دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده علوم و فنون نوین
گروه شبکه

استنتاج درخت فیلوژنی تومور سرطانی با استفاده از داده‌های تک‌سلولی و تغییرات تعداد تکرار

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات
گرایش سامانه‌های شبکه‌ای

افشین بزرگ‌پور

اساتید راهنما

دکتر سامان هراتی‌زاده و دکتر ابوالفضل مطهری

مرداد ۱۴۰۰


بسم الله الرحمن الرحيم

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده علوم و فنون نوین
گروه شبکه

استنتاج درخت فیلوژنی تومور سرطانی با استفاده از داده‌های تک‌سلولی و تغییرات تعداد تکرار

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات
گرایش سامانه‌های شبکه‌ای

افشین بزرگ‌پور

اساتید راهنما

دکتر سامان هراتی‌زاده و دکتر ابوالفضل مطهری

مرداد ۱۴۰۰

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده علوم و فنون نوین

گواهی دفاع از پایان‌نامه کارشناسی ارشد

هیأت داوران پایان‌نامهٔ کارشناسی ارشد آقای / خانم افشین بزرگ‌پور به شمارهٔ دانشجویی ۸۳۰۵۹۶۰۰۵ در رشتهٔ مهندسی فناوری اطلاعات – گرایش سامانه‌های شبکه‌ای را در تاریخ ..... با عنوان «استنتاج درخت فیلوژنی تومور سرطانی با استفاده از داده‌های تک‌سلولی و تغییرات تعداد تکرار»

	به عدد	به حروف
با نمرهٔ نهایی		

و درجهٔ [ ] ارزیابی کرد.

ردیف	مشخصات هیأت داوران	نام و نام خانوادگی	مرتبهٔ دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر سامان هراتی‌زاده	استادیار	دانشگاه تهران	
۲	استاد راهنما	دکتر ابوالفضل مطهری	استادیار	دانشگاه تهران	
۳	استاد داور داخلی	دکتر داور داخلی	دانشیار	دانشگاه تهران	
۴	استاد مدعو	دکتر داور خارجی	دانشیار	دانشگاه داور خارجی	
۵	نمایندهٔ تحصیلات تکمیلی دانشکده	دکتر نماینده	دانشیار	دانشگاه تهران	

نام و نام خانوادگی معاون تحصیلات تکمیلی و پژوهشی دانشکده / گروه:

تاریخ و امضا:

نام و نام خانوادگی معاون آموزشی و تحصیلات تکمیلی پردیس دانشکده‌های فنی:

تاریخ و امضا:

تعهدنامهٔ اصالت اثر

باسمه تعالی

اینجانب افشین بزرگ‌پور تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آن‌ها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتری ارائه نشده است.

نام و نام خانوادگی دانشجو: افشین بزرگ‌پور

تاریخ و امضای دانشجو:

تقدیم به:

همسر و فرزندانم

و

پدر و مادرم

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را به زیور عقل آراست.

در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ اساتید راهنمای خود، جناب آقای دکتر ... و ...، صمیمانه تشکر و قدردانی کنم که در طول انجام این پایان‌نامه با نهایت صبوری همواره راهنما و مشوق من بودند و قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید.

از جناب آقای دکتر ... که زحمت مشاوره، بازبینی و تصحیح این پایان‌نامه را تقبل فرمودند کمال امتنان را دارم.

با سپاس بی‌دریغ خدمت دوستان گران‌مایه‌ام، خانم‌ها ... و آقایان ... در آزمایشگاه ...، که با همفکری مرا صمیمانه و مشفقانه یاری داده‌اند.

و در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می‌کنم وجود مقدس‌شان را و تشکر می‌کنم از خانواده عزیزم به پاس عاطفه سرشار و گرمای امیدبخش وجودشان، که بهترین پشتیبان من بودند.

افشین بزرگ‌پور

مرداد ۱۴۰۰

چکیده

این راهنما، نمونه‌ای از قالبِ پروژه، پایان‌نامه و رسالهٔ دانشگاه تهران می‌باشد که با استفاده از کلاس tehran-thesis و بستهٔ زی‌پرشین در LaTeX تهیه شده است. این قالب به گونه‌ای طراحی شده است که مطابق با دستورالعمل نگارش و تدوین پایان‌نامه کارشناسی ارشد و دکتری، مورخ ۹۳/۰۶/۰۳ پردیس دانشکده‌های فنی دانشگاه تهران باشد و حروف‌چینی بسیاری از قسمت‌های آن، مطابق با استاندارد قالب‌های فارسی پایان‌نامه در لاتک، به طور خودکار انجام می‌شود.

چکیده بخشی از پایان‌نامه است که خواننده را به مطالعه آن علاقمند می‌کند و یا از آن می‌گریزاند. چکیده باید ترجیحاً در یک صفحه باشد. در نگارش چکیده نکات زیر باید رعایت شود. متن چکیده باید مزین به کلمه‌ها و عبارات سلیس، آشنا، بامعنی و روشن باشد. بگونه‌ای که با حدود ۳۰۰ تا ۵۰۰ کلمه بتواند خواننده را به خواندن پایان‌نامه راغب نماید. چکیده، جدای از پایان‌نامه باید به تنهایی گویا و مستقل باشد. در چکیده باید از ذکر منابع، اشاره به جداول و نمودارها اجتناب شود.تمیز بودن مطلب، نداشتن غلط‌های املایی یا دستور زبانی و رعایت دقت و تسلسل روند نگارش چکیده از نکات مهم دیگری است که باید درنظر گرفته شود. در چکیده پایان‌نامه باید از درج مشخصات مربوط به پایان‌نامه خودداری شود. چکیده باید منعکس‌کننده اصل موضوع باشد. در چکیده باید اهداف تحقیق مورد توجه قرار گیرد. تأکید روی اطلاعات تازه (یافته‌ها) و اصطلاحات جدید یا نظریه‌ها، فرضیه‌ها، نتایج و پیشنهادها متمرکز شود. اگر در پایان‌نامه روش نوینی برای اولین بار ارائه می‌شود و تا به حال معمول نبوده است، با جزئیات بیشتری ذکر شود. شایان ذکر است چکیده فارسی و انگلیسی باید حتماً به تأیید استاد راهنما رسیده باشد.

کلمات کلیدی در انتهای چکیده فارسی و انگلیسی آورده می‌شود. محتوای چکیده‌ها بر اساس موضوع و گرایش تحقیق طبقه‌بندی می‌شود و به همین جهت وجود کلمات شاخص و کلیدی، مراکز اطلاعاتی را در طبقه‌بندی دقیق و سریع پایان‌نامه یاری می‌دهد. کلمات کلیدی، راهنمای نکات مهم موجود در پایان‌نامه هستند. بنابراین باید در حد امکان کلمه‌ها یا عباراتی انتخاب شود که ماهیت، محتوا و گرایش کار را به وضوح روشن نماید.

واژگان کلیدی حداکثر ۵ کلمه یا عبارت، متناسب با عنوان، قالب پایان‌نامه، لاتک

فهرست مطالب

[1] Tumor Evolution
[2] Variant allele frequency
[3] Allelec dropout

[4]Markov Chain Monte Carlo (MCMC)
[5]Machine learning
[6]Reinforcement learning
[7]Recurrent Neural Network

فهرست تصاویر

[8] Convolution
[9] Convolutional neural network
[10] Kernel
[11] Activation function
[12] Sigmoid
[13] Dropout

فهرست جداول

فهرست الگوریتم‌ها

فهرست برنامه‌ها

فصل ۱

مقدمه

تومور[1] از رشد غیر طبیعی سلول با احتمال حمله یا گسترش به سایر قسمت‌های بدن تشکیل می‌شود. تومورهای بدخیم[2] معمولاً سرطان[3] نامیده می‌شوند. سرطان علل مختلفی از جمله تغییرات ژنتیکی، آلودگی محیط زیست یا انتخاب‌های نادرست در سبک زندگی دارد. یک تومور ممکن است از زیرجمعیت‌های سلولی با تغییرات ژنومی‌مشخص تشکیل شده باشد، این پدیده ناهمگنی تومور[4] نامیده می‌شود. ناهمگنی تومور احتمالاً برای درمان سرطان و کشف نشانگر زیستی، به ویژه در روش‌های درمانی هدفمند، تأثیراتی خواهد داشت [۲۱]. درمان‌های فعلی، سرطان را به عنوان یک بیماری همگن درمان می‌کنند [۴۳].

داروهای هدفمند در برابر زیرجمعیت‌های تک یا چند سلولی با انکوژن[5] جهش‌یافته که آن‌ها را هدف قرار می‌دهند، تولید شده اند، در حالی که آن دسته از زیرجمعیت‌های سلولی که هیچ گونه تاثیری از داروهای به واسطه جهش خود، نمی‌گیرند بدون درمان باقی مانده و ممکن است منجر به عود مجدد تومور یا عدم درمان تومور می‌شوند [۲۱]. این زیرجمعیت‌های سلولی بدون درمان ممکن است منجر به پیشرفت تومور پس از درمان دارویی شوند [۲۱]. به عنوان مثال، رشد مجدد سلول‌های تومورزا در سرطان روده بزرگ[6] سرطان پستان و گلیوبالستوم[7] پس از تابش یا درمان سیکلوفسفامید مشاهده شده است [۴۳]. بنابراین، مطالعه روند رشد تومور و ناهمگنی آن تأثیرات زیادی بر تشخیص و درمان سرطان دارد.

تومورها می‌توانند خوش‌خیم، بدخیم و دارای رفتاری نامشخص یا ناشناخته باشند [۲]. تومورهای خوش خیم

[1] Tumor
[2] Malignant tumor
[3] Cancer
[4] Tumor heterogeneity
[5] Oncogene
[6] Colorectal carcinoma
[7] Glioblastomas

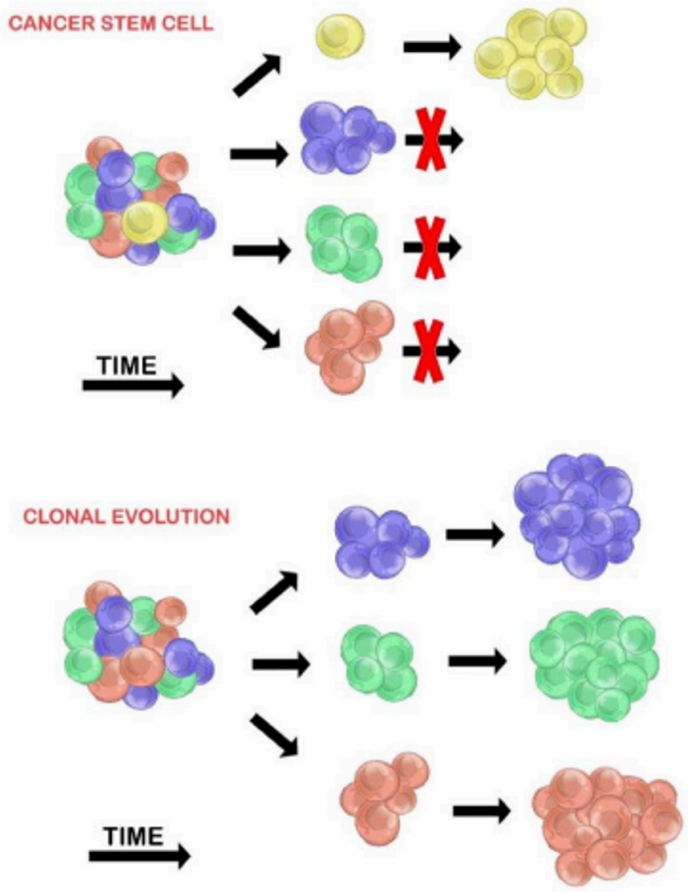
شامل فیبروییدهای رحمی[۸] و خال‌های ملانوسیتیک[۹] است. آن‌ها محدود و محلی[۱۰] هستند و به سرطان تبدیل نمی‌شوند [۴]. تومورهای بالقوه بدخیم[۱۱] شامل سرطان در محل[۱۲] هستند. آن‌ها به سایر بافت‌ها حمله نکرده و از بین نمی‌روند اما ممکن است به سرطان تبدیل شوند [۳]. تومورهای بدخیم را معمولاً سرطان می‌نامند. آن‌ها به بافت اطراف حمله کرده و از بین می‌روند، ممکن است متاستاز[۱۳] ایجاد کنند و اگر درمان نشوند یا به درمان پاسخ ندهند، کشنده خواهد بود [۳].

ناهمگنی تومور توضیح می‌دهد که تومور بیش از یک نوع سلول شامل می‌شود. انواع مختلف سلول‌های داخل تومور دارای ویژگی‌های مورفولوژیکی و فیزیولوژیکی متمایزی مانند گیرنده‌های سطح سلول، تکثیر[۱۴] و رگ‌زایی[۱۵] هستند. ناهمگنی تومور می‌تواند بین تومورها (ناهمگنی بین توموری) و یا درون تومورها (ناهمگنی درون توموری) رخ دهد. به طور گسترده‌ای پذیرفته شده است که توسعه تومور یک روند تکاملی است [۱۰]، و پیشرونده[۱۶] معمولاً از یک سلول منشأ می‌گیرند و گروهی از سلول‌ها را تشکیل می‌شوند که در نهایت یک توده را شکل می‌دهند.

دو مدل برای ناهمگنی تومور وجود دارد (شکل ۱.۱). یک مدل تشکیل سرطان از طریق سلول‌های بنیادی بوده که قابلیت ارث‌بری ندارند و مدل دیگر تشکیل سرطان از طریق تکامل کلونی[۱۷] بوده که قابلیت ارثبری دارد. [۱۰]. مفهوم سلول‌های بنیادی سرطانی بیان می‌کند که رشد و پیشرفت بسیاری از تومورها توسط کسری کمی‌از سلول‌ها کنترل می‌شود و اکثر سلول‌های موجود در تومور محصولات تمایز غیر طبیعی سلول‌های بنیادی سرطانی هستند [۱۰]. بنابراین، برای توصیف و از بین بردن سلول‌های بدخیم در تومورها، لازم است که بر بخش کوچکی از سلول‌های تومورزا تمرکز کنیم [۳۰]. مفهوم تکامل کلونی بیان می‌کند که تومور از یک سلول طبیعی ژنتیکی بوجود می‌آید که به تعداد زیادی سلول تبدیل می‌شود. در این تکامل، جهش‌های تصادفی به طور مداوم تولید می‌شوند و در نهایت تومور حاصل میلیاردها سلول بدخیم است که حاصل از تجمع تعداد زیادی جهش است[۲۷]. تکامل تومور به عنوان توالی پیدرپی گسترش کلونی توصیف می‌شود، که در آن در هر حالت جدید یک رویداد جهش اضافی ایجاد می‌شود [۱۰].

یکی از توالی‌های پی در پی گسترش کلونی، یک مدل خطی از جانشینی کلونی است، جایی که جهش‌های متوالی پیدرپی باعث ایجاد توالی خطی از مجموعه‌های گسترش کلون می‌شوند و منجر به رشد کلون می‌شوند

[8] Uterine fibroid
[9] Melanocytic nevi
[10] Local
[11] Potentially malignant tumor
[12] Carcinoma In Situ
[13] Metastases
[14] Proliferative
[15] Angiogenic
[16] Spontaneous
[17] Clonal



شکل ۱.۱: دو مدل برای ناهمگونی تومور

[۱۰]. مورد دیگر یک مدل چند کلونی از پیشرفت تومور است، که در آن یک سلول منفرد از طریق مکانیزم تقسیم به چندین زیرکلون گسترش می‌یابد [۳۴]. این مدل بیش از مدل خطی با ناهمگنی تومور مرتبط است. جهش‌های اکتسابی منجر به افزایش بی ثباتی ژنومی با هر نسل متوالی می‌شود [۱۴].

تومرهای ناهمگن[18] که متشکل از چندین کلون هستند، می‌توانند حساسیت‌های مختلفی را نسبت به داروهای سمیت سلولی[19] در نشان دهند. علاوه بر این، می‌زان ناهمگنی تومور می‌تواند خود به عنوان نشانگر زیستی[20] مورد استفاده قرار گیرد زیرا هر چقدر می‌زان ناهمگنی تومور بیشتر باشد، احتمال حضور کلون‌های مقاوم در برابر درمان بیشتر است [۴۶]. دلایل حساسیت‌های مختلف می‌تواند تعاملات بین کلون‌ها باشد که ممکن است اثر درمانی را مهار یا تغییر دهد [۱۰]. تومورهایی با ناهمگنی زیاد، با احتمال بیشتری از کلون‌های گوناگون تشکیل

[18] Heterogenetic
[19] Cytotoxic
[20] Biomarker

شده است که به درمان مقاوم هستند و ممکن است منجر به عدم موفقیت در درمان شوند. روش‌های نوین درمان تومورها با هدف شخصی‌سازی برنامه‌های درمانی از طریق هدف قرار دادن جمعیت‌های سلولی توموری موجود در یک بیمار، توسعه می‌یابند [۲۰]. ناهمگنی‌های توموری یکی از عوامل اصلی مقاومت در برابر دارو است و بنابراین، یک عامل بالقوه در شکست درمان محسوب می‌شود. [۲۰]. تومورها می‌توانند از راه‌های مختلف به طور همزمان به مقاومت دارویی دست یابند، بنابراین هدف قرار دادن فقط یک مکانیسم مقاومت برای غلبه بر نارسایی درمانی، می‌تواند مزیت درمان‌های هدفمند را محدود کند[۱۲]. بنابراین، ناهمگنی تومور می‌تواند برای درک توسعه تومور، پیچیدگی ایجاد کند و توسعه روش‌های موفقیت آمیز را با چالش روبرو کند [۲۰]. مطالعه ناهمگنی تومور می‌تواند منجر به پیشرفت و توسعه روش‌های درمانی شخصی سازی شده شوند و درک ما را از روابط عملکردی بین کلون‌ها در طول درمان افزایش دهند[۱۲]. برای مطالعه ناهمگنی تومور، بسیاری از ابزارهای محاسباتی موثر برای تجزیه و تحلیل اطلاعات کلونی تومور و تاریخچه تکامل آن تولید شده است. این ابزارها با استفاده از داده‌های تغییرپذیری ژنتیکی، تولید شده توسط فناوری‌های توالی یابی نسبتاً دقیق، قادر هستند تا ترکیب‌های کلونی تومور و رابطه اجداد بین کلون‌ها نتیجه دهند. این اطلاعات برای درک پیشرفت تومور و کمک به پیشرفت‌های درمانی کارآمد مهم است.

در ادامه مفاهیم حوزه تحقیق مثل مدل‌های ناهمگنی توموری، روش‌های مختلف توالی‌یابی، روش‌های مختلف ساخت درخت فیلوژنی تومور، مباحث مرتبط به یادگیری عمیق و یادگیری تقویتی به اختصار توضیح داده شد. در فصل سوم تحقیق پیشرو، به بررسی الگوریتم‌هایی که با استفاده از داده‌های توالی‌یابی تکسولی، درخت فیلوژنی تومور را استنباط کرده‌اند پرداخته شد. هر یک از این روش‌ها برای ساخت درخت فیلوژنی به همراه دادگان مورد استفاده، مورد ارزیابی قرار گرفت و در انت‌های فصل سوم مقایس‌های بین روش‌های مختلف صورت گرفت. در فصل چهارم روش پیشنهادی استنباط درخت فیلوژنی بر مبنای یادگیری تقویتی و داده‌های توالی‌یابی تکسولی به تفصیل بیان شده و در فصل پایانی نتایج بدست آمده و مقایسه آن با نتایج پیشین، گزارش شده است. در پایان موضوعات پیشنهادی که در کارهای آتی در راستای ادامه این پژوهش می‌تواند مورد بررسی قرار گیرند، توضیح داده شد.

فصل ۲

مبانی تحقیق

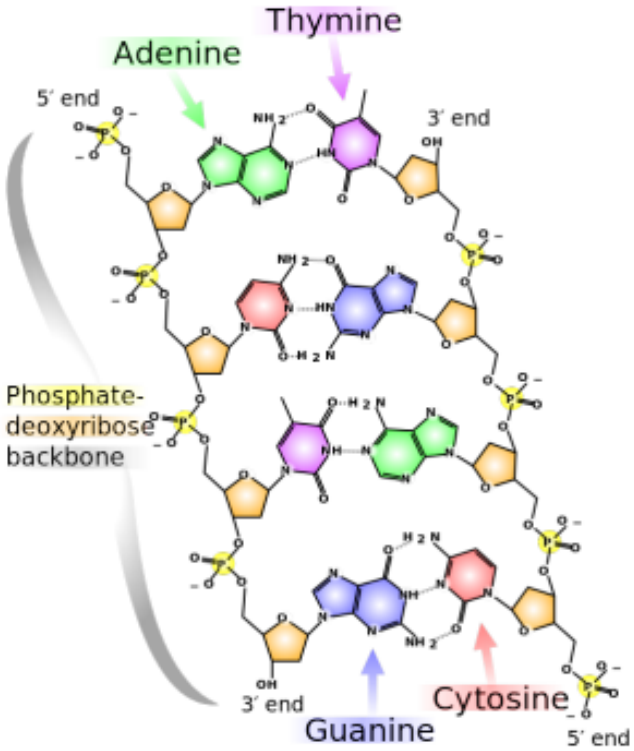
در این فصل ابتدا مفاهیم مورد نیاز جهت تعریف مسئله مانند مدل‌های ناهمگنی تومور، روش‌های یافتن درخت تکاملی تومور، روش‌های توالی‌یابی داده مورد بررسی قرار می‌گیرند. در ادامه مدل‌های مورد استفاده برای استنباط درخت تکاملی تومور معرفی می‌شوند. در پایان مفاهیم مرتبط با یادگیری ماشینی، یادگیری عمیق و یادگیری تقویتی به منظور استنباط درخت تکاملی تومور با رویکرد مبتنی بر داده[۱] توضیح داده می‌شوند.

۱.۲ تنوع ژنتیکی

دی‌ان‌ای[۲] یک مولکول بیولوژیکی است که توسط نوکلئوتیدها[۳] پلیمری شده است. در دی‌ان‌ای چهار نوع نوکلئوتید وجود دارد: آدنین[۴] ،(A) تیمین[۵] ،(T) سیتوزین[۶] (C) و گوانین[۷] (G). دی‌ان‌ای اساس توالی اسیدهای آمینه است که پروتئین را تشکیل می‌دهد. یک مولکول دی‌ان‌ای از دو رشته تشکیل شده است. که در موازات[۸] هم و درجهت‌های مخالف قرار دارند و ساختاری از مارپیچ دوتایی ایجاد می‌کنند. هر نوع نوکلئوتید روی یک رشته با نوع دیگری از نوکلئوتید در رشته دیگر مرتبط است: A با T ؛ C با G (شکل ۱.۲) [۶]. این به عنوان قانون پایه

[1]Data driven
[2]DNA
[3]Nucleotid
[4]Adenine
[5]Thymine
[6]Cytosine
[7]Guanine
[8]Antiparallel

جفت شدن نوکلئوتیدها در هر رشته از دی‌ان‌ای شناخته می‌شود.

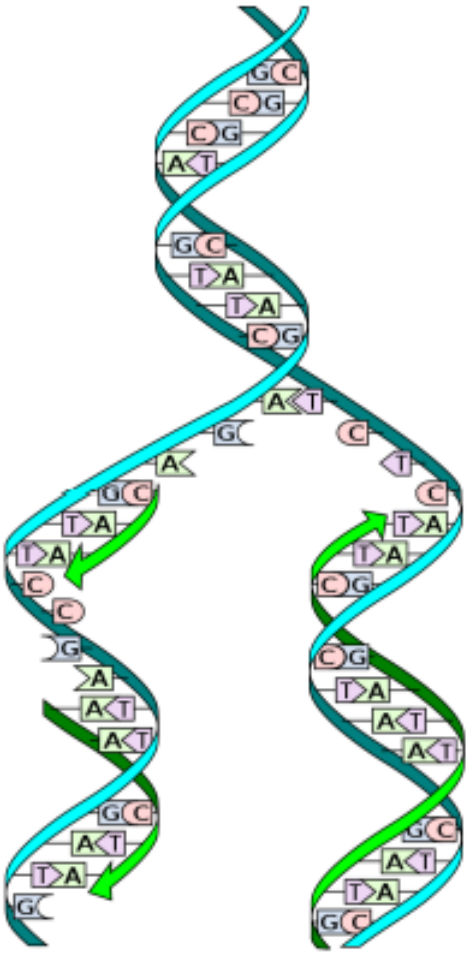


شکل ۱.۲: مارپیچ دوگانه دی‌ان‌ای

همانند سازی دی‌ان‌ای فرآیند تولید دو مولکول دی‌ان‌ای یکسان از مولکول دی‌ان‌ای اصلی است. وقتی تکثیر شروع می‌شود، دو رشته یک مولکول دی‌ان‌ای از یکدیگر جدا می‌شوند و هر رشته به عنوان الگویی برای ساخت نمونه مشابه خود عمل می‌کند. نوکلئوتیدها در هر موقعیت از یک رشته با نوع دیگری از نوکلئوتید مبتنی بر قانون پایه جفت شدن، به منظور سنتز همتای این رشته، متصل می‌شود. پس از همانند سازی، مولکول دی‌ان‌ای اصلی به دو مولکول یکسان تبدیل می‌شود (شکل ۲.۲) [۶].

ژن ناحیه‌ای از دی‌ان‌ای است و به عنوان مولکول واحد وراثت شناخته می‌شود. ژن‌های متعددی در ساختار دی‌ان‌ای با عملکردهای متفاوت وجود دارد. جهش به تغییر دائمی‌توالی هسته‌ای ژنوم اتلاق می‌شود. جهش‌ها می‌توانند در حین فرآیند تکثیر دی‌ان‌ای و با جفت‌گیری اشتباه در قسمت‌های مختلف دی‌ان‌ای ایجاد می‌شود. انواع مختلفی از جهش‌ها مانند جهش تک نوکلئوتیدی[۹](جهش نقطه‌ای[۱۰]) (شکل ۳.۲) و تغییرات ساختاری[۱۱]

[9]Single nucleotide mutation
[10]Point mutation
[11]Single variant

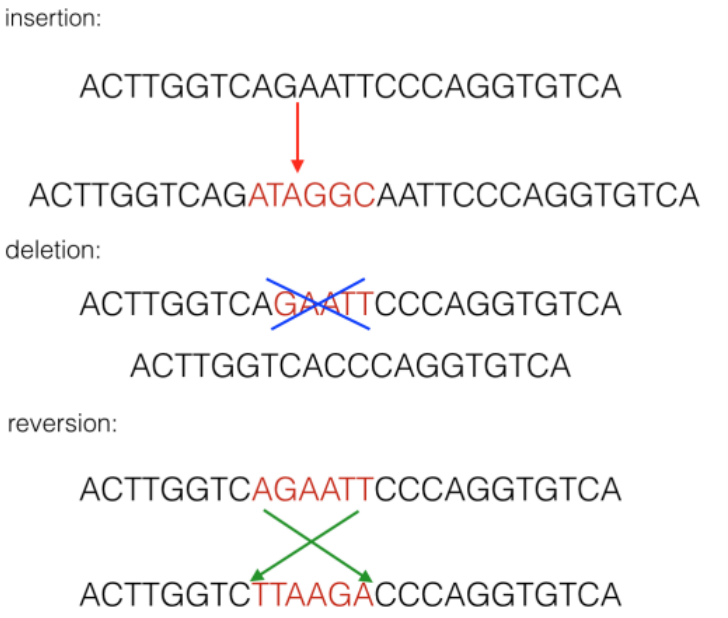


شکل ۲.۲: همانندسازی دی‌ان‌ای

شامل درج[۱۲] ، حذف[۱۳] و برگشت[۱۴] (شکل ۴.۲) وجود دارد. جهش‌های سلولی می‌توانند به بنا بر دلایلی چون مواد شیمیایی، سمیت یا ویروس ایجاد شوند. جهش در یک ژن می‌تواند محصولات آن را تغییر دهد (مانند ایجاد پروتئین متفاوت) یا از عملکرد صحیح ژن جلوگیری کند [۶].

[12] Insertion
[13] Deletion
[14] reversion

شکل ۳.۲: جهش تک‌نوکلئوتیدی



شکل ۴.۲: تغییرات ساختاری

۲.۲ تکامل تومور

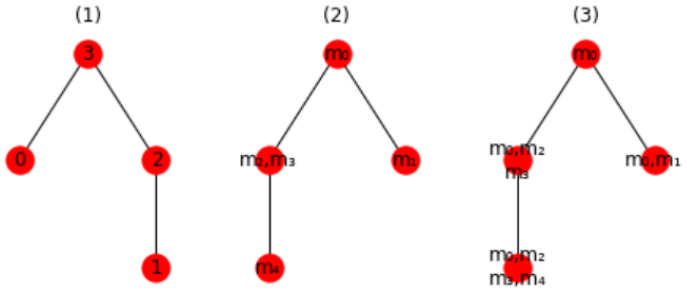
جهشی که در هر سلول از بدن اتفاق می‌افتد، به استثنای سلول‌های جنسی (اسپرم و تخمک)، جهش جسمی[15] نامیده می‌شود [۱]. تجمع جهش بدنی در طول زندگی یک فرد می‌تواند منجر به رشد کنترل نشده مجموعه‌ای از سلول(تومور) شود [۳۶] و می‌تواند باعث شکل‌گیری سرطان یا بیماری‌های دیگر شود [۱]. بدلیل تجمع سلول‌های گوناگون، بیش از یک نوع سلول در تومور وجود خواهد داشت. به گروه‌های سلول با مجموعه‌ای از جهش مشخص، کلون یا جمعیت سلولی تومور گفته می‌شود. کلون‌های موجود در تومور از نظر فیلوژنتیک با هم مرتبط هستند و رابطه آنها را می‌توان با یک درخت فیلوژنتیک نشان داد [۱۰]. درخت فیلوژنتیک رابطه تکاملی بین

[15] somatic

کلون و ترتیب وقوع هر جهش را نشان می‌دهد. به عنوان مثال، شکل۵.۲ :

- یک درخت فیلوژنتیک از یک تومور با چهار کلون با برچسب ۰ تا ۳ را نشان می‌دهد.
- جهش جدیدی را نشان می‌دهد که در هر کلون در طول تکامل این تومور رخ داده است.

همچنین هر کلون جهشی را در مسیر از کلون بالایی به سمت خود به ارث می‌برد. به عنوان مثال، کلون ۰ جهش‌های ،m۰ m۱ دارد. کلون ۱ دارای جهش ،m۰ ،m۲ ،m۳ m۴ است.



شکل ۵.۲: درخت فیلوژنیک تومور

۳.۲ تکنولوژی‌های توالی‌یابی و فراوانی تغییرات آلل

تعیین توالی دی‌ان‌ای روشی برای تشخیص ترتیب دقیق نوکلئوتیدها در یک رشته دی‌ان‌ای است. روش توالی‌یابی نسل بعدی[۱۶] از تعدادی فناوری مدرن توالی تشکیل شده است که امکان تعیین هزینه و زمان توالی‌یابی را به طور موثر فراهم می‌کند. با استفاده از نمونه بیولوژیکی به عنوان ورودی این تکنولوژی‌ها، توالی‌های کوتاه نوکلئوتیدی تولید می‌شود (که به آن خوانش[۱۷] گفته می‌شود). سپس خوانش با استفاده از الگوریتم هم‌ترازی[۱۸] متنوعی مانند الگوریتم تبدیل Burrows-Wheeler با ژنوم مرجع تراز می‌شوند. پس از ترازبندی، می‌توان با جمع‌آوری خوانش‌های همپوشانی[۱۹]، توالی اجماعی[۲۰] ایجاد کرد (شکل ۶.۲). در موقعیتی از توالی اجماع به دلیل همپوشانی خوانش ها، ممکن است بیش از یک نوع خوانش از نوکلئوتید تراز شده وجود داشته باشد (تعداد

[16] Next generation sequencing
[17] Read
[18] Alignment
[19] Overlapping read
[20] Consensus

کل قرائت مرتبط با یک نوع جهش، را پوشش خوانش[۲۱] نامیده می‌شود). نوکلئوتید موجود در این موقعیت به عنوان رایج‌ترین نوکلئوتید تراز شده، مشخص می‌شود. به عنوان مثال، در شکل ۶.۲، سه آدنین ،(A) یک گوانین (G) و یک تیمین (T) در موقعیت سوم توالی اجماع تراز می‌شوند، سپس نوکلئوتید در آن موقعیت به عنوان آدنین (A) تعیین می‌شود. پس از ایجاد توالی اجماع، نوکلئوتیدهای موجود در آن توالی، که متفاوت از ژنوم مرجع هستند، شناسایی شده و به عنوان تغییرات بدنی تک نوکلئوتیدی[۲۲] شناخته می‌شود. با استفاده از نمونه‌های متعدد استخراج شده از یک نمونه تومور، ما می‌توانیم تغییرات بدنی تک نوکلئوتیدی را در هر نمونه با فناوری تعیین توالی‌یابی تشخیص دهیم. نسبت تعداد سلول‌های موجود در یک نمونه حاوی تغییرات بدنی تک نوکلئوتیدی به کل سلول‌ها، فراوانی تغییرات آلل یک تغییر بدنی تک نوکلئوتیدی در این نمونه نامیده می‌شود. مقادیر فراوانی تغییرات آلل برای هر تغییر بدنی تک نوکلئوتیدی در هر نمونه تومور قابل محاسبه است. ابزارهای زیادی برای بازسازی درخت فیلوژنتیک تومور از مقادیر فراوانی تغییرات آلل تومور به عنوان ورودی الگوریتم استفاده می‌کنند.

شکل ۶.۲: تشخیص تغییر بدنی تک‌نوکلئوتیدی از طریق خوانش هم‌ترازی

۴.۲ ناهمگنی ژنومی تومور

سرطان بیماری‌ای است که بدلیل ایجاد ناهنجاری‌های اساسی در فرآیندهای بنیادی سلول مانند تکثیر[۲۳]، تمایز[۲۴] و مرگ[۲۵] سلول ایجاد می‌شود [۲۹]. این ناهنجاری منجر به رشد کنترل نشده تومور و به‌کارگیری بافت غیرسرطانی برای حمایت از این رشد می‌شود. علت اصلی این تغییرات جهش است. جهش یک اصطلاح گسترده است که چندین دسته از تغییرات ژنتیکی را پوشش می‌دهد. هنگام حاملگی، یک جنین دارای یک ژنوم خاص

[21] Read coverage
[22] Somatic single nucleotide variation
[23] Replication
[24] Differentiation
[25] Death

و منحصر به فرد است. این ژنوم که به ژنوم جوانه‌زنی[۲۶] معروف است، می‌تواند با ژنوم انسانی مرجع مقایسه شود. ژنوم انسانی مرجع یک نمونه از ژنوم انسان است و از دی‌ان‌ای چند نفر تشکیل شده است. تفاوت بین ژنوم جوانه‌زنی و ژنوم مرجع به عنوان جهش ژنوم جوانه زنی شناخته می‌شود. جهش‌های جوانه زنی می‌توانند مسئول افزایش خطر ابتلا به سرطان باشند [۴۱]، اما بندرت خود مسئول مستقیم توسعه تومور هستند.

معمولاً تومورها در اثر جهش‌های اکتساب شده پس از لقاح، که معروف به جهش‌های بدنی هستند، ایجاد می‌شوند. جهش‌های بدنی نتیجه اشتباهات در تکثیر دی‌ان‌ای [۹]، قرار گرفتن در معرض جهش‌های با منشأ داخلی یا خارجی یا واردشدن توالی‌های دی‌ان‌ای با منشأ بیرونی بدلیل قرار گرفتن در معرض ویروس است [۴۵]. غالباً در سرطان، جهش‌های بدنی باعث ایجاد اختلال در روند تکثیر دی‌ان‌ای یا ترمیم آن می‌شوند و حتی جهش‌های بدنی بیشتری ایجاد می‌کنند [۴۲]. نظریه کلونی بودن سرطان [۳۶] سرطان را به عنوان یک تک سلولی با منشأ غیرجنسی در نظر می‌گیرد که در اثر تولید مثل فراوان، یک توده متشکل از کلون‌های سلولی گوناگون را ایجاد می‌کند. در این مدل سلولهای توموری با یکدیگر در رقابت هستند و جهش‌های بدنی که مزیت رشد را ایجاد می‌کنند در جمعیت سلول‌های توموری از نسبت بیشتری برخوردار خواهند بود. جهش‌های بدنی که باعث رشد تومور شده و از سلولی به سلولی دیگر منتقل می‌شوند به عنوان جهش‌های راننده[۲۷] شناخته می‌شوند. اولین سلولی که دارای جهش راننده بوده و آن را به جهش‌های بعدی منتقل می‌کند به عنوان سلول بنیانگذار شناخته می‌شود. همه فرزندان این سلول بنیانگذار، جهش راننده و هر جهش دیگری را که سلول بنیانگذار قبل از به دست آوردن جهش راننده بدست آورده است، دارند. این جهش‌های دیگر، که مزیتی برای رشد و گسترش تنوع توموری ندارند، به عنوان جهش‌های مسافر[۲۸] شناخته می‌شوند. شایان ذکر است که تعریف جهش راننده و مسافر به زمینه ژنتیکی و محیطی بستگی دارد. به عنوان مثال، شیمی درمانی داروهای سمیت سلولی (سیتوتوکسیک) می‌تواند باعث تغییر جهش از مسافر به جهش راننده شود و عامل اصلی مقاومت در برابر درمان باشد. همچنین جهش ها را می‌توان بر اساس نوع تغییری که در دی‌ان‌ای ایجاد می‌شود، به طبقات متمایز تقسیم کرد. حذف و تغییر تک‌نوکلئوتیددها[۲۹] جهش‌هایی هستند که یک پایه در ژنوم را به پایه دیگری تغییر می‌دهند. ایندل[۳۰] درج یا حذف یک بخش دی‌ان‌ای است که می‌تواند کوتاه یا طولانی باشد. از ایندل کوتاه و تغییرات تک نوکلئوتیدی در مجموع به عنوان جهش‌های ساده بدنی[۳۱] یاد می‌شود. در همه قسمت‌های یک ژنوم، از جمله کل کروموزوم‌ها، قابلیت حذف یا کپی شدن قسمتی از ژنوم وجود دارد. تغییرات شماره کپی به جهشی اتلاق می‌شود که منجر به حذف یا کپی شدن قسمتی از ژنوم می‌شود. تغییرات شماره کپی[۳۲] نوعی تغییر ساختاری هستند که شامل وارونگی (وقتی

[26] Germline genome
[27] Driver mutation
[28] Passenger mutation
[29] Single nucleotide variants (SNV)
[30] Indel
[31] Single Somatic Mutation
[32] Copy number alteration

قسمت بزرگی از ژنوم معکوس شده باشد) و انتقال متعادل (جایی که دو بخش ژنومی مکان‌های خود را با یکدیگر تعویض می‌کنند) می‌باشند[۴۲]. این گونه‌های مختلف جهش مستقل از یکدیگر نیستند و می‌توانند در رابطه با یکدیگر اتفاق بیفتند (به عنوان مثال یک جهش می‌تواند منجر به تقویت یک وارونگی شود).

تکنیک توالی‌یابی نسل بعدی این امکان را فراهم کرده است تا با صرف هزینه بسیار کم و با استفاده از یک نمونه توموری، توالی‌یابی از دی‌ان‌ای صورت پذیرد و همین امر منجر به تحول گسترده‌ای در زمینه مطالعه تکامل تومور شده زیر امکان نمونه-برداری در تعداد بسیار بالا را از تومور فراهم می‌کند. نمونه‌گیری در حجم بالا این امکان را فراهم آورده است تا ناهمگنی تومور از نقطه منظر ژنتیکی مورد بررسی قرار گیرد و پاسخ به درمان بیماران سرطانی با جزییات بیشتری مورد ارزیابی قرار گیرد.

تقریباً همه نمونه‌های استخراج شده از تومور ترکیبی از سلول ها با ژنوتیپ‌های مختلف را شامل می‌شود. یک نمونه توموری به ندرت فقط شامل بافت سرطانی است زیرا شامل سلول‌های غیر سرطانی از استرومای اطراف[۳۳] یا سلول‌های ایمنی نفوذی[۳۴] است. مطالعات ژنومیک نشان داده است که حتی در میان سلولهای سرطانی، غالباً زیرجمعیت‌های متعدد سرطانی نیز وجود دارد. به عنوان مثال، در یک مطالعه مهم در سال ۲۰۱۲، گرلینگر و همکارانش [۲۵] توالی‌یابی ژنوم و تغییرات شماره کپی را از طریق نمونه‌های مکانی مجزا استخراج شده از سرطان کلیه اولیه و نقاط متاستاز ثانویه بدست آورده‌اند. با بررسی این نمونه‌های متعدد، مشخص شد که یک ناهمگنی ژنتیکی قابل توجهی در تومور وجود دارد. تعداد بسیار زیادی از جهش‌های شناسایی شده در همه سلول‌های توموری مشاهده نشدند و این بدان معناست که این جهش‌ها بیش از آن‌که یک ناحیه کلونی باشند، به صورت یک ناحیه زیر کلونی بوده‌اند. با استفاده از روش‌های پردازش غیراتوماتیک، تغییرات تک نوکلئوتیدی ها و تغییرات شماره کپی بر اساس نمونه‌هایی که از آن استخراج شده‌اند، به خوشه‌های مجزا دسته‌بندی شده و یک درخت فیلوژنی به آن‌ها نسبت داده شد. بازسازی درخت فیلوژنیک تومور این امکان را فراهم آورد تا سیر تکاملی تومور با استفاده از شاخه‌های مختلف درخت فیلوژنی شامل جهش‌هایی با عملکرد یکسان از سه ژن متفاوت مورد بررسی قرار گیرد.

در همان سال، یک مطالعه مهم دیگر، "تاریخچه زندگی ۲۱ سرطان پستان" [۳۵]، حضور ITH را نیز نشان داد. در این مطالعه آنها توالی‌یابی کامل ژنوم را در عمق متوسط 188X بر روی تومور پستان PD4120a انجام دادند. این عمق اجازه می‌دهد تا جمعیت‌های شیوع تا ۵٪ کم باشد. آنها مشاهده کردند که تغییرات تک نوکلئوتیدی‌ها در تعداد کمی از خوشه‌های مجزا مشاهده می‌شوند که با توجه به کسر نوع آلل (VAF) آنها مشاهده می‌شود، نسبت خواندن ها در یک مکان متفاوت شامل آلل نوع. علاوه بر این، آنها توانستند نشان دهند که برخی از این خوشه‌های مجزا را نمی‌توان با جهش‌های موجود در تمام جمعیت‌های سرطانی توضیح داد، که این نشان دهنده

[33] Surrounding stroma
[34] Infiltrating immune cell

حضور تغییرات تک نوکلئوتیدی‌های تحت کلونال است. در همان زمان، آنها دریافتند که بسیاری از جهش‌ها در تمام سلول‌های سرطانی موجود در نمونه وجود دارد، که نشان می‌دهد جد مشترک اخیر نسبتاً دیر در زمان تکامل رشد کرده است. مشاهده اینکه جهش‌های زیر کلونال به جای توزیع یکنواخت یا مطابق قانون قدرت در خوشه‌های متمایز پیدا شده است، شواهدی را نشان می‌دهد که این جهش‌های زیرکلونالی بیش از آنکه ناشی از تکامل خنثی یا مصنوعات فنی باشد، در زیرمجموعه‌های متمایز ناشی از فشارهای انتخابی یافت می‌شود. نویسندگان همچنین با تأیید اینکه جهش‌های زیر کلونال محدود به تغییرات تک نوکلئوتیدی نیستند، توانستند حضور تغییرات شماره کپی‌های کلونال و زیرکلونال را تأیید کنند. نویسندگان یک الگوریتم خوشه‌بندی غیر پارامتریک (یک مدل مخلوط فرآیند دیریشله (DPMM)) را با استدلال قابل توجه دستی برای استنباط فیلوژنی شاخه‌ای از چهار زیر جمعیت سرطانی در آن نمونه منفرد تومور ترکیب کردند. درک معماری ژنتیکی این زیرجمعیت‌ها می‌تواند به مطالعه زیست شناسی سرطان کمک کند و نشان داده شده است که در پیش‌بینی بقا در بسیاری از انواع سرطان مفید است [۸]. به عنوان مثال، زیرجمعیت‌های مختلف، که توسط مجموعه جهش‌های جسمی حمل شده تعریف می‌شوند، توانایی‌های مختلفی در مقاومت در برابر درمان و متاستاز دارند. برای انجام این کار، باید از یک یا تعداد کمی از نمونه‌های تومور فله، ژنوتیپ‌های موجود در نمونه را شناسایی کرد. این مسئله، تحت عنوان بازسازی ساب کلونال، موضوع اصلی این پایان‌نامه است. مطالعات پیشگام که نشان داد ITH برای انجام این بازسازی به استدلال دستی قابل توجهی نیاز دارد. استدلال دستی کند، مستعد خطا است و به تخصص قابل توجهی نیاز دارد. مزایای بازسازی کاملاً خودکار بدیهی است. این بخش پیش زمینه مشکل بازسازی زیر کلونال، چگونگی پرداختن به آن برای انواع مختلف جهش، خصوصیات اصلی الگوریتم‌های بازسازی زیر کلونال و خلاصه‌ای از کارهای موجود در این زمینه را توصیف می‌کند.

۵.۲ بازسازی زیر کلونال

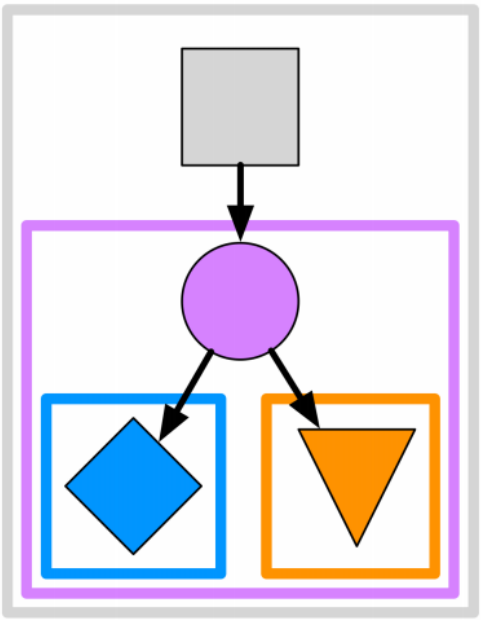
بازسازی ساب کلونال سعی دارد ژنوتیپ‌های موجود در تومور را از تعداد کمی از نمونه‌های توالی دی‌ان‌ای از آن تومور استنباط کند. تعداد ژنوتیپ‌های موجود در تومور از قبل مشخص نیست. این ژنوتیپ‌های زیر کلونال به طور معمول با جهش‌هایی که در مقایسه با ژنوم خط جوانه‌ای دارند، توصیف می‌شوند. ژنوم جوانه‌زنی علاوه بر نمونه(های) تومور، با تعیین توالی یک نمونه غیرسرطانی تعیین می‌شود. در حال حاضر در هنگام تعریف این جمعیت از دو نوع جهش به طور معمول استفاده می‌شود: جهش‌های ساده بدنی‌های متشکل از تعویض‌ها و درج / حذف کوچک (ایندل) و CNA حاصل از تغییرات ساختاری بزرگتر. مشاهده انواع جهش‌های دیگر، مانند مجموعه گسترده‌ای از SVها که شامل بازآرایی هستند، مشاهده آنها دشوارتر است و روش‌های شناسایی آنها در

مراحل اولیه رشد است.

به طور متوسط، حتی در شرایط ایده آل، هر سلول در هر بخش یک جهش پیدا می‌کند [۹]، به همین ترتیب، بیشتر سلول‌های تومور ژنوتیپ منحصر به فردی خواهند داشت. بنابراین، به طور دقیق، اکثر سلولهای تومور می‌توانند به طور بالقوه نمایانگر زیرجمعیت منحصر به فرد خود باشند. با این حال، به طور عملی، جهش‌هایی که مختص سلول‌های منفرد است یا فقط تعداد کمی از سلول‌ها آنها را به اشتراک می‌گذارد، در حین فراخوانی نوع شناسایی نمی‌شوند. تماس متغیر در بخش ۳.۵.۲ بیشتر مورد بحث قرار گرفته است. بعلاوه، سلول‌هایی که بخش عمده‌ای از جهش‌های خود را به اشتراک می‌گذارند، خصوصاً جهش‌های راننده، صفات مشابهی دارند. به همین ترتیب، من قرارداد گسترده‌ای را اتخاذ کرده و یک زیر جمعیت را به عنوان تمام سلول‌هایی که دارای زیر مجموعه یکسان جهش‌های بدنی در هنگام فراخوانی نوع هستند، تعریف می‌کنم.

یک گام مهم در بازسازی ساب کلونال محاسبه شیوع سلولی تبارهای زیر کلونال و سپس، در نهایت، زیرجمعیت‌های سرطانی است. شیوع سلولی یک زیرجمعیت، نسبت سلول‌های نمونه توالی شده متعلق به آن است. غالباً، شیوع سلولی با تقسیم بر خلوص نمونه، یعنی نسبت سلولهای سرطانی در نمونه، به بخش سلولهای سرطانی، نسبت سلولهای سرطانی، تبدیل می‌شود. هر سلول دقیقاً به یک زیرمجموعه تعلق دارد، بنابراین این شیوع باید در یک جمع باشد. به طور کلی، سلول‌های غیر سرطانی در یک زیرمجموعه واحد قرار می‌گیرند. با این حال، از آنجا که جهش‌ها اغلب در زیرجمعیت‌های متعدد وجود دارند، شیوع سلولی بسیاری از زیرجمعیت‌ها را نمی‌توان مستقیماً از جهش‌های آن استنباط کرد. برای پرداختن به این موضوع، ما یک نسب زیر کلونال برای یک جهش به عنوان مجموعه زیرجمعیت‌هایی که در آن وجود دارد، تعریف می‌کنیم. به طور رسمی، دودمان‌های زیر کلونال از زیر جمعیت بنیانگذار تشکیل می‌شود (جایی که جهش برای اولین بار ظاهر می‌شود) و همه زیرجمعیت‌های بعدی آن (که وراثت جهش) علاوه بر جهش‌های خاص خود، این زیرمجموعه‌های فرزندی حاوی تمام جهش‌های موجود در نژاد تعریف کننده زیر جمعیت هستند (به جز در صورت حذف محل منبع جهش، برای جزئیات بیشتر به فصل ۳ مراجعه کنید). نسب مربوط به یک زیر درخت (یا کلاد) از درخت کلون تومور است. شیوع سلولی یک تبار مجموع شیوع سلولی زیرجمعیت‌هایی است که متعلق به آن تبار هستند. از آنجا که سلول‌ها می‌توانند در چندین نژاد زیرکلونال وجود داشته باشند، شیوع نسب در یک جمع نیست.

شکل ۷.۲ تصویری از یک درخت کلون نمونه را ارائه می‌دهد. گره‌های موجود در درخت، همانطور که در بالا تعریف شد، نشان دهنده زیر جمعیت است. فلش‌ها از جمعیت والدین به سمت فرزندانشان هدایت می‌شوند. دودمانهای زیر کلونال به صورت مستطیل نشان داده می‌شوند و با توجه به زیرمجموعه بنیادی آنها که در ریشه تیغه یافت می‌شوند، رنگی هستند.



شکل ۷.۲: درخت کلون تومور

۶.۲ تغییرات تعداد کپی

بیشتر ژنوم انسان دیپلوئید است، به این معنی که دو نسخه از توالی دی‌ان‌ای ما در سلول‌های ما وجود دارد، یکی از پدر و دیگری از مادر. تغییرات شماره کپی این تغییر را می‌دهند، یا با تغییر در تعداد نسخه‌ها (مثلاً از طریق تکثیر کل ژنوم)، نسبت کپی‌های مادر به پدر (مثلاً از دست دادن خنثی هتروزیگوزیته در تعداد کپی‌ها، جایی که برای همان منطقه یک ژنوم والدین تکثیر می‌شود و دیگری حذف شده است) یا هر دو (به عنوان مثال کپی کروموزوم مادر). بیشتر این تغییرات (به استثنای تکثیر کل ژنوم) دامنه محدودی از ژنوم را تحت تأثیر قرار می‌دهد، اما می‌تواند از تأثیر یک ژن تا یک کروموزوم کامل باشد. این بخش از ژنوم تغییر یافته به عنوان یک بخش شناخته می‌شود.

تغییرات شماره کپی می‌توانند تعداد کپی کل یک بخش و / یا تعداد نسبی نسبی دو کروموزوم والدین را تغییر دهند. هر یک از این تغییرات توسط توالی‌یابی ژنومی هسته قابل تشخیص است. تغییر در تعداد کپی کل یک بخش را می‌توان تشخیص داد زیرا نسبت خواندن آن نقشه به آن بخش بین خط جوانه زنی و نمونه تومور متفاوت خواهد بود. بخش از یک قطعه نسبت ورود خوانده شده است که به یک قطعه در یک نمونه غیر سرطانی ترسیم شده است به نسبت خوانده شده که به یک بخش در یک نمونه سرطانی ترسیم شده است. از نسبت نسبت‌ها برای محاسبه این واقعیت استفاده می‌شود که تعداد کل قرائت‌ها اغلب بین توالی‌یابی سرطانی و غیرسرطانی متفاوت

است، در مناطق مختلف ژنوم عمق خواندن بیشتر یا پایین‌تر ناشی از محتوای GC یا نقشه برداری وجود دارد و تردستی یک تومور با بافت طبیعی متفاوت است. تکرر یک ژنوم، میانگین تعداد کپی از هر کروموزوم است که برای طول کروموزوم نرمال می‌شود.

با تغییر در کسر آلل می‌توان عدم تعادل در تعداد نسخه‌های مادری و پدری این بخش را تشخیص داد. در مناطق دیپلوئید ژنوم‌ها، اگر یک بازه بین کپی‌های مادر و پدر متفاوت باشد، موقعیت هتروزیگوت نامیده می‌شود. جهش‌های تک پایه، خط جوانه زنی همچنین به عنوان چند شکلی تک هسته‌ای نامیده می‌شوند. وقتی یک ژنوم توالی‌یابی شود، حدود نیمی از قرائت آن مکان هتروزیگوت حاوی هر یک از بازها خواهد بود، در نتیجه کسر آلل ۵۰ است. این امر تا زمانی که نسبتی برابر با نسخه‌های مادرانه و پدری وجود داشته باشد، صادق خواهد بود. اگر این نسبت تغییر کند، کسر آلل تمام پولیمورفیسم تک هسته‌ای در بخش آسیب دیده تغییر می‌کند. پولیمورفیسم تک هسته‌ای هتروزیگوت به طور متوسط هر ۱۵۰۰ باز [۱۳] رخ می‌دهد و بنابراین برای بخشهای طولانی بسیاری از پولیمورفیسم تک هسته ایی هتروزیگوت تحت تأثیر قرار می‌گیرند. توزیع کسر آلل S تمام پولیمورفیسم تک هسته‌ای در بخش، حالت دوگانه‌ای پیدا می‌کند که هر حالت نشان دهنده نسبت نسخه‌های آن بخش از هر والد است.

فراخوانی CNA چالش برانگیز است زیرا با مشاهده مستقل هر بخش، مسئله هنوز مشخص نشده است. حتی با فرض اینکه هر بخش فقط توسط یک CNA تحت تأثیر قرار گیرد، CNA موسوم به سه پارامتر (نسبت سلولهای حاوی CNA، تعداد کپی‌های مادر و تعداد کپی‌های پدری) وجود دارد و فقط دو مشاهده برای توضیح وجود دارد (و کسر آلل)

همه روش‌ها با فرض اینکه تعداد کمی از نژادهای زیرکلونال مسئول بیشتر یا تمام تغییرات شماره کپی هستند، این ابهام را برطرف می‌کنند. روشی که توسط الگوریتم باتنبرگ [۳۵] به کار رفته است، به بیشتر تغییرات شماره کپی وابسته به یک نژاد زیر کلونال منفرد و شایع به نام تبار کلونال متکی است. تحت این روش، شیوع این تبار، همراه با تعداد کپی اصلی و جزئی در تمام تغییرات تعداد کپیکلونال، می‌تواند با یک فرآیند دو مرحله‌ای تخمین زده شود. در گام اول، این روش با فرض شیوع نژاد کلون f_c آغاز می‌شود. شیوع تبار کلونال در بیشتر موارد با خلوص نمونه تومور برابر است. با توجه به شیوع کلونال، هر بخش پس از آن فقط دو متغیر برای توضیح دارد (تعداد کپی بزرگ و جزئی). از آنجا که هر بخش دارای دو مشاهدات است، اکنون مسئله هنوز به درستی تعیین نشده است و بهترین کپی اصلی و مینور متناسب است. سپس، ترکیب کلی مقدار Φ_c فرض شده با ترکیب مناسب در تمام بخشها تعیین می‌شود. الگوریتم با بهینه سازی این تناسب بهترین مقدار Φ_c را انتخاب می‌کند. سپس برای هر بخش، شماره کپی اصلی و جزئی با بهینه سازی متناسب بودن قطعه با بهترین مقدار Φ_c انتخاب می‌شود. این روش فرض می‌کند که تمام تغییرات شماره کپی به نژاد کلونال تعلق دارند، که همیشه درست نیست. در مرحله بعدی، بخشهایی که حاوی تغییرات تعداد کپیتحت کلونال هستند با جستجوی بخشهایی با اطلاعات مناسب

ضعیف با استفاده از Φ_c استنباط شده مشخص می‌شوند. در این بخش‌ها، روش به طور همزمان و مستقل از هر بخش دیگر، عدد Φ_i و عدد کپی بزرگ و جزئی را استنباط می‌کند.

از آنجا که سه متغیر وجود دارد و تنها دو مشاهده وجود دارد، راه حل‌های بسیاری با تناسب داده برابر وجود دارد که از نظر زیست شناختی برای این تغییرات تعداد کپی زیر کلونال قابل قبول است. این ابهام با انتخاب راه حلی که نزدیکترین شماره به شماره نسخه طبیعی است برطرف می‌شود، اما تعدادی از موارد متداول وجود دارد که این ابتکار عمل ناموفق است. سپس این روش‌ها انتساب تغییرات تعداد کپی زیرکلونال به دودمان و تمام استنباط‌های فیلوژنتیک را برای روش‌های پایین دست رها می‌کنند.

رویکرد عمده دیگر این است که فرض کنیم همه تغییرات شماره کپی از تعداد کمی تبار ساب کلونال به وجود می‌آیند. الگوریتم‌هایی که از این روش استفاده می‌کنند به طور مشترک شیوع این نژادها و تعداد کپی بزرگ و جزئی را برای هر بخش استنتاج می‌کنند (به عنوان مثال THetA[۴۹، ۴۷] و TITAN). تعداد دودمانهای زیر کلونال معمولاً با استفاده از احتمال جریمه شده‌ای مانند معیار اطلاعات بیزی (BIC) یا انواع BIC تعیین می‌شود (به عنوان مثال THetA از BIC اصلاح شده با پارامتر مقیاس گذاری استفاده می‌کند [۴۹]). بنابراین این روش‌ها هم تغییرات شماره کپیرا فراخوانی می‌کنند و هم آنها را به دودمان‌های زیرکلونال اختصاص می‌دهند. هیچ روش موجود این دودمان‌ها را در یک درخت فیلوژنتیک قرار نمی‌دهد

۷.۲ جهش‌های ساده بدنی

جهش‌های ساده بدنی جهش‌های کوچکی هستند که می‌توانند مستقیماً از طریق توالی‌یابی و نسبت کروموزوم‌های موجود در نمونه حاوی آنها از تعداد قرائت‌های حاوی جهش و تعداد کل خوانده‌ها در آن مکان، مشاهده شوند. نسبت قرائت حاوی جهش به کل قرائت به عنوان VAF جهش شناخته می‌شود. جهش‌های ساده بدنی‌ها معمولاً با بررسی مشترک ترازها و یک نمونه غیرسرطانی خوانده می‌شوند. این استنباط مشترک برای جداسازی انواع بدنی و ژرمینال مورد نیاز است.

این فرایند به دلیل انواع مختلف خطاها و تعصبات که در داده‌های NGS وجود دارد، دشوار می‌شود[۲۲]. یک مشکل اساسی در تشخیص جهش‌های ساده بدنی این است که به نظر می‌رسد خطاهای توالی جهش‌های ساده بدنی شیوع کمی دارند. به طور خاص، در Illumina Hiseq2000 که به طور گسترده استفاده می‌شود، از هر ۱۰۰۰ پایه یکی از آنها دارای یک خطا است (به طور معمول یک تعویض) [۳۷]. به همین ترتیب، در طول سه میلیارد پایه ژنوم انسانی، یک احتمال غیر قابل اغماض وجود دارد که در بعضی موقعیت‌ها، چندین بار خواندن دقیقاً شامل خطای توالی دقیقاً در همان موقعیت‌ها است. به نظر می‌رسد این خطاها شیوع کم جهش‌های ساده

بدنی دارند. تمایز بین این خطاها و شیوع کم واقعی جهش‌های ساده بدنی‌ها شامل یک معامله بین حساسیت و ویژگی و در حالت ایده آل، یک مدل نویز بسیار دقیق است. حل این مشکل امتداد طبیعی کار گسترده‌ای است که در زمینه فراخوانی جهش‌های جوانه‌زنی انجام شده است و الگوریتم‌های زیادی برای انجام این کار وجود دارد (به عنوان مثال [۲۲، ۱۶])

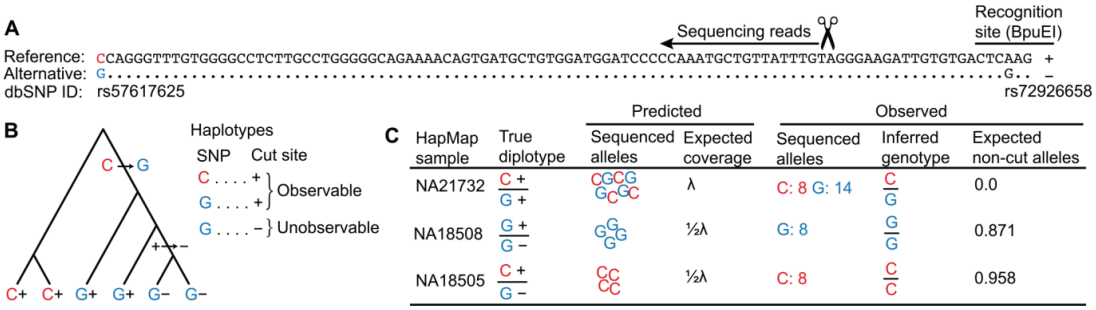
۸.۲ ترک آللی

اگرچه روش‌های تعیین توالی با بازدهی بالا [۳۲] ارزان هستند، اما تحت تاثیر مقدار بایاس هستند و مارکرهای ژنتیکی‌ای تولید می‌کنند که تقریباً به طور تصادفی در کل ژنوم تقسیم می‌شوند. این روشها با موفقیت در نگاشت[۳۵] صفات [۳۶، ۲۶]، ساخت مپ پیوندی [۳۸، ۱۹]، اسکن انتخاب [۱۷، ۴۸]، و برآورد تنوع ژنتیکی [۱۵] استفاده شده است. یکی از این روش‌ها، تعیین ژنوتیپ براساس توالی [۷] (GBS) است. در GBS، هدف توالی‌یابی فقط با اتصال آداپتورهای توالی به محل‌های برش آنزیم محدود کننده، به کمتر از ٪۵ از ژنوم کاهش می‌یابد (شکل زیر). قرائت GBS همچنین می‌تواند به صورت کانکت‌های کوتاه مونتاژ شود، که بدون نیاز به توالی ژنوم فراخوانی یک نوع تغییر تک هسته‌ای (تغییرات تک نوکلئوتیدی) را امکان پذیر می‌کند [۲۸]. از این رو، GBS یک روش محبوب در سیستم‌های غیر مدلی است که به طور معمول فاقد منابعی مانند مجموعه ژنوم و ریزآرایه‌ها است.

بر خلاف توالی‌یابی کل ژنوم (WGS)، GBS مستعد ابتلا به خطاهای مختلف تماس به دلیل محدودیت چندشکلی‌های سایت است (کاهش آللیک). کاهش آللیک در GBS می‌تواند برنامه‌هایی را که به فراخوانی دقیق تغییرات نادر، از جمله تخمین طیف فرکانس سایت در ژنتیک جمعیت متکی هستند، را دچار اختلال کند. یک رویکرد آماری سیستماتیک برای تشخیص کاهش آللیک در داده‌های توالی GBS، اجرا شده و در بسته نرم افزاری منبع باز GBStools وجود دارد. این روش مبتنی بر این واقعیت است که کاهش آللیک متناسب با تعداد آللهای سایت محدود کننده بدون برش که در آنجا حمل می‌کند، میزان خوانش نمونه را در یک سایت خاص کاهش می‌دهد. بنابراین GBStools پوشش هر نمونه را در یک سایت خاص به عنوان یک متغیر تصادفی پواسون مورد استفاده قرار می‌دهد که از توزیع با میانگین λ (آللیک‌های بدون برش صفر)، توزیع با میانگین λ½ (یک آللیک بدون برش)، یا با میانگین صفر (دو آللیک بدون برش). GBStools حداکثر احتمال پارامتر λ را با استفاده از تعداد واقعی آللیک‌های بدون برش در هر نمونه که به عنوان متغیرهای نهفته (مشاهده نشده) در نظر رفته می‌شود و از طریق حداکثر رساندن مقدار چشم انتظاری (EM)، محاسبه می‌کند . از مقادیر مورد انتظار این متغیرهای نهفته می‌توان برای تخمین اینکه کدام نمونه ها یک آللیک بدون برش دارند استفاده کرد. به طور همزمان، GBStools

[35] Mapping

فرکانس سایت آلل‌های SNP مرجع قابل مشاهده و جایگزین، φ_1 و φ_2 ، و آللیک بدون برش، φ_3 ، که در آن $\varphi_1+\varphi_2+\varphi_3=1$ برآورد می‌کند و در نهایت، آزمون نسبت احتمال با مقایسه فرضیه صفر $\varphi_3 = 0$ با فرضیه $\varphi_3 > 0$ جایگزین می‌کند. GBStools در اجرای فعلی خود نمی تواند ژنوتیپ‌های واقعی پنهان شده توسط کاهش آللیک را استنباط کند، اما می‌توان با فیلتر کردن سایت‌هایی که نسبت احتمال آنها زیاد است خطاها را حذف کند.



شکل ۸.۲: نمایی از تطابق ژنتیکی

در شکل بالا، آلل BpuEI بدون برش ناشی از SNP rs72926658 با برچسب "–" و آلل برش با "+" برچسب گذاری شده است. آلل "–" در هاپلوتیپ با آلل G مشتق شده بوجود آمده و باعث شده تا برخی از آلل‌های G توسط GBS قابل مشاهده نباشند. نمونه‌های نشان داده شده دارای سه دیپلوتیپ هتروزیگوت است. نتایج توالی با پیش‌بینی‌ها مطابقت داشت و نمونه NA18505 به اشتباه هموزیگوت نامیده می‌شد، اما انتظار می‌رود تعداد آلل‌های کاهشی محاسبه‌شده توسط GBStools (0.958) با تعداد واقعی (۱) مطابقت داشته باشد، و آن را به عنوان یک تماس اشتباه احتمالی مشخص کند.

۹.۲ مقدمه‌ای بر مدل‌سازی احتمالی

وظیفه اصلی یادگیری ماشین، یادگیری از داده‌ها است، کاری که به عنوان استنباط شناخته می‌شود. برای یادگیری از داده‌ها، باید فرضیاتی را مطرح کرد. توصیف رسمی فرضیات صورت گرفته به عنوان یک مدل ذکر می‌شود. یک مدل احتمالی مفروضات ارائه شده را تعریف می‌کند که اطلاعات آموخته شده را با استفاده از متغیرهای تصادفی و توزیع‌های احتمال به داده‌های مشاهده شده پیوند می‌دهد. توزیع‌های احتمال توابع ریاضی هستند که یک رویداد را ورودی می‌کنند و احتمال آن واقعه را بیرون می‌آورند. توزیع احتمال می‌تواند تابعی بیش از واقعه باشد و این متغیرهای اضافی به عنوان پارامترهای توزیع شناخته می‌شوند[۲۸]. رویکرد بیزی در یادگیری ماشین شامل استنباط احتمالی مقادیر پارامترهای منوط به مشاهدات است [۲۹]. چهار مولفه دارد:

- احتمال: احتمال مشاهده داده‌ها است، مشروط به تنظیم پارامتر P(data | parameters)
- پارامترهای احتمال
- پارامترهای قبلی
- داده‌های مشاهده شده

پارامترها خود مجموعه‌ای از متغیرهای تصادفی هستند که از توزیع قبلی (P (پارامترها)) گرفته شده اند، که باورهای ما را در مورد احتمال حالت‌های مختلف پارامتر در غیاب مشاهده مشاهده می‌کند. این اصطلاحات با استفاده از قانون بیز با هم ترکیب می‌شوند:

- P(parameters|data) = P(data|parameters) $*$ P(parameters) / P(data)
- Posterior $\propto$ likelihood $*$ prior

پس زمینه توزیع پارامترهای مشروط به مشاهده داده‌ها است و خروجی اصلی استنتاج بیزی است. از توزیع پسین می‌توان برای انجام کارهایی مانند پیش‌بینی مشاهدات آینده استفاده کرد.

۱.۹.۲ زنجیره مارکوف مونت کارلو

برای انجام استنتاج بیزی[۳۶]، ما اغلب می‌خواهیم در توزیع پسین ادغام شده، پیش‌بینی کنیم یا خلاصه‌هایی پیدا کنیم، به عنوان مثال میانگین پارامتر پسین. به طور کلی، انجام چنین ادغامی (جمع بندی در مورد متغیرهای گسسته) از نظر تحلیلی غیرقابل حل است. با این حال، می‌توان چنین ادغام‌هایی را با استفاده از نمونه‌هایی که از قسمت پسین ترسیم شده‌اند تقریبی داد:

$$E[f] = \int f(x)p(x)dx \approx ۱/N \sum_{۱..N} f\left(x_i\right) \tag{۱.۲}$$

که در آن x_i نمونه i از $p(x)$ و $p(x)$ و $f(x)$ به ترتیب توزیع و عملکرد مورد نظر ما است. به ندرت می‌توان مستقیماً از توزیع پسین نمونه برداری کرد. برای تولید موثر نمونه‌ها از توزیع، حتی در ابعاد بالا، می‌توان از تکنیک زنجیره ماکوف مونت کارلو استفاده کرد. زنجیره ماکوف مونت کارلو یک زنجیره مارکوف می‌سازد که در آن توزیع

[36] Bayesian

تعادل توزیع پسین است. سپس مقادیر زنجیره می‌تواند به عنوان نمونه از پسین با توجه به همگرایی کافی به توزیع تعادل مورد استفاده قرار گیرد. برای انجام زنجیره ماکوف مونت کارلو، تاز زمانی که بتوان $p \propto p(x)$ را محاسبه کرد، نیازی به محاسبه $p(x)$ نیست. این زنجیره ماکوف مونت کارلو را قادر می‌سازد تا از محاسبه ثابت‌های نرمال سازی، که اغلب غیرقابل حل هستند، خودداری کند. یک زنجیره مارکوف به عنوان یک سری متغیرهای تصادفی تعریف می‌شود که دارای ویژگی استقلال شرطی زیر هستند:

$$p\left(z^{N+1} \mid z^{1}..z^{N}\right) = p\left(z^{N+1} \mid z^{N}\right) \tag{۲.۲}$$

نمونه‌ای از الگوریتم زنجیره ماکوف مونت کارلو الگوریتم Metropolis-Hastings (MH) است [۳۱]. الگوریتم MH از حالت دلخواه Z^t شروع می‌شود. سپس یک حالت پیشنهادی z از توزیع پروپوزال $q(z|z^t)$ ترسیم می‌شود. این حالت پیشنهادی z با احتمال زیر پذیرفته می‌شود:

$$\min\left(1, \hat{p}\left(z^{*}\right) q\left(z^{t} \mid z^{*}\right) / \hat{p}\left(z^{t}\right) q\left(z^{*} \mid z^{t}\right)\right) \tag{۳.۲}$$

می توان نشان داد که الگوریتم MH تعادل دقیق را برآورده می‌کند و از این رو، $p(x)$ توزیع تعادل است [۱۱]. در حالی که توازن دقیق برای اثبات اینکه در محدوده نمونه‌های بی‌نهایت زنجیره به توزیع مورد نظر همگراست کافی است، اما در عمل فقط تعداد محدودی از نمونه‌ها را می‌توان ترسیم کرد. واضح است که نمونه‌های ابتدای زنجیره، که از یک مکان دلخواه در فضای حالت شروع می‌شوند، بعید است از توزیع تعادل باشد. این نمونه‌ها به عنوان نمونه‌های سوختنی کنار گذاشته می‌شوند. هرچه همگرایی زنجیره مارکوف سریعتر باشد، نمونه‌های کمتری باید کنار گذاشته شوند و می‌توان از تعداد بیشتری برای محاسبه انتظارات استفاده کرد. با بررسی اثری از مقادیر مهم پارامتر یا احتمال همگرایی می‌توان نظارت کرد، اما این امر ممکن است چند حالت را از دست بدهد. متأسفانه دانستن اینکه آیا همگرایی حاصل شده است غیرممکن است، فقط گاهی اوقات می‌توان همگرایی را رد کرد [۲۴]. گذشته از همگرایی، یکی دیگر از خصوصیات اصلی یک زنجیره مارکوف میزان اختلاط زنجیره است. با توجه به n نمونه مستقل از توزیع، واریانس میانگین پارامتر برآورد σ_n است که σ انحراف استاندارد توزیع خلفی پارامتر است. نمونه‌های گرفته شده از زنجیره مارکوف مستقل نیستند، زیرا به وضعیت فعلی زنجیره بستگی دارند (یعنی فقط از نظر شرطی مستقل هستند). برای تخمین اندازه نمونه موثر یک زنجیره مارکوف، یعنی تعداد نمونه‌های مستقل با همان خطای استاندارد همان زنجیره، می‌توان از معادله زیر استفاده کرد:

$$ESS = \frac{n}{1 + 2\sum_{0}^{\infty} \rho_j} \tag{۴.۲}$$

حاصل جمع بی نهایت محاسبه ESS را می‌توان با استفاده از برآوردگر پریودوگرام کوتاه تطبیقی Sokal [۳۹] تخمین زد.

۱۰.۲ یادگیری ماشین و یادگیری تقویتی

آنالیز داده‌های بالینی یک حوزه مهم تحقیقاتی در انفورماتیک، علوم کامپیوتر و پزشکی است که توسط محققان شاغل در دانشگاه‌ها، صنعت و مراکز بالینی انجام می‌شود. یکی از بزرگ‌ترین چالش‌ها در تجزیه و تحلیل داده‌های پزشکی، استخراج و تجزیه و تحلیل داده‌ها از تصاویر است. در چند سال اخیر روش‌های یادگیری ماشین انقلابی بزرگ در بینایی کامپیوتر[۳۷] به وجود آورده است که راه‌حل‌های جدید و کارآمدی را درمورد خیلی از مسائل و مشکلات موجود در آنالیز تصاویر که مدت زمان طولانی است حل نشده باقی مانده‌اند معرفی می‌کنند. برای اینکه این انقلاب وارد حوزه آنالیز تصاویر پزشکی شود شیوه و روش‌های اختصاصی‌ای باید طراحی شوند تا خاص بودن تصاویر پزشکی را در نظر گیرند. سیستم‌های کامپیوتری هوشمند چندین دهه است که در دنیا جایگاه برجسته‌ای پیدا کرده‌اند. در حال حاضر، به خاطر تکنیک‌های جدید هوش مصنوعی[۳۸]، قابلیت پردازش کامپیوتری بالا و رشد گسترده تصویربرداری و ذخیره‌سازی دیجیتالی داده، کاربرد هوش مصنوعی در حال انتقال به حوزه‌های گوناگون می‌باشد. در حوزه پزشکی، سیستم‌های هوش مصنوعی به منظور آشکارسازی بیماری، پیش‌بینی و به عنوان استراتژی پشتیبان در تصمیم‌گیری بالینی در حال توسعه، کاوش و ارزیابی هستند. در زمینه سرطان سینه[۳۹] از هوش مصنوعی به منظور آشکارسازی زودهنگام و تفسیر ماموگرام‌ها[۴۰] به منظور بهبود غربالگری سرطان پستان و کاهش تشخیص مثبت کاذب[۴۱] استفاده می‌شود و این امکان فراهم شده است تا متخصصانی مانند رادیولوژیست‌ها[۴۲] بتوانند بر اساس میلیون‌ها تصویر از بیماران قبلی که مشخصات مشابهی دارند، تصمیمات آگاهانه‌ای بگیرند. استفاده از هوش مصنوعی در شیوه‌های تشخیص سرطان سینه به مدالیته تصویربرداری[۴۳] و همچنین تفسیر آسیب‌شناسی[۴۴] نیز گسترش یافته است. یادگیری عمیق[۴۵] که زیر شاخه‌ای از یادگیری ماشین می‌باشد یکی از تکنیک‌های هوش مصنوعی است که در انواع مختلفی از مسائل کلینیکی و پردازش تصاویر

[37] Computer Vision
[38] Artificial Intelligence (AI)
[39] Breast cancer
[40] Mammogram
[41] False positive
[42] Radiologist
[43] Imaging modality
[44] Pathology
[45] Deep learning

پزشکی شامل آشکارسازی[۴۶]/شناسایی[۴۷]، قطعه‌بندی[۴۸] و تشخیص به کمک کامپیوتر[۴۹] به کار گرفته می‌شود.

یادگیری عمیق مجموعه‌ای از الگوریتم‌های ماشین است که قادر به مدل‌سازی الگوها به طور مستقیم از داده‌های خام می-باشد. الگوریتم‌های یادگیری عمیق از مجموعه‌ای از لایه‌های چندگانه با واحدهای پردازنده غیرخطی برای استخراج و تبدیل ویژگی استفاده می‌کنند. هر لایه از خروجی لایه قبل به عنوان ورودی استفاده می‌کند. این مفهوم با بسیاری از روش‌های دیگر یادگیری ماشین که نیاز به استخراج ویژگی دارند متفاوت است. به همین ترتیب این الگوریتم‌ها حتی در مسائلی که دانش بسیار کمی در موردشان وجود دارد، می‌توانند مورد استفاده قرار گیرند. اگرچه در دهه ۱۹۹۰ این الگوریتم‌ها در برخی از مطالعات مورد استفاده قرار گرفته اند، اما در چند سال اخیر شاهد نتایج بسیار چشمگیر این الگوریتم‌ها هستیم. با توجه به وجود داده‌های بیشتر و همچنین قدرت محاسباتی بالا، این روش‌ها در بسیاری از زمینه‌ها توانسته اند به عملکرد انسان یا بهتر از انسان دست یابند[۵]. شبکه‌های عصبی مصنوعی نوع خاصی از مدل‌های یادگیری عمیق هستند که برای کار با داده‌های از نوع تصویر مناسب هستند.

شبکه‌های عصبی مصنوعی مدل‌هایی هستند که در بسیاری از زمینه‌های تحقیقاتی از جمله یادگیری ماشین کاربرد دارند. یک شبکه عصبی مصنوعی از واحدهای ساده ای به نام نورون[۵۰] تشکیل شده است که در یک سیستم پیچیده سازمان یافته اند. هر نورون بر اساس ورودی‌های خود، یک خروجی (فعال‌سازی[۵۱]) را محاسبه می‌کند که می‌تواند فعالیت‌ها یا داده‌های سایر نورون باشد. متداول‌ترین نوع شبکه عصبی، شبکه عصبی کاملاً متصل شبکه عصبی کاملا متصل پیش‌خور[۵۲] است. این شبکه‌ها دارای ورودی (جایی که داده‌ها وارد می‌شوند) و خروجی هستند. به طور معمول، هدف از استفاده از این مدل‌ها حل رگرسیون[۵۳] یا طبقه بندی[۵۴]، توسط تقریب فعال‌سازی خروجی با مقدار هدف، برای هر داده ورودی است. این شبکه‌ها به صورت لایه[۵۵] متوالی سازماندهی شده اند که یک نورون(واحد) از لایه k تمام نورون لایه ۱ $-$ k را به عنوان ورودی دریافت می‌کند، ترکیبی خطی از این مقادیر را محاسبه کرده و آن را از طریق تابع غیر خطی عبور می‌دهد

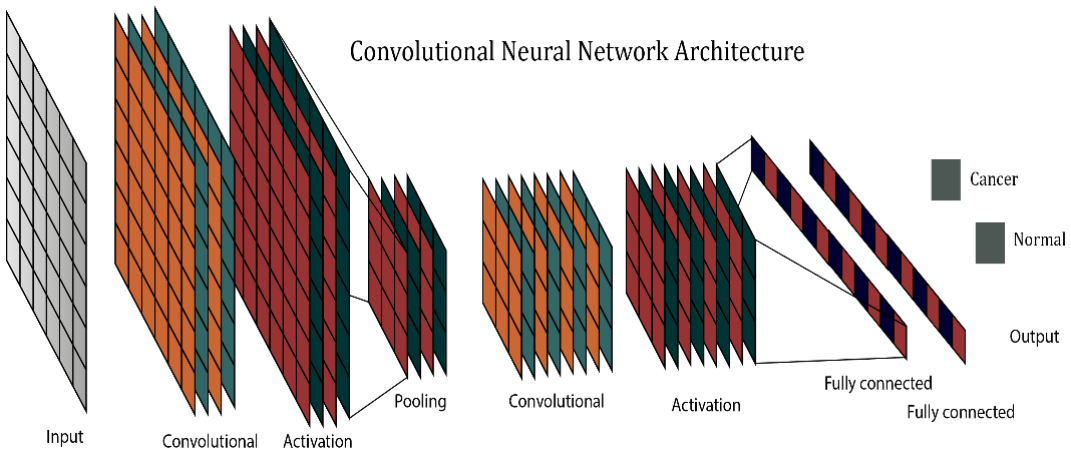
محاسبه خروجی نورون iام لایه k

$$O_{k,i} = \mathrm{actv}\left(\mathrm{W}_{k,\mathrm{i}} \cdot \mathrm{l}_{\mathrm{k-1}} + b_{k,i}\right) \qquad (۵.۲)$$

[46] Detection
[47] Recognition
[48] Segmentation
[49] Computer-aided diagnosis
[50] Neuron
[51] Activation
[52] Fully-connected feed forward neural network
[53] Regression
[54] Classification
[55] Layer

که $O_{k,i}$ واحد iام لایه k و l_{k-1} بردار تمام فعال‌سازهای لایه $k-1$ است. بردار $W_{k,i}$ و عدد $b_{k,i}$ پارامترهای ما هستند که اغلب به آنها وزن شبکه[۵۶] گفته می‌شود که برای یک وظیفه خاص آموخته می‌شوند. تابع فعال‌سازی غیرخطی actv می‌تواند اشکال مختلفی به خود بگیرد. هر مدل با یک لایه پنهان و تعداد مشخصی نورون اگر پارامترهای کافی داشته باشد می‌تواند هر تابع پیوسته ای را با خطا دلخواه تقریب بزند[۱۸].

شبکه‌های عصبی کانولوشنی یک نوع شبکه عصبی مصنوعی هستند که از نورون‌ها، لایه‌ها و وزن‌ها تشکیل شده اند. مطالعه ای که در سال ۱۹۶۸ میلادی صورت گرفت نشان داد که قشر بینایی مغز برای پردازش اطلاعات از تصاویر از الگوی پیچیده ای استفاده می‌نماید[۴۴]. نواحی ادراکی که قشر بینایی در آن قرار دارد، همانند فیلترهای محلی بر روی اطلاعات تصویر اعمال می‌شود. سلول‌های ساده‌تر برای تشخیص ویژگی‌های ادراکی سطح پایین‌تر در نواحی ادراکی مانند لبه‌ها کاربرد دارند، همچنین سلول‌های پیچیده قادر به تشخیص ویژگی‌های مهم‌تر و اختصاصی‌تر و در سطوح بالاتر می‌باشند. تشخیص ویژگی‌های اختصاصی‌تر نتیجه و ترکیبی از ویژگی‌های سطح پایین می‌باشد. این عملکرد مغز الهام بخش شبکه‌های عصبی عمیق امروزی می‌باشد. مفهوم شبکه کانولوشن نخستین بار در سال ۱۹۸۰ توسط فکوشیما مطرح گردید[۲۳]. اما به دلیل نیاز به سخت افزار ها و پردازشگرهای گرافیکی قوی استفاده از این شبکه ها برای تشخیص تا سال ۲۰۱۲ که به شکل اختصاصی برای تشخیص تصاویر ارایه و معرفی گریدی به تعویق افتاد[۳۳].



شکل ۹.۲: معماری یک شبکه عصبی کانولوشنی

همانطور که قبلاً بیان شد، شبکه‌های عصبی کانولوشنی مدل‌های شبکه عصبی کاملا متصل پیش‌خور هستند که از لایه‌های زیادی تشکیل شده اند. بسیاری از این مدل‌ها محدودیت‌های پارامتر و مکانی دارند که در ادامه توضیح داده خواهد شد. با این حال، آنها در تغییراتی که بر ورودی‌شان اعمال می‌کنند تفاوت دارند. در اینجا ما

[56] Network weight

تمام لایه‌های یک شبکه کانولوشنی و توابع مورد استفاده در آموزش آن‌ها را شرح می‌دهیم. یک معماری می‌تواند یاد بگیرد که مسائل بسیار متفاوتی را حل کند تا زمانی که پارامترها برای هر یک از مسائل به خوبی بهینه شوند.

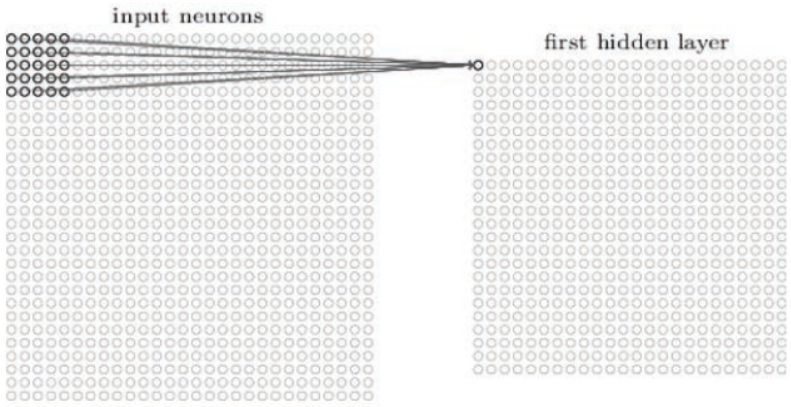
لایه ورودی فقط نمایشی از داده خام است که به مدل داده می‌شود که نیاز به شکل ورودی ثابت دارد. در رایج ترین حالت، یک تصویر به یک آرایه ۳ بعد[۵۷]ی تبدیل می‌شود با ابعاد $[w, h, ۳]$ که w و h عرض و ارتفاع هستند. بعد آخر به دلیل استفاده از تصاویر رنگی RGB[58] اغلب ۳ است. وقتی از تصاویر اشعه ایکس[۵۹] استفاده می‌کنیم چون دارای یک کانال[۶۰] شدت[۶۱] هستند بعد سوم برابر با ۱ است.

این لایه اصلی ترین لایه شبکه‌های عصبی کانولوشنی است و این شبکه‌ها نام خود را از این لایه‌ها دریافت می‌کنند. وظیفه این لایه استخراج ویژگی‌ها است. این لایه عملیات کانولوشن را بر روی داده ورودی اعمال می‌کند و خروجی‌هایی به نام نقشه ویژگی[۶۲] از این لایه به دست می‌آید. در نتیجه تمامی نورون‌ها در یک نقشه ویژگی، وزن‌ها و بایاس‌ها[۶۳] مشابه و مشترکی دارند که باعث می‌شود، ویژگی‌های تصویر در موقعیت‌های مختلف قابل شناسایی باشند. از طرف دیگر این اشتراک وزن‌ها باعث کاهش تعداد پارامترهای مورد نیاز برای آموزش می‌شود. در شبکه‌های کانولوشن اتصالات به صورت نواحی کوچک و محلی صورت می‌گیرد. به بیان دیگر هر نورون در نخستین لایه مخفی به ناحیه کوچکی از نورون‌های ورودی متصل می‌شود. برای مثال اگر این ناحیه ۵ × ۵ باشد این ناحیه کوچک ۲۵ پیکسلی ناحیه ادراک محلی[۶۴] یا کرنل کانولوشن نامیده می‌شود. با توجه به شکل ۱۰.۲ یک تصویر ورودی ۲۸ × ۲۸ داریم که یک کرنل ۵ × ۵ بر روی پیکسل‌های ورودی از چپ به راست حرکت می‌کند هر پنجره به نورونی در لایه مخفی متصل می‌شود. بنابراین همان طور که در شکل ۱۰.۲ مشخص است لایه مخفی شامل یک شبکه ۲۴ × ۲۴ نورونی خواهد بود.

در شکل ۱۰.۲ هر نورون لایه مخفی دارای یک بایاس و تعداد ۵ × ۵ وزن می‌باشد که به ناحیه ادراکی خود متصل شده است. تمامی نورون‌های لایه مخفی مذکور که دارای ابعاد ۲۴ × ۲۴ هستند، دارای وزن‌ها و بایاس‌های مشترکی می‌باشند. به عبارت دیگر خروجی نورون لایه کانولوشن $y_{w,h,m}$ در طول و عرض w, h و عمق m به صورت رابطه ۶.۲ است.

$$y_{w,h,m} = f\left(\sum_{i=(w-۱)S+۱}^{(w-۱)S+K}\sum_{j=(h-۱)S+۱}^{(h-۱)S+K}\sum_{k=۱}^{N} W_{k,m}(x_{i,j,k}) + b_m\right) \qquad (۶.۲)$$

[57] Dimension
[58] Red Green Blue
[59] X-ray
[60] Channel
[61] Intensity
[62] Feature map
[63] Bias
[64] Local receptive field



شکل ۱۰.۲: عملیات کانولوشن در یک شبکه عصبی کانولوشنی با کرنل ۵ × ۵

که در این رابطه f تابع فعالیت، b_m بایاس مشترک نورون‌ها، $W_{k,m}$ وزن‌های ۵ × ۵ مشترک نورون‌ها و $x_{i,j,k}$ ورودی در موقعیت i, j, k می‌باشد. بنابراین تمامی نورون‌های واقع در لایه مخفی اول به طور دقیق ویژگی‌های مشابهی را در نواحی مختلف تصویر شناسایی می‌کنند. در نهایت خروجی لایه ورودی یا نورون‌های لایه مخفی به عنوان نقشه ویژگی شناخته می‌شوند. ابعاد مربوط به ماتریس خروجی لایه کانولوشن $W_۲ \times H_۲ \times D_۲$ که از ماتریس ورودی با ابعاد $W_۱ \times H_۱ \times D_۱$ است، به صورت رابطه ۷.۲ به دست می‌آید.

$$W_۲ = \frac{W_۱ - F + ۲P}{S + ۱}, \qquad H_۲ = \frac{H_۱ - F + ۲P}{S + ۱}, \qquad D_۲ = K \qquad (۷.۲)$$

در روابط ۷.۲ که بیانگر نحوه محاسبه ابعاد ماتریس خروجی کانولوشن است، F, P, S و k به ترتیب نشان دهنده اندازه کرنل، مدار لایه‌گذاری صفر[۶۵]، اندازه اندازه گام[۶۶] و تعداد فیلترها می‌باشد. طبق این روابط به ازای هر فیلتر تعداد $F \times F \times D_۱$ وزن داریم و با توجه به تعداد k فیلتر موجود، در مجموع تعداد $k(F \times F \times D_۱)$ وزن و k بایاس ایجاد می‌شود. بنابراین تعداد پارامترهایی که شبکه در یک لایه کانولوشن خود می‌بایست آموزش ببیند زیاد است.

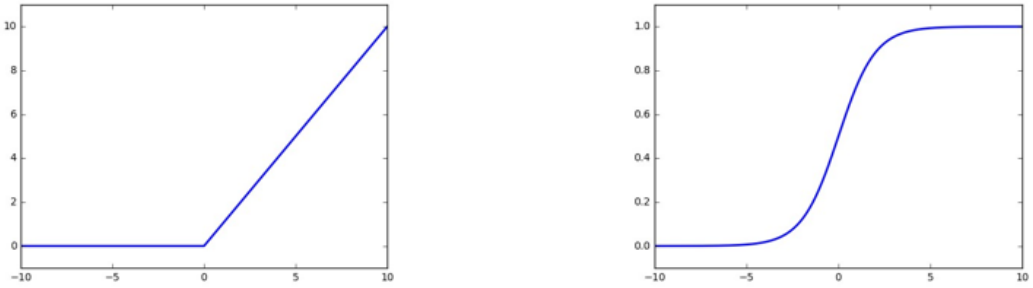
بکارگیری تابع فعالیت در لایه کانولوشن باعث ایجاد خصوصیات غیر خطی در خروجی می‌شود و باعث می‌شود عملکرد مدل متمایز کننده‌تر شود. این توابع با حفظ اندازه لایه، بدون نیاز به پارامترهای آموخته شده، یک عملکرد ساده عنصرگونه در مدل انجام می‌دهند. تابع تابع واحد اصلاح‌شده خطی[۶۷] متداول ترین تابع مورد

[65] Zero padding
[66] Stride
[67] Rectified linear unit (ReLU)

استفاده به خاطر آسان کردن مرحله آموزش است. مثال‌های دیگر شامل تابع سیگموید و هایپربولیک[۶۸] است.

$$\begin{aligned} \text{ReLU:} \quad & r_{m,n,c} = \max\{۰, l_{x,y,z}\} \\ \text{Sigmoid:} \quad & s_{m,n,c} = \frac{۱}{۱ - \exp(-l_{x,y,z})} \end{aligned} \tag{۸.۲}$$

در یک شبکه عصبی کانولوشن معمولا پس از هر لایه کانولوشن یک لایه pooling قرار می‌گیرد . این لایه از آن



(a) ReLU (b) Sigmoid

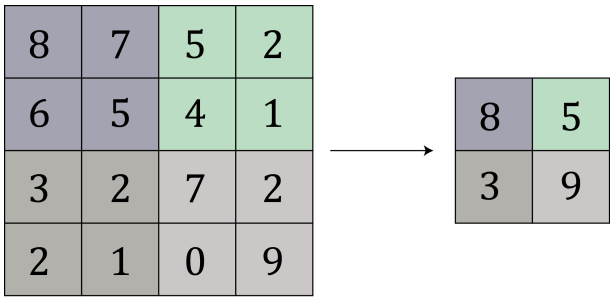
شکل ۱۱.۲: (a) تابع فعالیت ReLU و (b) تابع فعالیت سیگموید

جهت اهمیت دارد که باعث کاهش تعداد پارامترهایی می‌شود که باید آموزش ببینند. بنابراین با بکارگیری این لایه ضمن کاهش محاسبات مورد نیاز در بخش آموزش، باعث کنترل بیش‌پردازش[۶۹] احتمالی در شبکه می‌شود. این لایه بر روی هر عمق از ورودی اعمال می‌شود و اندازه آن را تغییر می‌دهد. دو تابع عملکردی معروف این لایه max-pooling و mean-pooling نام دارند که تابع اول دارای کاربرد بیشتری در شبکه‌های عصبی کانولوشنی است. طریقه عملکرد max-pooling به این صورت است که در هر پنجره بزرگترین پیکسل[۷۰] را به خروجی می‌فرستد. این پنجره بر روی تصویر مانند تابع کانولوشن از چپ به راست و از بالا به پایین با انداه گام‌های مشخص حرکت می‌کند و نتیجه را به خروجی می‌فرستد. به دلیل اینکه این عملیات بر روی تمامی عمق‌ها اعمال می‌گردد، عمق خروجی همان عمق ورودی به لایه pooling است. یک مثال از عمل max-pooling در شکل ۱۲.۲ به نمایش گذاشته شده است.

$$\text{with } l \in [s \times x, s \times x + m], j \in [s \times y, s \times y + m], \quad R_{x,y,x} = \max\{l_{i,j,z}\} \tag{۹.۲}$$

[68] Hyperbolic tangent
[69] Over-fitting
[70] Pixel

لایه کاملا متصل لایه آخر یک شبکه عصبی کانولوشنی محسوب می‌شود و اتصالات کاملی با خروجی لایه قبلی



شکل ۱۲.۲: تابع max-pooling بر روی آرایه دو بعدی کوچک $m = ۲$ و $s = ۲$

ایجاد می-کند. این لایه ورودی را دریافت و سپس خروجی را به صورت برداری با N مولفه تولید می‌کند که N تعداد کلاس‌هایی که شبکه باید طبقه بندی کند است. در واقع یک شبکه شبکه عصبی کانولوشنی جهت تولید یک بردار خروجی با N مولفه عددی طراحی می‌شود که هر عدد در این بردار خروجی درصد احتمال تعلق به کلاس مورد نظر را نشان می‌دهد. برای یک مسئله با تعداد k کلاس، k نورون خروجی داریم که هر احتمال را با تابع SoftMax محاسبه می‌کنند

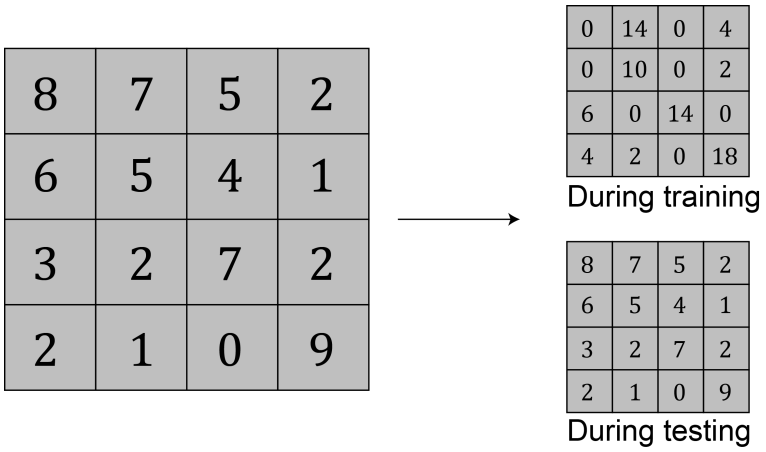
$$P(C)_j = \frac{e^{c_j}}{\sum_{k=۱}^{K} e^{c_k}} \tag{۱۰.۲}$$

اگر دو کلاس داشته باشیم می‌توانیم از تابع SoftMax با دو خروجی استفاده کنیم یا از یک نورون استفاده کنیم و تابع سیگموید را محاسبه کنیم. برای دو کلاس احتمال توسط معادله ؟؟ محاسبه می‌شود

$$P(۱) = \frac{۱}{۱ + e^{i}} \qquad P(۰) = ۱ - P(۱) \tag{۱۱.۲}$$

حذف تصادفی یک روش بسیار رایج برای جلوگیری از بیش‌پردازش شبکه عصبی مصنوعی از جمله مدل‌های یادگیری عمیق است[۴۰]. ایده این تکنیک این است که با جلوگیری از هماهنگی نورون‌ها ، ویژگی‌های قوی تری ایجاد شود. اجرای آن ساده است تنها نیاز به بهم چسباندن لایه‌های اضافی در شبکه معمولا پس از توابع فعال سازی است. این ماژول بطور تصادفی برخی از نقاط نقشه ویژگی ورودی را صفر می‌کند. هریک از ماژول‌ها دارای یک احتمال مستقل σ برای نگهداری نقاط هستند و در صورت بروز چنین اتفاقی ، توسط $\frac{۱}{\sigma}$ مقیاس بندی می‌شوند. نقاطی که نگهداری نمی‌شوند بر روی صفر تنظیم می‌شوند. این لایه فقط یک پارامتر σ دارد، که برای

آموزش در فاصله [۰, ۱] قرار دارد و برای آزمایش روی ۱ قرار می‌گیرد. به طور شهودی، می‌توان این فرآیند را به عنوان حذف برخی از نورون‌های شبکه عصبی ، به طور موقت ، همراه با اتصالات ورودی و خروجی آن تصور کرد. مکانیزم حذف، نورون‌هایی را که به اتصالات ورودی کمتری متکی هستند را در نظرمی‌گیرد. زیرا افت یک زیر مجموعه از ورودی‌ها در مقایسه با یک نورون که به بسیاری از ورودی‌ها متکی است، قابل توجه‌تر خواهد بود و به این ترتیب ویژگی‌های کلی‌تر مهم‌تر می‌شوند. شکل ۱۳.۲ یک مثال از لایه حذف تصادفی را نمایش می‌دهد. نرمال‌سازی دسته[۷۱] یک تکنیک جدید ولی خیلی کارامد است. در طی آموزش مدل‌های عمیق ، وزن‌ها در هر



شکل ۱۳.۲: لایه حذف تصادفی با $\sigma = ۰/۵$

تکرار[۷۲] به روز می‌شوند. یک اثر جانبی این امر این است که در هر لایه توزیع‌های ورودی تغییر می‌کند، پدیده ای که به آن تغییر همبستگی داخلی[۷۳]می‌گویند. این پدیده فرایند آموزش را کند می‌کند، به مقدار دهی دقیق‌تر وزن احتیاج دارد و مانع بهینه‌سازی[۷۴] مدل‌های غیرخطی اشباع، مانند مماس‌های سیگموید یا هایپربولیک می‌شود. برای حل این مشکل نرمال‌سازی دسته را پیشنهاد می‌شود که مشابه با حذف تصادفی، به عنوان لایه ای در شبکه با رفتارهای متفاوت در حین آموزش و آزمون پیاده سازی می‌شود. برای رفع مشکل تغییر کواریانس[۷۵] داخلی ، این لایه برای هر دسته آموزش با کم کردن میانگین و تقسیم بر انحراف استاندارد[۷۶] همه نورون‌های عمق مشابه، ورودی خود را نرمال می‌کند. به میانگین و انحراف استاندارد آمار mini-batch گفته می‌شود. برای اطمینان از اینکه مدل می‌تواند دقیقاً همان تابع را با یا بدون نرمال‌سازی دسته عادی نشان دهد ، دو وزن جدید قابل تمرین γ

[71] Batch normalization
[72] Iteration
[73] Internal covariate shift
[74] Optimization
[75] Covariance
[76] Standard deviation

و β اضافه می‌شوند که خروجی را اندازه گیری و جبران می‌کنند. بنابراین خروجی به صورت معادله ۱۲.۲ است.

$$\begin{aligned} &\text{در طی آموزش} : \quad I_c = \gamma \left(\frac{I_c - \text{mean}(I_c)}{\text{std}(I_c)} \right) + \beta \\ &\text{در طی آزمایش} : \quad I_c = \gamma \left(\frac{I_c - u_c}{v_c} \right) + \beta \end{aligned} \tag{۱۲.۲}$$

که u_c و v_c متوسط‌های در حال اجرا $\text{mean}(I_c)$ و $\text{std}(I_c)$ هستند. نشان داده شده است که نرمال‌سازی دسته باعث آهنگ یادگیری بالاتر می‌شود و مدل در تکرارهای کمتری همگرا خواهد شد. این روش دارای اثر رگولاریزیشن[۷۷] است. مدل با استفاده از تابع هزینه[۷۸] یاد می‌گیرد. این روشی است برای ارزیابی اینکه تا چه میزان خوب یک الگوریتم داده‌های مشاهده شده را می‌تواند مدل سازی کند. اگر پیش بینی‌ها بیش از حد از نتایج واقعی منحرف شوند ، تابع هزینه مقدار بالایی خواهد داشت. به تدریج، با کمک برخی توابع بهینه سازی، تابع هزینه می‌آموزد تا خطا در پیش بینی را کاهش دهد.

بهینه سازی مهمترین بخش در الگوریتم‌های یادگیری عمیق است. این کار با تعریف تابع هزینه شروع می‌شود و با به حداقل رساندن آن با استفاده از یک روش بهینه سازی به پایان می‌رسد. فرض کنید یک مجموعه داده D با تعداد I تصویر داریم. این تصاویر می‌توانند ضایعه باشند یا نباشند، بنابراین دارای برچسب $y \in \{۰, ۱\}$ هستند.

باید مدلی بسازیم که با توجه به یک تصویر ورودی I_i، یک احتمال $p(I_i)$ تولید کند که تا حد ممکن به برچسب مربوط به آن تصویر (y_i) نزدیک باشد. برای این منظور الگوریتم‌های بهینه سازی متفاوتی وجو دارد مانند Adam، SGD[۷۹] و Adadelta.

به حداقل رساندن تابع هزینه با کاهش گرادیان تقریباً رایج ترین الگوریتم برای بهینه سازی شبکه‌های عصبی است. اگر تابع هزینه آنتروپی متقاطع دودویی[۸۰] باشد و بخواهیم محاسبه کنیم که $p(I_i)$ تا چه حد خوب می‌تواند برچسب y_i را تقریب بزند از معادله ۱۳.۲ استفاده می‌شود.

$$L = \frac{۱}{|\mathcal{D}|} \sum_{i}^{|\mathcal{D}|} \Big(y_i \log \big(P(I_i) \big) + (۱ - y_i) \log \big(۱ - P(I_i) \big) \Big) \tag{۱۳.۲}$$

احتمال برای یک ورودی به وزن‌های آن (θ) بستگی دارد و با $p(I, \theta)$ نمایش داده می‌شود. با توجه به θ می‌توان $L(\theta)$ را با اجرای مدل بر روی مجموعه داده به دست آورد.

[77] Regularization
[78] Cost function
[79] Stochastic gradient descent
[80] Binary cross-entropy

بک‌پروپگیشن[۸۱] اساس آموزش شبکه عصبی است. این عمل تنظیم-دقیق وزن‌های یک شبکه عصبی بر اساس میزان خطا[۸۲] در هر دوره[۸۳] قبلی است که این امر با محاسبه مشتق‌های تابع خطا بر اساس وزن‌ها $\nabla_\theta L(\theta)$ در زمان آموزش امکان پذیراست. تنظیم مناسب وزن‌ها باعث کاهش میزان خطا می‌شود. در فرایند بک‌پروپگیشن ابتدا ورودی در سراسر شبکه انتشار داده می‌شود سپس $L(\theta)$ محاسبه شده و در نهایت این خطا از طریق تمام وزن‌ها در شبکه رو به عقب منتشر می‌شود. مشتق تابع هزینه از خروجی توسط معادله ۱۴.۲ محاسبه می‌شود.

$$\frac{\partial L}{\partial P} = \frac{\partial\Big(-\big(y_i \log(p) + (۱ - y)\log(۱ - P)\big)\Big)}{\partial P} = \frac{P - y}{P(۱ - P)} \qquad (۱۴.۲)$$

همچنین محاسبه مشتق تابع هزینه L از ورودی i به صورت معادله ۱۵.۲ محاسبه می‌شود.

$$\frac{\partial L}{\partial i} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial i} = P - y \qquad (۱۵.۲)$$

همچنین محاسبه مشتق تابع هزینه بر اساس وزن‌های لایه آخر w به صورت،

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial i}\frac{\partial i}{\partial w} = (P - y)a \qquad (۱۶.۲)$$

می‌باشد که a در آن برابر با ترکیب خطی از ورودی‌های لایه آخر است. این کار را می‌توان به راحتی به لایه‌های قبلی تعمیم داد، بنابراین می‌توان $\nabla_\theta L(\theta)$ را محاسبه کرد.

۱۱.۲ شبکه‌های عصبی بازگشتی

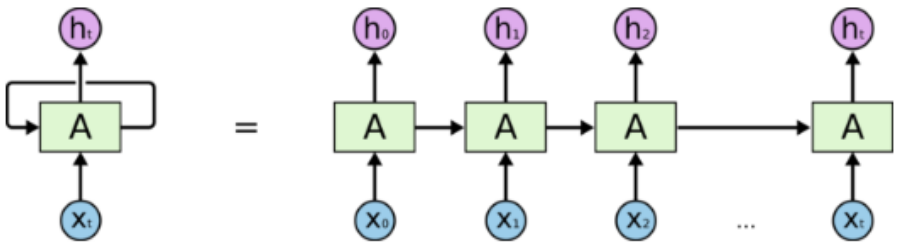
قبل از آشنا شدن با شبکه‌های عصبی بازگشتی بهتر است مروری بر مفهوم شبکه عصبی داشته باشیم. شبکه‌های عصبی مجموعه‌ای از الگوریتم‌ها هستند که شباهت نزدیکی به مغز انسان داشته و به منظور تشخیص الگوها طراحی شده‌اند. شبکه‌ی عصبی داده‌های حسی را از طریق ادراک ماشینی ، برچسب زدن یا خوشه بندی ورودی‌های خام تفسیر می‌کند. شبکه می‌تواند الگوهای عددی را شناسایی کند؛ این الگوها بردارهایی هستند که همه‌ی داده‌های دنیای واقعی (تصویر، صدا، متن یا سری‌های زمانی) برای تفسیر باید به شکل آن‌ها درآیند. شبکه‌های

[81] Back-propagation
[82] Loss
[83] epoch

عصبی مصنوعی از تعداد زیادی مؤلفه‌ی پردازشی (نورون) تشکیل شده‌اند که اتصالات زیادی بینشان وجود دارد و برای حل یک مسئله با یکدیگر همکاری دارند. شبکه‌ی عصبی مصنوعی معمولاً تعداد زیادی پردازشگر دارد که به صورت موازی کار می‌کنند و در ردیف‌هایی کنار هم قرار می‌گیرند. ردیف اول، همچون عصب‌های بینایی انسان در پردازش بصری، اطلاعات ورودی‌های خام را دریافت می‌کند. سپس هر کدام از ردیف‌های بعدی، به جای ورودی خام، خروجی ردیف قبلی را دریافت می‌کنند؛ در پردازش بصری نیز نورون‌هایی که از عصب بینایی فاصله دارند، سیگنال را از نورون‌های نزدیک‌تر می‌گیرند. ردیف آخر خروجی کل سیستم را تولید می‌کند.

۱.۱۱.۲ شبکه عصبی بازگشتی چیست؟

شبکه‌ی عصبی بازگشتی شکلی از شبکه‌ی عصبی پیشخور است که یک حافظه‌ی داخلی دارد. شبکه عصبی بازگشتی ذاتاً بازگشتی است، زیرا یک تابع یکسان را برای همه‌ی داده‌های ورودی اجرا می‌کند، اما خروجی داده‌ی (ورودی) فعلی به محاسبات ورودی قبلی بستگی دارد. خروجی بعد از تولید، کپی شده و مجدداً به شبکه‌ی بازگشتی فرستاده می‌شود. این شبکه برای تصمیم‌گیری، هم ورودی فعلی و هم خروجی که از ورودی قبلی آموخته شده را در نظر می‌گیرد. شبکه عصبی بازگشتی برخلاف شبکه‌های عصبی پیشخور می‌توانند از حالت (حافظه‌ی) درونی خود برای پردازش دنباله‌هایی از ورودی‌ها استفاده کنند. این خاصیت باعث می‌شود در مسائلی همچون تشخیص دست‌خط زنجیره‌ای یا تشخیص گفتار کاربرد داشته باشند. در سایر شبکه‌های عصبی، ورودی‌ها از یکدیگر مستقل هستند، اما در شبکه عصبی بازگشتی ورودی‌ها به هم مرتبط می‌باشند. به شکل ۱۴.۲ توجه کنید، این شبکه ابتدا X_0 را از دنباله‌ی ورودی‌ها گرفته و خروجی h_0 را تولید می‌کند که همراه با X_1 ورودی گام بعدی



An unrolled recurrent neural network.

شکل ۱۴.۲: یک نمونه بازشده شبکه عصبی بازگشتی

محسوب خواهند شد. یعنی h_0 و X_1 ورودی گام بعدی هستند. به همین صورت h_1 بعدی همراه با X_1 ورودی گام بعدی خواهند بود. شبکه عصبی بازگشتی بدین طریق می‌تواند هنگام آموزش زمینه را به خاطر داشته باشد.

فرمول حالت[84] کنونی به صورت رابطه ۱۷.۲ خواهد بود که در آن،

$$h_t = f(h_{t-1}, x_t) \tag{۱۷.۲}$$

خواهد بود که در آن h_t برابر است با،

$$h_t = \tanh(W_{hh}h_{t-1} + W_{hx}x_t) \tag{۱۸.۲}$$

در این فرمول W وزن، h تک‌بردار نهان، $W_h h$ وزن حالت نهان قبلی، W_{hx} وزن حالت ورودی کنونی و tanh تابع فعالیت است که با استفاده از تابعی غیرخطی، خروجی را فشرده می‌کند تا در بازه‌ی $[-1, 1]$ جای گیرند. در نهایت حالت خروجی Y_t از طریق رابطه ۱۹.۲ بدست می‌آید،

$$y_t = W_{hy}h_t \tag{۱۹.۲}$$

که در آن W_{hy} برابر وزن در حالت تولید شده را نشان می‌دهد.

۲.۱۱.۲ مزایای شبکه عصبی بازگشتی

شبکه عصبی بازگشتی می‌تواند دنباله‌ای از داده‌ها را به شکلی مدل‌سازی کند که هر نمونه وابسته به نمونه‌های قبلی به نظر برسد. شبکه عصبی بازگشتی را می‌توان با لایه‌های پیچشی نیز به کار برد تا گستره‌ی همسایگی پیکسلی را افزایش داد.

۳.۱۱.۲ معایب شبکه عصبی بازگشتی

- گرادیان کاهشی و مشکلات ناشی از آن
- آموزش بسیار دشوار
- ناتوانی در پردازش دنباله‌های طولانی از ورودی در صورت استفاده از تابع فعالیت tanh یا ReLU

[84] State

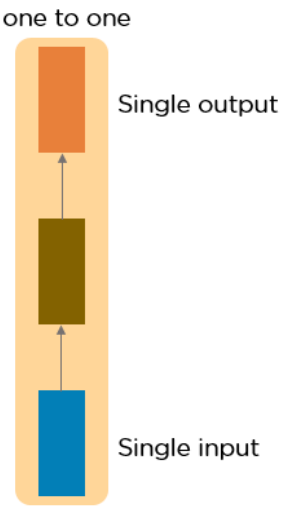
۴.۱۱.۲ کاربردهای شبکه عصبی بازگشتی

- شرح نویسی عکس[۸۵]: شبکه عصبی بازگشتی با تحلیل حالت کنونی عکس، برای شرح نویسی عکس به کار می‌رود
- پیش بینی سری‌های زمانی[۸۶]: هر مسئله سری زمانی مانند پیش بینی قیمت یک سهام در یک ماه خاص، با شبکه عصبی بازگشتی قابل انجام است
- پردازش زبان طبیعی[۸۷]: کاوش متن و تحلیل احساسات می‌تواند با استفاده از شبکه عصبی بازگشتی انجام شود
- ترجمه ماشینی[۸۸]: شبکه شبکه عصبی بازگشتی می‌تواند ورودی خود را از یک زبان دریافت و آن را به عنوان خروجی به زبان دیگری ترجمه کند

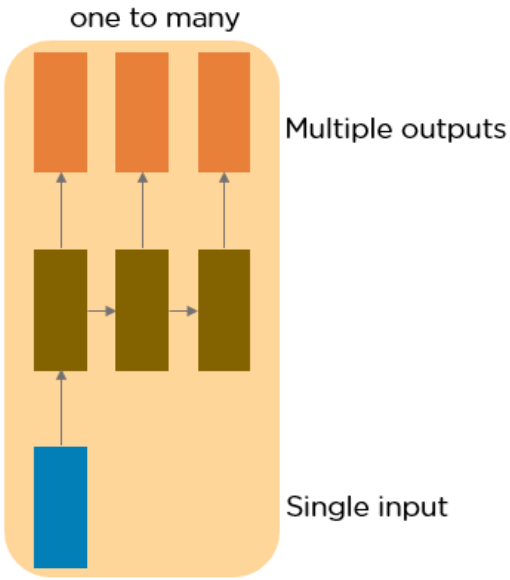
۵.۱۱.۲ انواع شبکه عصبی بازگشتی

به طور کلی ۴ نوع شبکه عصبی بازگشتی داریم:
یک به یک (one to one) : این نوع شبکه عصبی به عنوان شبکه عصبی وانیلی نیز شناخته می‌شود و برای مسائل یادگیری ماشین که یک ورودی و یک خروجی دارند به کار می‌رود.
یک به چند (one to many): این شبکه عصبی بازگشتی دارای یک ورودی و چند خروجی است. یک نمونه آن، شرح نویسی عکس است.
چند به یک (many to one): این نوع از شبکه عصبی بازگشتی، دنباله ایی از ورودی ها را می‌گیرد و یک خروجی تولید می‌کند. تحلیل احساسات مثال خوبی از این نوع شبکه است که یک جمله را به عنوان ورودی می‌گیرد و آن را با احساس مثبت یا منفی طبقه بندی می‌کند.
چند به چند (many to many): دنباله ایی از ورودی ها را می‌گیرد و دنباله ایی از خروجی ها را تولید می‌کند. ترجمه ماشینی نمونه ایی از این نوع شبکه است.

[85] Image Captioning
[86] Time Series Prediction
[87] Natural Language Processing
[88] Machine Translation



شکل ۱۵.۲: ساختار شبکه عصبی بازگشتی یک به یک

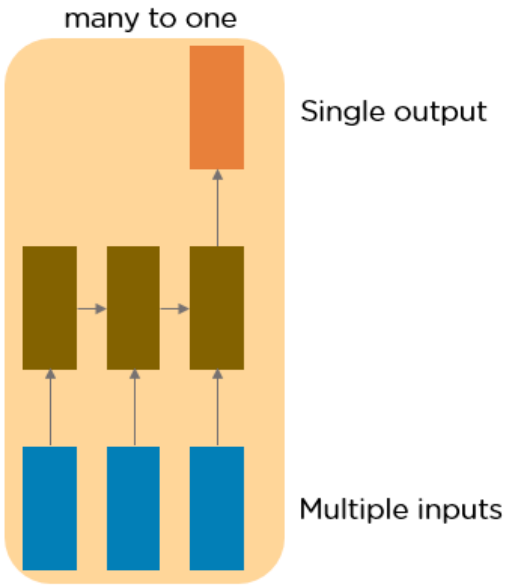


شکل ۱۶.۲: ساختار شبکه عصبی بازگشتی یک به چند

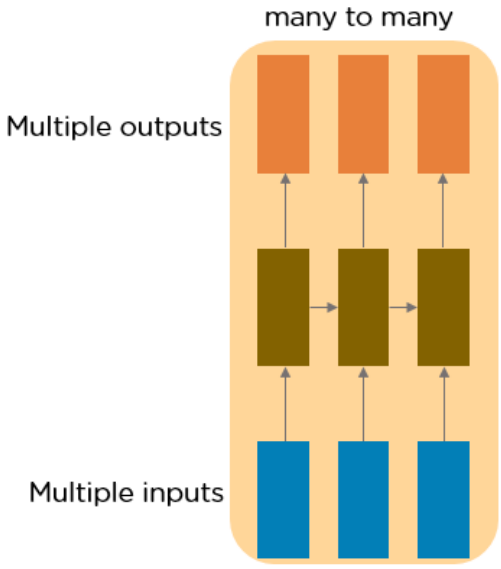
۶.۱۱.۲ حافظه‌ی کوتاه‌مدت بلند (LSTM)

شبکه‌های حافظه‌ی کوتاه‌مدت بلند[۸۹] یا LSTM نسخه‌ی تغییریافته‌ای از شبکه‌های عصبی بازگشتی هستند که یادآوری داده‌های گذشته در آن‌ها تسهیل شده است. مشکل گرادیان کاهشی که در شبکه عصبی بازگشتی وجود داشت نیز در این شبکه‌ها حل شده است. شبکه‌های LSTM برای مسائل رده‌بندی، پردازش و پیش‌بینی سری‌های

[89] Long Short Term Memory (LSTM)



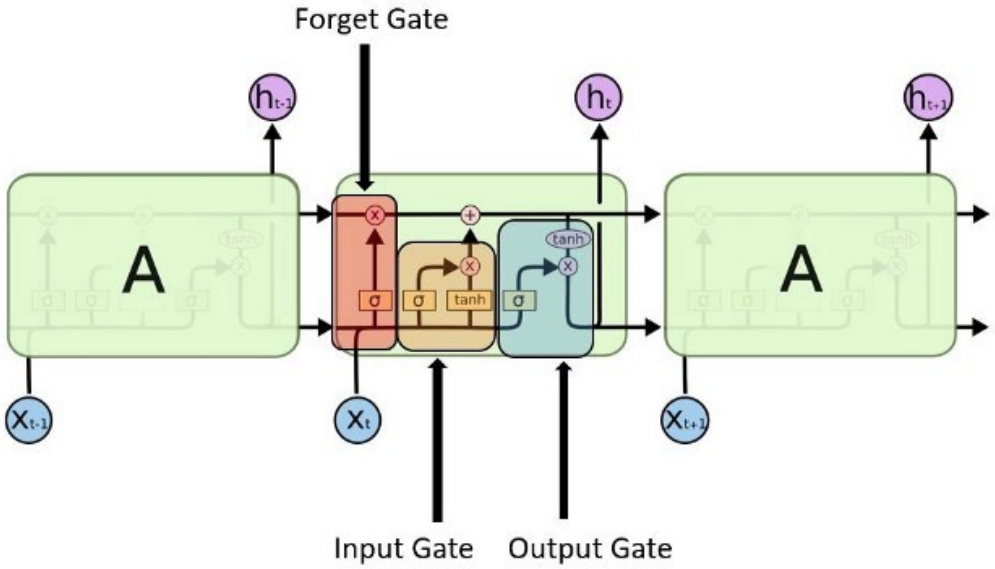
شکل ۱۷.۲: ساختار شبکه عصبی بازگشتی چند به یک



شکل ۱۸.۲: ساختار شبکه عصبی بازگشتی چند به چند

زمانی با استفاده از برچسب‌های زمانی مدت‌های نامعلوم مناسب هستند. این شبکه‌ها مدل را با استفاده از انتشار رو به عقب آموزش می‌دهند.

همان‌طور که در شکل ۱۹.۲ نمایش داده شده است، در یک شبکه‌ی LSTM سه دریچه وجود دارد:



شکل ۱۹.۲: ساختار LSTM

دریچه‌های LSTM

۱) **دریچه‌ی ورودی:** با استفاده از این دریچه می‌توان دریافت کدام مقدار از ورودی را باید برای تغییر حافظه به کار برد. تابع سیگموید تصمیم می‌گیرد مقادیر بین ۰ و ۱ اجازه‌ی ورود دارند و تابع tanh با ضریب‌دهی (بین -۱ تا $+۱$) به مقادیر، در مورد اهمیت آن‌ها تصمیم می‌گیرد.

$$i_t = \sigma(W_i \cdot [h_{t-۱}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-۱}, x_t] + b_C) \quad (۲۰.۲)$$

۲) **دریچه‌ی فراموشی:** از طریق این دریچه می‌توان جزئیاتی را که باید از بلوک حذف شوند، تشخیص داد. تصمیم‌گیری در این مورد برعهده‌ی تابع سیگموید است. این تابع با توجه به حالت قبلی $h_{t-۱}$ و ورودی محتوا X_t، عددی بین ۰ تا ۱ به هرکدام از اعداد موجود در حالت سلولی $C_{t-۱}$ اختصاص می‌دهد؛ ۰ نشان‌دهنده‌ی حذف

آن عدد و ۱ به معنی نگه داشتن آن است.

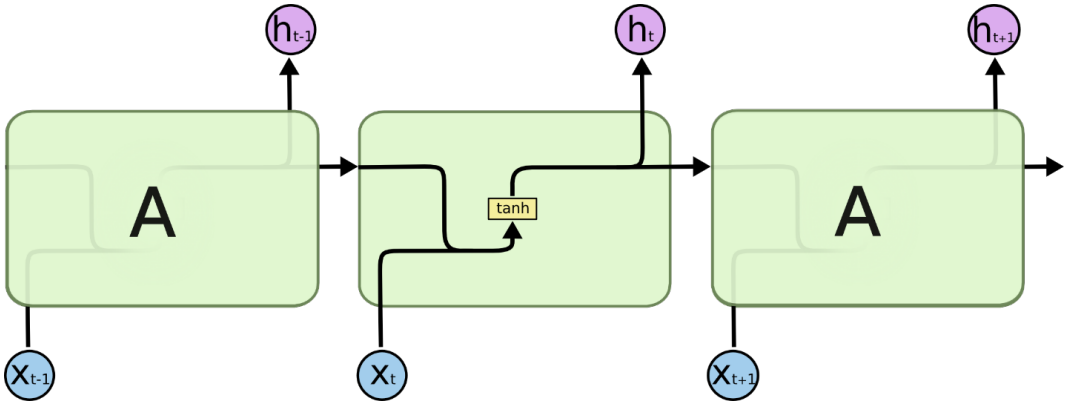
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{۲۱.۲}$$

۳) **دریچه‌ی خروجی**: ورودی و حافظه‌ی بلوک برای تصمیم‌گیری در مورد خروجی مورد استفاده قرار می‌گیرند. تابع سیگموئید تصمیم می‌گیرد مقادیر بین ۰ و ۱ اجازه‌ی ورود دارند و تابع tanh با ضریب‌دهی (بین ۱− تا ۱+) به مقادیر و ضرب آن‌ها در خروجی تابع سیگموید در مورد اهمیت آن‌ها تصمیم‌گیری می‌کند.

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{۲۲.۲}$$

در حقیقت هدف از طراحی شبکه‌های LSTM، حل کردن مشکل وابستگی بلندمدت بود. به این نکته مهم توجه کنید که به یاد سپاری اطلاعات برای بازه‌های زمانی بلند مدت، رفتار پیش‌فرض و عادی شبکه‌های LSTM است و ساختار آن‌ها به صورتی است که اطلاعات خیلی دور را به خوبی یاد می‌گیرند که این ویژگی در ساختار آن‌ها نهفته است.

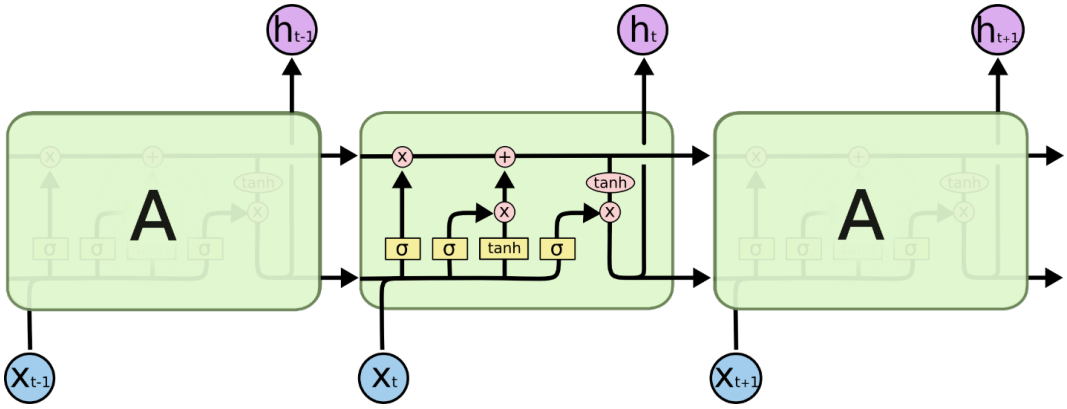
همه شبکه‌های عصبی بازگشتی به شکل دنباله‌ای (زنجیره‌ای) تکرار شونده از ماژول‌های (واحدهای) شبکه‌های عصبی هستند. در شبکه‌های عصبی بازگشتی استاندارد، این ماژول‌های تکرار شونده ساختار ساده‌ای دارند، برای مثال تنها شامل یک لایه تانژانتِ هایپربولیک (tanh) هستند.



شکل ۲۰.۲: ماژول‌های تکرار شونده در شبکه‌های عصبی بازگشتی استاندارد فقط دارای یک لایه هستند.

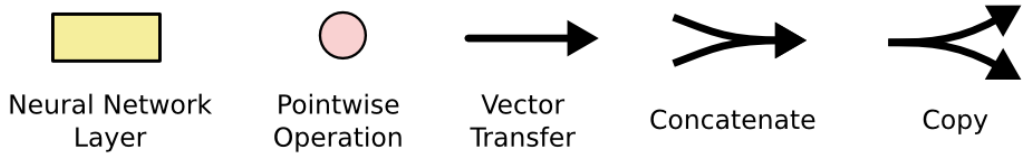
شبکه‌های LSTM نیز چنین ساختار دنباله یا زنجیره‌مانندی دارند ولی ماژولِ تکرار شونده ساختار متفاوتی دارد. به جای داشتن تنها یک لایه شبکه عصبی، ۴ لایه دارند که طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط

هستند. در ادامه قدم به قدم ساختار شبکه‌های حافظه‌ی کوتاه‌مدت بلند را توضیح خواهیم داد. اما در ابتدا معنی



شکل ۲۱.۲: ماژول‌های تکرار شونده در LSTMها دارای ۴ لایه هستند که با هم در تعامل می‌باشند.

هر کدام از شکل و علامت‌هایی را که از آن‌ها استفاده خواهیم کرد توضیح می‌دهیم. در شکل ۲۲.۲، هر خط

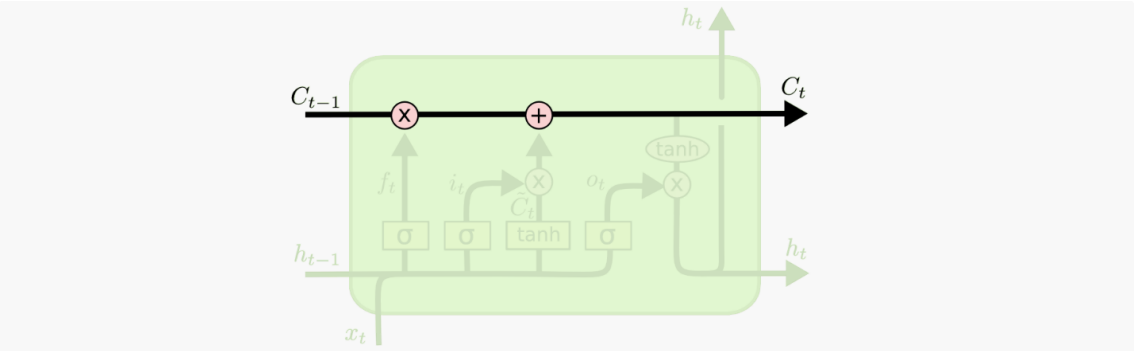


شکل ۲۲.۲: اشکال از راست به چپ به ترتیب برابر هستند با: کپی کردن، وصل کردن، بردار انتقال، عملیات نقطه به نقطه، یک لایه‌ی شبکه عصبی.

یک بردار را به صورت کامل از خروجی یک گره به ورودی گره دیگر انتقال می‌دهد. دایره‌های صورتی نمایش دهنده عملیات‌های نقطه به نقطه مانند «جمع کردن دو بردار» هستند. مستطیل‌های زرد، لایه‌های شبکه‌های عصبی هستند که شبکه پارامترهای آن‌ها را یاد می‌گیرد. خط‌هایی که با هم ادغام می‌شوند نشان‌دهنده الحاق[۹۰] و خط‌هایی که چند شاخه می‌شوند نشان دهنده‌ای این موضوع است که محتوای آن‌ها کپی و به بخش‌های مختلف ارسال می‌شود.

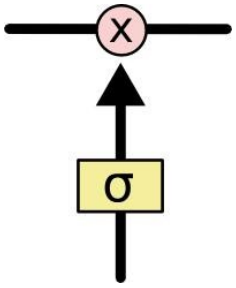
عنصر اصلی LSTMها سلول حالت[۹۱] است که در حقیقت یک خط افقی است که در بالای شکل ۲۳.۲ قرار دارد. سلول حالت را می‌توان به صورت یک تسمه نقاله تصور کرد که از اول تا آخر دنباله یا همان زنجیره با تعاملات خطیِ جزئی در حرکت است (یعنی ساختار آن بسیار ساده است و تغییرات کمی در آن اتفاق می‌افتد).
LSTM این توانائی را دارد که اطلاعات جدیدی را به سلول حالت اضافه یا اطلاعات آن را حذف کنید. این کار

[90] Concatenation
[91] Cell state



شکل ۲۳.۲: سلول حالت در ماژول LSTM

توسط ساختارهای دقیقی به نام دروازه‌ها[92] انجام می‌شود. دروازه‌ها راهی هستند برای ورود اختیاری اطلاعات. آن‌ها از یک لایه شبکه عصبیِ سیگموید به همراه یک عملگر ضرب نقطه به نقطه تشکلیل شده‌اند.



شکل ۲۴.۲: نمایی از نحوه تاثیر و ورود اطلاعات به سلول حالت

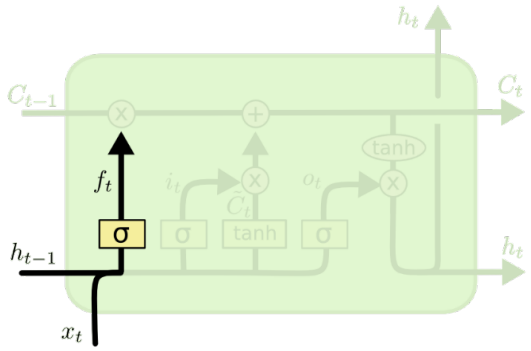
خروجی لایه سیگموید عددی بین صفر و یک است، که نشان می‌دهد چه مقدار از وروی باید به خروجی ارسال شود. مقدار صفر یعنی هیچ اطلاعاتی نباید به خروجی ارسال شود در حالی که مقدار یک یعنی تمام ورودی به خروجی ارسال شود!

LSTM دارای ۳ دروازه مشابه برای کنترل مقدار سلول حالت است که در ادامه به بررسی قدم به قدمِ آن‌ها از لحظه ورود تا خروج اطلاعات خواهیم پرداخت.

قدم اول در LSTM تصمیم در مورد اطلاعاتی است که می‌خواهیم آن‌ها را از سلول حالت پاک کنیم. این تصمیم توسط یک لایه سیگموید به نام «دروازه فراموشی[93]» انجام می‌شود. این دروازه با توجه به مقادیر h_{t-1} و x_t، برای هر عدد، مقدار صفر یا یک را در سلول حالتِ C_{t-1} به خروجی می‌برد. مقدار یک یعنی به صورت کامل مقدار حال حاضرِ سلول حالت C_{t-1} را به C_t انتقال داده شود و مقدار صفر یعنی به صورت کامل اطلاعات سلول حالت

[92] Gate
[93] Forget gate

کنونی C_{t-1} را پاک شود و هیچ مقداری از آن به C_t برده نشود. بیاید به مثال قبلی‌مان که یک مدل زبانی‌ای بود که در آن تلاش داشتیم کلمه بعدی را بر اساس همه کلمه‌های قبلی حدس بزنیم، برگردیم. در چنین مسأله‌ای، سلول حالت ممکن است دربردارنده جنسیت فاعل کنونی باشد، که با توجه به آن می‌توانیم تشخیص دهیم از چه ضمیری باید استفاده کنیم. زمانی که یک فاعل جدید در جمله ظاهر می‌شود، می‌بایست جنسیت فاعل قبلی حذف شود.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

شکل ۲۵.۲: قدم اول در پاک کردن اطلاعات از سلول حالت در وضعیت ورودی

قدم بعدی این است که تصمیم بگیریم چه اطلاعات جدیدی را می‌خواهیم در سلول حالت ذخیره کنیم. این تصمیم دو بخشی است. ابتدا یک لایه سیگموید به نام دروازه ورودی[۹۴] داریم که تصمیم می‌گیرد چه مقادیری به‌روز خواهند شد. مرحله بعدی یک لایه تانژانت هایپربولیک است که برداری از مقادیر به نام $\tilde{C}_t$ می‌سازد که می‌توان آن‌ها را به سلول حالت اضافه کرد. در مرحله بعد، ما این دو مرحله را با هم ترکیب می‌کنیم تا مقدار سلول حالت را به‌روز کنیم.

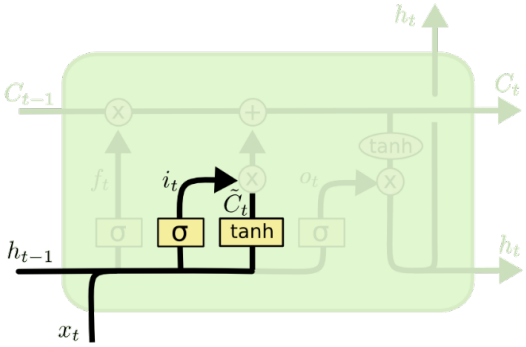
در مثال مدل زبانی‌ای که پیش‌تر داشتیم، قصد داریم جنسیت فاعل جدید را به سلول حالت اضافه کنیم تا جایگزین جنسیت فاعل قبلی شود که در مرحله قبلی تصمیم گرفتیم آن را فراموش کنیم.

حال زمان آن فرا رسیده است که سلول حالت قدیمی یعنی C_{t-1} را سلول حالت جدید یعنی C_t به‌روز کنیم. در مراحل قبلی تصمیم گرفته شد که چه کنیم و در حال حاضر تنها لازم است تصمیماتی را که گرفته شد عملی کنیم.

ما مقدار قبلی سلول حالت را در f_t ضرب می‌کنیم که یعنی فراموش کردن اطلاعاتی که پیش‌تر تصمیم گرفتیم آن‌ها را فراموش کنیم. سپس $i_t * \tilde{C}_t$ را به آن اضافه می‌کنیم. در حال حاضر مقادیر جدید سلول حالت با توجه به تصمیماتی که پیش‌تر گرفته شده بود بدست آمده‌اند. در مثال مدل زبانی، اینجا دقیقاً جائی است که اطلاعاتی که در مورد جنسیت قبلی داشتیم را دور می‌ریزیم و اطلاعات جدید را اضافه می‌کنیم.

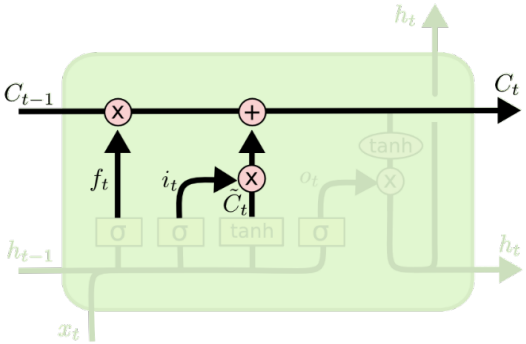
در نهایت باید تصمیم بگیریم قرار است چه اطلاعاتی را به خروجی ببریم. این خروجی با در نظر گرفتن مقدار سلول حالت خواهد بود، ولی از فیلتر مشخصی عبور خواهد کرد. در ابتدا، یک لایه سیگموید داریم که تصمیم

[94] Input gate



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

شکل ۲۶.۲: قدم دوم در اضافه کردن اطلاعات جدید به سلول حالت

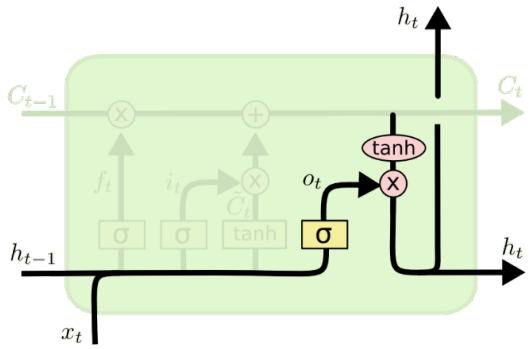


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

شکل ۲۷.۲: به‌روز رسانی اطلاعات در سلول حالت

می‌گیرد چه بخشی از سلول حالت قرار است به خروجی برده شود. سپس مقدار سلول حالت (پس از به‌روز شدن در مراحل قبلی) را به یک لایه تانژانت هایپربولیک (تا مقادیر بین ۱− و ۱+ باشند) می‌دهیم و مقدار آن را در خروجی لایه سیگموید قبلی ضرب می‌کنیم تا تنها بخش‌هایی که مد نظرمان است به خروجی برود.

در مثال مدل زبانی، با توجه به اینکه تنها فاعل را دیده‌است، در صورتی که بخواهیم کلمه بعدی را حدس بزنیم، ممکن است بخواهد اطلاعاتی در ارتباط با فعل را به خروجی ببرد. برای مثال ممکن است اینکه فاعل مفرد یا جمع است را به خروجی ببرد، که ما با توجه به آن بدانیم فعل به چه فرمی خواهد بود.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

شکل ۲۸.۲: قدم نهایی برای تولید خروجی ماژول LSTM

۱۲.۲ یادگیری تقویتی

۱.۱۲.۲ مقدمه و بیشینه تاریخی

ادوارد ثورندایک[۹۵] پدر روانشناسی مدرن در سال ۱۸۷۴ میلادی در ایالت ماساچوست آمریکا متولد شد. وی در اوایل قرن ۲۰ میلادی آزمایشی انجام داد که باعث ارائه قانون اثر شد. او برای این آزمایش، گربه ای را در جعبه ای موسوم به جعبه معما قرار داد. هر کوشش درستی، از این گربه برای نجات از جعبه صورت می گرفت، باعث میشد ثورندایک به عنوان پاداش به او غذا بدهد. به تدریج گربه به کارهای درست خود پی برد و آنها را تکرار کرد، تا جایی که دیگر هیچ کار اشتباهی نمی کرد و بلاخره موفق به خروج از جعبه شد. ثورندایک در سال ۱۹۱۲ به ریاست انجمن روانشناسان، در سال ۱۹۱۷ به عضویت انجمن علوم، در سال ۱۹۳۴به ریاست انجمن علوم پیشرفته نایل آمد و در سال ۱۹۴۷ در سن ۷۴ سالگی، بدرود حیات گفت. در سال ۲۰۰۲ رتبه ای از برترین روانشناسان تاریخ ارائه شد که ثورندایک جزء ۱۰ روانشناس برتر تاریخ قرار گرفت. می توان مهم ترین کشف وی را، اثبات وجود یادگیری تقویتی در روانشناسی دانست.

شاید ریچارد بلمن[۹۶] (مخترع الگوریتم بلمن–فورد) را بتوان اولین کسی دانست که یادگیری تقویتی را وارد هوش مصنوعی ساخت. در اوایل دهه ۱۹۵۰ بلمن مسئله ای با عنوان «کنترل بهینه» را مطرح ساخت که با استفاده از روش های پویا در برنامه ریزی پویا کنترل کننده ها را به سمت نتیجه بهینه رهنمون می شد. در اواخر دهه ۵۰ میلادی مینسکی در پایان نامه دکتری خود روش های محاسبات آزمون و خطا توسط مفهوم یادگیری تقویتی را مطرح نمود و الگوریتم های یادگیری تقویتی را پایه ریزی کرد. در کل دهه ۵۰ میلادی را میتوان دهه تشکیل الگوریتم های محاسباتی اولیه یادگیری تقویتی دانست. در دهه ۶۰ میلادی اولین کابرد های یادگیری تقویتی به

[95] Edward Thorndike
[96] Richard E. Bellman

وقوع پیوستند. در اولین تلاش ها فارلی و کلارک از یادگیری تقویتی برای تشخیص الگو استفاده کردند بدین صورت که هر بار برنامه نتیجه بهتری به دست می آمد او را تشویق می کردند. در اواخر دهه ۶۰ میلادی، یادگیری نظارتی از یادگیری تقویتی ، مشتق شد. در یادگیری نظارتی طراح نتیجه نهایی را در دست دارد و از هوش مصنوعی می خواهد هر بار مسیر بین ورودی و نتیجه را طراحی کرده و هربار که برنامه، مسیر بهتری به دست می آورد، تشویق می شود. همچنین طراح نظارت مستقیم بر عملکرد عامل دارد.

فصل ۳

روش‌های پیشین

فصل ۴

روش پیشنهادی

۱.۴ مقدمه

پس از آشنایی با روش‌های پیشین که برای حل مسئله مشابه مورد استفاده قرار گرفته‌اند، حال می‌توانیم به معرفی و تشریح روش‌های پیشنهادی خود برای حل مسئله پیش رو بپردازیم. در این فصل ابتدا داده‌های ورودی مسئله را همراه با فرضیات در نظر گرفته شده بیان می‌کنیم و پس از آن دو روش پیشنهادی متفاوت را بیان خواهیم نمود. در روش اول که به رویکردهای پیشین نزدیک‌تر است با تغییری از جنس روش‌های نوین در مراحل میانی به یک روش جدید می‌رسیم که به علت افزایش سرعت همگرایی می‌توان فرض و داده‌های جدیدی را از طریق CNV به آن افزود و پاسخ گرفت. اما روش دوم کاملا متفاوت بوده و با رویکردی جدید در حوزه یادگیری ماشین همراه است که به کمک یادگیری تقویتی به حل مسئله مورد نظر می‌پردازد.

۲.۴ معرفی دادگان ورودی

قبل از وارد شدن به بخش روش‌های پیشنهادی نیاز است تا دادگان ورودی را مشخص و معرفی نماییم تا در قسمت‌های بعدی بتوانیم از نمادهای معرفی شده در این بخش استفاده نماییم.
دادگان مورد استفاده

۳.۴ روش پیشنهادی اول (درخت‌بازی)

۱.۳.۴ پیش‌پردازش

قبل از شروع باید بر روی داده‌ها یک پیش‌پردازش اعمال کنیم که وابسته به سیاست درنظر گفته شده می‌تواند باعث تغییر در پاسخ نهایی نیز شود. به این منظور داده‌هایی که miss شده‌اند با روش‌های زیر می‌تواند برای ورود به مرحله بعد تخمین زده شود.

۱.۱.۳.۴ تصادفی

پر کردن کاملا تصادفی میس‌ها

جدول ۱.۴: اندیس‌های به کار رفته در مدل ریاضی

I, J	بیماران
k	مرحله زمان‌بندی (بستری، اتاق عمل، ریکاوری)
L_k	ماشین (تخت یا اتاق عمل) در مرحله k
n	جراح

جدول ۲.۴: پارامترهای مدل ریاضی

t_{ik}	زمان خدمت‌دهی به بیمار در مرحله kام
$\tilde{t}_{ik}$	زمان فاری خدمت‌دهی به بیمار در محله kام
t^p_{ik}	مقدار بدبینانه (حداکثر) برای زمان خدمت‌دهی به بیمار در مرحله kام
t^m_{ik}	محتمل‌ترین مقدار برای زمان خدمت‌دهی به بیمار در مرحله kام
t^o_{ik}	مقدار خوشبینانه (حداقل) برای زمان خدمت‌دهی به بیمار در مرحله kام

جدول ۳.۴: متغیرهای مدل ریاضی

X_{ild_k}	متغیر صفر-یک تخصیص بیمار به تخت/اتاق عمل
S_{ild_k}	زمان شروع خدمت‌دهی به بیمار
Y_{ijkl_k}	متغیر صفر-یک توالی بیماران
V_{ni}	متغیر صفر-یک تخصیص جراح به بیمار

فصل ۵

نتایج تجربی

فصل ۶

بحث و نتیجه‌گیری

مراجع

[1] Nci dictionary of cancer terms: somatic mutation definition. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/somatic-mutation?redirect=true.

[2] Ii neoplasms. 19 June 2014.

[3] Cancer - activity 1 - glossary. page page 4 of 5, 2008.

[4] Abrams, Gerald. Neoplasia i. 23 January 2012.

[5] Akselrod-Ballin, Ayelet, Karlinsky, Leonid, Hazan, Alon, Bakalo, Ran, Horesh, Ami Ben, Shoshan, Yoel, and Barkan, Ella. Deep learning for automatic detection of abnormal findings in breast mammography. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 321–329. Springer, 2017.

[6] Alberts, Bruce, Johnson, Alexander, Lewis, Julian, Raff, Martin, Roberts, Keith, and Walter, Peter. Molecular biology of the cell 4th edition. *New York: Garland Science*, 1463, 2002.

[7] Anderson, Kristina, Lutz, Christoph, Van Delft, Frederik W, Bateman, Caroline M, Guo, Yanping, Colman, Susan M, Kempski, Helena, Moorman, Anthony V, Titley, Ian, Swansbury, John, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356–361, 2011.

[8] Andor, Noemi, Graham, Trevor A, Jansen, Marnix, Xia, Li C, Aktipis, C Athena, Petritsch, Claudia, Ji, Hanlee P, and Maley, Carlo C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine*, 22(1):105–113, 2016.

[9] Behjati, Sam, Huch, Meritxell, van Boxtel, Ruben, Karthaus, Wouter, Wedge, David C, Tamuri, Asif U, Martincorena, Iñigo, Petljak, Mia, Alexandrov, Ludmil B, Gundem, Gunes, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, 2014.

[10] Birbrair, Alexander, Zhang, Tan, Wang, Zhong-Min, Messi, Maria Laura, Olson, John D, Mintz, Akiva, and Delbono, Osvaldo. Type-2 pericytes participate in normal and tumoral angiogenesis. *American Journal of Physiology-Cell Physiology*, 307(1):C25–C38, 2014.

[11] Bishop, Christopher M. Pattern recognition. *Machine learning*, 128(9), 2006.

[12] Burrell, Rebecca A and Swanton, Charles. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular oncology*, 8(6):1095–1111, 2014.

[13] Chen, Rui, Mias, George I, Li-Pook-Than, Jennifer, Jiang, Lihua, Lam, Hugo YK, Chen, Rong, Miriami, Elana, Karczewski, Konrad J, Hariharan, Manoj, Dewey, Frederick E, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.

[14] Cooper, Geoffrey M. *Elements of human cancer*. Jones & Bartlett Learning, 1992.

[15] de Visser, J Arjan GM and Rozen, Daniel E. Clonal interference and the periodic selection of new beneficial mutations in escherichia coli. *Genetics*, 172(4):2093–2100, 2006.

[16] Demicheli, R, Retsky, MW, Hrushesky, WJM, Baum, M, and Gukas, ID. The effects of surgery on tumor growth: a century of investigations. *Annals of oncology*, 19(11):1821–1828, 2008.

[17] Dentro, Stefan C, Leshchiner, Ignaty, Haase, Kerstin, Tarabichi, Maxime, Wintersinger, Jeff, Deshwar, Amit G, Yu, Kaixian, Rubanova, Yulia, Macintyre, Geoff, Vázquez-García, Ignacio, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *BioRxiv*, page 312041, 2018.

[18] Dhungel, Neeraj, Carneiro, Gustavo, and Bradley, Andrew P. Fully automated classification of mammograms using deep residual neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 310–314. IEEE, 2017.

[19] Fearon, Eric R and Vogelstein, Bert. A genetic model for colorectal tumorigenesis. *cell*, 61(5):759–767, 1990.

[20] Fedele, Clare, Tothill, Richard W, and McArthur, Grant A. Navigating the challenge of tumor heterogeneity in cancer therapy. *Cancer discovery*, 4(2):146–148, 2014.

[21] Fisher, Rosie, Pusztai, Lazos, and Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[22] Friedl, Peter and Wolf, Katarina. Plasticity of cell migration: a multiscale tuning model. *Journal of Cell Biology*, 188(1):11–19, 2010.

[23] Fukushima, Kunihiko. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.

[24] Gelman, Andrew, Shirley, Kenneth, et al. Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, 6:163–174, 2011.

[25] Gerlinger, Marco, Rowan, Andrew J, Horswell, Stuart, Larkin, James, Endesfelder, David, Gronroos, Eva, Martinez, Pierre, Matthews, Nicholas, Stewart, Aengus, Tarpey, Patrick, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 366:883–892, 2012.

[26] Greaves, Mel and Maley, Carlo C. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[27] Halford, S, Rowan, A, Sawyer, E, Talbot, I, and Tomlinson, Ian. O6-methylguanine methyltransferase in colorectal cancers: detection of mutations, loss of expression, and weak association with g: C> a: T transitions. *Gut*, 54(6):797–802, 2005.

[28] Hanahan, Douglas and Weinberg, Robert A. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[29] Hanahan, Douglas and Weinberg, Robert A. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[30] Handa, Osamu, Naito, Yuji, and Yoshikawa, Toshikazu. Redox biology and gastric carcinogenesis: the role of helicobacter pylori. *Redox Report*, 16(1):1–7, 2011.

[31] Hastings, W Keith. Monte carlo sampling methods using markov chains and their applications. 1970.

[32] Hugo, Honor, Ackland, M Leigh, Blick, Tony, Lawrence, Mitchell G, Clements, Judith A, Williams, Elizabeth D, and Thompson, Erik W. Epithelial—mesenchymal and mesenchymal—epithelial transitions in carcinoma progression. *Journal of cellular physiology*, 213(2):374–383, 2007.

[33] LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015.

[34] Lee, Kyung-Hwa, Lee, Ji-Shin, Nam, Jong-Hee, Choi, Chan, Lee, Min-Cheol, Park, Chang-Soo, Juhng, Sang-Woo, and Lee, Jae-Hyuk. Promoter methylation status of hmlh1, hmsh2, and mgmt genes in colorectal cancer associated with adenoma–carcinoma sequence. *Langenbeck's archives of surgery*, 396(7):1017–1026, 2011.

[35] Nik-Zainal, Serena, Van Loo, Peter, Wedge, David C, Alexandrov, Ludmil B, Greenman, Christopher D, Lau, King Wai, Raine, Keiran, Jones, David, Marshall, John, Ramakrishna, Manasa, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[36] Nowell, Peter C. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[37] Sabeh, Farideh, Shimizu-Hirota, Ryoko, and Weiss, Stephen J. Protease-dependent versus-independent cancer cell invasion programs: three-dimensional amoeboid movement revisited. *Journal of Cell Biology*, 185(1):11–19, 2009.

[38] Sakr, WA, Haas, GP, Cassin, BF, Pontes, JE, and Crissman, JD. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *The Journal of urology*, 150(2):379–385, 1993.

[39] Sokal, Alan. Monte carlo methods in statistical mechanics: foundations and new algorithms. In *Functional integration*, pages 131–192. Springer, 1997.

[40] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[41] Stewart, BWKP and Wild, CP. World cancer report 2014. health, 2017.

[42] Stratton, Michael R, Campbell, Peter J, and Futreal, P Andrew. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[43] Sun, Xiao-xiao and Yu, Qiang. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica*, 36(10):1219–1227, 2015.

[44] Sutherland, NS. Outlines of a theory of visual pattern recognition in animals and man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):297–317, 1968.

[45] Talbot, Simon J and Crawford, Dorothy H. Viruses and tumours–an update. *European Journal of Cancer*, 40(13):1998–2005, 2004.

[46] Truninger, Kaspar, Menigatti, Mirco, Luz, Judith, Russell, Anna, Haider, Ritva, Gebbers, Jan-Olaf, Bannwart, Fridolin, Yurtsever, Hueseyin, Neuweiler, Joerg, Riehle, Hans-Martin, et al. Immunohistochemical analysis reveals high frequency of pms2 defects in colorectal cancer. *Gastroenterology*, 128(5):1160–1171, 2005.

[47] Vander Heiden, Matthew G, Cantley, Lewis C, and Thompson, Craig B. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

[48] Waclaw, Bartlomiej, Bozic, Ivana, Pittman, Meredith E, Hruban, Ralph H, Vogelstein, Bert, and Nowak, Martin A. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264, 2015.

[49] Zhu, Aizhi, Lee, Daniel, and Shim, Hyunsuk. Metabolic positron emission tomography imaging in cancer detection and therapy response. In *Seminars in oncology*, volume 38, pages 55–69. Elsevier, 2011.

Abstract

This thesis studies on writing projects, theses and dissertations using tehran-thesis class. It ...

Keywords SNV, CNV, Phylogenetic, Tree, Q-learning, Deep learning

University of Tehran
College of Engineering
Faculty of New Science and Technology
Network

Inference of Phylogenetic Tree for Inter Tumor using Single Cell Mutations and CNV

A Thesis submitted to the Graduate Studies Office
In partial fulfillment of the requirements for
The degree of Master of Science
in Information Technology - Network Science

By:

Afshin Bozorgpour

Supervisors:

Dr. Saman Haratizadeh and Dr. Abolfazl Motahari

Jul 2021