مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره اول. ص ۳۹-۳۰، ۱۳۹۹

بهبود تخمین اثر بیولوژیکی ملکول‌های مهارکننده پروتئین کیناز، با استفاده از شبکه عصبی و مینیمم خطای جزئی

رویا آرین[۱]، علیرضا مهری دهنوی[۲]، فهیمه قاسمی[۳]*

• دریافت مقاله: ۱۳۹۸/۲/۲ • پذیرش مقاله: ۱۳۹۸/۸/۱۱

مقدمه: پروتئین کیناز عامل ایجاد بسیاری از بیماری‌ها از جمله سرطان است؛ بنابراین مهار آن ها در درمان بسیاری از بیماری‌ها نقش بسزایی ایفا می‌کند. کشف داروهای جدید با روش‌های آزمایشگاهی، از جمله موضوعات هزینه بردار و زمان‌بر می‌باشد؛ یافتن مدل‌های محاسباتی قابل اطمینان برای شناسایی مهارکننده‌ها می‌تواند هزینه‌ها را به حداقل برساند. هدف از این مطالعه به کارگیری روش شبکه عصبی جهت طبقه‌بندی ترکیبات در دو گروه فعال و غیر فعال و مدل رگرسیون خطی مینیمم خطای جزئی به منظور تخمین میزان اثر بیولوژیکی آن‌ها است.

روش: در این پژوهش، پس از استخراج توصیفگرها از داده‌ها، به منظور جلوگیری از بیش برازش مدل‌ها، کاهش ابعاد داده از طریق الگوریتم ژنتیک صورت پذیرفت. همچنین جهت طبقه‌بندی داده‌ها در کلاس فعال و غیر فعال از مدل شبکه عصبی و جهت تخمین مقادیر اثر بیولوژیکی ریزملکول‌ها از مدل رگرسیون خطی مینیمم خطای جزئی استفاده شد.

نتایج: نتایج نشان داد بعد از کاهش بعد توصیفگرهای ملکولی، صحت مدل شبکه عصبی از ۷۴/۴۵٪ به ۸۶/۷٪ تغییر یافت. این مدل در تعداد گره‌های لایه پنهان برابر با ۶، صحت ۸۶/۷٪، حساسیت ۸۳/۴٪، اختصاصی بودن ۸۹/۶٪ و ضریب همبستگی متیو ۷۳/۲٪ را ارائه می‌دهد. مدل رگرسیون خطی مینیمم خطای جزئی نیز با میزان همبستگی متوسط ۸۵/۸٪ مقادیر بیولوژیکی را تخمین می‌زند.

نتیجه‌گیری: مدل طبقه‌بندی شبکه عصبی و مدل رگرسیون خطی مینیمم خطای جزئی تا میزان قابل قبولی می‌توانند مهارکننده‌های پروتئین کیناز را پیش‌بینی کنند و الگوریتم کاهش بعد ژنتیک عملکرد این مدل‌ها را بهبود می‌بخشد.

کلید واژه‌ها: پروتئین کیناز، طبقه‌بندی، شبکه عصبی، رگرسیون، مینیمم خطای جزئی

•ارجاع: آرین رویا، علیرضا دهنوی مهری، قاسمی فهیمه. بهبود تخمین اثر بیولوژیکی ملکول‌های مهارکننده پروتئین کیناز، با استفاده از شبکه عصبی و مینیمم خطای جزئی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۱): ۹-۳۰.

۱. کارشناسی ارشد بیوالکتریک، گروه بیوالکتریک، دانشکده فناوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲. دکترای تخصصی بیوالکتریک، استاد، گروه بیوالکتریک، دانشکده فناوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۳. دکترای تخصصی بیوالکتریک، استادیار، گروه بیوانفورماتیک، دانشکده فناوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران

*** نویسنده مسئول:** فهیمه قاسمی

آدرس: اصفهان، خیابان هزارجریب، دانشگاه علوم پزشکی اصفهان

•شماره تماس: ۰۳۱ -۳۷۹۲۳۸۶۵ • **Email:** f_ghasemi@amt.mui.ac

مقدمه

پروتئین کیناز یک گروه مهم از آنزیم‌های کیناز است که پروتئین‌ها را با افزودن گروه فسفات (PO_4) فسفوریله می‌کند. بیش از سی درصد از پروتئین‌های بدن تحت تأثیر این آنزیم‌ها فسفوریله می‌شوند. گروه‌های فسفات می‌توانند فعالیت پروتئین را دچار اختلال کنند و موجب ایجاد بسیاری از بیماری‌ها از جمله سرطان گردند. مهار پروتئین کیناز امری حیاتی در درمان بعضی از سرطان‌ها و بیماری‌های التهابی است؛ بنابراین مهارکننده‌های پروتئین کیناز می‌توانند به عنوان دارو عمل کنند [۱].

کشف داروهای جدید در آزمایشگاه‌ها و به روش کلینیکالی بسیار پر هزینه و زمان‌بر می‌باشد. تحقیقات نشان داده است که از روی شکل ظاهری مولکول‌ها و ترکیبات می‌توان به اطلاعاتی دست یافت که با مقایسه این اطلاعات به ویژگی‌های هر ترکیب پی خواهیم برد. به همین جهت بررسی رابطه بین ساختار مولکولی ریزملکول‌ها (لیگاندها) و میزان فعالیت آن و نیز شناسایی ریز مولکول‌های تأثیرگذار بر پروتئین خاص بر اساس رفتار ریز مولکول‌های مشابه با آن که قبلاً مورد بررسی قرار گرفته‌اند، در طراحی و کشف نمونه‌های جدید دارویی از اهمیت بسیار زیادی برخوردار می‌باشد [۲].

طراحی کامپیوتری دارو با استفاده از روش‌های ریاضی و یا آماری برای هدف مشخص، رهیافتی نوین در طراحی و غربالگری داروهای جدید با منشأ شیمیایی یا گیاهی، جهت کاهش چشمگیر هزینه و وقت است. داده‌های به دست آمده در روش‌های کامپیوتری به منظور برپاکردن مدل‌هایی مورد استفاده قرار می‌گیرند که قادر به پیش‌بینی قدرت اثر بیولوژیکی و یا دسته‌بندی مولکول‌ها از نظر فعالیت بیولوژیکی هستند. هرچه مدل از قدرت بیشتری جهت طبقه‌بندی یا تخمین برخوردار باشد با دقت بهتری می‌تواند نمونه‌های دارویی جدید را پیش‌بینی و پیشنهاد کند. این روش مجازی تخمین اثر بیولوژیکی ترکیبات که سریع‌تر و مقرون به صرفه‌تر از روش‌های تجربی است، به دو دسته مختلف غربالگری بر اساس لیگاند (Quantitative Structure Activity Relationship) QSAR و غربالگری بر اساس ساختار، تقسیم می‌شوند. در روش‌های QSAR در ابتدا سعی می‌شود مدلی جهت ارتباط دادن توصیفگرهای ساختاری یک دسته مولکول و میزان فعالیت آن‌ها ارائه شود، سپس بهترین مولکول که با کمترین غلظت، بیشترین و اختصاصی‌ترین اثر را دارد انتخاب می‌شود [۳].

استفاده از روش‌های یادگیری ماشینی از جمله ماشین بردار پشتیبان SVM(Support Vector Machine) [۴]، K نزدیک‌ترین همسایه KNN(K Nearest Neighbors) [۵]، K نزدیک‌ترین همسایه فازی Fuzzy KNN [۶-۸]، Naïve Bayesian [۹]، RF [۱۰]، NN (Neural Network) [۱۱]، PLS(Partial Least Square) [۱۲] و شبکه‌های عمیق [۹] علمی است که در دهه‌های اخیر در همه حوزه‌ها از جمله علم داروسازی وارد شده است و در طبقه‌بندی ترکیبات و تخمین مقادیر اثر بیولوژیکی بر اساس توصیفگرهای مولکولی به کار گرفته شده است [۱۳]. توصیفگرهای مولکولی به کلیه اطلاعات مستخرج از ساختار ترکیبات گفته می‌شود که از جمله آن‌ها می‌توان ویژگی‌های فیزیکوشیمیایی و بارهای فضایی اطراف لیگاند را نام برد؛ اما مشکل عمده‌ایی که با آن مواجه خواهیم شد تعداد زیاد توصیفگرهای مولکولی در مقایسه با تعداد کم ترکیبات می‌باشد که باعث به وجود آمدن مشکل بیش برازش مدل می‌شود که روش‌های کاهش بعد، به کمک این مسئله آمده‌اند [۲].

هدف از این مطالعه به‌کارگیری روش شبکه عصبی LVQ(Learning Vector Quantization Neural Network) به منظور طبقه‌بندی ترکیبات در دو گروه فعال (مهارکننده پروتئین کیناز) و غیرفعال (غیرمهار کننده پروتئین کیناز) و مدل PLS به منظور تخمین میزان اثر بیولوژیکی و بررسی عملکرد آن‌ها است. در این راستا ابتدا جهت ارزیابی مدل‌ها، داده‌ها به دو دسته داده تست و آموزش به روش K-fold تقسیم شدند. به منظور رفع مشکل بیش برازش مدل به علت تعداد بالای توصیفگرهای استخراج شده، با استفاده از روش کاهش بعد الگوریتم ژنتیک تعداد توصیفگرهای مولکولی به ۶ توصیفگر کاهش داده و عملکرد آن بر روی مدل بررسی شد. نتایج به دست آمده حاکی از آن است که این الگوریتم در نتیجه مدل بسیار مؤثر است. در نهایت با کلاسیفایر LVQ ریز مولکول‌ها طبقه‌بندی و با مدل رگرسیون PLS اثر بیولوژیکی آن‌ها تخمین زده شد.

در این مطالعه ریز مولکول‌های (لیگاندهای) (مقدار غلظت دارو یا لیگاند (بر حسب مولار) که برای مهار فعالیت یک پروتئین، آنزیم یا ... و کاهش فعالیت آن به نصف میزان اولیه، نیاز است.) زیستی کاندیدای مهارکنندگی پروتئین کیناز مورد مطالعه قرار گرفتند. تعداد ۹۰ ریز مولکول بررسی شده Saghaie و همکاران [۱۴]، با میزان مهارکنندگی یا (Half maximal Inhibitory Concentration) IC_{50} مشخص، به عنوان

ورودی مدل در نظر گرفته شد. تمامی ریز مولکول‌ها در نرم‌افزار ChemDraw به صورت دوبعدی ترسیم و سپس شکل سه بعدی آن‌ها با استفاده از نرم‌افزار HyperChem Professional که شبیه‌سازی فضایی را انجام می‌دهد بهینه شد. سپس به منظور استخراج ویژگی، نرم‌افزار Dragon مورد استفاده قرار گرفت [۱۵]. نرم‌افزار دراگون نخستین بار توسط گروه QSPR و کمومتریکس دانشگاه میلانو در سال ۱۹۹۴ طراحی شده است. این نرم‌افزار توانایی استخراج بیش از ۱۶۰۰ توصیفگر یک‌بعدی، دوبعدی و سه‌بعدی را دارد که به ۲۰ دسته اصلی تقسیم می‌شوند [۱۶]. در نهایت خروجی به فرم زیر به دست آمد که این ماتریس، همان ماتریس ورودی مدل است.

$$\text{Input Matrix} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

در اینجا m تعداد ریز مولکول‌ها یعنی همان ۹۰ و n تعداد توصیفگرها می‌باشد که در این مطالعه ۷۸۲ توصیفگر از نرم‌افزار استخراج گشت. تعداد کل توصیفگرهای استخراج شده بسیار زیاد است که جهت بهبود عملکرد مدل‌ها بایستی این تعداد با استفاده از روش‌های کاهش بعد مناسب تقلیل یابد. در آنالیزهای همبستگی می‌توان توصیفگرهایی را که همبستگی زیادی با توصیفگر دیگر یا همبستگی پایینی با کلاس فعال دارد را حذف نمود. مستقل بودن و نداشتن همپوشانی با سایر توصیفگرها از جمله ویژگی‌های مطلوبی است که توصیفگر باقی مانده باید دارا باشد. از میان دو توصیفگری که همبستگی بالایی با یکدیگر دارند، توصیفگری حذف می‌گردد که کمترین همبستگی با کلاس هدف دارد [۱۷]. در این مطالعه جهت کاهش ابعاد و جلوگیری از رخداد بیش برازش از روش الگوریتم ژنتیک استفاده و توصیفگرهای بهینه، برگزیده شدند. الگوریتم ژنتیک نوعی الگوریتم تکاملی است و بر اساس تکنیک وراثت و جهش، عمل می‌کند و نخستین بار در سال ۱۶۶۷ توسط جان هالند، معرفی شد. در این الگوریتم مجموعه‌های کوچک از ویژگی‌ها انتخاب می‌گردد و در هر تکرار تأثیر یکی از آن‌ها بر مدل مورد آزمایش قرار می‌گیرد. مجموعه‌هایی که مدل با آن‌ها عملکرد مناسبی دارد با هم ترکیب شده و به عنوان یک مجموعه جدید به کار برده می‌شوند. در واقع الگوریتم ژنتیک بر دو اصل استوار است. اول آن که توصیفگرهایی که قوی‌تر و مؤثرترند فرصت بیشتری برای باقی ماندن دارند و دوم آن که از ترکیب دو توصیفگر (یا دو مجموعه توصیفگر)، ممکن است توصیفگر مؤثرتری، حاصل شود. الگوریتم چندین بار اجرا می‌شود و در نهایت بهترین‌ها، بعد از تمامی اجراها، کشف و انتخاب می‌شوند [۱۸].

بهتر است قبل از انتخاب ویژگی‌های بهینه، ابتدا آن‌ها نرمال شود تا تفاوت محدوده داده‌ای بین آن‌ها از بین برود. در این پژوهش جهت دستیابی به این هدف از معادله (۱) زیر استفاده شد.

$$Xnormji = (Xji - Mini)/(Maxi - Mini)$$

در اینجا، $Xnorm_{ji}$: توصیفگر نرمال شده (jام) از ستون (iام) ، X_{ji} توصیفگر (jام) از ستون (iام)، Max_i ماکزیمم داده ستون (iام) و Min_i مینیمم داده ستون (iام)، است.

در این مطالعه، در مرحله طبقه‌بندی مولکول‌ها، با آستانه گذاری مناسب مقدار IC_{50} می‌توان دریافت که ریز مولکول‌هایی با $IC_{50}<7\mu m$، توانایی مهار مؤثر مولکول کیناز را دارند (فعال) و مولکول‌هایی با $IC_{50}>=7\mu m$ به خوبی از این توانایی برخوردار نمی‌باشند (غیرفعال)، بر این اساس از ۹۰ داده اولیه، ۴۳ داده به کلاس فعال با برچسب 1 و ۴۷ داده به کلاس غیرفعال با برچسب 1- تعلق گرفت. مدل به کار گرفته‌شده در این مقاله به منظور طبقه‌بندی ریز مولکول‌های موجود در دو کلاس فعال (مهارکننده پروتئین کیناز) و غیرفعال(غیر مهارکننده پروتئین کیناز)، شبکه عصبی LVQاست. LVQ یک روش طبقه‌بندی است که در آن هرکدام از خروجی‌ها، نمایش دهنده یک کلاس است. بردار وزن هرکدام از کلاس‌ها، مقداردهی اولیه شده و سپس توسط الگوریتم‌های یادگیری با نظارت بهینه می‌شوند، درنهایت شبکه LVQ ورودی را با توجه به وزن‌ها و بایاس‌های نهایی به یکی از کلاس‌ها نسبت می‌دهد.

الگوریتم

X: بردار ورودی، Wj: بردار وزن jامین کلاس،Cj: کلاس برای jامین واحد خروجی، T: کلاس صحیح برای بردار

آموزشی، ||X-Wj||: فاصله اقلیدسی بین بردار ورودی و jامین خروجی)

۱. انتخاب اولیه بردار وزن و نرخ یادگیری (α)
۲. مقدار j را به نحوی انتخاب کن که مقدار ||X-Wj|| حداقل شود.
۳. به‌روزرسانی وزن‌ها

Wj(new)=Wj(old)+α[X-Wj(old)] (if Cj=T)

Wj(new)=Wj(old)-α[X-Wj(old)] (if Cj≠T)

۴. کاهش نرخ یادگیری
۵. تا زمانی که به‌شرط خاتمه نرسیده است الگوریتم از مرحله (۲) تکرار شود، (شرط خاتمه می‌تواند تعداد تکرار معین و یا کمتر شدن مقدار خطا از میزانی مشخص باشد [۱۹]).

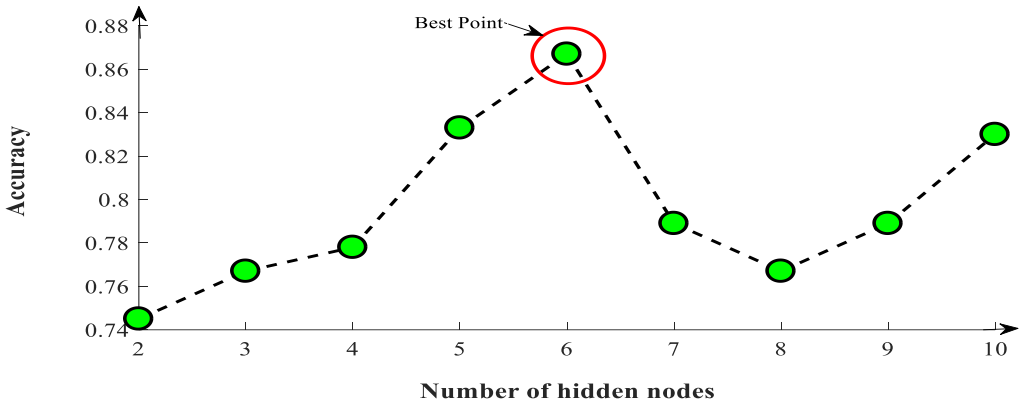
در این پژوهش همچنین تخمین دقیق اثر بیولوژیکی لیگاندها، با استفاده از روش خطی مینیمم خطای جزئی یا همان PLS(Partial Least Squares Regression) صورت گرفت. به‌طورکلی برای برآورد پارامترهای یک مدل دو رویکرد از جمله رویکرد مبتنی بر کوواریانس و رویکرد مبتنی بر واریانس یا PLS، وجود دارد. در PLS متغیرهای پنهان به‌گونه‌ای برآورد خواهند شد که ترکیب خطی از مقادیر مطلوب خود باشند. در این روش وزن‌ها به‌گونه‌ای محاسبه می‌شوند که برای پیش‌بینی متغیرهای وابسته از روی متغیرهای مستقل، بیشترین واریانس را داشته باشد. روش حداقل مربعات جزئی سازگاری بسیاری با نمونه‌ها و مجموعه‌های کوچک دارد.

در این پژوهش که تمامی روش‌ها در نرم‌افزار Matlab پیاده‌سازی شده‌است، ۹۰ داده به روش cross validation K-fold به داده‌های تست و آموزش تقسیم شدند که مقدار K در اینجا با توجه به تعداد داده‌ها، ۵ انتخاب شد. هر بار یک فولد به عنوان داده تست و باقی فولدها به عنوان داده آموزش به مدل داده شده و پارامترهای ACC، SE,SP و MCC محاسبه و در نهایت میانگین پارامترهای تمامی فولدها به عنوان مقادیر نهایی گزارش شد.

پس از پیش‌پردازش داده‌ها و استخراج ویژگی‌ها، توصیفگرهای بهینه از طرق الگوریتم ژنتیک برگزیده شدند. پس از چندین مرتبه اجرای الگوریتم ژنتیک و به دنبال آن، باقی مدل‌های طبقه‌بندی و پیش‌بینی کننده، به این نتیجه رسیدیم که مدل‌ها با ۶ توصیفگر بهینه نتایج قابل قبولی خواهند داشت.

سپس داده‌ها به روش شبکه عصبی LVQ طبقه‌بندی شدند. ابتدا مدل طراحی شده یک بار برای داده‌ها بدون اعمال الگوریتم ژنتیک و کاهش بعد و سپس برای داده‌هایی که به روش ژنتیک کاهش بعد یافته بودند، اجرا شد تا تأثیر کاهش بعد بر عملکرد مدل شفاف گردد. برای پی بردن به تعداد گره لایه پنهان که مدل با آن بهترین عملکرد را دارد نیز، این تعداد از ۲ تا ۱۰ تغییر داده و میزان صحت در هر تکرار بررسی شد که نتایج در شکل ۱ گزارش شد. همان‌گونه که در این شکل مشاهده شد، بهترین تعداد گره از نظر هزینه محاسبات، زمان و دقت، شش بود.

نتایج



شکل ۱: تغییرات صحت مدل LVQ به نسبت تغییرات تعداد گره‌های لایه پنهان شبکه

جهت بررسی عملکرد کلاسیفایر (Classifier)، پارامترهای صحت (Accuracy) ACC، حساسیت (Sensitivity) SE، ویژگی (Specificity) SP و ضریب همبستگی متیو (Mathew correlation coefficient) MCC (معادلات ۲-۵) محاسبه گردیدند که در آن TP (True Positive)، مثبت درست، TN (True Negative)، منفی درست، FP (False Positive)، مثبت نادرست و FN(False Negative)، منفی نادرست می‌باشد.

معادله (۲) $$SE = \frac{TP}{TP + FN}$$

معادله (۳) $$SP = \frac{TN}{TN + FP}$$

معادله (۴) $$ACC = \frac{TN + TP}{TP + TN + FP + FN}$$

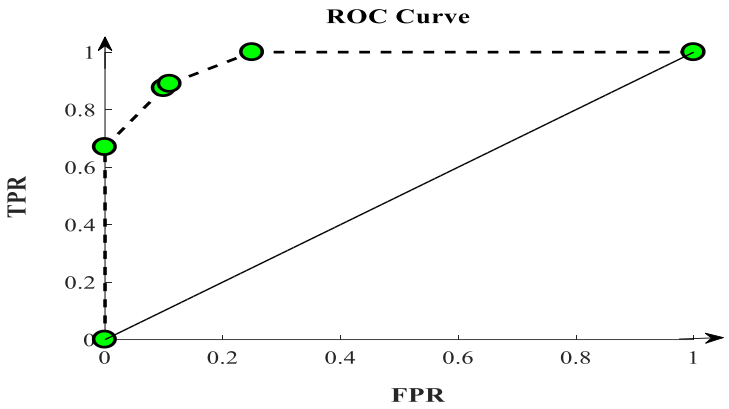
معادله (۵) $$MCC = \frac{TNTP - FPFN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

به منظور بررسی عملکرد الگوریتم ژنتیک نتایج مدل در دو حالت مختلف، پیش از کاهش بعد و بعد از آن، مورد ارزیابی قرار گرفت، همان‌گونه که مشاهده شد، پیاده‌سازی الگوریتم ژنتیک تأثیر بسزایی در بهبود عملکرد مدل به همراه داشت (جدول ۲).

جدول ۲: مقادیر صحت، حساسیت، ویژگی و ضریب همبستگی متیو در مدل طبقه‌بندی کننده LVQ برای داده‌ها قبل و بعد از انجام کاهش بعد

نوع داده برده شده	صحت (٪)	حساسیت (٪)	ویژگی (٪)	ضریب همبستگی متیو (٪)
پیش از کاهش بعد	۷۴/۴۵	۷۱/۴۳	۷۷/۰۸	۴۸/۵
بعد از کاهش بعد به روش الگوریتم ژنتیک	۸۶/۷	۸۳/۴	۸۹/۶	۷۳/۲

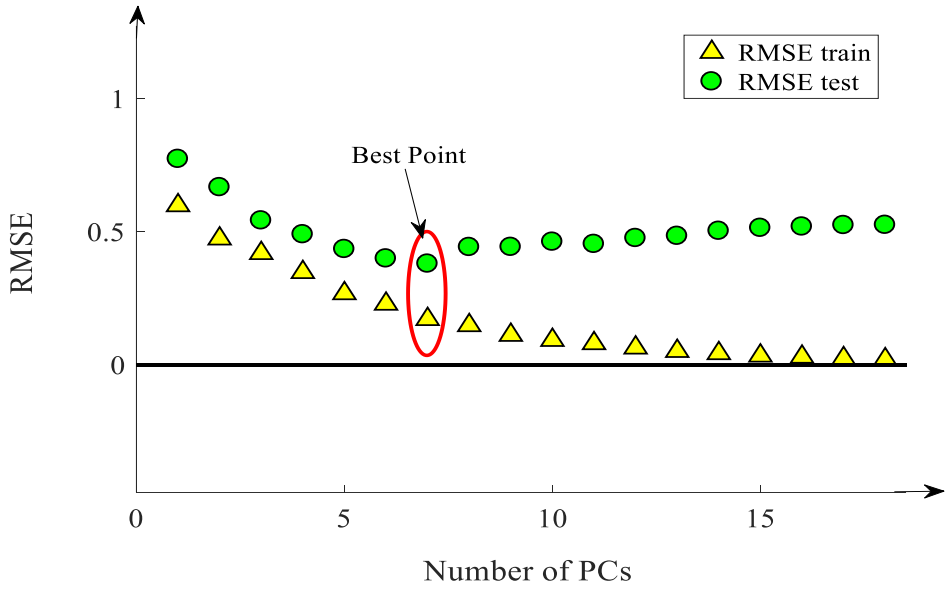
از طرف دیگر، جهت ارزیابی دقیق‌تر مدل، نمودار (Receiver Operating Characteristic)ROC مدل که منحنی با نمودار عمودی برحسب تعداد موارد تخمین زده شده مثبت که به‌درستی تخمین زده‌شده‌اند (TPR) و نمودار افقی برحسب تعداد موارد تخمین زده شده مثبت که به‌درستی تخمین زده نشده‌اند (FPR) نیز، به ازای تکرار مدل در فولدهای مختلف رسم و در شکل ۲ آورده شد.



شکل ۲: منحنی ROC کلاسیفایر LVQ

در پایان، جهت تخمین میزان اثر بیولوژیکی هر ریز مولکول، از مدل آماری PLS استفاده شد. جهت ارزیابی مدل PLS از روش اعتبارسنجی RMSE (معادله ۶)، میزان میانگین مینیمم خطای مربع را مشخص می‌کند، استفاده شد. $RMSE_{TRAIN}$ جهت بررسی خطای خروجی داده‌های آموزش و $RMSE_{TEST}$ جهت بررسی داده‌های تست به کار رفته است. به منظور انتخاب صحیح تعداد مؤلفه‌های بهینه و مؤثر در مدل PLS، برای یک مجموعه داده، تعداد اجزا، بین صفر تا ۱۸، تعداد داده‌های تست آن مجموعه، تغییر داده و مقادیر RMSE محاسبه شد که نتایج مربوطه در شکل ۳ قابل مشاهده است.

معادله (۶)
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_k - y'_k)^2}{n}}$$



شکل-۳: تغییرات RMSE برای داده‌های تست و آموزش به نسبت تغییرات تعداد اجزای مدل PLS

در این معادله، y_k خروجی مطلوب، y'_k خروجی تخمینی مدل و n تعداد داده‌ها است. همان‌گونه که در شکل ۳ مشاهده شد، بهترین تعداد مؤلفه جهت تخمین خروجی، از نظر هزینه محاسباتی و دقت مدل، هفت است. از طرفی، به منظور ارزیابی دقیق‌تر مدل PLS میزان همبستگی بین مقادیر خروجی تخمین زده شده توسط مدل و مقدار واقعی IC_{50} ها نیز مورد مطالعه قرار گرفت که این پارامتر از طریق معادله (۷) محاسبه گردید.

معادله (۷)
$$corr(y_k, y'_k) = \frac{cov(y_k, y'_k)}{\sigma_{y_k}\sigma_{y'_k}}$$

که در آن corr همبستگی، cov کوواریانس، y_k خروجی مطلوب، y'_k خروجی مدل و σ انحراف از معیار است. نتایج حاصله برای هر بار تکرار مدل طراحی شده در فولدهای مختلف در جدول ۳ آورده شد. در واقع، مدل PLS توانسته است به طور متوسط با همبستگی حدود ۰/۸۶ میزان IC_{50} ترکیبات را تخمین بزند.

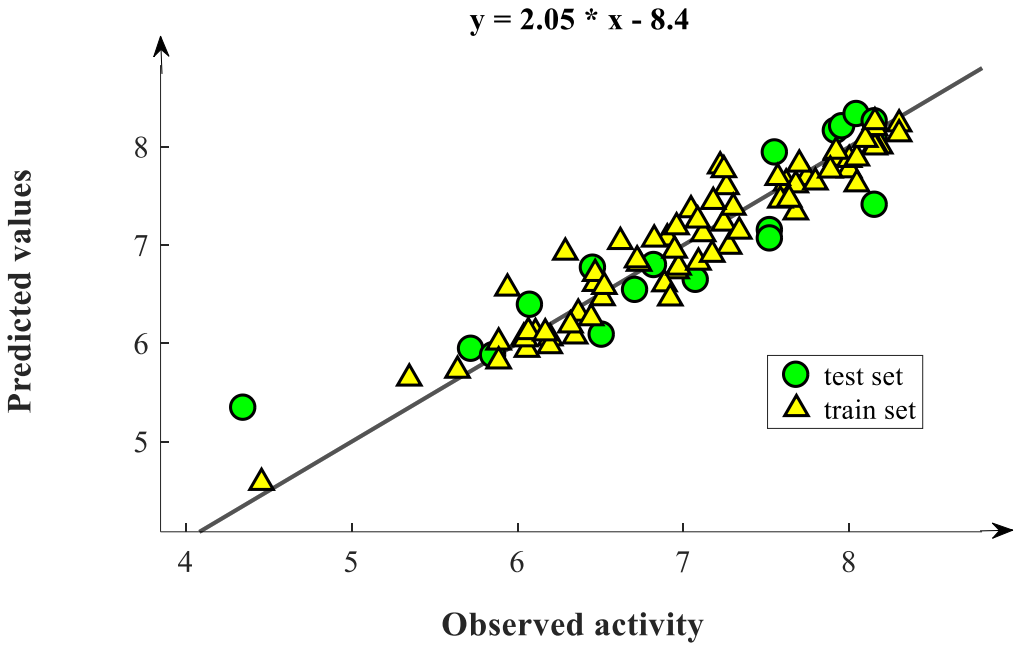
جدول ۳: میزان همبستگی میان مقادیر مطلوب و مقادیر پیش‌بینی‌شده توسط مدل PLS در فولدهای مختلف

شماره تکرار مدل	۱	۲	۳	۴	۵	میانگین
میزان همبستگی	۰/۸۹۴	۰/۸۱۵	۰/۸۶	۰/۹۱	۰/۸۰۹	۰/۸۵۸

جهت ارزیابی دقیق‌تر عملکرد مدل تخمینی PLS، منحنی پراکندگی مقادیر خروجی تخمین زده شده بر حسب IC_{50}های واقعی برای نمونه‌های آموزش و تست ترسیم شده است (شکل ۴). با توجه به نتایج به دست آمده به نظر می‌رسد رابطه بین مقادیر تخمینی و مقادیر حقیقی از معادله (۸) تبعیت می‌کند.

معادله (۸)
$$y = 2.05 * x - 8.4$$

که در آن x و y به ترتیب نشانگر مقادیر حقیقی و تخمینی می‌باشد.



شکل ۴: مقایسه مقادیر مطلوب میزان فعالیت ریز مولکول‌ها و مقادیر به دست آمده توسط مدل PLS برای داده‌های تست و آموزش، *مثلث‌های بنفش نمونه‌های آموزشی و دایره‌های سبز نمونه‌های تست می‌باشد.

بحث و نتیجه‌گیری

همان‌گونه که می‌دانید، هر بیماری حاصل به هم خوردن نظم طبیعی یکی از چرخه‌های فعالیت بدن است؛ بنابراین با دانستن این‌که کدام چرخه بدن از نظم طبیعی خود خارج شده است، می‌توان دارویی را به کار برد که دوباره آن را به حالت اولیه بازگرداند؛ اما ازآنجایی‌که کشف داروهای جدید در آزمایشگاه‌ها و به روش کلینیکالی بسیار پرهزینه و زمان‌بر می‌باشد؛ لذا شناسایی ریز مولکول‌های تأثیرگذار، بر اساس رفتار سایر ریز مولکول‌های مشابه باآنکه قبلاً مورد بررسی قرار گرفته‌اند، در کاهش هزینه‌های طراحی و کشف نمونه‌های جدید دارویی

بسیار مؤثر می‌باشد. از طرف دیگر، با توجه به رشد روزافزون سرطان در جوامع بشری و تأثیر بسیار زیاد آن از نظر روحی و روانی بر مردم جامعه، طراحی دارو برای گیرنده‌های مرتبط با رشد سلول‌های سرطانی از اهمیت ویژه‌ای در این حوزه برخوردار است که یکی از مهم‌ترین این گیرنده‌ها پروتئین‌های کیناز می‌باشد. پروتئین کیناز یک گروه مهم از آنزیم‌های کیناز است که پروتئین‌ها را با افزودن گروه فسفات (PO_4) فسفوریله می‌کند. بیش از سی درصد از پروتئین‌های بدن تحت تأثیر این آنزیم‌ها فسفوریله می‌شوند. گروه‌های فسفات می‌توانند فعالیت پروتئین را دچار اختلال کنند و باعث ایجاد بسیاری از بیماری‌ها از جمله سرطان گردند. مهار پروتئین کیناز امری حیاتی در درمان بعضی از سرطان‌ها و بیماری‌های التهابی است؛ بنابراین مهارکننده‌های پروتئین کیناز می‌توانند به عنوان دارو عمل کنند.

در این پژوهش، مدل طبقه‌بندی شبکه عصبی LVQ به منظور تمایز قائل شدن میان مهارکننده‌های پروتئین کیناز از غیر مهارکننده‌ها و مدل رگرسیون خطی PLS جهت تخمین مقادیر اثر بیولوژیکی (IC_{50}) هر ریز مولکول کاندید مهارکنندگی پروتئین کیناز، به کار گرفته شد. توصیفگرهای مولکولی که از نرم‌افزار دراگون استخراج می‌شوند، نقش بسیار مهمی در عملکرد مدل‌ها ایفا می‌کنند. به منظور کاهش ابعاد و جلوگیری از بیش برازش مدل‌ها، در این پژوهش از روش ژنتیک بهره گرفته شد. نتایج به دست آمده از اجرای مدل LVQ قبل و بعد از کاهش بعد، حاکی از آن است که این روش در عملکرد مدل نقش بسزایی ایفا می‌کند و توانسته است صحت مدل را از ۷۱٪/۴۵ به ۸۶/۷٪ برساند. واضح است که مدل LVQ بهترین عملکرد خود را در تعداد گره‌های لایه پنهان برابر با ۶، با صحت ۸۶/۷٪، حساسیت ۸۳/۴٪، اختصاصی بودن ۸۹/۶٪ و ضریب همبستگی متیو ۷۳/۲٪ ایفا خواهد کرد.

نتایج حاصل حاکی از آن است که مدل رگرسیون PLS نیز در تعداد کامپونت‌های برابر با ۷ کمترین مقدار RMSE را در داده‌های تست دارد. در تعداد بیش از ۶ کامپونت، RMSE داده‌های آموزش همچنان کاهش می‌یابند؛ اما RMSE داده تست رو به افزایش می‌رود که مطلوب نیست. مدل PLS با ارائه میزان همبستگی ۸۵/۸٪ بین مقادیر مطلوب و مقادیر به دست آمده از مدل، توانسته است تا حد قابل‌قبولی در تخمین اثر بیولوژیکی مهارکننده‌های پروتئین کیناز مفید واقع شود. لازم به ذکر است داده‌های این پژوهش از مطالعه Saghaie و همکاران [۱۴]، استخراج شد که در آن با استفاده از ویژگی‌های مستخرج از ساختمان مولکول‌ها سعی بر تخمین اثر بیولوژیکی ترکیبات با استفاده از روش PLS نموده و به همبستگی ۹۵٪ رسیده است؛ بنابراین به نظر می‌رسد در این دسته ترکیبات استفاده از ویژگی‌های ساختاری جهت تخمین اثر بیولوژیک بهتر از توصیفگرهای فیزیکو-شیمیایی عمل می‌کند. همچنین با توجه به این که مدل طبقه‌بندی PLS به کار برده شده در این مطالعه، در مقالات دیگری نیز مورد بررسی قرار گرفته و مطلوب بودن عملکرد آن تأیید شده است [۲۲-۲۰]، بنابراین واضح است تفاوت نتایج در این مطالعه و مطالعه مذکور تنها در نوع انتخاب ویژگی‌ها بوده است و مدل طبقه‌بندی PLS عملکرد مناسبی را به اجرا گذاشته است.

با توجه به اهمیت مسئله کاهش بعد در روش‌های یادگیری ماشین، لازم است در تحقیقات آینده روش‌های بیشتری در کنار روش الگوریتم ژنتیک مورد تجزیه‌وتحلیل قرار گیرد.

تشکر و قدردانی

این مطالعه مستخرج از طرح تحقیقاتی، مصوب با شماره ۲۹۷۱۷۴ و با حمایت معاونت آموزشی و پژوهشی دانشگاه علوم پزشکی اصفهان انجام شد.

تعارض منافع

نویسندگان تعارض منافع نداشتند.

References

1. Lu Z, Hunter T. Metabolic kinases moonlighting as protein kinases. Trends Biochem Sci 2018;43(4):301-10. doi: 10.1016/j.tibs.2018.01.006.

2. Ghasemi F, Mehridehnavi A, Fassihi A, Pérez-Sánchez H. Deep neural network in QSAR studies using deep belief network. Applied Soft Computing. 2018;62:251-8. doi.org/10.1016/j.asoc.2017.09.040

3. Mostashari-Rad T, Arian R, Sadri H, Mehridehnavi A, Mokhtari M, Ghasemi F, et al. Study of $CXCR_4$ chemokine receptor inhibitors using QSPR and molecular docking methodologies. Journal of Theoretical and Computational Chemistry 2019;18(04):1950018. doi: 10.1142/S0219633619500184

4. Hughes ZE, Walsh TR. Structural Disruption of an Adenosine-Binding DNA Aptamer on Graphene: Implications for Aptasensor Design. ACS Sens. 2017;2(11):1602-11. doi: 10.1021/acssensors.7b00435.

5. Tahir M, Hayat M, Kabir M. Sequence based predictor for discrimination of enhancer and their types

by applying general form of Chou's trinucleotide composition. Computer Methods and Programs in Biomedicine 2017;146:69-75. doi.org/10.1016/j.cmpb.2017.05.008

6. Tiwari AK, Srivastava R. An Efficient Approach for Prediction of Nuclear Receptor and Their Subfamilies Based on Fuzzy k-Nearest Neighbor with Maximum Relevance Minimum Redundancy. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 2018;88(1):129-36. doi: 10.1007/s40010-016-0325-6

7. Derevyanko G, Grudinin S, Bengio Y, Lamoureux G. Deep convolutional networks for quality assessment of protein folds. Bioinformatics 2018;34(23):4046-53. doi: 10.1093/bioinformatics/bty494

8. Geng H, Lu T, Lin X, Liu Y, Yan F. Prediction of Protein-Protein Interaction Sites Based on Naive Bayes Classifier. Biochemistry Research International 2015:1-7.

9. Ghasemi F, Mehridehnavi A, Fassihi A, Pérez-Sánchez H. Deep neural network in QSAR studies using deep belief network. Applied Soft Computing 2018;62:251-8. doi.org/10.1016/j.asoc.2017.09.040

10. Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou KC, Webb GI. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. Journal of Theoretical Biology. 2018;443:125-37. doi.org/10.1016/j.asoc.2017.09.040

11. Wei L, Xing P, Tang J, Zou Q. PhosPred-RF: A Novel Sequence-Based Predictor for Phosphorylation Sites Using Sequential Information Only. IEEE Trans Nanobioscience 2017;16(4):240-7. doi: 10.1109/TNB.2017.2661756.

12. Alladio E, Giacomelli L, Biosa G, Corcia DD, Gerace E3, Salomone A, et al. Development and validation of a Partial Least Squares-Discriminant Analysis (PLS-DA) model based on the determination of ethyl glucuronide (EtG) and fatty acid ethyl esters (FAEEs) in hair for the diagnosis of chronic alcohol abuse. Forensic Sci Int 2018;282:221-30. doi: 10.1016/j.forsciint.2017.11.010.

13. Alghamedy F, Bopaiah J, Jones D, Zhang X, Weiss HL, Ellingson SR. Incorporating Protein Dynamics Through Ensemble Docking in Machine Learning Models to Predict Drug Binding. AMIA Jt Summits Transl Sci Proc 2018;2017:26 - 34.

14. Saghaie L, Shahlaei M, Madadkar-Sobhani A, Fassihi A. Application of partial least squares and radial basis function neural networks in multivariate imaging analysis-quantitative structure activity relationship: study of cyclin dependent kinase 4 inhibitors. Journal of Molecular Graphics and Modelling. 2010;29(4):518-28. doi.org/10.1016/j.jmgm.2010.10.001

15. Ghasemi F, Fassihi A, Pérez-Sánchez H, Mehri Dehnavi A. The role of different sampling methods in improving biological activity prediction using deep belief network. J Comput Chem 2017;38(4):195-203. doi: 10.1002/jcc.24671.

16. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. MATCH Communications in Mathematical and in Computer Chemistry 2006;56(2):237-48.

17. Fang J, Yang R, Gao L, Zhou D, Yang S, Liu AL, et al. Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. J Chem Inf Model 2013;53(11):3009-20. doi: 10.1021/ci400331p.

18. Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM. Advances with support vector machines for novel drug discovery. Expert Opin Drug Discov 2019;14(1):23-33. doi: 10.1080/17460441.2019.1549033.

19. Korkmaz S, Zararsiz G, Goksuluk D. MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. PLoS One 2015;10(4):e0124600. doi: 10.1371/journal.pone.0124600.

20. Ahmadi M, Shahlaei M. Quantitative structure-activity relationship study of P2X7 receptor inhibitors using combination of principal component analysis and artificial intelligence methods. Res Pharm Sci 2015;10(4):307-25.

21. Shahlaei M, Bahrami G, Abdolmaleki S, Sadrjavadi K. Majnooni MB. Application of unfolded principal component analysis–radial basis function neural network for determination of celecoxib in human serum by three-dimensional excitation–emission matrix fluorescence spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2015; 138.675-83. doi.org/10.1016/j.saa.2014.12.007

22. Shahlaei M, Madadkar-Sobhani A, Fassihi A, Saghaie L, Arkan E. QSAR study of some CCR5 antagonists as anti-HIV agents using radial basis function neural network and general regression neural network on the basis of principal components. Medicinal Chemistry Research 2012;21(10):3246-62. doi: 10.1007/s00044-011-9863-2

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 7(1): 30-39

Improving biological activity prediction of protein kinase inhibitors using artificial neural network and partial least square methods

Arian Roya[1], Mehri Dehnavi Alireza[2], Ghasemi Fahimeh[3]*

• **Received:** 22 Apr, 2019 • **Accepted:** 2 Nov, 2019

Introduction: Protein kinase causes many diseases, including cancer; therefore, inhibiting them plays an important role in the treatment of many diseases. Traditional discovery inhibitors of this enzyme is a time-consuming and costly process. Finding a reliable computer-aided drug discovery tools which can detect the inhibitors will reduce the cost. In this study, it is attempted to separate kinase inhibitors into two groups, active and inactive, using artificial neural network and finally predict biological activities of the predicted active compounds by partial least square .

Method: In this study, after extracting the molecular descriptors in order to avoid overfitting problem, dimensional reduction was applied using Genetic algorithm. Moreover, artificial neural network was applied to distinguish active compounds from inactive ones and the biological activities of the small molecules were predicted using partial least square linear regression.

Results: The results show that accuracy of the Neural networkmodel was improved from 74.45% to 86.7%, after reducing molecular descriptor dimensions. . The number of hidden nodes of this model was six with 86.7% accuracy, 83.4% sensitivity, 89.6% specificity and 73.2% Mathew's correlation coefficient. Moreover the partial least square linear regression model predicts the biological activity valuesby 85.8% correlation.

Conclusion: The Neural network model and the partial least square linear regression model can sufficiently predict Kinase inhibitors and Genetic algorithm will improve the models performance.

Keywords: Protein kinase, Classification, Neural network, Regression, partial least square

• **Citation:** Arian R, Mehri Dehnavi AR, Ghasemi F. Improving biological activity prediction of protein kinase inhibitors using artificial neural network and partial least square methods. Journal of Health and Biomedical Informatics 2020; 7(1): 30-9. [In Persian

1. M.Sc. in Bioelectrics, Bioelectrics Dept., School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
2. Ph.D. in Bioelectrics, Professor, Bioelectrics Dept., School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
3. Ph.D. in Bioelectrics, Assistant Professor, Bioinformatics Dept., School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
***Corresponding Author**: Fahimeh Ghasemi
Address: Isfahan University of Medical Sciences Hezar Jarib St., Isfahan
• **Tel:** 031-37923865 • **Email:** f_ghasemi@amt.mui.ac.ir