخی از آنها اشاره می‌گردد[۳۴]:

۱- Dependent Variable

۲- Independent Variable(s)

۲-۲-۵-۱- ورود اجباری^۱

روشی است که طی آن، تمامی متغیرهای مستقل در یک مرحله وارد مدل می‌شوند. این روش چندان مطلوب نیست زیرا در آن از اثرات مثبت و منفی توصیف‌کننده‌ها و میزان همبستگی آنها با متغیر وابسته برای ورود به مدل صرف‌نظر می‌شود.

۲-۲-۵-۲- روش برازش مرحله‌ای

در این روش متغیرهای مستقل به صورت مرحله‌ای انتخاب شده که در ابتدا، متغیری که بیشترین همبستگی با متغیر وابسته داشته باشد وارد مدل می‌گردد؛ سپس در قدم بعد، متغیر دوم با بیشترین همبستگی از بین متغیرهای باقی‌مانده انتخاب و به همین ترتیب ادامه می‌یابد. در هر مرحله برای سنجش معنادار بودن ضریب همبستگی متغیرهای وارد شده در مدل، آزمون F انجام می‌گیرد (توضیح آماره F در ادامه خواهد آمد) و در صورتی که F از نظر آماری برای متغیری معنادار باشد، آن متغیر در مدل باقی‌مانده در غیر این صورت حذف خواهد شد. در نهایت به کمک متغیرهایی که وارد معادله‌ی رگرسیون شده‌اند مدل مناسب ساخته می‌شود. لازم به ذکر است در این پایان‌نامه برای انتخاب متغیر از این روش استفاده شده است.

۲-۲-۵-۳- حذف پس‌رو^۲

در این تکنیک تمام توصیف‌کننده‌ها وارد معادله شده و سپس به طور متوالی حذف می‌گردند. در مرحله‌ی اول، متغیری که کمترین همبستگی را با متغیر وابسته دارد شناسایی و در صورتی که بر اساس آزمون F ضریب همبستگی آن بی‌معنا باشد، حذف خواهد شد. سپس از بین متغیرهای باقی‌مانده در معادله، متغیری که کوچکترین همبستگی را با متغیر وابسته دارد برای حذف در نظر

۱- Enter

۲- Backward Elimination

گرفته می‌شود و این فرآیند ادامه می‌یابد تا زمانی که هیچ متغیری در معادله از ضوابط موجود برای حذف پیروی نکند.

۲-۲-۴- انتخاب پیش‌رو^۱

یک رویه انتخاب متغیر است که در آن متغیرها به طور متوالی وارد مدل شده و اولین متغیری که برای ورود به مدل شناسایی می‌شود دارای بالاترین هم بستگی مثبت یا منفی با متغیر وابسته است. متغیر مربوطه تنها زمانی می‌تواند وارد مدل گردد که ضریب همبستگی آن از لحاظ آزمون F معنادار باشد. هنگامی که دیگر متغیر دارای شرط ورود وجود نداشته باشد، این روند متوقف خواهد شد.

۲-۳- مدل سازی با استفاده از روش MLR

MLR روشی است که برای مدل سازی ارتباط خطی بین متغیر وابسته و چند متغیر مستقل به کار می‌رود. معادله‌ی این نوع رگرسیون به این صورت است:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + e \quad (\text{معادله ۲-۱})$$

که (X_i) ها متغیرهای مستقل (توصیف کننده‌های عددی) و مقادیر b_0 تا b_n ضرایب متغیرهای مستقل اند و e نیز مقدار خطای رگرسیون است که باقی مانده نامیده می‌شود. Y نیز همان متغیر وابسته یا به عبارت دیگر خاصیت یا فعالیت مورد اندازه گیری است. مقدار ضریب متغیر مستقل در معادله، بیانی از میزان اهمیت آن متغیر در پیش بینی مقادیر متغیر وابسته است. این روش تنها زمانی می‌تواند به کار رود که تعداد مولکول‌های مورد بررسی بیش از تعداد توصیف کننده‌ها باشد و متغیرها با یکدیگر همبستگی زیادی نداشته باشند.

۱- Forward Selection

۲-۴- شبکه عصبی مصنوعی

شبکه‌های عصبی زیر مجموعه‌ای از هوش مصنوعی هستند. این شبکه‌ها در واقع الگویی از سیستم‌های عصبی زیستی هستند و مانند آنها از عناصر عملیاتی ساده‌ای به نام نرون^۱ که به صورت موازی عمل میکنند، ساخته شده‌اند. نرون‌ها با مکانیسم فراگیری و آموزش و با تأثیرپذیری و تأثیرگذاری متقابل بر دستگاه‌های دیگر بدن، ارتباط هدفمند جان‌دار را با محیط اطرافش ممکن می‌سازند. دریافت پیام‌های محیطی و تحلیل و پاسخ‌دهی به آنها منوط به چینش صحیح و عملکرد سریع نرون‌ها می‌باشد که این شرایط از طریق تنظیم چگونگی اتصالات میان آنها حاصل می‌شود. مدل‌های مصنوعی شبکه‌های عصبی نیز بر اساس همین الگو بنا شده‌اند و هر کدام برای کاربردی ویژه مانند تشخیص الگو یا طبقه‌بندی اطلاعات از طریق فرآیند آموزش با مثال، صورت‌بندی گردیده‌اند. تحقیقات نشان می‌دهد که مغز اطلاعات را همانند الگوها ذخیره می‌کند که فرآیند ذخیره‌سازی اطلاعات به صورت الگو و تجزیه و تحلیل آن الگو اساس روش نوین محاسباتی را تشکیل می‌دهند. این حوزه از دانش محاسباتی از شبکه‌های بزرگی که به صورت موازی آرایش یافته و آموزش می‌بینند بهره می‌برد. پردازش موازی به معنای اجرای برنامه‌هایی است که در هر بار بیش از یک عمل را انجام می‌دهند. کامپیوترهای معمولی اعمال محاسباتی را یکی پس از دیگری (به صورت ترتیبی) به انجام می‌رسانند. در شبکه‌های عصبی اطلاعات همانند مغز انسان‌ها پردازش می‌شوند؛ به همین دلیل روش برخورد چنین مدل‌هایی با مسائل، مغایر با روش‌های محاسباتی معمول در سیستم‌های ریاضی متداول است که دانسته‌های مربوط به آنها به صورت قواعد روشن و از پیش مشخص شده در سیستم وجود دارند. در یک شبکه‌ی عصبی مصنوعی نیز با تنظیم مقادیر هر اتصال تحت عنوان وزن اتصال^۲، می‌توان نحوه‌ی ارتباط میان اجزای آن را تعیین نمود. پس از تنظیم یا آموزش شبکه بر مبنای تطابق^۳ و هم‌سنجی میان ورودی و هدف، خروجی شبکه بر هدف منطبق می‌گردد. استفاده از زوج داده‌های

۱- Neuron

۲- Connection Weight

۳- Adaptive

ورودی و هدف بیان گر یادگیری نظارت شده^۱ می باشد [۳۵]. برای آشنایی با شبکه های عصبی مصنوعی، ابتدا سیستم عصبی زیستی را به اختصار بررسی می کنیم.

۲-۴-۱- سیستم عصبی زیستی

مغز در بدن، به عنوان یک سیستم پردازش اطلاعات با ساختار موازی و کاملاً پیچیده است که دو درصد وزن بدن را تشکیل می دهد. مغز متشکل از ۱۰۰ تریلیون نرون می باشد که با بیش از ۱۰^{۱۶} ارتباط به هم متصل شده اند. نرون ها ساده ترین واحد ساختاری سیستم های عصبی هستند و اصلی ترین عنصر پردازش^۲ اطلاعات بشمار می آیند. غالب نرون ها از سه قسمت اصلی دندریت^۳، جسم سلولی^۴ (شامل هسته و بخش های دیگر) و آکسون^۵ تشکیل شده اند. آکسون هر سلول عصبی در مکانی به نام سیناپس^۶ با دندریت های سلول های دیگر تلاقی می کند. در واقع سیناپس ها محل ارتباط مابین نرون ها می باشند. چگونگی برقراری این ارتباط به میزان مواد انتقال دهنده نرونی که در انتهای آکسون ها ذخیره شده اند، بستگی دارد. دندریت ها به عنوان مناطق دریافت سیگنال ها، شبکه هایی متشکل از فیبرهای سلولی هستند که دارای سطحی نامنظم و شاخه های انشعابی بی شمار می باشند. دندریت ها پس از دریافت اطلاعات به شکل سیگنال آن را به هسته ی سلول که در جسم سلولی قرار دارد، هدایت می کنند. جسم سلولی ضمن فراهم کردن انرژی لازم برای فعالیت نرون، عملیات لازم را بر روی سیگنال های دریافتی انجام می دهد که این عملیات شبیه یک عمل جمع ساده بر روی کل سیگنال ها و مقایسه با یک سطح آستانه می باشد. آکسون برخلاف دندریت ها از سطحی هموارتر و تعداد شاخه های کمتر و طول بیشتری برخوردار است. نقش این قسمت، انتقال سیگنال های

۱- Supervised Learning or Supervised Training

۲- Processing Element

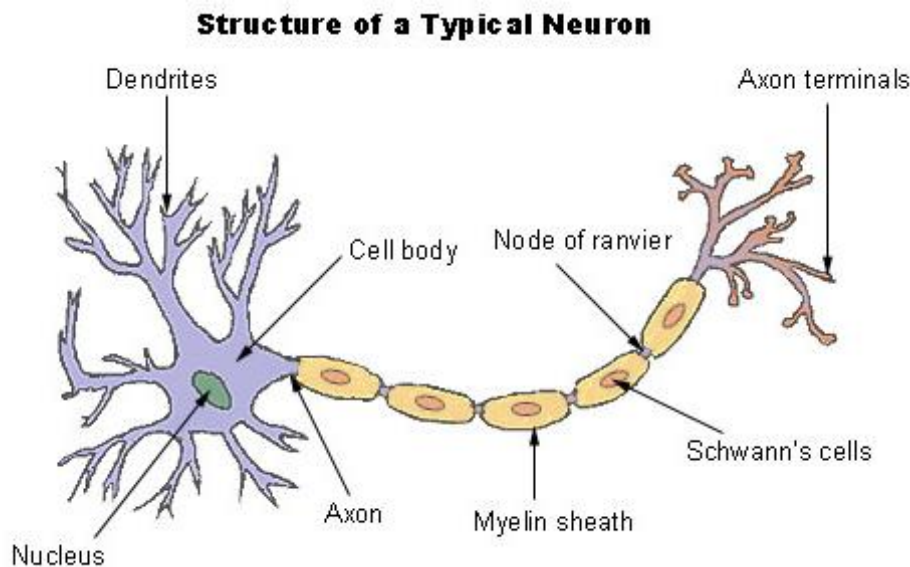
۳- Dendrite

۴- Cell Body (Soma)

۵- Axon

۶- Synaps

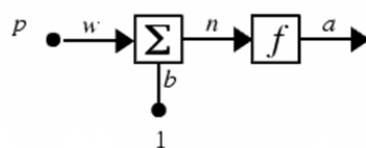
الکتروشیمیایی دریافتی از جسم سلولی از طریق سیناپس‌ها به نرون‌های دیگر است و بدین ترتیب فعالیت‌های مغزی انجام می‌گیرد [۳۶]. شکل ۲-۲ ساختار یک نرون زیستی را نشان می‌دهد.



شکل ۲-۲ سلول عصبی زیستی [۳۷]

۲-۴-۲- سیستم عصبی مصنوعی

می‌توان به تبعیت از شبکه‌های طبیعی یک ساختار مصنوعی ساخت و همان‌طور که گفته شد با تنظیم مقادیر وزن اتصال نحوه ارتباط میان اجزای شبکه را تعیین نمود. شکل ۲-۳، الگویی از یک نرون محاسباتی با توجه به نحوه عملکرد نرون طبیعی ارائه 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 ۲-۳- مدل ریاضی نرون [۳۸]

در این شکل از علامت p برای نشان دادن یک سیگنال ورودی استفاده شده و a نیز نشان‌دهنده‌ی خروجی نرون است. در واقع در این مدل یک سیگنال ورودی پس از تقویت (یا تضعیف) شدن با وزن w ، pw را به وجود می‌آورد. میزان تأثیر ورودی p روی خروجی a ، به وسیله‌ی w مشخص می‌شود. ورودی دیگر که یک مقدار ثابت^۱ است در بایاس^۱ ضرب شده و حاصل آن با pw جمع می‌گردد. سپس تابع انتقال^۲ نرون (f) روی این مقدار یعنی n اعمال می‌شود.

$$n = pw + b \quad \text{معادله (۲-۲)}$$

پس به این ترتیب خروجی نرون این‌گونه محاسبه می‌شود:

$$a = f(pw + b) \quad \text{معادله (۳-۲)}$$

w و b 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 [۳۹].

برخی از اجزای یک شبکه‌ی عصبی مصنوعی عبارتند از:

ورودی‌ها: به حالت خام در اولین لایه به صورت داده‌های عددی و رقمی، متون ادبی، فنی، تصویر و یا شکل بوده و یا می‌توانند خروجی سایر لایه‌ها باشند.

وزن‌ها: همان‌طور که گفته شد، میزان تأثیر ورودی p بر خروجی a ، توسط وزن اندازه‌گیری می‌شود که با وارد شدن بایاس b مقدار وزن‌دار شده برای تطبیق بیشتر ورودی-خروجی نرون افزایش یا کاهش می‌یابد.

خروجی: منظور همان پاسخ مسأله است.

۱- Bias

۲-Transfer Function

تابع جمع‌کننده^۱: در شبکه‌های تک نرونی، خروجی مسأله و در شبکه‌های چند نرونی میزان سطح فعالیت نرون در لایه‌های درونی را مشخص می‌کند.

تابع تبدیل یا تابع انتقال^۲: وظیفه محدودسازی خروجی گره^۳ 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 [۴۰]. انواع متفاوتی از توابع انتقال از قبیل پله‌ای، تکه‌ای - خطی^۴ و سیگموئیدی^۵ 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 (وزن‌ها) تنظیم می‌گردند. از توابع انتقال می‌توان به تابع انتقال خطی^۶، لگاریتم سیگموئیدی^۷ و تانژانت سیگموئیدی^۸ اشاره نمود. تابع انتقال خطی همان مقدار ورودی را به عنوان خروجی برمی‌گرداند ولی تابع لگاریتم سیگموئید مقادیر ورودی را در محدوده‌ی $-\infty$ تا $+\infty$ دریافت 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 [۴۱]. در شکل ۲-۴، سه نوع تابع انتقال معرفی شده، به صورت شماتیک نمایش داده شده‌اند.

۱- Additional Function

۲- Transfer Function

۳- Node

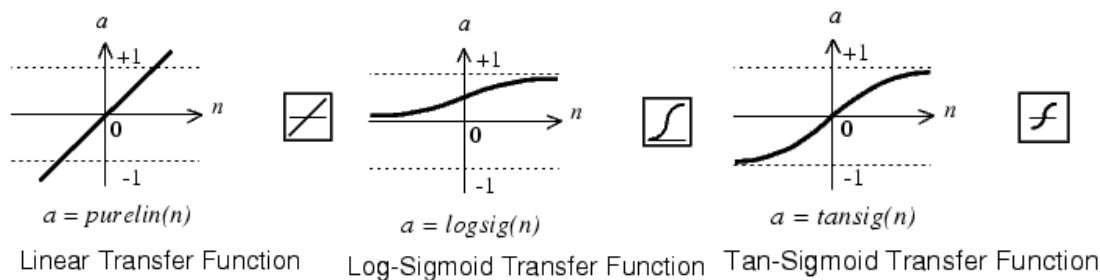
۴- Linear-Segment

۵- Sigmoid

۶- Linear

۷- Log-Sigmoid (logsig)

۸- Tan-Sigmoid (tansig)

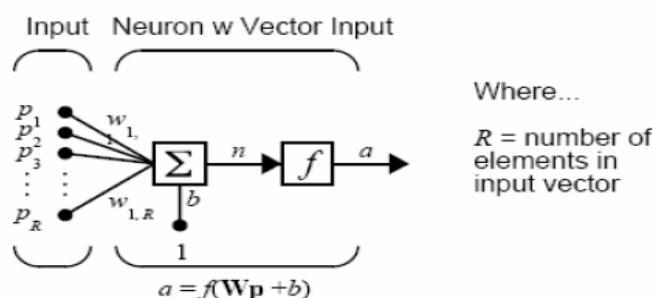


شکل ۴-۲- برخی از انواع توابع انتقال در شبکه‌های عصبی [۳۸]

۵-۲- معماری شبکه‌های عصبی^۱

در صورتی که ورودی نرون به شکل یک بردار باشد، نرون تک لایه را به صورت شکل ۵-۲

خواهیم داشت:



شکل ۵-۲- نمایش نرون تک لایه با یک بردار ورودی [۳۸]

در اینجا به تعداد R ، ورودی‌های تک عنصری p وجود دارد یا به عبارت دیگر $p_1, p_2, p_3, \dots, p_R$

که این‌ها بردار p را تشکیل می‌دهند. این مقادیر در وزن‌های مربوطه $(w_{1,1}, w_{1,2}, w_{1,3}, \dots, w_{1,R})$ ضرب

شده و حاصل آن یعنی n به محل تابع جمع‌کننده خورانده می‌شود. n در واقع حاصل ضرب ماتریسی

بردار p و w است. همان‌طور که در شکل ۵-۲ توضیح داده شده، R برابر با تعداد المان‌ها در بردار

ورودی می‌باشد. ورودی تابع انتقال f (یعنی n) و مقدار خروجی را به ترتیب زیر خواهیم داشت:

$$n = w_{1,1}p_1 + w_{1,2}p_2 + \dots + w_{1,R}p_R + b \quad \text{معادله (۴-۲)}$$

$$a = f(w_{1,1}p_1 + w_{1,2}p_2 + \dots + w_{1,R}p_R + b) \quad \text{معادله (۵-۲)}$$

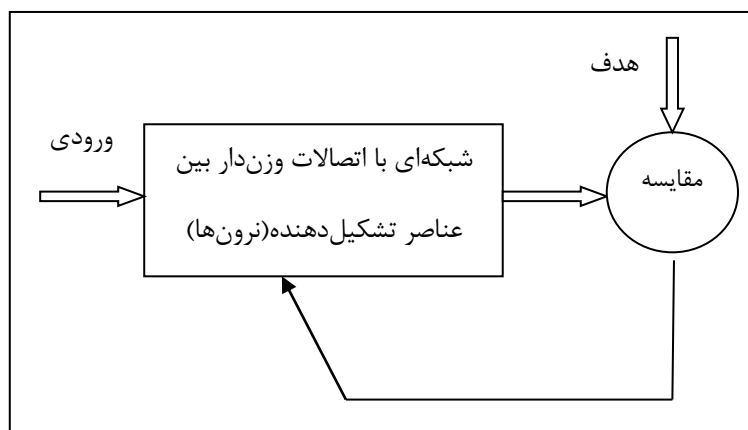
b همان مقدار بایاس است که به حاصلضرب ماتریسی w در p اضافه شده است.

به ترکیب وزن‌ها، بایاس، تابع انتقال و عملیات ضرب و جمع انجام شده یک لایه از شبکه می‌گویند؛ باید توجه داشت که در بیشتر نوشتارها، بردار ورودی یک لایه محسوب نمی‌گردد. دو یا چند نرون می‌توانند در قالب یک لایه با هم ترکیب شده و یک شبکه نیز می‌تواند به نوبه خود از یک یا چند لایه تشکیل شود. هر لایه در شبکه دارای ماتریس وزن‌ها، بردار بایاس و خروجی مختص به خود است. در شبکه‌های چندلایه، خروجی هر نرون وارد تک‌تک نرون‌های لایه‌ی بعدی خواهد شد [۴۲].

۶-۲- آموزش شبکه عصبی

معمولاً شبکه‌های عصبی تنظیم شده یا آموزش می‌یابند تا اعمال یک ورودی خاص منجر به دریافت خروجی تا حد ممکن نزدیک به مقدار هدف شود؛ به عبارت دیگر، مقدار پارامترها برای رسیدن به خروجی مطلوب‌تر تنظیم می‌شود. در نخستین مرحله‌ی آموزش، به w و b مقادیر تصادفی داده می‌شود زیرا تا این پارامترها مقدار نداشته باشند شبکه عصبی قابل استفاده نخواهد بود. همانطور که در شکل ۶-۲ دیده می‌شود شبکه بر مبنای تطابق و هم‌سنجی بین ورودی و هدف سازگار شده تا اینکه خروجی شبکه و هدف بر هم منطبق گردند. عموماً تعداد زیادی از این زوج‌های ورودی و خروجی به کار گرفته می‌شوند تا در این روند که یادگیری تحت نظارت نامیده می‌شود، شبکه آموزش داده شود. در یادگیری تحت نظارت قاعده یادگیری با استفاده از مجموعه‌ای از مثال‌ها (مجموعه آموزشی) به دست می‌آید و شبکه را آموزش می‌دهد. اگر زوج داده‌های زیر در نظر گرفته شوند:

$$\{p_1, t_1\}, \{p_2, t_2\}, \dots, \{p_q, t_q\}$$



شکل ۲-۶- شیوه‌ی یادگیری با ناظر [۴۳]

که در آن، p_Q یک ورودی شبکه و t_Q هدف مورد نظر متناظر با هر ورودی می‌باشد. زمانی که ورودی به شبکه اعمال شد، خروجی آن با هدف مقایسه گردیده و سپس قواعد یادگیری برای تنظیم وزن‌ها و بایاس‌ها به کار برده می‌شود تا خروجی شبکه را به هدف نزدیک نماید. اما در یادگیری بدون ناظر یا نظارت نشده^۱ ورودی‌ها و بایاس‌ها تنها در مقابل ورودی شبکه اصلاح می‌شوند و در واقع هیچ هدفی وجود ندارد. این نوع الگوریتم‌ها غالباً برای دسته‌بندی سری داده‌ها مورد استفاده قرار می‌گیرند. عموماً برای آموزش شبکه‌های عصبی از قواعد یادگیری نظارت شده استفاده می‌شود اما می‌توان شبکه‌ها را با روش آموزش بدون ناظر نیز آموزش داد [۴۳].

۲-۷- انواع شبکه‌های عصبی

۲-۷-۱- شبکه‌های پیش‌خور^۲

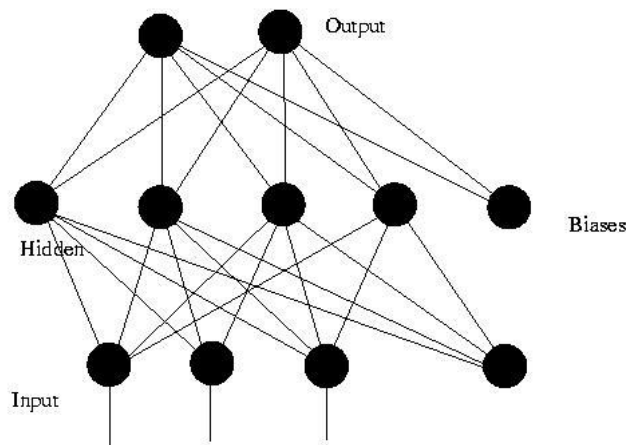
شبکه‌های پیش‌خور، شبکه‌هایی هستند که مسیر پاسخ در آن‌ها همواره رو به جلو پردازش شده و به نرون‌های همان لایه یا لایه قبل باز نمی‌گردد. در این نوع شبکه‌ها به سیگنال اجازه داده می‌شود که از مسیر یک‌طرفه، یعنی از ورودی تا خروجی عبور کند و در نتیجه بازخوردی^۳ وجود

۱- Unsupervised Training

۲- Feedforward

۳- Recurrent

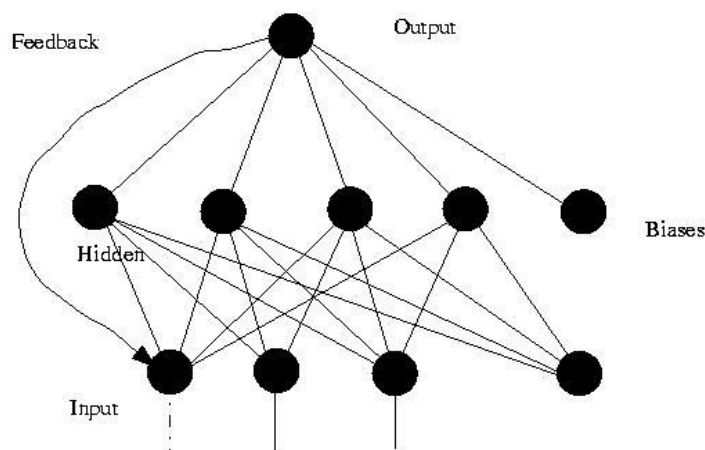
نداشته و خروجی هر لایه تأثیری بر همان لایه و همچنین لایه‌های قبلی ندارد. شکل ۷-۲ نمونه‌ای از شبکه پیش‌خور را ارائه می‌دهد.



۷-۲- شماتیک شبکه عصبی مصنوعی از نوع پیش‌خور [۴۵]

۷-۲-۲- شبکه‌های پس‌خور یا برگشتی^۱

در شبکه‌های برگشتی حداقل یک سیگنال از یک نرون به همان نرون یا نرون‌های دیگر همان لایه و یا نرون‌های لایه قبل برگشت می‌کند؛ یا به عبارتی در این نوع شبکه‌ها بازخورد وجود دارد. شکل ۸-۲ نمونه‌ای از یک شبکه برگشتی را نمایش می‌دهد [۳۹، ۴۴].



۸-۲- شماتیک شبکه عصبی مصنوعی از نوع پس‌خور [۴۵]

۸-۲- طراحی شبکه پیش‌خور

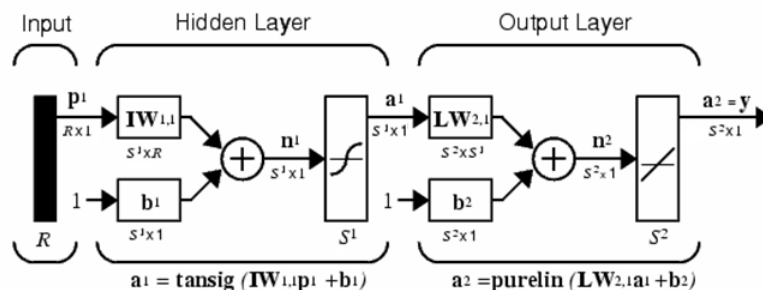
در شبکه‌های پیش‌خور، نرون‌ها در چند لایه سازمان‌دهی می‌شوند، بدین ترتیب که اولین و آخرین لایه به ترتیب لایه ورودی^۱ و لایه خروجی^۲ بوده، در حالی که لایه‌های بین این دو لایه را لایه‌های پنهان^۳ می‌نامند. در این نوع شبکه‌ها، ابتدا اطلاعات ورودی از محیط خارج به نرون‌های لایه ورودی ارسال می‌شود، اما در این لایه هیچ پردازشی بر روی ورودی‌ها صورت نمی‌گیرد. در واقع لایه ورودی صرفاً نقش دریافت‌کننده و انتقال‌دهنده 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 ۹-۲، شبکه‌های پیش‌خور اغلب از یک یا چند لایه مخفی از نرون‌های سیگموئید و از یک لایه پایانی خطی استفاده می‌کنند. شبکه چند لایه از نرون‌ها با یک تابع انتقال 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 ∞ باشد.

۱- Input Layer
۲- Output Layer
۳- Hidden Layer

تا $+\infty$ قرار گیرد. البته اگر بخواهیم خروجی در دامنه محدودی قرار گیرد، می‌توان از توابع

محدودساز سیگموئید در لایه خروجی استفاده کرد [۴۱].



شکل ۹-۲- شبکه پیش‌خور tansig/purelin [۳۸]

۹-۲- آموزش شبکه‌های پیش‌خور

به طور کلی مراحل این فرایند شامل ایجاد شبکه، تعیین مجموعه داده‌های آموزش، مقدار دهی

آغازین به وزن‌ها، آموزش شبکه و شبیه‌سازی^۱ پاسخ شبکه می‌باشد [۴۱].

۹-۲-۱- ایجاد یک شبکه پیش‌خور

اولین مرحله در طراحی یک شبکه پیش‌خور، ایجاد شبکه می‌باشد که تابع newff نیز یک شبکه پیش‌خور ایجاد می‌کند. این تابع چهار ورودی داشته و شبکه تهیه شده را به عنوان خروجی برمی‌گرداند. اولین پارامتر ماتریس بردار ورودی، دومین پارامتر ماتریس مقادیر مطلوب خروجی‌ها، سومین پارامتر نوع تابع انتقال و آخرین پارامتر نیز نوع تابع آموزش شده می‌باشد.

۹-۲-۲- مقداردهی آغازین به وزن‌ها

قبل از آموزش یک شبکه پیش‌خور، مقادیر ابتدایی وزن‌ها و بایاس‌ها باید تعیین گردند. تابع newff به صورت خودکار به وزن‌ها مقدار می‌دهد ولی ممکن است با دستور init آن را به صورت غیرخودکار نیز انجام داد.

۲-۹-۳- آموزش شبکه

پس از مقداردهی به وزن‌ها و بایاس‌ها نوبت به آموزش شبکه می‌رسد. شبکه می‌تواند برای تقریب توابع، تشخیص الگو و یا طبقه‌بندی الگوها مورد استفاده قرار گیرد. فرایند آموزش به یک سری مثال‌ها از رفتار مورد انتظار شبکه نیاز دارد که شامل ورودی شبکه (p) و هدف (t) می‌باشد. در طول فرایند آموزش وزن‌ها و بایاس‌ها تنظیم شده تا تابع کارایی^۱ شبکه، بهینه گردد. تابع کارایی پیش‌فرض برای شبکه‌های پیش‌خور، میانگین مجموع مربعات خطاها (MSE) می‌باشد که طبیعتاً بایستی مینیمم شود.

$$mse = \frac{1}{N} \sum_{i=1}^N (e_i)^2 = \frac{1}{N} \sum_{i=1}^N (t_i - a_i)^2 \quad \text{معادله (۲-۶)}$$

۲-۹-۴- شبیه‌سازی

تابع sim یک شبکه را شبیه‌سازی می‌کند. این تابع شبکه و بردار ورودی را به عنوان ورودی دریافت و خروجی شبکه را بر می‌گرداند [۳۳].

۲-۱۰- الگوریتم‌های آموزشی شبکه عصبی

در شبکه‌های عصبی برای بدست آوردن ارتباط بین متغیرهای ورودی و خروجی از فرایند یادگیری با یک الگوی آموزشی استفاده می‌شود. قدرت شبکه‌های عصبی در همین آموزش‌پذیری آنهاست که آنها را قادر می‌سازد به طور نسبی به تعمیم آموخته‌های خویش بپردازند؛ چرا که آنچه یادگیری را از حفظ کردن متفاوت و بالاتر می‌سازد همین قدرت تعمیم آموخته‌ها و اطلاعات اولیه است [۳۹]. برای آموزش شبکه‌های عصبی از الگوریتم‌های یادگیری متفاوتی مانند الگوریتم

۱- Performance Function Modification

نزول گرادینانی^۱، الگوریتم پسانتشار گرادیان مزدوج^۲ و الگوریتم لونبرگ-مارکوارت^۳ استفاده می‌شود که همگی به عنوان زیرمجموعه‌ی یک تکنیک آموزشی به نام پسانتشار^۴ یا انتشار خطا به شمار می‌روند. انتخاب هر الگوریتم بر سرعت یادگیری و دقت شبکه مؤثر است.

پسانتشار خطا یک روش متداول آموزش با ناظر برای شبکه‌های 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 دارای 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 اتصالاتی که اولین بار برای به دست آوردن خروجی استفاده شده بودند، بازپراکنده (باز پخش) می‌شود. اگر خروجی شبکه با هدف همانندی مناسبی داشته باشد، اتصال واحدهایی که در خروجی مؤثر بوده‌اند تقویت می‌شود؛ اگر چنین نباشد، اتصال بین 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.

دو روش مختلف برای ارائه الگوهای آموزشی و پیاده‌سازی الگوریتم وجود دارد:

در روش اول که آموزش گام به گام^۵ نام دارد، وزن‌ها و بایاس‌ها بعد از اعمال هر ورودی به روز^۶ می‌شوند در حالی که در روش آموزش دسته‌ای^۷، پس از اعمال تمام ورودی‌ها (اعضای مجموعه آموزشی) عملیات به روزرسانی وزن‌ها انجام می‌گیرد. بدین صورت که شیب‌های محاسبه شده برای هر ورودی با هم جمع می‌گردد تا در نهایت وزن‌ها و بایاس‌ها از طریق آن به روز شوند [۴۱]. در این تحقیق شیوه آموزش دسته‌ای مورد استفاده قرار گرفته است.

۱- Gradient Descent

۲- Conjugate Gradient (CG)

۳- Levenberg - Marquardt (LM)

۴- Backpropagation

۵- Incremental Training

۶- Update

۷- Batch Training

به طور کلی آموزش به کمک تکنیک پس‌انتشار بر طبق مراحل زیر انجام می‌شود [۳۵]:

(الف) اختصاص ماتریس وزن تصادفی به هریک از اتصالات

(ب) انتخاب بردار ورودی و خروجی متناسب با آن و انتشار ورودی‌ها از نرون‌های ورودی به سمت

نرون‌های خروجی

(ج) محاسبه خروجی نرون در هر لایه و در نتیجه محاسبه خروجی نرون‌ها در لایه‌ی خروجی

(د) مقایسه خروجی‌های شبکه با مقادیر واقعی (مقادیر هدف) و محاسبه‌ی خطای شبکه

(ه) به روزسازی و اصلاح وزن‌ها به روش انتشار خطای شبکه به لایه‌های قبل که خطای یاد شده

ناشی از اختلاف بین خروجی واقعی و خروجی محاسبه شده است.

(و) ارزیابی عملکرد شبکه آموزش دیده به کمک تابع کارایی تعریف شده مانند میانگین مجموع

مربعات خطاها (MSE)

(ز) و سرانجام برگشت به قسمت پایان آموزش.

مراحل فوق تا زمانی تکرار می‌شود که به حداکثر تکرار^۱ مجاز رسیده یا مقدار تابع کارایی از

مقداری که تعیین شده کمتر باشد.

به دلیل فرایندی که در طی این تکنیک رخ می‌دهد، این روش پس‌انتشار نامیده می‌شود زیرا خطای

خروجی برای اصلاح وزن‌ها به شبکه برگشت داده می‌شود بدین ترتیب که ابتدا وزن‌های ارتباطی بین

لایه خروجی و لایه پنهان و سپس وزن‌های ارتباطی بین لایه پنهان و لایه ورودی اصلاح می‌گردند.

در ادامه به بررسی دو الگوریتم نزول گرادیانی و لوبنبرگ-مارکوارت می‌پردازیم.

۲-۱۰-۱- الگوریتم آموزشی نزول گرادیانی

متداول‌ترین الگوریتمی که در الگوی آموزشی پس‌انتشار به کار می‌رود، الگوریتم آموزشی نزول

گرادیانی است. در این الگوریتم آموزشی، پارامترهای شبکه از قبیل وزن‌ها و بایاس‌ها در جهتی که

^۱ - Epoch

تابع کارایی حداقل شود، تغییر می‌کنند. به روز رسانی وزن‌ها در گام آموزشی $n+1$ به صورت روابط ۷-۲ و ۸-۲ می‌باشد:

$$w_{ji,n+1} = w_{ji,n} + \alpha \Delta w_{ji,n} \quad \text{معادله (۷-۲)}$$

$$\Delta w_{ji,n} = - \frac{\partial E}{\partial w_{ji,n}} \quad \text{معادله (۸-۲)}$$

که در روابط فوق $w_{ji,n}$ وزن ارتباطی بین نرون i ام و نرون j ام در گام آموزشی n ام، α سرعت یادگیری و E نیز تابع خطاست. سرعت یادگیری، اندازه گام‌های برداشته شده در اصلاح وزن‌ها را نشان می‌دهد. سرعت یادگیری زیاد باعث نوسان‌های زیاد در حین آموزش شده در حالی که مقدار کم آن باعث آهستگی سرعت هم‌گرایی می‌شود. علامت منفی در رابطه نشان می‌دهد که حرکت به سمتی است که تابع خطا حداقل شود و یا به بیان دیگر هدف از تصحیح وزن‌ها کاهش مقدار خطاست [۳۳].

۲-۱۰-۲- الگوریتم آموزشی لونیبرگ-مارکوارت

با وجود اینکه نزول گرادیانی، الگوریتمی متداول است؛ دو ضعف عمده و اساسی دارد: اول اینکه سرعت همگرایی آن پایین است و دومین ضعف این روش متوقف شدن پارامترهای شبکه به دلیل به دام افتادن آنها در نقاط مینیمم محلی^۱ قبل از پایان یادگیری کامل شبکه می‌باشد [۴۶، ۴۷]. الگوریتم دیگری که در الگوی آموزشی پس‌انتشار استفاده می‌شود، الگوریتم آموزشی لونیبرگ-مارکوارت است. روش آموزشی لونیبرگ-مارکوارت نسبت به روش پس انتشار صدها برابر سریع‌تر است. ویژگی دیگر این روش کاهش شدید تعداد نرون‌های لایه مخفی در ساختار شبکه است. این الگوریتم آموزشی برای اولین بار توسط لونیبرگ در سال ۱۹۹۴ پیشنهاد شد و در آن هم از مشتق اول تابع خطا نسبت به پارامترهای شبکه (گرادیان) و هم از مشتق دوم آن موسوم به هسیان^۲ برای اصلاح وزن‌ها استفاده می‌شود. این روش در حقیقت ترکیبی از دو روش نزول گرادیان و

۱- Local Minimum

۲- Hessian

گوس-نیوتن^۱ است که در آن از دقت بالای روش اول به همراه سرعت روش دوم در تعیین پارامترهای بهینه توابع غیرخطی استفاده می‌شود [۴۸]. در این روش آموزشی ماتریس‌های گرادیان و هسیان طبق روابط ۹-۲ و ۱۰-۲ تقریب زده می‌شوند:

$$H(x) = \nabla^2 F(x) \approx 2J^T(x)J(x) \quad \text{معادله (۹-۲)}$$

$$g(x) = \nabla F(x) \approx 2J^T(x) E(w) \quad \text{معادله (۱۰-۲)}$$

در این معادلات، x بیان گر وزن‌های شبکه عصبی و H ماتریس هسیان است که به صورت معادله ۱۱-۲ بیان می‌شود:

$$H = \nabla^2 F(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(x) & \frac{\partial^2}{\partial x_1 \partial x_2} F(x) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} F(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(x) & \frac{\partial^2}{\partial x_2^2} F(x) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} F(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} F(x) & \frac{\partial^2}{\partial x_n \partial x_2} F(x) & \dots & \frac{\partial^2}{\partial x_n^2} F(x) \end{bmatrix}_{n \times n} \quad \text{معادله (۱۱-۲)}$$

$E(w)$ نیز بردار خطای شبکه و J ماتریس ژاکوبین^۲ می‌باشد که شامل مشتقات مرتبه اول خطای شبکه نسبت به پارامترهای شبکه است و به صورت معادله ۱۲-۲ محاسبه می‌شود:

۱- Gauss-Newton
۲- Jacobian Matrix

$$J(x) = \begin{bmatrix} \frac{\partial e_1(x)}{\partial x_1} & \frac{\partial e_1(x)}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial e_1(x)}{\partial x_n} \\ \frac{\partial e_2(x)}{\partial x_1} & \frac{\partial e_2(x)}{\partial x_2} & \cdot & \cdot & \cdot & \frac{\partial e_2(x)}{\partial x_n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \frac{\partial e_n(x)}{\partial x_1} & \frac{\partial e_n(x)}{\partial x_2} & & & & \frac{\partial e_n(x)}{\partial x_n} \end{bmatrix}_{n \times n} \quad \text{معادله (۱۲-۲)}$$

طبق این الگوریتم، تغییر پارامترها در هر گام آموزشی برابر است با:

$$x_{k+1} = x_k - [J^T(x_k)J(x_k) + \mu_k I]^{-1} * J^T(x_k)E(x_k) \quad \text{معادله (۱۳-۲)}$$

این الگوریتم حالت بینابین روش نزول گرادیانی و روش گوس-نیوتن است که در آن I ماتریس یکانی و μ_k ضریب یادگیری مارکوارت و عددی غیر منفی می باشد. زمانی که مقدار عددی μ_k صفر باشد، این تابع تبدیل به یک روش گوس-نیوتن برای تقریب ماتریس هسیان می شود؛ و زمانی که μ_k بزرگ باشد، این تابع مبدل به روش نزول گرادیانی می شود. بنابراین μ_k بعد از هر گام موفق (کاهش در تابع کارایی) کاهش یافته و تنها زمانی که گام آزمایشی موجب افزایش تابع کارایی شود، افزایش پیدا می کند. به این ترتیب تابع کارایی همیشه در حال کاهش است [۴۷، ۴۸].

پارامترهای آموزشی برای الگوریتم لونبرگ-مارکوارت عبارتند از:

⇐ دوره ها (epochs)

⇐ نمایش (show)

⇐ هدف (goal)

⇐ زمان (time)

⇐ مینیمم گرادیان (min-grad)

⇐ ماکزیمم رد شدن (max-fail)

⇐ پارامتر μ

⇐ کاهش μ (mu-dec)

⇐ افزایش μ (mu-inc)

⇐ ماکزیمم μ (mu-max)

⇐ کاهش حافظه (mem-reduc)

پارامتر μ مقدار آغازین μ می‌باشد و در هر گامی که تابع کارایی کاهش یابد در mu-dec و هر زمان که افزایش یابد در mu-inc ضرب می‌شود. اگر μ از mu-max بزرگ‌تر شود، الگوریتم متوقف می‌گردد. پارامتر mem-reduc مقدار حافظه مورد استفاده الگوریتم را کنترل می‌کند. این روش یکی از سریعترین روش‌های ارائه شده در نرم‌افزار MATLAB می‌باشد و برای یک شبکه متوسط (با چند صد پارامتر مؤثر) دارای کارایی بسیار بالا می‌باشد [۴۱].

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 (بیش‌آموزی^۲ یا بیش‌برازش^۳). بدین معنی که در هنگام آموزش شبکه، خطا به مقدار قابل قبول می‌رسد ولی هنگام ارزیابی داده‌های جدید، خطای شبکه به مراتب از خطای داده‌های آموزشی بیشتر گردیده که گفته می‌شود شبکه تعمیم برای حالت‌های جدید را یاد نگرفته است. علت این امر آن است که شبکه کم‌کم 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 شبکه مورد استفاده کوچک باشد، توانایی لازم برای انطباق داده‌ها را نخواهد داشت. متأسفانه از ابتدا نمی‌توان پیش‌بینی کرد که اندازه شبکه

۱- Improving Generalization

۲- Overtraining

۳- Overfitting

برای یک کاربرد خاص تا چه میزان باید بزرگ باشد. در جعبه ابزار شبکه عصبی برای بهبود تعمیم و جلوگیری از بیش‌آموزی شبکه دو راه حل تنظیم^۱ و توقف زودهنگام^۲ در نظر گرفته شده است [۴۹،۵۰].

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 ۲-۱۴).

$$F = \beta E_D + \alpha E_W \quad \text{معادله (۲-۱۴)}$$

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 بوده و طبق روابط ۲-۱۵ و ۲-۱۶ تعریف می‌شوند.

$$E_D = \frac{1}{N} \sum_{i=1}^N (e_i)^2 = \frac{1}{N} \sum_{i=1}^N (y_i - t_i)^2 \quad \text{معادله (۲-۱۵)}$$

$$E_W = \frac{1}{N} \sum_{j=1}^N (w_j)^2 \quad \text{معادله (۲-۱۶)}$$

که e_i بیان‌گر خطای مطلق، t_i مقادیر پارامتر تجربی، y_i مقادیر محاسبه شده برای پارامتر تجربی مدّ نظر و N نشان‌دهنده‌ی تعداد مولکول‌ها می‌باشد.

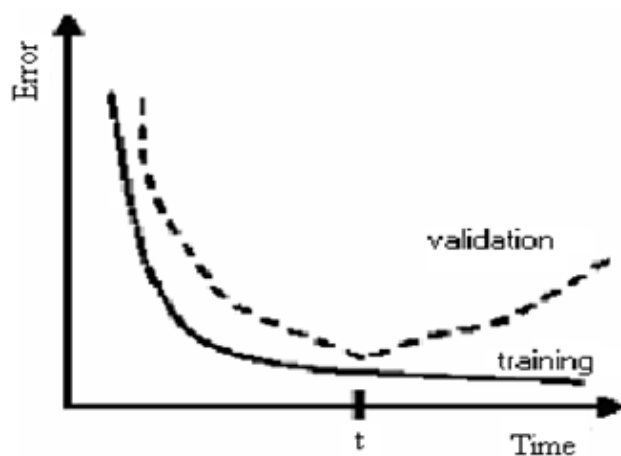
۱- Regularization

۲- Early Stopping

قابل ذکر است که در الگوریتم آموزشی تنظیم بایزین، اصلاح وزن‌ها و بایاس‌ها همانند الگوریتم لونبرگ-مارکوارت انجام می‌شود با این تفاوت که تابع خطا، همان تابع کارایی اصلاح شده است؛ یعنی تابع خطا ترکیبی از میانگین مربع خطاها و میانگین مربع وزن‌هاست. هدف از آموزش به روش پس انتشار در این الگوریتم نیز، کاهش این تابع کارایی اصلاح شده است [۵۱].

۲-۱۱-۲- توقف زودهنگام

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 به طور طبیعی در طول فاز اولیه آموزش کاهش می‌یابد، همان گونه که خطای سری آموزشی کاهش می‌یابد؛ لازم به ذکر است که خطای سری ارزیابی همواره از خطای سری آموزش بیشتر است. زمانی که شبکه شروع به برآزش 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 ۲-۱۰ روند آموزش می‌بایست در زمان t متوقف گردد.



۲-۱۰- تغییرات خطای سری آموزش و سری ارزیابی [۴۱]

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 [۴۱].

۱۲-۲- ارزیابی مدل

برای اطمینان از اینکه مدل به دست آمده، مدل مناسبی بوده و قادر است نمونه‌های مختلفی را در یک جمعیت پیش‌بینی کند، باید مدل را ارزیابی نمود. 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، عبارتند از [۳۳]:

ضریب همبستگی^۲ یا آماره R: محاسبه‌ی ضریب همبستگی، ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر است. ضریب همبستگی دو متغیر y و x طبق رابطه ۲۰-۲ تعریف می‌شود:

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (20-2)$$

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.

ضریب تعیین^۳ یا آماره R^۲: به عنوان یک شاخص برای بیان دقت خط رگرسیون به کار رفته و مقدار آن بین صفر و یک تغییر می‌کند. این آماره نشان‌دهنده‌ی میزان تغییرات متغیر وابسته توسط متغیر مستقل است. ضریب تعیین با استفاده از رابطه ۲۱-۲ محاسبه می‌گردد:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (21-2)$$

۱- Statistical Indices

۲- Correlation Coefficient

۳- Determination Coefficient

که SSR^1 (مجموع مجذورات رگرسیون) طبق رابطه ۲-۲۲ مبین مجموع مربعات انحراف مقادیر

پیش‌بینی شده‌ی متغیر وابسته از میانگین مقادیر آن است:

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (22-2)$$

SST^2 (مجموع مجذورات کل) طبق رابطه ۲-۲۳ نشان‌دهنده‌ی مجموع مربعات انحراف مقادیر

تجربی متغیر وابسته از میانگین مقادیر آن می‌باشد:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (23-2)$$

SSE^3 (مجموع مجذورات خطاهای رگرسیون یا باقی‌مانده‌ها) نیز بیان‌گر مجموع مربعات انحراف

مقادیر تجربی متغیر وابسته از مقادیر پیش‌بینی شده برای آن است:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (24-2)$$

بنابراین با توجه به روابط فوق می‌توان معادله‌ی ۲-۲۱ را به شکل زیر بازنویسی نمود:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (25-2)$$

بر اساس رابطه ۲-۲۵ وقتی مقدار R^2 برابر یک می‌شود که به ازای تمامی نقاط $y_i = \hat{y}_i$ باشد.

هرگونه انحراف از این حالت باعث می‌شود مقدار R^2 از یک کوچکتر شود.

آزمون فیشرف یا آماره F: آزمون F در واقع آزمون معنی‌دار بودن آماری در تحلیل رگرسیون ساده و

چند متغیره است و برابر با نسبت میانگین مربعات رگرسیون (MSR) به میانگین مربعات باقی‌مانده‌ها

(MSE) است. بیان ریاضی آن به صورت رابطه ۲-۲۷ می‌باشد:

۱- Sum Square Regression

۲- Sum Square Total

۳- Sum Square Error

$$F = \frac{MSR}{MSE} = \frac{SSR/df_m}{SSE/df_{res}} \quad (27-2)$$

$$df_m = p \quad df_{res} = n - p - 1$$

df_m درجه آزادی مدل، p تعداد متغیرهای مستقل آن، n تعداد کل ترکیبات مربوط به مدل و df_{res} درجه آزادی باقی مانده ها می باشد.

مجموع مربعات باقی مانده ها^۱ (PRESS): برابر مجموع مجذورات تفاوت بین مقدار کمیت مشاهده شده (y_i) و مقدار تخمین زده شده ($\hat{y}_i$) است.

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (28-2)$$

میانگین مربع خطاها^۲ (MSE): توسط رابطه ۲۹-۲ تعریف می گردد که پارامترهای آن قبلاً تعریف شده اند.

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (29-2)$$

خطای استاندارد پیش بینی^۳ (SEP):

$$SEP = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (30-2)$$

خطای نسبی پیش بینی^۴ (REP):

$$REP(\%) = \frac{100}{\bar{y}} \times SEP = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (31-2)$$

۱- Predictive Residual Sum of Squares

۲- Mean Square Error

۳- Standard Error of Prediction

۴- Relative Error of Prediction

خطای مطلق میانگین^۱ (MAE):

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}|}{n} \quad (۳۲-۲)$$

۱۳-۲- تکنیک‌های اعتبارسنجی

این تکنیک‌ها برای ارزیابی اعتبار مدل‌های به دست آمده و بررسی قدرت پیش‌بینی 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 [۵۲]. 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 نزدیک‌تری را برای تعداد زیادی از پارامترها و متغیرها در مدل به دست می‌دهند (SE کوچک‌تر و R^2 بزرگ‌تر)، در اعتبارسنجی مدل کافی نیست؛ زیرا این پارامترها ارتباطی به توانایی مدل برای ایجاد پیش‌بینی‌های مطمئن بر روی داده‌های بعدی ندارند. مسائل دیگر برای اعتبارسنجی مدل‌ها زمانی پیش می‌آیند که لازم است مدلی با متغیرهای اندک، به کمک فرآیندهای بر پایه انتخاب متغیر به دست آیند [۵۳]. از طرف دیگر زمانی که مجموعه‌ای شامل تعداد زیادی توصیف‌کننده برای انتخاب از بین آنها در دسترس باشد، مدل‌های به دست آمده برازش خوبی دارند ولی همراه با همبستگی شانس^۲ هستند. اگر در مدلی متغیرهای مستقل به صورت تصادفی با متغیرهای پاسخ همبستگی داشته باشند، گفته می‌شود مدل دارای همبستگی شانس است بدین معنی که منحنی مقادیر نظری محاسبه شده از مدل بر حسب مقادیر تجربی در یک خط مستقیم واقع شده بدون این که توانایی پیش‌بینی داشته

۱- Mean Absolute Error

۲- Chance Correlation

باشد. برای جلوگیری از ایجاد مدل‌های با همبستگی شانس، باید یک اعتبارسنجی متفاوت انجام گیرد. برای این منظور بهتر است یک سری ارزیابی مناسب از ترکیبات که در فرایند آموزش و بهینه‌سازی دخالت نداشته باشد به نام سری ارزیابی بیرونی^۱، در دسترس باشد [۵۴]. تعدادی از تکنیک‌های اعتبارسنجی در زیر فهرست شده‌اند.

۲-۱۳-۱- ارزیابی متقاطع یا اعتبارسنجی تقاطعی^۲

رایج‌ترین تکنیک اعتبارسنجی 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 کنارگذاشته شده از روی این مدل پیش‌بینی می‌شود [۵۵]. این روش‌ها به ترتیب به نام‌های حذف تک‌تک داده‌ها^۳ و حذف گروهی از داده‌ها^۴ نامیده می‌شوند.

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

در این تکنیک مجموعه داده به یک سری آموزش^۵ و یک سری ارزیابی^۶ تقسیم‌بندی می‌شود. مدل از سری آموزش به دست آمده و قدرت تعمیم و پیش‌بینی توسط سری ارزیابی بررسی می‌گردد. تقسیم‌بندی داده‌ها به صورت کاملاً تصادفی انجام گرفته و نتایج نیز تا حد زیادی به این تقسیم‌بندی وابسته‌اند [۵۴].

۱- External Validation Set

۲- Cross Validation

۳- Leave One Out

۴- Leave Group Out

۵- Train Series

۶- Validation Series

۲-۱۳-۳- اعتبارسنجی بیرونی^۱

تکنیکی است که در آن یک سری تست^۲ نیز برای انجام یک بررسی اضافی بر روی قدرت پیش‌بینی مدل به دست آمده از سری آموزش و با قدرت پیش‌بینی بهینه شده توسط سری ارزیابی، در نظر گرفته می‌شود. سری تست نباید دخالتی در روند آموزش و انتخاب مدل داشته باشد [۵۴].

۲-۱۳-۴- آزمون Y-تصادفی^۳

این تکنیک برای مطالعه همبستگی‌های شانس مدل غیرخطی طراحی شده است. در این آزمون، مقادیر تجربی (که در اینجا بردار Y نامیده می‌شود) 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 معمولاً R^2 است، مورد بررسی قرار می‌گیرد. اختلاف زیاد بین شاخص آماری به دست آمده از این روش و شاخص آماری به دست آمده از مدل اصلی، نشان‌دهنده عدم وجود همبستگی شانس می‌باشد. به طور معمول این فرایند چندین بار تکرار می‌گردد [۵۶].

۲-۱۴- نرم‌افزارهای مورد استفاده در این تحقیق

در این تحقیق از بسته‌های نرم‌افزاری HyperChem08 برای رسم ساختار ترکیبات مورد بررسی، Dragon2/1 برای محاسبه توصیف‌کننده‌های مولکولی، SPSS17 برای انجام آنالیزهای آماری رگرسیون‌های خطی چندگانه و MATLAB R2008a به منظور برنامه‌نویسی شبکه عصبی مصنوعی استفاده شد، که در ادامه به معرفی آنها می‌پردازیم. برای رسم نمودارهای سه‌بعدی نیز نرم‌افزار SigmaPlot مورد استفاده قرار گرفت.

۱- External Validation

۲- Test Series

۳- Y-Randomization

۲-۱۴-۱- نرم افزار Hyperchem

یک نرم افزار مدل سازی و شبیه سازی است که کاربران را قادر می سازد از آن برای رسم مولکول ها و بهینه سازی ساختار آنها با استفاده از روش های کوانتومی و مکانیکی و همچنین به منظور محاسبات پیچیده ی شیمیایی بهره برداری کنند. به کمک این برنامه می توان طول پیوند، زاویه پیوندی، زوایای پیچشی، بار اتم ها و انرژی تشکیل مولکول و غیره را در مولکول تعیین نمود. این بسته نرم افزاری، قابلیت ترسیم ساختار ترکیبات پیچیده را داشته و در کتابخانه ی آن ساختار نوکلئیک اسیدها و اسیدهای آمینه نیز موجود است [۲۷]. داده های حاصل از این نرم افزار را می توان به عنوان ورودی به سایر نرم افزارها نظیر Dragon معرفی نمود.

۲-۱۴-۲- نرم افزار Dragon

اولین ویرایش این بسته نرم افزاری در سال ۱۹۹۷ توسط گروه QSAR و کمومتریکس دانشگاه میلانو به منظور محاسبه ی توصیف کننده های مولکولی طراحی گردید. نرم افزار Dragon قابلیت محاسبه ۱۴۸۱ توصیف کننده برای هر مولکول با کمتر از ۱۵۰ اتم را داراست. جهت محاسبه توصیف کننده توسط این نرم افزار لازم است ساختار هندسی بهینه مولکول مورد استفاده قرار گیرد. این توصیف کننده ها برای ارزیابی روابط ساختار- فعالیت یا ساختار-ویژگی مولکولی به کار رفته و به ۱۸ دسته اصلی تقسیم می شوند که در جدول ۲-۱ آورده شده است [۵۷].

جدول ۱-۲- انواع توصیف‌گرهای محاسبه شده توسط نرم افزار Dragon

۱- توصیف‌کننده‌های زیر ساختاری ^۱	۲- توصیف‌کننده‌های توپولوژیکی ^۲
۳- شمارنده‌های مولکولی مورس ^۳	۴- توصیف‌کننده‌های BCUT ^۴
۵- شاخص‌های بار ^۵	۶- خود ارتباطی‌های دو بعدی ^۶
۷- توصیف‌کننده‌های بار ^۷	۸- شاخص‌های آروماتیسیت ^۸
۹- پروفایل‌های مولکولی راندیک ^۹	۱۰- توصیف‌کننده‌های هندسی ^{۱۰}
۱۱- توصیف‌کننده‌های RDF ^{۱۱}	۱۲- توصیف‌کننده‌های سه بعدی ^{۱۲}
۱۳- توصیف‌کننده‌های WHIM ^{۱۳}	۱۴- توصیف‌کننده‌های GETAWAY ^{۱۴}
۱۵- گروه‌های عاملی ^{۱۵}	۱۶- اجزای میان اتمی ^{۱۶}
۱۷- توصیف‌کننده‌های تجربی ^{۱۷}	۱۸- خصوصیات مولکولی ^{۱۸}

۱- Constitutional Descriptors

۲- Topological Descriptors

۳- Molecular Walk Counts

۴- BCUT Descriptors

۵- Charge Indices

۶- 2D Autocorrelations

۷- Charge Descriptors

۸- Aromatic Indices

۹- Rancid Molecular Profiles

۱۰- Geometrical Descriptors

۱۱- Radial Distribution Function Descriptors

۱۲- 3D Molecule Representation of Structure Backed on Electron Diffraction Descriptors (3D-MoRSE Descriptors)

۱۳- Weighted Holistic Invariant Molecular Descriptors

۱۴- Geometry, Topology and Atom Weights Assembly

۱۵- Functional Groups

۱۶- Atom-Centered Fragments

۱۷- Empirical Descriptors

۱۸- Properties

۲-۱۴-۳- نرم افزار SPSS^۱

این بسته‌ی نرم افزاری، به منظور تجزیه و تحلیل داده‌ها تدوین شده است. یادگیری آسان، سادگی استفاده و قدرت بالا در انجام محاسبات پیچیده، SPSS را به یکی از رایج‌ترین بسته‌های نرم‌افزاری مورد استفاده در تجزیه و تحلیل آماری مبدل نموده است. در این پروژه از این نرم افزار، برای بدست آوردن بهترین مدل خطی استفاده شده است [۳۴].

۲-۱۴-۴- نرم افزار MATLAB^۲

MATLAB یکی از جامع‌ترین نرم‌افزارها با کارایی بالا برای محاسبات تخصصی بوده و به عنوان نرم‌افزاری قوی و در عین حال ساده، امکان حل مسائل پیچیده علی‌الخصوص مسائلی که به صورت ماتریسی هستند را فراهم می‌کند. یکی از قابلیت‌های این نرم‌افزار برنامه‌نویسی در محیطی ساده بوده که در آن، مسائل و راه‌حل‌ها با علائم ریاضی مشابه تعریف می‌شوند. کاربردهای نوعی آن شامل ریاضیات و الگوریتم محاسباتی، شبیه‌سازی، مدل‌سازی و آنالیز اطلاعات و... می‌باشد. از جمله کاربردهای مهم آن که مورد استقبال شیمی‌دانان قرار گرفته است، امکان استفاده برای آنالیز آماری و مدل‌سازی داده‌های شیمیایی به روش شبکه عصبی مصنوعی است که در تحقیق حاضر مورد استفاده قرار گرفته است. اطلاعات شیمیایی ترکیبات به عنوان ورودی، به شکل ماتریس به نرم‌افزار ارائه می‌شوند. در محیط MATLAB با استفاده از فرامین، آرایه‌ها و امکانات ویژه، مدل‌سازی غیرخطی فعالیت‌ها و خواص شیمیایی ترکیبات با ساختار آنها امکان‌پذیر است. سایر روش‌های کمومتریکس مانند آنالیز رگرسیون اجزای اصلی، روش کمترین مربعات جزئی، الگوریتم ژنتیک و غیره نیز توسط این نرم افزار قابل اجرا هستند [۳۸].

۱- Statistical Package for the Social Science

۲- Matrix Laboratory

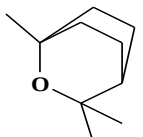
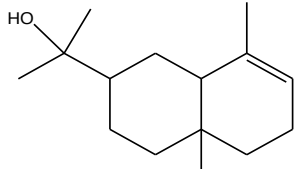
فصل سوم

مدل سازی QSPR شاخص بازداري برخي از
اجزای فرّار روغن های ضروري، با استفاده از
روش رگرسیون خطی چندگانه (MLR) و
شبکه عصبی مصنوعی (ANN)

۳-۱- سری داده‌ها

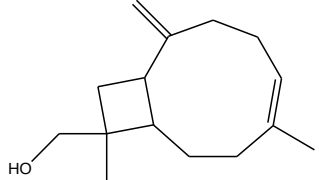
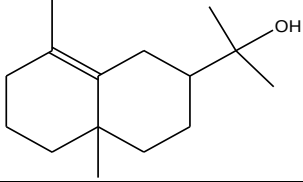
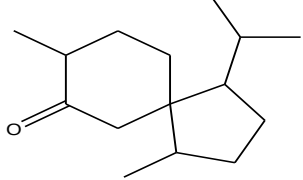
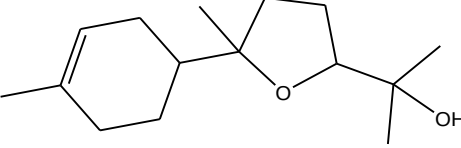
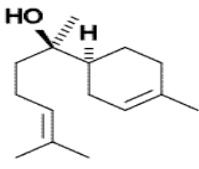
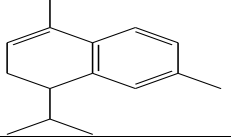
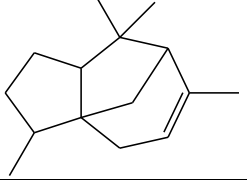
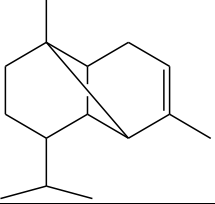
در این قسمت، اندیس بازداری ۷۷ روغن ضروری با گروه‌های عاملی گوناگون، به عنوان سری داده‌ها مورد استفاده قرار گرفت. آنالیز این مواد با استفاده از سیستم GC-MS انجام گرفت که دربرگیرنده‌ی گاز-کروماتوگراف CP-3900 و یک اسپکترومتر جرمی به‌دام‌انداز یون^۱ بود. جداسازی‌ها توسط یک ZB-5 (۵٪ فنیل و ۹۵٪ دی‌متیل‌پلی‌سیکلوپنتان) در ستون کاپیلاری سیلیکای مذاب^۲ (۰/۲۵ mm i. d. × ۳۰ m و ضخامت فیلم ۰/۲۵ μm) انجام گرفت. دمای آون از ۵۰ تا ۲۵۰°C با سرعت ۳ °C min⁻¹ تغییر داده شده و گاز هلیوم ۹۹/۹۹٪ به عنوان گاز حامل با یک سرعت ثابت ۱/۰ ml min⁻¹ و حجم تزریق ۱ μl به کار برده شد [۴]. نام، ساختار مولکولی و اندیس بازداری این ترکیبات در جدول ۳-۱ آمده است. داده‌های جدول ۳-۱ بیان‌گر این است که نوع استخلاف‌ها، تعداد شاخه‌ها و نحوه‌ی آرایش اتم‌ها در مولکول بر روی شاخص بازداری مؤثر است. لذا در این تحقیق سعی شده مدلی ارائه گردد که بتواند شاخص بازداری یک ترکیب خاص را با هر نوع استخلاف و در هر موقعیتی پیش‌بینی کند بدون اینکه نیاز به سنتز و آزمایش ترکیب موردنظر باشد.

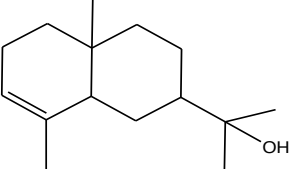
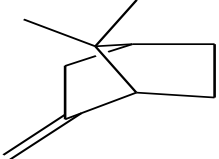
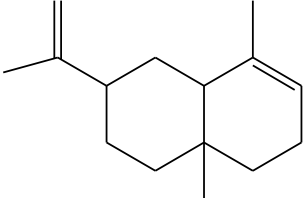
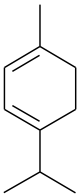
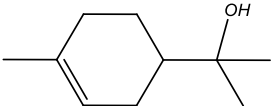
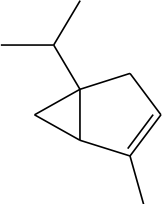
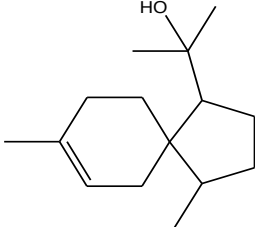
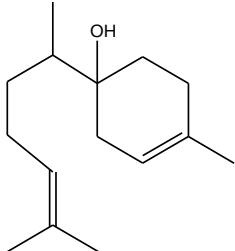
جدول ۳-۱- سری داده‌های به کار رفته در مدل‌سازی

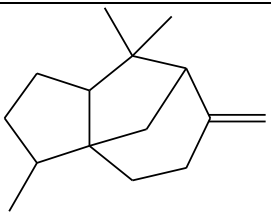
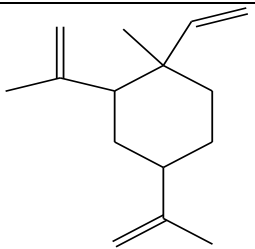
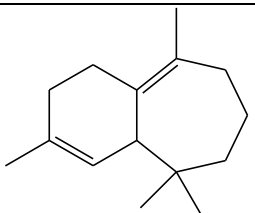
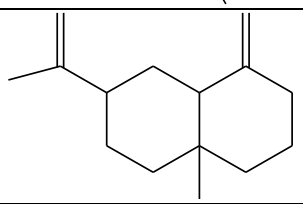
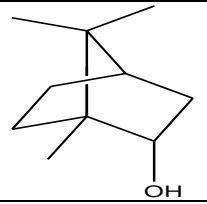
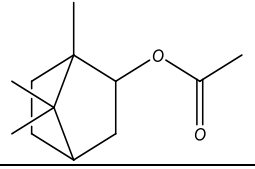
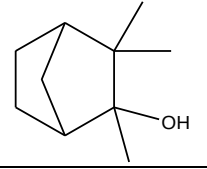
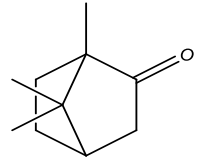
سری*	شاخص بازداری	ساختار ترکیب	نام ترکیب	شماره ترکیب
tr	۱۰۳۳		1-8-Cineole	۱
tr	۱۶۵۹		7-epi-α-Eudesmol	۲

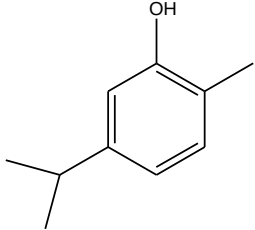
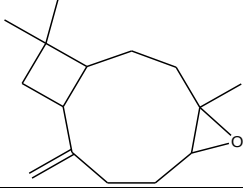
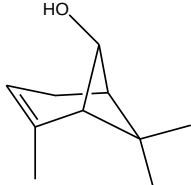
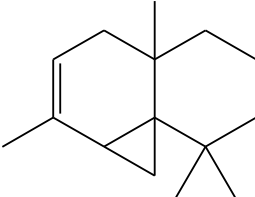
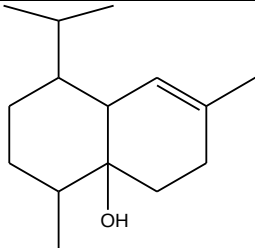
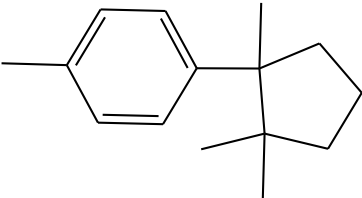
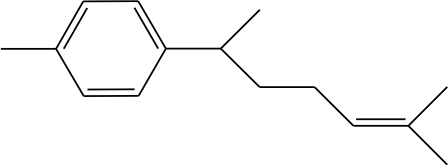
۱- Ion-Trap

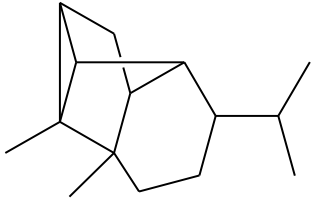
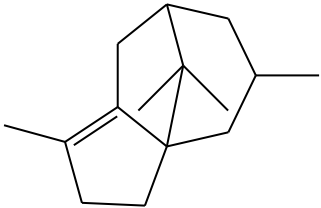
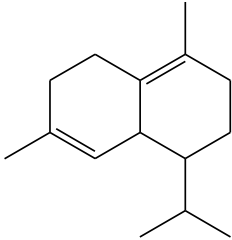
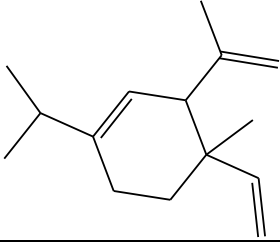
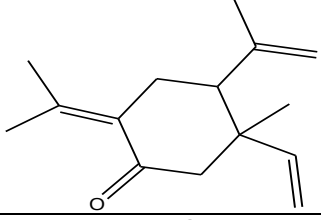
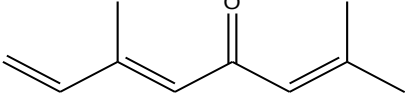
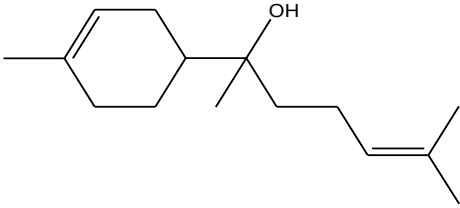
۲- Fused-Silica

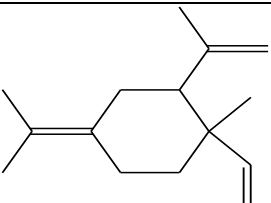
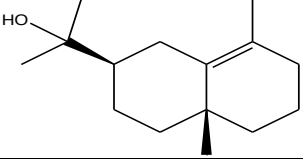
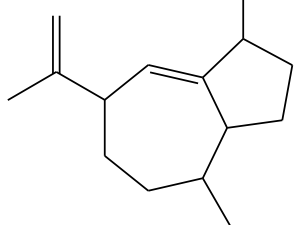
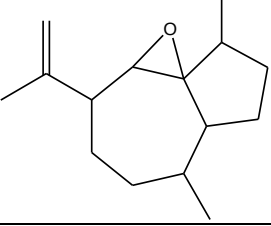
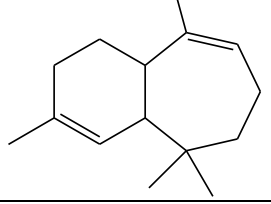
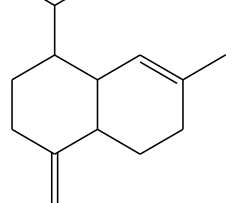
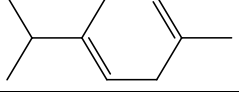
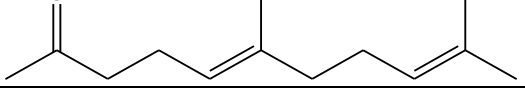
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۳	9-epi- α -Caryophyllene		۱۴۶۵	t
۴	10-epi- γ -Eudesmol		۱۶۱۸	v
۵	Acorenone		۱۶۸۹	tr
۶	α -Bisabolol oxid B		۱۶۵۵	tr
۷	α -Bisabolol		۱۶۸۴	v
۸	α -Calacorene		۱۵۴۳	t
۹	α -Cedrene		۱۴۰۸	tr
۱۰	α -Copaene		۱۳۷۵	tr

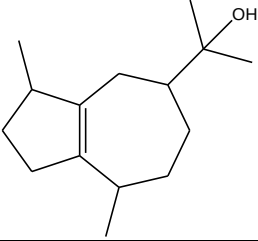
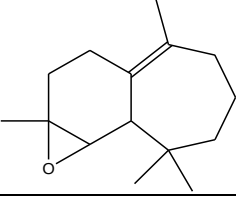
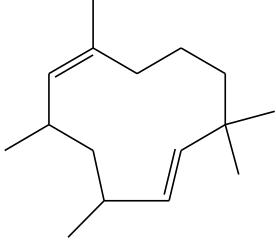
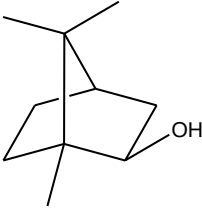
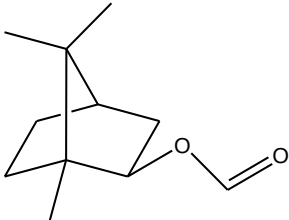
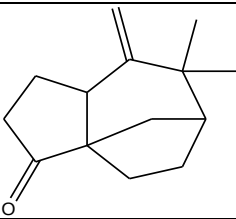
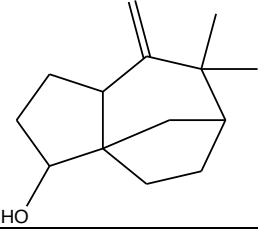
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۱۱	α -Eudesmol		۱۶۵۲	tr
۱۲	α -Fenchene		۹۵۱	tr
۱۳	α -Selinene		۱۴۹۳	t
۱۴	α -Terpinene		۱۰۱۸	tr
۱۵	α -Terpineol		۱۱۸۹	v
۱۶	α -Thujene		۹۳۱	t
۱۷	β -Acorenol		۱۶۳۴	v
۱۸	β -Bisabolol		۱۶۷۰	tr

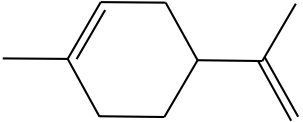
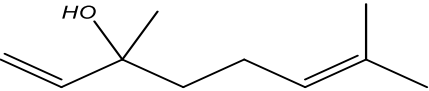
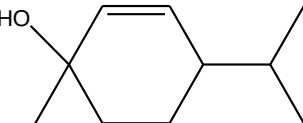
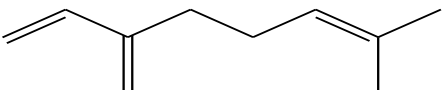
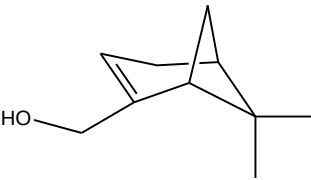
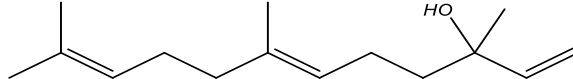
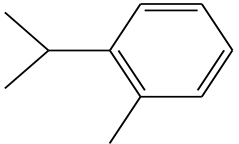
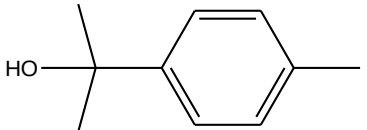
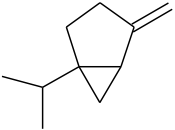
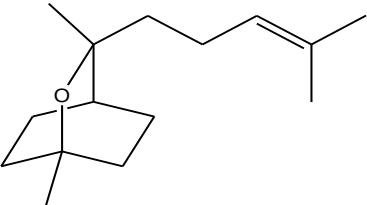
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۱۹	β -Cedrene		۱۴۲۰	tr
۲۰	β -Elemene		۱۳۹۲	tr
۲۱	β -himachalene		۱۵۰۰	v
۲۲	β -Selinene		۱۴۸۶	tr
۲۳	Borneol		۱۱۶۴	t
۲۴	Bornyl acetate		۱۲۸۵	tr
۲۵	Camphene hydrate		۱۱۴۶	tr
۲۶	Camphor		۱۱۴۲	tr

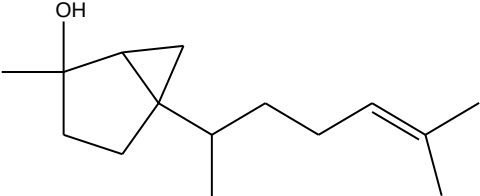
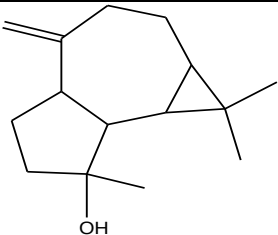
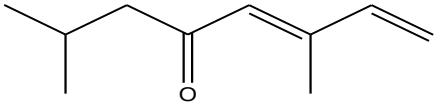
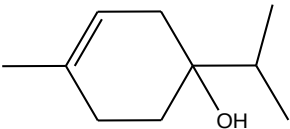
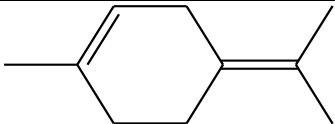
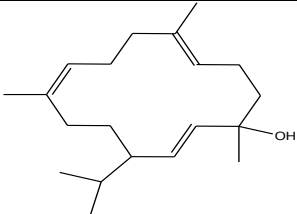
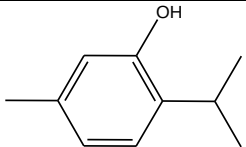
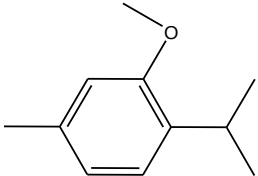
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۲۷	Carvacrol		۱۳۰۰	v
۲۸	Caryophyllen oxide		۱۵۸۳	tr
۲۹	cis-Crysanthenol		۱۱۶۱	t
۳۰	cis-Thujopsene		۱۴۲۴	tr
۳۱	Cubenol		۱۶۴۲	t
۳۲	Cuparene		۱۵۰۵	tr
۳۳	Curcumene		۱۴۸۳	tr

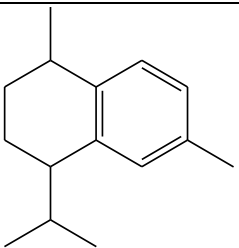
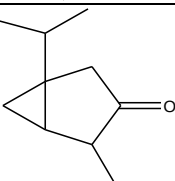
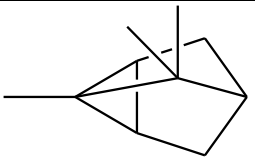
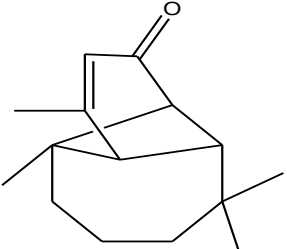
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۳۴	Cyclosativene		۱۳۶۷	v
۳۵	Cyperene		۱۳۹۹	tr
۳۶	δ -Cadinene		۱۵۲۴	tr
۳۷	δ -Elemene		۱۳۳۷	v
۳۸	Elemenone		۱۵۹۷	tr
۳۹	E-Ocimenone		۱۲۳۴	t
۴۰	epi- α -Biabolol		۱۶۸۵	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۴۱	γ -Elemene		۱۴۳۰	tr
۴۲	γ -Eudesmol		۱۶۲۵	t
۴۳	γ -Gurjunene		۱۴۷۰	tr
۴۴	γ -Gurjunenepoxide		۱۶۷۳	tr
۴۵	γ -Himachlene		۱۴۷۴	v
۴۶	γ -Muurolene		۱۴۷۷	tr
۴۷	γ -Terpinene		۱۰۶۲	tr
۴۸	Geranyl acetone		۱۴۵۲	v

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۴۹	Guaiol		۱۶۰۳	t
۵۰	Himachalene oxide		۱۶۰۹	tr
۵۱	Humulane-1,6-dien-3-ol		۱۶۰۶	v
۵۲	Isoborneol		۱۱۵۵	tr
۵۳	Isobornyl formate		۱۲۲۷	tr
۵۴	Khusimone		۱۵۹۲	tr
۵۵	Khusinol		۱۶۷۶	t

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۵۶	Limonene		۱۰۳۱	tr
۵۷	Linalool		۱۰۹۹	v
۵۸	Menth-2-en-1-ol		۱۱۲۰	tr
۵۹	Myrcene		۹۹۱	tr
۶۰	Myrtenol		۱۱۹۵	t
۶۱	Nerolidol		۱۵۶۴	tr
۶۲	o-Cymene		۱۰۲۲	tr
۶۳	p-Cymene-8-ol		۱۱۸۳	v
۶۴	Sabinene		۹۷۶	t
۶۵	Sesquicineole		۱۵۱۴	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۶۶	Sesquisabinene hydrate		۱۵۷۹	v
۶۷	Spathulenol		۱۵۷۷	tr
۶۸	Tagetone		۱۱۵۲	tr
۶۹	Terpinen-4-ol		۱۱۷۵	tr
۷۰	Terpinolene		۱۰۸۷	t
۷۱	Thunbergol		۱۸۲۵	tr
۷۲	Thymol		۱۲۹۱	t
۷۳	Thymol methyl ether		۱۲۳۰	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۷۴	Trans-Calamenene		۱۵۳۴	tr
۷۵	trans-Thujone		۱۱۱۵	v
۷۶	Tricyclene		۹۲۶	tr
۷۷	Vulgarone-B		۱۶۴۷	tr

*v:validation series, t:test series, tr:train series.

۳-۲- رسم و بهینه‌سازی ساختمان هندسی و به دست آوردن توصیف‌کننده‌ها

در این مرحله از مطالعه، ابتدا ساختار مولکول‌ها به وسیله نرم افزار HyperChem رسم و ساختار هندسی آنها با احتساب اتم‌های هیدروژن و با استفاده از روش AM1 بهینه شد. بهینه‌سازی تا زمانی ادامه یافت که گرادیان به 0.001 کیلو کالری بر مول برسد. سپس ساختارهای بهینه شده به عنوان ورودی به نرم‌افزار Dragon وارد شدند. محاسبه توصیف‌کننده‌ها توسط این نرم‌افزار صورت گرفت و حدود ۱۴۸۱ توصیف‌کننده از هیجده گروه مختلف برای هر ترکیب محاسبه گردید.

۳-۳- انتخاب بهترین توصیف‌کننده‌ها و ایجاد مدل نهایی با روش رگرسیون

خطی چندگانه

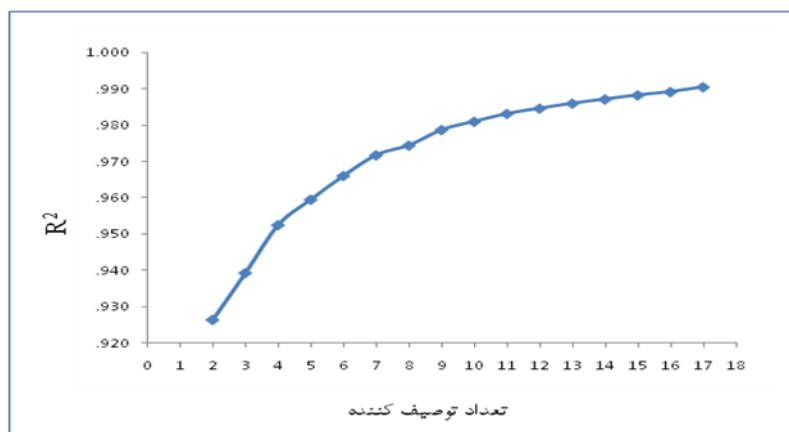
توصیف‌کننده‌های به دست آمده از مرحله قبل، به عنوان متغیر مستقل، و اندیس بازداري (RI) به عنوان متغیر وابسته وارد نرم افزار SPSS شدند. سپس به منظور جلوگیری از طولانی شدن فرایند برازش و مدل‌سازی و برای انتخاب توصیف‌کننده‌های مهم، روش‌های کاهش متغیر به کار گرفته شد. برای نیل به این هدف، تعدادی از توصیف‌کننده‌ها 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 (بیش از ۰/۹) بودند؛ پارامتری که همبستگی کمتری با پارامتر وابسته موردنظر داشت حذف گردید. برای این کار برنامه‌ای در نرم‌افزار MATLAB نوشته و اجرا شد. در نهایت با استفاده از روش برازش مرحله‌ای در انتخاب متغیر، تعدادی از توصیف‌کننده‌ها که بیشترین همبستگی را با پارامتر وابسته داشتند، انتخاب گردیدند. این مراحل به این صورت انجام گرفت که در منوی آنالیز در نرم‌افزار SPSS، گزینه رگرسیون خطی انتخاب شد و بعد از وارد کردن متغیرهای مستقل و متغیر وابسته، روش مرحله‌ای (stepwise) برای مدل‌سازی استفاده شد. چندین مدل مختلف به دست آمد که در بین مدل‌های ایجاد شده مدلی که تعداد توصیف‌کننده‌های کمتر و ضریب تعیین (R^2) و آماره‌ی F بیشتری داشت، به عنوان بهترین مدل انتخاب شد. در جدول ۲-۳ توصیف‌کننده‌های وارد شده در مدل و همچنین پارامترهای آماری مربوطه آورده شده است.

جدول ۳-۲- مقایسه آماره‌های چند مدل به دست آمده از روش MLR

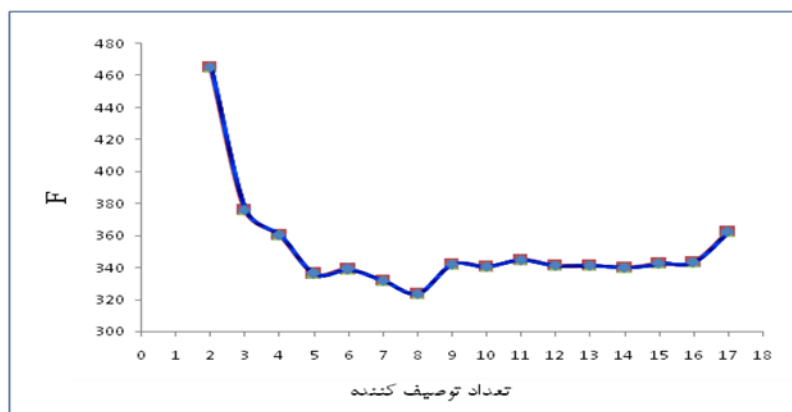
مدل	توصیف‌کننده‌ها	R^2	F	SE
۲	Dz, Mor29p	۰/۹۲۶	۴۶۵/۲۱۶	۰/۶۵۰
۳	Dz, Mor29p, Whetp	۰/۹۳۹	۳۷۶/۲۹۵	۰/۵۹۴
۴	Dz, Mor29p, Whetp, Mor28e	۰/۹۵۲	۳۶۰/۷۱۴	۰/۵۲۹
۵	Dz, Mor29p, Whetp, Mor28e, RDF095e	۰/۹۵۹	۳۳۶/۲۳۳	۰/۴۹۲
۶	Dz, Mor29p, Whetp, Mor28e, RDF095e, VEA2	۰/۹۶۶	۳۳۸/۸۹۸	۰/۴۵۳
۷	Dz, Mor29p, Whetp, Mor28e, RDF095e, VEA2, G2v	۰/۹۷۲	۳۳۱/۹۴۰	۰/۴۱۷

همان‌طور که در فصل دوم شرح داده شد، R^2 که ضریب تعیین نامیده می‌شود نشان‌دهنده‌ی این است که چه میزان از تغییرات متغیر وابسته می‌تواند توسط متغیر مستقل توضیح داده شود. آماره SE خطای استاندارد رگرسیون و آماره F مربوط به آزمون فیشر است. این پارامترها به کمک نرم‌افزار SPSS محاسبه شده‌اند.

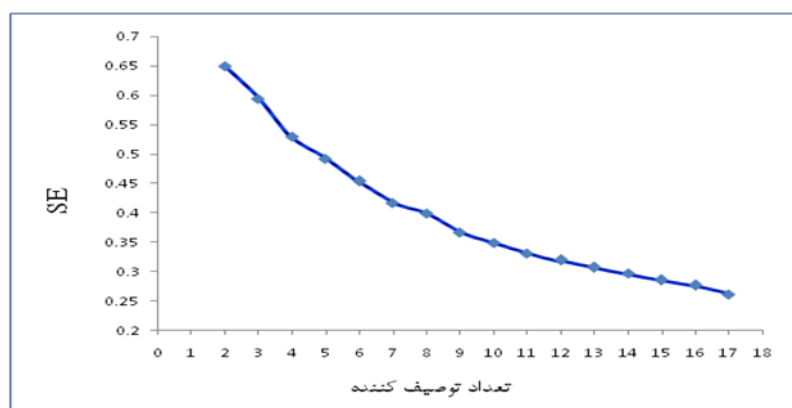
در شکل ۳-۱ تأثیر تعداد توصیف‌کننده‌ها بر مقادیر R^2 ، F و SE نشان داده شده است.



(الف)



(ب)



(ج)

شکل ۳-۱- تأثیر تعداد توصیف کننده‌ها بر مقادیر R^2 (الف)، F (ب) و SE (ج) در مدل‌های MLR

با توجه به نمودارهای داده شده، مدل ششم مدل مناسبی شناخته شد. لازم به ذکر است مدل شماره ۷ مقدار خطای کمتری دارد ولی، با توجه به اینکه میزان آماره F در مدل ۶ بیشتر بوده و اختلاف مقادیر R^2 و SE نیز در این دو مدل زیاد نیست؛ لذا مدل ششم که دارای تعداد توصیف کننده کمتری می‌باشد به عنوان مدل برتر انتخاب گردید. جدول ۳-۳ ضریب رگرسیون، گروه و نام کامل توصیف کننده‌ها و اثر متوسط آنها در مدل منتخب را نشان می‌دهد.

اثر متوسط^۱ یک متغیر مستقل با استفاده از فرمول ۳-۱ به دست می‌آید:

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

۱- MeanEffect

که در آن، β_x ضریب متغیر مستقل x در مدل MLR، X_n مقدار متغیر مستقل X مورد نظر برای ترکیب n ام و RI نیز اندیس بازداري تجربي می باشد.

جدول ۳-۳- توصیف کننده های وارد شده در مدل MLR منتخب

شماره	علامت	ضریب رگرسیون	اثر متوسط	گروه	نام کامل
۱	Dz	۹۵/۷۱۵	۱۳/۷۷۸	topological descriptors	Pogliani index
۲	Mor29p	۱۸۵/۶۹۰	۴/۵۹۸	3D-MoRSE descriptors	3D-MoRSE - signal 29 / wheighted by atomic polarizabilities
۳	Whetp	-۱/۴۶۱	-۷/۹۹۰	topological descriptors	Wiener-type index from polarizability weighted distance matrix
۴	Mor28e	۱۱۹/۳۱۸	۴/۱۶۱	3D-MoRSE descriptors	3D-MoRSE - signal 28 / wheighted by atomic Sanderson electronegativities
۵	RDF095e	۲۴/۱۶۳	۵/۱۶۹	RDF descriptors	Radial Distribution Function - 9.5 / wheighted by atomic Sanderson electronegativities
۶	VEA2	۲۶۲۴/۱۷۹	۳/۶۷۸	topological descriptors	Average eigenvector coefficient sum from adjacency matrix

جدول ۳-۴ ماتریس همبستگی توصیف کننده های وارد شده در مدل منتخب را نشان می دهد. همانطور که ملاحظه می شود مقدار همبستگی بین پارامترها قابل قبول (کمتر از ۰/۹) می باشد.

جدول ۳-۴- ماتریس همبستگی توصیف کننده های به کار رفته در مدل نهایی

توصیف کننده	Dz	Mor29p	Whetp	Mor28e	RDF095e	VEA2
Dz	۱					
Mor29p	-۰/۲۷۱	۱				
Whetp	۰/۸۴۸	۰/۲۱۹	۱			
Mor28e	-۰/۲۰۳	۰/۰۳۳	-۰/۱۲۲	۱		
RDF095e	۰/۳۹۳	۰/۱۲۶	۰/۵۸۴	۰/۱۰۱	۱	
VEA2	۰/۸۵۸	۰/۲۶۶	-۰/۸۰۲	۰/۲۳۲	-۰/۴۵۲	۱

به این ترتیب معادله خطی نهایی به صورت زیر به دست آمد:

$$RI_{cal.} = -1453/123 + 95/715Dz + 185/690Mor29p - 1/461Whetp + 119/318Mor28e + 24/163RDF095e + 2624/179VEA2 \quad (\text{معادله ۲-۳})$$

۳-۴- مدل سازی با استفاده از شبکه عصبی مصنوعی (ANN)

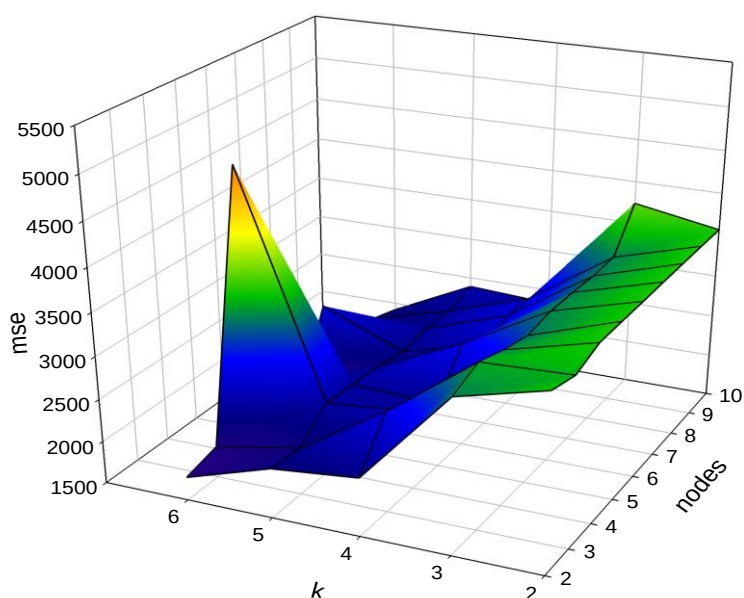
با توجه به احتمال وجود روابط غیرخطی بین متغیرهای مستقل و متغیر وابسته و از آنجا که ممکن است مدل خطی بهترین مدل برای توصیف رفتار سری داده‌ها نباشد، علاوه بر مدل سازی به روش رگرسیون خطی چندگانه، شبکه عصبی مصنوعی نیز برای مدل سازی غیرخطی داده‌ها مورد استفاده قرار گرفت. شبکه عصبی که در این تحقیق مورد استفاده قرار گرفت، یک شبکه پیش‌خور با تکنیک پس‌انتشار است که برنامه نویسی و اجرای آن، در محیط نرم افزار MATLAB نسخه 2008a صورت گرفت. مقدار عددی توصیف کننده‌های وارد شده در مدل خطی به عنوان ورودی و مقادیر تجربی اندیس بازداري (RI) نیز به عنوان هدف به شبکه عصبی داده شدند تا پاسخ شبکه با آنها سنجیده شود. داده‌ها به صورت تصادفی به سه سری آموزش (۴۷ ترکیب=۶۰٪ داده‌ها)، ارزیابی (۱۵ ترکیب=۲۰٪ داده‌ها) و تست (۱۵ ترکیب=۲۰٪ داده‌ها) تقسیم شدند. سری آموزش برای آموزش شبکه و سری ارزیابی برای بررسی قدرت تعمیم‌پذیری مدل به دست آمده از سری آموزش به کار گرفته شد. سری تست نیز به عنوان داده‌های خارجی برای ارزیابی نهایی شبکه در نظر گرفته شد. برای آموزش موفق و در نتیجه بهتر شدن قدرت پیش‌بینی شبکه، می‌بایست عوامل مؤثر در آموزش شبکه عصبی مصنوعی بررسی و بهینه شود. این پارامترها شامل تعداد متغیرهای ورودی شبکه (تعداد توصیف کننده‌ها= k)، نوع تابع آموزش، نوع تابع انتقال، تعداد گره‌ها در لایه پنهان (n)، پارامتر μ و تعداد دوره‌های آموزش (epoch) می‌باشد. در این روش تابع کارایی شبکه، میانگین مربع خطای استاندارد (MSE) می‌باشد.

۳-۴-۱- بهینه‌سازی

هر شبکه دارای حداقل یک لایه ورودی، یک لایه خروجی و تعدادی لایه پنهان می‌باشد. تعداد ورودی‌ها (توصیف‌کننده‌ها) با تعداد نرون‌های لایه ورودی برابر است و لایه خروجی نیز دارای یک نرون است که نشان دهنده‌ی RI متناظر با هر ترکیب می‌باشد [۵۸]. لازم به ذکر است که هیچ راهنمای مناسبی برای انتخاب تعداد لایه‌های پنهان در دست نیست و ساختار شبکه‌های عصبی، اغلب به شکل آزمون و خطا ایجاد می‌شود. با این وجود در بیشتر موارد به نظر می‌رسد که یک لایه پنهان مناسب باشد [۳۳، ۴۰]. بنابراین در این تحقیق نیز یک لایه پنهان مورد استفاده قرار گرفت، ولی تعداد (گره) نرون‌های لایه پنهان باید بهینه شود. برای بهینه‌سازی و شناسایی بهترین پارامترهای شبکه، شبکه‌هایی با ورودی‌های از ۲ تا ۶ توصیف‌کننده ایجاد شدند که این ورودی‌ها، همان توصیف‌کننده‌های به دست آمده از روش MLR بودند. هر شبکه با دو الگوریتم آموزشی لونبرگ-مارکوارت (trainlm) و تنظیم بایزین (trainbr) و تعداد گره‌های از ۲ تا ۱۰ در لایه پنهان آموزش داده شد. برای به دست آوردن بهترین تابع انتقال در لایه پنهان مدل‌های شبکه عصبی، از توابع لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) و خطی (purelin)، استفاده گردید. به این ترتیب ۲۷۰ ($3 \times 9 \times 2 \times 5$) شبکه مختلف طراحی شد و در روند بهینه‌سازی شبکه از میان تمامی موارد ذکر شده، به حداقل رساندن مقدار میانگین مربعات خطای شبکه به عنوان معیار انتخاب گردید. در ابتدا، در تمام شبکه‌ها تعداد دوره‌های آموزشی یکسان و برابر ۲۰ در نظر گرفته شد و از تابع انتقال خطی (purelin) در لایه خروجی استفاده گردید. در ادامه، جداول ۳-۵ تا ۳-۱۰ میزان میانگین مربعات خطای شبکه را برای ترکیب‌های مختلف توابع انتقال و الگوریتم‌های آموزشی نشان می‌دهند. نمودارهای سه بعدی نیز برای به دست آوردن دامنه‌ی خطاها و شکل گرافیکی گسترش خطاها سودمند می‌باشند. به همین دلیل در شکل‌های ۳-۲ تا ۳-۷ داده‌های جداول مذکور به صورت سه بعدی نشان داده شده‌اند.

جدول ۳-۵- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

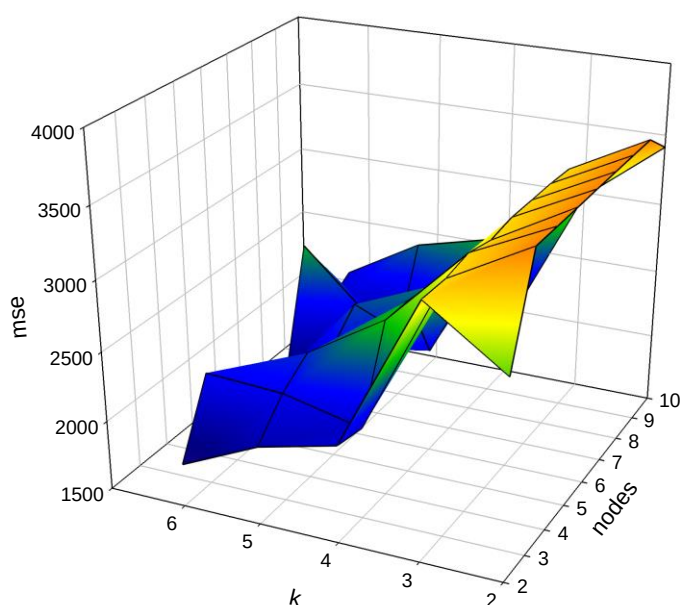
		تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k					
		۲	۳	۴	۵	۶	
تعداد گره‌ها	۲	۳۵۴۰	۳۲۸۶	۲۱۸۳	۲۰۸۳	۱۷۷۲	
	۳	۳۴۵۴	۳۴۱۵	۲۶۷۷	۲۰۵۱	۱۸۶۸	
	۴	۳۵۶۵	۳۳۱۰	۲۴۲۴	۲۳۲۹	۴۸۸۰	
	۵	۳۵۶۶	۳۲۳۶	۲۲۴۵	۲۳۰۷	۲۱۳۰	
	۶	۳۵۶۵	۳۲۸۹	۲۷۳۶	۲۲۸۴	۱۷۱۷	
	۷	۳۵۶۷	۳۲۸۵	۲۴۸۰	۲۲۲۶	۲۶۲۹	
	۸	۳۵۶۷	۳۲۶۵	۲۶۱۶	۲۲۹۲	۱۹۰۷	
	۹	۳۵۶۶	۳۲۸۰	۲۶۰۹	۲۲۸۷	۲۰۱۳	
	۱۰	۳۵۶۴	۳۷۳۹	۲۳۲۹	۲۳۸۵	۱۹۱۲	



شکل ۳-۲- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل لگاریتم سیگموئید

جدول ۳-۶- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

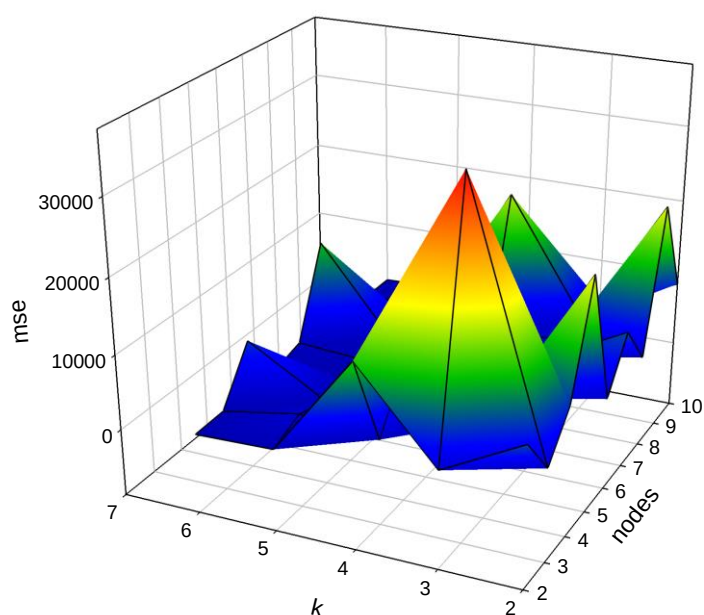
		تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k					
		۲	۳	۴	۵	۶	
تعداد گره‌ها	۲	۲۸۹۴	۳۲۷۷	۲۱۹۴	۲۰۶۰	۱۸۰۷	
	۳	۳۵۷۵	۳۲۷۳	۲۱۵۲	۲۲۸۱	۲۳۱۳	
	۴	۳۵۷۴	۳۳۱۱	۲۷۵۹	۲۴۱۹	۲۰۲۸	
	۵	۳۵۷۴	۳۲۶۶	۲۷۵۶	۲۳۰۴	۱۸۳۱	
	۶	۳۵۷۵	۳۲۹۸	۲۲۵۱	۲۴۸۸	۱۷۳۰	
	۷	۳۵۷۵	۳۲۹۶	۲۶۲۵	۲۱۹۱	۲۶۹۴	
	۸	۳۵۷۵	۳۲۹۶	۲۶۱۰	۱۹۴۳	۲۱۰۶	
	۹	۳۵۷۴	۳۲۸۵	۲۶۷۹	۲۵۴۲	۲۲۲۸	
	۱۰	۳۴۱۵	۲۹۷۶	۱۸۷۱	۲۰۶۷	۱۸۵۷	



شکل ۳-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل تانژانت سیگموئید

جدول ۳-۷- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

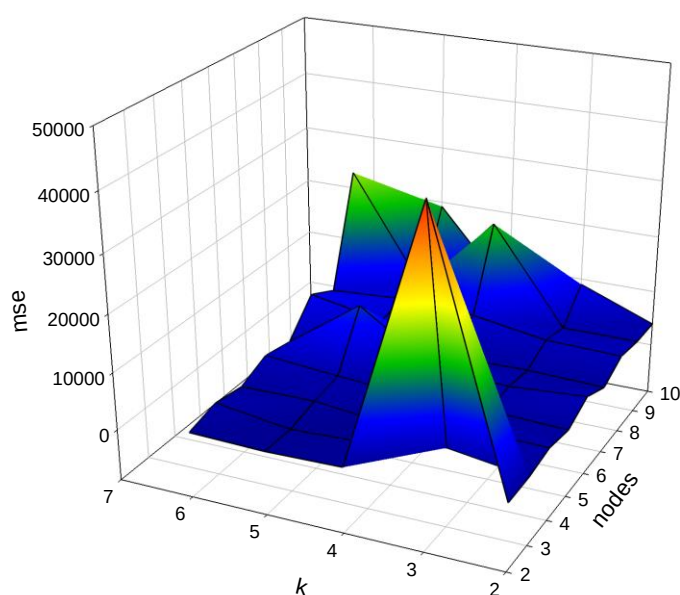
تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k						
تعداد گره‌ها		۲	۳	۴	۵	۶
	۲	۹۶۲۵	۴۰۱۶	۱۵۴۸۵	۱۹۸۲	۱۷۶۰
	۳	۳۴۴۸	۳۷۴۷۳	۲۴۷۹	۳۶۱۱	۱۷۶۵
	۴	۸۴۵۵	۴۰۲۶	۲۴۵۹	۲۴۳۸	۸۴۷۴
	۵	۲۲۲۱۱	۴۰۱۵	۲۴۵۹	۱۹۷۴	۱۷۵۹
	۶	۳۴۱۶	۴۰۱۷	۲۵۱۵	۲۰۲۲	۲۷۴۱
	۷	۹۵۵۹	۴۰۱۵	۲۴۵۷	۱۹۸۰	۱۴۲۷۵
	۸	۳۴۱۷	۴۰۲۸	۲۴۹۱	۱۹۷۲	۱۷۶۰
	۹	۲۱۴۸۵	۴۰۲۶	۱۹۸۵۱	۱۹۷۳	۱۷۶۴
	۱۰	۸۴۸۰	۴۰۱۶	۵۰۶۶	۱۹۷۹	۱۷۶۲



شکل ۳-۴- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل خطی

جدول ۳-۸- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لونیبرگ- مارکوارت (trainlm) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k		۲	۳	۴	۵	۶
تعداد گره‌ها	۲	۲۷۱۴	۴۷۱۸	۲۴۳۹	۲۱۲۷	۲۴۵۴
	۳	۲۸۳۶	۴۷۲۶	۲۵۷۰	۲۰۰۹	۳۸۲۱
	۴	۳۶۸۵	۳۹۵۷	۱۷۷۵	۳۱۷۷	۳۰۰۶
	۵	۲۸۱۵	۳۳۵۳	۳۰۹۸	۳۹۷۸	۵۰۶۰
	۶	۴۸۵۶	۳۹۲۷	۴۴۵۴	۱۳۱۷۰	۴۰۷۰
	۷	۲۷۹۱	۳۲۲۹	۲۹۶۶	۷۱۷۱	۹۵۵۷
	۸	۴۸۶۸	۵۲۰۴	۷۲۶۴	۷۹۱۰	۷۰۸۰
	۹	۴۰۶۳	۴۰۷۲	۲۰۹۴۰	۶۹۵۲	۲۶۱۴۴
	۱۰	۳۹۵۲	۹۰۱۰	۴۹۵۱	۱۹۱۸۲	۷۱۹۴

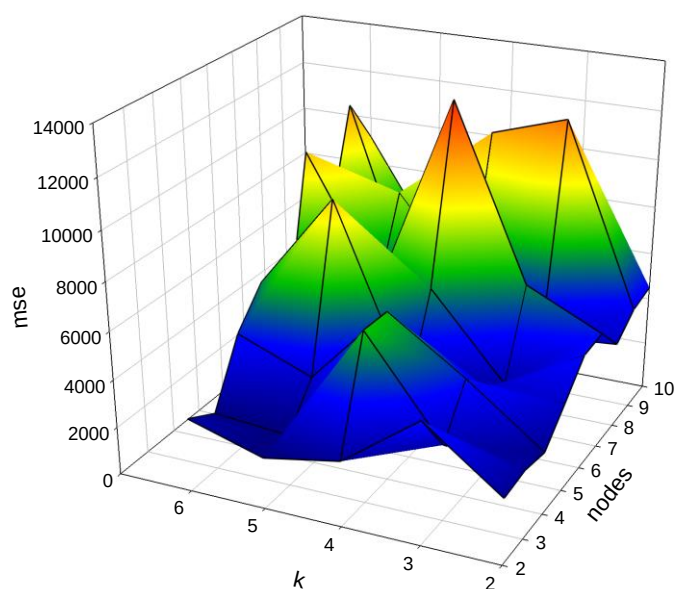


شکل ۳-۵- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونیبرگ- مارکوارت و تابع تبدیل لگاریتم

سیگموئید

جدول ۳-۹- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ - مارکوارت (trainlm) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k							
تعداد گره‌ها		۲	۳	۴	۵	۶	
	۲	۲۷۱۴	۵۰۴۷	۲۷۲۱	۲۱۳۳	۳۰۶۲	
	۳	۲۸۳۷	۳۱۴۳	۷۰۷۱	۲۱۳۰	۲۳۷۶	
	۴	۲۵۸۹	۴۹۲۸	۶۹۹۸	۳۶۸۴	۴۸۹۱	
	۵	۳۶۷۱	۳۶۷۵	۳۰۸۰	۱۰۱۸۶	۶۲۷۱	
	۶	۴۸۰۵	۲۹۴۲	۶۳۸۵	۳۶۳۴	۳۴۰۰	
	۷	۴۸۷۹	۶۳۳۹	۱۳۳۱۰	۵۱۷۲	۱۰۳۳۸	
	۸	۳۶۰۲	۳۳۳۷	۴۹۰۹	۸۴۷۳	۶۶۶۰	
	۹	۴۳۱۰	۱۱۸۱۰	۱۰۸۳۰	۲۶۴۷	۱۱۱۱۸	
	۱۰	۴۵۰۸	۴۰۳۴	۵۷۸۴	۴۵۴۶	۹۱۵۹	

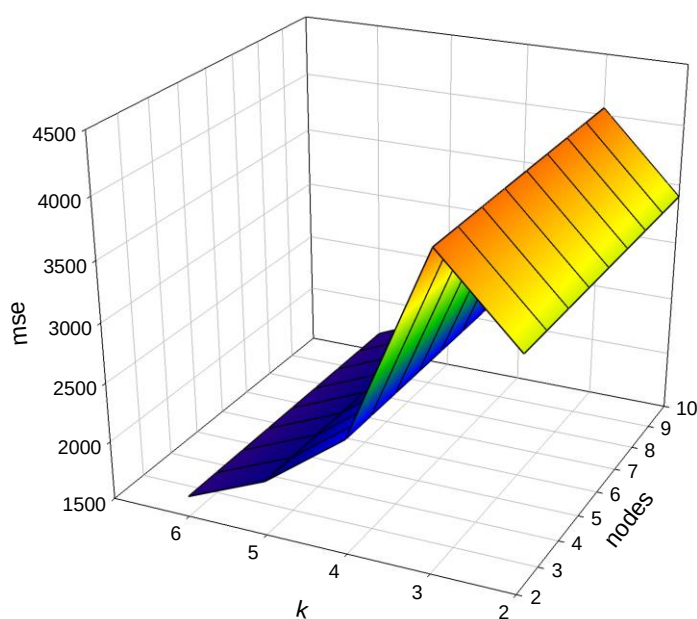


شکل ۳-۶- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لوببرگ- مارکوارت و تابع تبدیل تانژانت

سیگموئید

جدول ۳-۱۰- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لونبرگ -
مارکوارت (trainlm) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

		تعداد متغیرهای ورودی به شبکه یا توصیف‌کننده‌ها k				
		۲	۳	۴	۵	۶
تعداد گره‌ها	۲	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۳	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۴	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۵	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۶	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۷	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۸	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۹	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴
	۱۰	۳۴۰.۱	۴۰.۵۷	۲۴۵۹	۱۹۶۰	۱۶۷۴



شکل ۳-۷- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لونبرگ -مارکوارت و تابع تبدیل خطی

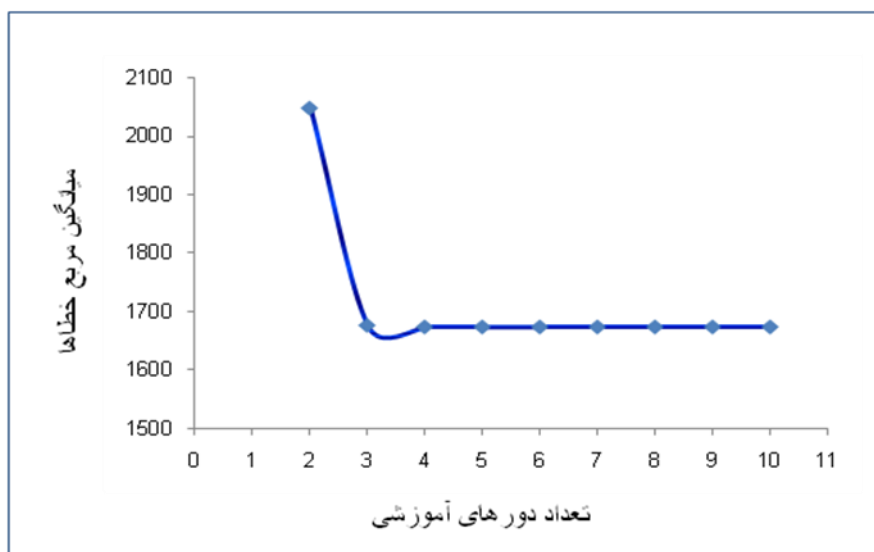
با توجه به نتایج موجود در جداول ۳-۵ تا ۳-۱۰، شبکه مورد نظر با الگوریتم آموزشی trainlm و تابع انتقال purelin، مقدار بهینه برابر با ۲ برای تعداد نرون‌ها در لایه پنهان و میزان بهینه برای تعداد متغیرهای ورودی، برابر ۶ به دست آمد.

۳-۴-۱-۱- بهینه سازی تعداد چرخه‌های آموزشی

تعداد چرخه‌های آموزشی پارامتری است که برای جلوگیری از آموزش بیش از حد شبکه بهینه‌سازی می‌شود. برای این منظور، با استفاده از مقادیر بهینه‌ی پارامترها در شبکه‌ی بهینه شده شامل تعداد نرون‌ها، تعداد متغیرهای ورودی، تابع آموزش و تابع انتقال، تعداد چرخه‌های آموزشی از ۲ تا ۷ تغییر داده شد و میزان کمترین خطا برای تعداد دور ۴ در مدل به دست آمد. منحنی تعداد چرخه‌ها بر حسب خطای مجذور میانگین شبکه در شکل ۳-۸ رسم شده است و در جدول ۳-۱۱ نیز نتایج در محدوده‌ی کمترین خطاها آورده شده است.

جدول ۳-۱۱- مقادیر خطای شبکه با تغییر تعداد چرخه‌های آموزشی

تعداد دوره‌های آموزشی	میانگین مربع خطا
۲	۲۰۴۶/۸
۳	۱۶۷۵/۵
۴	۱۶۷۴/۱
۵	۱۶۷۴/۲
۶	۱۶۷۴/۲
۷	۱۶۷۴/۲



شکل ۳-۸- منحنی مقدار مربع میانگین خطاها در برابر تغییر تعداد چرخه های آموزشی

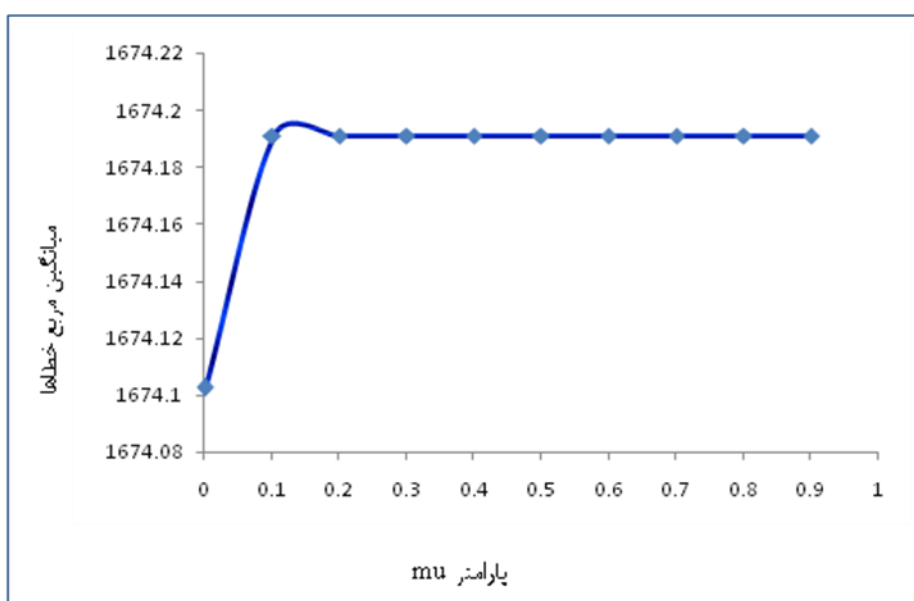
پس با توجه به نتایج سری ارزیابی، تعداد بهینه چرخه های آموزشی ۴ انتخاب گردید.

۳-۴-۱-۲- بهینه سازی پارامتر μ

برای بهینه سازی μ ، مقدار این پارامتر در شبکه بهینه شده بین ۰/۰۰۱ تا ۱/۰ با فاصله ی ۰/۱ تغییر داده شد و برای هر مورد مقدار mse سری های ارزیابی و تست محاسبه گردید. سپس منحنی خطا بر حسب μ رسم و نقطه ای که کمترین خطا را برای سری آموزش داشت به عنوان مقدار بهینه انتخاب گردید. مقدار μ نیز در این شرایط برابر با مقدار پیش فرض آن در تابع آموزشی لونیبرگ-مارکوارت یعنی ۰/۰۰۱ به دست آمد (جدول ۳-۱۱). منحنی تغییرات میزان مربع خطاها نیز بر حسب μ در شکل ۳-۹ رسم شده است.

جدول ۳-۱۲ - مقادیر خطای شبکه با تغییر پارامتر μ

میانگین مربع خطا	مقدار پارامتر μ
۱۶۷۴/۱	۰/۰۰۱
۱۶۷۴/۲	۰/۱۰۱
۱۶۷۴/۲	۰/۲۰۱
۱۶۷۴/۲	۰/۳۰۱
۱۶۷۴/۲	۰/۴۰۱
۱۶۷۴/۲	۰/۵۰۱
۱۶۷۴/۲	۰/۶۰۱
۱۶۷۴/۲	۰/۷۰۱
۱۶۷۴/۲	۰/۸۰۱
۱۶۷۴/۲	۰/۹۰۱



شکل ۳-۹ - منحنی مقادیر خطای شبکه در مقابل پارامتر μ

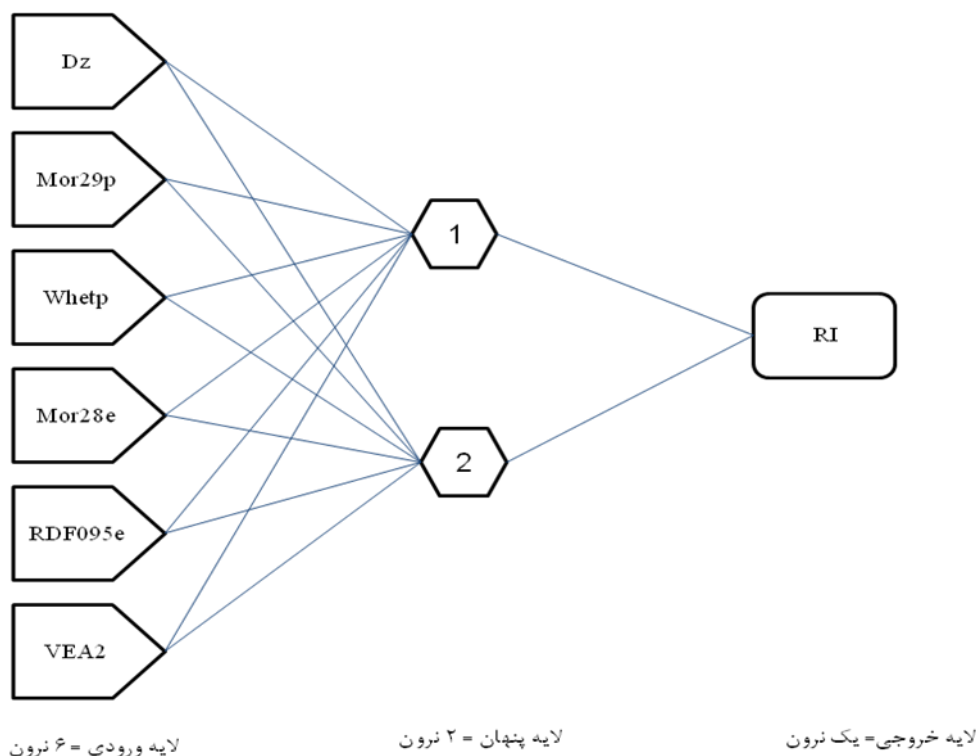
۳-۴-۱-۳- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، مشخصات شبکه بهینه در جدول ۳-۱۳ ارائه شده است.

جدول ۳-۱۳- مشخصات شبکه بهینه

تابع آموزش	trainlm
تابع انتقال لایه پنهان	purelin
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۲
تعداد متغیرهای ورودی	۶
تعداد چرخه‌های آموزشی	۴
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.



شکل ۳-۱۰- تصویر شماتیک ساختار شبکه عصبی مصنوعی به دست آمده پس از بهینه‌سازی

۳-۵- ارزیابی مدل‌های خطی و غیرخطی

مرحله مهمی که در مطالعات رابطه کمی ساختار- خاصیت مدنظر قرار می‌گیرد، ارزیابی صحت مدل‌های به دست آمده و قدرت پیش‌بینی آنها می باشد. زیرا همان‌طور که می‌دانیم شبکه باید قادر باشد به طور رضایت‌بخشی خواص مولکول‌هایی را که در سری آموزش وجود ندارند، پیش‌بینی نماید. به این منظور در ادامه، سه روش مختلف ارزیابی شامل ارزیابی با استفاده از سری‌های ارزیابی و تست، ارزیابی به کمک حذف مرحله‌ای تک‌تک داده‌ها و آزمون Y-تصادفی ارائه می‌شوند.

۳-۵-۱- ارزیابی مدل‌های MLR و ANN به کمک سری‌های ارزیابی و تست

در MLR داده‌ها با همان اسلوب شبکه و به‌طور تصادفی به سه سری آموزش (۴۷ ترکیب)، ارزیابی و تست (هر یک ۱۵ ترکیب) تقسیم‌بندی شدند. با قرار دادن مقادیر عددی توصیف‌کننده‌های مربوط به هر سری در معادله ۳-۲، می‌توان مقادیر پیش‌بینی شده را محاسبه و با مقادیر تجربی مقایسه کرد. قدرت پیش‌بینی شبکه بهینه نیز به کمک ۱۵ ترکیب که در آموزش به کار گرفته نشدند (سری تست) مورد بررسی قرار گرفت. پس در مدل خطی با استفاده از ضرایب مدل برتر سری آموزش و در مدل غیرخطی با استفاده از شبکه عصبی بهینه شده، مقادیر RI این ترکیبات پیش‌بینی شد.

در جداول ۳-۱۴ تا ۳-۱۷ مقادیر محاسبه شده، مقادیر تجربی و درصد خطای نسبی به ترتیب برای دو سری ارزیابی و تست آورده شده‌اند. درصد خطای نسبی طبق معادله ۳-۴ به دست‌آمد که در آن، E خطای مطلق بوده و از معادله ۳-۳ محاسبه می‌شود. همچنین y_i مقدار محاسبه شده برای پارامتر مورد نظر، y_t مقدار تجربی پارامتر و E_r درصد خطای نسبی می‌باشد.

$$E = y_i - y_t \quad \text{معادله (۳-۳)}$$

$$E_r = \frac{E}{y_t} \times 100 \quad \text{معادله (۴-۳)}$$

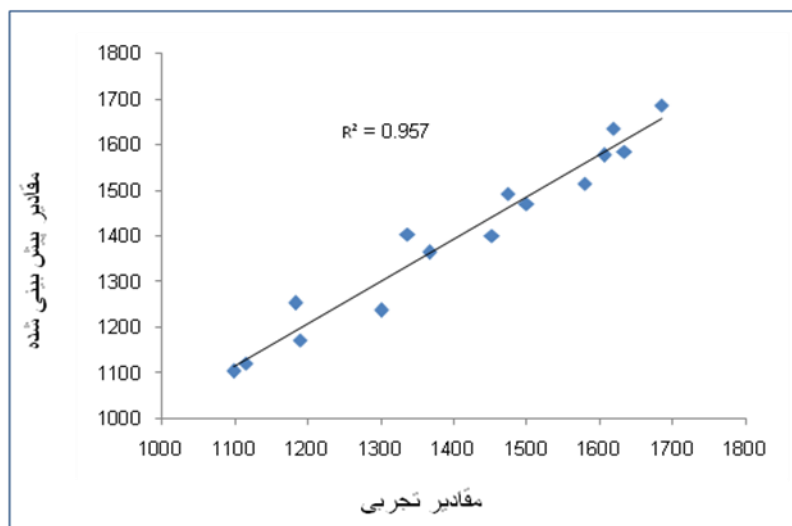
مقایسه خطاهای نسبی برای سری ارزیابی و سری تست در دو مدل، نشان‌دهنده این است که مقادیر محاسبه شده در روش ANN به مقادیر تجربی نزدیک‌تر بوده و این مدل از قدرت پیش‌بینی بالاتری برخوردار می‌باشد. نمودارهای مقادیر پیش‌بینی شده در مقابل مقادیر تجربی اندیس بازداری نیز، برای دو سری ارزیابی و تست به ترتیب در شکل‌های ۱۱-۳ تا ۱۴-۳ رسم شدند. نتایج حاصله بیان‌گر آن هستند که میزان R^2 در مدل ANN برای هر دو سری مقادیر بهتری را نسبت به مدل خطی نشان می‌دهند.

جدول ۱۴-۳ - مقایسه مقادیر تجربی و پیش‌بینی شده‌ی شاخص بازداری برای سری ارزیابی در روش MLR

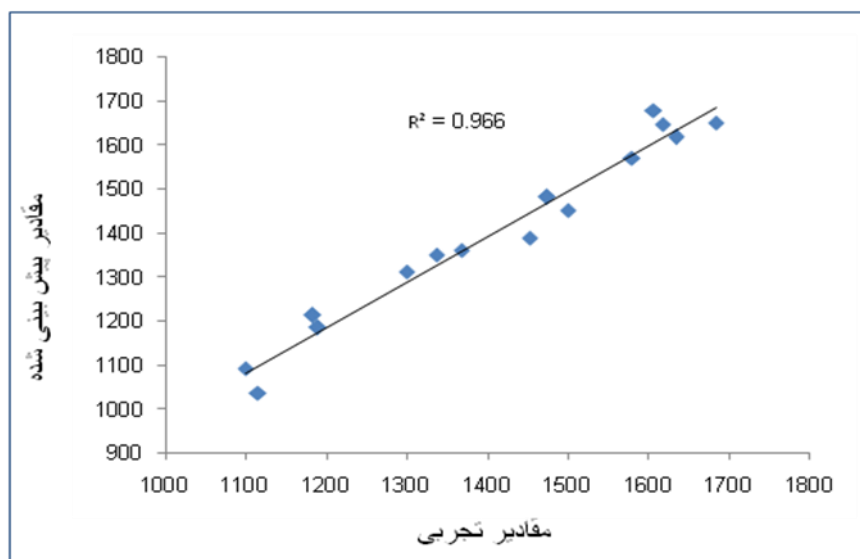
شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطای نسبی
۴	۱۶۱۸	۱۶۳۶	۱/۱
۷	۱۶۸۴	۱۶۸۷	۰/۲
۱۵	۱۱۸۹	۱۱۷۳	-۱/۳
۱۷	۱۶۳۴	۱۵۸۵	-۳/۱
۲۱	۱۵۰۰	۱۴۷۱	-۱/۹
۲۷	۱۳۰۰	۱۲۳۷	-۴/۸
۳۴	۱۳۶۷	۱۳۶۵	-۰/۲
۳۷	۱۳۳۷	۱۴۰۵	۵/۱
۴۵	۱۴۷۴	۱۴۹۴	۱/۴
۴۸	۱۴۵۲	۱۴۰۲	-۳/۵
۵۱	۱۶۰۶	۱۵۷۷	-۱/۸
۵۷	۱۰۹۹	۱۱۰۴	۰/۵
۶۳	۱۱۸۳	۱۲۵۳	-۵/۶
۶۶	۱۵۷۹	۱۵۱۶	-۴/۰
۷۵	۱۱۱۵	۱۱۲۱	۰/۵

جدول ۳-۱۵- مقایسه مقادیر تجربی و پیش‌بینی شده‌ی اندیس بازداري برای سری ارزیابی در روش ANN

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطای نسبی
۴	۱۶۱۸	۱۶۴۸	۱/۹
۷	۱۶۸۴	۱۶۵۰	-۲/۰
۱۵	۱۱۸۹	۱۱۸۵	-۰/۳
۱۷	۱۶۳۴	۱۶۱۸	-۱/۰
۲۱	۱۵۰۰	۱۴۵۳	-۳/۱
۲۷	۱۳۰۰	۱۳۱۳	۱/۰
۳۴	۱۳۶۷	۱۳۶۱	-۰/۴
۳۷	۱۳۳۷	۱۳۵۰	۱/۰
۴۵	۱۴۷۴	۱۴۸۲	۰/۶
۴۸	۱۴۵۲	۱۳۹۰	-۴/۳
۵۱	۱۶۰۶	۱۶۷۸	۴/۵
۵۷	۱۰۹۹	۱۰۹۴	-۰/۵
۶۳	۱۱۸۳	۱۲۱۴	۲/۶
۶۶	۱۵۷۹	۱۵۷۰	-۰/۵
۷۵	۱۱۱۵	۱۰۳۶	-۷/۱



شکل ۳-۱۱- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری ارزیابی



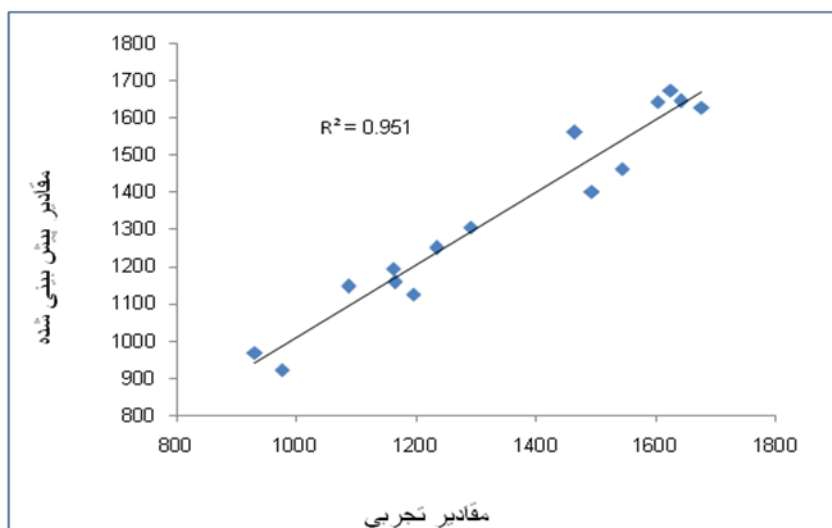
شکل ۳-۱۲- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش ANN برای سری ارزیابی

جدول ۳-۱۶ - مقایسه مقادیر تجربی و پیش‌بینی شده‌ی اندیس بازداري برای تست با روش MLR

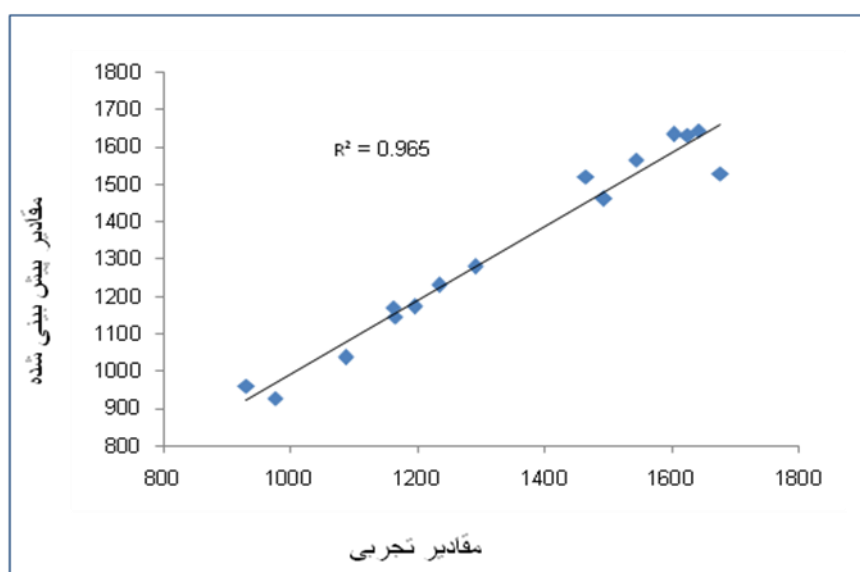
شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۳	۱۴۶۵	۱۵۶۰	۶/۵
۸	۱۵۴۳	۱۴۶۱	-۵/۳
۱۳	۱۴۹۳	۱۴۰۲	-۶/۱
۱۶	۹۳۱	۹۶۸	۴/۰
۲۳	۱۱۶۴	۱۱۵۸	-۰/۶
۲۹	۱۱۶۱	۱۱۹۳	۲/۸
۳۱	۱۶۴۲	۱۶۴۴	۰/۱
۳۹	۱۲۳۴	۱۲۵۰	۱/۲
۴۲	۱۶۲۵	۱۶۷۴	۳/۰
۴۹	۱۶۰۳	۱۶۴۳	۲/۵
۵۵	۱۶۷۶	۱۶۲۸	-۲/۹
۶۰	۱۱۹۵	۱۱۲۶	-۵/۸
۶۴	۹۷۶	۹۲۳	-۵/۴
۷۰	۱۰۸۷	۱۱۴۶	۵/۴
۷۲	۱۲۹۱	۱۳۰۵	۱/۱

جدول ۳-۱۷ - مقایسه مقادیر تجربی و پیش‌بینی شده‌ی اندیس بازداري برای تست با روش ANN

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۳	۱۴۶۵	۱۵۲۰	۳/۷
۸	۱۵۴۳	۱۵۶۵	۱/۴
۱۳	۱۴۹۳	۱۴۶۰	-۲/۲
۱۶	۹۳۱	۹۵۹	۳/۰
۲۳	۱۱۶۴	۱۱۴۴	-۱/۸
۲۹	۱۱۶۱	۱۱۷۱	۰/۸
۳۱	۱۶۴۲	۱۶۴۲	$۶/۱e^{-۵}$
۳۹	۱۲۳۴	۱۲۲۹	-۰/۴
۴۲	۱۶۲۵	۱۶۳۱	۰/۴
۴۹	۱۶۰۳	۱۶۳۳	۱/۹
۵۵	۱۶۷۶	۱۵۲۹	-۸/۸
۶۰	۱۱۹۵	۱۱۷۲	-۱/۹
۶۴	۹۷۶	۹۲۶	-۵/۱
۷۰	۱۰۸۷	۱۰۳۶	-۴/۷
۷۲	۱۲۹۱	۱۲۷۹	-۰/۹



شکل ۳-۱۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای سری تست توسط روش MLR



شکل ۳-۱۴- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای سری تست توسط روش ANN

۳-۵-۲- ارزیابی مدل‌های خطی و غیر خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

در روش حذف مرحله‌ای تک‌تک داده‌ها که روش دیگری برای ارزیابی مدل به دست آمده است، یکی از ترکیبات از سری داده‌ها کنار گذاشته شده و مدل‌سازی با ۷۶ ترکیب باقی‌مانده صورت گرفت. سپس مدل به دست آمده برای پیش‌بینی اندیس بازداری ترکیب کنارگذاشته‌شده به کار گرفته شد. این فرایند برای تمام ترکیبات در سری داده‌ها تکرار گردید که نتایج حاصله برای دو روش MLR و ANN در جداول ۳-۱۸ و ۳-۱۹ و منحنی مقادیر پیش‌بینی‌شده RI بر حسب مقادیر تجربی در شکل‌های ۳-۱۵ و ۳-۱۶ آورده شده است. نتایج این ارزیابی توانایی هر دو روش را در پیش‌بینی شاخص بازداری نشان می‌دهد ولی در عین حال نتایج مدل ANN مطلوب‌تر است. در نمودارها نیز مقادیر R^2 برای هر دو مدل داده شده است و مشاهده می‌شود شبکه نتایج بهتری ارائه داده‌است.

جدول ۳-۱۸- ارزیابی مدل خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۱۴/۲	۱۱۸۰	۱۰۳۳	۱
-۳/۰	۱۶۰۹	۱۶۵۹	۲
۴/۱	۱۵۲۴	۱۴۶۵	۳
۱/۳	۱۶۳۹	۱۶۱۸	۴
-۶/۸	۱۵۷۴	۱۶۸۹	۵
-۲/۲	۱۶۱۹	۱۶۵۵	۶
۱/۶	۱۷۱۱	۱۶۸۴	۷
۱/۲	۱۵۶۱	۱۵۴۳	۸
۳/۷	۱۴۶۰	۱۴۰۸	۹
۳/۲	۱۴۱۹	۱۳۷۵	۱۰
-۲/۷	۱۶۰۷	۱۶۵۲	۱۱
۰/۹	۹۶۰	۹۵۱	۱۲
-۲/۲	۱۴۶۱	۱۴۹۳	۱۳
-۰/۸	۱۰۱۰	۱۰۱۸	۱۴
-۰/۹	۱۱۷۸	۱۱۸۹	۱۵
۳/۲	۹۶۱	۹۳۱	۱۶

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۲/۸	۱۵۸۸	۱۶۳۴	۱۷
-۰/۷	۱۶۵۸	۱۶۷۰	۱۸
۱/۰	۱۴۳۵	۱۴۲۰	۱۹
۲/۴	۱۴۲۵	۱۳۹۲	۲۰
-۱/۶	۱۴۷۶	۱۵۰۰	۲۱
-۲/۵	۱۴۴۹	۱۴۸۶	۲۲
-۰/۷	۱۱۵۵	۱۱۶۴	۲۳
۳/۵	۱۳۳۰	۱۲۸۵	۲۴
-۱/۳	۱۱۳۱	۱۱۴۶	۲۵
۰/۷	۱۱۵۰	۱۱۴۲	۲۶
-۵/۲	۱۲۳۲	۱۳۰۰	۲۷
-۰/۸	۱۵۷۰	۱۵۸۳	۲۸
۱/۲	۱۱۷۵	۱۱۶۱	۲۹
۰/۱	۱۴۲۵	۱۴۲۴	۳۰
-۰/۱	۱۶۴۱	۱۶۴۲	۳۱
-۴/۳	۱۴۴۰	۱۵۰۵	۳۲
۳/۷	۱۵۳۸	۱۴۸۳	۳۳
۰/۱	۱۳۶۹	۱۳۶۷	۳۴
۸/۳	۱۵۱۵	۱۳۹۹	۳۵
-۲/۴	۱۴۸۷	۱۵۲۴	۳۶
۵/۵	۱۴۱۱	۱۳۳۷	۳۷
۲/۲	۱۶۳۲	۱۵۹۷	۳۸
-۰/۲	۱۲۳۲	۱۲۳۴	۳۹
۰/۷	۱۶۹۷	۱۶۸۵	۴۰
-۰/۴	۱۴۲۵	۱۴۳۰	۴۱
۰/۴	۱۶۳۱	۱۶۲۵	۴۲
-۰/۶	۱۴۶۲	۱۴۷۰	۴۳
-۴/۰	۱۶۰۶	۱۶۷۳	۴۴
۱/۷	۱۴۹۹	۱۴۷۴	۴۵
۰/۵	۱۴۸۴	۱۴۷۷	۴۶
-۷/۱	۹۸۷	۱۰۶۲	۴۷

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۳/۵	۱۴۰۲	۱۴۵۲	۴۸
۲/۲	۱۶۳۹	۱۶۰۳	۴۹
۱/۹	۱۶۳۹	۱۶۰۹	۵۰
-۱/۶	۱۵۸۰	۱۶۰۶	۵۱
-۱/۰	۱۱۴۳	۱۱۵۵	۵۲
۷/۸	۱۳۲۳	۱۲۲۷	۵۳
۰/۸	۱۶۰۵	۱۵۹۲	۵۴
-۸/۸	۱۵۲۸	۱۶۷۶	۵۵
-۲/۷	۱۰۰۳	۱۰۳۱	۵۶
-۰/۲	۱۰۹۶	۱۰۹۹	۵۷
۴/۳	۱۱۶۸	۱۱۲۰	۵۸
۲/۲	۱۰۱۳	۹۹۱	۵۹
-۱/۸	۱۱۷۳	۱۱۹۵	۶۰
-۰/۷	۱۵۵۲	۱۵۶۴	۶۱
-۱/۵	۱۰۰۷	۱۰۲۲	۶۲
۶/۴	۱۲۵۹	۱۱۸۳	۶۳
-۵/۴	۹۲۳	۹۷۶	۶۴
۰/۸	۱۵۲۶	۱۵۱۴	۶۵
-۵/۲	۱۴۹۷	۱۵۷۹	۶۶
۰/۶	۱۵۸۶	۱۵۷۷	۶۷
۱/۷	۱۱۷۲	۱۱۵۲	۶۸
۰/۵	۱۱۸۱	۱۱۷۵	۶۹
-۴/۸	۱۰۳۵	۱۰۸۷	۷۰
۶/۴	۱۹۴۲	۱۸۲۵	۷۱
-۱/۴	۱۲۷۳	۱۲۹۱	۷۲
-۰/۴	۱۲۲۵	۱۲۳۰	۷۳
۰/۶	۱۵۴۳	۱۵۳۴	۷۴
۰/۸	۱۱۲۴	۱۱۱۵	۷۵
۰/۸	۹۳۴	۹۲۶	۷۶
-۰/۱	۱۶۴۶	۱۶۴۷	۷۷

جدول ۳-۱۹- ارزیابی مدل غیر خطی به روش حذف مرحله‌ای تک تک داده‌ها

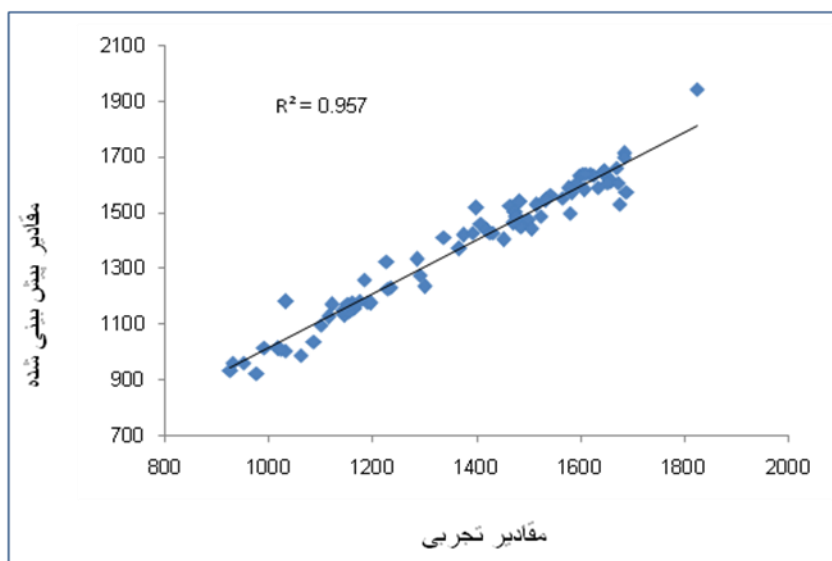
درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۴/۵	۱۰۷۹	۱۰۳۳	۱
-۳/۰	۱۶۱۰	۱۶۵۹	۲
۴/۱	۱۵۲۵	۱۴۶۵	۳
۱/۳	۱۶۳۹	۱۶۱۸	۴
-۶/۸	۱۵۷۴	۱۶۸۹	۵
-۲/۲	۱۶۱۹	۱۶۵۵	۶
۴/۰	۱۷۵۱	۱۶۸۴	۷
۱/۲	۱۵۶۱	۱۵۴۳	۸
۳/۶	۱۴۶۰	۱۴۰۸	۹
۳/۲	۱۴۱۹	۱۳۷۵	۱۰
-۲/۷	۱۶۰۷	۱۶۵۲	۱۱
۹/۲	۹۶۰	۹۵۱	۱۲
-۲/۲	۱۴۶۰	۱۴۹۳	۱۳
-۰/۸	۱۰۱۰	۱۰۱۸	۱۴
-۰/۹	۱۱۷۸	۱۱۸۹	۱۵
۳/۲	۹۶۱	۹۳۱	۱۶
-۳/۱	۱۵۸۴	۱۶۳۴	۱۷
-۰/۷	۱۶۵۸	۱۶۷۰	۱۸
۱/۰	۱۴۳۵	۱۴۲۰	۱۹
۲/۴	۱۴۲۵	۱۳۹۲	۲۰
-۱/۶	۱۴۷۶	۱۵۰۰	۲۱
-۲/۵	۱۴۴۹	۱۴۸۶	۲۲
-۰/۷	۱۱۵۵	۱۱۶۴	۲۳
۳/۵	۱۳۳۰	۱۲۸۵	۲۴
-۱/۳	۱۱۳۱	۱۱۴۶	۲۵
۰/۶	۱۱۴۹	۱۱۴۲	۲۶
-۵/۲	۱۲۳۳	۱۳۰۰	۲۷
-۰/۸	۱۵۷۰	۱۵۸۳	۲۸
۱/۲	۱۱۷۵	۱۱۶۱	۲۹

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۰	۱۴۲۵	۱۴۲۴	۳۰
-۰/۱	۱۶۴۱	۱۶۴۲	۳۱
-۴/۳	۱۴۴۰	۱۵۰۵	۳۲
۳/۷	۱۵۳۸	۱۴۸۳	۳۳
۱/۵	۱۳۶۹	۱۳۶۷	۳۴
۸/۳	۱۵۱۵	۱۳۹۹	۳۵
-۲/۴	۱۴۸۷	۱۵۲۴	۳۶
۵/۵	۱۴۱۱	۱۳۳۷	۳۷
۲/۲	۱۶۳۲	۱۵۹۷	۳۸
-۰/۲	۱۲۳۲	۱۲۳۴	۳۹
۰/۷	۱۶۹۷	۱۶۸۵	۴۰
-۰/۴	۱۴۲۵	۱۴۳۰	۴۱
۰/۴	۱۶۳۱	۱۶۲۵	۴۲
-۰/۶	۱۴۶۲	۱۴۷۰	۴۳
-۴/۰	۱۶۰۶	۱۶۷۳	۴۴
۱/۶	۱۴۹۹	۱۴۷۴	۴۵
۰/۵	۱۴۸۴	۱۴۷۷	۴۶
-۷/۱	۹۸۷	۱۰۶۲	۴۷
-۳/۵	۱۴۰۲	۱۴۵۲	۴۸
۲/۲	۱۶۳۸	۱۶۰۳	۴۹
۱/۹	۱۶۳۹	۱۶۰۹	۵۰
-۱/۶	۱۵۸۰	۱۶۰۶	۵۱
-۱/۰	۱۱۴۳	۱۱۵۵	۵۲
۷/۸	۱۳۲۳	۱۲۲۷	۵۳
۰/۸	۱۶۰۵	۱۵۹۲	۵۴
-۸/۸	۱۵۲۸	۱۶۷۶	۵۵
-۲/۷	۱۰۰۳	۱۰۳۱	۵۶
-۰/۲	۱۰۹۶	۱۰۹۹	۵۷
۴/۳	۱۱۶۸	۱۱۲۰	۵۸

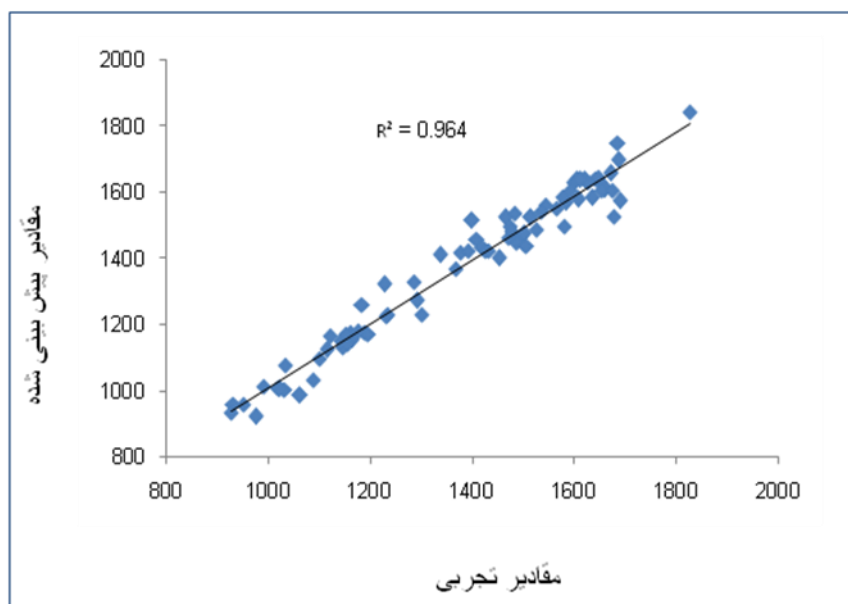
ادامه جدول ۳-۱۹

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۲/۲	۱۰۱۳	۹۹۱	۵۹
-۱/۸	۱۱۷۳	۱۱۹۵	۶۰
-۰/۷	۱۵۵۲	۱۵۶۴	۶۱
-۱/۴	۱۰۰۷	۱۰۲۲	۶۲
۶/۴	۱۲۹۵	۱۱۸۳	۶۳
-۵/۴	۹۲۳	۹۷۶	۶۴
۰/۸	۱۵۲۶	۱۵۱۴	۶۵
-۵/۲	۱۷۱۰	۱۵۷۹	۶۶
۰/۶	۱۴۹۷	۱۵۷۷	۶۷
۱/۷	۱۱۷۱	۱۱۵۲	۶۸
۰/۵	۱۱۸۰	۱۱۷۵	۶۹
-۴/۸	۱۰۳۵	۱۰۸۷	۷۰
۰/۹	۱۸۴۲	۱۸۲۵	۷۱
-۱/۴	۱۲۷۳	۱۲۹۱	۷۲
-۰/۴	۱۲۲۵	۱۲۳۰	۷۳
۰/۶	۱۵۴۳	۱۵۳۴	۷۴
۰/۸	۱۱۲۴	۱۱۱۵	۷۵
۰/۸	۹۳۴	۹۲۶	۷۶
-۰/۱	۱۶۴۶	۱۶۴۷	۷۷

در ادامه، نمودار مقادیر محاسبه شده‌ی اندیس بازداري بر حسب مقادير تجربي آن برای دو مدل در ارزیابی به روش حذف تک‌تک داده‌ها، در شکل‌های ۳-۱۶ و ۳-۱۷ آورده شده است. همانطور که مشاهده می‌شود، روش ANN قدرت پیش‌بینی بهتری را در این ارزیابی نسبت به مدل خطی نشان می‌دهد.

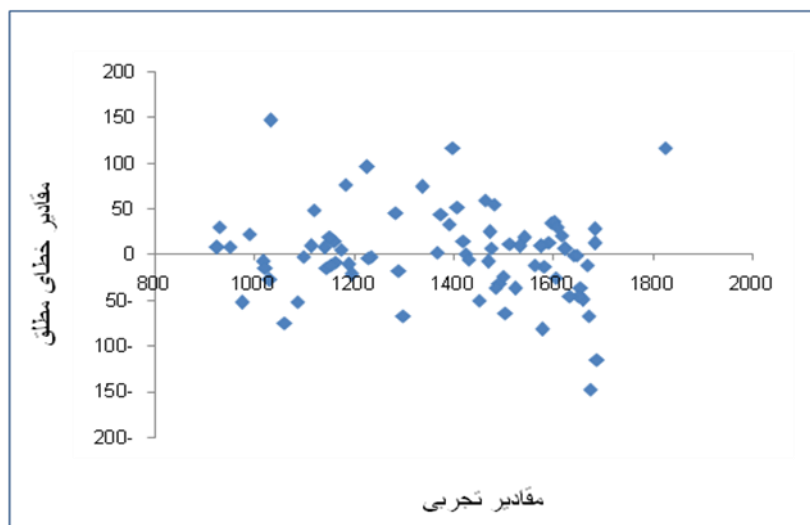


شکل ۳-۱۵- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

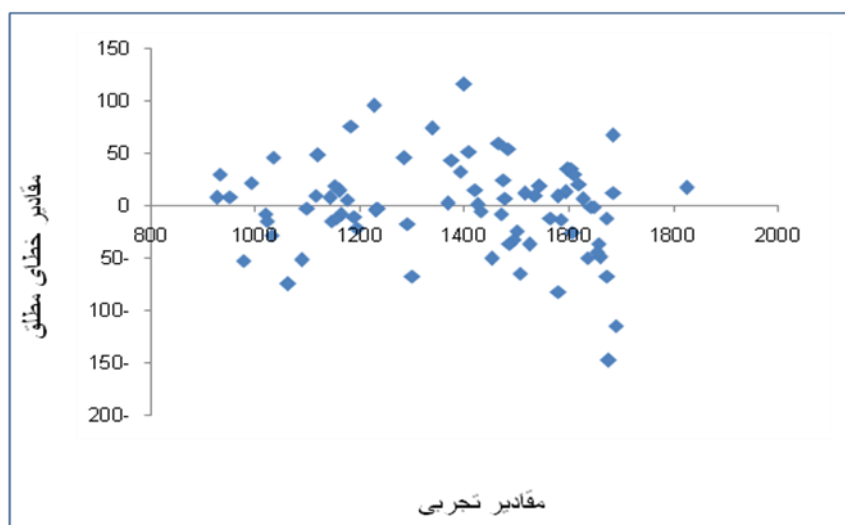


شکل ۳-۱۶- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل غیرخطی به روش حذف مرحله‌ای تک‌تک داده‌ها

همچنین نمودار مقادیر خطای مطلق محاسبه شده یا نمودار باقی مانده‌ها در ارزیابی مدل‌های MLR و ANN به روش رد مرحله‌ای تک‌تک داده‌ها برحسب مقادیر تجربی در شکل‌های ۳-۱۷ و ۳-۱۸ رسم شده است. این نمودارها از تفاضل مقدار تجربی شاخص بازداري با مقدار محاسبه شده‌ی آن، بر حسب مقدار تجربی به دست می‌آیند و معیاری برای شایستگی مدل به دست آمده هستند که در آن تقارن نسبی داده‌ها حول خط صفر نشان‌دهنده عدم وجود خطای معین می‌باشد.



شکل ۳-۱۷- رسم منحنی مقادیر باقیمانده بر حسب مقادیر تجربی اندیس بازداري در ارزیابی مدل MLR



شکل ۳-۱۸- رسم منحنی مقادیر باقیمانده در مقابل مقادیر تجربی شاخص بازداري در ارزیابی مدل ANN

۳-۵-۳- ارزیابی شبکه به روش Y-تصادفی

برای بررسی وجود ارتباطات تصادفی بین داده‌ها، آزمون Y-تصادفی با استفاده از برنامه‌ای در نرم‌افزار MATLAB نوشته و اجرا شد. در این آزمون مقادیر متغیر وابسته در محدوده‌ی همان مقادیر به طور تصادفی تغییر داده شدند و مدل‌های جدیدی به کمک این مقادیر تصادفی ایجاد شده و مقادیر اصلی متغیرهای مستقل ساخته شدند. تفاوت قابل توجه بین مقادیر ضریب تعیین در مدل اصلی و مدل‌های جدید، بیان‌گر عدم وجود ارتباط تصادفی در مدل اصلی است. نتایج حاصل از ۱۰ بار اجرای آزمون Y-تصادفی در جدول ۳-۲۰ نشان داده شده است. مقادیر پایین R^2 در این آزمون نشان‌دهنده‌ی تصادفی نبودن مدل اصلی می‌باشد.

جدول ۳-۲۰- مقادیر R^2 برای سری ارزیابی و تست در روش Y-Randomization

شماره تکرار آزمون	R^2 سری تست	R^2 سری ارزیابی
۱	۰/۲۱۲۲	۰/۰۹۲۱
۲	۰/۰۸۵۵	۰/۲۸۲۶
۳	۰/۲۶۷۱	۰/۰۵۵۵
۴	۰/۶۵۱۶	۰/۳۹۹۷
۵	۰/۰۶۵۲	۰/۰۰۷۸
۶	۰/۱۷۰۹	۰/۳۰۷۷
۷	۰/۲۹۸۵	۰/۰۳۴۰
۸	۰/۲۷۸۰	۰/۱۵۰۸
۹	۰/۴۸۶۶	۰/۱۳۸۵
۱۰	۰/۲۰۵۴	۰/۰۰۵۹

۳-۵-۴- ارزیابی مدل‌های برتر خطی و غیرخطی با استفاده از پارامترهای آماری

مطابق جدول ۳-۲۱ پنج پارامتر آماری، جهت ارزیابی توانایی پیش‌گویی مدل‌های ساخته شده به روش خطی (MLR) و غیرخطی (ANN) به کار گرفته شد. مقادیر این پارامترها نشان می‌دهد که هر

دو مدل ارائه شده از قدرت پیشگویی مناسبی برخوردار هستند ولی در عین حال نتایج حاصل از مدل غیر خطی دقیق تر بوده و این مدل توانایی پیش‌بینی بالاتری دارد.

جدول ۳-۲۱- پارامترهای آماری برای دو روش ANN و MLR

پارامتر	سری تست (n=۱۵)		سری ارزیابی (n=۱۵)		کل داده ها (n=۷۷)	
	MLR	ANN	MLR	ANN	MLR	ANN
MAE	۴۶/۳۰۰	۳۲/۸۴۷	۳۲/۷۳۳	۲۸/۵۴۰	۳۴/۷۱۳	۳۲/۶۷۸
MSE	۲۹۶۲/۳۹۵	۲۲۸۴/۹۷۴	۱۶۷۴/۳۰۵	۱۶۶۱/۳۸۶	۲۳۲۹/۲۲۲	۱۹۵۸/۶۹۱
SEP	۵۴/۴۲۸	۴۷/۸۰۱	۴۰/۹۱۸	۴۰/۷۶۰	۴۸/۲۶۲	۴۴/۲۵۷
REP(%)	۴/۰۶۵	۳/۵۷۰	۲/۹۰۴	۲/۸۹۳	۳/۴۹۲	۳/۲۰۳
R ²	۰/۹۵۱	۰/۹۶۵	۰/۹۵۷	۰/۹۶۶	۰/۹۵۸	۰/۹۶۵

۳-۶- بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اندیس بازداری

ترکیبات

در این قسمت با توجه به توصیف‌کننده‌های وارد شده در مدل‌های ANN و MLR یک بررسی اجمالی روی اثرات مختلف موجود در شاخص بازداری ترکیبات مورد مطالعه صورت خواهد گرفت. بهترین مدل انتخاب شده شامل ۶ توصیف‌کننده است که هرکدام بیان‌گر خصوصیات متفاوتی از مولکول‌های مورد بررسی است. این توصیف‌کننده‌ها عبارت بودند از:

Dz, Mor29p, Whetp, Mor28e, RDF095e, VEA2

۳-۶-۱- توصیف‌کننده‌های Dz و Whetp و VEA2

این توصیف‌کننده‌ها، جزء گروه توصیف‌کننده‌های توپولوژیکی هستند که از گراف مولکولی تهی از هیدروژن^۱ محاسبه می‌شوند. گراف مولکولی را با علامت

$$G = G(V, E) \quad \text{معادله (۳-۵)}$$

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 می‌شود. تعداد یال‌های متصل به یک رأس را درجه آن رأس می‌نامند. یک گراف مولکولی که بدون در نظر گرفتن اتم‌های هیدروژن به دست بیاید، گراف مولکولی تهی‌شده از هیدروژن نامیده می‌شود. با استفاده از نظریه گراف و با در نظر گرفتن 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 پرشاخه‌تر شدن مولکول پیچیده‌تر می‌شود [۴۳]. توصیف‌کننده Dz که به نام اندیس Pogliani نامیده می‌شود، طبق تعریفی عبارت است از مجموع تمامی اتم‌های غیر هیدروژن یک درجه رأس که به طریق زیر برآورد می‌گردد:

$$Dz = \sum_{i=1}^A \delta_i^z = \sum_{i=1}^A \frac{Z_i^v}{L_i} \quad \text{معادله (۳-۶)}$$

که در آن Z_i^v تعداد الکترون‌های ظرفیت، L_i عدد کوانتومی اصلی و δ_i^z z-delta number نام دارد [۵۹، ۶۰]. تأثیر این توصیف‌کننده بر اندیس بازداري مثبت است (۱۳/۷۷۸) بنابراین احتمالاً با افزایش درجه رأس در یک مولکول باید شاهد افزایش شاخص بازداري آن باشیم.

توصیف‌کننده Whetp یا اندیس واینر^۲، برای تعداد مشخصی از اتم‌ها در یک مولکول، در مورد ساختارهای خطی به ماکسیمم مقدار و برای ساختارهای حلقوی و پرشاخه به مینیمم اندازه‌ی خود

۱- H-depleted Molecular Graph or Hydrogen-Suppressed Molecular Graph

۲- Wiener-type Index or Wiener Index

می‌رسد [۶۱]. از آنجا که اثر این توصیف‌کننده بر شاخص بازداري منفی (۷/۹۹۰-) است انتظار می‌رود با افزایش این اندیس، شاخص بازداري کاهش یابد. مثلاً در مولکول شماره ۵۹ با نام میرسین^۱ با ساختار خطی مقدار اندیس واینر زیاد و شاخص بازداري آن به نسبت کم است (۹۹۱)؛ درحالی که در مولکول شماره ۷۰ که ترپینولن نام دارد^۲ با ساختمان حلقوی و شاخه‌دار شاخص بازداري ۱۰۸۷ می‌باشد.

توصیف‌کننده VEA2 از جمله توصیف‌کننده‌های اندیس VEA است که توسط ضرایب l_{iA} از یک بردار مشخصه تعریف می‌شود که این بردار با بزرگترین مقدار مشخصه منفی در ارتباط است؛ به عبارت دیگر، l_{iA} مقدار مشخصه A از توالی رو به کاهش این مقادیر است. VEA2 به صورت زیر محاسبه می‌گردد:

$$VEA2 = \frac{\sum_{i=1}^A l_{iA}}{A} \quad \text{معادله (۷-۳)}$$

A بیان‌گر تعداد اتم‌ها در مولکول است [۶۳، ۶۲]. میزان تأثیر این توصیف‌کننده بر شاخص بازداري ۳/۶۷۸ می‌باشد.

۳-۶-۲- توصیف‌کننده‌های Mor28e و Mor29p

جزء توصیف‌کننده‌های 3D-MoRSE به معنای نمایش سه بعدی ساختار مولکول براساس تفرق الکترون هستند. هدف توصیف‌کننده‌های 3D-MoRSE، این است که ساختار سه بعدی مولکول را به یک کد ثابت تبدیل کنند. این عمل، توسط معادله‌ای که به منظور فراهم آوردن منحنی‌های نظری تفرق در مطالعات پراش الکترونی به کار رفته انجام می‌گیرد [۶۴].

$$I(s) = \sum_{i=2}^N \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \quad \text{معادله (۸-۳)}$$

۱- Myrcene

۲- Terpinolene

در معادله ۳-۸، I شدت الکترون پراکنده شده، A_i و A_j خاصیت اتمی اتم‌های i و j، s زاویه پراکندگی، N تعداد کل اتم‌ها و r_{ij} فاصله اقلیدسی^۱ بین دو اتم i و j را نشان می‌دهد. چون این توصیف‌کننده‌ها آرایش سه‌بعدی اتم‌ها را بدون اینکه به اندازه مولکول ارتباط داشته باشند بیان می‌کنند، بنابراین برای تعداد زیادی مولکول با تفاوت‌های ساختاری زیاد قابل کاربرد هستند [۶۵]. Mor29p با توانایی قطبش پذیری اتمی وزن دار شده است و اثر آن مثبت و برابر با (۴/۵۹۸) است. Mor28e نیز توسط الکترونگاتیویته‌ی ساندerson اتمی وزن دار گردیده است و میزان تأثیر آن نیز بر شاخص بازداري مثبت و قابل توجه است (۴/۱۶۱).

۳-۶-۳- توصیف‌کننده RDF095e

توصیف‌کننده‌های گروه RDF بر پایه توزیع فاصله در نمایش هندسی یک مولکول و تشکیل یک تابع توزیع شعاعی^۲ که خصوصیات مشترکی را با کد 3D-MoRSE نشان می‌دهد، بنا شده‌اند. این توابع مربوط به یک دسته از اتم‌ها بوده و معادل احتمال یافتن یک اتم در فضای کروی به شعاع R است. معادله ۳-۹ نحوه محاسبه این توصیف‌کننده‌ها را نشان می‌دهد.

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i w_j e^{-\beta \cdot (R - r_{ij})^2} \quad \text{معادله (۳-۹)}$$

که f یک فاکتور مقیاس و A تعداد اتم‌های مولکول است. همچنین، r_{ij} فاصله بین اتمی میان اتم‌های i و j باشد و w یک ویژگی یا خاصیت اتمی نظیر الکترونگاتیوی، جرم اتمی یا قطبش پذیری مربوط به اتم‌های i و j است. β یک فاکتور تسهیل‌کننده است که احتمال توزیع یک فاصله بین اتمی منحصربه‌فرد را مشخص می‌کند و می‌توان آن را به فاکتور دما برای تعریف تحرکات اتمی تعبیر نمود. $g(R)$ در تعدادی از نقاط گسسته با فواصل معین محاسبه می‌گردد. توصیف‌کننده‌های RDF ساختار سه بُعدی مولکول را بیان می‌کنند. با داشتن ویژگی w یا خواص اتمی مربوط به اتم‌های i و j، کد

۱- Euclidean Distance

۲- Radial Distribution Function

RDF می‌تواند در موارد متفاوتی به اقتضای نیازمندی‌های اطلاعات ارائه شده به کار برده شود. این ویژگی‌های اتمی قادرند میان اتم‌های یک مولکول، برای تقریباً 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 سه بُعدی روبرو می‌شود: این تابع مستقل از تعداد اتم‌ها یا به عبارت دیگر اندازه مولکول است؛ و با توجه به آرایش سه بُعدی اتم‌ها تابعی منحصر به فرد است، ولی در برابر حرکات انتقالی و چرخش کل مولکول نامتغیر است. کد RDF می‌تواند در انواع خاصی از اتم‌ها یا بُردهایی از فاصله محدود شود تا اطلاعات مشخصی را درباره یک ساختار فضایی سه بُعدی معین به دست دهد. توصیف‌کننده‌های RDF با به کارگیری مجموعه قوانین ساده قابل تفسیرند و از این رو امکان تبدیل کد را به ساختار سه بُعدی 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 غیرمسطح و انواع اتم‌ها به دست می‌دهد [۶۶].

۳-۷- بررسی میزان مشارکت توصیف‌کننده‌های منتخب توسط شبکه عصبی

میزان مشارکت توصیف‌کننده‌های منتخب به صورت زیر تعیین شد:

⇐ وزن‌های مربوط به توصیف‌کننده موردنظر از شبکه بهینه حذف شد.

⇐ مقدار متغیر وابسته با استفاده از توصیف‌کننده‌های باقی‌مانده برای هر ترکیب سری آموزش پیش‌بینی گردید.

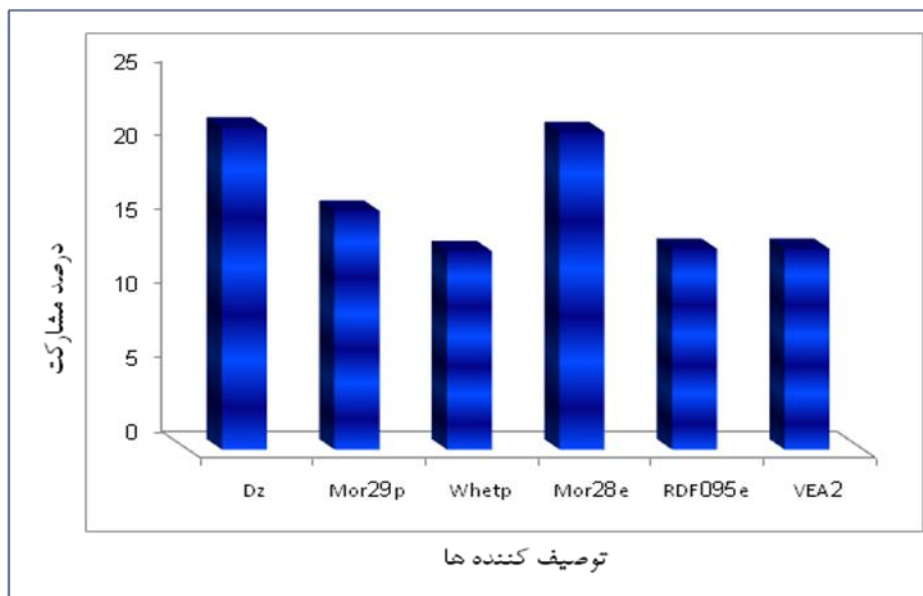
⇐ میانگین خطای مطلق (MAE) حاصل از ارزیابی تقاطعی ترکیبات سری آموزش محاسبه شد.

⇐ مراحل بالا برای دیگر توصیف‌کننده‌ها نیز تکرار شد.

⇐ سرانجام درصد مشارکت هر توصیف‌کننده توسط رابطه (۳-۱۰) محاسبه گردید [۳۳].

$$c_i = 100 \frac{\Delta m_i}{\sum_{i=1}^N \Delta m_i} \quad \text{معادله (۳-۱۰)}$$

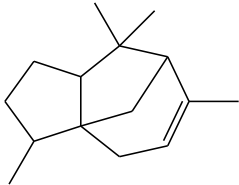
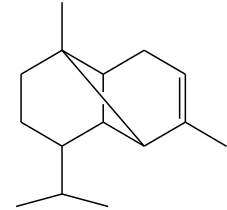
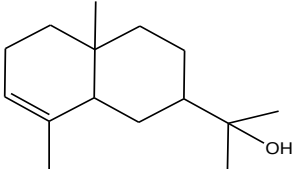
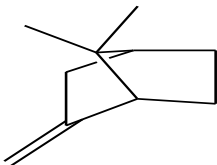
در این رابطه c_i درصد مشارکت توصیف‌کننده حذف شده Δm_i ، N تعداد توصیف‌کننده‌ها در مدل و Δm_i میانگین خطای مطلق حاصل از ارزیابی تقاطعی سری آموزش در غیاب توصیف‌کننده Δm_i را نشان می‌دهد. بنابراین درصد مشارکت توصیف‌کننده‌های منتخب در ترکیبات مورد بررسی مطابق شکل ۳-۱۹ می‌باشد.



شکل ۳-۱۹- نمایش میزان مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینه

همان‌طور که مشاهده می‌شود، درصد مشارکت توصیف‌کننده‌های Dz و Mor28e از بقیه بیشتر است. Dz یک توصیف‌کننده توپولوژیکی است که از جمله اثرات آن بر اندیس بازداري بر اساس تعداد اتم‌های موجود در مولکول می‌باشد. در جدول ۳-۲۲ نمونه‌هایی از این اثرات نشان داده شده است.

جدول ۳-۲۲- مثال‌هایی از اثر توصیف‌کننده Dz بر شاخص بازداری

RI	مقدار Dz	تعداد اتم‌ها (به جز هیدروژن)	ترکیب
۱۴۰۸	۳۰	۱۵	α -Cedrene 
۱۳۷۵	۳۰	۱۵	α -Copaene 
۱۶۵۲	۳۳	۱۶	α -Eudesmol 
۹۵۱	۲۰	۱۰	α -Fenchene 

توصیف‌کننده Mor28e نیز در برگیرنده اطلاعات سه بعدی مولکول می‌باشد که این اطلاعات با صورت‌بندی ترکیب در ارتباط است. پس می‌توان گفت که اندیس بازداری این ترکیبات وابستگی قابل ملاحظه‌ای به تغییرات صورت‌بندی آنها دارد.

۳-۸- نتیجه‌گیری

در این تحقیق از مدل‌های خطی و غیرخطی برای پیش‌بینی اندیس بازداری یک سری روغن‌های ضروری فرار استفاده شده است که بر اساس این مدل‌ها می‌توان تأثیر دسته خاصی از توصیف‌کننده‌ها را به طور ویژه روی مکانیسم بازداری مورد بررسی قرار داد که با صرف هزینه و وقت کمتر، پیش از سنتز ترکیب مورد نظر و به کمک مدل معتبر، این امر محقق می‌گردد.

فصل چهارم

مدل سازی QSPR اندیس بازداري (RI)

روغن های ضروري با استفاده از روش

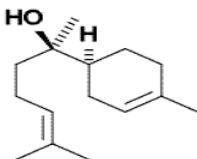
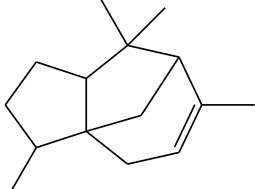
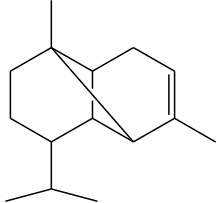
رگرسیون خطی چندگانه (MLR) و شبکه

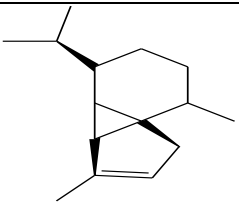
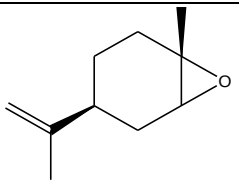
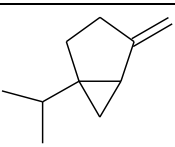
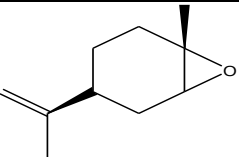
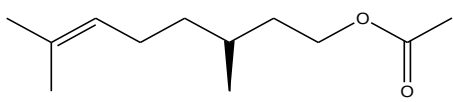
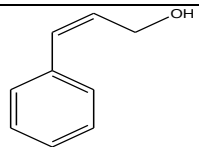
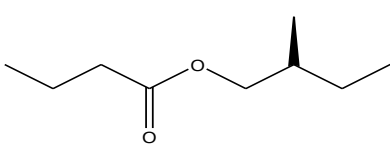
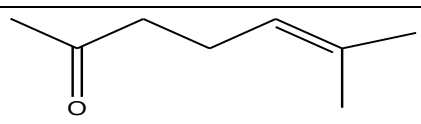
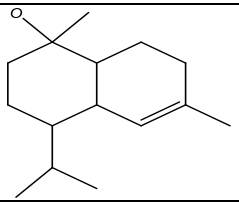
عصبی مصنوعی (ANN)

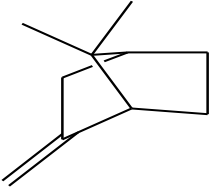
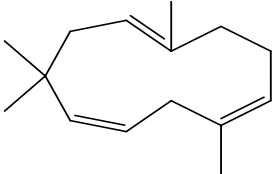
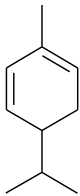
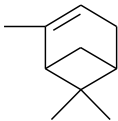
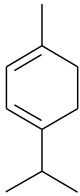
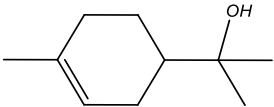
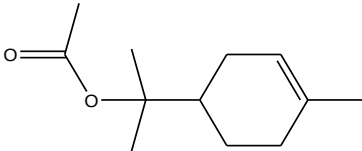
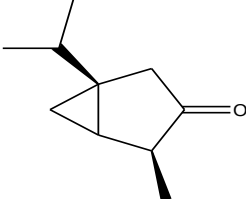
۴-۱- سری داده‌ها

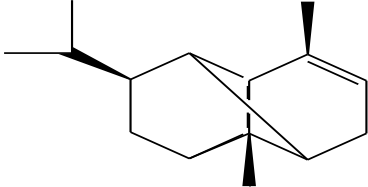
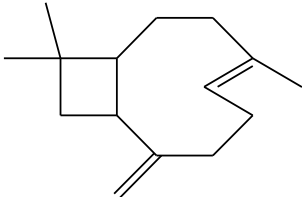
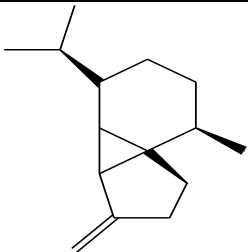
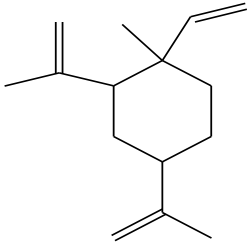
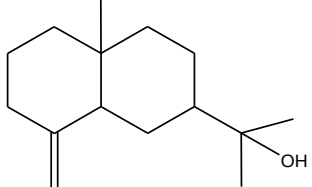
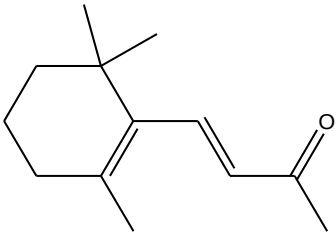
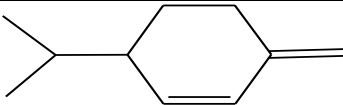
در ابتدا، شاخص بازداری (RI) ۱۱۹ روغن ضروری متنوع که دارای گروه‌های عاملی گوناگون می‌باشند، به عنوان سری داده‌ها مورد استفاده قرار گرفت. آنالیز این ترکیبات با استفاده از یک گاز-کروماتوگراف GC 17-A در اتصال با یک اسپکترومتر جرمی با ولتاژ یونیزاسیون جرمی ۷۰ eV، ولتاژ شتاب‌دهنده یون ۱۵۰۰ V و دمای مخزن یون ۲۵۰ °C انجام گرفت؛ همچنین این سیستم به یک ستون کاپیلاری سیلیکای مذاب DB-Wax (mm i. d.) ۰/۲۵ × ۶۰ m و ضخامت فیلم ۰/۲۵ μm مجهز بوده و برنامه دمایی ستون از ۷۰ تا ۲۳۰ °C با سرعت ۲ °C min⁻¹ تنظیم شد. نمونه روغن با حجم ۰/۲ μl تزریق شده و هلیوم به عنوان گاز حامل با سرعت جریان ثابت ۱/۰ ml min⁻¹ انتخاب گردید [۵]. نام، ساختمان و اندیس بازداری این ترکیبات در جدول ۴-۱ آمده است.

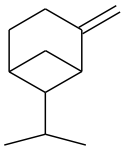
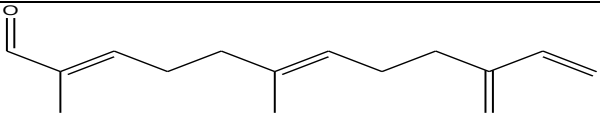
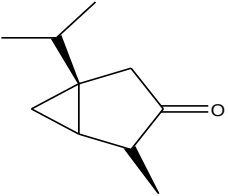
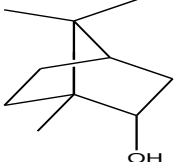
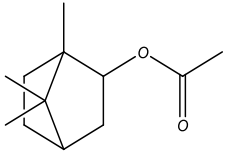
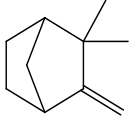
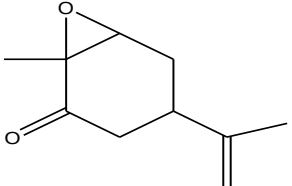
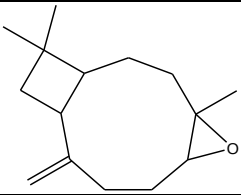
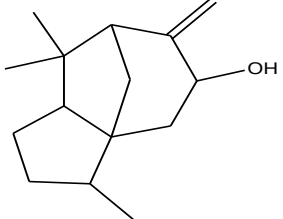
جدول ۴-۱- مجموعه داده‌های به کار گرفته شده در مدل‌سازی

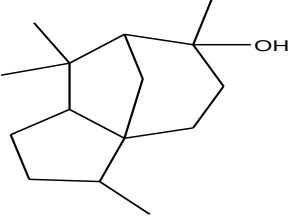
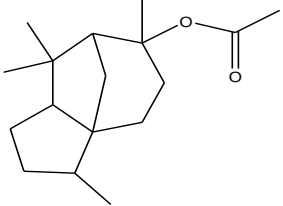
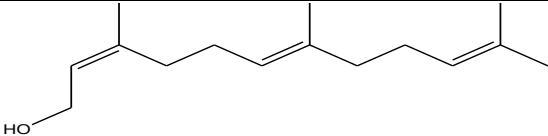
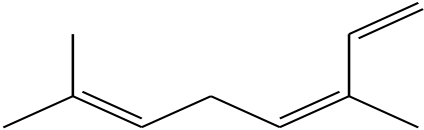
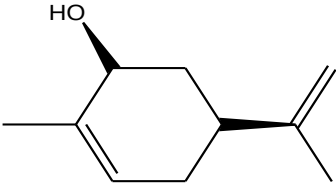
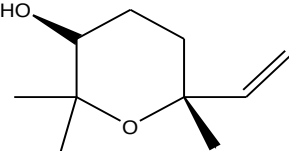
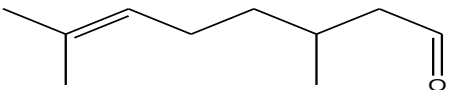
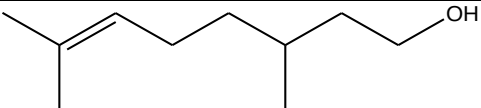
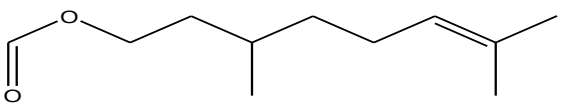
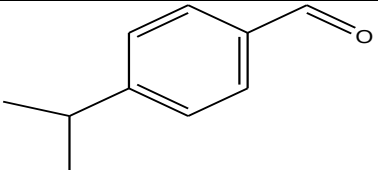
سری*	شاخص بازداری	ساختار ترکیب	نام ترکیب	شماره ترکیب
tr	۲۲۲۹		(-)-α-Bisabolol	۱
tr	۱۵۷۴		(-)-α-Cedrene	۲
v	۱۵۰۰		(-)-α-Copaene	۳

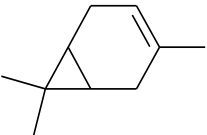
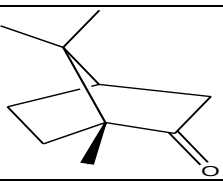
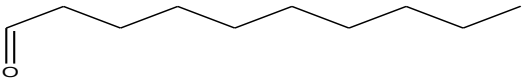
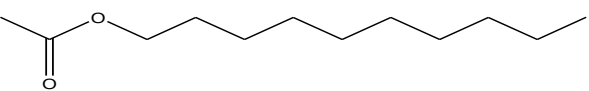
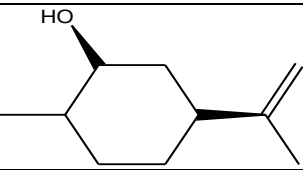
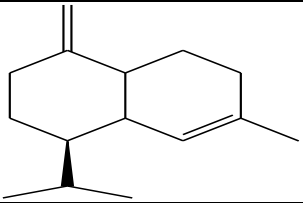
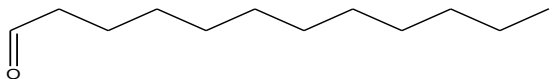
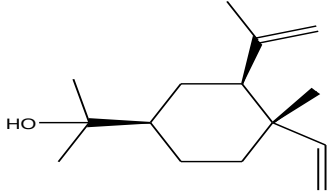
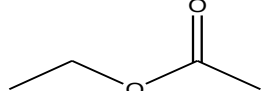
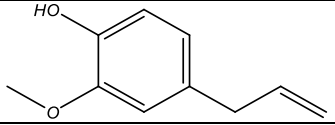
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۴	(-)- α -Cubebene		۱۴۶۶	t
۵	(+)-cis-Limonene oxide		۱۴۵۸	tr
۶	(+)-Sabinene		۱۱۳۵	tr
۷	(+)-trans-Limonene oxide		۱۴۷۰	t
۸	($\pm$)-Citronellyl acetate		۱۶۶۸	v
۹	(z)-Cinnamyl alcohol		۲۳۰۳	tr
۱۰	2-Methylbutylbutyrate		۱۲۸۷	tr
۱۱	6-Methyl-hept-5-en-2-one		۱۳۶۱	tr
۱۲	α -Cadinol		۲۲۱۹	tr

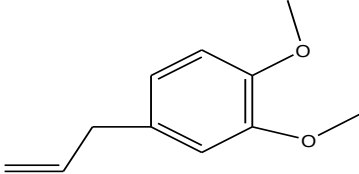
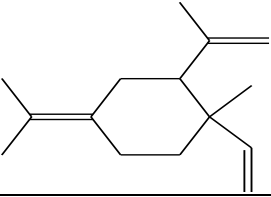
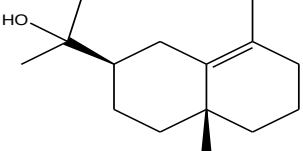
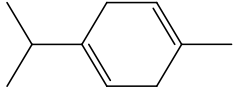
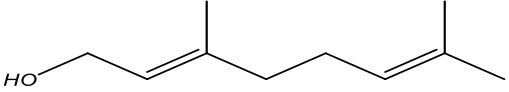
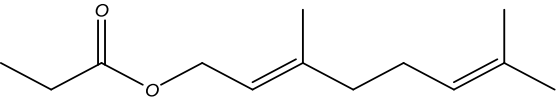
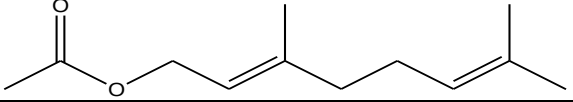
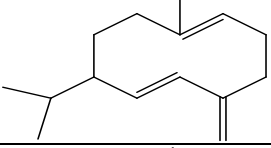
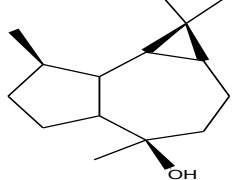
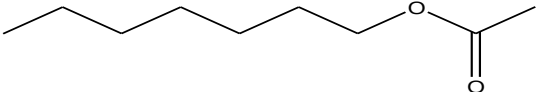
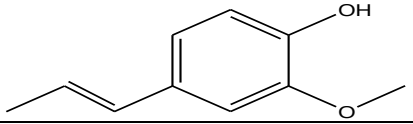
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۱۳	α -Fenchene		۱۰۷۶	v
۱۴	α -Humulene		۱۶۷۸	tr
۱۵	α -Phellandrene		۱۱۷۹	t
۱۶	α -Pinene		۱۰۳۹	v
۱۷	α -Terpinene		۱۱۹۵	t
۱۸	α -Terpineol		۱۷۰۸	tr
۱۹	α -Terpinyl acetate		۱۷۰۱	tr
۲۰	α -Thujone		۱۴۳۳	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۲۱	α -Ylangene		۱۴۹۴	t
۲۲	β -Caryophyllene		۱۶۰۶	tr
۲۳	β -Cubebene		۱۵۵۲	v
۲۴	β -Elemene		۱۵۹۶	tr
۲۵	β -Eudesmol		۲۲۴۶	tr
۲۶	β -Ionone		۱۹۵۳	tr
۲۷	β -Phellandrene		۱۲۳۳	t

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۲۸	β -Pinene		۱۱۲۶	tr
۲۹	β -Sinensal		۲۲۳۹	v
۳۰	β -Thujone		۱۴۴۹	tr
۳۱	Borneol		۱۵۴۸	v
۳۲	Bornyl acetate		۱۵۹۲	tr
۳۳	Camphene		۱۰۸۵	tr
۳۴	Carvone oxide		۱۸۱۷	t
۳۵	Caryophyllen oxide		۲۰۰۱	tr
۳۶	Cedrenol		۲۱۴۲	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۳۷	Cedrol		۲۱۱۲	t
۳۸	Cedryl acetate		۲۱۵۰	tr
۳۹	cis,trans-Farnesol		۲۳۵۲	v
۴۰	cis-β-Ocimene		۱۲۴۳	tr
۴۱	cis-Carveol		۱۸۴۵	tr
۴۲	cis-Linalool pyranoxide		۱۷۵۵	v
۴۳	Citronellal		۱۴۸۶	tr
۴۴	Citronellol		۱۷۷۲	tr
۴۵	Citronellyl formate		۱۶۲۸	t
۴۶	Cuminaldehyde		۱۷۸۵	tr

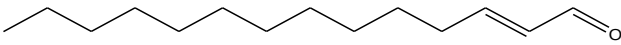
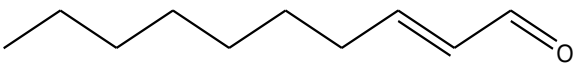
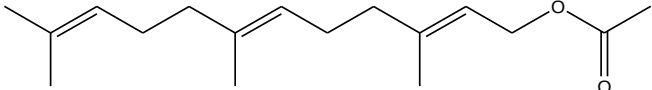
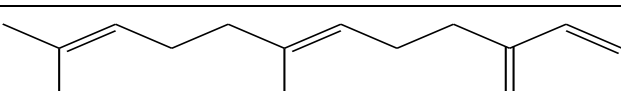
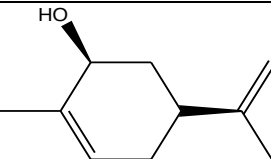
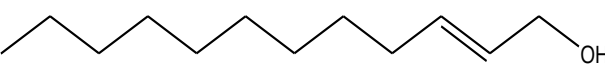
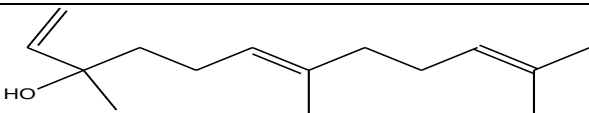
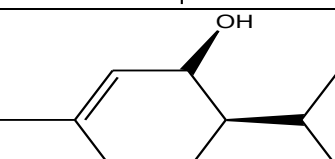
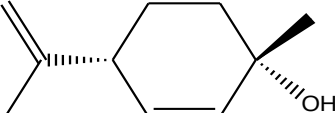
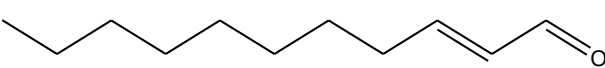
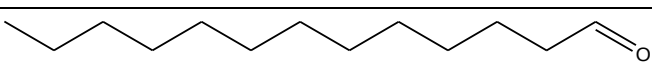
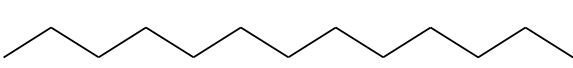
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۴۷	δ -3-Carene		۱۱۶۴	tr
۴۸	d-Camphor		۱۵۲۹	t
۴۹	Decanal		۱۵۰۹	v
۵۰	Decyl acetate		۱۶۹۰	tr
۵۱	Dehydrocarveol		۱۹۴۳	t
۵۲	δ -Muurolene		۱۶۸۳	tr
۵۳	Dodecanal		۱۷۱۸	tr
۵۴	Elemol		۲۰۹۰	tr
۵۵	Ethyl acetate		۹۰۴	v
۵۶	Eugenol		۲۱۷۲	tr

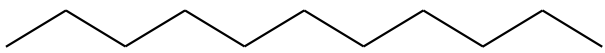
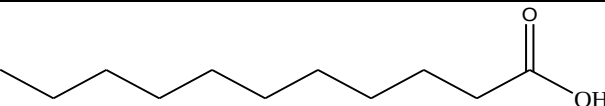
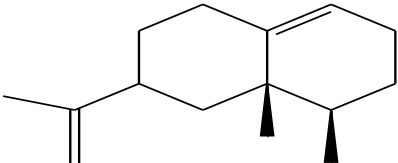
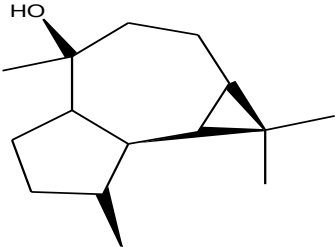
شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۵۷	Eugenol methylether		۲۰۳۱	t
۵۸	γ -Elemene		۱۶۳۳	tr
۵۹	γ -Eudesmol		۲۲۱۰	tr
۶۰	γ -Terpinene		۱۲۶۳	v
۶۱	Geraniol		۱۸۵۹	tr
۶۲	Geranyl propanoate		۱۸۳۰	tr
۶۳	Geranyl acetate		۱۷۶۶	t
۶۴	Germacrene D		۱۷۲۴	v
۶۵	Globulol		۲۰۶۲	tr
۶۶	Heptyl acetate		۱۳۹۲	t
۶۷	Isoeugenol		۲۲۷۱	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۶۸	Isothymol		۲۲۲۵	tr
۶۹	l-carvone		۱۷۴۴	tr
۷۰	Limonene		۱۲۳۱	v
۷۱	Linalool		۱۵۵۷	tr
۷۲	Linalyl acetate		۱۵۷۰	v
۷۳	l-Menthol		۱۶۵۲	tr
۷۴	Menthone		۱۴۷۶	tr
۷۵	Methyl Laurate		۱۸۱۲	t
۷۶	Myrcene		۱۱۷۰	tr
۷۷	Myrcenol		۱۶۲۵	t
۷۸	Neral		۱۶۹۷	v

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۷۹	Nerol		۱۸۵۱	tr
۸۰	Nerol oxide		۲۳۸۱	tr
۸۱	Neryl acetate		۱۷۳۴	v
۸۲	Nonanoic acid		۲۱۹۴	tr
۸۳	Nonyl acetate		۱۵۸۱	tr
۸۴	Octadecanal		۲۳۹۰	tr
۸۵	Octadecane		۱۸۰۹	t
۸۶	Octanal		۱۲۹۶	tr
۸۷	Octanoic acid		۲۰۸۴	tr
۸۸	Octanol		۱۵۶۲	v
۸۹	p-Cymene		۱۲۸۱	t
۹۰	Pentadecane		۱۵۰۴	tr
۹۱	Perillaldehyde		۱۷۹۳	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداری	سری
۹۲	Perillyl acetate		۱۹۱۲	v
۹۳	p-Mentha-1,8-dien-10-ol		۲۳۱۴	tr
۹۴	p-Mentha-1-en-9-ol		۱۹۴۸	tr
۹۵	p-Mentha-1-en-9-yl acetate		۱۸۳۲	t
۹۶	Sesquiphellandrene		۱۷۷۷	v
۹۷	Spathulenol		۲۱۲۱	tr
۹۸	Terpinen-4-ol		۱۶۱۳	t
۹۹	Terpinolene		۱۲۹۳	tr
۱۰۰	Tetradec-1-ene		۱۴۲۸	tr
۱۰۱	Tetradecanal		۱۹۲۹	t
۱۰۲	Tetradecane		۱۴۰۰	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۱۰۳	Tetradecenal		۱۹۶۴	v
۱۰۴	Trans-2-decenal		۱۶۴۸	tr
۱۰۵	Trans-2-dodecenal		۱۸۶۷	tr
۱۰۶	trans,trans-Farnesyl acetate		۲۲۷۶	v
۱۰۷	trans-β-Farnesene		۱۶۷۳	tr
۱۰۸	trans-Carveol		۱۸۷۷	tr
۱۰۹	trans-Dodec-2-enol		۲۰۳۶	t
۱۱۰	trans-Nerolidol		۲۰۵۰	v
۱۱۱	trans-Piperitol		۱۶۹۱	tr
۱۱۲	trans-p-Mentha-2,8-diene-1-ol		۱۶۴۱	t
۱۱۳	trans-2-Undecenal		۱۷۶۱	tr
۱۱۴	Tridecanal		۱۸۲۲	tr
۱۱۵	Tridecane		۱۳۱۱	tr

شماره ترکیب	نام ترکیب	ساختار ترکیب	شاخص بازداري	سری
۱۱۶	Undacane		۱۱۱۰	t
۱۱۷	Undecanoic acid		۲۴۱۷	tr
۱۱۸	Valencene		۱۷۲۸	v
۱۱۹	Viridiflorol		۲۱۰۸	tr

*v:validation series, t:test series, tr:train series.

۴-۲- رسم و بهینه‌سازی ساختار هندسی ترکیبات

در مرحله دوم، ساختمان مولکولی ترکیبات در نرم‌افزار HyperChem 8.0 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^۱ (بر حسب کیلوکالری بر مول)، به ۰/۰۰۱

برسد.

۴-۳- محاسبه‌ی توصیف‌کننده‌های مولکولی

ساختارهای بهینه شده، به عنوان ورودی به نرم افزار Dragon وارد شدند. محاسبه

توصیف‌کننده‌ها توسط این نرم‌افزار صورت پذیرفت و حدود ۱۴۸۱ توصیف‌کننده از هیجده گروه

مختلف برای هر ترکیب محاسبه گردید.

۱- Root Mean Square Gradient of Energy

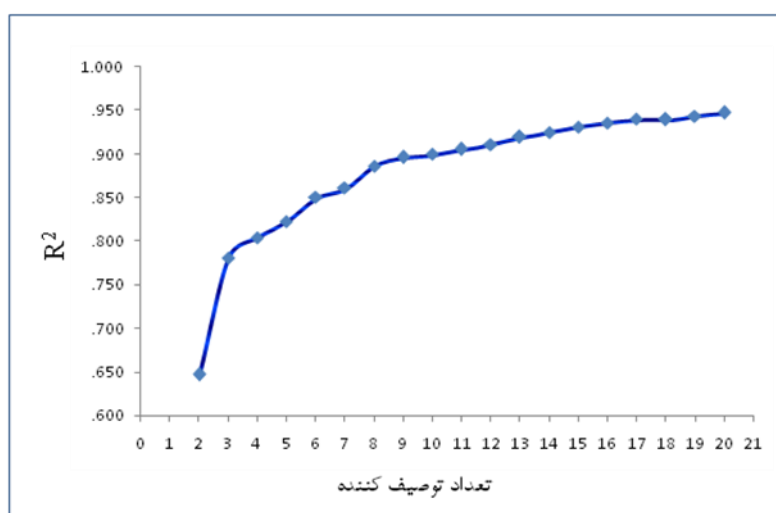
۴-۴- انتخاب بهترین توصیف‌کننده‌ها به روش رگرسیون خطی چندگانه

توصیف‌کننده‌های به دست آمده از مرحله قبل به عنوان متغیر مستقل و شاخص بازداری ترکیبات به عنوان متغیر وابسته وارد نرم افزار SPSS شدند. سپس همانطور که در دو فصل قبل (بخش ۲-۲-۴) توضیح داده شد؛ و با به‌کارگیری تکنیک‌های کاهش متغیر و روش برازش مرحله‌ای برای انتخاب متغیر، بهترین توصیف‌کننده‌ها انتخاب شدند. به این ترتیب که با استفاده از منوی آنالیز در نرم‌افزار، گزینه رگرسیون خطی انتخاب شد و بعد از وارد کردن متغیرهای مستقل و متغیر وابسته، روش مرحله‌ای برای مدل‌سازی گزینش گردید. از بین مدل‌های ایجاد شده مدلی که تعداد توصیف‌کننده‌های آن کمتر و ضریب تعیین (R^2) و آماره‌ی F بیشتری داشت، به عنوان بهترین مدل انتخاب شد. توصیف‌کننده‌های وارد شده در مدل و همچنین پارامترهای آماری مربوطه در جدول ۴-۲ آورده شده است.

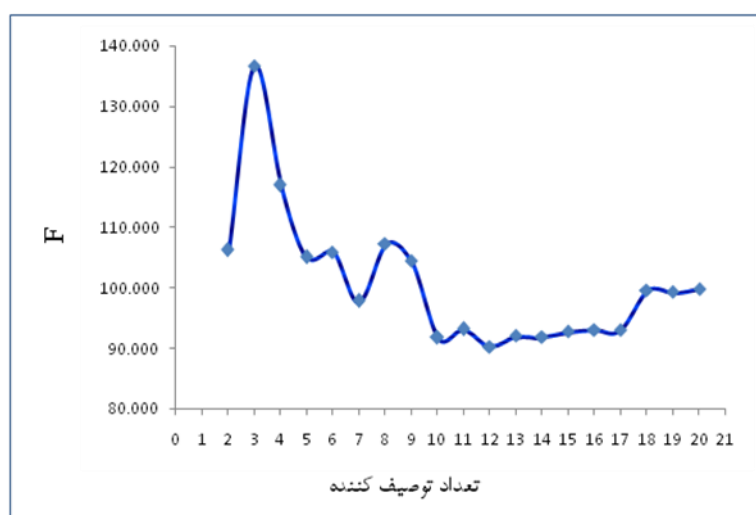
جدول ۴-۲- مقایسه آماره‌ی مدل‌های به دست آمده از روش MLR

مدل	توصیف‌کننده‌ها	R^2	F	SE
۲	Ss, MATS1e	۰/۶۴۷	۱۰۶/۲۳۴	۰/۲۱۰
۳	Ss, MATS1e, BAC	۰/۷۸۱	۱۳۶/۶۴۷	۰/۱۶۶
۴	Ss, MATS1e, BAC, RDF020m	۰/۸۰۴	۱۱۶/۹۵۵	۰/۱۵۸
۵	Ss, MATS1e, BAC, RDF020m, MATS1p	۰/۸۲۳	۱۰۵/۰۷۵	۰/۱۵۰
۶	Ss, MATS1e, BAC, RDF020m, MATS1p, R3v_A	۰/۸۵۰	۱۰۵/۹۰۴	۰/۱۳۹
۷	Ss, MATS1e, BAC, RDF020m, MATS1p, R3v_A, MLOGP	۰/۸۶۱	۹۷/۸۵۳	۰/۱۳۵
۸	Ss, MATS1e, BAC, RDF020m, MATS1p, R3v_A, MLOGP, Mor01v	۰/۸۸۶	۱۰۷/۲۸۸	۰/۱۲۲
۹	Ss, MATS1e, BAC, RDF020m, MATS1p, R3v_A, MLOGP, Mor01v, Mor07p	۰/۸۹۶	۱۰۴/۴۵۶	۰/۱۱۷

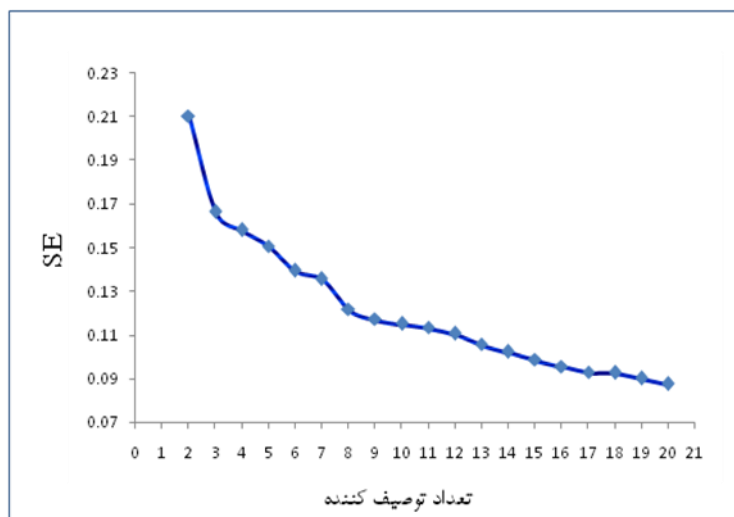
در شکل ۴-۱ تأثیر تعداد توصیف‌کننده‌ها بر مقادیر R^2 ، F و SE نشان داده شده است. با توجه به نمودارهای داده شده و همانطور که در جدول ۴-۲ مشاهده می‌شود، مقادیر R^2 و SE در دو مدل ۸ و ۹ تفاوت چندانی ندارند. آماره F نیز در مدل هشتم بیشتر است؛ و از آنجا که مدل شماره ۸ تعداد توصیف‌کننده کمتری دارد به عنوان مدل نهایی انتخاب گردید.



(الف)



(ب)



(ج)

شکل ۴-۱- تأثیر تعداد توصیف کننده‌ها بر مقادیر R^2 (الف)، F (ب) و SE (ج) در مدل‌های MLR

نام کامل توصیف کننده‌های وارد شده در این مدل به همراه ضریب رگرسیون و اثر متوسط

آنها در جدول ۴-۳ آورده شده است.

جدول ۴-۳- توصیف کننده‌های وارد شده در مدل MLR

شماره	علامت	ضریب رگرسیون	اثر متوسط	گروه	نام کامل
۱	Ss	۱۰۷۸/۳۹۲	۸/۶۰۰	constitutional descriptors	sum of Kier-Hall electrotopological states
۲	MATS1e	-۱۶۸۳/۲۹۴	-۵/۱۹۴	2D autocorrelation descriptors	Moran autocorrelation - lag 1 / weighted by atomic Sanderson electronegativities
۳	BAC	-۱۳/۴۹۶	-۷/۱۵۷	topological descriptors	Balaban centric index
۴	RDF020m	۶۲۳/۴۳۶	۴/۴۳۲	RDF descriptors	Radial Distribution Function – 2.0 / weighted by atomic masses
۵	MATS1p	۲۱۴۹/۴۹۴	۷/۶۶۸	2D autocorrelation descriptors	Moran autocorrelation – lag 1 / weighted by atomic polarizabilities
۶	R3v_A	-۷۶۷/۹۳۵	-۵/۴۷۹	GETAWAY descriptors	R maximal autocorrelation of lag 3 / weighted by atomic van der Waals volumes
۷	MLOGP	-۱۵۰/۱۶۷	-۵/۰۱۳	properties	Muriguchi octanol-water partition coefficient (logP)
۸	Mor01v	۲/۲۱۰	۴/۱۳۱	3D-MoRSE descriptors	3D-MoRSE – signal 01 / weighted by atomic van der Waals volumes

پارامترهای وارد شونده به معادله خطی باید حتی المقدور مستقل از یکدیگر باشند. جدول ۴-۴ ماتریس همبستگی توصیف کننده‌های وارد شده در مدل منتخب را نشان می‌دهد. به طوری که ملاحظه می‌شود مقدار همبستگی بین پارامترها قابل قبول (کمتر از ۰/۹) است.

جدول ۴-۴- ماتریس همبستگی توصیف‌کننده‌های به‌کار رفته در مدل نهایی

توصیف کننده	Ss	MATS1e	BAC	RDF020m	MATS1p	R3v_A	MLOGP	Mor01v
Ss	۱							
MATS1e	-۰/۰۶۰	۱						
BAC	۰/۷۶۲	-۰/۱۵۴	۱					
RDF020	۰/۵۷۵	-۰/۰۱۱	۰/۴۳۷	۱				
MATS1p	-۰/۱۴۰	-۰/۱۰۰	-۰/۳۱۰	-۰/۱۴۹	۱			
R3v_A	-۰/۰۴۷	۰/۲۵۰	-۰/۲۶۱	-۰/۱۲۰	۰/۷۵۳	۱		
MLOGP	۰/۱۴۵	۰/۲۱۷	۰/۲۱۷	۰/۲۴۲	۰/۳۳۲	۰/۳۴۷	۱	
Mor01v	۰/۶۵۹	۰/۰۸۷	۰/۵۶۴	۰/۵۱۴	۰/۱۵۳	۰/۲۳۵	۰/۷۷۷	۱

به این ترتیب، معادله خطی نهایی به صورت زیر به دست آمد:

$$RI_{cal.} = 52/285 + 1078/392Ss - 1683/294MATS1e - 13/496BAC + 623/436RDF020m + 2149/494MATS1p - 767/935R3v_A - 150/16VMLOGP + 2/210Mor01v$$

معادله (۴-۱)

۴-۵- مدل‌سازی با استفاده از شبکه عصبی مصنوعی (ANN)

در این مرحله، شبکه عصبی مصنوعی برای مدل‌سازی غیرخطی داده‌ها به منظور پیش‌بینی RI ترکیبات موجود در روغن‌های ضروری مورد استفاده قرار گرفت. مقادیر عددی توصیف‌کننده‌های وارد شده در مدل خطی به عنوان ورودی به شبکه داده شد و مقادیر تجربی اندیس بازداري (RI) به دنبال آن و به عنوان هدف وارد شبکه عصبی شدند. داده‌ها به صورت تصادفی به سه سری آموزش (شامل ۷۱ ترکیب)، سری ارزیابی (۲۴ ترکیب) و سری تست (۲۴ ترکیب) تقسیم‌بندی شدند.

برای به دست آوردن بهترین معادله و کمترین خطا، پارامترهای شبکه (تعداد متغیرهای ورودی شبکه، نوع توابع آموزش و انتقال، تعداد گره‌ها در لایه پنهان، μ و تعداد دورهای آموزش) بهینه‌سازی شدند.

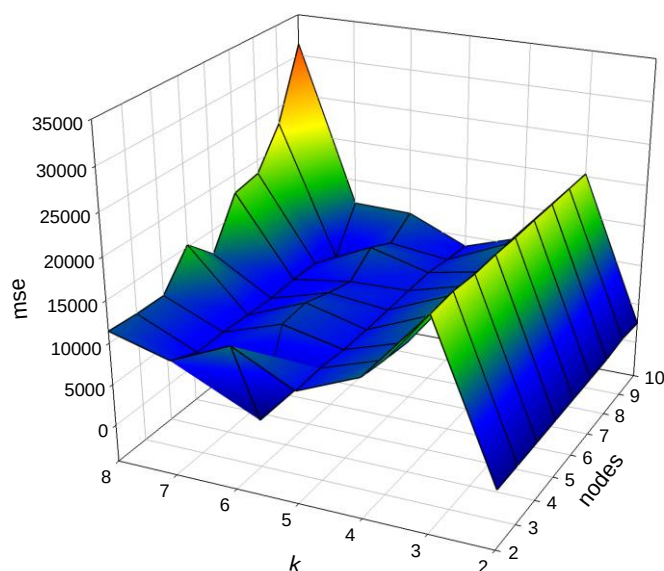
۴-۵-۱- بهینه‌سازی

تعداد گره‌ها از ۲ تا ۱۰ و تعداد متغیرهای ورودی از ۲ تا ۸ برای سه تابع انتقال tansig ، purelin و logsig با دو تابع آموزش trainlm و trainbr تغییر داده شد (جداول ۴-۵ تا ۴-۱۰). نقطه‌ای که کمترین خطا را برای سری ارزیابی داشت به عنوان نقطه بهینه انتخاب گردید. همانطور که مشاهده می‌شود، شبکه‌هایی با تابع آموزش trainbr و تابع انتقال purelin با تعداد گره ۹ نتایج بهتری را نشان می‌دهند.

با توجه به نتایج به دست آمده، الگوریتم آموزشی تنظیم بایزین نسبت به الگوریتم آموزشی لونبرگ-مارکوات دارای میانگین مربع خطاهای کمتری می‌باشد. این الگوریتم علاوه بر سرعت آموزش بسیار زیاد دارای قدرت تعمیم‌پذیری بالا می‌باشد، به این معنی که قادر است میان ترکیباتی که در آموزش به کار گرفته شده و آنهایی که در آموزش به کار گرفته نشده، مطابقت قابل قبولی برقرار نماید. همانطور که در جداول ۴-۵ تا ۴-۱۰ و شکل‌های ۴-۲ تا ۴-۷ ملاحظه می‌شود، در این تحقیق شبکه با تابع آموزش trainbr ، تابع انتقال purelin ، ۵ متغیر ورودی و تعداد ۹ نرون در لایه مخفی، کمترین میانگین خطاها را نشان داده و به عنوان شبکه بهینه انتخاب شد.

جدول ۴-۵- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

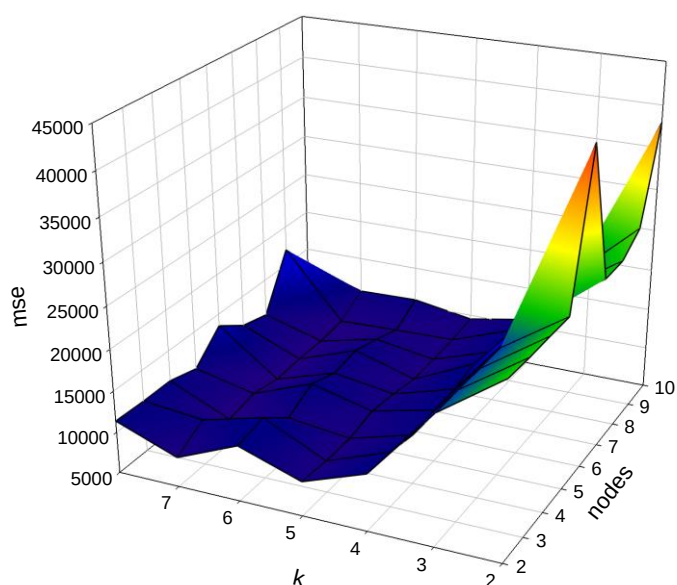
		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k						
		۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۵۸۲۰	۲۰۴۹۶	۱۲۰۹۲	۸۹۵۲	۱۲۶۲۰	۹۴۱۳	۱۱۴۰۳
	۳	۲۵۵۸۷	۲۰۷۰۱	۱۱۶۰۶	۹۱۱۲	۱۲۲۷۲	۷۹۷۳	۱۱۱۷۸
	۴	۲۵۵۱۴	۲۰۴۱۸	۱۱۲۳۷	۸۸۸۹	۱۰۴۸۵۷	۸۰۶۶	۱۱۱۵۰
	۵	۲۵۵۲۸	۲۰۴۸۴	۱۲۰۵۰	۸۹۴۱	۱۰۸۲۱	۷۸۸۷	۱۵۲۴۲
	۶	۲۵۵۵۸	۲۰۳۸۳	۱۲۲۰۳	۹۰۸۰	۹۸۳۲	۷۷۳۹	۱۲۰۴۱
	۷	۲۵۳۹۶	۲۰۴۰۰	۱۱۲۵۸	۹۲۶۱	۹۰۶۵	۷۹۹۴	۱۷۷۵۴
	۸	۲۵۴۵۴	۲۰۳۴۵	۱۰۸۲۸	۹۳۸۷	۱۱۳۱۰	۷۹۷۲	۱۸۳۶۷
	۹	۲۸۰۸۴	۲۰۳۴۲	۱۲۲۲۷	۸۹۶۷	۹۸۳۴	۷۷۲۹	۲۲۹۲۷
	۱۰	۲۵۳۸۴	۲۰۳۰۷	۱۱۰۴۰	۸۶۶۰	۱۱۷۸۱	۱۱۹۴۲	۳۱۳۷۷



شکل ۴-۲- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل لگاریتم سیگموئید

جدول ۴-۶- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

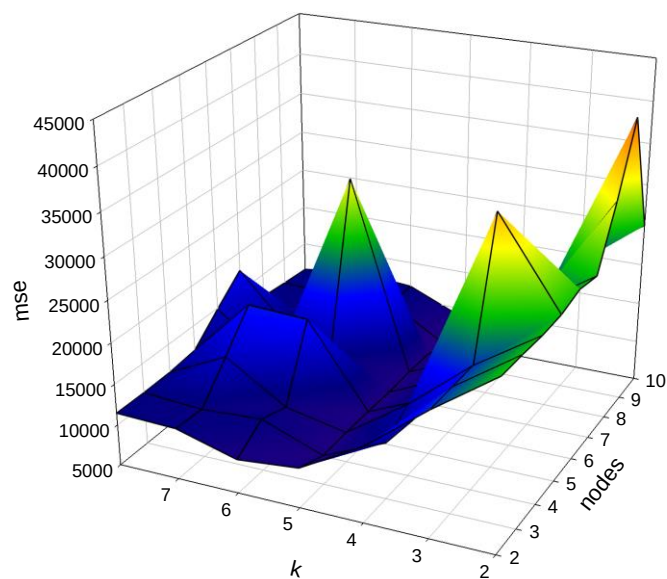
		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k							
			۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۵۸۳۶	۲۰۴۸۶	۱۱۸۳۶	۹۱۴۶	۱۱۷۱۰	۸۵۸۰	۱۱۳۷۷	
	۳	۲۵۶۳۷	۲۰۴۸۰	۱۱۶۰۰	۸۸۳۹	۱۲۴۰۰	۱۰۵۱۴	۱۱۲۳۴	
	۴	۲۵۵۸۴	۲۰۳۹۸	۱۰۹۵۸	۸۸۶۶	۹۷۷۱	۷۹۶۴	۱۱۲۵۹	
	۵	۲۵۴۵۴	۲۰۳۹۹	۱۱۳۶۲	۹۶۲۳	۱۱۴۶۷	۸۶۶۲	۱۰۳۱۷	
	۶	۴۲۵۳۶	۲۰۳۶۰	۱۱۴۱۵	۹۲۳۵	۱۰۳۴۱	۸۰۲۷	۱۳۵۰۴	
	۷	۲۵۳۷۲	۲۰۳۷۳	۱۲۱۰۸	۸۶۲۸	۱۰۸۵۲	۸۲۴۱	۱۱۲۴۰	
	۸	۲۵۳۶۱	۲۰۳۵۷	۱۱۲۸۵	۸۷۷۱	۹۹۴۳	۸۲۵۱	۱۰۳۷۰	
	۹	۲۷۲۲۹	۲۰۳۲۴	۱۱۶۹۵	۸۹۹۴	۸۸۵۶	۸۶۵۶	۱۶۹۷۶	
	۱۰	۳۷۹۷۰	۲۰۳۳۹	۱۰۸۴۴	۹۴۹۳	۱۰۷۰۹	۱۰۵۴۳	۱۲۸۱۴	



شکل ۴-۳- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل تانژانت سیگموئید

جدول ۴-۷- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی بایزین (trainbr) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

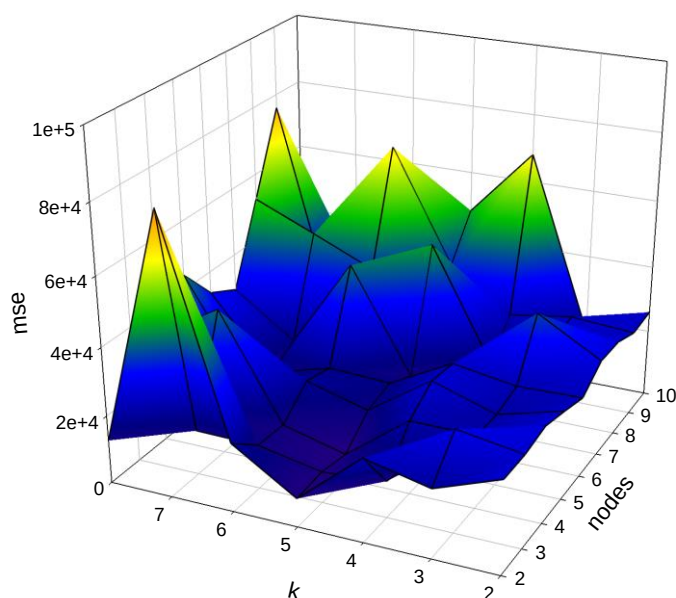
		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k						
		۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۵۳۷۱	۲۰۰۴۶	۱۴۱۲۹	۹۸۵۱	۹۱۳۲	۱۱۲۱۷	۱۱۴۷۱
	۳	۲۵۲۲۲	۲۰۰۴۷	۱۱۸۳۰	۸۵۵۹	۱۱۳۵۲	۱۰۹۳۶	۱۱۴۲۲
	۴	۲۴۹۵۸	۲۰۰۴۰	۱۳۷۹۸	۸۶۷۵	۱۰۲۲۱	۱۴۹۴۶	۱۱۴۹۴
	۵	۲۴۹۹۸	۳۵۳۰۸	۱۱۸۸۲	۸۶۶۰	۱۸۷۱۵	۱۹۰۸۷	۱۱۸۷۱
	۶	۲۵۶۸۵	۲۰۰۴۷	۱۱۹۸۳	۸۵۷۳	۱۲۱۸۸	۹۴۱۲	۱۴۸۷۰
	۷	۲۵۰۴۹	۲۰۰۵۳	۱۱۸۵۸	۸۷۱۰	۳۱۷۰۲	۱۰۱۸۹	۱۷۷۱۴
	۸	۳۱۸۰۲	۲۰۰۶۳	۱۱۸۶۲	۸۶۱۶	۱۱۹۹۵	۹۲۰۱	۱۲۰۷۱
	۹	۳۹۷۸۴	۲۰۱۰۰	۱۲۰۰۸	۷۲۹۵	۱۱۶۳۸	۱۱۱۹۸	۱۱۷۲۴
	۱۰	۲۴۹۶۱	۲۰۱۰۰	۱۱۸۲۰	۸۶۷۴	۱۱۷۸۱	۱۲۳۷۵	۱۱۴۳۶



شکل ۴-۴- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی بایزین و تابع تبدیل خطی

جدول ۴-۸- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت (trainlm) و تابع تبدیل لگاریتم سیگموئید (logsig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k						
		۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۷۱۳۷	۲۰۱۷۰	۲۴۴۳۱	۸۳۹۲	۲۰۲۰۱	۸۱۴۵۰	۱۲۸۳۳
	۳	۲۷۰۷۶	۳۰۰۵۰	۱۰۸۳۹	۹۱۸۲	۱۲۸۸۸	۱۳۱۶۳	۲۳۹۸۱
	۴	۲۸۴۲۰	۲۱۶۲۳	۲۱۷۹۳	۷۵۷۲	۱۲۳۹۱	۴۳۰۱۹	۳۲۵۶۸
	۵	۲۶۱۸۸	۲۱۱۳۹	۲۷۱۸۸	۱۴۳۹۶	۱۹۴۴۵	۱۵۴۴۱	۴۴۳۴۶
	۶	۲۳۶۶۱	۲۳۱۸۹	۲۶۲۳۵	۱۶۴۳۷	۱۷۳۲۳	۱۲۰۶۴	۳۳۵۸۳
	۷	۲۸۵۰۱	۳۹۰۴۸	۱۲۳۰۸	۱۲۵۱۴	۴۳۲۹۴	۱۶۲۷۲	۲۹۱۴۳
	۸	۲۸۲۳۴	۲۴۷۶۲	۱۴۱۶۸	۴۷۵۵۷	۳۶۸۶۲	۴۴۷۲۸	۵۲۴۰۵
	۹	۲۵۲۲۸	۲۶۶۵۸	۲۳۷۸۷	۳۶۱۴۴	۶۹۳۶۷	۱۰۶۷۳	۷۶۰۳۶
	۱۰	۲۶۰۲۳	۱۶۳۲۰	۶۸۱۹۶	۴۸۲۰۵	۴۰۱۰۸	۲۲۶۰۸	۴۲۴۰۲

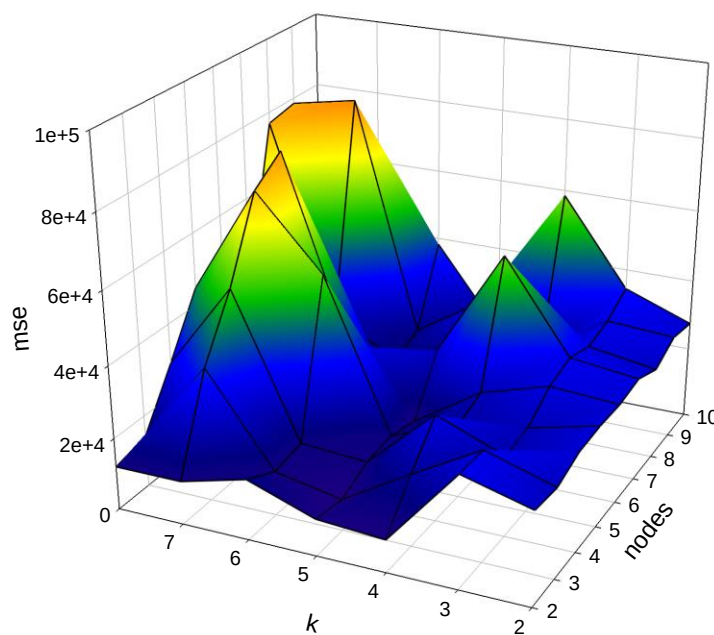


شکل ۴-۵- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت و تابع تبدیل لگاریتم

سیگموئید

جدول ۴-۹- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ - مارکوارت (trainlm) و تابع تبدیل تانژانت سیگموئید (tansig) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k						
		۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۵۹۳۱	۳۱۰۲۹	۹۰۷۶	۱۰۲۹۲	۱۶۸۹۲	۱۲۱۰۳	۱۲۰۵۵
	۳	۲۴۹۳۸	۳۰۶۹۰	۱۹۹۸۵	۸۶۶۸	۱۲۳۴۸	۳۷۱۷۲	۱۴۱۶۷
	۴	۲۷۷۲۱	۲۲۳۹۸	۲۹۷۰۴	۱۲۵۲۲	۱۳۷۷۴	۵۳۱۴۹	۲۹۰۰۹
	۵	۲۸۱۶۴	۲۳۷۲۱	۲۰۷۱۷	۱۳۴۶۸	۵۴۹۳۵	۷۴۳۰۹	۴۴۶۸۴
	۶	۲۷۷۵۶	۲۹۰۵۶	۲۳۴۶۱	۱۱۲۹۰	۲۹۴۳۷	۸۰۴۲۰	۴۱۳۱۴
	۷	۲۸۸۲۷	۳۱۱۱۶	۵۶۲۱۰	۲۶۰۹۷	۲۱۶۵۵	۱۶۷۲۸	۳۵۷۶۷
	۸	۲۵۵۷۶	۲۸۵۵۱	۲۱۸۰۲	۱۶۵۷۰	۱۳۹۲۸	۲۳۲۱۸	۷۶۹۱۸
	۹	۲۹۰۷۸	۳۰۳۶۱	۱۸۰۱۴	۲۸۸۸۳	۱۷۱۶۲	۸۱۷۸۲	۷۸۵۲۷
	۱۰	۲۷۵۰۶	۳۴۴۷۳	۵۸۵۶۴	۱۵۱۳۳	۳۸۲۰۷	۲۴۶۷۵	۲۴۲۸۷

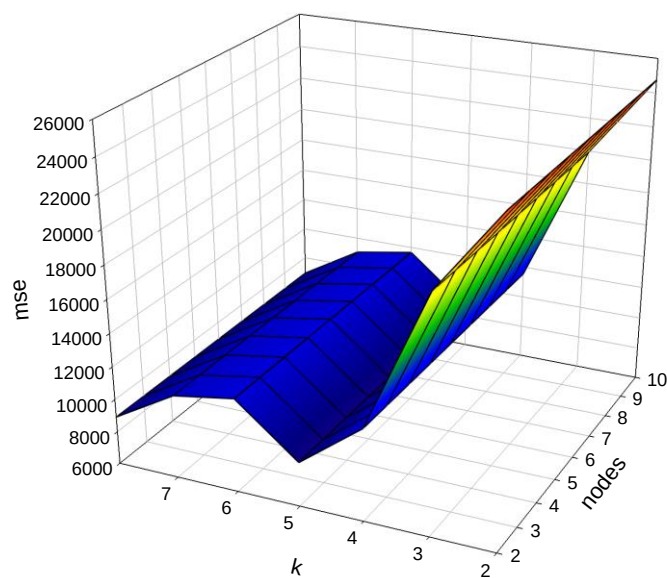


شکل ۴-۶- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لوببرگ- مارکوارت و تابع تبدیل تانژانت

سیگموئید

جدول ۴-۱۰- مقادیر میانگین مربع خطاها برای شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت (trainlm) و تابع تبدیل خطی (purelin) با تعداد متغیرهای ورودی متفاوت و تعداد گره‌های مختلف در لایه پنهان

		تعداد متغیرهای ورودی به شبکه یا توصیف کننده‌ها k						
		۲	۳	۴	۵	۶	۷	۸
تعداد گره‌ها	۲	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۳	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۴	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۵	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۶	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۷	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۸	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۹	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷
	۱۰	۲۴۷۴۸	۱۹۹۴۹	۱۱۵۴۵	۸۶۷۵	۱۱۶۶۴	۱۱۰۳۷	۸۹۱۷



شکل ۴-۷- نمایش سه بعدی شبکه‌هایی با الگوریتم آموزشی لوببرگ-مارکوارت و تابع تبدیل خطی

۴-۵-۱-۱- بهینه سازی تعداد چرخه‌های آموزشی

بدین منظور، ابتدا مقدار بهینه پارامترهای شبکه شامل ۹ نرون در لایه مخفی، ۵ توصیف‌کننده در لایه ورودی و با توابع آموزش و انتقال انتخاب شده، در شبکه‌ی مورد نظر قرار داده شد و با تغییر تعداد چرخه‌های آموزشی، مقادیر خطای مدل‌ها به دست آمد. این نتایج در محدوده‌ی کمترین خطاها در جدول ۴-۱۱ آورده شده و در شکل ۴-۸ منحنی آن رسم گردیده است.

جدول ۴-۱۱- مقادیر خطای شبکه با تغییر تعداد چرخه‌های آموزشی

میانگین مربع خطا	تعداد دوره های آموزشی
۸۴۱۴	۱۶
۹۱۲۲	۱۷
۷۸۶۵	۱۸
۸۶۱۳	۱۹
۷۲۹۵	۲۰
۸۳۷۹	۲۱
۷۸۲۳	۲۲
۸۸۰۲	۲۳
۸۴۸۸	۲۴
۸۹۱۴	۲۵
۸۷۲۴	۲۶
۹۰۹۶	۲۷
۸۹۲۴	۲۸
۹۲۴۲	۲۹
۸۸۷۲	۳۰



شکل ۴-۸- منحنی مقادیر خطای شبکه در مقابل تعداد چرخه‌های آموزشی

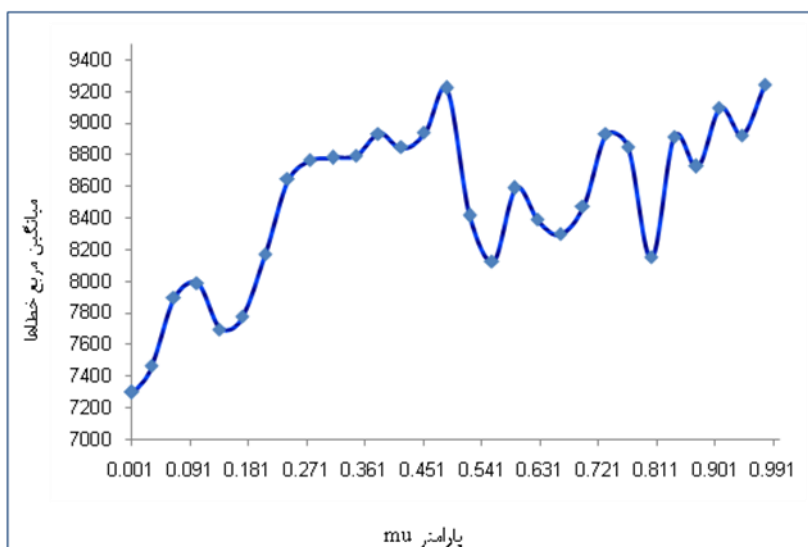
با توجه به نتایج به دست آمده و منحنی رسم شده برای میانگین مربعات خطای سری ارزیابی، تعداد بهینه چرخه‌های آموزشی ۲۰ انتخاب گردید.

۴-۵-۱-۲- بهینه‌سازی پارامتر μ

برای بهینه‌سازی پارامتر μ مقدار آن در شبکه بهینه شده بین ۰/۰۰۱ تا ۱/۰ تغییر داده شد و سپس MSE برای هر مورد در سری ارزیابی محاسبه و منحنی مقادیر آن بر حسب μ رسم گردید (جدول ۴-۱۲ و شکل ۴-۹). نتایج نشان داد که مقدار ۰/۰۰۱ برای پارامتر μ یعنی همان مقدار پیش‌فرض الگوریتم بایزین، می‌تواند به عنوان نقطه بهینه در نظر گرفته شود.

جدول ۴-۱۲ - مقادیر خطای شبکه با تغییر میزان μ

میانگین مربع خطا	مقدار پارامتر μ	میانگین مربع خطا	مقدار پارامتر μ
۸۴۱۴	۰/۵۲۱	۷۲۹۵	۰/۰۰۱
۸۱۲۲	۰/۵۵۶	۷۴۶۶	۰/۰۳۱
۸۵۸۶	۰/۵۹۱	۷۸۹۷	۰/۰۶۶
۸۳۸۷	۰/۶۲۶	۷۹۸۶	۰/۱۰۱
۸۲۹۹	۰/۶۶۱	۷۶۹۶	۰/۱۳۶
۸۴۷۲	۰/۶۹۶	۷۷۷۵	۰/۱۷۱
۸۹۳۲	۰/۷۳۱	۸۱۷۰	۰/۲۰۶
۸۸۴۶	۰/۷۶۶	۸۶۴۴	۰/۲۴۱
۸۱۵۰	۰/۸۰۱	۸۷۶۸	۰/۲۷۶
۸۹۱۴	۰/۸۳۶	۸۷۸۰	۰/۳۱۱
۸۷۲۵	۰/۸۷۱	۸۷۹۲	۰/۳۴۶
۹۰۹۶	۰/۹۰۶	۸۹۳۲	۰/۳۸۱
۸۹۲۴	۰/۹۴۱	۸۸۴۶	۰/۴۱۶
۹۲۴۲	۰/۹۷۶	۸۹۴۱	۰/۴۵۱
		۹۲۱۹	۰/۴۸۶



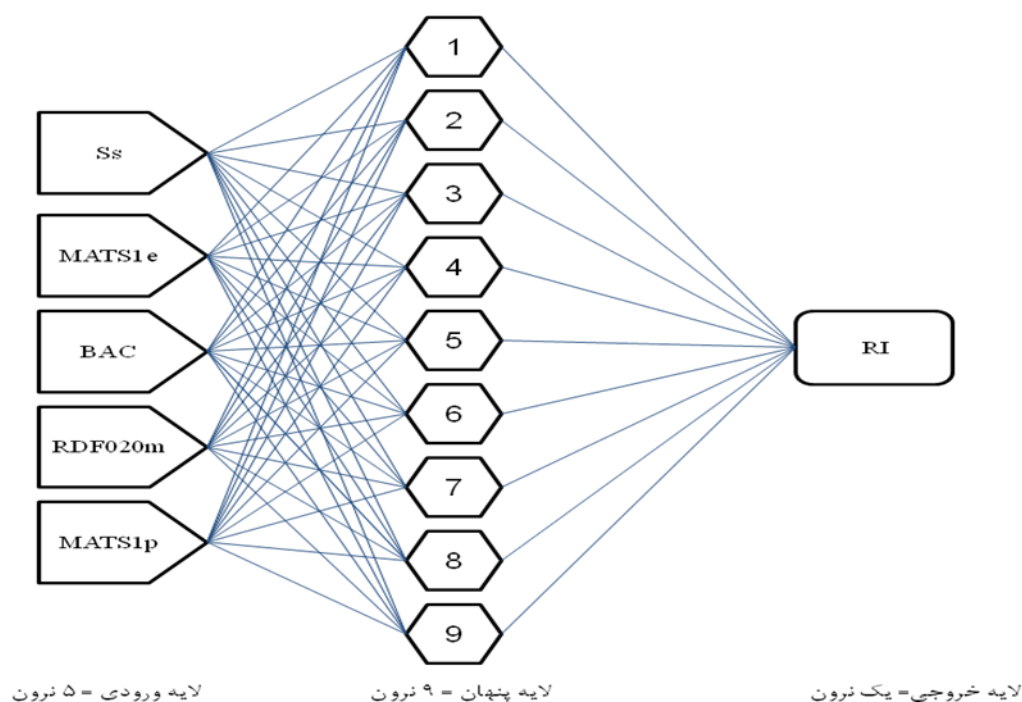
شکل ۴-۹- منحنی مقادیر خطای شبکه در مقابل پارامتر μ

۴-۵-۱-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده در بهینه‌سازی شبکه، یک شبکه عصبی سه لایه پس‌انتشار با مشخصات ذکر شده در جدول ۴-۱۳ به عنوان شبکه بهینه انتخاب شد که ساختار آن نیز به طور شماتیک در شکل ۴-۱۰ نشان داده شده است.

جدول ۴-۱۳- مشخصات شبکه بهینه

تابع آموزش	trainbr
تابع انتقال لایه پنهان	purelin
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد نرون های لایه پنهان	۹
تعداد متغیرهای ورودی	۵
تعداد چرخه‌های آموزشی	۲۰
پارامتر μ	۰/۰۰۱



شکل ۴-۱۰- تصویر شماتیک ساختار شبکه عصبی مصنوعی به دست آمده پس از بهینه سازی

۴-۶- ارزیابی مدل‌ها

۴-۶-۱- ارزیابی مدل‌های MLR و ANN به کمک سری‌های ارزیابی و تست

با استفاده از روش‌های ANN و MLR مقادیر اندیس بازداری برای سری‌های ارزیابی و تست پیش‌بینی شدند. در جداول ۴-۱۴ تا ۴-۱۷ مقادیر محاسبه شده و درصد خطای نسبی برای سری ارزیابی و سری تست در روش‌های خطی و غیرخطی آورده شده است. نمودارهای مقادیر پیش‌بینی شده در مقابل مقادیر تجربی شاخص بازداری برای دو سری ارزیابی و تست مربوط به هر دو روش به ترتیب در شکل‌های ۴-۱۱ تا ۴-۱۴ رسم شده است. ضرایب تعیین برای سری ارزیابی در روش‌های خطی و غیرخطی به ترتیب ۰/۹۳۷ و ۰/۹۶۱ و برای سری تست ۰/۸۸۵ و ۰/۹۱۵ می‌باشد.

جدول ۴-۱۴- مقایسه مقادیر تجربی و پیش‌بینی شده در روش MLR برای سری ارزیابی

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطای نسبی
۳	۱۵۰۰	۱۶۴۶	۸/۹
۸	۱۶۶۸	۱۷۰۵	۲/۲
۱۳	۱۰۷۶	۱۰۱۵	-۶/۰
۱۶	۱۰۳۹	۱۱۸۸	۱۲/۵
۲۳	۱۵۵۲	۱۶۱۰	۳/۶
۲۹	۲۲۳۹	۲۲۲۱	-۰/۸
۳۱	۱۵۴۸	۱۶۶۳	۶/۹
۳۹	۲۳۵۲	۲۱۷۳	-۸/۲
۴۲	۱۷۵۵	۱۸۰۷	۲/۹
۴۹	۱۵۰۹	۱۵۵۴	۲/۹
۵۵	۹۰۴	۶۱۶	-۳۱/۸
۶۰	۱۲۶۳	۱۲۷۶	۱/۰
۶۴	۱۷۲۴	۱۷۵۸	۱/۹
۷۰	۱۲۳۱	۱۲۲۸	-۰/۳
۷۲	۱۵۷۰	۱۵۷۸	۰/۵
۷۸	۱۶۹۷	۱۶۶۹	-۱/۷
۸۱	۱۷۳۴	۱۸۴۷	۶/۱
۸۸	۱۵۶۲	۱۵۴۵	-۱/۱
۹۲	۱۹۱۲	۱۹۸۲	۳/۵
۹۶	۱۷۷۷	۱۷۲۶	-۳/۰
۱۰۳	۱۹۶۴	۲۰۰۳	۱/۹

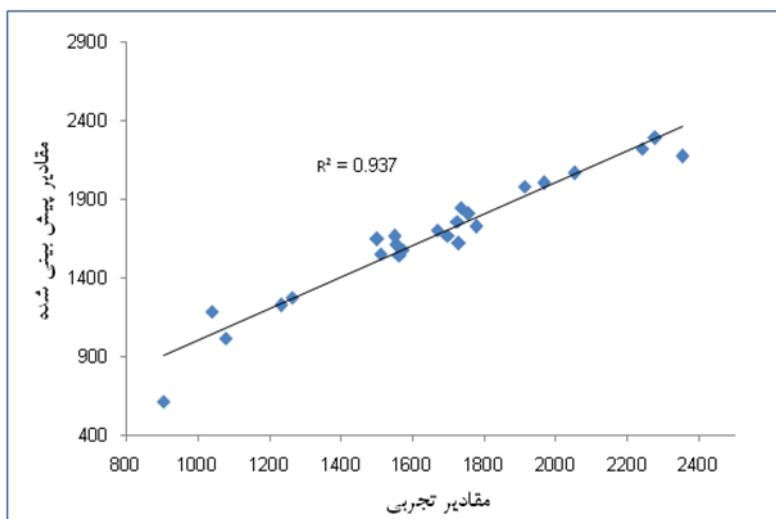
ادامه جدول ۴-۱۴

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۱۰۶	۲۲۷۶	۲۲۸۹	۰/۶
۱۱۰	۲۰۵۰	۲۰۶۵	۰/۷
۱۱۸	۱۷۲۸	۱۶۲۱	-۶/۲

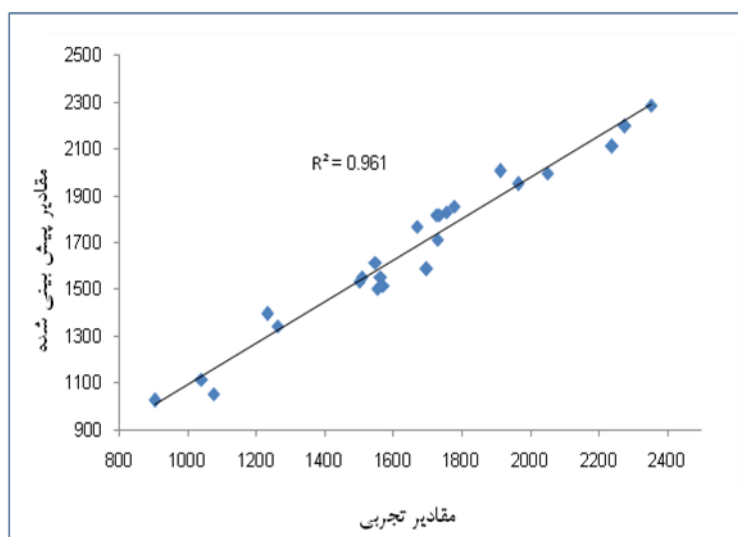
جدول ۴-۱۵ - مقایسه مقادیر تجربی و پیش‌بینی شده در روش ANN برای سری ارزیابی

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۳	۱۵۰۰	۱۵۳۵	۲/۳
۸	۱۶۶۸	۱۷۶۸	۶/۰
۱۳	۱۰۷۶	۱۰۵۵	-۲/۰
۱۶	۱۰۳۹	۱۱۱۶	۷/۴
۲۳	۱۵۵۲	۱۵۰۴	-۳/۱
۲۹	۲۲۳۹	۲۱۱۲	-۵/۷
۳۱	۱۵۴۸	۱۶۱۳	۴/۲
۳۹	۲۳۵۲	۲۰۸۶	-۱۱/۳
۴۲	۱۷۵۵	۱۸۳۰	۴/۳
۴۹	۱۵۰۹	۱۵۵۱	۲/۸
۵۵	۹۰۴	۱۰۲۵	۱۳/۴

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۶۰	۱۲۶۳	۱۳۴۴	۶/۵
۶۴	۱۷۲۴	۱۸۱۹	۵/۵
۷۲	۱۵۷۰	۱۵۱۵	-۳/۵
۷۸	۱۶۹۷	۱۵۸۵	-۶/۶
۸۱	۱۷۳۴	۱۸۱۸	۴/۸
۸۸	۱۵۶۲	۱۵۵۰	-۰/۷
۹۲	۱۹۱۲	۲۰۰۶	۴/۹
۹۶	۱۷۷۷	۱۸۵۴	۴/۳
۱۰۳	۱۹۶۴	۱۹۵۱	-۰/۷
۱۰۶	۲۲۷۶	۲۱۹۷	-۳/۵
۱۱۰	۲۰۵۰	۱۹۹۷	-۲/۶
۱۱۸	۱۷۲۸	۱۷۱۴	-۰/۸



شکل ۴-۱۱- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری ارزیابی



شکل ۴-۱۲- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ANN برای سری ارزیابی

جدول ۴-۱۶- مقایسه مقادیر تجربی و پیش‌بینی شده‌ی سری تست در MLR

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطای نسبی
۴	۱۴۶۶	۱۵۹۶	۸/۱
۷	۱۴۷۰	۱۳۶۵	-۷/۷
۱۵	۱۱۷۹	۱۲۲۰	۳/۴
۱۷	۱۱۹۵	۱۲۴۴	۴/۰
۲۱	۱۴۹۴	۱۶۵۷	۹/۸
۲۷	۱۲۳۳	۱۲۱۶	-۱/۴
۳۴	۱۸۱۷	۱۸۵۹	۲/۳
۳۷	۲۱۱۲	۲۰۲۳	-۴/۴
۴۵	۱۶۲۸	۱۶۷۴	۲/۷
۴۸	۱۵۲۹	۱۴۲۸	-۷/۰
۵۱	۱۹۴۳	۱۸۲۹	-۶/۲
۵۷	۲۰۳۱	۱۸۶۷	-۸/۸
۶۳	۱۷۶۶	۱۸۸۴	۶/۲
۶۶	۱۳۹۲	۱۳۸۲	-۰/۸
۷۵	۱۸۱۲	۱۹۵۱	۷/۰
۷۷	۱۶۲۵	۱۶۶۸	-۲/۶
۸۵	۱۸۰۹	۱۷۵۳	-۳/۲
۹۵	۱۸۳۲	۱۹۵۵	۶/۳
۹۸	۱۶۱۳	۱۷۵۰	۷/۸
۸۹	۱۲۸۱	۱۲۹۳	۰/۹

ادامه جدول ۴-۱۶

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۱۰۱	۱۹۲۹	۱۹۳۳	۰/۲
۱۰۹	۲۰۳۶	۱۹۶۶	-۳/۶
۱۱۲	۱۶۴۱	۱۸۵۴	۱۱/۵
۱۱۶	۱۱۱۰	۱۰۹۹	-۱/۰

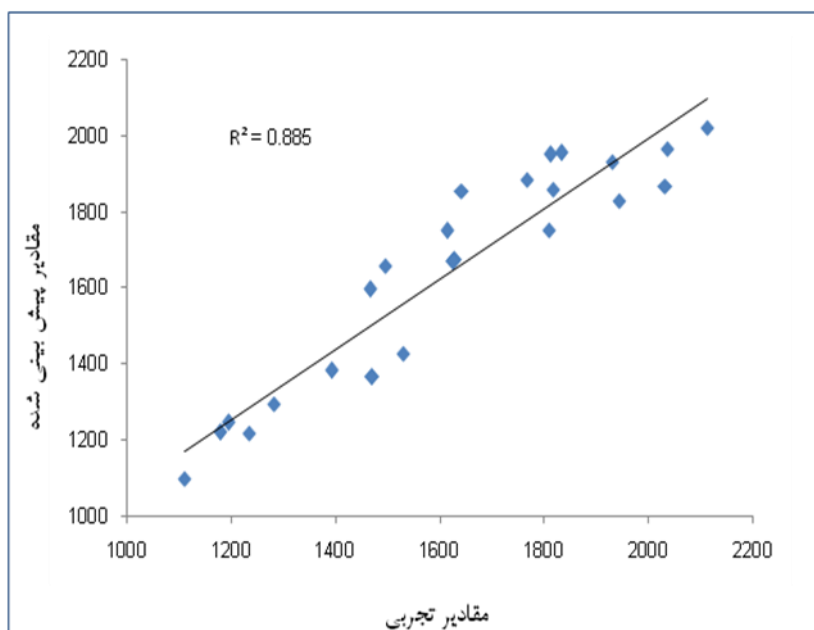
جدول ۴-۱۷- مقایسه مقادیر تجربی و پیش‌بینی شده‌ی سری تست در ANN

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۴	۱۴۶۶	۱۴۶۴	-۰/۱
۷	۱۴۷۰	۱۳۴۶	-۸/۴
۱۵	۱۱۷۹	۱۲۳۲	۴/۴
۱۷	۱۱۹۵	۱۱۲۰	-۶/۲
۲۱	۱۴۹۴	۱۶۵۰	۱۰/۴
۲۷	۱۲۳۳	۱۳۸۸	۱۲/۶
۳۴	۱۸۱۷	۱۸۴۹	۱/۸
۳۷	۲۱۱۲	۲۰۵۸	-۲/۶
۴۵	۱۶۲۸	۱۷۰۷	۴/۸
۴۸	۱۵۲۹	۱۴۹۲	-۲/۴

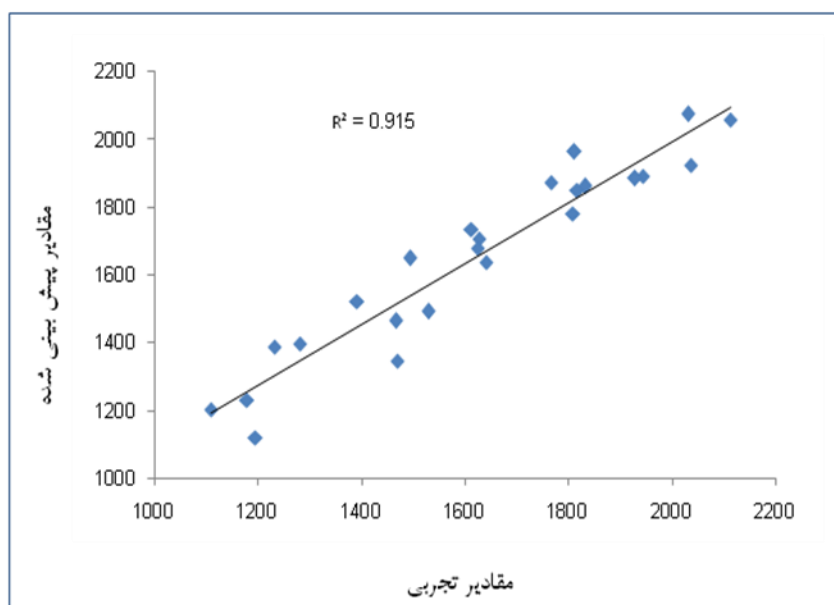
ادامه جدول ۱۷-۴

ادامه جدول ۱۷-۴

شماره ترکیب	Retention Index		
	مقدار تجربی	مقدار پیش بینی شده	درصد خطای نسبی
۵۱	۱۹۴۳	۱۸۹۱	-۲/۷
۵۷	۲۰۳۱	۲۰۷۴	۲/۱
۶۳	۱۷۶۶	۱۸۷۳	۶/۱
۶۶	۱۳۹۲	۱۵۲۳	۹/۴
۷۵	۱۸۱۲	۱۹۶۴	۸/۴
۷۷	۱۶۲۵	۱۶۷۹	۳/۳
۸۵	۱۸۰۹	۱۷۸۲	-۱/۵
۸۹	۱۲۸۱	۱۳۹۷	۹/۱
۹۵	۱۸۳۲	۱۸۶۲	۱/۶
۹۸	۱۶۱۳	۱۷۳۳	۷/۵
۱۰۱	۱۹۲۹	۱۸۸۵	-۲/۳
۱۰۹	۲۰۳۶	۱۹۲۵	-۵/۴
۱۱۲	۱۶۴۱	۱۶۳۷	-۰/۲
۱۱۶	۱۱۱۰	۱۲۰۲	۸/۳



شکل ۴-۱۳- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش MLR برای سری تست



شکل ۴-۱۴- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در روش ANN برای سری تست

۴-۶-۲- ارزیابی مدل‌های خطی و غیر خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

برای ارزیابی مدل، روش حذف مرحله‌ای تک‌تک داده‌ها مورد استفاده قرار گرفت و نتایج حاصل از آن با مقادیر تجربی برای سری داده‌ها توسط مدل‌های خطی و غیرخطی مقایسه گردید. مقدار واقعی، پیش‌بینی شده، درصد خطای نسبی برای سری داده‌ها در جداول ۴-۱۸ و ۴-۱۹ مشاهده می‌شود. نزدیک بودن این نتایج به دست آمده از مدل منتخب نشان‌دهنده صحت مدل می‌باشد.

جدول ۴-۱۸- ارزیابی مدل خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۰/۹	۲۲۴۹	۲۲۲۹	۱
-۱۲/۲	۱۳۸۳	۱۵۷۴	۲
۸/۸	۱۶۳۲	۱۵۰۰	۳
۷/۲	۱۵۷۱	۱۴۶۶	۴
-۲/۱	۱۴۲۸	۱۴۵۸	۵
-۲/۲	۱۱۱۰	۱۱۳۵	۶
-۳/۰	۱۴۲۶	۱۴۷۰	۷
۲/۵	۱۷۱۰	۱۶۶۸	۸
۱/۶	۲۳۴۰	۲۳۰۳	۹
۴/۱	۱۳۴۰	۱۲۸۷	۱۰
۳/۴	۱۴۰۸	۱۳۶۱	۱۱
۱/۷	۲۲۵۷	۲۲۱۹	۱۲
-۷/۷	۹۹۳	۱۰۷۶	۱۳
۰/۶	۱۶۸۹	۱۶۷۸	۱۴
۴/۰	۱۲۲۶	۱۱۷۹	۱۵
۱۵/۳	۱۱۹۸	۱۰۳۹	۱۶
۴/۷	۱۲۵۱	۱۱۹۵	۱۷
۳/۹	۱۷۷۵	۱۷۰۸	۱۸
۰/۲	۱۷۰۴	۱۷۰۱	۱۹
۸/۴	۱۵۵۳	۱۴۳۳	۲۰
۱۰/۲	۱۶۴۶	۱۴۹۴	۲۱

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۳/۳	۱۶۵۹	۱۶۰۶	۲۲
۱/۳	۱۵۷۲	۱۵۵۲	۲۳
-۱۰/۵	۱۴۲۸	۱۵۹۶	۲۴
-۶/۶	۲۰۹۹	۲۲۴۶	۲۵
-۵/۵	۱۸۴۶	۱۹۵۳	۲۶
-۱/۳	۱۲۱۷	۱۲۳۳	۲۷
۶/۵	۱۱۹۹	۱۱۲۶	۲۸
-۳/۹	۲۱۵۲	۲۲۳۹	۲۹
۹/۱	۱۵۸۰	۱۴۴۹	۳۰
۷/۵	۱۶۶۴	۱۵۴۸	۳۱
۳/۸	۱۶۵۲	۱۵۹۲	۳۲
-۸/۳	۹۹۵	۱۰۸۵	۳۳
۵/۰	۱۹۰۸	۱۸۱۷	۳۴
-۱/۳	۱۹۷۴	۲۰۰۱	۳۵
-۱/۵	۲۱۰۹	۲۱۴۲	۳۶
-۵/۵	۱۹۹۴	۲۱۱۲	۳۷
۵/۸	۲۲۷۵	۲۱۵۰	۳۸
-۸/۶	۲۱۴۹	۲۳۵۲	۳۹
۵/۴	۱۳۱۰	۱۲۴۳	۴۰
۶/۴	۱۹۶۴	۱۸۴۵	۴۱
۷/۰	۱۸۷۸	۱۷۵۵	۴۲
۳/۴	۱۵۳۷	۱۴۸۶	۴۳
-۲/۶	۱۷۲۶	۱۷۷۲	۴۴
۴/۰	۱۶۹۳	۱۶۲۸	۴۵
-۱/۸	۱۷۵۲	۱۷۸۵	۴۶
-۵/۱	۱۱۰۵	۱۱۶۴	۴۷
-۸/۲	۱۴۰۴	۱۵۲۹	۴۸
۶/۰	۱۵۹۹	۱۵۰۹	۴۹
۴/۲	۱۷۶۲	۱۶۹۰	۵۰
-۶/۰	۱۸۲۸	۱۹۴۳	۵۱

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۱/۷	۱۶۵۴	۱۶۸۳	۵۲
۳/۳	۱۷۷۴	۱۷۱۸	۵۳
-۱۰/۴	۱۸۷۲	۲۰۹۰	۵۴
-۱۲/۰	۷۹۶	۹۰۴	۵۵
۱۰/۴	۲۳۹۸	۲۱۷۲	۵۶
-۷/۱	۱۸۸۸	۲۰۳۱	۵۷
۰/۴	۱۶۳۹	۱۶۳۳	۵۸
-۳/۴	۲۱۳۴	۲۲۱۰	۵۹
۱/۴	۱۲۸۱	۱۲۶۳	۶۰
-۰/۴	۱۸۵۲	۱۸۵۹	۶۱
-۲/۶	۱۷۸۳	۱۸۳۰	۶۲
۵/۸	۱۸۶۹	۱۷۶۶	۶۳
۱/۱	۱۷۴۲	۱۷۲۴	۶۴
۴/۶	۲۱۵۸	۲۰۶۲	۶۵
۳/۴	۱۴۴۰	۱۳۹۲	۶۶
۱/۸	۲۳۱۲	۲۲۷۱	۶۷
۱۴/۸	۱۸۵۹	۲۲۲۵	۶۸
-۵/۴	۱۶۴۹	۱۷۴۴	۶۹
$-۸/۱e^{-۵}$	۱۲۳۱	۱۲۳۱	۷۰
۷/۸	۱۶۷۸	۱۵۵۷	۷۱
-۲/۵	۱۵۳۰	۱۵۷۰	۷۲
۶/۰	۱۷۵۱	۱۶۵۲	۷۳
۳/۲	۱۵۲۲	۱۴۷۶	۷۴
۱۰/۵	۲۰۰۳	۱۸۱۲	۷۵
۱۲/۴	۱۳۱۵	۱۱۷۰	۷۶
۱/۹	۱۶۵۶	۱۶۲۵	۷۷
-۱/۷	۱۶۶۸	۱۶۹۷	۷۸
۰/۱	۱۸۵۳	۱۸۵۱	۷۹
-۳/۸	۲۲۹۱	۲۳۸۱	۸۰
۶/۱	۱۸۳۹	۱۷۳۴	۸۱

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۱۰/۰	۱۹۷۶	۲۱۹۴	۸۲
۳/۹	۱۶۴۳	۱۵۸۱	۸۳
۳/۱	۲۴۶۵	۲۳۹۰	۸۴
-۰/۷	۱۷۹۷	۱۸۰۹	۸۵
۱۱/۸	۱۴۴۸	۱۲۹۶	۸۶
-۹/۲	۱۸۹۲	۲۰۸۴	۸۷
۳/۸	۱۶۲۱	۱۵۶۲	۸۸
$7/8e^{-5}$	۱۲۸۱	۱۲۸۱	۸۹
-۳/۱	۱۴۵۸	۱۵۰۴	۹۰
-۲/۶	۱۷۴۷	۱۷۹۳	۹۱
۲/۷	۱۹۶۳	۱۹۱۲	۹۲
-۱۲/۷	۲۰۲۱	۲۳۱۴	۹۳
-۱/۸	۱۹۱۳	۱۹۴۸	۹۴
۶/۹	۱۹۵۸	۱۸۳۲	۹۵
-۴/۵	۱۶۹۸	۱۷۷۷	۹۶
۴/۹	۲۲۲۶	۲۱۲۱	۹۷
۹/۴	۱۷۶۵	۱۶۱۳	۹۸
۰/۹	۱۳۰۵	۱۲۹۳	۹۹
۱۰/۴	۱۵۷۷	۱۴۲۸	۱۰۰
۲/۴	۱۹۷۴	۱۹۲۹	۱۰۱
-۴/۴	۱۳۳۹	۱۴۰۰	۱۰۲
۳/۱	۲۰۲۵	۱۹۶۴	۱۰۳
۰/۷	۱۶۶۰	۱۶۴۸	۱۰۴
-۲/۰	۱۸۲۹	۱۸۶۷	۱۰۵
-۰/۴	۲۲۶۶	۲۲۷۶	۱۰۶
۰/۶	۱۶۸۳	۱۶۷۳	۱۰۷
-۱/۳	۱۸۵۳	۱۸۷۷	۱۰۸
-۲/۱	۱۹۹۴	۲۰۳۶	۱۰۹
-۱/۱	۲۰۲۸	۲۰۵۰	۱۱۰
۱۱/۰	۱۸۷۶	۱۶۹۱	۱۱۱

ادامه جدول ۴-۱۸

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۱۲/۸	۱۸۵۱	۱۶۴۱	۱۱۲
-۱/۸	۱۷۳۰	۱۷۶۱	۱۱۳
۱/۹	۱۸۵۷	۱۸۲۲	۱۱۴
-۴/۱	۱۲۵۷	۱۳۱۱	۱۱۵
-۱/۶	۱۰۹۲	۱۱۱۰	۱۱۶
-۱۲/۵	۲۱۱۴	۲۴۱۷	۱۱۷
-۸/۲	۱۵۸۷	۱۷۲۸	۱۱۸
۲/۰	۲۱۴۹	۲۱۰۸	۱۱۹

جدول ۴-۱۹- ارزیابی مدل غیر خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۱/۹	۲۲۷۲	۲۲۲۹	۱
-۸/۱	۱۴۴۶	۱۵۷۴	۲
۳/۴	۱۵۵۱	۱۵۰۰	۳
-۱/۸	۱۴۴۰	۱۴۶۶	۴
-۵/۱	۱۳۸۳	۱۴۵۸	۵
-۱۰/۳	۱۰۱۸	۱۱۳۵	۶
-۶/۱	۱۳۸۱	۱۴۷۰	۷
۳/۵	۱۷۲۶	۱۶۶۸	۸
-۲/۳	۲۲۵۰	۲۳۰۳	۹
۸/۱	۱۳۹۲	۱۲۸۷	۱۰
-۹/۷	۱۲۲۹	۱۳۶۱	۱۱
۰/۳	۲۲۲۵	۲۲۱۹	۱۲
-۵/۳	۱۰۱۹	۱۰۷۶	۱۳
۶/۴	۱۷۸۶	۱۶۷۸	۱۴
۷/۵	۱۲۶۷	۱۱۷۹	۱۵
۷/۳	۱۱۱۵	۱۰۳۹	۱۶

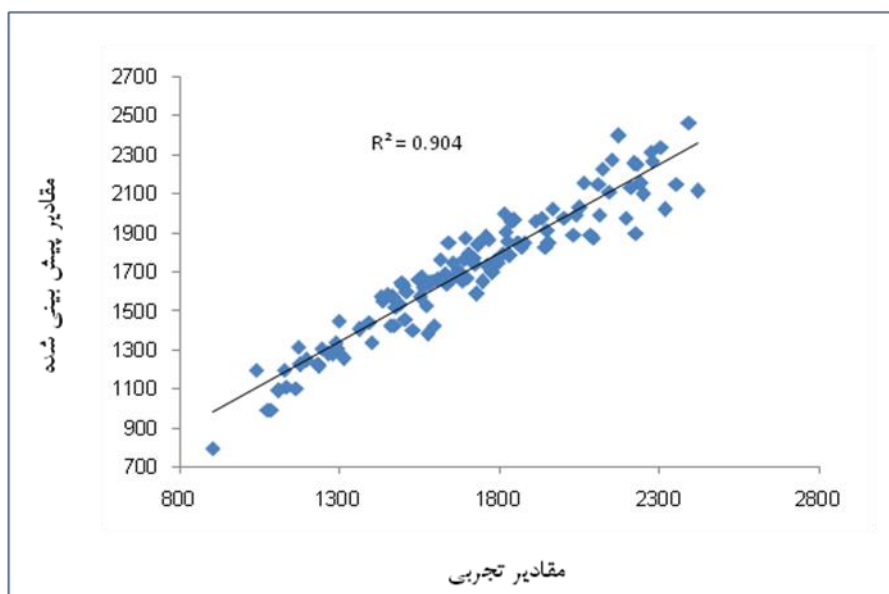
درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۴/۵	۱۲۴۹	۱۱۹۵	۱۷
۶/۹	۱۸۲۶	۱۷۰۸	۱۸
۸/۹	۱۸۵۳	۱۷۰۱	۱۹
۶/۳	۱۵۲۴	۱۴۳۳	۲۰
۴/۶	۱۵۶۳	۱۴۹۴	۲۱
۶/۵	۱۷۱۰	۱۶۰۶	۲۲
-۴/۹	۱۴۷۶	۱۵۵۲	۲۳
-۰/۴	۱۵۸۹	۱۵۹۶	۲۴
-۴/۲	۲۱۵۱	۲۲۴۶	۲۵
-۵/۵	۱۸۴۵	۱۹۵۳	۲۶
۶/۳	۱۳۱۰	۱۲۳۳	۲۷
۲/۵	۱۱۵۵	۱۱۲۶	۲۸
-۴/۴	۲۱۴۱	۲۲۳۹	۲۹
۵/۳	۱۵۲۵	۱۴۴۹	۳۰
۱۲/۹	۱۷۴۸	۱۵۴۸	۳۱
۱۲/۹	۱۷۹۸	۱۵۹۲	۳۲
-۶/۴	۱۰۱۵	۱۰۸۵	۳۳
۰/۳	۱۸۲۳	۱۸۱۷	۳۴
-۳/۷	۱۹۲۶	۲۰۰۱	۳۵
۵/۶	۲۲۶۳	۲۱۴۲	۳۶
۰/۷	۲۱۲۶	۲۱۱۲	۳۷
۸/۸	۲۱۳۹	۲۱۵۰	۳۸
-۱۰/۰	۲۱۱۷	۲۳۵۲	۳۹
۶/۹	۱۳۲۸	۱۲۴۳	۴۰
۵/۸	۱۹۵۱	۱۸۴۵	۴۱
۴/۰	۱۸۲۵	۱۷۵۵	۴۲
-۰/۹	۱۴۷۲	۱۴۸۶	۴۳
-۷/۴	۱۶۴۱	۱۷۷۲	۴۴
۵/۹	۱۷۲۳	۱۶۲۸	۴۵
-۰/۷	۱۷۷۲	۱۷۸۵	۴۶

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۱۶/۶	۹۷۰	۱۱۶۴	۴۷
-۰/۵	۱۵۲۱	۱۵۲۹	۴۸
-۲/۰	۱۴۷۹	۱۵۰۹	۴۹
۱/۹	۱۷۲۲	۱۶۹۰	۵۰
-۴/۶	۱۸۵۳	۱۹۴۳	۵۱
۱/۷	۱۷۱۱	۱۶۸۳	۵۲
-۳/۰	۱۶۶۷	۱۷۱۸	۵۳
-۶/۵	۱۹۵۵	۲۰۹۰	۵۴
۳/۵	۹۳۶	۹۰۴	۵۵
۶/۵	۲۳۱۲	۲۱۷۲	۵۶
-۲/۹	۱۹۷۲	۲۰۳۱	۵۷
۵/۵	۱۷۲۲	۱۶۳۳	۵۸
۰/۴	۲۲۲۰	۲۲۱۰	۵۹
-۰/۱	۱۲۶۲	۱۲۶۳	۶۰
-۶/۹	۱۷۳۱	۱۸۵۹	۶۱
-۱/۰	۱۸۱۱	۱۸۳۰	۶۲
۳/۴	۱۸۲۶	۱۷۶۶	۶۳
۴/۹	۱۸۰۸	۱۷۲۴	۶۴
۴/۳	۲۱۵۰	۲۰۶۲	۶۵
۲/۴	۱۴۲۶	۱۳۹۲	۶۶
-۱/۳	۲۲۴۲	۲۲۷۱	۶۷
-۹/۸	۲۰۰۷	۲۲۲۵	۶۸
-۲/۳	۱۷۰۴	۱۷۴۴	۶۹
۶/۴	۱۳۰۹	۱۲۳۱	۷۰
۶/۱	۱۶۵۱	۱۵۵۷	۷۱
۱/۵	۱۵۹۴	۱۵۷۰	۷۲
۵/۴	۱۷۴۱	۱۶۵۲	۷۳
۴/۶	۱۵۴۴	۱۴۷۶	۷۴
۶/۱	۱۹۲۳	۱۸۱۲	۷۵
۸/۴	۱۲۶۸	۱۱۷۰	۷۶

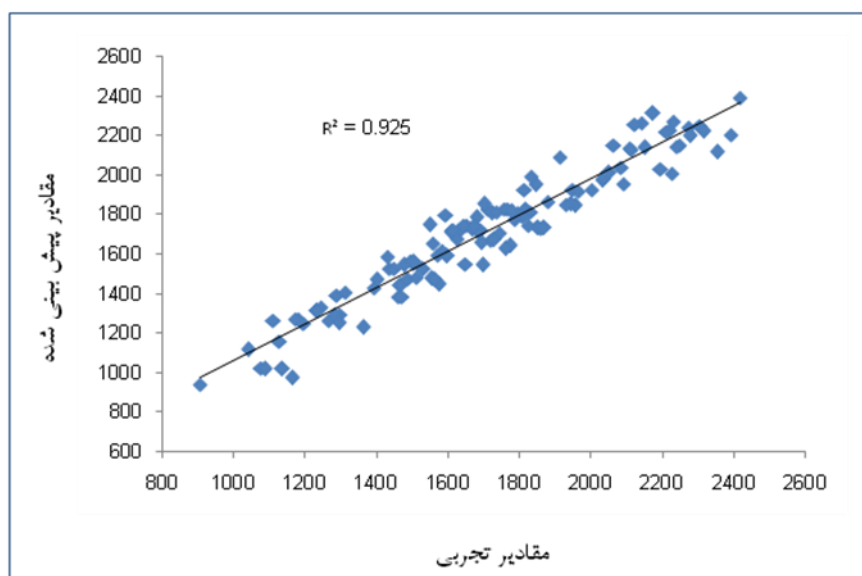
درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۳/۲	۱۶۷۷	۱۶۲۵	۷۷
-۹/۰	۱۵۴۴	۱۶۹۷	۷۸
-۶/۵	۱۷۳۱	۱۸۵۱	۷۹
۰/۴	۲۳۹۱	۲۳۸۱	۸۰
۴/۴	۱۸۱۰	۱۷۳۴	۸۱
-۷/۴	۲۰۳۲	۲۱۹۴	۸۲
۲/۲	۱۶۱۶	۱۵۸۱	۸۳
-۷/۸	۲۲۰۴	۲۳۹۰	۸۴
-۱/۲	۱۷۸۷	۱۸۰۹	۸۵
-۰/۵	۱۲۸۹	۱۲۹۶	۸۶
-۲/۳	۲۰۳۶	۲۰۸۴	۸۷
-۶/۰	۱۴۶۸	۱۵۶۲	۸۸
۰/۵	۱۲۸۷	۱۲۸۱	۸۹
۳/۷	۱۵۶۰	۱۵۰۴	۹۰
۰/۴	۱۸۰۰	۱۷۹۳	۹۱
۹/۲	۲۰۸۸	۱۹۱۲	۹۲
-۴/۰	۲۲۲۱	۲۳۱۴	۹۳
-۱/۳	۱۹۲۲	۱۹۴۸	۹۴
۸/۷	۱۹۹۲	۱۸۳۲	۹۵
۲/۱	۱۸۱۵	۱۷۷۷	۹۶
۶/۳	۲۲۵۴	۲۱۲۱	۹۷
۶/۶	۱۷۲۰	۱۶۱۳	۹۸
-۳/۴	۱۲۴۹	۱۲۹۳	۹۹
۱۱/۱	۱۵۸۶	۱۴۲۸	۱۰۰
-۴/۲	۱۸۴۸	۱۹۲۹	۱۰۱
۴/۱۹	۱۴۶۸	۱۴۰۰	۱۰۲
-۲/۳	۱۹۱۹	۱۹۶۴	۱۰۳
-۶/۱	۱۵۴۸	۱۶۴۸	۱۰۴
-۷/۲	۱۷۳۴	۱۸۶۷	۱۰۵
-۳/۴	۲۲۰۰	۲۲۷۶	۱۰۶

درصد خطای نسبی	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
۴/۳	۱۷۴۴	۱۶۷۳	۱۰۷
-۰/۷	۱۸۶۳	۱۸۷۷	۱۰۸
-۲/۴	۱۹۸۶	۲۰۳۶	۱۰۹
-۱/۹	۲۰۱۲	۲۰۵۰	۱۱۰
-۲/۰	۱۶۵۷	۱۶۹۱	۱۱۱
۶/۳	۱۷۴۵	۱۶۴۱	۱۱۲
-۷/۶	۱۶۲۷	۱۷۶۱	۱۱۳
-۴/۳	۱۷۴۳	۱۸۲۲	۱۱۴
۷/۲	۱۴۰۶	۱۳۱۱	۱۱۵
۱۳/۵	۱۲۶۰	۱۱۱۰	۱۱۶
-۱/۰	۲۳۹۲	۲۴۱۷	۱۱۷
-۳/۳	۱۶۷۲	۱۷۲۸	۱۱۸
۱/۲	۲۱۳۴	۲۱۰۸	۱۱۹

نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی داده‌ها نیز در ارزیابی دو مدل ANN و MLR به روش حذف مرحله‌ای تک‌تک داده‌ها، رسم گردید که نتایج به دست آمده در شکل‌های ۴-۱۵ و ۴-۱۶ آورده شده است. همانطور که مشاهده می‌گردد، روش ANN قدرت پیش‌گویی بهتری را در این ارزیابی نسبت به مدل خطی نشان می‌دهد.

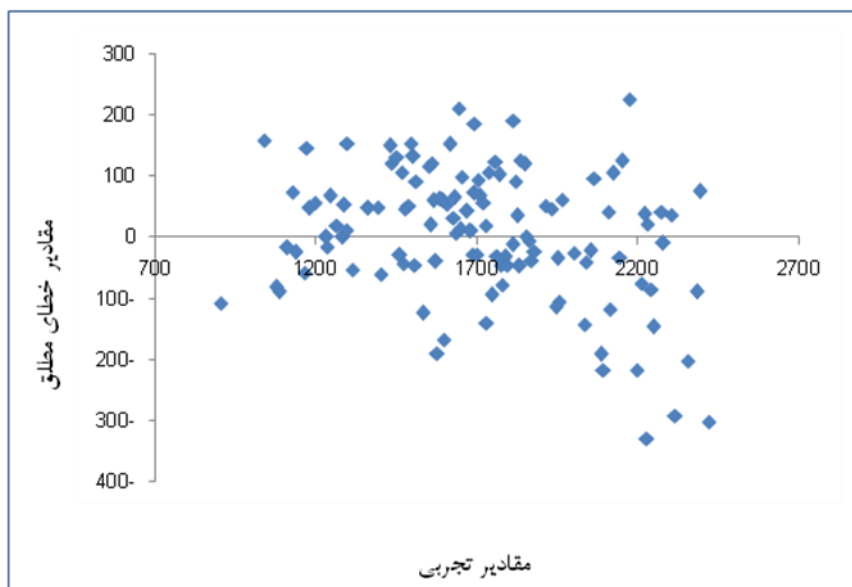


شکل ۴-۱۵- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل خطی به روش حذف مرحله‌ای تک‌تک داده‌ها

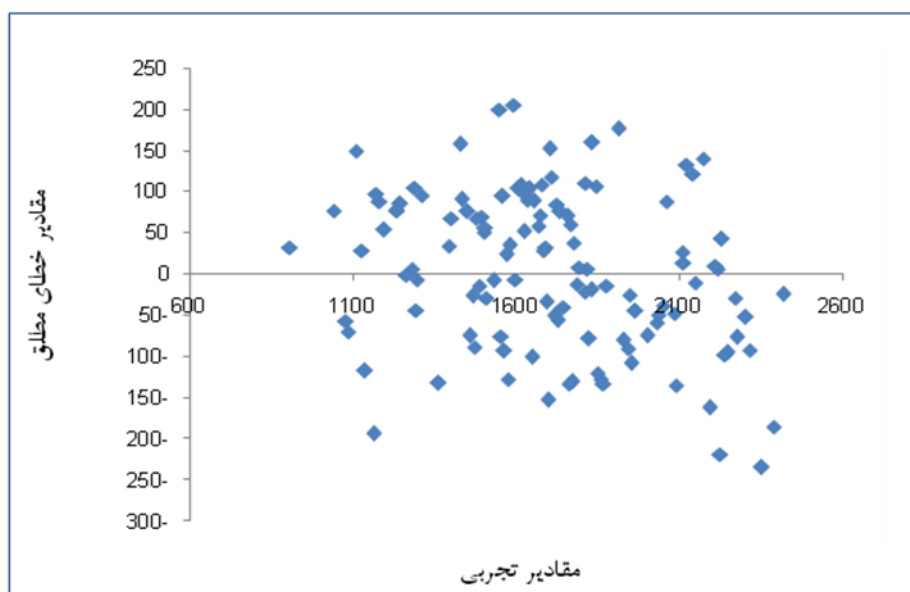


شکل ۴-۱۶- رسم مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در ارزیابی مدل ANN به روش حذف مرحله‌ای تک‌تک داده‌ها

همچنین نمودار مقادیر باقی مانده‌ی محاسبه شده بر حسب مقادیر تجربی شاخص بازداری، در ارزیابی مدل‌های ANN و MLR در شکل‌های ۱۷-۴ و ۱۸-۴ رسم شد. تقارن تقریبی داده‌ها در اطراف خط صفر نشان‌دهنده عدم وجود خطای معین می‌باشد.



شکل ۱۷-۴ - رسم مقادیر خطای مطلق در ارزیابی مدل MLR در مقابل مقادیر تجربی شاخص بازداری



شکل ۱۸-۴ - منحنی مقادیر خطای مطلق محاسبه شده بر حسب مقادیر اندیس بازداری در روش ANN

۴-۶-۳- ارزیابی شبکه به روش 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 جدول ۴-۲۰ مقادیر ضریب تعیین برای سری ارزیابی و تست آورده شده است. مقادیر پایین R^2 نشان‌دهنده‌ی عدم وجود همبستگی شانسی در مدل اصلی هستند.

جدول ۴-۲۰- مقادیر R^2 برای سری ارزیابی و تست در روش Y-Randomization

شماره تکرار آزمون	R^2 سری تست	R^2 سری ارزیابی
۱	۰/۰۲۵۴	۰/۰۱۷۲
۲	۰/۰۹۱۵	۰/۰۴۳۸
۳	۰/۱۶۳۲	۰/۱۱۷۰
۴	۰/۲۲۷۵	۰/۰۷۴۷
۵	۰/۰۴۱۰	۰/۰۵۲۳
۶	۰/۱۹۶۳	۰/۲۰۲۶
۷	۰/۳۳۰۷	۰/۱۵۴۶
۸	۰/۲۳۱۳	۰/۱۷۴۲
۹	۰/۰۴۸۷	۰/۰۰۳۳
۱۰	۰/۱۰۱۹	۰/۰۵۶۹

۴-۶-۴- ارزیابی مدل‌های برتر خطی و غیرخطی با استفاده از پارامترهای آماری

مطابق جدول ۴-۲۱ پنج پارامتر آماری، جهت ارزیابی توانایی پیش‌گویی مدل‌های ساخته شده به روش خطی (MLR) و غیرخطی (ANN) آورده شده‌اند. مقادیر این پارامترها نشان می‌دهد که مدل غیرخطی ارائه شده به مراتب از قدرت پیش‌گویی بهتری نسبت به مدل خطی برخوردار است.

جدول ۴-۲۱- مقایسه برخی پارامترهای آماری دو روش MLR و ANN

پارامتر	سری تست (n = ۲۴)		سری ارزیابی (n = ۲۴)		کل داده ها (n = ۱۱۹)	
	MLR	ANN	MLR	ANN	MLR	ANN
MAE	۸۳/۱۷۹	۷۷/۰۷۹	۶۹/۱۱۷	۷۱/۲۲۱	۸۳/۶۲۸	۷۸/۹۰۷
MSE	۱۰۱۲۲/۵۴۰	۸۱۲۰/۲۵۵	۹۲۳۱/۶۲۵	۷۲۱۹/۱۶۷	۱۱۳۶۸/۰۷۱	۸۸۹۰/۳۲۲
SEP	۱۰۰/۶۱۱	۹۰/۱۱۲	۹۶/۰۹۵	۸۴/۹۶۶	۱۰۶/۶۲۱	۹۴/۲۸۸
REP(%)	۶/۲۰۰	۵/۴۴۳	۵/۷۹۶	۵/۱۴۶	۶/۲۲۹	۵/۵۲۷
R ²	۰/۸۸۵	۰/۹۱۵	۰/۹۳۸	۰/۹۶۲	۰/۹۰۵	۰/۹۲۶

۴-۷- بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اندیس بازداری

ترکیبات

بهترین مدل انتخاب شده شامل ۸ توصیف‌کننده است که هر کدام بیان‌گر خصوصیات متفاوتی

از مولکول‌های مورد بررسی می‌باشد. این توصیف‌کننده‌ها عبارت بودند از:

Ss, MATS1e, BAC, RDF020m, MATS1p, R3v_A, MLOGP, Mor01v

۴-۷-۱- توصیف‌کننده Ss

این توصیف‌کننده از گروه توصیف‌کننده‌های ساختاری می‌باشد. توصیف‌کننده‌های ساختاری

ساده‌ترین و رایج‌ترین توصیف‌کننده‌ها هستند که منعکس‌کننده ساختار تشکیل‌دهنده یک ترکیب

می‌باشند. تعداد اتم‌ها، تعداد پیوندها، تعداد حلقه‌ها و وزن مولکولی از توصیف‌کننده‌های زیرمجموعه

این گروه هستند. این گروه از توصیف‌کننده‌ها به تفاوت‌های پیکربندی حساس نیستند و بنابراین بین

ایزومرهای فضایی یک ترکیب تفاوتی قائل نمی‌شوند. این توصیف‌کننده‌ها، صفربعدی یا یک‌بعدی

هستند.

حالت الکتروتوپولوژیکی اتم i ام در یک مولکول که با S_i نشان داده می‌شود، اطلاعاتی راجع به حالات توپولوژیکی و الکترونی اتم در مولکول در اختیار می‌گذارد و به صورت زیر تعریف می‌شود:

$$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^A \frac{I_i - I_j}{(d_{ij} + 1)^k} \quad \text{معادله (۲-۴)}$$

در معادله ۲-۴، I_i حالت الکتروتوپولوژیکی ذاتی اتم i ام و ΔI_i اثر میدان روی اتم i ام است که به صورت انحراف حالت الکتروتوپولوژیکی اتم i ام توسط سایر اتم‌های مولکول محاسبه می‌گردد. d_{ij} فاصله الکتروتوپولوژیکی بین i امین و j امین اتم است و A نیز تعداد اتم‌ها در مولکول است. K پارامتری است که اثر اتم‌های مجاور یا غیرمجاور را برای مطالعات خاص اصلاح می‌کند و معمولاً برابر با ۲ در نظر گرفته می‌شود. حالت الکتروتوپولوژیکی ذاتی i امین اتم به صورت زیر محاسبه می‌شود:

$$I_i = \frac{\left(\frac{2}{L_i}\right)^2 \cdot \delta_i^V + 1}{\delta_i} \quad \text{معادله (۳-۴)}$$

در معادله ۳-۴، L_i عدد کوانتومی اصلی (۲ برای اتم‌های C، N، O و F و ۳ برای Si، S، Cl و...)، δ_i^V تعداد الکترون‌های لایه ظرفیت و δ_i تعداد الکترون‌های سیگمای i امین اتم در گراف ملکولی تهی شده از هیدروژن است. حالت الکتروتوپولوژیکی ذاتی یک اتم را می‌توان به صورت ساده نسبت جفت الکترون‌های تنها و σ به تعداد پیوندهای π در گراف مولکولی، برای اتم مورد نظر بیان نمود. بنابراین حالت الکتروتوپولوژیکی ذاتی یک اتم، منعکس‌کننده امکان پخش اثر الکترون‌های غیر سیگما در اطراف اتم مورد نظر است. هرچه پخش شدن اثر الکترون کمتر باشد، الکترون‌های لایه ظرفیت برای برهمکنش‌های بین مولکولی بیشتر در دسترس هستند. از تعریف حالت الکتروتوپولوژیکی ذاتی و اثرات میدان می‌توان استنباط نمود که مقادیر مثبت و بزرگ S_i به اتم‌هایی با الکترون‌گاتیویته بالا و یا اتم‌های انتهایی مربوط می‌شوند. مقادیر کوچک یا منفی S_i به اتم‌هایی مربوط می‌شوند که فقط دارای الکترون‌های σ هستند و یا داخل مولکول محبوس شده و یا در نزدیکی اتم‌هایی با الکترون‌گاتیویته بالاتر قرار گرفته‌اند. بنابراین، حالت الکتروتوپولوژیکی منعکس‌کننده میزان در دسترس بودن الکترون‌های اتم می‌باشد و می‌توان آن را به صورت معیاری از احتمال برهمکنش با سایر مولکول‌ها نیز

تفسیر کرد. به همین دلیل انتظار می‌رود افزایش مقادیر این توصیف‌کننده باعث افزایش برهمکنش مولکول با فاز ساکن و افزایش شاخص بازداری گردد. اثر متوسط مثبت برای این توصیف‌کننده (۶/۸) این فرضیه را تأیید می‌کند [۴۳،۶۷].

۴-۷-۲- توصیف‌کننده‌های MATS1e و MATS1p

جزء توصیف‌کننده‌های خودارتباطی دو بعدی (2D autocorrelation) می‌باشند که توسط خواص فیزیکی- شیمیایی اتم وزن‌دار شده و بر اساس تابع خودارتباطی (AC_1) محاسبه می‌گردند. این تابع به صورت زیر محاسبه می‌گردد:

$$AC_1 = \int_a^b f(x).f(x+l).dx \quad \text{معادله (۴-۴)}$$

که $f(x)$ هر تابع از متغیر x است و l ، lag یا پس‌افت می‌باشد که نشان‌دهنده‌ی یک بازه از x است. a و b بازه‌ی کلی مورد مطالعه‌ی تابع را تعریف می‌کنند. تابع $f(x)$ معمولاً یا یک تابع وابسته به زمان (مانند یک سیگنال الکتریکی وابسته به زمان) و یا یک تابع وابسته به فضا (مانند چگالی جمعیت در فضا) است [۶۸]. MATS1e با الکترونگاتیویته‌ی ساندerson وزن‌دار شده است. تأثیر این توصیف‌کننده روی شاخص بازداری منفی (۵/۱۹۴-) است. MATS1p نیز با توانایی قطبش‌پذیری اتمی وزن‌دار گردیده که دارای تأثیر مثبت (۷/۶۶۸) بر اندیس بازداری می‌باشد.

۴-۷-۳- توصیف‌کننده BAC

توصیف‌کننده‌های توپولوژیکی نه تنها ارتباط بالایی با خواص مختلف مولکولی از خود نشان می‌دهند بلکه آنها ارتباط چشمگیری با دیگر توصیف‌کننده‌های مولکولی، به خصوص توصیف‌کننده‌های شیمیایی نظیر حجم مولکولی و خصلت چربی‌دوستی^۱ دارند [۷]. BAC یا اندیس مرکزی بالابان، یک

^۱- Lipophilicity

اندیس توپولوژیکی است که برای گراف‌های غیر حلقوی بر پایه اصلاح گراف^۱ تعریف می‌شود. که این اصلاح عبارت از برازش مرحله‌ای برای پاک‌سازی رئوس انتهایی مثلاً رأس‌های با درجه‌ی یک ($\delta_i = 1$) و زوایای تلاقی کننده متناظر از گراف مولکولی تهی از هیدروژن^{۳-۷-۱} G می‌باشد. n_g رئوس و زوایای پاک‌سازی شده در g امین مرحله می‌باشد و R تعداد کلیه‌ی مراحل برای پاک‌سازی تمامی رئوس است. BAC به صورت زیر برآورد می‌شود:

$$B = \sum_{g=1}^R n_g^2 \quad \text{معادله (۴-۵)}$$

این شاخص، امکان اندازه‌گیری شاخه‌دارشدگی مولکولی را فراهم می‌کند؛ بیشتر بودن مقدار B به معنی تعداد بیشتر شاخه در درخت یا tree مولکول است. اندیس مزبور به این دلیل مرکزی نامیده می‌شود که منعکس‌کننده‌ی توپولوژی درخت مولکولی از میانه‌ی آن می‌باشد [۶۹]. اثر توصیف‌کننده BAC روی شاخص بازداری ($-7/157$) کم و منفی بوده و نشان‌دهنده‌ی وجود تعداد نسبتاً کم شاخه در مولکول‌های مورد بررسی است.

۴-۷-۴- توصیف‌کننده RDF020m

این توصیف‌کننده ساختار سه بُعدی مولکول را توصیف می‌کند و در اینجا توسط جرم اتمی وزن‌دار شده است. توضیحات مربوط به توصیف‌کننده‌های گروه RDF در بخش (۳-۷-۳) آمده است.

۴-۷-۵- توصیف‌کننده R3v_A

این توصیف‌کننده، از گروه توصیف‌کننده‌های GETAWAY می‌باشد. دو نوع توصیف‌کننده‌ی GETAWAY طراحی شده‌اند، یک گروه از این توصیف‌کننده‌ها به نام H-GETAWAY از یک

۱- Pruning of the Graph

نمایش جدید از ساختار مولکولی به نام ماتریس تأثیر مولکولی^۱ 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. هرچه الکترون‌گاتیویته، اندازه اتمی و فاصله بین اتم و مرکز مولکول بیشتر باشد، مقدار این توصیف‌کننده بیشتر است.

گروه دیگر، توصیف‌کننده‌های R-GETAWAY هستند که R3v_A نیز از این نوع است و از ماتریس اثر-فاصله (R) گرفته شده‌اند و در آن عناصر ماتریس تأثیر مولکولی با عناصر ماتریس هندسی^۲ ترکیب شده‌اند [۴۳]. این توصیف‌کننده با حجم‌های واندروالس اتمی وزن دار شده است و اثر آن روی شاخص بازدارنده منفی (۵/۴۷۹-) است.

۴-۷-۶- توصیف‌کننده MLOGP

MLOGP یا ضریب توزیع آب-اکتانول (جدول ۴-۳) بیان‌گر ضریب توزیع مورینگوچی^۳ می‌باشد [۷۰]. این توصیف‌کننده، از گروه توصیف‌کننده‌های ویژگی^۴ است که به طور خاص، در اینجا ویژگی چربی‌دوستی را مطرح می‌کند. چربی‌دوستی، بیانی از توزیع یک ترکیب بین دو فاز آلی و آبی بوده و معمولاً با علامت P نمایش داده می‌شود. P به عنوان نسبت غلظت‌های یک ترکیب در فازهای آلی و آبی یک سیستم دو جزئی تحت شرایط تعادل تعریف می‌شود:

$$P = \frac{[C]_{org}}{[C]_{aq}} \quad \text{معادله (۴-۶)}$$

که در آن، $[C]_{org}$ و $[C]_{aq}$ به ترتیب غلظت‌های حل‌شونده در فازهای آلی و آبی هستند. ضرایب توزیع معمولاً به صورت لگاریتمی بیان می‌شوند:

۱- Molecular Influence Matrix

۲- Geometry Matrix

۳- Muriguchi Octanol-water Partition Coefficient (logP)

۴- Properties

$$\log P = \log \frac{[C]_{org}}{[C]_{aq}} = \log [C]_{org} - \log [C]_{aq} \quad \text{معادله (۷-۴)}$$

به منظور اجتناب از تجمعات ممکن حل‌شونده در فاز آلی، ضرایب توزیع باید در غلظت‌های پایین یا به صورت برون‌یابی شده در یک غلظت رقیق شده از حل‌شونده اندازه‌گیری شوند. اندازه‌گیری‌های بسیاری از تمایل نسبی مولکول، با توجه به دو فاز و بسته به پدیده‌های جذب، انتقال و توزیع انجام گرفته است. ترکیباتی که $P > 1$ یا $\log P > 0$ چربی‌دوست یا لیپوفیل نام دارند؛ و ترکیباتی که $P < 1$ یا $\log P < 0$ هیدروفیل یا آب‌دوست می‌باشند. به طور خاص، چربی‌دوستی به اندازه‌ی حل‌شونده، قطبیت آن و اثرات پیوند هیدروژنی بستگی دارد. توصیف‌کننده‌ی مولکولی که بیشترین استفاده را در مورد این خاصیت دارد، ضریب توزیع آب-اکتانول است که با K_{ow} نشان داده می‌شود. به این ترتیب، ضریب توزیع میان ۱-اکتانول و آب به شکل زیر بیان می‌گردد:

$$\log P = \log K_{ow} = \log \frac{[C]_{1-octanol}}{[C]_{water}} = \log [C]_{1-octanol} - \log [C]_{water} \quad \text{معادله (۸-۴)}$$

ضرایب توزیع مورد استفاده‌ی دیگر، به طور معمول برای سیستم‌های n -آلکانی تعریف می‌شوند [۷۱].

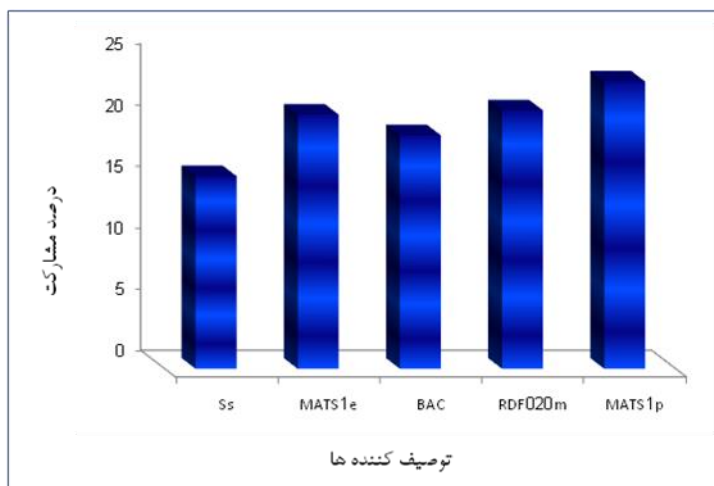
۷-۷-۴- توصیف‌کننده Mor01v

این توصیف‌کننده، جزء گروه 3D-MoRSE است که هدف آنها، به دست آوردن اطلاعات از مختصات اتمی سه بعدی با استفاده از انتقال به کار رفته در مطالعات پراش الکترونی به منظور فراهم آوردن منحنی‌های نظری تفرق می‌باشد (بخش ۳-۷-۲). این توصیف‌کننده با حجم‌های واندروالس اتمی وزن دار شده است و اثر آن روی شاخص بازداري ۴/۱۳۱ است.

۸-۴- بررسی میزان مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی

بر اساس روشی که در بخش ۸-۳ بیان شد، میزان مشارکت توصیف‌کننده‌های انتخاب شده

در پیش‌بینی اندیس بازداري موردنظر برآورد گردید که نتایج آن در شکل ۴-۱۹ ارائه شده است.



شکل ۴-۱۹- نمایش میزان مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینه

بر اساس شکل ۴-۱۹، توصیف‌کننده MATS1p بیشترین اثر مشارکت را دارا می‌باشد. توصیف‌کننده‌های دیگری که درصد مشارکت بالایی دارند، توصیف‌کننده RDF020m و MATS1e هستند.

با توجه به تأثیر مثبت توصیف‌کننده MATS1p بر شاخص بازداري، مشخص می‌گردد که بین این دو رابطه مستقیم وجود داشته که با افزایش این توصیف‌کننده، اندیس بازداري زیاد می‌شود در حالی که میان توصیف‌کننده MATS1e و پارامتر تجربی رابطه عکس برقرار است. توصیف‌کننده RDF020m نیز اطلاعات مهمی درباره فواصل بین اتمی در کل مولکول، فاصله‌ی پیوند، نوع حلقه‌ها، سیستم‌های مسطح یا غیرمسطح و... به دست می‌دهد. مقدار مثبت اثر این توصیف‌کننده بر اندیس بازداري، بیان‌کننده تأثیر مستقیم آن بر پارامتر تجربی می‌باشد.

۴-۹- نتیجه‌گیری

با توجه به تنوع گسترده روغن‌های ضروری و کاربرد وسیع آنها مطالعه ترکیبات این روغن‌ها در زمینه‌های مختلف امری اجتناب ناپذیر به نظر می‌رسد. از آنجا که اندازه‌گیری و بررسی این مواد و ترکیبات مشابه آنها به صورت تجربی و آزمایشگاهی وقت‌گیر و هزینه‌بر است، بنابراین با صرف زمان و هزینه‌ی کمتر می‌توان با استفاده از مدل‌های معتبر خطی و غیر خطی شاخص بازداري را برای یک دسته از روغن‌های ضروری با درصد اطمینان مشخص تخمین زد و سپس آن را برای سری مشابهی از ترکیبات که تاکنون سنتز نشده‌اند پیش بینی نمود.

آینده‌نگری

✦ از آنجا که اندیس بازداري ترکیبات مختلف^۱ در شرایط متفاوتی از ستون کروماتوگرافی نظیر انواع گوناگون فازهای ثابت یا برنامه‌های دمایی متفاوت ستون اندازه‌گیری می‌شود؛ لذا برای مدل‌سازی اندیس بازداري یک سری ترکیب خاص مثل روغن‌های ضروری و تعریف دسته‌ای از توصیف‌کننده‌ها برای آنها، بررسی شرایط ستون اندازه‌گیری‌کننده لازم است؛ در عین حال نتایج مطالعات مختلفی که در این زمینه انجام پذیرفته است، تا حدی قابلیت تعمیم و مقایسه با یکدیگر را دارند. با توجه به موارد گفته شده، از نتایج این تحقیقات می‌توان برای پیش‌بینی شاخص بازداري ترکیباتی که در این مدل‌سازی به کار نرفته یا حتی هنوز سنتز نشده‌اند، استفاده نمود.

✦ در این تحقیق برای ساختن مدل خطی از روش رگرسیون خطی چندگانه استفاده شده و سپس توصیف‌کننده‌های حاصل از آن به عنوان ورودی به شبکه عصبی مصنوعی خورانده شدند. چون در روش‌های رگرسیون خطی چندگانه تنها برهم‌کنش‌های خطی بین پارامترها لحاظ می‌شود، بنابراین برای انتخاب توصیف‌کننده‌های مهم می‌توان از روش‌های دقیق‌تری مثل الگوریتم ژنتیک، بهینه‌سازی اجتماع مورچگان^۱ و طرح‌ریزی پیاپی^۲ بهره برد.

✦ می‌توان از نرم‌افزارهایی با قابلیت‌های بیشتر از HypeChem برای رسم ترکیبات یا Dragon به منظور محاسبه‌ی توصیف‌کننده‌ها بهره گرفت.

۱- Ant Colony Optimization

۲- Successive Projection Algorithm (SPA)

✦ می‌توان از روشهای خطی دیگر مانند حداقل مربعات جزئی (PLS) و آنالیز اجزای اصلی (PCA) به جای روش رگرسیون خطی چندگانه استفاده نمود.

✦ روشهایی مانند KPLS, poly-PLS, SVM و LS-SVM^۱ را می‌توان به جای شبکه عصبی مصنوعی یا در کنار آن برای مدل‌سازی غیرخطی استفاده کرده و نتایج را با هم مقایسه نمود.

^۱- Least Square Supported Vector Machine

فهرست منابع

- [۱]- Manssnat D. L., Vandeginste B. G., Deming S. N. and Kaufman L., (1998), "*Chemometrics, A Text Book*", Elsevier, Amsterdam.
- [۲]- Burt s., (2004), "Essential Oils: Their Antibacterial Properties and Potential Application in Foods- a Review" *International J. Food Microb.*, 94, pp 223-253.
- [۳]- Sellar W., (2001), " *The Directory of Essential Oils*" (Reprinted.). Essex: The C.W. Daniel Company, Ltd.
- [۴]- Simionatto E., Bonani V. F. L., Morel A. F., Poppi N., Junior J. L. R., Stuker C. Z., Peruzzo, G. M., Peres M. T. L. P. and Hess S. C., (2007), " Chemical Composition and Evaluation of Antibacterial and Antioxidant Activities of the Essential Oil of Croton Urucurana Baillon (Euphorbiaceae) Stem Bark", *J. Braz. Soc.*, 18, No. 5, pp 879-885.
- [۵]- Choi H. and Sawamura M., (2000), "Composition of the Essential Oil of Citrustamurana Hort.ex Tanaka (Hyuganatsu)", *J. Agric. FoodChem.*, 48, pp 4868-4873.
- [۶]- روود، دین، (۱۳۸۳)، "روش‌های کروماتوگرافی: راهنمای عملی استفاده، نگهداری و رفع عیب از دستگاه‌های کروماتوگرافی گازی با ستون موئینه" امامی س. رحیمی ح.، چاپ اول، انتشارات نورپردازان تهران، ص ۵۲.
- [۷]- Hemmateenejad B., Javadnia K. and Elyasi M., (2007), " Quantitative Structure-Retention Relationship for the Kovats Retention Indices of a Large Set of Terpenes: A combined Data Splitting-Feature Selection Strategy", *Anal. Chim. Acta*, 592, pp 72-81.
- [۸]- Riahi S. Ganjali M. R., Pourbasheer S., Norouzi P.,(2008), "QSRR Study of GC Retention Indices of Essential Oil Compounds by Multiple Linear Regression with a Genetic Algorithm", *Chromatographia*, 67, pp 917-922.

- [٩]- Riahi S., Ganjali M. R., Pourbasheer S., Norouzi P., (2009), "Investigation of Different Linear and Nonlinear Chemometric Methods for Modeling of Retention Index of Essential Oil Components: Concerns to Support Vector Machine" *J. Hazardous Materials* ,166 , pp 853–859.
- [١٠]- Noorizadeh H., Farmany A., (2010), "QSRR Models to Predict Retention Indices of Cyclic Compounds of Essential Oils", *Chromatographia*,72, pp 563-569.
- [١١]- Noorizadeh H., Farmany A., (2010), "Investigation of Retention Behaviors of Essential Oils by Using QSRR", *J. Chinese Chem. Soc.*, 57, pp 982-991.
- [١٢]- Noorizadeh H., Farmany A., Khosravi A., (2010), "Exploration of Linear and Nonlinear Modeling Techniques to Predict of Retention Index of Essential Oils", *J. Chinese Chem. Soc.*, 57, pp 1268-1277.
- [١٣]- Noorizadeh H., Farmany A., Noorizadeh M., (2011), "Application of GA-PLS and GA-KPLS Calculations for the Prediction of the Retention Indices of Essential Oils ", *Quim. Nova*, XY, pp 1-7.
- [١٤]- Wold S. (1995), "Chemometrics; What Do We Mean with It and What Do We Want from It", *J. Chemolab.*, 30, pp 109-115.
- [١٥]- <http://www.wordiq.com/definition/chemometrics>
- [١٦]- Brereton R.G. (1990), "Chemometrics", Ellis Horwood, Chichester, pp 308.
- [١٧]- Todeschini R. and Consonni V. (2000), "Handbook of Molecular Descriptors", John Wiley-VCH, New York, pp 59.
- [١٨]- Howery D.G., Hirsch R.F., (1983), "Chemometrics in the Chemistry Curriculum", *J. Chem. Ed.* 60, pp 656-659.
- [١٩]- Brown S. D., "Has the 'Chemometrics Revolution' Ended? Some Views on the Past, Present and Future of Chemometrics", *papers on Chemometrics: Philosophy, History and Directions of InCINC'94*. <http://www.emsl.pnl.gov:2080/docs/hncinic/homepage.html>
- [٢٠]- Hemmatinejad B. (2005), "Chemometrics in Iran", *Chemometrics and Intelligent Laboratory Systems*, 81, pp 202-208.

- [۲۱]- Jalali-Heravi M., Garakani-Nejad Z., (1993), "Prediction of Gas Chromatographic Retention Indices of Some Benzene Derivatives", *J. Chroma.*, 648, pp 389-393.
- [۲۲]- Charton M., Charton B. (2002), "Advances in Quantitative Structure-Property Relationship", University of Bristol, UK, 3, pp 140.
- [۲۳]- Katritzky A. R., Maran U., Lobanov V. S., Karelson M., (2000), "Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties", *J. Chem. Inf. Comp. Soc.*, 40, pp 1-18.
- [۲۴]- Katritzky A. R., Dobchev D. A., Karelson M. (2006), "Physical, Chemical and Technological Property Correlation with Chemical Structure: Potential of QSPR", *Z. Naturforsch. B.*, 61, pp 373-384.
- [۲۵]- Gharagheizi F., Sattari M., (2008), "Prediction of Some Important Physical Properties of Sulfur Compounds Using QSPR Models", *Mol. Divers.*, 12, pp 143-155.
- [۲۶]- Todeschini R. and Consonni V. (2000), "Handbook of Molecular Descriptors", John Wiley-VCH, New York, pp 98.
- [۲۷]- HyperChem8.0 for Windows, (2007), Toronto, Canada: Hypercube Inc, <http://www.hyper.com>.
- [۲۸]- www.iasbs.ac.ir/chemistry/Chemometrics/.../8th/qsar_introduction.ppt
- [۲۹]- Todeschini R. and Consonni V. (2000), "Handbook of Molecular Descriptors", John Wiley-VCH, New York, pp 307.
- [۳۰]- Arab Chamjangali M., (2009), "Modeling of Cytotoxicity Data (CC₅₀) of Anti-HIV 1-[5-Chlorophenyl) Sulfonyl]-1H-Pyrrole Derivatives Using Calculated Molecular Descriptors and Levenberg- Marquardt Artificial Neural Network", *J. Chem. Bio. Drug. Des.*, 73, pp 456-465.
- [۳۱]- Habibi A., Danandeh M., (2007), "Prediction of Acidity Constant for Substituted Acetic acids in Water Using Artificial Neural Networks", *Indian J. Chem.*, 46B, pp 478-487.
- [۳۲]- Topliss J. G. and Edward R. P., (1979), "Chance Fraction in Studies of Quantitative Structure-Activity Relationships", *J. Med. Chem.*, 22, pp 1224-1238.

[۳۳]- اشرفی م.، (۱۳۸۹)، پایان نامه کارشناسی ارشد: "مطالعه ارتباط کمی ساختار- فعالیت مشتقات تیوکربامات ها به عنوان دسته ی جدیدی از بازدارنده های غیر نوکلئوزیدی HIV"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[۳۴]- SPSS for windows, (Aug. 23, 2008), Statistical Package for IBM PC, SPSS Inc., <http://www.SPSS.com>.

[۳۵]- فردوسی م. ع.، (۱۳۸۹)، پایان نامه کارشناسی ارشد: "پیش بینی ثابت های هنری بعضی از ترکیبات آلی با استفاده از روش های خطی و غیر خطی QSPR"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[۳۶]- جنی ر.، بینز و.، "هوش مصنوعی از الف تا ی"، بیگدلی قمی س.، محسنی م.، بدیع ک.، دفتر تحقیقات یاسین.

[۳۷]- www.psychology.wikia.com

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

[۳۹]- منهاج م.، (۱۳۸۷)، "مبانی شبکه های عصبی (هوش محاسباتی)"، جلد اول، چاپ پنجم، مرکز نشر دانشگاه صنعتی امیرکبیر، تهران.

[۴۰]- Zupan J. and Gasteiger J., (1993), "Neural Networks for Chemists", VCH Publishers, New York, Chapter 2.

[۴۱]- کیا م.، (۱۳۸۷)، "شبکه های عصبی در MATLAB"، چاپ دوم، انتشارات کیان رایانه سبز.

[۴۲]- هاگان ت. ه.، دیموث ه.، بیل م.، "طراحی شبکه های مصنوعی"، کیا م.، انتشارات کیان رایانه سبز.

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

[۴۴]- <http://foram.takdownload.ir/threads/25029؟-چيست-شبکه-های-مصنوعی>

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

- [۴۶]- Kermani B. G., Schiffman S. S., Troy Nagle H., (2005), "Performance of the Levenberg-Marquardt Neural Network Training Method in Electronic Nose Applications", *sensors and Actuators B*, 110, pp 13-22.
- [۴۷]- Wang X. G., Tang Z., Tamura H., Ishii M., (2004), "A Modified Error Function for the Back Propagation Algorithm", *Neurocomputing*, 57, pp 477- 484.
- [۴۸]- Hagan M. T. and Menhaj M., (1994), "Training Feed forward Networks with the Marquardt Algorithm", *IEEE Trans. Neural Networks*, 5, No. 6, pp 989-993.
- [۴۹]- Jalali-Heravi M., Mani-Varnosfaderani A., (2009), "QSAR Modeling of 1-(3,3-diphenylpropyl)-Piperidinylamides as CCR₅ Modulators Using Multivariate Adaptive Regression Spline and Bayesian Regularization Genetic Neural Networks", *J. QSAR Comb. Sci.*, 9, pp 946-958.
- [۵۰]- Caballero J., Fernandez M., (2006), "Linear and Nonlinear Modeling of Antifungal Activity of Some Heterocyclic Ring Derivatives Using Multiple Linear Regression and Bayesian-Regularized Neural Networks"; *J. Mol. Model*, 12, pp 168-181.
- [۵۱]- Caballero J., Fernandez L., Abreu J. I., Fernandez M., (2006), "Aminoacid Sequence Autocorrelation Vectors and Ensembles of Bayesian-Regularized Genetic Neural Networks for Prediction of Conformational Stability of Human Lysozyme Mutant", *J. Chem. Inf. Model.*, 46, pp 1255-1268.
- [۵۲]- Rawlings J. O., (1988), "Applied Regression Analysis", Wadsworth&Brooks/Cole, Pacific Grove (CA).
- [۵۳]- Allen D. M., (1971), "Mean Square Error of Prediction as a Criterion for Selecting Variables", *Technometrics*, 13, pp 469-475.
- [۵۴]- غفاری‌راد ح.، (۱۳۸۸)، پایان‌نامه کارشناسی ارشد: "مدل‌سازی QSPR فاکتور بازداري ترکیبات آلي بر روی فازهای ثابت تری‌فلوئوروپروپیل"، دانشکده شیمی، دانشگاه صنعتی شاهرود.
- [۵۵]- 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)", *J. Cemom.*, 6, pp 335-346.

- [Δϫ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 461.
- [ΔΥ]- Todeschini R., (2001), Milano Chemometrics and QSAR Group, Italy, Dragon software version 2.1, <http://www.disat.unimib.it/vhml>.
- [ΔΛ]- 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.
- [ΔϹ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 476.
- [ϫ•]- <http://Bioinformatics.oxfordjournals.org>
- [ϫ\]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 499-501.
- [ϫΥ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 132.
- [ϫϣ]- http://www.strandls.com/sarchitect/documents/manual_html/desctheory.html
- [ϫϤ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 513-514.
- [ϫΔ]- Schuur J. H., Selzer P., Gasteiger J., (1996), “The Coding of the Three Dimensional Structure of Molecules by Molecular Transform and Its Application to Structure-Spectra Correlation and Studies of Biological Activity”, *J. Chem. Inf. Comput. Sci* , 36, pp 334-344.
- [ϫϤ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 366-367.
- [ϫΥ]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 159-163.

[୧୮]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 17-20.

[୧୯]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 42.

[୧୦]- Moriguchi I., Hirano S., Liu Q., Nakagome I., Matsushita Y., (1992), “Simple Method of Calculating Octanol/Water Partition Coefficient”, *Chem. Pharm. Bull.*, 40, pp 127-130.

[୧୧]- Todeschini R. and Consonni V. (2000), “*Handbook of Molecular Descriptors*”, John Wiley, New York, pp 269-270.

Abstract

Herbaceous essential oils and related compounds have a broad application as for flavoring food, folk medicine and perfume and medical industries. In this project, the quantitative structure-property relationship studies was taken for retention indices of 77 and 119 essential oils in first and second parts, respectively. For modeling the retention indices of these compounds, two methods containing linear (multiple linear regression; MLR) and nonlinear (artificial neural network; ANN) were applied. The stepwise regression method was used for descriptor selection. The models were validated using external test set, as well as leave one out and Y-Randomization techniques. Both linear and nonlinear methods have good prediction ability, although ANN model has more accurate results. The correlation coefficient (R) of the test set obtained by MLR and ANN models in the first part of this study were 0.975 and 0.982 respectively. Also in the second part, the correlation coefficient of the test set obtained by MLR and ANN models were 0.941 and 0.956 respectively.

Keywords: Quantitative structure-property relationship (QSPR), Multiple linear regression (MLR), Artificial neural network (ANN), Essential oil, Retention index (RI)



Shahrood University of Technology

Faculty of Chemistry

Prediction of Retention Indices of Some Essential Oils Using Linear and Nonlinear QSPR Methods

Hanie Yazdandoost

Supervisor:

Dr. N. Goodarzi

Advisor:

Dr. Z. Kalantar

Date: July – 2011