مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هشتم شماره سوم. ص ۲۷۰-۲۸۱، ۱۴۰۰

استخراج ویژگی چندمتغیره برای پیش‌بینی ماتریس بیان ژنی آینده

پریناز اسکندریان[۱]، جمشید باقرزاده محاسفی[۲]*، حبیب‌اله پیرنژاد[۳]، زهرا نیازخانی[۴]

• دریافت مقاله: ۱۴۰۰/٤/۱۵ • پذیرش مقاله: ۱٤۰۰/۹/۲٤

مقدمه: ویژگی‌های یک سلول را می‌توان با بررسی ماتریس بیان ژنی مربوط به آن سلول تعیین کرد. اگر بتوان ماتریس‌های بیان ژنی مربوط به سلول‌های فرزند آینده را پیش‌بینی کرد، در حقیقت ویژگی‌های سلول‌های آینده پیش‌بینی شده‌اند. هدف مطالعه حاضر، طراحی یک شبکه عصبی مصنوعی برای پیش‌بینی ماتریس‌های بیان ژنی برای سلول‌های فرزندی است که از تقسیم/تمایز سلول‌های بنیادی هماتوپویتیک در آینده به دست خواهند آمد.

روش: شبکه عصبی طراحی‌شده ماتریس بیان ژنی یک سلول بنیادی هماتوپویتیک والد را به عنوان ورودی می‌گیرد و ماتریس‌های بیان ژنی مربوط به سلول‌های فرزند آینده آن را تولید می‌کند. یک کدگذار زمانی برای کدگذاری سری زمانی اصلی و یک کدگذار مکانی برای کدگذاری سری‌های زمانی ثانویه پیشنهاد می‌شود.

نتایج: برای آن که پیش‌بینی قابل پذیرشی انجام شود، باید ماتریس‌های بیان ژنی مربوط به دست‌کم چهار مرحله اولیه از تقسیم/تمایز مشخص باشند. شبکه عصبی طراحی‌شده از نظر خطای پیش‌بینی و تعداد مراحل تقسیم/تمایز که به درستی پیش‌بینی شده باشند، نسبت به شبکه‌های عصبی موجود بهتر عمل می‌کند. طرح پیشنهادی این مطالعه می‌تواند پیش‌بینی را برای صدها مرحله از تقسیم/تمایز سلولی انجام دهد. خطای طرح پیشنهادی برای پیش‌بینی ۱، ۴، ۱۶، ۶۴ و ۱۲۸ مرحله از تقسیم/تمایز به ترتیب برابر با ۳/۰۴، ۳/۷۶، ۵/۵، ۷/۸۳، و ۱۱/۰۶ درصد بوده است.

نتیجه‌گیری: با داشتن ماتریس بیان ژنی مربوط به یک سلول هماتوپویتیک والد می‌توان ماتریس‌های بیان ژنی مربوط به فرزندان آن را تا صدها مرحله از تقسیم/تمایز پیش‌بینی کرده و در صورت لزوم، به موقع چاره‌ای برای روبه‌رو شدن با مشکلات ژنتیکی آینده اندیشید.

کلیدواژه‌ها: سلول بنیادی هماتوپویتیک، شبکه عصبی، سری زمانی چندمتغیره، ماتریس بیان ژنی، پیش‌بینی

•**ارجاع:** اسکندریان پریناز، باقرزاده محاسفی جمشید، پیرنژاد حبیب اله، نیازخانی زهرا. استخراج ویژگی چندمتغیره برای پیش‌بینی ماتریس بیان ژنی آینده. مجله انفورماتیک سلامت و زیست پزشکی ۱۴۰۰؛ ۸(۳): ۸۱-۲۷۰.

۱. دانشجوی دکتری مهندسی کامپیوتر، گروه کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

۲. دکتری مهندسی نرم‌افزار، دانشیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه ارومیه، ارومیه، ایران

۳. دکتری تخصصی انفورماتیک پزشکی، دانشیار، مرکز تحقیقات ایمنی بیمار، پژوهشکده تحقیقات بالینی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران

۴. دکتری تخصصی انفورماتیک پزشکی، دانشیار، مرکز تحقیقات نفرولوژی و پیوند کلیه، پژوهشکده تحقیقات بالینی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران

*** نویسنده مسئول:** جمشید باقرزاده محاسفی

آدرس: ارومیه، کیلومتر ۳ جاده سلماس، دانشگاه آزاد اسلامی واحد ارومیه

• **شماره تماس:** ۰۴۴۱-۳۱۸۰۳۰۰۰ • **Email:** j.bagherzadeh@urmia.ac.ir

مقدمه

سلول‌های بنیادی هماتوپویتیک (Stem Hematopoietic Cells) HSC به عنوان سلول‌های چندتوان (Multipotent Cells) در نظر گرفته می‌شوند که امکان تکثیر به نوع خود و یا تمایز به چند سلول هماتوپویتیک والد (Hematopoietic Progenitor Cells) HPC را دارند. تولید سلول‌های خون از روی سلول‌های بنیادی هماتوپویتیک شامل چند گام تقسیم /تمایز سلولی است که از طریق این گام‌ها چند سلول بنیادی هماتوپویتیک چندتوان به صورت اجداد تک‌فرزند یا دوفرزند و در پایان به صورت سلول‌های خون متمایز می‌شوند [۱]. مسیر تمایزی که سلول‌های بنیادی هماتوپویتیک برمی‌گزینند هنوز به طور کامل شناخته نشده است [۱]؛ اما نوع سلول فرزندی که توسط یک سلول والد تولید می‌شود، از بیان ژنی سلول تأثیر می‌گیرد که می‌تواند تحت تأثیر سیگنال‌های برون‌سلولی قرار گیرد [۲].

وضعیت ژن‌ها در یک سلول را می‌توان با اندازه‌گیری ماتریس بیان ژنی (Gene Expression Profile Matrix) مربوط به آن سلول بررسی کرد. ماتریس بیان ژنی برای یک سلول مشخص توسط یک ماتریس P قابل نمایش است (شکل ۱). هر سطر از این ماتریس متناظر با یک ژن است. فرض می‌کنیم به تعداد M ژن در هر سلول وجود دارد. مقدار p_g (واقع در سطر g از ماتریس) نشان دهنده سطح بیان ژن شماره g است ($g \in \{1,2,\dots,M\}$).

$$P=\begin{pmatrix} p_1 \\ p_2 \\ \dots \\ p_M \end{pmatrix}\begin{matrix} Gene1 \\ Gene2 \\ \dots \\ GeneM \end{matrix}$$

شکل ۱: یک ماتریس بیان ژنی با ابعاد Mx1

در این پژوهش، این واقعیت در نظر گرفته شد که نوع، رفتار و ویژگی‌های یک سلول را می‌توان با استفاده از ماتریس بیان ژنی مربوط به آن سلول تعیین کرد. پس می‌توان یک ماشین یادگیرنده طراحی نمود که ماتریس بیان ژنی متعلق به یک سری سلول بنیادی هماتوپویتیک والد و سلول‌های فرزند آن‌ها را بخواند. ماشین با این کار یاد می‌گیرد که ماتریس بیان ژنی چگونه در طول هر تقسیم/تمایز سلولی تغییر می‌کند؛ بنابراین پس از یک تعداد کافی از گام‌های آموزش، آن ماشین می‌تواند ماتریس بیان ژنی مربوط به یک سلول بنیادی هماتوپویتیک والد را بگیرد و ماتریس‌های بیان ژنی مربوط به سلول‌های فرزند آینده آن را پیش‌بینی کند. به کمک چنین ماشین یادگیرنده می‌توان وضعیت سلول‌های آینده را پیش‌بینی کرد و چاره‌ای برای روبه‌رو شدن با مشکل‌های ژنتیکی آن‌ها برداشت. برای نمونه، ممکن است این ماشین پیش‌بینی کند که این سلول‌های بنیادی با این شرایط ژنتیکی که اکنون دارند، در آینده سبب یک بیماری ژنتیکی در فرد بیمار خواهند شد.

فرض می‌کنیم سلول‌های بنیادی هماتوپویتیک در محیط یک ظرف کشت داده می‌شوند. آن‌ها در حالی که تعدادی سلول فرزند تولید می‌کنند، تقسیم یا متمایز می‌شوند. این سلول‌های فرزند ممکن است بیشتر تقسیم/متمایز شوند و نسل‌های آینده سلول‌ها را با ژنوتایپ‌ها و ویژگی‌های فنوتایپ خاص‌تر تولید کنند. یک زیست‌شناس ماتریس بیان ژنی سلول‌های موجود در ظرف را در نقاط زمانی مختلف اندازه می‌گیرد. فرض می‌شود زمان سپری‌شده بین دو نقطه زمانی پشت‌سرهم ثابت است که آن را یک برش زمانی (Timeslot) می‌نامیم. پس برش زمانی t را می‌توان برای $t = 1,2,\dots,N$ در نظر گرفت، جایی که N تعداد برش‌های زمانی است.

زیست‌شناس ماتریس بیان ژنی مربوط به نقاط زمانی پشت‌سرهم را گردآوری کرده و آن‌ها را داخل یک ماتریس M در N قرار می‌دهد (شکل ۲). که آن ماتریس (Gene Expression Sequence) GES نامیده می‌شود. هر ستون از ماتریس GES یک ماتریس بیان ژنی است که در یک نقطه زمانی از ظرف اندازه‌گیری می‌شود. ستون اول متعلق به سلول‌های بنیادی اولیه است. در ماتریس GES سطر g نشان دهندهٔ سطوح بیان ژن g است، که $g \in \{1,2,\dots,M\}$. سطر g را می‌توان به عنوان یک سری زمانی در نظر گرفت که نشان می‌دهد سطح بیان مربوط به ژن شماره g چگونه در طول تقسیم/ تمایز سلول تغییر می‌کند.

$$Y=\begin{pmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,N} \\ y_{2,1} & y_{2,2} & \dots & y_{2,N} \\ \dots & \dots & \dots & \dots \\ y_{M,1} & y_{M,2} & \dots & y_{M,N} \end{pmatrix}\begin{matrix} Gene1 \\ Gene2 \\ \dots \\ GeneM \end{matrix}$$
$$\begin{matrix} timeslot1 & timeslot2 & \dots & timeslotN \end{matrix}$$

شکل ۲: یک ماتریس GES با اندازه M در N

مسئله Hematopoietic Gene Expression Prediction) HGEP [۳] تعدادی مجموعه داده آموزشی به ما داده می‌شود که هر کدام دارای یک ماتریس GES کامل مانند شکل ۲ هستند. همه ماتریس‌های GES در مجموعه‌های داده دارای M سطر هستند که تخصیص ژن یکسانی دارند؛ اما تعداد ستون‌های آن‌ها متفاوت است. همه مقادیر در این ماتریس‌های GES آموزشی شناخته شده هستند. اکنون یک ماتریس GES ناقص با ابعاد M در N به نام Y در نظر گرفته می‌شود که در آن تنها مقادیر k ستون اول شناخته شده‌اند ($k < N$). بر پایه ماتریس‌های GES آموزشی داده شده، مقدارهای موجود در ستون ($k + 1$) از ماتریس Y را پیش‌بینی کنید.

یک رویکرد استدلال مبتنی بر مثال (Case-based reasoning) برای پیش‌بینی نتیجه تمایز سلول بنیادی این است که ویژگی‌های سلول بنیادی که باید پیش‌بینی شود را با ویژگی‌های یک پایگاه داده از سلول‌های بنیادی که در گذشته متمایز شده‌اند، مقایسه شود. اگر یک سلول بنیادی در پایگاه داده خود یافت شود که از نظر ژنتیکی (در شرایط آزمایشگاهی یکسان) مانند سلول‌های بنیادی موردنظر جهت پیش‌بینی باشد، آن گاه به این نتیجه می‌رسیم که سلول بنیادی موردنظر به طور احتمالی، نتایج تمایز سلول بنیادی یافت‌شده را خواهند داشت. این رویکرد می‌تواند یک دید بلندمدت از آینده سلول بنیادی که باید پیش‌بینی شوند را به ما بدهد.

برخی پژوهشگران با استفاده از رویکرد استدلال مبتنی بر مثال، تعدادی سرنخ و ویژگی را کشف نموده‌اند که اگر آن‌ها در یک سلول بنیادی مشاهده شوند، آن گاه آن سلول با یک احتمال بالا به یک نوع مشخص از سلول‌ها متمایز می‌شود. برای نمونه، پتانسیل تمایز سلول بنیادی جنین انسان به سمت سلول‌های غدد جنسی نر را می‌توان با آنالیز کردن شرایط رشد و آنالیز بیان ژنی پیش‌بینی کرد [۴]. Yanagihara و همکاران [۵] یک روش برای پیش‌بینی گرایش سلول‌های بنیادی پرتوان در تمایز به سمت سلول‌های کبدی ارائه کرده‌اند. هر کدام از این مطالعات بر یک حالت ویژه از تمایز سلولی تمرکز کرده‌اند و یک پتانسیل تمایز بسیار ویژه را یافته‌اند؛ بنابراین این روش‌ها نمی‌توانند برای آن گونه از سلول‌های بنیادی پیش‌بینی انجام دهند که در پایگاه داده محدود آن‌ها وجود ندارند.

Bhattacharjee و Vishwakarma بتازگی یک روش [۶] ارائه داده‌اند که داده بیان ژنی آینده را با استفاده از داده بیان ژنی مشاهده‌شده پیش‌بینی می‌کند. برای این کار، از محاسبات آماری استفاده کرده و میانگین و واریانس را برای داده بیان ژنی آینده حدس می‌زنند. آن روش تنها متکی بر نقاط زمانی قبلی است و مشخصات سلول را در نظر نمی‌گیرد. در این مطالعه طرح Bhattacharjee2019 نامیده شد و مورد ارزیابی قرار گرفت.

مسئله HGEP به شکل یک مسئله پیش‌بینی سری زمانی چندمتغیره مدل‌سازی شده است. روش‌های مختلفی برای حل این مسئله وجود دارد. براساس روش‌های موجود پیش‌بینی سری‌های زمانی چندمتغیره بررسی شد و نتایج نشان داد این این روش‌ها که بر پایه شبکه عصبی مصنوعی طراحی شده‌اند، نمی‌توانند ویژگی‌های عصبی را به درستی از سری‌های زمانی ورودی استخراج کنند. در نتیجه، نمی‌توانند به دقت پیش‌بینی بالایی دست یابند.

تعداد اندکی مطالعه در زمینه پیش‌بینی سری زمانی چندمتغیره غیرخطی وجود دارد که از ماشین‌های یادگیرنده استفاده می‌کنند. برای پیش‌بینی عنصر بعدی سری زمانی اصلی با استفاده از یک شبکه عصبی، پیشنهاد شده است که نه تنها از سری زمانی اصلی بلکه از سری‌های زمانی ثانویه هم استفاده شود [۷]. پرسپترون چندلایه (Multi Layer Perceptron) MLP برای پیش‌بینی روزانه تشعشعات خورشیدی به عنوان سری زمانی اصلی استفاده شده است [۸]. Rao و Srivastava از حالات توییتر به عنوان سری زمانی ثانویه جهت پیش‌بینی قیمت بازار به عنوان سری زمانی اصلی، استفاده کرده‌اند [۱۰]. از شبکه عصبی فازی خودسازماندهی برای پیش‌بینی سری زمانی چندمتغیره استفاده شده است که با شامل نمودن سری‌های زمانی مربوط به حالت‌های توییتر، دقت پیش‌بینی افزایش خواهد یافت [۹]. طرح‌های DA-RNN [۱۱] و EA-LSTM [۱۲] برای پیش‌بینی سری زمانی چندمتغیره ارائه شدند. آن‌ها از شبکه‌های توجه (Attention Networks) برای کشف وابستگی‌ها بین سری‌های زمانی استفاده می‌کنند. Bu و Cho یک مکانیزم توجه چندمسیره ارائه داده‌اند که آن مکانیزم همراه با لایه‌های کانولوشنی (Convolution) و لایه‌های LSTM (Long Short Term Memory) برای پیش‌بینی مصرف انرژی به کار رفته است [۱۳]. Yuan و همکاران یک شبکه توجه چندمسیره که کیفیت اطلاعات ارائه شده توسط هر مسیر را در نقطه‌های زمانی مختلف بررسی می‌کند را ارائه داده‌اند [۱۴]. Hu و Zheng یک

پارامتر برای محاسبه تأثیر مختلف برای هر سری زمانی ثانویه پیشنهاد داده‌اند [۱۵]. یک شبکه کانولوشنی زمانی بر پایه توجه ارائه شده است که آن نقطه‌های زمانی ورودی را در نظر می‌گیرد که در خروجی‌های آینده بیشترین تأثیرگذاری را دارند [۱۶]. دو شبکه عصبی به نام‌های HSCP و EHSCP در [۳] پیشنهاد شده‌اند. طرح HSCP از سه گره LSTM سری تشکیل شده است و تنها سری زمانی اصلی را در نظر می‌گیرد. طرح EHSCP بر پایه HSCP طراحی شده است، با این تفاوت که هم سری زمانی اصلی و هم سری‌های زمانی ثانویه را در محاسبات خود در نظر می‌گیرد.

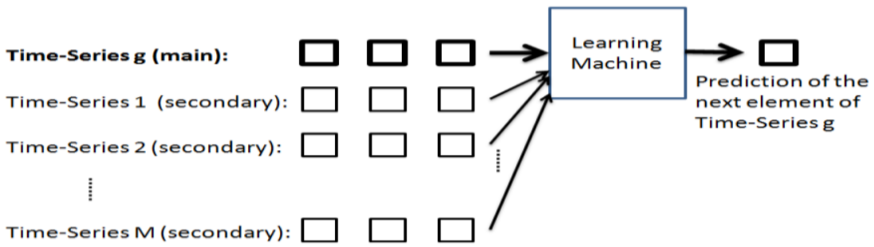
در این پژوهش برای افزایش دقت پیش‌بینی در مسئله HGEP، یک شبکه عصبی مصنوعی طراحی شد که روش جدیدی را ارائه می‌کند که بتوان یک ویژگی چندمتغیره (Multivariate Feature) را با ترکیب سری‌های زمانی ثانویه استخراج کرد. این ویژگی سبب می‌شود که طرح این مطالعه بتواند تأثیر سری‌های زمانی ثانویه بر سری زمانی اصلی را محاسبه کند. در نتیجه، شبکه عصبی این مطالعه می‌تواند پیش‌بینی دقیق‌تری ارائه کند. نوآوری‌های دیگر در این پژوهش عبارت اند از:

- در شبکه عصبی پیشنهادی، مسیرهای داخلی اختصاصی به عنصرهای مهم از سری‌های زمانی ورودی اختصاص پیدا می‌کند.
- شبکه عصبی پیشنهادی به این صورت طراحی شد که می‌تواند هزاران سری زمانی ثانویه را در نظر بگیرد در حالی که روش‌های موجود نمی‌توانند بیش از صد سری زمانی ثانویه را در نظر بگیرند.
- تعدادی پارامتر داخلی در شبکه عصبی پیشنهادی در نظر گرفته شد که با پیکربندی آن‌ها می‌توان خطای پیش‌بینی را تنظیم نمود.
- شبکه عصبی پیشنهادی با استفاده از ۱۵۵ مجموعه داده آموزش داده و در ۳۹ مجموعه داده ارزیابی شد.
- ارزیابی‌های پژوهش حاضر مقدار بهینه برای پارامترهای داخلی طرح پیشنهادی را مشخص کرده و نشان می‌دهند شبکه عصبی طراحی‌شده از نظر خطای پیش‌بینی و تعداد مراحل تقسیم/تمایز که به درستی پیش‌بینی شده باشند، نسبت به روش‌های موجود بهتر عمل می‌کند.
- ارزیابی‌های این مطالعه نشان می‌دهند طرح پیشنهادی می‌تواند پیش‌بینی را برای صدها مرحله از تقسیم/تمایز سلولی انجام دهد.
- بر پایه ارزیابی‌هایی که انجام شد، اگر بخواهیم یک مرحله از تقسیم/تمایز پیش‌بینی شود، طرح پیشنهادی اندکی از نظر دقت پیش‌بینی روش‌های موجود را بهبود داده است. اگر تعداد زیادی از مرحله‌های تقسیم/تمایز باید پیش‌بینی شوند، آن‌گاه طرح پیشنهادی به طور قابل‌توجهی دقت پیش‌بینی را در مقایسه با طرح‌های موجود بهبود می‌دهد.

روش

در این بخش، یک ماشین یادگیرنده جدید به نام Input-efficient Attention-based Multivariate) IAMTSP (Time Series Prediction ارائه شد. الگوریتم‌های پیش‌بینی و آموزش در این طرح همانند الگوریتم‌های پیش‌بینی و آموزش در HSCP است [۳].

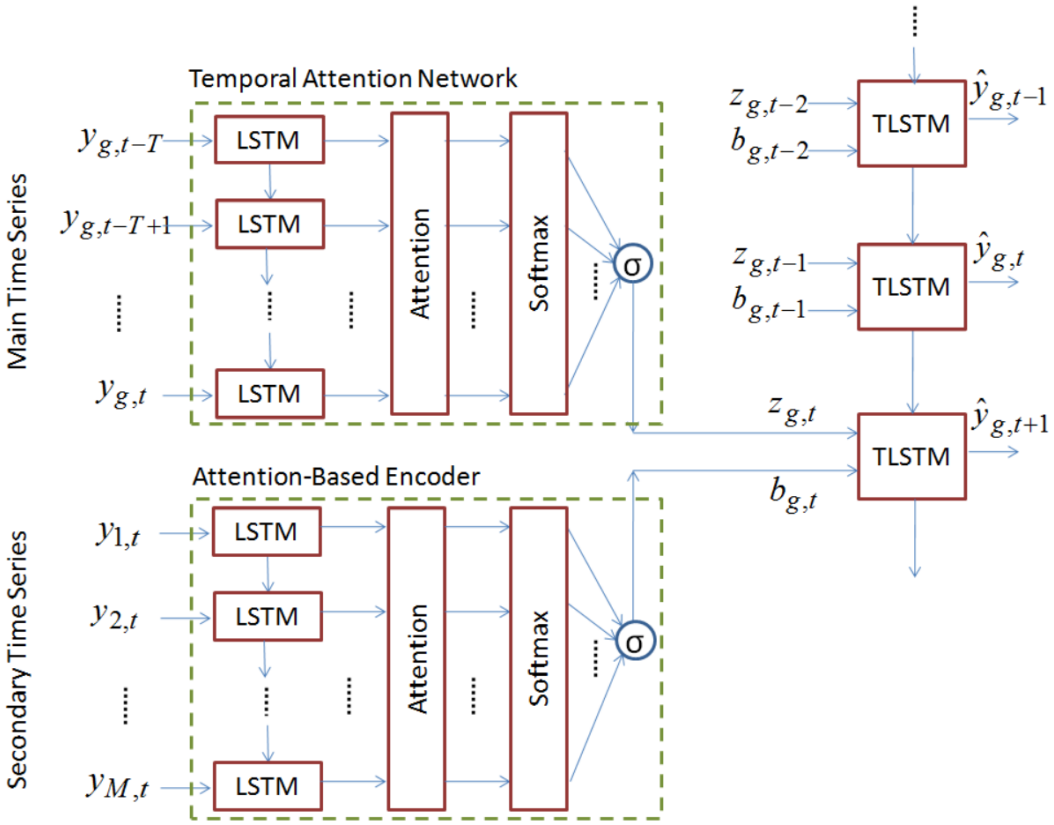
ماشین یادگیرنده در طرح IAMTSP به صورت حالت کلی نمایش داده شده در شکل ۳ طراحی شد. یک ماشین یادگیرنده به هر ژن اختصاص داده شد، به طوری که ماشین g عنصر بعدی در سری زمانی g را پیش‌بینی می‌کند ($g = \{1,2, \dots, M\}$). برای ژن g در ماشین g، سری زمانی g به عنوان سری زمانی ورودی اصلی و (M-1) سری زمانی دیگر (متعلق به ژن‌های دیگر) به عنوان سری‌های زمانی ورودی ثانویه در نظر گرفته می‌شوند.



شکل ۳: طراحی کلی ماشین شماره g در طرح پیشنهادی

طراحی ماشین یادگیرنده

یک ماشین یادگیری طراحی شد که در شکل ۴ مشاهده می‌شود. دو مجموعه مسیر جداگانه به سری‌های زمانی اصلی و ثانویه اختصاص داده شد. در برش زمانی t، ماشین $\hat{y}_{g,t+1}$ را پیش‌بینی می‌کند که مقدار پیش‌بینی شده برای $y_{g,t+1}$ است. در هر برش زمانی، به تعداد T عنصر از سری زمانی اصلی و یک عنصر از هر سری زمانی ثانوی وارد ماشین می‌شوند. آن‌ها از بخش‌های داخلی شبکه عبور می‌کنند و در پایان $\hat{y}_{g,t+1}$ تولید خواهد شد. T یک پارامتر قابل پیکربندی در ماشین است که می‌توان آن را با یک عدد دلخواه مقداردهی کرد. اگر $y_{g,t+1}$ در ماتریس Y وجود داشته باشد، آنگاه از $\hat{y}_{g,t+1}$ چشم‌پوشی می‌شود و $y_{g,t+1}$ در برش زمانی بعدی برای ورود به ماشین استفاده می‌شود.



شکل ۴: ساختار داخلی ماشین شماره g برای برش زمانی t در طرح پیشنهادی

این ماشین دارای دو مرحله است:

- مرحله ۱: این مرحله شامل تعدادی لایه عصبی برای کدگذاری سری‌های زمانی ورودی به شکلی است که توسط مرحله دوم قابل استفاده باشد. همچنین، این مرحله عنصرهای قبلی از سری‌های زمانی ورودی را به خاطر می‌سپارد.
- مرحله ۲: این مرحله شامل یک لایه عصبی برای استخراج ویژگی‌های چندمتغیره از سری‌های زمانی ورودی و سپس پیش‌بینی عنصر بعدی از سری زمانی اصلی است.

این ماشین حاوی سه بخش اصلی است:

- کدگذار زمانی: به تعداد T عنصر از سری زمانی اصلی در نقطه زمانی t وارد این بخش می‌شوند. این عنصرهای ورودی شامل $y_{1,t-T}, y_{1,t-T+1}, \ldots, y_{1,t}$ هستند. این بخش ورودی‌های سری قبلی را به خاطر می‌سپارد و $z_{g,t}$ را تولید می‌کند که می‌توان آن را به عنوان پیش‌بینی اولیه برای $y_{g,t+1}$ لحاظ کرد. کدگذار مکانی: یک عنصر از هر سری زمانی ثانویه در نقطه زمانی t وارد این بخش می‌شود. همه این عناصر ورودی متعلق به نقطه زمانی t هستند. این بخش مقادیر فعلی سری‌های زمانی ثانویه را می‌گیرد و آن‌ها را به شکل یک ویژگی چندمتغیره یکسان با نام $b_{g,t}$ کدگذاری می‌کند.

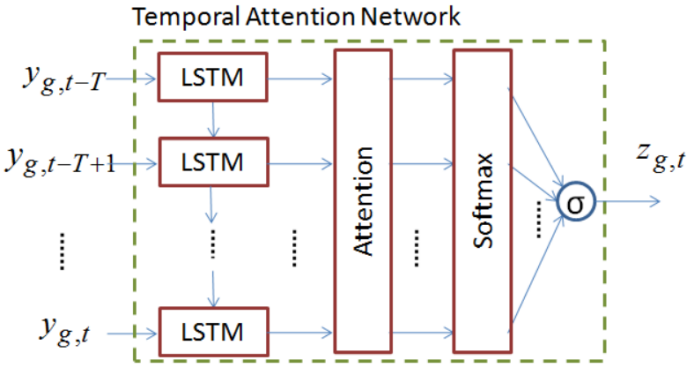
- TLSTM: این واحد تأثیرات $y_{g,t}$ و $b_{g,t}$ بر روی $z_{g,t}$ برای پیش‌بینی $\hat{y}_{g,t+1}$ را در نظر می‌گیرد.

در این مطالعه از واحد LSTM استاندارد و واحد Attention افزایشی استفاده شده است [۱۷، ۱۱]. TLSTM یک واحد Tree-LSTM Child-sum است که تنها یک فرزند سمت چپ و یک فرزند سمت راست دارد [۱۸].

پارامتر $b_{g,t}$ را می‌توان به عنوان یک فاکتور در TLSTM در نظر گرفت که مقدار افزایش یا کاهش در $z_{g,t}$ برای تولید $\hat{y}_{g,t+1}$ را نشان می‌دهد. در طول آموزش، ماشین یاد می‌گیرد که در هر نقطه زمانی چه چیزی را به $b_{g,t}$ تخصیص دهد تا $z_{g,t}$ افزایش یا کاهش یابد. به عنوان مثال، $b_{g,t}$ می‌تواند منجر به افزایش شدید در $z_{g,t}$ شود. با آموزش، کدگذار عصبی مصنوعی یاد می‌گیرد که برخی از سری‌های زمانی ثانویه تأثیر قابل‌توجهی بر $\hat{y}_{g,t+1}$ ندارند و وزن‌های کمی به آن‌ها تخصیص می‌دهد. همچنین، یاد می‌گیرد چه وزن‌هایی به سری‌های زمانی ثانویه دیگر بدهد که تأثیرات بیشتری بر $\hat{y}_{g,t+1}$ داشته باشند.

کدگذاری سری زمانی ورودی

با توجه به سری زمانی اصلی در شکل ۵ در نقطه زمانی t، به تعداد T عنصر آخر از سری زمانی اصلی وارد ماشین می‌شوند. لایه LSTM این عنصرها را به T مقدار میانی تبدیل می‌کند. همچنین، عنصرهای قبلی را در حافظه داخلی خود به یاد می‌سپارد. سپس لایه Attention/Softmax هر کدام از این مقدارهای میانی را در یک وزن مناسب ضرب می‌کند. سرانجام، مقدار $z_{g,t}$ تولید می‌شود که نشان‌دهنده عنصرهای سری زمانی اصلی است که یک ضریب وزنی به عنوان درجه اهمیت در هر یک از آن‌ها ضرب شده است. در این طراحی، به تعداد T عنصر از سری زمانی اصلی مسیرهای اختصاصی به سمت مرحله دوم از ماشین را به دست می‌آورند. این برخلاف روش‌های موجود مانند طرح *DA-RNN* است که از تعداد کمی مسیر مشترک برای سری‌های زمانی اصلی و ثانویه استفاده می‌کنند.



شکل ۵: کدگذاری سری زمانی اصلی برای برش زمانی t در ماشین g

هزاران عنصر از سری‌های زمانی ثانویه می‌آیند. برخی از این عنصرها اطلاعات مهمی در مورد سری‌های زمانی اصلی ارائه می‌دهند که می‌توان در مرحله ۲ برای پیش‌بینی دقیق از آن استفاده کرد؛ اما عنصرهای زیادی نیز بین آن‌ها وجود دارند که دارای اطلاعات سودمندی برای این هدف نیستند. کدگذار مکانی عنصرهای بی‌ارزش را از عنصرهای ورودی حذف می‌کند تا عنصرهای ورودی به یک ویژگی تبدیل شوند.

ویژگی‌های طرح

طراحی این مطالعه در شکل ۴ از ویژگی‌های زیر برخوردار است:

- سری‌های زمانی اصلی و ثانویه را در پیش‌بینی لحاظ می‌کند.
- وابستگی‌های درون سری‌های زمانی و بین سری‌های زمانی را در پیش‌بینی لحاظ می‌کند.
- دو مسیر جداگانه برای سری‌های زمانی اصلی و ثانویه در شبکه عصبی خود لحاظ می‌کند؛ یکی برای T عنصر آخر از سری زمانی اصلی و مسیر دیگر برای عناصر آخر از M سری زمانی ثانویه. با این کار، طرح ما اولویت بیشتری به سری زمانی اصلی نسبت به سری‌های زمانی ثانویه می‌دهد.

- طرح ما به تعداد T عنصر آخر از سری زمانی اصلی را در نظر می‌گیرد تا $z_{g,t}$ را تولید کند. با این کار، تعداد کافی از عناصر سری زمانی اصلی امکان تأثیر بر $z_{g,t}$ را پیدا می‌کنند.
- برای تولید $\hat{y}_{g,t+1}$ در TLSTM، همه سری‌های زمانی ورودی در نظر گرفته می‌شوند.
- در روش‌های قبلی (بخش ۱)، همه عناصر ورودی تقریباً دارای شانس یکسانی برای گذر از تعدادی مسیر مشترک هستند. این یعنی یک عنصر ورودی مهم تنها با به دست آوردن وزن‌های داخلی بالا در مؤلفه‌های شبکه عصبی می‌تواند یک مسیر به خروجی به دست آورد. برای به دست آوردن یک مسیر با وزن‌های داخلی یک آموزش انحصاری مورد نیاز است. برعکس، IAMTSP با تخصیص مسیرهای اختصاصی به سمت خروجی برای عناصر ورودی مهم، این مشکل را حل کرده است. به همین دلیل است که IAMTSP سریع‌تر از روش‌های دیگر به یک آموزش بهینه می‌رسد.

نتایج

در این بخش، درصد خطای پیش‌بینی‌هایی این طرح در شرایط گوناگون ارزیابی شد. برای مقایسه با ماشین پیشنهادی، دو طرح DA-RNN [۱۱] و EHSCP [۳] انتخاب شد که از بهترین و جدیدترین ماشین‌های یادگیرنده برای پیش‌بینی سری‌های زمانی چندمتغیره هستند. همچنین، این طرح با روش آماری Bhattacharjee2019 [۶] مقایسه شد. همچنین این طرح به زبان Python نسخه ۳ و با استفاده از کتابخانه Keras در سیستم عامل Linux پیاده‌سازی شد. برای واحد LSTM از LSTM استاندارد پیاده‌سازی شده در Keras استفاده شد.

تعریف: پارامتر D به عنوان فاصله ستونی که باید پیش‌بینی شود با ستون k در ماتریس Y تعریف شد. این یعنی این که D تعداد مراحل تقسیم/تمایز سلولی است که پس از برش زمانی k در ماتریس Y پیش‌بینی شد.

خطای پیش‌بینی در نمودارهای خود به کمک دو معیار MAE (Mean Absolute Error) و MAPE (Mean Absolute Percentage Error) نمایش داده شد. هر چه این دو معیار به صفر نزدیک‌تر باشند، درستی پیش‌بینی افزایش می‌یابد. این دو معیار به صورت (۱) و (۲) محاسبه می‌شوند جایی که n تعداد عناصر پیش‌بینی شده در همه مجموعه‌های داده ارزیابی است. $\hat{A}_i$ مقدار پیش‌بینی‌شده و A_i مقدار واقعی است.

(۱)
$$MAE = (1/n).\sum_{i=1}^{n}|A_i - \hat{A}_i|$$

(۲)
$$MAPE = (100\%/n).\sum_{i=1}^{n}|(A_i - \hat{A}_i)/A_i|$$

مجموعه‌های داده

تعداد ۱۹۴ مجموعه داده از پایگاه‌های GEO و ArrayExpress گردآوری گردید که شامل ماتریس بیان ژنی مربوط به سلول‌های بنیادی هماتوپویتیک بودند که در بیش از یک نقطه زمانی اندازه‌گیری شده بودند. برای نمونه، می‌توان به شناسه‌های مجموعه داده مانند *GSE4655*، *GSE10584*، *GSE12407* از پایگاه *GEO* اشاره کرد. مجموعه‌های داده برگزیده به طور تقریبی در شرایط آزمایشی همانند فراهم شده‌اند. همه مجموعه‌های داده که استفاده شد، متعلق به گونه Homo-Sapiens بوده و با استفاده از تکنولوژی Affymetrix-GeneChip تولید شده‌اند.

از ۸۰ درصد از مجموعه‌های داده برگزیده برای آموزش، ۱۰ درصد برای اعتباربخشی و ۱۰ درصد برای ارزیابی استفاده شد. همه مجموعه‌های داده برگزیده در محدوده عددی [۰،۱] نرمال‌سازی گردید. طول برش زمانی برابر با ۵ ساعت تعیین شد. اگر چه این طول برای تمایز سلول کمی کوتاه است، اما سبب شده است مجموعه‌های داده پس از درج ستون، یک تعداد کافی از ستون‌ها را به دست آورند. پس از انتخاب طول برش زمانی، این نیاز وجود دارد که تعدادی ستون به برخی از مجموعه‌های داده افزوده شود [۳].

سناریوهای ارزیابی

برای ارزیابی طرح پیشنهادی، چهار سناریو تعریف کرده و پارامترهای موردنیاز برای اجرای هر یک از این سناریوها تعیین شد. جدول ۱ این پارامترها را نشان می‌دهد. در سناریوی یک، پارامتر k در هر بار اجرای برنامه تغییر داده شد. در سناریوی دو، پارامتر D در هر بار اجرای برنامه تغییر داده شد تا تأثیر تعداد مرحله‌های پیش‌بینی در عملکرد طرح پیشنهادی و درست بودن خروجی آن سنجیده شود. در سناریوی سه، در طرح IAMTSP، پارامتر T در هر بار اجرای برنامه تغییر داده

شد تا تأثیر این پارامتر را بر روی درستی خروجی در این طرح پیشنهادی سنجیده شود. در سناریوی چهار در طرح IAMTSP، همانند سناریوی دو رفتار شد، ولی تعداد مختلفی از سری‌های زمانی در نظر گرفته شد تا تأثیر تعداد سری‌های زمانی در عملکرد طرح پیشنهادی و درست بودن خروجی آن سنجیده شود.

جدول ۱: پارامترهای ارزیابی در سناریوهای تعریف‌شده

	Scenario I	Scenario II	Scenario III	Scenario IV
k	۱،۲،۴،۸،۱۶	۸	۸	۸
M	۱۰۰	۱۰۰	۱۰۰	۱۰،۱۰۰،۱۰۰۰
D	۱	۱،۴،۱۶،۶۴،۱۲۸	۱	۱،۴،۱۶
T	۱۰	۱۰	۱،۲،۴،۸،۱۶،۳۲	۱۰

نتایج در سناریوی یک

در آغاز، روش این مطالعه در سناریوی یک ارزیابی شد که در آن، تعداد مختلفی از ستون‌های نخست در ماتریس GES که باید پیش‌بینی شود، شناخته شده هستند. اگر k بزرگ باشد، ماشین می‌تواند شرایطی را که سلول‌ها (یا سری‌های زمانی) در آن قرار دارند بهتر درک کند؛ بنابراین پیش‌بینی دقیق‌تری برای ستون $(k+1)$ ارائه دهد.

جدول ۲ نتایج سناریوی یک را نشان می‌دهد. همه روش‌های ارزیابی‌شده برای $k \geq 4$ دقت قابل قبولی دارند. از هیچ طرحی انتظار نمی‌رود که برای $k \leq 2$ دقت بالایی از خود نشان دهد. دلیل این است که اگر تعداد کافی از ستون‌های اولیه از ماتریسی که باید پیش‌بینی شود، معلوم نباشند، آنگاه ماشین نمی‌تواند به شناخت کافی از سری زمانی ورودی برسد. برای $k = 8$، طرح EHSCP به $MAPE = 3.07\%$ دست یافته است. طرح IAMTSP در مقایسه با چهار روش دیگر برای $k \geq 4$ خطای کمتری در پیش‌بینی‌ها داشته است. این نتایج ثابت می‌کنند IAMTSP می‌تواند برخی اطلاعات ارزشمند درباره عنصرهای سری‌های زمانی ورودی را به دست آورد که DA-RNN از آن‌ها چشم‌پوشی می‌کند. همچنین، IAMTSP تلاش می‌کند ویژگی‌های چندمتغیره بیشتری از سری‌های زمانی ورودی استخراج کند، در حالی که بر سری زمانی اصلی تأکید می‌کند.

جدول ۲: خطای پیش‌بینی در برابر k برای پیش‌بینی یک مرحله‌ای

	MAPE				MAE			
k	EHSCP	IAMTSP	DA-RNN	Bhattacharjee 2019	EHSCP	IAMTSP	DA-RNN	Bhattacharjee 2019
۱	۱۲/۶۳	۱۱/۴۹	۱۱/۰۵	۱۸/۱۷	۰/۰۶۷	۰/۰۶۴	۰/۰۵۹	۰/۰۹۶
۲	۷/۴۱	۷/۵	۸/۸۲	۱۲/۱۵	۰/۰۳۸	۰/۰۴۰	۰/۰۴۷	۰/۰۶۳
۴	۳/۸۴	۳/۷۶	۳/۹۲	۵/۱۴	۰/۰۲۰	۰/۰۲۰	۰/۰۲۰	۰/۰۲۶
۸	۳/۰۷	۳/۰۴	۳/۲۹	۳/۸۵	۰/۰۱۶	۰/۰۱۶	۰/۰۱۷	۰/۰۱۹۰
۱۶	۳/۰۲	۲/۹۱	۳/۱	۳/۴۴	۰/۰۱۵	۰/۰۱۵	۰/۰۱۶	۰/۰۱۷

نتایج در سناریوی دو

در سناریوی دو، طرح پیشنهادی با بررسی تعداد مراحل تقسیم/تمایزی که به درستی پیش‌بینی کرده است، ارزیابی شد. به تعداد $k = 8$ ستون نخست از ماتریس Y معلوم هستند. به ازای $l = 1,2,\dots,D$، برای پیش‌بینی ستون $(k+l)$، در واقع به تعداد $(k+l-1)$ ستون نخست به صورت سریال به ماشین داده شد.

جدول ۳ نتایج سناریوی دو را نشان می‌دهد. با افزایش D، خطای پیش‌بینی به تدریج در طرح‌های ارزیابی‌شده افزایش می‌یابد، ولی این افزایش در EHSCP و IAMTSP نسبت به دو روش دیگر کمتر است. این نشان می‌دهد که این دو طرح توانایی بیشتری در کارهای زیر دارند:

- شناسایی وضعیت سری زمانی اصلی
- به یاد سپردن عنصرهای پیشین در سری زمانی اصلی
- شناسایی وضعیت سری‌های زمانی ثانویه

جدول ۳: خطای پیش‌بینی در برابر فاصله ستون پیش‌بینی‌شده از ستون k

D	MAPE EHSCP	MAPE IAMTSP	MAPE DA-RNN	MAPE Bhattacharjee 2019	MAE EHSCP	MAE IAMTSP	MAE DA-RNN	MAE Bhattacharjee 2019
۱	۳/۰۷	۳/۰۴	۳/۲۹	۳/۸۵	۰/۰۱۶	۰/۰۱۶	۰/۰۱۷	۰/۰۲۰
۴	۳/۸۹	۳/۷۶	۴/۰۶	۵/۵۱	۰/۰۲۰	۰/۰۲	۰/۰۲۱	۰/۰۲۹
۱۶	۵/۶۲	۵/۵	۷/۵۲	۱۱/۰۷	۰/۰۲۹	۰/۰۳۱	۰/۰۴۰	۰/۰۶۱
۶۴	۸/۴۴	۷/۸۳	۱۱/۷۵	۱۸/۴۴	۰/۰۵۶	۰/۰۴۵	۰/۰۶۶	۰/۱۰۹
۱۲۸	۱۲/۲۸	۱۱/۰۶	۱۶/۳۸	۳۱/۲۹	۰/۰۷۴	۰/۰۶۸	۰/۰۹۵	۰/۲۰۵

دو طرح EHSCP و IAMTSP در سناریوی یک با $D = 1$ نسبت به روش‌های دیگر کمی بهتر عمل می‌کنند؛ اما سناریوی دو (جدول ۳) نشان می‌دهد که دو طرح EHSCP و IAMTSP در حالتی که D از ۱۶ بزرگ‌تر باشد، بهبود چشم‌گیری در مقایسه با طرح‌های موجود ارائه می‌دهند؛ یعنی تعداد بیشتری از مراحل تمایز/تقسیم سلولی می‌تواند پیش‌بینی شود. با این حال، طرح IAMTSP خطای کمتری نسبت به طرح EHSCP در پیش‌بینی‌ها داشته است.

نتایج در سناریوی سه

در این سناریو، مقدارهای مختلفی برای T استفاده شده‌اند ولی M ثابت است. می‌خواهیم بررسی کنیم که آیا انتخاب یک T بزرگ‌تر منجر به بهبود دقت پیش‌بینی می‌شود یا خیر؟

در جدول ۴ مشاهده می‌شود هنگامی که T افزایش می‌یابد، خطای پیش‌بینی کاهش پیدا می‌کند؛ زیرا با انتخاب یک T بزرگ‌تر، اطلاعات عنصرهای بیشتری از سری زمانی اصلی در ماشین استفاده می‌شوند. آن گاه ماشین می‌تواند دانش بیشتری درباره سری زمانی اصلی داشته باشد و پیش‌بینی دقیق‌تری از $y_{g,k+1}$ ارائه دهد؛ اما مشاهده شد که MAPE در نقطه $T = 8$ به یک مقدار نزدیک به بهینه دست می‌یابد، به طوری که افزایش بیشتر در T منجر به کاهش چشم‌گیری در MAPE نمی‌شود.

جدول ۴: خطای پیش‌بینی در برابر T در طرح IAMTSP

T	MAPE in IAMTSP
۰	۹/۱۲
۲	۵/۴۴
۴	۳/۷۳
۸	۳/۰۴
۱۶	۲/۹۱
۳۲	۲/۸۹

با افزایش در پارامتر T، تعداد عنصرهایی از سری‌های زمانی ورودی افزایش داده شد که در برش زمانی جاری، مسیر مستقیم به خروجی نهایی ماشین دارند. این به آن معنی نیست که تنها T عنصر آخر از سری زمانی اصلی می‌توانند بر خروجی نهایی ماشین تأثیر بگذارند؛ زیرا واحدهای LSTM در مسیرهای سری زمانی اصلی و سری‌های زمانی ثانویه جای داده شد. هرکدام از این واحدهای LSTM دارای یک سلول حافظه داخلی هستند که می‌تواند همه عنصرهای قبلی از سری زمانی ورودی خود را به صورت یک مقدار تکی به خاطر بسپارد. این یعنی این که همه عنصرهای موجود در همه سری‌های زمانی ورودی می‌توانند بر خروجی نهایی ماشین در برش زمانی جاری تأثیر بگذارند.

نتایج در سناریوی چهار

طرح‌های گذشتگان در پیش‌بینی سری زمانی چندمتغیره به گونه‌ای طراحی شده‌اند که به ازای هر سری زمانی ورودی، یک LSTM جداگانه در نظر می‌گیرند؛ بنابراین نمی‌توانند هزاران سری زمانی ورودی داشته باشند. اکنون در سناریوی چهار، طرح پیشنهادی در حالتی که تعداد سطرهای ماتریس‌ها

برابر با ۱۰، ۱۰۰ و ۱۰۰۰ باشد، ارزیابی شد.

جدول ۵ نتایج سناریوی چهار را نشان می‌دهد. با افزایش D، خطای پیش‌بینی به تدریج در طرح‌های ارزیابی‌شده افزایش می‌یابد، اما برای یک M بزرگ‌تر، سرعت این افزایش کمتر است. علت آن است که برای یک M بزرگ‌تر، تعداد سری‌های زمانی ثانویه بیشتر است. در نتیجه، اطلاعات بیشتری برای پیش‌بینی سری زمانی اصلی وجود دارد.

جدول ۵: خطای پیش‌بینی در برابر فاصله ستون پیش‌بینی‌شده از ستون *k*

MAPE in IAMTSP			
D	M=۱۰	M=۱۰۰	M=۱۰۰۰
۱	۳/۱۶	۳/۰۴	۲/۹۵
۴	۳/۸۵	۳/۷۶	۳/۶۸
۱۶	۶/۱۹	۶/۰۲	۵/۸۳

در حالت M=۱۰۰۰ نسبت به حالت M=۱۰، مقدار M صد برابر شده است، ولی زمان موردنیاز برای آموزش شبکه عصبی در طرح IAMTSP حدود ۶/۵ برابر افزایش یافته است. دلیل این است که با افزایش مقدار M، تعداد ورودی‌های شبکه عصبی با همان نسبت افزایش پیدا می‌کند، ولی تعداد گره‌های موردنیاز در شبکه عصبی با نسبت بسیار کمتری افزایش می‌یابد.

بحث و نتیجه‌گیری

با توجه به ناکارآمدی‌های طرح‌های موجود در حالت چندمتغیره، ماشین IAMTSP طراحی شد که قدرت بیشتری در استخراج ویژگی‌های سری‌های زمانی ورودی دارد. طرح IAMTSP تلاش می‌کند تأثیر سری‌های زمانی ثانویه بر روی سری زمانی اصلی را در نظر گیرد. این تأثیر به صورت یک ویژگی از سری‌های زمانی ثانویه تولید می‌شود. طرح IAMTSP سری‌های زمانی ورودی را به دو بخش تقسیم می‌کند: یک بخش برای کدگذاری زمانی سری زمانی اصلی و بخش دیگر برای کدگذاری سری‌های زمانی ثانویه. این طرح وابستگی درون سری زمانی و بین سری زمانی را به شکل کارآمدی در نظر می‌گیرد. برخلاف روش‌های موجود که با گذراندن ویژگی‌های چندمتغیره از یک مسیر یکسان آن‌ها را سرکوب می‌کنند، طرح IAMTSP تلاش می‌کند مسیرهای داخلی اختصاصی به عنصرهای مهم از سری زمانی اصلی اختصاص می‌دهد.

طرح IAMTSP در پیش‌بینی سری‌های زمانی اصلی را بر سری‌های زمانی ثانویه ترجیح می‌دهد. مرحله ۱ از ماشین پیشنهادی، به تعداد T عنصر از سری زمانی اصلی را در نظر می‌گیرد. با این کار تعداد کافی از عنصرهای سری زمانی اصلی امکان تأثیر بر $\hat{y}_{g,t+1}$ را خواهند داشت. برای تولید $\hat{y}_{g,t+1}$، همه عنصرهای سری‌های زمانی ورودی در پیش‌بینی در نظر گرفته می‌شوند، اگر چه برخی از این عنصرها مورد تأکید هستند.

در روش‌های موجود، همه عنصرهای ورودی دارای شانس برابری برای گذر از تعدادی مسیر مشترک دارند. این یعنی این که یک عنصر ورودی مهم می‌تواند تنها با به دست آوردن وزن‌های داخلی بالا در بخش‌های شبکه عصبی، یک مسیر به سمت خروجی به دست آورد. برای به دست آوردن یک مسیر با استفاده از وزن‌های داخلی، به یک آموزش طاقت‌فرسا نیاز است. به جای آن، طرح IAMTSP با تخصیص دادن مسیرهای اختصاصی به سمت خروجی برای عنصر ورودی مهم، این فرآیند را آسان کرده است. به همین دلیل است که انتظار می‌رود طرح IAMTSP سریع‌تر از روش‌های موجود به یک نقطه آموزش بهینه برسد.

اگر چه برای سلول‌های بنیادی هماتوپویتیک از طرح پیشنهادی استفاده شد، می‌توان این طرح را برای پیش‌بینی انواع دیگر سلول‌های بنیادی نیز به کار برد، مانند سلول‌های بنیادی جنین و یا سلول‌های Induced Pluripotent Stem Cell. همچنین، به عنوان یک موضوع پژوهشی در آینده می‌توان آموزش طرح پیشنهادی توسط مجموعه‌های داده متعلق به همه انواع سلول‌های بنیادی را بررسی کرد، به طوری که طرح پیشنهادی به یک ماشین جامع تبدیل شود که بتواند تقسیم/تمایز همه انواع سلول‌ها را پیش‌بینی کند.

همچنین، از آنجا که طرح پیشنهادی تنها ماتریس بیان ژنی را پیش‌بینی می‌کند، می‌توان چند پژوهش در زمینه ترجمه یک ماتریس بیان ژنی پیش‌بینی شده به نوع و مشخصات سلول آینده انجام داد. برای نمونه، پژوهشگران می‌توانند از طرح پیشنهادی برای پیش‌بینی تعداد گلبول‌های قرمز یا گلبول‌های سفیدی که انتظار می‌رود یک سلول بنیادی ویژه تولید کند، استفاده نمایند. برای نمونه دیگر، پژوهشگران می‌توانند از طرح پیشنهادی برای پیش‌بینی احتمال تولید سلول‌های ضعیف توسط یک سلول بنیادی استفاده کنند.

تعارض منافع

بدین‌وسیله نویسندگان تصریح می‌نمایند که در مورد پژوهش حاضر هیچ‌گونه تضاد منافعی وجود ندارد.

References

1. Scala S, Aiuti A. In vivo dynamics of human hematopoietic stem cells: novel concepts and future directions. Blood Adv 2019;3(12):1916-24. doi: 10.1182/bloodadvances.2019000039.

2. Barbosa CM, Fock RA, Hastreiter AA, Reutelingsperger C, Perretti M, Paredes-Gamero EJ, Farsky SH. Extracellular annexin-A1 promotes myeloid/granulocytic differentiation of hematopoietic stem/progenitor cells via the Ca 2+/MAPK signalling transduction pathway. Cell Death Discovery 2019;5(1):1.

3. Eskandarian P, Mohasefi JB, Pirnejad H, Niazkhani Z. Prediction of future gene expression profile by analyzing its past variation pattern. Gene Expr Patterns 2021;39:119166. doi: 10.1016/j.gep.2021.119166.

4. Kjartansdóttir KR, Reda A, Panula S, Day K, Hultenby K, Söder O, et al. A combination of culture conditions and gene expression analysis can be used to investigate and predict hES cell differentiation potential towards male gonadal cells. PloS one 2015;10(12):e0144029.
https://doi.org/10.1371/journal.pone.0144029

5. Yanagihara K, Liu Y, Kanie K, Takayama K, Kokunugi M, Hirata M, et al. Prediction of differentiation tendency toward hepatocytes from gene expression in undifferentiated human pluripotent stem cells. Stem Cells Dev 2016; 25(24): 1884–97. doi: 10.1089/scd.2016.0099

6. Bhattacharjee A, Vishwakarma GK. Time-course data prediction for repeatedly measured gene expression. International Journal of Biomathematics 2019;12(04):1950033.

7. Chakraborty K, Mehrotra K, Mohan CK, Ranka S. Forecasting the behavior of multivariate time series using neural networks. Neural Networks 1992;5(6):961-70.
https://doi.org/10.1142/S1793524519500335

8. Voyant C, Muselli M, Paoli C, Nivet ML. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy 2011;36(1):348-59.

9. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science 2011;2(1):1-8.
https://doi.org/10.1016/j.jocs.2010.12.007

10. Rao T, Srivastava S. Modeling movements in oil, gold, forex and market indices using search volume index and twitter sentiments. InProceedings of the 5th annual ACM Web Science Conference; 2013 May 2; New York NY, United States: Association for Computing Machinery; 2013. p. 336-45. https://doi.org/10.1145/2464464.2464521

11. Qin Y, Song D, Chen H, Cheng W, Jiang G, Cottrell G. A dual-stage attention-based recurrent neural network for time series prediction. International Joint Conference on Artificial Intelligence 2017.

12. Li Y, Zhu Z, Kong D, Han H, Zhao Y. EA-LSTM: Evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems 2019;181:104785.
https://doi.org/10.1016/j.energy.2020.119692

13. Bu SJ, Cho SB. Time series forecasting with multi-headed attention-based deep learning for residential energy consumption. Energies 2020;13(18):4722. https://doi.org/10.3390/en13184722

14. Yuan Y, Xun G, Ma F, Wang Y, Du N, Jia K, et al. Muvan: A multi-view attention network for multivariate temporal data. 18th IEEE International Conference on Data Mining, ICDM; 2018 Nov 17; Singapore, Singapore: IEEE; 2018. p. 717-26.

15. Hu J, Zheng W. Multistage attention network for multivariate time series prediction. Neurocomputing 2020;383:122-37.

16. Pantiskas L, Verstoep K, Bal H. Interpretable Multivariate Time Series Forecasting with Temporal Attention Convolutional Neural Networks. EEE Symposium Series on Computational Intelligence; 2020 Dec 1; Canberra, Australia: IEEE; 2020. p. 1687-94.

17. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2013 May 26-31; Vancouver, BC, Canada: IEEE; 2013. p. 6645-49. doi: 10.1109/ICASSP.2013.6638947

18. Li J, Liu C, Gong Y. Layer trajectory LSTM. arXiv preprint arXiv:1808.09522. 2018 Aug 28.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2021; 8(3):270-281

Multivariate Feature Extraction for Prediction of Future Gene Expression Profile

Eskandarian Parinaz[1], Bagherzadeh Mohasefi Jamshid[2]*, Pirnejad Habibollah[3], Niazkhani Zahra[4]

• **Received:** 6 Jul 2021 • **Accepted:** 15 Dec 2021

Introduction: The features of a cell can be extracted from its gene expression profile. If the gene expression profiles of future descendant cells are predicted, the features of the future cells are also predicted. The objective of this study was to design an artificial neural network to predict gene expression profiles of descendant cells that will be generated by division/differentiation of hematopoietic stem cells.

Method: The developed neural network takes the parent hematopoietic stem cell's gene expression profile as input and generates the gene expression profiles of its future descendant cells. A temporal attention was proposed to encode the main time series and a spatial attention was also provided to encode the secondary time series.

Results: To make an acceptable prediction, the gene expression profiles of at least four initial division/differentiation steps must be known. The designed neural network surpasses the existing neural networks in terms of prediction accuracy and number of correctly predicted division/differentiation steps. The proposed scheme can predict hundreds of division/differentiation steps. The proposed scheme' error in prediction of 1, 4, 16, 64, and 128 division/differentiation steps was 3.04, 3.76, 5.5, 7.83, and 11.06 percent, respectively.

Conclusion: Based on the gene expression profile of a parent hematopoietic stem cell, the gene expression profiles of its descendants can be predicted for hundreds of division/differentiation steps and if necessary, solutions must be sought to encounter future genetic disorders.

Keywords: Hematopoietic Stem Cell, Neural Network, Multivariate Time Series, Gene Expression Profile, Prediction

• **Citation:** Eskandarian P, Bagherzadeh Mohasefi J, Pirnejad H, Niazkhani Z. Multivariate Feature Extraction for Prediction of Future Gene Expression Profile. Journal of Health and Biomedical Informatics 2021; 8(3): 270-81. [In Persian]

1. Ph.D. Candidate in Computer Engineering, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
2. Ph.D. in Software Engineering, Associate Professor, Department of Electrical and Computer Engineering, Urmia University, Urmia, Iran
3. Ph.D. in Medical Informatics, Associate Professor, Patient Safety Research Center, Clinical Research Institute, Urmia University of Medical Sciences, Urmia, Iran
4. Ph.D. in Medical Informatics, Associate Professor, Nephrology and Kidney Transplant Research Center, Clinical Research Institute, Urmia University of Medical Sciences, Urmia, Iran
*** Corresponding Author:** Jamshid Bagherzadeh Mohasefi
Address: Urmia Branch, Islamic Azad University, Urmia, Iran
• **Tel:** 0441-31803000 • **Email:** j.bagherzadeh@urmia.ac.ir