دانشکده مهندسی

گروه مهندسی کامپیوتر

درجه کارشناسی ارشد گرایش هوش مصنوعی

داده‌کاوی در داده‌های ریز آرایه جهت تشخیص سرطان

نگارش:

فاطمه امین‌زاده

استاد راهنما:

جناب آقای دکتر علیرضا عصاره

استاد مشاور:

سرکار خانم دکتر بیتا شادگار

۱۳۸۸

به نام خداوند بخشنده مهربان

دانشکده مهندسی

گروه مهندسی کامپیوتر

درجه کارشناسی ارشد گرایش هوش مصنوعی

داده‌کاوی در داده‌های ریز آرایه جهت تشخیص سرطان

نگارش:

فاطمه امین‌زاده

استاد راهنما:

جناب آقای دکتر علیرضا عصاره

استاد مشاور:

سرکار خانم دکتر بیتا شادگار

۱۳۸۸

نظر هیئت داوران

استاد راهنما : جناب آقای دکتر عصاره

تاریخ و امضا

استاد داور : جناب آقای

تاریخ و امضا

تعهد نامه

بدینوسیله اینجانب فاطمه امین‌زاده قصرالدشتی تعهد می‌نمایم که مطالب ارائه شده در این پایان نامه حاصل کار پژوهشی و تحقیق اینجانب می‌باشد و قبلاً برای احراز هیچ مدرک دیگری ارائه نشده است. رجوع به دست آوردهای دیگران که در این پایان نامه از آن‌ها استفاده شده مطابق مقررات ارجاع داده شده است. کلیه حقوق مادی و معنوی این اثر متعلق به دانشکده مهندسی دانشگاه شهید چمران اهواز می‌باشد.

نام و نام خانوادگی دانشجو: فاطمه امین‌زاده

تاریخ: امضاء:

صفحه تقدیم

صفحه تقدیر و تشکر

چکیده

تکنولوژی ریزآرایه باعث تولید حجم انبوهی داده دسته‌بندی در بسیاری از زمینه‌ها شده است. تحلیل داده‌های ریزآرایه وکلاس‌بندی آن‌ها، نشان می‌دهد این روش در تشخیص بیماری‌ها و سرطان تأثیر بسزایی دارد. با توجه به تحقیقات بسیاری که در مورد کلاس‌بندی داده‌های ریزآرایه صورت گرفته است، اعمال روش‌های معمول یادگیری ماشین دارای معایبی ذاتی برای رسیدن به یک کلاس‌بندی پایدار و دقیق است. بنابراین مطلوب‌تر است که از ترکیب دسته‌بندی‌کننده‌های خبره، به‌جای تکیه بر نتیجه‌ی یک دسته‌بندی‌کننده منفرد استفاده شود. در این تحقیق، دسته-بندی داده‌های ریزآرایه‌ای با استفاده از روش‌های موثر انتخاب ژن‌ها و دسته‌بندی‌کننده‌ی ترکیبی RotBoost انجام شده است. این دسته‌بندی‌کننده از ترکیب تکنیک‌های Rotation Forest و AdaBoost به‌وجود می‌آید. با توجه به تحقیقات قبلی انجام شده در زمینه‌ی ریزآرایه‌ها، این نخستین باری است که تکنیک RotBoost بر روی دسته‌بندی‌کننده‌ی داده‌های ریزآرایه اعمال گردیده است. این تحقیق بر روی ۸ مجموعه‌داده‌ی ریزآرایه واقعی، پیاده‌سازی شده و از درخت تصمیم به عنوان دسته‌بندی‌کننده‌ی پایه استفاده شده است. بررسی نتایج به‌دست آمده نشان می‌دهد که در اکثر این ریزآرایه‌ها، روش پیشنهادی از صحت بالاتری نسبت به تکنیک‌هایی نظیر Rotation Forest و AdaBoost برخوردار است.

واژه‌های کلیدی: داده‌های ریزآرایه DNA، دسته‌بندی‌کننده ترکیبی، دسته بندی سرطان، راپر، فیلتر.

فهرست مطالب

صفحه

فصل چهارم

فصل پنجم

فهرست اشکال

فصل اول

فصل سوم

فصل چهارم

فهرست جداول

فصل اول

فصل چهارم

فصل اول

مقدمه

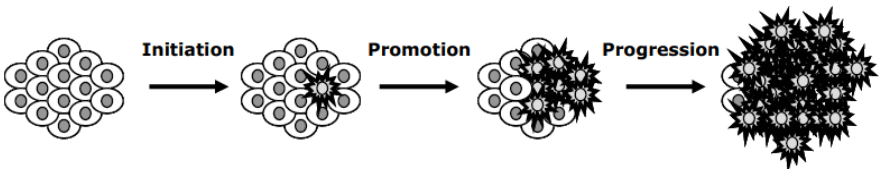
۱- مفاهیم مقدماتی

سالیان متمادی است که بشر می‌کوشد و علت وجودی آن‌ها را دریابد. بیماری‌های جسمی و روانی از جمله مسائل مهم انسانی هستند، که توجه بسیاری از متخصصین را در هر عصر و زمانی به خود معطوف

نموده‌اند. از کوشش برای ارتباط دادن بسیاری از بیماری‌ها به نیروهای ماوراءالطبیعه در سالیان دور تا کشف پنی‌سیلین و در سال‌های اخیر ظهور تکنولوژی مولکولی، جملگی تلاش‌هایی است که بشر در این راستا انجام داده است.

۱-۱- بیولوژی سرطان

سرطان چهارمین بیماری شایع و یکی از علل بسیار مهم مرگ‌ومیر در سراسر جهان است که هر ساله منجر به بیش از ۶ میلیون مورد مرگ می‌شود[۱]. سرطان بار مالی سنگینی به سیستم‌های پزشکی و ضایعات بسیار مهم‌تری به بیماران و خانواده‌های آن‌ها وارد می‌کند. سرطان توسط یک سری عوامل که باعث تغییر شکل سلول‌های نرمال به غیرنرمال و در نتیجه رشد و تقسیم غیرقابل کنترل آن‌ها می‌گردد، ایجاد می‌شود. همان‌گونه که شکل ۱-۱ نشان می‌دهد، این فرآیند شامل سه مرحله آغاز[۱]، پیشرفت[۲] و پیشرفت تصاعدی[۳] است.



شکل ۱-۱ مراحل پیشرفت سرطان

در مرحله‌ی اول، عواملی مانند ویروس، مشکلات محیطی (آلودگی‌هوا، مصرف دخانیات، تشعشعات و غیره) باعث جهش DNA[۴] می‌شود. مرحله‌ی پیشرفت شامل پروسه‌ی رشد و ازدیاد سلول‌های جهش‌یافته‌ای است که منجر به تومورهای کوچک می‌گردد. در مرحله‌ی پیشرفت تصاعدی این تومورها دچار رشد بیشتر و در نتیجه تخریب بافت اصلی می‌شوند [۱].

علی‌رغم پیشرفت‌های تکنولوژی طی یک دهه‌ی گذشته در زمینه‌ی سرطان موفقیت‌های محدودی برای مقابله با این بیماری ویران‌گر حاصل شده است. بنابراین تشخیص دقیق، دسته‌بندی و پیش‌بینی به‌موقع آن، از اهمیت به‌سزایی

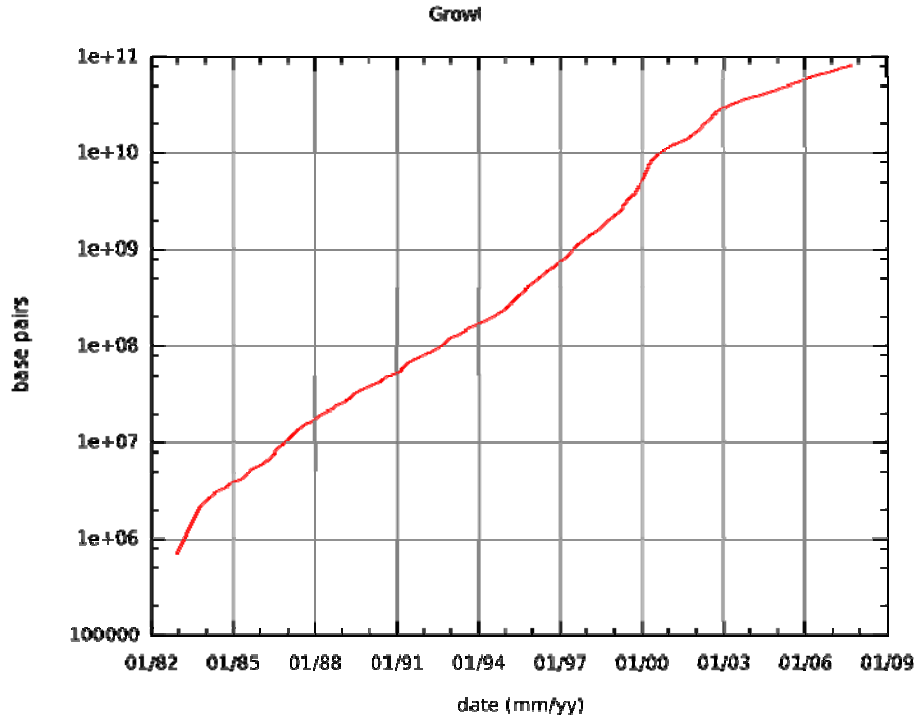
[1] Initiation
[2] Promotion
[3] Progression
[4] Deoxyribo Nucleic Acid

در مطالعات و تحقیقات برخوردار است. لذا به‌کارگیری مناسب تکنولوژی جدید به‌منظور پیش‌بینی و دسته‌بندی سرطان از اهمیت روزافزونی برخوردار شده است.

۱-۲- بیولوژی مولکولی

بیولوژی مولکولی به‌سرعت به سوی نظام‌های مبتنی بر داده در حرکت است و در نتیجه بر نقش ابزارها والگوریتم-های محاسباتی نیز روزبه‌روز افزوده شده تا علاوه بر بالا رفتن سرعت، با کاهش تعداد آزمایش‌ها برای یافتن پاسخ مسأله، در هزینه‌های آزمایشگاهی صرفه‌جویی شود. به‌تدریج روش‌های هوشمند با توجه به فضای جست‌وجوی بسیار بزرگ در برخی مسائل، به‌کارگرفته شده و امروزه شاهد استفاده از این الگوریتم‌ها در اغلب مسائل از جمله تشخیص و درمان بیماری‌ها به‌ویژه سرطان می‌باشیم که این کاربردها، مدام در حال افزایش است [۴].

بیولوژی مولکولی محاسباتی از شاخه‌های نسبتاً جدیدی است که با وجود نو بودن، توجه مشتاقان زیادی را به خود معطوف نموده است. اگرچه درک مکانیزم‌های حیات، از عرصه‌هایی است که همه‌روزه مرزهای آن فراتر می‌رود، اما با توجه به تاثیر عمیق محاسبات در تمامی شئون زندگی امروز، کنترل فرآیندهای حیات در سطح مولکولی، هنوز پتانسیل بالایی جهت پیشرفت دارد. همان‌گونه که اشاره شد، بیولوژی مولکولی به‌سوی نظام‌های مبتنی برداده درحرکت است و در نتیجه نقش ابزارها و الگوریتم‌های محاسباتی هر روزه پررنگ‌تر می شود. بیش از ده سال پیش، دانشمندان زیستی شروع به جمع‌آوری داده‌های بیولوژیک در سطح وسیعی نمودند و هم‌اکنون پروژه‌های توالی‌یابی ژنی و پروتئینی، میزان قابل توجهی از این داده‌ها را تشکیل می‌دهد. به‌عنوان مثال تا انتهای سال ۲۰۰۴، بیش از ۴۴میلیارد نوکلئوتید در ۴۰ میلیون رکورد از توالی‌های ژنی درGenBank لیست شده است و این میزان تقریباً هر۱۴ماه دو برابر می‌شود [۲]. شکل ۱-۲ نشان‌دهنده‌ی این موضوع است.



شکل۱-۲ نمودار رشد داده‌ها در پایگاه Gen Bank در سال‌های اخیر رشد بسیار بالا بوده و به‌طور متوسط هر ۱۴ ماه دو برابر شده است.

تحلیل دقیق این داده‌های حجیم و در حال افزایش، می‌تواند به درک صحیح فعالیت‌های زیستی در سطح مولکولی بیانجامد. سیستم‌های کامپیوتری با فراهم نمودن امکان ذخیره‌سازی و به اشتراک گذاشتن داده‌ها و سود جستن از سرعت بالا در محاسبات، نقش کلیدی را در این تحقیقات ایفا می‌کند. اگر چه با پیشرفت‌های اخیر در تکنولوژی، سرعت و ظرفیت این سیستم‌ها به‌طور عمده‌ای افزایش یافته است، لیکن تحلیل و ارائه این حجم از داده‌های زیستی و توانایی پیش‌بینی در مورد داده‌های آینده، ابزارهای قدرتمند و پیچیده‌تری را طلب می‌کند. در اینجا الگوریتم-های پیشنهادی هوش‌مصنوعی به مدد روش‌های سنتی آمده و راه‌حل‌هایی مؤثر برای مسائلی ارائه نموده که حل برخی از آن‌ها پیش از این (چه به‌دلیل کمبود دانش و چه به‌دلیل هزینه‌های بالای زمانی و حافظه‌ای در محاسبات) غیر عملی می‌نمود [۴].

۱-۲-۱- DNA، عامل انتقال اطلاعات وراثتی

قدم اول در بیولوژی مولکولی، درک صحیح فرآیند توارث است. بشر همیشه بر این باور بوده که در هنگام تولید-مثل، خصوصیات ظاهری از والد یا والدین به مولود منتقل می‌شود. حتی بسیاری از پیشینیان عقیده داشتند که فرزند، علاوه بر خصوصیات جسمی، خصوصیات روحی را نیز به ارث می برد [۲].

سلول کوچک‌ترین واحد زنده یک موجود زنده است و کلیه موجودات، طی فرآیند میتوز[1]، از تقسیم سلولی یک سلول به‌وجود می‌آیند. بدن متشکل از سلول‌های متفاوت با کارکردهای مختلف است. سلول، از غشاء سیتوپلاسمی، اندامک‌ها و هسته تشکیل شده است. کروموزوم‌ها درون هسته‌ی سلول قرار دارند که حاوی اطلاعات ژنتیکی (ژنوم یا مجموعه ژن‌ها) و متشکل از دو DNA می‌باشند. ژن‌ها تعیین‌کننده خصوصیات فیزیکی موجودات هستند به‌طوری‌که تغییرات کوچک در ژن‌ها باعث ایجاد تغییرات کوچک در ویژگی‌های فیزیکی می‌شود و محل قرار گرفتن آن‌ها در رشته DNA، loci یا مکان هندسی[2] نامیده می‌شود. به مجموعه کلیه اطلاعات ژنتیکی موجود در ژنوتیپ یک موجود، ژنوم گفته می‌شود [۲].

در قرون معاصر، Mendel اولین شخصی بود که پدیده‌ی فرآیند توارث را به‌شیوه جدید مورد بررسی قرار داد. وی در سال ۱۸۶۵یک مدل ریاضی جهت ارث‌بری ارائه داد و در آن، واحد اساسی توارث را ژن نام نهاد. کار مندل بعدها تقریباً به فراموشی سپرده شد تا اینکه در اوایل قرن بیستم با توجه به رشد سریع ریاضیات، دوباره بررسی‌هایی بر روی آن انجام شد. با تمام این‌ها، چیستی ژن تا سال ۱۹۴۴ که دانشمندان پی‌بردند ژن از ماده‌ای به‌نام DNA تشکیل شده است، هنوز برهمه پوشیده بود. در سال ۱۹۵۳ واتسون و کریک، ساختار مارپیچ دورشته‌ای معروف DNA را پیشنهاد دادند. این ساختار از آن جهت مورد توجه قرار گرفت که توانسته بود یک مدل فیزیکی برای بیان چگونگی تقسیم یک مولکول DNA به دو مولکول مشابه ارائه دهد. امروزه مشخص شده است که مولکولDNA، پلیمری از مولکول‌های کوچک‌تری به‌نام نوکلئوتیدها[3] است که اطلاعات مربوط به توارث، ناشی از ترتیب قرار گرفتن آن‌ها در کنار یکدیگر می-باشد (شکل ۱-۳). مارپیچ DNA درون یک سلول یک موجود زنده؛ این رشته پس از تا خوردن فراوان به‌صورت کروموزوم درون هسته سلول قرار می‌گیرد [۵].

[1] Mitoses
[2] Locus
[3] Nucleotide