JACR

بررسی معیارهای متفاوت برای منظم کردن اجزاهای اصلی به منظور ایجاد یک مدل QSPR برای پیش بینی نقطه های ذوب

ولی زارع شاه آبادی[1,*] و فاطمه عباسی تبار[2]

۱- دانشکده شیمی و مهندسی شیمی، دانشگاه آزاد اسلامی واحد ماهشهر، ماهشهر، ایران

۲- دانشگاه آزاد اسلامی واحد مرودشت، مرودشت، ایران

چکیده: براساس اهمیت پیش بینی نقطه های ذوب ترکیب ها، در این مقاله سعی شد که برای گروه وسیعی از ترکیب ها مدل مناسبی که توانایی پیش بینی نقطه های ذوب را در حد مطلوبی داشته باشد، ارایه شود. برای این منظور ۴۱۷۳ ترکیب شیمیایی با ساختارهای متنوع گزارش شده در مقاله های قبلی، انتخاب و برای توصیف ساختار آن ها از یک دسته ۲۰۲ تایی از توصیف‌کننده های 2D و 3D استفاده شد. این دسته داده ها به دو دسته آموزش و دسته تست به ترتیب با اندازه های ۲۵۷۳ و ۱۶۰۰ تقسیم بندی شد. ارزیابی بیشتر مدل ایجاد شده به کمک یک دسته داده دیگر با اندازه ۲۷۷ صورت گرفت. برای کاهش حجم ماتریس توصیف کننده ها از تجزیه اجزای اصلی کمک گرفته شد و از شبکه عصبی برای ایجاد مدل استفاده شد. بردارهای ویژه به دست آمده از تجزیه اجزای اصلی بر اساس معیارهای متفاوتی مرتب و به عنوان ورودی شبکه مورد استفاده قرار گرفتند. معیارهای استفاده شده عبارت بودند از بزرگی مقدار ویژه، میزان همبستگی با نقطه ذوب و میزان قدرت پیشبینی کنندگی. بهترین مدل در حالتی بدست آمد که بردارهای ویژه براساس میزان قدرت پیش بینی کنندگی اشان مرتب و به عنوان ورودی استفاده بودند. در پایان پارامترهای شبکه از قبیل تعداد لایه های پنهان، تعداد گره در هر لایه، سرعت یادگیری و مومنتوم بهینه سازی شدند. شبکه با ساختار [۲۵ ۱۰ ۸ ۱] و سرعت یادگیری ۰/۷ و مومنتوم ۰/۱۶ به عنوان شبکه بهینه انتخاب شد.

واژه های کلیدی: ارتباط کمی میان ساختار و خاصیت؛ آنالیز اجزای اصلی؛ مرتب سازی اجزاهای اصلی؛ شبکه عصبی مصنوعی

مقدمه

به دست آوردن یک مدل کمی که ارتباط میان ساختار شیمیایی ترکیب های و ویژگی یا فعالیت آنهـــا را مشخص می سازد (QSPR/QSAR)[1] در شیمی و بیوشیمی بسیار مهم و حایز اهمیت است. تهیه یک مـــدل QSPR دارای چند مرحـــله است [۱]:

۱- انتخاب دسته کاملی از مولکولها و گردآوری اطلاعات در مورد ویژگی آنها. این انتخاب بسیار مهم بوده و کیفیت مدل QSPR به دست آمده وابسته به آن می باشد. هر چه تعداد مولکول هایی که در ساخت مدل به کار می روند بیشتر باشند، مدل به دست دارای دامنه کاری وسیع تری است و می توان از آن برای پیش بینی خاصیت تعـــداد بیشتر از مولکول ها استفاده نمود. البته باید توجه داشت که

*عهده دار مکاتبات: E mail: valizare@gmail.com

اگر مولکول ها دارای تنوع ساختاری زیادی باشند و یا به عبارت دیگر متعلق به دسته های مختلفی باشند، ممکن است مدل به دست آمده کیفیت مناسبی نداشته باشد. ۲-محاسبه توصیف کننده های ساختار مولکولی. ۳-انتخاب[1] بهترین توصیف کننده ها و ایجاد مدل مناسب میان ویژگی مولکولی موردنظر و این توصیف کننده ها. ۴-ارزیابی مدل به دست آمده.

شاید به توان مشکلترین مرحله در بدست آوردن یک مدل QSPR را انتخاب توصیف کننده های مناسب دانست. همین قضیه سبب آن شده که محققین بسیاری برای حل این مشکل وقت صرف کنند و مقاله های زیادی در همین راستا به چاپ برسد [۲-۹۸۷۶۵۴۳]. آنالیز اجزاهای اصلی (PCA) یک[2] راه حل برای این مشکل است [۱]. در این روش از ترکیب خطی متغیرهای قبلی (توصیف کننده های مولکولی)، متغیرهای جدیدی به نام اجـــزاهای اصلی استخـــراج می شوند به نحوی که کل اطلاعات در تعداد اندکی از این اجزاها ذخیره می شود. دو نکته قابل ذکر آن است که اول اینکه اجزاهای اصلی بردارهای ویژه ماتریس توصیف کننده ها هستند و هیچ وابستگی بین آنها وجود ندارد زیرا آنها در فضا بر یکدیگر عمود هستند و دوم اینکه آنکه میزان اطلاعات ذخیره شده در اجزاهای اصلی از اولین آنها به دومین و الی آخر کاهش مییابد. با اینکه میزان اطلاعات ذخیره شده در اولین بردار ویژه بیشتر از دومین آنها است اما ممکن است که این اطلاعات هیچ وابستگی با خاصیت مورد مطالعه نداشته باشد و لذا این پرسش مطرح می شود که باید از کدام بردارهای ویژه[3] و با چه ترتیبی برای ساخت مدل QSPR استفاده کرد تا بهترین مدل بدست آید؟

چندین معیار برای مرتب کردن بردارهای ویژه در نظر گرفته شده است [۰۱]. یکی از این معیارها مقدار ویژه[4] هر جزء اصلی است. بردارهایی که دارای مقدار ویژه بزرگتری هستند حاوی اطلاعات بیشتری میباشند. بر همین اساس جزء نخست بزرگترین مقدار ویژه را دارد. متاسفانه همانطور که اشاره شد ممکن است که اطلاعات ذخیره شده در یک بردار ویژه ارتباطی با خاصیت مورد مطالعه نداشته باشد و لذا به کار بردن آن در مدل باعث تضعیف مدل می شود. معیار دیگری که استفاده شده میزان همبستگی میان یک بردار ویژه با ویژگی مورد بحث است. در اینجا هر چه میزان همبستگی بیشتر باشد، بردار ویژه مربوطه با اهمیت تر بوده و لذا زودتر وارد مدلسازی می شود.

نقطه ذوب یکی از ویژگیهای شاخص یک ماده شیمیایی است. نقطه ذوب دمایی است که در آن فاز جامد یک ترکیب با فاز مایع آن در فشار اتمسفری در حال تعادل است. برای تشخیص یک ماده و اثبات خلوص آن نقطه ذوب معیار مناسبی است. بسیاری از ویژگی فیزیکی و شیمیایی ماده، مانند حلالیت و نقطه جوش، با نقطه ذوب ارتباط نزدیکی دارند [۱۱-۳۱۲۱]. میزان حلالیت یک ماده در صنایع داروسازی بسیار مهم بوده زیرا که میزان در دسترس بودن یک دارو در بدن را مشخص می کند. نقطه جوش یک ماده نیز در مقوله های مربوط به محیط زیست اهمیت پیدا میکند. بنابراین اندازه گیری نقطه ذوب بسیار با مهم است. اما باید به این مطلب توجه داشت که تعیین نقطه ذوب تمام ترکیبات از طریق تجربی امکانپذیر نیست زیرا که بعضی از ترکیب ها پایداری حرارتی مناسبی نداشته و قبل از ذوب شدن تجزیه می شوند. به همین دلیل پیش بینی نقطه ذوب از طریق یک مدل QSPR یکی از مقوله های بسیار مهم است.

بررسی اثر ساختار مولکولی روی نقطه ذوب در چندین مقاله انجام گرفته است [۴۱-۰۲۹۱۸۱۷۱۶۱۵۱]. بیشتر این مطالعات روی دسته خاصی از ترکیبهای شیمیایی صورت گرفته است. سیگماورا و همکارانش معادلهای را برای بیش بینی دسته کوچکی از ترکیبهای آروماتیک غیر قابل انعطاف به دست آوردند [۱۲]. کرزیزیانیک[5] و همکارانش معادله ای را برای بیش بینی نقطه ذوب ترکیبات آلیفاتیک بدون پیوند هیدروژنی بدست آوردند [۲۲]. در سال ۱۹۹۴ یالکووسکی[6] و همکارانش رابطه ای بین نقطه ذوب و نقطه جوش حدود ۱۰۰۰ ترکیب بدون پیوند هیدروژنی برقرار کردند. ریشه توان دوم خطا در حدود ۳۶ °C گزارش شده است [۳۲]. کاترتزکی و همکارانش [۱۷] در مقاله اشان دسته وسیعی از ترکیبات بنزنی منو- و دی- استخلافی را بررسی کردند. برای پیشبینی نقطه ذوب این ترکیب های آنها یک مدل خطی ارایه کردند و به این نتیجه رسیدند که مهم ترین پارامتر در پیش بینی نقطه ذوب توصیف کننده پیوند هیدروژنی است.

در این مقاله مدل QSPR بر پایه ۴۱۷۳ ترکیب شیمیایی بنا خواهد شد. از تجزیه اجزاهای اصلی برای کوچک کردن فضای متغیرهای وابسته (توصیف کننده ها) استفاده می شود. از مناسب ترین بردارهای ویژه بدست آمده به عنوان ورودی برای شبکه عصبی استفاده کرده و با استفاده از آن سعی می شود که مدل مناسبی برای پیش بینی نقاط ذوب ایجاد شود.

بخش تجربی

داده ها و توصیف کننده های مولکولی

اطلاعات مربوط به نقطه ذوب ۴۱۷۳ ترکیب استفاده شده در این مقاله از مقاله کاتریکیان و همکارانش [۴۲] استخراج شد. ساختار این ترکیب ها، به همراه ۲۷۷ ترکیب دیگر که به عنوان دسته ارزیابی مورد استفاده قرار گرفتند، به وسیله کاتریکیان و همکارانش با استفاده از نرم افزار HyperChem و با روش نیمه تجربی AM۱ بهینهسازی شد. آنها با استفاده از نرم افزار Dragon برای هر ساختار ۸۰۰ توصیف کننده محاسبه کردند که پس از حذف آن هایی که با هم وابستگی خطی داشتند، ۲۰۲ توصیف کننده برای هر ساختار باقی ماند. داده های مربوط به این داده ها به همراه توصیف کننده های محاسبه شده و نام آن ها به عنوان داده های تکمیلی در مقاله کاتریکیان و همکارانش آورده شده است [۲۴].

کل مولکولها به دو دسته تقسیم شدند. ۲۵۷۳ ترکیب به عنوان دسته آموزش و ۱۶۰۰ ترکیب به عنوان دسته آزمون (دسته ارزیابی) مورد استفاده قرار گرفتند. این دسته بندی بدون ترتیب صورت گرفت. قبل از انجام PCA بر روی ماتریس داده ها عمل صفر کردن میانگین و یکسان سازی[۱] انحراف استاندارد[۲] صورت گرفت. برای ارزیابی بیشتر مدل ایجاد شده از یک دسته داده دیگر با اندازه ۲۷۷ استفاده شد.

شبکه عصبی مصنوعی (ANN)

الگوریتم شبکه عصبی از سامانه شبکه عصبی انسان الگو برداری شده است. الگوریتم شبکه های عصبی مصنوعی یک سامانه پردازش اطلاعات است که از تعداد زیادی واحدهای پردازشی (نورون) مرتبط با هم تشکیل شده اند. یک شبکه عصبی مصنوعی از فن‌های مورد استفاده انسان در یادگیری از روشی استناد به مثاله ایی از حل مسائل استفاده می‌کند. هر نورون ورودی های متعددی را پذیرا است که با یکدیگر به روشی جمع می شوند، بنابراین میتوان گفت که فعالیت هر نورون عبارت است از گرفتن مجموعه ای از یک یا چند ورودی، انجام یک عملیات ریاضی روی آنها و تولید خروجی های مناسب. عملکرد اساسی این مدل مبتنی بر جمع کردن ورودی ها و به دنبال آن به وجود آمدن یک خروجی است . ورودی های نورون از طریق دندریت هایی که به خروجی نورون های دیگر از طریق سیناپس متصل شدهاند، وارد می شوند . بدنه سلولی تمام این ورودی ها را دریافت می کند و چنانچه جمع این مقادیر از مقداری که به آن آستانه گفته میشود بیشتر شود در اصطلاح برانگیخته شده یا آتش می گیرد و در غیر این صورت خروجی نورون خاموش خواهد شد. در کـــل الگوریتم های شبکه عصبی بر اساس چند لایه ای بودن، نوع روش آموزش، نوع عملیات صورت گرفته در نورونها و نوع ورودی هایی که می پذیرند دسته بندی میشوند و برای انجام داده پردازی های مختلفی به کار گرفته شده اند. [۲۵].

در این مقاله از شبکه عصبی با تغذیه مستقیم با پس انتشار خطا (ANN-BP-FF)[۳] استفاده می شود [۲۵]. برای هر دسته ورودی این شبکه آموزش داده میشود. بدین معنی که بعد از انجام تجزیه اجزاهای اصلی بر روی دسته آموزش، بردارهای ویژه ی به دست آمده مرتب شده و یکی یکی وارد مدل می شوند. در هر مرحله شبکه عصبی بهینه سازی میشود به طوریکه کمترین خطا را در پیش بینی دسته آموزش و دسته تست داشته باشد. مرتب سازی بردارهای ویژه براساس محتوی اطلاعات (مقدار واریانس) و یا براساس میزان همبستگی هر بردار ویژه با مقادیر نقاط ذوب و یا براساس میزان قدرت پیش بینی کنندگی اشان صورت می گیرد.

لازم به ذکر است که تمامی محاسبات در محیط نرم افزار مطلب [۶۲] انجام گرفته و برای ساخت شبکه عصبی از جعبه ابزار شبکه عصبی مصنوعی آن بهره برداری شده است.

نتیجه ها و بحث

تجزیه اجزای اصلی

از آنالیز اجزای اصلی روی ماتریس توصیف کننده ها، ماتریس بردارهای ویژه به دست می آید. این بردارها بر یکدیگر عمودند و به هر کدام یک مقدار ویژه تخصیص داده میشود که نشان دهنده میزان محتوی اطلاعات بردار مربوطه میباشد. نخستین بردار ویژه دارای بزرگترین مقدار ویژه است. شکل ۱ جمع تراکمی مقادیر واریانس را در کل بردارهای ویژه نمایش میدهد. از روی شکل به راحتی میتوان فهمید که بیش از ۸۰ درصد اطلاعات در ۵۰ بردار ویژه اول جمع شده است. این حقیقت به یکی از مهمترین مزایای تجزیه اجزای اصلی که کاهش ابعاد داده هاست اشاره دارد. بنابراین میتوان به جای کل توصیف کننده ها که تعدادشان به ۲۲۰ میرسد تنها ۵۰ بردار ویژه اول به دست آمده را استفاده کرد.

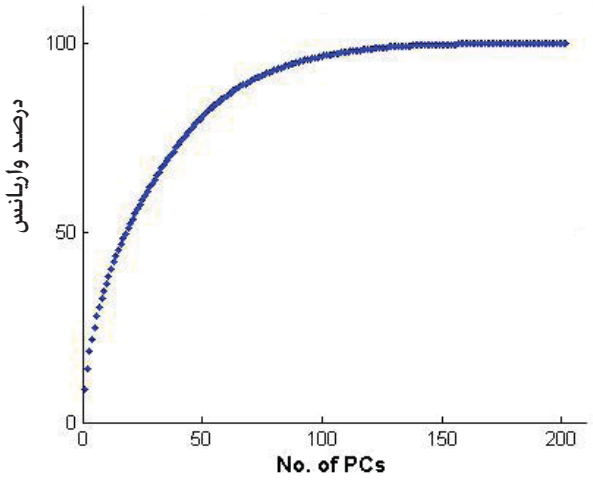
مدلسازی با استفاده از شبکه عصبی

از آنجا که میزان نقطه ذوب یک ماده به عوامل متعددی بستگی دارد، وجود رابطـه ی غیر خطـــی میان نقـــاط ذوب و توصیف کننده های مولکولی بسیار محتمل است. در این مقاله سعی شده است که با استفاده از الگوریتم شبکه مصنوعی (ANN) یک مدل با قدرت پیش بینی بالا ایجاد شود. در پایان نتیجه های به دست آمده با یک روش خطی مانند رگرسیون چندگانه[1] خطی مقایسه خواهد شد.

برای اجرای شبکه عصبی، سه دسته داده نیاز است: دسته آموزش (کالیبراسیون)، دسته تست (پیش بینی) و دسته ارزیابی. برای ایجاد یک شبکه بهینه از دسته آموزش استفاده شده و سپس از همین شبکه برای ارزیابی مقادیر نقطه ذوب مولکولها در دسته تست استفاده می شود. برای ارزیابی بیشتر شبکه ایجاد شده و اطمینان از صحت آن از دسته ارزیابی استفاده می شود.

انتخاب نوع ورودی شبکه عصبی بر میزان کارآیی آن تاثیرگذار است. در اینجا به جای استفاده از ماتریس توصیف کننده ها به عنوان ورودی شبکه عصبی از بردارهای ویژه استفاده شد زیرا که حجم زیادی از اطلاعات در تعداد کمی از بردارهای ویژه فشرده شده است. اما در اینجا نیز یک مسئله مطرح میشود و آن اینکه کدام بردارهای ویژه برای ایجاد شبکه مناسب ترند؟ در این مقاله از سه معیار برای مرتب کردن بردارهای ویژه و انتخاب آنها برای استفاده در شبکه عصبی استفاده شد:

۱. مرتب کردن بردارهای ویژه براساس مقدارهای ویژه آنها
۲. مرتب سازی براساس میزان همبستگی بردارهای ویژه با



شکل ۱ جمع تراکمی مقدارهای واریانس (اطلاعات) در هر بردار ویژه

مقدارهای نقاط ذوب

۳. مرتب سازی براساس میزان قدرت پیش بینی کنندگی (مدلسازی) هر بردار ویژه؛ برای تعیین این قدرت برای هر بردار، تک تک بردارها را به عنوان ورودی به شبکه عصبی وارد نموده (در هر مرحله یک بردار) و پس از آموزش شبکه ایجاد شده، از آن برای پیش بینی نقاط ذوب مولکول های دسته آزمون استفاده میشود. از مجذور همبستگی (R^2) میان مقادیر پیش بینی شده ومقدارهای واقعی به عنوان معیاری برای قدرت پیش بینی کنندگی هر بردار استفاده شد.

بعد از مرتب کردن بردارهای ویژه براساس هر معیاری، تعدادی از آن ها را که مسوول ایجاد بهترین شبکه هستند، انتخاب و به شبکه به عنوان ورودی داده می شود. سپس خود شبکه عصبی از نظر تعداد لایه های پنهانی، تعداد گره در هر لایه، میزان سرعت یادگیری و مومنتوم بهینه میشود. برای قسمتهایی که در ادامه برای انتخاب نوع ورودی می آیند با تعدادی آزمایشات اولیه و سعی و خطا شبکه عصبی با ۵ لایه (لایه ورودی، ۳ لایه میانی یا پنهان و لایه خروجی) با تعداد گره های ۱۰، ۱۰ و ۸ در لایه های میانی انتخاب شد.

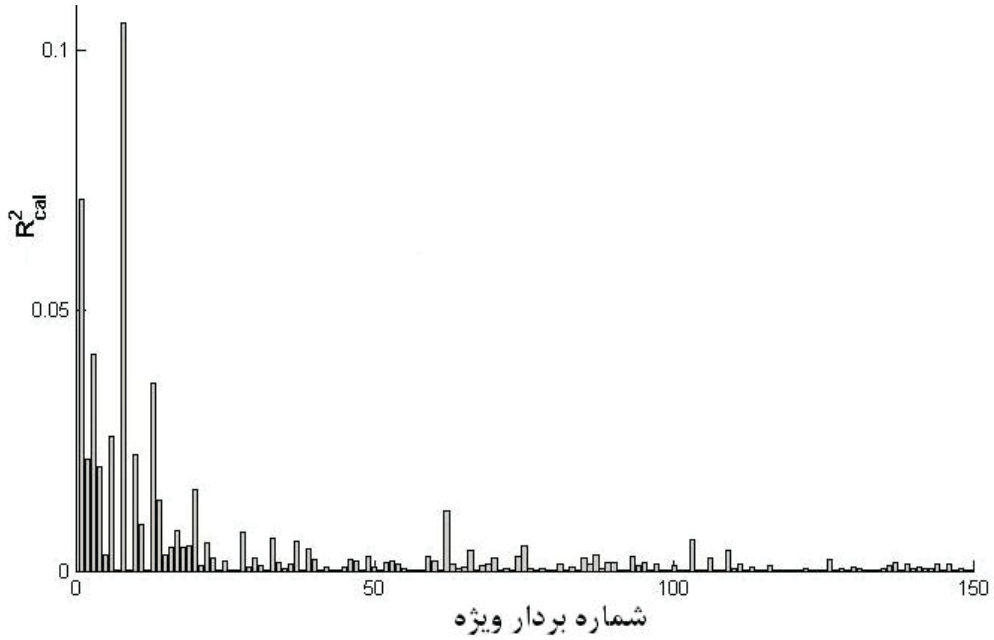
بردارهای ویژهای که از تجزیه اجزاهای اصلی ایجاد می شوندبراساس میزان واریانسشان مرتب شده اند و به ترتیب نیز به شبکه عصبی وارد می شوند. افزودن تعداد بردارها تا آنجا ادامه می یابد که دیگر بهبودی محسوسی در کیفیت مدل تولید شده ایجاد نشود. بیشترین میزان R^2_{cal} حاصل شده برابر ۰/۶۰ و حداکثر میزان R^2_{pred} برابر ۰/۴۹ بدست آمد. لازم به ذکر است که R^2_{cal} و R^2_{pred} به ترتیب میزان مجذور همبستگی میان مقادیر پیشبینی شده و مقادیر واقعی نقاط ذوب در مورد دسته آموزش و دسته تست را نشان می دهند.

لازم به ذکر است که نتیجه رگرسیون چندگانه خطی، مدلی با کارآیی و قدرت پیشبینی کنندگی کم است به طوری که R^2_{cal} برابر با ۰/۵۳ و R^2_{pred} برابر با ۰/۴۷ بدست میدهد. این نتایج با استفاده از ۴۵ توصیف کننده که از بین ۲۰۲ توصیف کننده با روش گام به گام انتخاب شده بودند، به دست آمد. ایجاد چنین مدل ضعیفی با استفاده از یک مدل خطی گواهی بر رابطه غیر خطی میان متغیرهای مستقل (توصیف کننده ها) و وابسته غیر مستقل (نقاط ذوب) می باشد.

همان طور که اشاره شد مطالعات قبلی نشان داده است که الزاماً نخستین بردارهای ویژه که شامل بیشترین اطلاعات هستند، برای مدلسازی یک خاصیت مفید نیستند [۱۰]. یک راه برای انتخاب مناسبترین بردارهای ویژه در یک مسئله، مرتب کردن آنها براساس یک معیار مناسب است. یک معیار مناسب میتواند میزان همبستگی هر بردار ویژه با خاصیت مورد بحث در مسئله باشد. شکل ۲ میزان همبستگی هر بردار ویژه با مقادیر نقطه های ذوب را به نمایش میگذارد. . از روی این شکل به سرعت میتوان دریافت که بیشترین همبستگی با نقاط ذوب را هشتمین بردار ویژه دارست. به عبارت دیگر، بررسی نتایج ارایه شده در این شکل نمایانگر این حقیقت است که اولین بردار ویژه الزاماً بیشترین همبستگی را با خاصیت مورد بحث ندارد.

برای افزایش کیفیت و قدرت پیشبینی مدلی که شبکه عصبی ارایه میکند، این بار بردارهای ویژه به ترتیب و براساس میزان همبستگی– اشان با مقادیر نقاط ذوب به شبکه داده شدند. در هر مرحله شبکه عصبی ایجاد شده بعد از بهینه سازی برای پیش بینی نقاط ذوب مولکولهای دسته آموزش و دسته تست استفاده شد. بیشترین مقدار برای R^2_{cal} و R^2_{pred} به ترتیب برابر با ۰/۷۲ و ۰/۶۳ حاصل شد. مقایسه این مقدارها با مقدارهای به دست آمده در قسمت قبل (به ترتیب ۰/۶۰ و ۰/۴۹) نمایانگر این مطلب است که میزان همبستگی PCها با نقاط ذوب معیار مناسبی برای مرتب کردن بردارهای ویژه است.

توجه به این نکته ضروری است که شبکه عصبی یک مدل غیر خطی ارایه میکند، حال آنکه میزان همبستگی یک بردار ویژه با خاصیت مورد بحث یک معیار خطی میباشد. لذا علیرغم آنکه بهبودی در نتایج با استفاده از این معیار بدست آمد، تصور میرود که اگر معیار دیگری که کمتر به رابطه خطی میان خاصیت مورد نظر و بردارهای ویژه وابسته است به کار گرفته شود، نتایج بهتری حاصل می شود. معیاری که چنین خصوصیتی داشته باشد می تواند میزان قدرت پیش بینی بردارهای ویژه باشد. بدین معنی که هر بردار ویژه بطور جداگانه به شبکه عصبی داده شده و بعد از بهینه



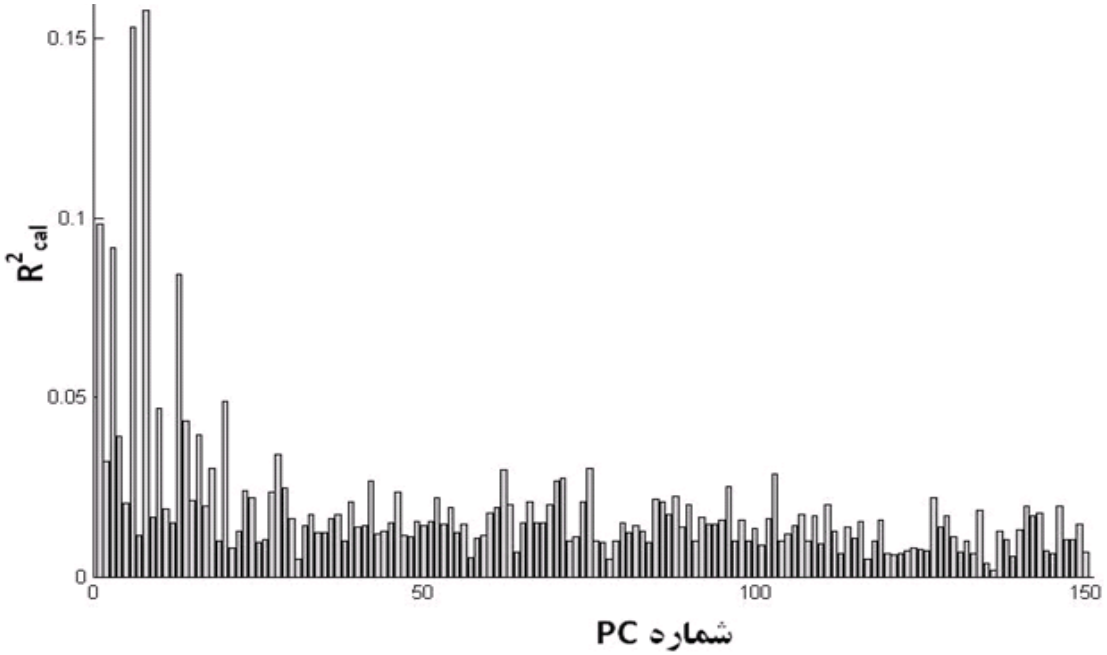
شکل ۲ ضریب های همبستگی هر بردار ویژه با مقدارهای نقطه های ذوب

سازی شبکه، مقدارهای نقاط ذوب مولکولهای دسته آموزش با آن ارزیابی می شوند. شکل ۳ میزان قدرت پیشبینی هر بردار ویژه را به نمایش می گذارد. مقایسه این شکل و شکل ۲ نشان دهنده دو مطلب است: اول آنکه ترتیب خوب و مناسب بودن بردارهای ویژه در هر دو شکل یکسان نیست و دوم آنکه به راستی هر بردار ویژه که دارای همبستگی بالایی با نقاط ذوب است، دارای قدرت پیش بینی خوبی نیست.

برای ارزیابی معیار جدید ارایه شده و بررسی میزان تأثیر آن بر کیفیت مدل شبکه عصبی، بردارهای ویژهای که براساس این معیار مرتب شده اند به ترتیب به عنوان ورودی به شبکه عصبی وارد می شوند. ماکزیموم R^2_{cal} و R^2_{pred} به دست آمده به ترتیب برابر ۰/۷۸ و ۰/۷۲ بودند. واضح است که این معیار تأثیر به سزایی در کیفیت مدل شبکه عصبی داشته است.

در آخرین مرحله سعی می شود که با استفاده از مناسب ترین معیار و انتخاب بهترین ورودیها، شبکه عصبی را از نظر تعداد لایه های پنهان، میزان سرعت یادگیری و مومنتوم بهینه کرد. برای این منظور شبکه هایی با تعداد لایه های پنهان یک، دو و سه و با تعداد نودهای یک تا سی در هر لایه مورد بررسی قرار گرفتند. بعد از انجام این سری بررسی های طولانی شبکه با ساختار [۲۵ ۱۰ ۸ ۱] به عنوان بهترین شبکه انتخاب شد. برای این شبکه مقادیر سرعت یادگیری و مومنتوم با سعی و خطا بهینه سازی شد و مقادیر ۰/۷۰ و ۰/۱۶ به ترتیب برای سرعت یادگیری و مومنتوم به عنوان مقادیر بهینه انتخاب شدند. با استفاده از این شبکه بهینه سازی شده و با استفاده از ده بردار ویژه اول (براساس شکل ۳)، مقادیر R^2_{cal} و R^2_{pred} برابر با ۰/۷۱ و ۰/۷۰ به دست آمدند.

به منظور ارزیابی بیشتر شبکه ایجاد شده و اطمینان از صحت آن از یک دسته داده دیگر شامل ۲۷۷ ترکیب جدید بهره گرفته شد. نقاط ذوب این دسته داده با استفاده از شبکه ایجاد شده پیش بینی شد. همبستگی مقادیر پیش بینی شده با مقادیر واقعی بسیار خوب بود (R^2_{val} = ۰/۶۸) که این حاکی از کارا بودن شبکه می باشد .نزدیکی مقادیر R^2_{cal} ، R^2_{pred} و R^2_{val} [۱] بیانگر آن است که شبکه ایجاد شده مشکل بیشتر جفت و جور شدن را ندارد.



شکل ۳ مقدار R^2_{cal} بدست آمده از شبکه ایجاد شده برای هر بردار ویژه. بالا بودن مقدار R^2_{cal} بیانگر مفید بودن بردار مربوطه میباشد

نتیجه گیری

در بسیاری از شاخه های شیمی با حجم انبوهی از داده ها روبه رو هستیم که برای کم کردن حجم و ابعاد آنهـــا مجبور به انجـــام روش هایی مانند تجزیه اجزاهای اصلی هستیم. این تجزیه منجر به تولید اجزاهای اصلی (بردارهای ویژه) می شود که در آنها اطلاعات ذخیره شده است. در این پروژه، سه معیار برای مرتب کردن اجزاهای اصلی و یافتن مهمترین آن ها ارایه شد. این معیارها عبارتند از میزان مقدارهای ویژه، میزان همبستگی با مقدارهای نقاط ذوب و میزان قدرت پیش بینی. نتیجه ها نشان می دهند که هنگامی که از شبکه عصبی برای ایجاد مدل استفاده میشود، مرتب سازی براساس میزان قدرت پیش بینی بسیار بهتر از سایر معیارها است. مدل ارایه شده با قدرت بسیار بالایی می –توانست مقادیر نقاط ذوب مولکول های جدید را پیش بینی نماید.

تقدیر و تشکر

مقاله حاضر از طرحِپوهشی با عنوان «بررسی معیارهای مختلف آماری برای مرتب کردن صحیح تر فاکتورهای اصلی یک دسته داده که با روش آنالیز فاکتورها محاسبه شده اند» استخراج شده است. نویسندگان مقاله از دانشگاه آزاد اسلامی واحد ماهشهر که هزینه مالی این پژوهش را تامین نموده است، کمال تشکر را دارند.

مراجع

[1] H. van de Waterbeemd. VCH, New York, 1995.
[2] Yao, S.W., Lopes, V.H.C., Fernandez, F., Garcia-Mera, X., Morales, M., Rodriguez-Borges, J.E., Cordeiro, M.N.D.S., Synthesis and QSAR Bioorg. Med. Chem. 11 4999–5006 (2003).
[3] Shijin, R., Chemosphere 53 1053–1065 (2003).

[4] Liang, Y.Z., Xie, Y.L., Yu, R.Q., Chim. Acta 222 347-357 (1989).

[5] Kalivas, J.H., Roberts, N., Sutter, M.J., Anal. Chem. 61 2024-2030 (1989).

[6] Lucasius, C.B., Beckers, M.L.M., Kateman, G., Anal. Chim. Acta 286 135-153 (1994).

[7] Araujo, M.C.U., Saldanha, T.C.B., Galvao, R.K.H., Oneyama, T., Chame, H.C., Visani, V., The Chemometr. Intell. Lab. Syst. 57 65-73 (2001).

[8] Leardi, R., . J. Chemomert. 14 643-655 (2000).

[9] Niculescu, S.P., (Theochem) 622 71–83 (2003).

[10] Hemmateenejad, B., Optimal QSAR PCR. J. Chemom. 18 475–485 (2004).

[11] Yalkowsky, S. H., Valvani, S. C.,. J. Pharm. Sci. 69 912–922 (1980).

[12] Ran, Y., Yalkowsky, S.H., (GSE). J. Chem. Inf. Comput. Sci. 41 354–357 (2001).

[13] Gavezzotti, A., Molecular symmetry. J. Chem. Soc., Perkin Trans; 2 1399–1404 (1995).

[14] Dearden, J.C., Sci. Total Environ. 109 59–68 (1991).

[15] Charton, M., Charton, B., J. Phys. Org. Chem. 1994, 7, 196–206.

[16] Katritzky, A.R., Maran, U., Karelson, M., J. Chem. Inf. Comput. Sci. 37 913–919 (1997).

[17] Charton, M., J. Comput.-Aided Mol. Des. 17 197–209 (2003).

[18] Bergström, C.A.S., Norinder, U., Luthman, K., Artursson, P., J. Chem. Inf. Comput. Sci. 43 1177–1185 (2003).

[19] Ma, L., Cheng, C., J. Chemom. 16 75–80 2002.

[20] Burch, K.J., J. Chem. Eng. Data 49 858–863 (2004).

[21] Simamora, P., Miller, A.H., Yalkowsky, S.H., J. Chem. Inf. Comput. Sci. 33 و437–440 (1993).

[22] Krzyzaniak, J.F., Myrdal, P.B., Simamora, P., Yakowsky, S.H., Ind. Eng. Chem. Res. 34 2530–2535 (1995).

[23] Simmamora, P., Yalkowsky, S.H., Ind Eng Chem Res 33 1405–1409 (1994).

[24] Karthikeyan, M., Glen, R.C., Bender, A., J. Chem. Inf. Model. 45 581–590 (2005).

[25] J., Gasteiger, J., Wiley-VCH, Weinheim, 1999.

[26] MATLAB, version 7.6.0.324, Math Work, Inc.

JACR

Investigation of various criteria to rank PCs in order to develop a QSPR model for predicting melting points

Vali Zare-Shahabadi[1]* and Fatemeh Abbasitabar[2]

1-Department of Chemistry, Islamic Azad University – Mahshahr Branch, Mahshahr, Iran.
2-Islamic Azad University – Marvdasht Branch, Marvdasht, Iran

Abstract: Due to the importance of melting points, we tried to develop a useful model for the prediction of them. In this way we used 4173 compounds with large diverse structures which reported previously in the literature. The chemical structures of all compounds were described by a set of 202 descriptors including 2D and 3D descriptors. This data set was divided into training set and test set with size of 2573 and 1600, respectively. Additional validation was performed on an external validation set consisting of 277 chemicals. Dimensionality reduction is performed by principal component analysis (PCA), while a feed-forward back-propagation artificial neural network (ANN) is employed for model generation. Principal components (PCs), obtained from PCA, were ranked based on different criteria and fed to the ANN as inputs. These criteria included the magnitude of eigenvalues, correlation values, and prediction ability. Optimal model was resulted in the case that PCs have been ranked based upon their prediction abilities. Finally, the net parameters such as number of hidden layers, number of nodes in each layer, learning rate, and momentum were adjusted. The net with structure of [25 10 8 1], learning rate of 0.7, and momentum value of 0.16 were chosen as optimum.

Keywords: Quantitative structure property relationship; principal component analysis; PC ranking; artificial neural network.

*E mail: valizare@gmail.com