استخراج و ترکیب ویژگی‌های کارآمد از توالی پروتئین به‌منظور دسته‌بندی پروتئین بر اساس جنگل چرخش

جمشید پیرگزی[1]، علی قنبری سرخی[2*]، مجید ایرانپور مبارکه[3]

[1و2] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه علم و فناوری مازندران، بهشهر، ایران

[3] استادیار، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، تهران، ایران

چکیده

پیش‌بینی عملکرد پروتئین یکی از چالش‌های اصلی در بیوانفورماتیک است که کاربردهای زیادی دارد. در سال‌های اخیر در پژوهش‌های بسیاری از روش‌های یادگیری ماشین در این زمینه استفاده شده‌است. در این روش‌ها ابتدا باید از توالی پروتئین ویژگی‌های مختلف استخراج و بر اساس ویژگی‌های استخراج‌شده عمل دسته‌بندی انجام شود. اغلب روش‌های استخراج ویژگی براساس خصوصیات فیزیکی و شیمایی توالی پروئتین است؛ بنابراین استخراج ویژگی‌هایی مناسب از توالی پروتئین باعث افزایش و بهبود عملکرد روش‌های یادگیری ماشین می‌شود. در این مقاله، یک مجموعه جدید از ویژگی‌ها بر اساس روش‌های PSSM، PsePSSM، K-gram، AAC و روش نوین TFCRF که تاکنون در این کاربرد استفاده نشده برای استخراج ویژگی‌های مناسب پیشنهاد شده‌است. ویژگی‌های استخراج‌شده با استفاده از این روش قدرت تمایزکنندگی خوبی بین داده‌ها در دسته‌ها، به مدل‌های یادگیری ماشین می‌دهد. در روش TFCRF وزن‌دهی ویژگی‌ها علاوه بر توجه به چگونگی توزیع آنها در توالی‌های مختلف به چگونگی توزیع آنها در طبقات مختلف نیز توجه می‌شود. در مرحله بعد با استفاده از ویژگی‌های استخراج‌شده با استفاده از روش جنگل چرخ عمل دسته‌بندی انجام می‌شود. روش پیشنهادی با دسته‌بندهای مختلف و روش‌های متفاوت مقایسه شده‌است. نتایج حاصل نشان‌دهنده کارایی مناسب روش پیشنهادی نسبت به سایر روش‌های نوین در این کاربرد است.

واژگان کلیدی: توالی پروتئین، استخراج ویژگی، TFCRF، جنگل چرخش، فاکتور ارتباط.

Extracting and combination efficient feature from protein sequence for classify protein based on rotation forest

Jamshid Pirgazi[1], Ali Ghanbari Sorkhi[2*], Majid Iranpour Mobarakeh[3]

[1,2] Faculty of Electrical and Computer Engineering, University of Science and Technology of Mazandaran, Behshahr, Iran.

[3] Department of Computer Engineering and IT, Payam Noor University, Tehran, Iran.

Abstract

Protein function prediction is one of the main challenges in bioinformatics, which has many applications. In recent years, many researches in this field have been used machine learning methods. In these methods, First, different features should be extracted from the protein sequence and classification should be done based on the extracted features. The feature extraction methods are based on the physical and chemical properties of the protein sequence. Therefore, extracting suitable features from protein sequence increases and improves the performance of machine learning methods. In this paper, usage of a new set of features based on Position-Specific Scoring Matrix (PSSM), Pseudo-Position Specific Scoring Matrix (PsePSSM), K-gram, Amino Acid Composition (AAC) and the new Term Frequency and Category Relevancy Factor (TFCRF) method, which has not been used in this application so far, is proposed to extract suitable features.

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۲/۴/۱۰ • تاریخ پذیرش: ۱۴۰۲/۱۲/۶ • تاریخ انتشار: ۱۴۰۲/۴/۱۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.1.15]

In the PSSM method for protein BLAST searches, a scoring matrix is used, in which amino acid substitution scores are given separately for each position in a multi-sequence protein alignment. The PsePSSM feature is described by considering different ranking correlation factors along a protein sequenc to preserve information about the amino acid sequence. The normalized occurrence frequency of a certain number of amino acids in the protein is calculated by the ACC method. An K-gram is a set of K successive items in a protein that include amino acid.
In the TFCRF weighting method, in addition to paying attention to how these are distributed in different sequences, how these are distributed in different classes is also paid attention to.The features extracted using this method give machine learning models a good discriminating power between data in classes. In the next step, classification is done using the extracted features using the rotation forest method. This classifier is a successful ensemble method for a wide range of data mining applications. In this method, the feature space is changed through Principal Component Analysis (PCA), which increases the power of this classifier. The proposed method has been compared to different classifiers. The results show that the efficiency of the proposed method is much better than other state-of–the-art methods in this application.

Keywords Protein sequence, feature extraction, TFCRF, rotation forest, relevancy factor.

۱- مقدمه

پیش‌بینی عملکرد پروتئین شامل روش‌هایی است که پژوهش‌گران حوزه بیوانفورماتیک استفاده می‌کنند تا نقش بیولوژیکی یا بیوشیمیایی را به پروتئین‌ها نسبت دهند. دسته‌بندی عملکرد پروتئین یکی از چالش‌های بیوانفورماتیک است؛ بنابراین کاربردهای زیادی از حاشیه‌نویسی ژنومی[1] تا کاربردهای بالینی، مانند پیش‌بینی مقاومت دارویی در ویروس نقص ایمنی انسانی (HIV) برای درمان‌های شخصی [۱] دارد. درهمین اواخر استفاده از روش‌های محاسباتی برای دسته‌بندی عملکرد استفاده شده‌است. یکی از روش‌های محبوب در این حوزه روش‌های یادگیری ماشین است. در مطالعات قبلی، از روش‌های یادگیری ماشین مانند ماشین بردار پشتیان (SVM)، شبکه های عصبی (NN)، جنگل تصمیم (DT) و غیره استفاده شده‌است [۲]. برای دسته‌بندی عملکرد پروتئین با استفاده از روش‌های یادگیری ماشین، ابتدا توالی پروتئین به‌صورت برداری از ویژگی‌ها تبدیل می‌شود. این روش‌ها اعلب به دو دسته روش‌های مبتنی بر توالی و روش‌های مبتنی بر ساختار تقسیم می‌شود. روش‌های استخراج ویژگی مبتنی بر توالی شامل رمزنگاری پراکنده، ترکیب اسیدآمینه ها، الفبای اسیدآمینه کاهش‌یافته خواص فیزیک و شیمیایی یا تبدیل فوریه است [۳, ۴]. روش‌های مبتنی بر ساختار شامل رابطه فعالیت کمی ساختار[2] (QSAR)، بدنه الکترواستاتیک یا مثلث‌سازی دلونی اشاره کرد [۵]. در [۶] روش‌های مختلف استخراج ویژگی از توالی پروتئین مورد مطالعه و بررسی شده‌است. بعد از استخراج ویژگی، این ویژگی‌ها برای آموزش مدل‌های مختلف یادگیری ماشین استفاده می‌شود.

استخراج ویژگی‌های مختلف از پروتئین باعث می‌شود که جنبه‌های مختلف پروتیئن در نظر گرفته شود، اما این‌کار باعث می‌شود که طول بردار ویژگی استخراج‌شده زیاد شود. زمانی که تعداد ویژگی زیاد باشد و تعداد نمونه‌ها کم، بیش‌تر روش‌های یادگیری ماشین دچار مشکل بیش‌برازش می‌شود. به بیان دیگر، در این حالت همه ویژگی در دسته‌بندی عملکرد پروتئین تأثیر نمی‌گذارند. بعضی از ویژگی‌ها غیرمرتبط و بعضی از ویژگی‌ها زائد و افزونه هستند. این دو نوع ویژگی باید از بردار ویژگی حذف شوند. برای این منظور در بسیاری از پژوهش‌ها بر روی روش‌های انتخاب ویژگی تمرکز کرده‌اند [۷]. در [۸] با استفاده از نمایش فضای عدم شباهت سعی شده‌است ویژگی‌هایی که شباهتی بین آنها نیست از مجموعه ویژگی‌های استخراج‌شده از توالی پروتیئن حذف شوند. به‌منظور انتخاب ویژگی‌های مرتبط استفاده از روش‌های مبتنی بر فیلتر و بسته‌بندی[3] در [۹] مورد توجه قرار گرفته است. در [۱۰] به‌منظور پیش‌بینی عملکرد پروتئین از روش تکاملی الگوریتم ژنتیک چندهدفه استفاده شده‌است. در این روش علاوه‌بر اینکه سعی شده‌است ویژگی‌هایی انتخاب شوند که دقت مدل‌های دسته‌بندی بالا باشد، سعی شده، تعداد ویژگی‌های انتخابی کمینه باشد. روش‌های مبتنی بر الگوریتم‌های تکاملی باعث می‌شوند که بتوان فضای جستجو برای پیداکردن زیرمجموعه بهینه از ویژگی‌ها را جستجو کرد؛ ولی این روش‌ها دارای پیچیدگی زمانی بالایی هستند؛ بنابراین در [۱۱] روشی مبتنی بر همبستگی برای انتخاب ویژگی‌های مؤثر ارائه شده‌است. در این روش همبستگی بین ویژگی‌ها و برچسب داده‌ها محاسبه می‌شود و ویژگی‌هایی که همبستگی زیادی دارند از مجموعه ویژگی‌ها حذف

[1] Genome Annotation
[2] Quantitative Structure–Activity Relationship
[3] wrapper

[DOI: 10.61186/jsdp.21.1.15]

می‌شوند. در [۱۲, ۱۳] روش‌های انتخاب ویژگی به صورت دقیق بیان شده‌است و برنامه کاربردی برای پیش‌بینی عملکرد پروتئین ارائه شده‌است.

بعد از استخراج و انتخاب ویژگی، ویژگی‌های انتخاب‌شده برای آموزش مدل‌های مختلف یادگیری ماشین استفاده می‌شوند. بسیاری از پژوهش‌گران سعی کرده‌اند مدلی مناسب مبتنی بر روش‌های یادگیری ماشین به‌منظور پیش‌بینی عملکرد پروتئین ارائه دهند [۱۴].

روش‌های مورد استفاده برای پیش‌بینی عملکرد پروتیئن به دو دسته مدل‌های یادگیری ماشین کلاسیک و روش‌های یادگیری عمیق تقسیم می‌شوند. در [۱۵] به‌منظور پیش‌بینی عملکرد پروتیئن‌ها در گیاهان از ابزار قدرتمند مبتنی بر روش‌های یادگیری ماشین استفاده و در این روش، ابتدا از توالی پروتئین ویژگی‌های ساختاری استخراج شده‌است. سپس با استفاده از جنگل تصمیم عملکرد پروتیئن پیش‌بینی می‌شود. در این مقاله انواع مختلف مدل‌های طبقه‌بندی ارزیابی شده‌است. به‌منظور افزایش نرخ تشخیص مدل‌های طبقه‌بند استفاده از روش‌های ترکیبی مورد توجه قرار گرفته شده‌است. در [۱۶] با استفاده از ترکیب روش‌های شبکه‌های عصبی و درخت تصمیم نرخ پیش‌بینی عملکرد پروتئین افزایش یافته است. روش NetGO برای پیش‌بینی عملکرد پروتئین در شبکه‌های با مقیاس بزرگ ارائه شده‌است. ویژگی‌های مختلف شامل ویژگی‌های مبتنی بر توالی، ویژگی‌های ساختاری و غیره از پروتئین استخراج و سپس با استفاده از روش‌های یادگیری ماشین پیش‌بینی عملکرد انجام می‌شود [۱۷].

مشکلی که روش‌های کلاسیک یادگیری ماشین دارند این است که باید مرحله مهندسی ویژگی یا استخراج ویژگی به بهترین شکل ممکن انجام شود. درواقع در این نوع روش‌ها باید روش‌های استخراج ویژگی، ویژگی‌های متمایزکننده را استخراج کنند تا عملکرد مدل‌های یادگیری ماشین قابل قبول باشد؛ بنابراین اخیرا استفاده از روش‌های یادگیری عمیق مورد توجه قرار گرفته شده‌است. زیرا این روش‌ها نیاز به مرحله‌ی استخراج ویژگی ندارند [۱۸].

در [۱۹] برای استخراج ویژگی‌های مناسب جهت پیش‌بینی عملکرد پروتئین برپایه توالی پروتئین از روش‌های یادگیری عمیق بازگشتی استفاده شده‌است. روش DeepGoPlus به‌منظور بهبود پیش‌بینی عملکرد پروتئین ارائه شده‌است. در این روش از شبکه‌های عمیق پیچشی به همراه ماتریس شباهت توالی استفاده و روش پیشنهادی بر روی مجموعه‌داده‌های مختلف ارزیابی شده‌است. نتایج نشان می‌دهد روش پیشنهادی دارای عملکرد قابل قبولی است [۲۰]. در [۲۱] روشی مبتنی بر آنتولوژی ژن و شبکه‌های عمیق رو به جلو ارائه شده‌است. در این روش ابتدا با استفاده از آنتولوژی ژن ویژگی‌های مناسب از پروتئین‌ها استخراج، سپس با استفاده از شبکه‌های عمیق چندلایهٔ روبه‌جلو عملکرد پروتئین‌ها پیش‌بینی می‌شود. استفاده از آنتولوژی ژن‌ها در [۲۲] نیز مورد توجه قرار گرفته شده‌است. در این روش بر اساس تعاملات بین ژن‌های تشکیل‌دهنده یک پروتیئن از شبکه‌های پیچشی گراف استفاده شده‌است.

هر چند روش‌های یادگیری عمیق محبوبیت بالایی دارند، اما این روش‌ها زمانی که تعداد داده‌ها کم باشد درعمل قابل استفاده نیستند؛ زیرا با توجه به تعداد زیاد پارامترهای که این مدل‌ها باید بر اساس داده‌ها تنظیم کنند، زمانی که تعداد داده‌ها کم باشد این پارامترها به‌درستی تنظیم نمی‌شوند و باعث بیش‌برازش مدل‌های مبتنی بر یادگیری عمیق می‌شود.

به همین منظور در این مقاله ابتدا با استفاده از روش‌های مختلف مانند PSSM، PsePSSM، K-gram، AAC[1] ویژگی‌های مناسب از توالی پروتئین استخراج می‌شود. علاوه‌براین در این مقاله ویژگی جدیدی از توالی پروتئین بر اساس روش TFCRF استخراج می‌شود. این روش بر خلاف سایر روش‌ها در وزن‌دهی علاوه بر اینکه به چگونگی توزیع آنها در توالی مختلف توجه می‌کند به چگونگی توزیع آنها در طبقات مختلف نیز توجه می‌کند. این امر سبب می‌شود ویژگی‌ها در دسته‌های مختلف اهمیت و قدرت تمایزکنندگی متفاوتی داشته باشند. این نوع ویژگی‌ها باعث افزایش نرخ تشخیص مدل‌های یادگیری ماشین می‌شود. در مرحله بعد با استفاده از روش ترکیبی جنگل چرخش عملکرد پروتئین‌ها پیش‌بینی می‌شود.

مقاله حاضر شامل چهار بخش است. در بخش دوم روش پیشنهادی و قسمت‌های مختلف روش پیشنهادی به‌تفصیل بیان می‌شود. در بخش سوم، تحلیل نتایج و معرفی پایگاه داده و مقایسات انجام‌شده با پژوهش‌های دیگر ارائه شده‌است. درنهایت در بخش چهارم، نتیجه‌گیری و کارهای آینده تشریح می‌شود.

[1] Amino acid composition (AAC)

[DOI: 10.61186/jsdp.21.1.15]

۲-روش پیشنهادی

در این مقاله به‌منظور پیش‌بینی توالی پروتئین، ابتدا داده‌ها به دو قسمت آموزش و آزمون تقسیم، سپس از داده‌های هر دو قسمت ویژگی‌های مختلف استخراج می‌شود. این ویژگی باید قدرت متمایزکنندگی زیادی داشته باشند. برای این منظور ویژگی‌های مختلف از توالی پروتئین استخراج می‌شود. برای این منظور، از چهار روش استخراج ویژگی به همراه روش پیشنهادی برای استخراج ویژگی استفاده شده‌است، سپس در مرحله بعد مقدار ویژگی‌ها نرمال می‌شوند. این کار باعث بهبود مدل‌های دسته‌بندی می‌شود؛ درنهایت با استفاده از روش‌های دسته‌بندی عمل دسته‌بندی توالی پروتئین انجام می‌شود. مدل دسته‌بندی استفاده‌شده در این مقاله جنگل چرخش است. برای این منظور در ادامه روش استخراج ویژگی پیشنهادی و روش جنگل چرخش با جزئیات بیان می‌شود.

۲-۱-استخراج ویژگی با استفاده از روش PSSM و PsePSSM

برای نشان‌دادن ویژگی‌های توالی اسیدآمینه (AA) برای توالی‌های پروتئین، ویژگی‌های ماتریس امتیازدهی شبه موقعیت خاص[1] (PsePSSM) [۲۳] استفاده می‌شود. این روش تکامل و اطلاعات توالی پروتئین را رمزگذاری می‌کند که به‌طور گسترده در پژوهش‌های بیوانفورماتیک استفاده شده‌است [۲۴].

برای هر دنباله هدف P با L اسیدآمینه، PSSM به عنوان توصیف‌کننده آن پیشنهادشده توسط [۲۵] استفاده می‌شود. ماتریس امتیازدهی ویژه موقعیت[2] (PSSM) با ابعاد L×20 را می‌توان به‌صورت رابطه (۱) تعریف کرد:

$$P_{PSSM} = \begin{bmatrix} M_{1\to1} & M_{1\to2} & \dots & M_{1\to20} \\ M_{2\to1} & M_{2\to2} & \dots & M_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{i\to1} & M_{i\to2} & \dots & M_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{L\to1} & M_{L\to2} & \dots & M_{L\to20} \end{bmatrix} \quad (۱)$$

در این رابطه $M_{i\to j}$ امتیاز اسیدآمینه را در موقعیت iام از توالی پروتئین را نشان می‌دهد که در طول فرآیند تکامل به اسیدآمینه نوع j جهش یافته است. در اینجا برای ساده‌سازی فرمول از کدهای عددی ۱، ۲، . . .، ۲۰ برای نشان‌دادن ۲۰ نوع اسیدآمینه بر اساس ترتیب حروف الفبای کدهای تک‌نویسه‌ای آن‌ها استفاده شده‌است [۲۶].

هر عنصر در ماتریس PSSM اصلی با استفاده از رابطه (۲) به بازه (۰، ۱) نرمال شد:

$$\bar{M}_{i\to j} = \frac{1}{1+\exp\left(-M_{i\to j}\right)} \quad (۲)$$

به‌دلیل طول‌های مختلف در توالی‌های هدف، ساختن توصیفگر PSSM به‌عنوان یک نمایش یک‌نواخت می‌تواند مفید باشد، یکی از نمایش‌های ممکن از نمونه پروتئین P به‌صورت رابطه (۳) است:

$$\bar{P}_{PSSM} = [\bar{M}_1.\bar{M}_2.\dots.\bar{M}_{20}] \quad (۳)$$

اگر در روند تکامل از P_{PSSM} رابطه (۱) استفاده شود تمام اطلاعات ترتیب توالی از بین می‌رود. برای جلوگیری ازدست‌دادن کامل اطلاعات ترتیب توالی، مفهوم ترکیب اسیدآمینه کاذب توسط [۲۷] معرفی شده‌است، که PsePSSM برای نشان‌دادن پروتئین P به‌صورت رابطه (۴) معرفی می‌شود:

$$P^{\lambda}_{psePSSM} = [\bar{M}_1.\bar{M}_2.\dots.\bar{M}_{20}.G^1_1.G^1_2.\dots.G^1_{20}.G^{\lambda}_1.G^{\lambda}_2.\dots.G^{\lambda}_{20}]^T \quad (۴)$$

که در آن

$$G^{\lambda}_j = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\left[\bar{M}_{i\to j} - \bar{M}_{(i+\lambda)\to j}\right]^2 \quad (j = 1.2.\dots.20;\ 0 \le \lambda \le L) \quad (۵)$$

در این رابطه G^{λ}_j نشان‌دهنده ضریب همبستگی j-امین اسیدآمینه و λ فاصله پیوسته در امتداد دنباله پروتئین است؛ بنابراین، یک توالی پروتئین را می‌توان به مشابه معادله (۳) با استفاده از PsePSSM تعریف کرد و یک بردار ویژگی ۲۰+۲۰×λ بعدی تولید کرد. در این مطالعه، λ روی ۱۰ تنظیم شده‌است. بعد بردار ویژگی هر پروتئین هدف برای توصیف‌گر PsePSSM، ۲۲۰ است.

۲-۲-استخراج ویژگی با استفاده از روش K-gram

در حوزه‌های پردازش متن و احتمالات، k-gram دنباله‌ای پیوسته از k قلم در یک دنباله معین از متن است. در این مقاله، اقلام، اسیدآمینه های موجود در توالی پروتئین هستند. k می‌تواند مقادیر متفاوتی داشته باشد که در این مقاله ۱ و ۲ در نظر گرفته شده‌است. فرکانس نسبی هر ۲۱ نوع اسیدآمینه (۲۰ اسیدآمینه استاندارد و کد غیر واقعی O وقتی طول پروتئین‌ها برابر نباشد) با

[1] pseudo-position specific scoring matrix
[2] position-specific scoring matrix (PSSM)

[DOI: 10.61186/jsdp.21.1.15]

استفاده از رابطه (۶) در 1-gram محاسبه می‌شود:

$$f(r) = \frac{N_r}{N} \qquad r = 1.2. \ldots .21 \qquad (۶)$$

که در آن N_r تعداد اسیدآمینه r را مشخص و N طول مقطع را مشخص می‌کند. در نتیجه، یک بردار ۲۱ بعدی برای هر بخش به دست می‌آید. 2-gram فرکانس نسبی همه دی‌پپتیدهای احتمالی را در دنباله محاسبه می‌کند. بردار ویژگی صورت رابطه (۷) بیان می‌شود:

$$f(r.s) = \frac{N_{rs}}{N-1} \qquad r.s = 1.2. \ldots .21 \qquad (۷)$$

در این رابطه، N_{rs} تعداد دی پپتید rs، N طول مقطع و N-1 تعداد کل دی‌پپتیدها را در بخش رمزگذاری‌شده نشان می‌دهد.

۲-۳-استخراج ویژگی با استفاده از روش AAC

ترکیب اسیدآمینه [۲۸] بردار بیست‌بعدی است که فرکانس هر ۲۰ اسیدآمینه طبیعی (یعنی "ACDEFGHIKLMNPQRSTVWY") را به‌صورت رابطه (۸) محاسبه می‌کند:

$$f_t = \frac{N(t)}{N} \qquad t \in \{A.C.D. \ldots . Y\} \qquad (۸)$$

۲-۴- استخراج ویژگی با استفاده از روش TFCRF

در این مقاله به‌منظور استخراج ویژگی‌های موثر از توالی پروتئین روش TFCRF[1] ارائه شده‌است. در این روش برای وزن‌دهی به ویژگی‌ها که در اینجا تعداد تکرار مربوط به هر اسیدآمینه در توالی پروتئین است از دو عامل، عامل ارتباط مثبت (positiveRF) و عامل ارتباط منفی (negativeRF) بر اساس مرجع [۲۹] که بر روی متون پیشنهاد شد، استفاده شده‌است. شکل (۱)، مراحل استخراج وزن مبتنی بر روش TFCRF را نشان می‌دهد. در عامل positiveRF نسبت تعداد اسیدآمینه‌ها را در یک توالی پروتئین (c_j) که شامل یک ویژگی مشترک (t_k) هستند، به کل تعداد اسیدآمینه های توالی پروتئین نشان می‌دهد که توسط رابطه (۹) محاسبه می‌شود:

$$PosotiveRF(t_k.c_i) = \frac{|D(t_k.c_j)|}{|D(c_j)|} \qquad (۹)$$

علاوه‌بر این عامل negativeRF، نسبت مجموع تعداد اسیدآمینه‌هایی را که در سایر توالی پروتئین به‌جز (c_j) که حاوی ویژگی مشترک (t_k) هستند، به کل تعداد اسیدآمینه‌های توالی پروتئین به غیر از پروتئین (c_j) نشان می‌دهد. که توسط رابطه (۱۰) محاسبه می‌شود:

$$NegativeRF(t_k.c_i) = \frac{\sum_{m-1.m\neq j}^{|c|} |D(t_k.c_m)|}{\sum_{m-1.m\neq j}^{|c|} |D(c_m)|} \qquad (۱۰)$$

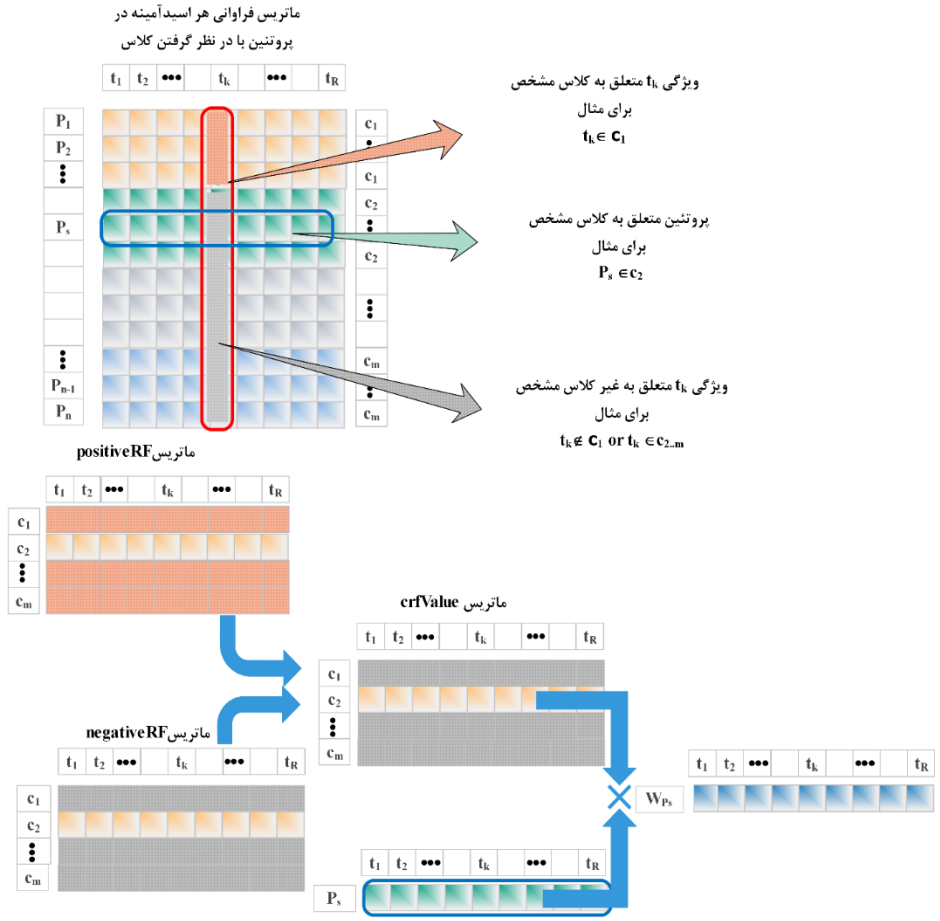
در روابط بالا $|D(c_j)|$ برابر با تعداد اسیدآمینه‌هایی از پروتئین c_j و $|D(t_k,c_j)|$ برابر با تعداد اسیدآمینه‌هایی از مجموعه (D) و پروتئین c_m که دارای ویژگی مشترک t_k هستند، است. همان‌طورکه در شکل (۱) نشان داده شده، خروجی این دو عامل، دو ماتریس با ابعاد تعداد دسته در تعداد ویژگی (m×R) است. در ادامه برای بررسی ارزش هر ویژگی در دسته متناظر با ورودی، مقدار ارزش عامل ارتباط هر طبقه (crfValue) به صورت رابطه (۱۱) تعریف می‌شود:

$$crfValue(t_k.c_j) = \frac{PosotiveRF(t_k.c_j)}{NegativeRF(t_k.c_j)} \qquad (۱۱)$$

بیش‌تر روش‌های وزن‌دهی ویژگی مبتنی بر دسته به‌مانند روش IDF در ابتدا برای کاربردهای بازیابی اطلاعات و طبقه‌بندی مستندات به کار گرفته شده‌اند و در کاربرد طبقه‌بندی توالی پروتئین استفاده نشده‌اند. به همین دلیل برخی از مسائلی که در این حوزه به آن‌ها باید توجه شود نادیده گرفته شده‌است. روش‌های مبتنی بر IDF هر چه تعداد توالی که دارای یک ویژگی خاص هستند بیشتر باشد قدرت آن ویژگی در متمایزکردن توالی‌ها از یکدیگر پایین‌تر بوده و درنتیجه وزن کمتری به آن ویژگی اختصاص داده می‌شود. اگر چه این فرض در حوزه بازیابی صحیح است، اما در حوزه طبقه‌بندی نیازمند اعمال اصلاحاتی است. معیار crfValue معرفی‌شده در TFCRF راه حلی برای مسأله بالا ارائه می‌دهد؛ زیرا در آن وزن هر ویژگی در هر توالی رابطه مستقیم با تعداد توالی‌هایی دارد که از طبقه آن توالی بوده و رابطه معکوسی با تعداد توالی‌هایی دارد که از طبقه غیر از طبقه آن توالی هستند. مشاهده می‌شود که در این روش تأثیر طبقه‌ای که هر ویژگی در آنها ظاهر می‌شود نیز در نظر گرفته شده‌است. گفتنی است عامل crfValue مستقل از تعداد توالی‌های موجود در هر طبقه است، به همین دلیل کارایی طبقه‌بندی کننده توالی‌ها را تا حد قابل توجهی افزایش می‌دهد.

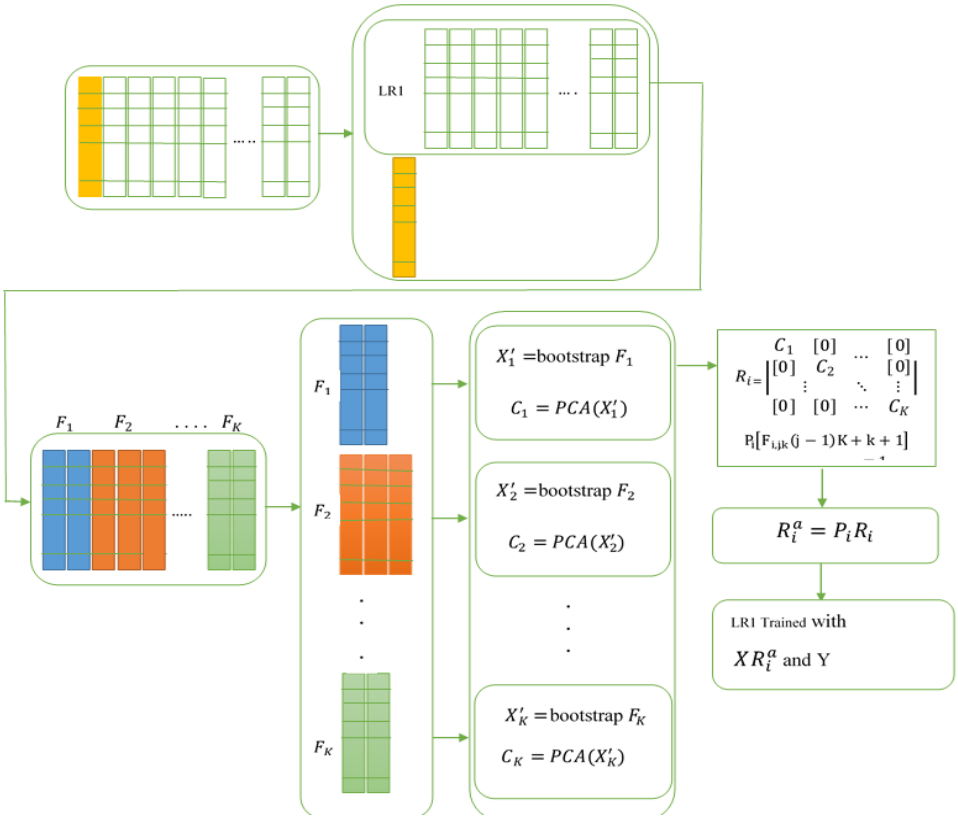
[1] Term Frequency and Category Relevancy Factor

[DOI: 10.61186/jsdp.21.1.15]



(شکل- ۱) مراحل روش استخراج ویژگی با استفاده از روش TFCRF

(Figure- 1) The steps of the feature extraction method using the TFCRF method



(شکل- ۲) مراحل الگوریتم جنگل چرخش [30]

(Figure - 2) Steps of rotation forest algorithm

[DOI: 10.61186/jsdp.21.1.15]

	NNRTI			NRTI					PI					
	DLV	EFV	NVP	3TC	ABC	AZT	D4T	DDI	APV	IDV	LPV	NFV	RTV	SQV
مثبت	455	447	415	195	179	322	336	306	424	384	223	303	349	457
منفی	263	274	318	429	440	299	285	317	278	374	278	472	379	304

(جدول- ۱) جزئیات مربوط به مجموعه‌داده استفاده شده

(Table-1) Details of the dataset used

در جنگل چرخش، روشی برای تولید مجموعه‌های طبقه‌بندی‌کننده بر اساس استخراج ویژگی پیشنهاد شده‌است. برای ایجاد داده‌های آموزشی برای یک طبقه‌بندی کننده‌پایه، مجموعه ویژگی به‌طور تصادفی به زیرمجموعه‌های Kتایی تقسیم می‌شود و تجزیه و تحلیل مؤلفه اصلی[1] برای هر زیر مجموعه اعمال می‌شود. تمام اجزای اصلی به‌منظور حفظ اطلاعات تغییرپذیری در داده‌ها، حفظ می‌شوند. بنابراین، چرخش‌های محور K برای تشکیل ویژگی‌های جدید برای یک طبقه‌بندی‌کننده پایه اتفاق می‌افتد. ایده رویکرد چرخشی تشویق هم‌زمان دقت و تنوع فردی در ترکیب را سبب می‌شود. تنوع از طریق استخراج ویژگی برای هر طبقه‌بندی‌کننده پایه ارتقا می‌یابد. درختان تصمیم در اینجا انتخاب شدندT زیرا به چرخش محورهای ویژگی حساس هستند. دقت با حفظ تمام مؤلفه‌های اصلی و همچنین استفاده از کل مجموعه‌داده‌ها برای آموزش هر طبقه‌بندی‌کننده‌پایه جستجو می‌شود.

جنگل چرخش یک روش ترکیبی است که دارای قابلیت تعمیم خوبی است. فرض شود، $x=[x_1, x_2, \ldots, x_n]$ مجموعه داده با n ویژگی باشد و X مجموعه آموزش با N نمونه و $Y=[y1, y2, \ldots, yn]$ بردار خروجی یا برچسب داده‌ها و F مجموعه ویژگی‌ها باشد. اندازه ترکیب L یک ابر پارامتر جنگل چرخش است. ابرپارامتر دیگر جنگل چرخش K است که تعداد زیرمجموعه ویژگی‌ها است [۳۰] (اندازه هر زیر مجموعه). هر طبقه‌بند D_i در ترکیب بر اساس مجموعه‌داده انتقال‌یافته ساخته می‌شود. برای هر مجموعه‌داده یک ماتریس چرخش R_i^a ایجاد می‌شود. ابتدا مجموعه ویژگی‌ها به k زیرمجموعه تقسیم می‌شود. برای هر زیرمجموعه ویژگی F_{ij}، داده‌های $X_{i,j}$ از مجموعه‌داده X را شکل می‌دهند که تنها ویژگی‌های F_{ij} را شامل می‌شوند. در جنگل چرخش برای دسته‌بندی، نمونه‌هایی از زیرمجموعه مناسب از دسته‌ها حذف، سپس یک خودراه‌انداز بر روی نمونه‌های $X_{i,j}$ اعمال می‌شود و مجموعه‌داده X'_{ij} به‌دست می‌آید. در شکل (۲) جنگل چرخش با جزئیات نشان داده شده‌است. آنالیز مؤلفه‌های اصلی بر روی X'_{ij} اعمال می‌شود تا ماتریس چرخش C_{ij} به‌دست آید. هر مولفه اصلی در ستون این ماتریس است. اندازه این ماتریس $M\times M_{i,j}$ است. M اندازه $F_{i,j}$ و $M_{i,j} \le M$ است که معمولا $M_{i,j}$ به‌اندازه M است، زیرا از M ویژگی می‌توان M مؤلفه (یا کمتر اگر بعضی از مؤلفه‌ها صفر باشد) به‌دست آورد. ماتریس C_{ij} به‌صورت قطری در ماتریس چرخش R_i قرار می‌گیرد و به‌صورت رابطه (۱۴) نشان داده می‌شود.

$$R_i = \begin{bmatrix} C_{i.1} & [0] & \cdots & [0] \\ [0] & C_{i.2} & & [0] \\ \vdots & & \ddots & \vdots \\ [0] & [0] & \cdots & C_{i.k} \end{bmatrix} \qquad (۱۴)$$

ماتریس R_i نمی‌تواند برای چرخش مستقیم مجموعه داده X به‌کار رود زیرا ویژگی‌ها در مرتبه مورد نظر نمی‌باشند. برای K گروه با M ویژگی این مرتبه به‌صورت رابطه (۱۵) است:

$$F_{i.1.1}.F_{i.1.2}.\cdots.F_{i.1.M}.F_{i.2.1}.F_{i.2.2}.\cdots.F_{i.2.M}.\cdots.F_{i.1.k}.\cdots.F_{i.k.M} \qquad (۱۵)$$

فرض شود ماتریس P_i یک ماتریس نگاشت n*n باشد. در هر سطر و ستون فقط یک عدد یک وجود دارد و مابقی مؤلفه‌ها صفر است. برای به‌دست‌آوردن چنین مرتبه‌ای برای هر $F_{i.1.k}(j = 1.\ \cdots.K.k = 1.\cdots.M.F_{i.1.k} \in \{1.\cdots.n\})$ رابطه (۱۶) وجود دارد:

$$P_i\left[F_{i.j.k}(j-1)K + k + 1\right] = 1 \qquad (۱۶)$$

سپس XP_i شامل مجموعه‌داده‌ای با ویژگی‌های دوباره مرتب شده‌است. اگر فرض شود $R_i^a = P_i R_i$ یک ماتریس چرخش دوباره مرتب شده باشد. طبقه‌بند D_i با مجموعه XR_i^a و Y آموزش داده شده‌است. در فاز پیش‌بینی، برای پیش‌بینی نمونه x، این نمونه با ماتریش R_i^a متناظر چرخش می‌یابد و برای هر طبقه‌بند D_i در ترکیب پیش‌بینی طبقه‌بند $D_i R_i^a$ خواهد بود. پیش‌بینی نهایی بر اساس میانگین‌گیری بین طبقه‌بندها در ترکیب خواهد بود.

[1] Principal Component Analysis (PCA)

[DOI: 10.61186/jsdp.21.1.15]

۳-ارزیابی نتایج

در این قسمت، جزئیات پیاده‌سازی روش پیشنهادی و تحلیل نتایج آزمایشات بیان می‌شود. برای این منظور، زبان‌برنامه نویسی پایتون و رایانه‌ای با مشخصات، کارت گرافیک GTX 1080، حافظه 8G، پردازنده Core i7 4790k-4GHz استفاده شده است.

۳-۱-مجموعه‌داده

ویروس HIV-1 به دلیل نرخ جهش زیاد آن، شناخته می‌شود که به ویروس فرصتی برای تکامل سریع در مقاومت دارویی می‌دهد. بنابراین پیش‌بینی مقاومت دارویی برای درمان شخص بیمار بسیار مهم و ضروری است.

بنابراین توالی‌های پروتئینی پروتئاز HIV-1 و رونوشت معکوس (RT) از سویه‌های زیرگروه B با داده‌های هفت گروه PI در نظر گرفته می شود.

(RTV: Ritonavir , IDV: Indinavir , SQV: Saquinavir , NFV: Nelfinavir, APV: Amprenavir, ATV: Atazanavir, LPV: Lopinavir)

و سه NNRTIs شامل:

(NVP: Nevirapine, EFV: Efavirenz, DLV: Delavirdine)

پنج NRTIs شامل:

(3TC: Lamivudine, ABC: Abacavir, AZT: Zidovudine, D4T: Stavudine, DDI: Didanosine)

با نسبت IC50 از پایگاه داده مقاومت دارویی HIV جمع‌آوری شده است. اطلاعات مربوط به مجموعه‌داده استفاده شده در جدول (۱) نشان داده شده است. لازم به ذکر است، توالی‌هایی را از مجموعه داده مانند ATV که هیچ اطلاعات مقاومتی برای آنها در دسترس نبود حذف شد زیرا تعداد زیادی از توالی‌ها فاقد اطلاعات IC50 بودند [۳۱].

۳-۲-معیارهای ارزیابی

برای ارزیابی روش پیشنهادی، از روش اعتبارسنجی متقابل با مقدار ۱۰[1] برای تعیین پارامترهای مدل براساس مجموعه داده‌های آموزشی استفاده شده است و از مجموعه داده مستقل برای ارزیابی عملکرد مدل استفاده می‌شود. جهت ارزیابی روش پیشنهادی از حساسیت (Sn)، ویژگی (Sp)، دقت (Pre)، صحت (ACC) و ضریب همبستگی Matthew's (MCC) استفاده شده است [٤]. این شاخص‌ها به صورت روابط (۱۷-۱۹) تعریف می‌شوند:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (۱۷)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (۱۸)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱۹)$$

در این روابط TP درستی مثبت می‌باشد، یعنی تعداد برچسب‌های مثبتی که به درستی توسط طبقه‌بندی کننده مثبت پیش‌بینی شده است، TN درستی منفی می‌باشد، یعنی تعداد برچسب‌های منفی که توسط طبقه‌بندی کننده به درستی منفی پیش‌بینی شده است، FP نادرست مثبت، تعداد برچسب‌های منفی که توسط طبقه‌بندی کننده به اشتباه مثبت پیش‌بینی شده است و FN نشان دهنده نادرست منفی می‌باشد، یعنی تعداد برچسب‌های مثبت که توسط طبقه‌بندی‌کننده به اشتباه منفی پیش‌بینی شده است اشاره می‌کند.

۳-۳-نتایج حاصل از انواع ویژگی‌ها

همانطور که بیان شده به منظور پیش‌بینی توالی پروتئین‌ها، از این توالی‌ها ویژگی‌های مختلف استخراج می‌شود. با توجه به اینکه در این مقاله هدف استخراج ویژگی‌های موثر از توالی پروتئین‌ها می‌باشد. در این بخش ویژگی‌های استخراجی مورد تجزیه و تحلیل قرار گرفته شده‌اند. در این مقاله پنج روش استخراج ویژگی بر روی توالی‌های پروتئین‌ها اعمال شد. به منظور ارزیابی ویژگی‌های استخراجی توسط هر روش، مدل جنگل چرخش با استفاده از هر یک از ویژگی‌های PSSM، PsePSSM، K-gram، AAC و TFCRF بر اساس اعتبارسنجی متقابل با مقدار k=10 آموزش داده می‌شود. تشخیص مثبت یا منفی‌بودن تأثیر در جدول (۲) نشان داده شده‌است.

(جدول- ۲) نتایج روش RF با ویژگی‌های مختلف
(Table -2) The results of the RF method with different features

	AAC	PsePSSM	PSSM	K-gram	TFCRF
Accuracy	77.73	93.86	92.9	82.61	**97.23**
Sensitivity	77.27	92.11	91.67	81.49	**96.17**
Specificity	80.04	96.32	94.49	84.12	**98.92**
AUC	0.912	0.9645	0.9621	0.9023	**0.9911**

همان‌طورکه از نتایج حاصل از این آزمایش مشخص است، ویژگی‌های استخراج‌شده توسط TFCRF دارای قدرت متمایز کنندگی بیشتری هستند و در تمامی مجموعه‌داده دارای نرخ تشخیص بهتری و علاوه‌براین روش‌های PSSM

[1] 10-fold cross-validation

[DOI: 10.61186/jsdp.21.1.15]

و PsePSSM نیز دارای عملکرد قابل قبولی هستند. هر یک از این ویژگی‌ها، الگویی از داده‌ها را نشان می‌دهد که باعث می‌شود مدل طبقه‌بندی به‌خوبی تأثیر مثبت و منفی دارو بر روی پروتئین را شناسایی کند.

جهت مقایسه بین ویژگی‌های استخراج‌شده با استفاده از روش‌های مختلف نمودار ROC در شکل (۳) برای پنج نوع ویژگی رسم شده‌است. همان‌طورکه مشاهده می‌شود ویژگی TFCRF از سایر ویژگی‌ها بهتر عمل کرده است و مساحت زیر نمودار بالاتری دارد. روش K-gram نسبت به سایر روش‌ها دارای عملکرد پایین‌تری بوده است. بر اساس نتایج شکل (۳) و جدول (۲) می‌توان نتیجه گرفت که ترکیب ویژگی‌های مختلف باعث بهبود عملکرد مدل طبقه‌بندی جهت شناسایی تعاملات بین داروها و پروتئین‌ها می‌شود.

به‌منظور ارزیابی بهتر روش پیشنهادی، معیار AUC دسته‌بندهای مختلف با استفاده از ویژگی‌های TFCRF برای توالی‌های مختلف در جدول (۳) نشان داده شده‌است. نتایج حاصل از این جدول با جدول شماره (۲) این تفاوت را دارد که در جدول شماره (۲) نتایج دسته‌بندی مثبت و منفی در همه توالی‌ها بررسی شده‌است. در این حالت تعداد داده‌ها در دو دسته زیاد است. ولی در جدول (۳) برای هر گروه دسته‌بندی از نظر مثبت و منفی انجام شده‌است. در این حالت تعداد داده‌ها نسبت به حالت قبل خیلی کاهش می‌یابد؛ بنابراین در این حالت نمی‌توان به‌خوبی مثل حالت قبل (جدول ۲) الگوهای بین داده‌های دو دسته را شناسایی کرد.

(جدول- ۳) مقایسه AUC طبقه‌بندهای مختلف برای ویژگی TFCRF

(Table -3) AUC comparison of different classifications for TFCRF feature

	NNRTI			NRTI					PI					
	DLV	EFV	NVP	3TC	ABC	AZT	D4T	DDI	APV	IDV	LPV	NFV	RTV	SQV
SVM	0.97	0.95	0.94	0.97	0.95	0.94	0.93	0.88	0.96	0.97	0.96	0.96	0.97	0.95
MLP	0.91	0.95	0.95	0.96	0.92	0.91	0.88	0.83	0.89	0.91	0.92	0.91	0.94	0.92
DT	0.90	0.96	0.96	0.96	0.93	0.93	0.89	0.85	0.91	0.95	0.95	0.95	0.96	0.93
RF	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**	**0.92**	**0.98**	**0.99**	**0.99**	**0.99**	**1**	**0.99**

(جدول- ۴) نتایج روش پیشنهادی بر روی داده آزمون و داده‌های مستقل

(Table -4) The results of the proposed method on test data and independent data

	NNRTI			NRTI					PI					
	DLV	EFV	NVP	3TC	ABC	AZT	D4T	DDI	APV	IDV	LPV	NFV	RTV	SQV
Test data	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**	**0.92**	**0.98**	**0.99**	**0.99**	**0.99**	**1**	**0.99**
Independent data	0.96	0.97	0.96	0.98	0.96	096	0.96	0.91	0.96	0.97	0.97	0.98	0.98	0.98

(جدول- ۵) مقایسه معیارهای کارایی روش پیشنهادی و نتایج گزارش‌شده در مقالات معتبر

(Table -5) Comparison of the performance criteria of the proposed method and the results reported in the authoritative papers

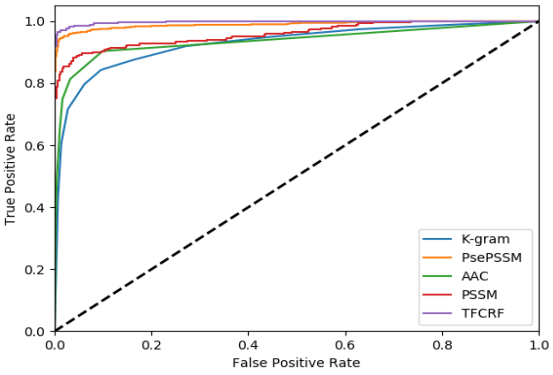
	NNRTI			NRTI					PI					
	DLV	EFV	NVP	3TC	ABC	AZT	D4T	DDI	APV	IDV	LPV	NFV	RTV	SQV
Rhee et al. [31]	84	87	91	90	84	84	78	75	84	79	81	82	89	84
Heider et al.[32]	87	88	87	87	88	87	84	79	88	93	92	91	95	89
Hou et al. [33]	-	-	-	-	-	-	-	-	89	86	91	87	93	89
Löchel et al. [34]	92	94	**96**	97	95	94	87	85	91	97	98	**96**	97	90
Proposed Method	**93**	**96**	**96**	**98**	**97**	**96**	**90**	**89**	**94**	**99**	**99**	**96**	**98**	**92**

همان‌طورکه از نتایج مشخص است دسته‌بند جنگل چرخش نسبت به سایر دسته‌بندها دارای عملکرد بهتری است؛ علاوه‌براین با توجه به قدرت متمایزکنندگی ویژگی‌های TFCRF عملکرد سایر دسته‌بندها نیز دارای عملکرد قابل قبولی است. با توجه به اینکه دسته‌بند MLP تعداد پارامترهای بیشتری نسبت به سایر روش‌ها دارد و با توجه به تعداد کم نمونه‌ها در دو دسته، یادگیری این پارامترها به درستی انجام نمی‌شود؛ بنابراین نرخ تشخیص کمتری نسبت به سایر روش‌ها دارد.

به‌منظور ارزیابی بهتر و بررسی واریانس خطای روش پیشنهادی، نمودار خطا برای دسته‌بندهای مختلف بر اساس ویژگی‌های مختلف در شکل (۴) نشان داده

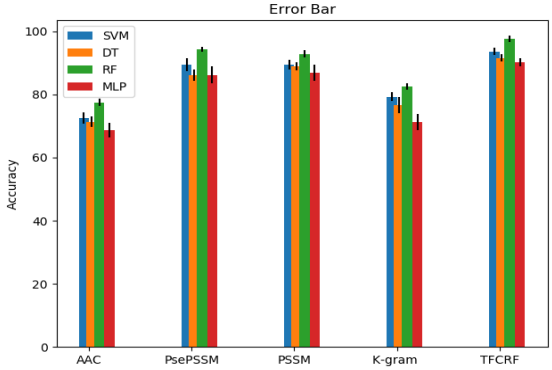
[DOI: 10.61186/jsdp.21.1.15]

شده‌است. همان‌طورکه از نتایج مشخص است بر اساس ویژگی‌های استخراج‌شده، دسته‌بند جنگل‌چرخش دارای عملکرد بهتری نسبت به سایر روش‌ها است و نرخ خطای کمتری نسبت به سایر دسته‌بندها دارد. این امر به این دلیل است که ویژگی‌های استخراجی دارای خاصیت متمایزکننده خوبی هستند.



(شکل- ۳) نمودار ROC برای مقایسه بین ویژگی‌های مختلف استخراج شده از توالی

(Figure- 3) ROC diagram for the comparison between different features extracted from the sequence



(شکل- ۴) نمودار خطا در طبقه بندهای مختلف بر روی ویژگی‌های متفاوت استخراج شده

(Figure- 4) Error diagram in different categories on different extracted features

علاوه‌براین طبقه‌بند جنگل چرخش با توجه به این‌که ویژگی‌ها مناسب‌تر را برای ساخت درخت‌ها انتخاب می‌کند باعث می‌شود که قابلیت تعمیم خوبی داشته باشد. علاوه‌براین، عملکرد سایر دسته‌بندها با استفاده از ویژگی‌های استخراجی TFCRF نسبت به سایر ویژگی‌ها بهتر است و همچنین نرخ خطا نسبت به سایر ویژگی‌ها کمتر است؛ علاوه‌بر این به بررسی صحت روش پیشنهادی، نتایج حاصل از ارزیابی روش پیشنهادی، بر روی داده‌های آزمون و آزمون مستقل ارزیابی شده‌است. همان‌طورکه از نتایج جدول (۴) مشخص است نتایج بر روی داده‌های آزمون و آزمون مستقل خیلی با یکدیگر تفاوت ندارند. این نشان می‌دهد که مدل بر روی داده‌های مشاهده‌نشده دارای کارایی مناسبی می باشد.

۴-۳-مقایسه با سایر روش‌ها

به‌منظور ارزیابی بهتر، روش پیشنهادی با سایر روش‌های موجود که در سال‌های اخیر در این کاربرد استفاده شده‌اند، مورد مقایسه قرار گرفته شده‌است. گفتنی است نتایج بر روی مجموعه‌داده یکسان مورد ارزیابی قرار گرفته است. نتایج حاصل از این آزمایش در جدول (۵) نشان داده شده‌اند. روش‌های مورد مقایسه ویژگی‌های مختلفی از توالی پرتئین استخراج و از دسته‌بندهای مختلف استفاده کرده‌اند. همان‌طورکه مشخص است، مقادیر نرخ صحت روش پیشنهادی بهتر از سایر روش‌ها است. در بیشتر موارد روش پیشنهادی از سایر روش‌ها عملکرد بهتری دارد. روش پیشنهادی با توجه به این‌که ویژگی‌های متمایز و مؤثر در هر دسته را شناسایی و استخراج می‌کند و روش جنگل چرخش برای ساخت درخت‌ها، ویژگی‌ها مناسب‌تر را انتخاب می‌کند سبب بهبود عملکرد روش پیشنهادی نسبت با سایر روش‌ها می‌شود.

۴-نتیجه‌گیری

در این مقاله به‌منظور پیش‌بینی توالی پروتئین روشی مبتنی بر یادگیری ماشین ارائه شده‌است. در این روش ابتدا از این توالی، ویژگی‌های مختلف استخراج، سپس با استفاده از روش‌های دسته‌بندی عمل پیش‌بینی انجام می‌شود. با توجه به این‌که هر چقدر ویژگی‌های استخراجی در مرحله نخست مناسب‌تر باشد، عملکرد روش‌های دسته‌بندی بهبود می‌یابد. در این مقاله روشی مبتنی بر TFCRF ارائه شده‌است تا بتوان ویژگی‌هایی که دارای قدرت متمایزکنندگی خوبی هستند از توالی پروتئین استخراج شود. در مرحله دسته‌بندی نیز جنگل چرخش ارائه شده‌است و عمل دسته‌بندی بر اساس آن انجام می‌شود. نتایج حاصل از پیاده‌سازی نشان می‌دهد روش پیشنهادی دارای نرخ تشخیص قابل قبولی است.

۵- مراجع 5-Refrence

1. Heider, D., et al., A computational approach for the identification of small GTPases based on preprocessed amino acid sequences. Technology in cancer research & treatment, 2009. 8(5): p. 333-341.
2. Löchel, H.F., et al., SCOTCH: subtype A coreceptor tropism classification in HIV-1. Bioinformatics, 2018. 34(15): p. 2575-2580.

[DOI: 10.61186/jsdp.21.1.15]

3. Heider, D. and D. Hoffmann, Interpol: An R package for preprocessing of protein sequences. BioData mining, 2011. 4(1): p. 1-6.
4. Armano, G. and A. Giuliani, A two-tiered 2d visual tool for assessing classifier performance. Information Sciences, 2018. 463: p. 323-343.
5. Yu, X., I. Weber, and R. Harrison. Sparse representation for HIV-1 protease drug resistance prediction. in Proceedings of the 2013 SIAM international conference on data mining. 2013. SIAM.
6. Spänig, S. and D. Heider, Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. BioData Mining, 2019. 12(1): p. 1-29.
7. Zhang, J., et al., Variable selection from a feature representing protein sequences: a case of classification on bacterial type IV secreted effectors. BMC bioinformatics, 2020. 21: p. 1-15.
8. De Santis, E., et al. Dissimilarity space representations and automatic feature selection for protein function prediction. in 2018 International joint conference on neural networks (IJCNN). 2018. IEEE.
9. Bonetta, R. and G. Valentino, Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics, 2020. 88(3): p. 397-413.
10. Rizzo, R., et al. Classification experiments of DNA sequences by using a deep neural network and chaos game representation. in Proceedings of the 17th International Conference on Computer Systems and Technologies 2016. 2016.
11. Qian, W., et al., Feature selection for label distribution learning via feature similarity and label correlation. Information Sciences, 2022. 582: p. 38-59.
12. Abu Khurma, R., et al., A review of the modification strategies of the nature inspired algorithms for feature selection problem. Mathematics, 2022. 10(3): p. 464.
13. Törönen, P. and L. Holm, PANNZER—a practical tool for protein function prediction. Protein Science, 2022. 31(1): p. 118-128.
14. Lv, Z., C. Ao, and Q. Zou, Protein function prediction: from traditional classifier to deep learning. Proteomics, 2019. 19(14): p. 1900119.
15. Mahood, E.H., L.H. Kruse, and G.D. Moghe, Machine learning: a powerful tool for gene function prediction in plants. Applications in Plant Sciences, 2020. 8(7): p. e11376.
16. Martino, A., A. Rizzi, and F.M.F. Mascioli. Supervised approaches for protein function prediction by topological data analysis. in 2018 International joint conference on neural networks (IJCNN). 2018. IEEE.
17. You, R., et al., NetGO: improving large-scale protein function prediction with massive network information. Nucleic acids research, 2019. 47(W1): p. W379-W387.
18. Lai, B. and J. Xu, Accurate protein function prediction via graph attention networks with predicted structure information. Briefings in Bioinformatics, 2022. 23(1): p. bbab502.
19. Liu, X., Deep recurrent neural network for protein function prediction from sequence. arXiv preprint arXiv:1701.08318, 2017.
20. Kulmanov, M. and R. Hoehndorf, DeepGOPlus: improved protein function prediction from sequence. Bioinformatics, 2020. 36(2): p. 422-429.
21. Sureyya Rifaioglu, A., et al., DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Scientific reports, 2019. 9(1): p. 7344.
22. Li, M., et al., A deep learning framework for predicting protein functions with co-occurrence of GO terms. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022. 20(2): p. 833-842.
23. Shen, H.-B. and K.-C. Chou, Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering, Design & Selection, 2007. 20(11): p. 561-567.
24. Akbar, S., et al., iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach. Chemometrics and Intelligent Laboratory Systems, 2020. 204: p. 104103.
25. Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 1999. 292(2): p. 195-202.
26. Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 1997. 25(17): p. 3389-3402.
27. Chou, K.C., Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics, 2001. 43(3): p. 246-255.
28. Bhasin, M. and G.P. Raghava, Classification of nuclear receptors based on amino acid composition and dipeptide composition. Journal of Biological Chemistry, 2004. 279(22): p. 23262-23266.
29. Sorkhi, A.G., J. Pirgazi, and V. Ghasemi, A hybrid feature extraction scheme for efficient malonylation site prediction. Scientific Reports, 2022. 12(1): p. 5756.
30. Pirgazi, J., A.R. Khanteymoori, and M. Jalilkhani, GENIRF: An algorithm for gene regulatory network inference using rotation forest. Current Bioinformatics, 2018. 13(4): p. 407-419.
31. Rhee, S.-Y., et al., Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proceedings of the National Academy of Sciences, 2006. 103(46): p. 17355-17360.
32. Heider D, Verheyen J, Hoffmann D. Machine learning on normalized protein sequences. BMC research notes. 2011 Dec;4:1-0.
33. Hou, T. Zhang, W. Wang, J. Wang, W., Predicting drug resistance of the HIV-1 protease using molecular interaction energy components. Proteins: Structure, Function, and Bioinformatics. 2009 Mar; 74(4): p. 837-46.
34. Löchel HF, Eger D, Sperlea T, Heider D. Deep learning on chaos game representation for proteins. Bioinformatics. 2020 Jan 1;36(1): p. 272-9.

[DOI: 10.61186/jsdp.21.1.15]

جمشید پیرگزی، دارای مدرک دکترای مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه زنجان است. وی در حال حاضر استادیار دانشگاه علم و فناوری مازندران در گروه مهندسی کامپیوتر و سرپرست آزمایشگاه تحقیقاتی یادگیری و بینایی ماشین است. علایق پژوهشی ایشان مدل های زبانی بزرگ، شبکه‌های عصبی عمیق، یادگیری ماشین، پردازش داده‌ای زیستی و بیوانفورماتیک است.

نشانی رایانامه ایشان عبارت است از:

j.pirgazi@mazust.ac.ir

علی قنبری سرخی، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۸۹ از دانشگاه علم و صنعت ایران و مدرک کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی به‌ترتیب در سال‌های ۱۳۹۱ و ۱۳۹۷ از دانشگاه صنعتی شاهرود دریافت کرده است. وی در حال حاضر استادیار دانشگاه علم و فناوری مازندران در گروه مهندسی کامپیوتر است. زمینه پژوهش ایشان عبارتند از: هوش مصنوعی، پردازش تصویر، شبکه های عصبی عمیق، بیوانفورماتیک.

نشانی رایانامه ایشان عبارت است از:

ali.ghanbari@mazust.ac.ir

مجید ایرانپور مبارکه، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار در سال ۱۳۸۳ از دانشگاه آزاد اسلامی واحد مبارکه اصفهان و مدرک کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی به‌ترتیب در سال‌های ۱۳۸۷ و ۱۳۹۵ از دانشگاه علم و صنعت ایران و دانشگاه صنعتی شاهرود دریافت کرده است. وی در حال حاضر استادیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه پیام نور است. زمینه پژوهش ایشان عبارتند از: پردازش تصویر و بینایی ماشین، پردازش اسناد تصویری، پردازش متن، شبکه‌های عصبی عمیق، بیوانفورماتیک.

نشانی رایانامه ایشان عبارت است از:

Iranpour@pnu.ac.ir

[DOI: 10.61186/jsdp.21.1.15]