DOR :20.1001.1.2538161.1401.7.2.1.8

دریافت مقاله: ۱۴۰۰/۱۲/۲۰
پذیرش مقاله:۱۴۰۱/۰۳/۰۹

نوع مقاله: پژوهشی

شبکه‌های عصبی پیچشی حساس به هزینه برای طبقه‌بندی زیرگروه‌های سرطان

راضیه هاشمی عالم
کارشناس ارشد مهندسی کامپیوتر، نرم‌افزار، دانشکده برق و کامپیوتر، صنعتی قم، قم، ایران
پست الکترونیکی: hashemialam.r@qut.qc.ir

محبوبه شمسی*
استادیار، دانشکدهٔ برق و کامپیوتر، صنعتی قم، قم، ایران
پست الکترونیکی:shamsi@qut.qc.ir

مجید آقایی
مربی، دانشکدهٔ برق و کامپیوتر، صنعتی قم، قم، ایران
پست الکترونیکی: aghaee@qut.qc.ir

چکیده

طبقه‌بندی زیرگروه‌های سرطان وظیفه بسیار مهمی در تشخیص و پیش بینی سرطان هاست. در سال‌های اخیر، روش‌های یادگیری عمیق به همین دلیل محبوبیت قابل‌توجهی به دست آورده‌اند. بااین‌حال، تعیین ساختار شبکه عصبی دشوار است زیرا عملکرد شبکه عمیق تا حد زیادی به ساختار آن بستگی دارد. علاوه بر این، تعداد بالای ژن‌ها در مجموعه داده بیان ژن و عدم تعادل داده‌ها بین رده‌های مختلف تأثیر مستقیمی بر پیچیدگی و عملکرد مدل‌های طبقه‌بندی زیرگروه سرطان دارد. برای پرداختن به مشکل داده‌های نامتعادل، یک مدل شبکه عصبی پیچشی با استفاده از یک راهبرد حساس به هزینه برای افزایش دقت مدل در شناسایی رده‌های اقلیت پیشنهادشده است. از سوی دیگر، از سه تکنیک نسبت فیشر، مجموعه‌های ناهنجار و ترکیبی برای کاهش ژن‌ها در مرحله پیش‌پردازش استفاده شده است. در روش حساس به هزینه، ماتریس هزینه بر اساس توزیع رده‌ها ایجاد شده و سپس از این ماتریس در مرحله تابع هزینه شبکه پیچشی برای محاسبه میزان خطا استفاده شده است. دو مجموعه از مجموعه داده‌های سرطان برای ارزیابی روش پیشنهادی استفاده شده است. نتایج با استفاده از چهار معیار دقت، فراخوانی، صحت و F1-Score مقایسه شده است. نتایج نشان می‌دهد که انتخاب ژن‌های مناسب و استفاده از یادگیری حساس به هزینه توانسته است عملکرد روش پیشنهادی را نسبت به مدل پیچشی بدون انتخاب ویژگی و یادگیری حساس به هزینه در حدود ۱۱٪، ۱۰٪، ۱۸٪، ۲۱٪، به ترتیب برای دقت، فراخوانی، صحت و F1-Score افزایش دهد.

واژه‌های کلیدی: دسته‌بندی، داده‌های نامتوازن، زیرگروه‌های سرطان، داده‌های بیان ژن، شبکه عصبی پیچشی، راهبرد حساس به هزینه

* نویسندهٔ مسئول

۱ – مقدمه

طبقه‌بندی زیرگروه‌های سرطان بر اساس داده‌های بیان ژن ابزار بسیار قدرتمندی جهت تحلیل رفتار هم‌زمان هزاران ژن است. وجود ژن‌های بسیار زیاد در طبقه‌بندی داده‌های حاصل از بیان ژن، باعث به وجود آمدن مشکلات زیادی در تحلیل این داده‌ها شده است. مشکلاتی از قبیل وجود نوفه و اطلاعات غیرمفید در برخی از ژن‌ها که نه‌تنها در تحلیل اطلاعات مفید نیستند، بلکه باعث طبقه‌بندی نادرست اطلاعات نیز می‌شوند. همچنین، اطلاعات مشابه در برخی از ژن‌ها باعث عدم اعتبار بعضی از روش‌های تحلیل داده‌ها می‌شود. علاوه بر این، مدل‌سازی شبکه‌های عصبی با داده‌های با ابعاد بالا باعث افزایش هزینه محاسباتی و پیچیدگی روش‌های طبقه‌بندی می‌شود. بااین‌حال، تنها مجموعه کوچکی از ژن‌ها عامل بیماری هستند، بنابراین حضور ژن‌های بسیار زیاد باعث کم‌رنگ شدن اثر ژن‌های عامل بیماری خواهند شد.

در سال‌های اخیر، روش‌های یادگیری ماشین و یادگیری عمیق برای یافتن ژن‌های مفید و همچنین طبقه‌بندی داده‌های بیان ژن مورداستفاده قرارگرفته‌اند. بااین‌حال، به دلیل اندازه بزرگ داده‌های بیان ژن، بسیاری از روش‌های قبلی از روش‌های آماری برای فیلتر کردن ژن‌ها استفاده می‌کردند. درواقع قبل از استفاده از روش‌های یادگیری ماشینی و یادگیری عمیق، ژن‌های نامربوط با تکنیک‌هایی مانند آزمون t حذف می‌شوند و تنها ژن‌هایی که طبقه‌بندی خوبی ارائه می‌دهند برای ساخت مدل استفاده می‌شوند. مسئله عدم توازن بر روی عملکرد دسته‌بندها تأثیر بسیاری دارد. در این میان، الگوریتم‌هایی که مسئلهٔ عدم توازن رده را در نظر نمی‌گیرند، تمایل دارند که توسط رده اکثریت تحت پوشش قرار داده شوند و در مقابل توسط رده اقلیت نادیده گرفته شوند. در مسائلی که سطح عدم توازن در آن‌ها زیاد است، برای طراحی یک دسته‌بند خوب می‌بایست سطح عدم توازن با دقت مدیریت شود. تکنیک‌هایی که مشکلات مربوط به مجموعه داده‌های نامتعادل را در برمی‌گیرند می‌توانند به دو گروه تقسیم شوند. دسته اول روش‌های سطح داده است که بر روی مجموعه آموزشی کار می‌کنند و توزیع رده‌ها را تغییر می‌دهند. دسته دیگر روش‌های سطح دسته‌بند (الگوریتمی) را در برمی‌گیرد. این روش‌ها مجموعه داده‌های آموزشی را بدون تغییر نگه می‌دارند و الگوریتم‌های آموزشی با توزیع داده‌ها تنظیم می‌شوند.

راهبرد یادگیری حساسیت به هزینه یکی از تکنیک‌های مبتنی بر سطح دسته‌بند است که هزینههای مرتبط با دسته‌بندی نادرست نمونه‌ها را در نظر می‌گیرد. در تشخیص سرطان، تفاوت هزینه بین خطاهای طبقه‌بندی اشتباه بسیار زیاد است. در یک سیستم تشخیص سرطان که در آن هر رده نشان‌دهنده این است که آیا یک فرد مبتلا به سرطان است یا نه، طبقه‌بندی اشتباه یک بیمار به‌عنوان یک فرد سالم منجر به هزینه بسیار بیشتری در مقایسه با طبقه‌بندی یک فرد سالم به‌عنوان یک بیمار خواهد شد. به این دلیل که تشخیص اشتباه ممکن است باعث تأخیر در درمان یا مرگ بیمار شود. یادگیری حساس به هزینه یک راهبرد برای به حداقل رساندن هزینه کلی یادگیری است که باعث می‌شود یک مدل یادگیری به‌گونه‌ای باشد که روند آموزش نسبت به رده‌هایی که هزینه کمتری دارند حساس‌تر باشد.

هنگام استفاده از یادگیری حساس به هزینه در مدل‌های یادگیری عمیق، فرایند آموزش نسبت به رده‌هایی که هزینه بالاتری دارند حساس‌تر است. برخی از تلاش‌های تحقیقاتی هزینه‌های خاص رده را در طبقه‌بندی‌کننده‌های یادگیری عمیق بررسی کرده‌اند. اولین بار یادگیری عمیق حساس به هزینه توسط چونگ[1] و همکاران معرفی شد [۸]، که هزینه‌ها را در تابع خطا[2] مرحله قبل از آموزش DNN[3] و CNN[4] ادغام کرد. وانگ[5] و همکاران [۹] با در نظر گرفتن میانگین خطا در هر رده ، عملکرد از دست دادن میانگین

1- Chung
2- Loss function
3- Deep Neural Network
4- Convolutional Neural Network
5- Wang

خطای مربع (MSE)[6] برای DNN را بهبود بخشید. خان و همکاران [۱۰] یک روش ابتکاری را برای اختصاص دادن هزینه به‌طور خودکار به هر رده با توجه به توزیع داده‌های رده، فرموله کرد. برخی از توابع ضرر که به‌طور گسترده مورداستفاده قرار می‌گیرند ، مانند MSE، کراس آنتروپی و SVM Hinge، با استفاده از یادگیری حساس به هزینه بهبود یافتند. تلیکانی و همکاران [۱۱] یک خو رمزگذار پشته شده (SAE)[7] حساس به هزینه ایجاد کرد که در آن هزینه‌ها در عملکرد از بین رفتن آنتروپی متقابل قرار گرفتند. برتری این روش نسبت به سایر رویکردهای یادگیری عمیق حساس به هزینه این است که نیازی به استفاده از ماتریس هزینه دست‌ساز نیست زیرا هزینه‌ها از طریق آمار داده‌ها تعیین می‌شود.

باقیمانده مقاله به شرح زیر سازماندهی شده است: بخش ۲ به‌طور خلاصه کارهای مرتبط در انتخاب ژن و رویکردهای طبقه‌بندی مبتنی بر یادگیری عمیق برای بیان ژن را خلاصه می‌کند. در بخش ۳، یک معماری CNN با استفاده از یادگیری حساس به هزینه معرفی شده است. بخش ٤ نتایج ارزیابی کارایی را ارائه می‌دهد و سپس نتیجه‌گیری در بخش ٥ استنتاج می‌شود و در بخش ٦ کارهای آتی پیشنهاد شد.

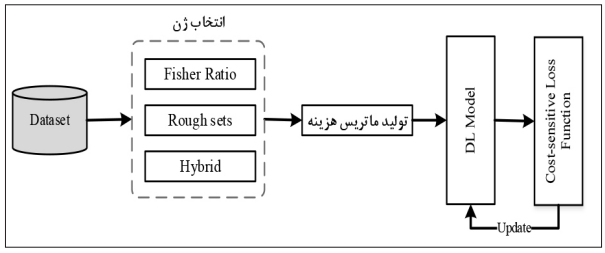
۲ – کارهای مرتبط

تحلیل داده بیان ژن، از جمله روش‌هایی است که در سال های اخیر جهت تنظیم پردازش‌های سلولی در سیستم‌های زیستی مورد توجه قرار گرفته است. در همین راستا با توجه به پیچیدگی‌های بیان ژن استفاده از روش‌های آماری و تکنیک‌های یادگیری ماشین و یادگیری عمیق بیش از پیش مورد توجه بوده است. از جمله این روش‌ها می‌توان به روش‌های یافتن رده نمونه‌های زیستی مرتبط، انتخاب ویژگی به منظور یافتن ژن‌های حاوی اطلاعات مفید و روش‌های طبقه‌بندی جهت تخصیص رده به نمونه‌های سلول با شرایط زیستی نامشخص اشاره نمود.

سه دسته برای انتخاب ژن وجود دارد که شامل فیلتر، بسته‌بندی و جاسازی شده است. در [۲] از آزمون t برای غلبه بر مشکل پراکندگی برای انتخاب ژن استفاده شد. لیائو و همکاران [۳] از آزمون مجموع رتبه ویلکاکسون به همراه ماشین بردار پشتیبان برای ارزیابی اهمیت ژن‌ها استفاده کرد. رویکردهای Wrapper از یک طبقه‌بندی کننده برای ارزیابی عملکرد یک زیرمجموعه ویژگی استفاده می‌کنند. K–نزدیک‌ترین همسایه [٤]، شبکه عصبی [٥] و ماشین بردار پشتیبان [٦] طبقه‌بندی‌کننده‌های پرکاربرد برای روش لفاف هستند. هو و همکاران [۷] مدل مجموعه خشن همسایگی را برای پردازش مجموعه داده‌های بیان ژن گسسته و پیوسته پیشنهاد کرد. یک مدل جنگل عصبی در [۱]، مجموعه‌ای از مدل درخت عصبی برای طبقه‌بندی زیرگروه‌های سرطان پیشنهاد شد. مدل جنگل پیشنهادی یک مسئله چند طبقه‌بندی را به بسیاری از مسائل طبقه‌بندی دودویی برای هر جنگل تبدیل می‌کند. یک رویکرد انتخاب ژن با ترکیب نسبت فیشر و مجموعه خشن همسایگی ایجاد شد. روش DeepGene [۷] یک شبکه عصبی عمیق بهبودیافته برای بیان ژن است. ابتدا از تکنیک Clustered Gene Filtering برای حذف ژن‌های نامربوط استفاده شد. سپس، طبقه‌بندی‌کننده DNN برای استخراج ویژگی‌های سطح بالا برای طبقه‌بندی استفاده شد.

در سال‌های اخیر، فناوری داده‌های بیان ژن، به طور گسترده‌ای در تشخیص‌های بالینی مورد استفاده قرار گرفته است. از مهم‌ترین مشکلات این داده‌ها، تعداد محدود نمونه‌ها در برابر ابعاد بالای آن (تعداد زیاد ژن‌ها) می‌باشد. بسیاری از این ژن‌ها جهت تشخیص بیماری‌ها مفید نیستند حتی ممکن است مانع از تشخیص درست نیز گردند. بنابراین یافتن ژن‌های حاوی اطلاعات مفید در بهبود تشخیص ناهنجاری‌های ژنتیکی با اهمیت است. بدین منظور، روش‌های داده‌کاوی، یادگیری ماشین و آماری به طور گسترده‌ای برای یافتن ژن‌های حاوی اطلاعات مفید به کار رفتند. بیشتر

6- Mean Squared Error (MSE)
7- Stacked Auto-Encoder (SAE)



شکل ۱: روش پیشنهادی

الگوریتم‌های ژنتیک و روش‌های استفاده شده در تحلیل داده‌های بیان ژن، تنها برای رتبه‌بندی اهمیت ژن‌ها مورد استفاده قرار می‌گرفتند و تعداد ژن‌های مورد نیاز را پیشنهاد نمی‌کردند. در این مقاله از روش شبکه‌های عصبی پیچشی حساس به هزینه برای طبقه‌بندی زیرگروه‌های سرطان استفاده شد. در واقع مشکلات بیش‌برازش و پیچیدگی محاسباتی بالا در روش‌های قبل و همچنین یک اجرای بازگشتی برای کاهش پیچیدگی‌های محاسباتی پیشنهاد شد.

این مقاله هم راهبردهای یادگیری حساس به هزینه و هم راهبرد کاهش ویژگی را برای رسیدگی به مشکل عدم تعادل طبقاتی و مشکلات ابعادی در طبقه‌بندی زیرگروه سرطان ادغام می‌کند. سه تکنیک نسبت فیشر، مجموعه‌های ناهنجاری و ترکیبی برای حذف ژن‌های نامربوط و غیرمفید در مرحله پیش‌پردازش استفاده می‌شوند. علاوه بر این، عملکرد تلفات متقابل آنتروپی با استفاده از راهبرد یادگیری حساس به هزینه با ادغام هزینه‌های مربوط به رده هنگام محاسبه ارزش تلفات در طول آموزش CNN، بهبود می‌یابد. در این راهبرد هزینه‌ها بر اساس آمار داده‌های دسته‌های سرطان تعریف می‌شود. این رویکرد باعث می‌شود که مدل CNN نسبت به رده‌های سرطان با فرکانس پایین حساس باشد و عملکرد مدل بیان ژن را در این نوع سرطان‌ها افزایش می‌دهد. آزمایش‌های مختلفی روی مجموعه داده GBM [8] انجام می‌شود و نتایج از نظر دقت، فراخوانی، صحت و F1-Score ارزیابی می‌شوند. نتایج نشان می‌دهد که چارچوب پیشنهادی ما می‌تواند عملکرد مدل CNN را برای بیان ژن، به‌ویژه برای سرطان‌های با فرکانس پایین، بهبود بخشد.

8- Glioblastoma Multiforme

۳ – متدولوژی پیشنهادی

در این بخش، یک مدل CNN مبتنی بر یادگیری حساس به هزینه معرفی می‌شود که شامل چهار مرحله است: انتخاب ژن، تولید ماتریس هزینه، مدل یادگیری ژرف، و تابع زیان حساس به هزینه. شکل ۱ مراحل روش پیشنهادی را نشان می‌دهد و در ادامه هریکی از مراحل به‌طور مفصل بحث می‌شود.

۳-۱- روش‌های انتخاب ژن: داده‌های بیان ژن به‌طورکلی از هزاران ژن تشکیل‌شده است، درحالی‌که تعداد نمونه‌های موجود اغلب اندک است. در میان هزاران ویژگی در داده‌های بیان ژن، تنها چند ژن درواقع با زیرگروه‌های سرطان همراه هستند درحالی‌که بقیه ممکن است به‌عنوان ویژگی‌های زائد یا عامل اغتشاش در نظر گرفته شوند. بنابراین، انتخاب ژن می‌تواند به‌عنوان یک مشکل کاهش ابعاد در نظر گرفته شود که سعی می‌کند ضمن حفظ دقت طبقه‌بندی ژن‌های اصلی، ژن‌های مهم را نیز انتخاب کند.

۳-۱-۱- ضریب فیشر

نسبت فیشر نسبت فواصل رده با فواصل طبقه‌بندی‌شده است. اگر دو طبقه‌بندی در یک مجموعه داده وجود داشته باشد، هر نمونه می‌تواند به‌عنوان $Y\epsilon\{-1, +1\}$ نشان داده شود چون داده‌های بیان ژن می‌توانند به عنوان $x_i = \{x_1^i, ..., x_n^i\}$ باشند برای هر ژن میزان انحراف استاندارد $\sigma_i^+(resp, \sigma_i^-)$ و انحراف میانگین $\mu_i^+(resp, \mu_i^-)$ محاسبه‌شده و میزان ضریب فیشر نیز از رابطه ۱محاسبه می‌شود:

$$F_i = \frac{(\mu_i^+ - \mu_i^-)^2}{(\sigma_i^+) + (\sigma_i^-)^2} \qquad (۱)$$

ژن با بالاترین مقدار F_i آموزنده‌ترین است و میزان بیشترین تفاوت داده‌های بیان ژن را در دو رده به‌طور متوسط نشان می‌دهد درحالی‌که رده‌های مربوطه به ژن‌هایی که دارای انحراف کوچک هستند، بیشترین اختلاف را نشان می‌دهند. سپس ژن‌هایی با مقادیر F_iبالا به‌عنوان ویژگی‌های برتر انتخاب می‌شوند.

۳-۱-۲- مجموعه ناهنجاری‌های مجاور

یک مدل مجموعه ژن‌های ناهنجار مجاور NRS[9]، می‌تواند برای پردازش مجموعه داده‌های گسسته و مداوم در عین حفظ اطلاعات لازم برای طبقه‌بندی دقیق داده‌ها استفاده شود. با توجه به مجموعه‌ای از نمونه‌ها $U = \{x_1, x_2, ..., x_n\}$، با مجموعه‌ای از ویژگی‌های نوع واقعی توصیف U و D یک شاخص تصمیم‌گیری است. اگر A یک خانواده از ژن‌های مجاورش را در دامنه تولید کند، به آن $NDT = \{U, A, D\}$ که یک سیستم تصمیم‌گیری مجاور است گفته می‌شود. اگر D تقسیم کند U را نسبت به N طبقه هم‌ارز داریم:

$X_1, X_2, ..., X_N, \forall B \subseteq A$، سپس تقارن‌های پایین و بالا تصمیم D با توجه به B می‌تواند به شرح زیر باشد:

$$\underline{N}_B D = \bigcup_{I=1}^{N} \underline{N}_B X_i$$

$$\bar{N}_B D = \bigcup_{I=1}^{N} \bar{N}_B X_i$$

که در آن $\underline{N}_B X = \{x_i \mid \delta_B(x_i) \subseteq X, x_i \varepsilon U\}$ $\bar{N}_B X = \{x_i \mid \delta_B(x_i) \cap X \neq \phi, x_i \varepsilon U\}$ $\delta_B(x_i)$ است یک بخش از اطلاعات مجاور که تولید می‌شود با شاخص B و اندازهٔ۱. تقریب پایین‌تر از تصمیم D، دامنه تصمیم مثبت نیز نامیده می‌شود که به‌عنوان POSB(D) مشخص می‌شود.

اندازه دامنه مثبت میزان جداشدگی مشکل طبقه‌بندی را در یک فضای مشخص نشان می‌دهد. دامنه مثبت هر چه بزرگ‌تر باشد، همپوشانی مرزهای رده کمتر خواهد بود. این عمل با استفاده از مجموعه ویژگی‌های انتخاب شده، توصیف بهتری از یک مسئله طبقه‌بندی را تضمین می‌کند. بنابراین، وابستگی شاخص تصمیم D به ویژگی شرط B شناخته می‌شود به عنوان:

$$\gamma_B(D) = Card(\underline{N}_B D) / Card(U)$$

با توجه به یک سیستم تصمیم گیری مجاور $NDT = \{U, A, D\} B \subseteq A, \forall_a \varepsilon A - B$، اهمیت یک نسبت به B را می‌توان تعریف کرد:

$$SIG(a, B, D) = \gamma_B \cup a(D) - \gamma_B(D)$$

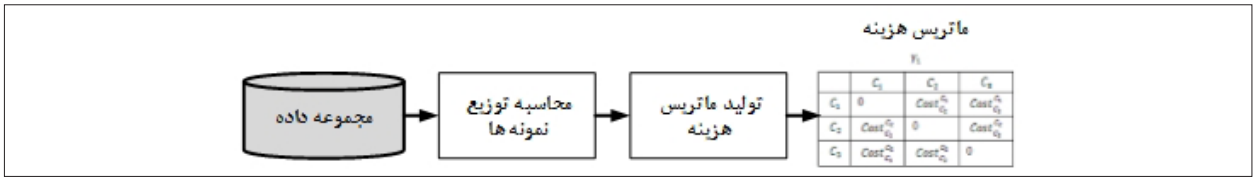
بر اساس شاخص اهمیت ویژگی، از یک شاخص گردکردن در الگوریتم کاهشی استفاده می‌شود. این الگوریتم مجموعه خالی را به عنوان نقطه شروع در نظر می‌گیرد، هر بار شاخص اهمیت همه خصوصیات باقیمانده را محاسبه می‌کند و ویژگی را با بزرگ‌ترین مقدار شاخص اهمیت برای پیوستن به مجموعه کاهنده انتخاب می‌کند. این فرآیند تا زمانی که تمام خصوصیات باقیمانده دارای اهمیت ۰ باشند تکرار می‌شود. این بدان معنی است که اگر خصوصیات جدیدی اضافه شوند، مقادیر عملکرد وابسته به سیستم دیگر تغییر نمی کنند. الگوریتم جستجوی پیش رونده اطمینان می‌دهد که ابتدا ویژگی‌های مهم به مجموعه کاهش اضافه می‌شوند، بنابراین ویژگی‌های مهم حذف نمی‌شوند.

۳-۱-۳- روش ترکیبی

روش نسبت فیشر می‌تواند به طور موثری با ناهنجاری در داده‌های بیان ژن مقابله کند و ژن‌های ناهنجار را با توجه به سهم خود در طبقه‌بندی، متلاطم کند و بنابراین به طور مؤثر به شناسایی ژن‌های زیرگروه سرطان کمک می‌کند. مجموعه ژن‌های ناهنجار مجاور دارای ویژگی‌های عدم نیاز به تفسیر داده‌های مداوم است و از دست دادن اطلاعات ناشی از گسسته سازی داده‌ها جلوگیری می‌کند، که می‌تواند ژن‌های اضافی را از بین ببرد. اگر فقط نسبت فیشر را به عنوان روش انتخاب ژن استفاده کنیم، ویژگی‌های رویه K انتخاب می‌شوند. نسبت فیشر رابطه بین ژن‌ها را در نظر نمی‌گیرد و ممکن است ژن‌های زائد با همبستگی زیاد را انتخاب کند، که نه تنها میزان محاسبه را افزایش می‌دهد بلکه منجر به نتایج نادرست طبقه‌بندی نیز می‌شود. هنگامی که مجموعه ژن‌های ناهنجار مجاور مستقیماً برای از بین بردن ژن‌های زائد استفاده می‌شود، با افزایش تعداد ژن‌ها، می‌توان هزینه محاسباتی الگوریتم را بالاتر برد.

۳-۲- تولید ماتریس هزینه: برای آموزش مدل CNN با استفاده از هزینه‌های مربوط به دسته‌بندی‌های مختلف، ایجاد یک ماتریس هزینه ضروری است. این ماتریس در تابع هزینه برای محاسبه مقدار خطای دسته‌بندی استفاده

9 -Neighborhood Rough Set



شکل۲: فرآیند تولید ماتریس هزینه

می‌شود. برخلاف بسیاری از روش‌های قبلی برای تولید ماتریس هزینه که به‌صورت دستی و از طریق کاربر/ متخصص وزن مربوط به هر دسته‌بندی تعیین می‌شود، روش پیشنهادی از یک مکاشفه استفاده می‌کند تا هزینه‌ها به‌صورت خودکار و بدون دخالت کاربر مشخص شوند. این هزینه‌ها با در نظر گرفتن توزیع رده‌ها تعیین می‌شوند. شکل ۲ فرآیند تولید ماتریس هزینه ۷ را نشان می‌دهد.

در مرحله اول، توزیع هر رده در مجموعه داده محاسبه می‌شود تا برای تولید ماتریس هزینه استفاده شود. برای تولید ماتریس هزینه، یک فرموله سازی مبتنی بر توزیع داده‌ها انجام می‌شود. هزینه بالاتر طبقه‌بندی نادرست برای رده‌های اقلیت در نظر گرفته می‌شود درحالی‌که هزینه طبقه‌بندی پایین‌تری برای رده‌های اکثریت تعیین می‌شود. هزینه طبقه‌بندی نادرست رده i در رده j با استفاده از رابطه ۲ محاسبه می‌شود

$$\begin{cases} \gamma_{i,j} = \frac{\alpha_i}{\alpha_i+\alpha_j} \; i,j = 1,2,\dots,C \\ \quad subject\ to\ i \neq j \end{cases} \qquad (۲)$$

$$\begin{cases} \gamma_{i,j} = \frac{\alpha_i}{\alpha_i+\alpha_j} \; i,j = 1,2,\dots,C \\ \quad subject\ to\ i \neq j \end{cases}$$

در یک ماتریس هزینه، سطر مورب ماتریس به‌عنوان بردار سودمندی شناخته می‌شود. این بردار طبقه‌بندی‌های صحیح را نشان می‌دهد و به صفر تنظیم می‌شود. همچنین، تمام هزینه‌ها غیرمنفی هستند، یعنی $\gamma_{i,j} > 0$. در این رابطه، α_i و α_j به ترتیب تعداد نمونه‌های رده‌های i و j هستند. جدول ۱ مثالی از ماتریس هزینه را برای طبقه‌بندی سه طبقه نشان می‌دهد. یک ماتریس ۳×۳ به‌گونه‌ای ایجاد می‌شود که تمام سلول‌های ماتریس بزرگ‌تر از صفر است غیر از آن‌هایی که در سطر مورب هستند که همیشه صفر هستند. این بدان معنی است که وقتی الگوریتم نمونه را به‌درستی طبقه‌بندی می‌کند، هیچ هزینه‌ای وجود ندارد. در

جدول ۱: یک نمونه از ماتریس هزینه با سه رده

	Predicted C_1	Predicted C_2	Predicted C_3
Actual C_1	۰	$\gamma_{1,2}$	$\gamma_{1,3}$
Actual C_2	$\gamma_{2,1}$	۰	$\gamma_{2,3}$
Actual C_3	$\gamma_{3,1}$	$\gamma_{3,2}$	۰

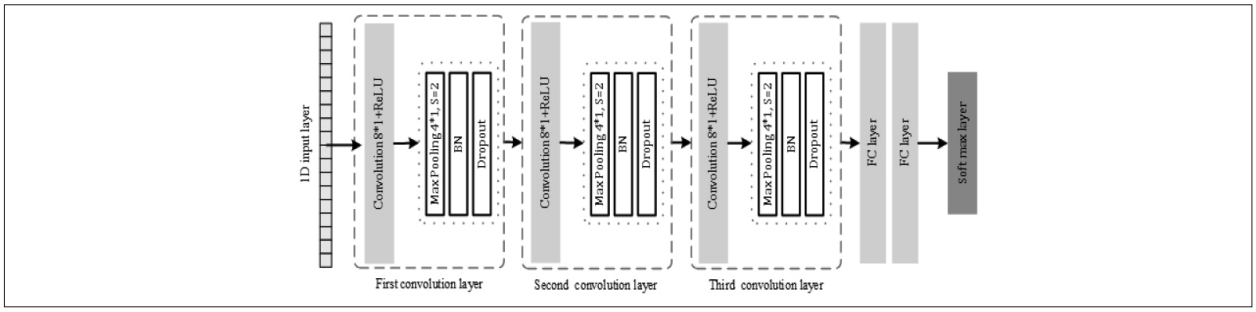
غیر این صورت، الگوریتم ما بسته به هزینه اختصاص داده‌شده در ماتریس هزینه یک هزینه برای طبقه‌بندی نادرست اختصاص می‌دهد.

در این رابطه، α_i و α_j به ترتیب تعداد نمونه‌های رده‌های i و j هستند. شبه کد مربوط به مرحله تولید ماتریس هزینه در الگوریتم ۱ نشان داده شده است.

Algorithm 1: Cost matrix generation

Input: y_train, n_classes
Output: cost_matrix γ
1: Begin
2: γ ← Initialize with zeros
3: α ← Compute frequency of classes
4: For each i ∈ labels
5: For each j ∈ labels
6: if i≠j
7: $\gamma_{i,j} = \frac{\alpha_i}{\alpha_i+\alpha_j}$

۳-۳-معماری شبکه عصبی پیچشی: این بخش معماری CNN برای روش پیشنهادی را تشریح می‌کند (شکل ۳) که دارای یک لایه ورودی یک‌بعدی و سه لایه پیچشی است که هر یک از آن‌ها دارای یک پیچش و به دنبال آن لایه‌های تابع فعال‌ساز ReLU و ادغام حداکثری است. اندازه فیلتر برای لایه پیچش ۱×۸ و stride = 1 است و هر لایه ادغام حداکثری یک ورودی ۱×٤ را با stride = 2 پردازش می‌کند. بعد از هر لایه ReLu، از نرمال‌سازی دسته‌ای و Dropout با نسبت ۰/۰۵ استفاده می‌شود. بعد از لایه‌های پیچش، دولایه اتصال کامل برای طبقه‌بندی ژن استفاده شد.



شکل۳: معماری CNN حساس به هزینه

۳-۴- تابع هزینه حساس به هزینه: در این بخش یک تابع هزینه حساس به هزینه پیشنهاد می‌شود که نسبت به طبقه‌بندی نادرست رده‌های اقلیت حساس‌تر است. در طول آموزش، روش یادگیری پیشنهادی به‌طور مشترک هزینه‌های وابسته به رده و پارامترهای شبکه عصبی را بهینه می‌کند. در مقایسه با رویکردهای سطح داده (نمونه‌برداری مجدد)، روش پیشنهادی توزیع داده‌های اصلی را تغییر نمی‌دهد، که درنتیجه هزینه‌های محاسباتی پایین‌تر در طول فرآیند آموزش به دست می‌دهد. علاوه بر این، برخلاف روش‌های حساس به هزینه که از یک ماتریس هزینه دستی که بر اساس نظر یک متخصص تعیین می‌شود، در روش پیشنهادی هزینه‌های مرتبط با هر رده به‌صورت خودکار با استفاده از توزیع داده‌ها در طول فرآیند یادگیری تنظیم می‌شود.

هدف ما مجازات کردن انواع خطاهای طبقه‌بندی بر اساس برخی هزینه‌های تعیین‌شده است. این مقدار جریمه برای زمانی که نمونه اقلیت به‌عنوان رده اکثریت طبقه‌بندی می‌شود بیشتر از زمانی است که نمونه اکثریت به‌اشتباه به‌عنوان رده اقلیت طبقه‌بندی می‌شود. همان‌طور که در بخش قبلی اشاره کردیم، رده‌های اقلیت و اکثریت تعیین می‌شوند و فقط باید هزینه مربوطه را از ماتریس هزینه پیدا کنیم. برتری الگوریتم ما این است که تعیین نوع رده‌ها ازنظر اقلیت یا اکثریت لازم نیست. درواقع، هزینه‌ها فقط بر اساس توزیع رده‌ها اختصاص می‌یابد. این ویژگی کمک میکند تا الگوریتم در هر مجموعه داده استفاده شود.

این رویکرد قصد دارد با در نظر گرفتن مقادیر هزینه مربوط به هر نوع طبقه‌بندی نادرست ، تابع هزینه cross–entropy اصلاح کند. این روش باعث حساسیت بیشتر مدل CNN نسبت به طبقه‌بندی نادرست رده‌های اقلیت می‌شود. درواقع، خروجی لایه Softmax که به شکل احتمالات است، به‌عنوان ورودی تابع هزینه در نظر گرفته می‌شود تا مقدار زیان حساس به هزینه محاسبه شود. دلیل انتخاب cross-entropy این است که می‌تواند در بیشتر موارد نسبت به توابع هزینه دیگر عملکرد بهتری داشته باشد. علاوه بر این، cross-entropy می‌تواند از کاهش سرعت یادگیری که یکی از مشکلات تابع میانگین خطای مربع (MSE) در یادگیری است جلوگیری کند.

قبل از تشریح راهبرد تابع زیان حساس به هزینه پیشنهادی، نحوه عملکرد لایه Softmax توضیح داده می‌شود. فرض کنید لایه خروجی به‌صورت $\{X,Y\}=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_C)\}$ باشد، که $x_i\in\mathbb{R}^{d\times1}$ و $y_i\in\mathbb{R}^{C\times1}$ هستند. اصطلاح d اندازه لایه خروجی و C تعداد رده‌ها هستند. تابع Softmax احتمال اینکه نمونه i (x_i) متعلق به یک رده باشد را محاسبه می‌کند:

$$f_\theta(x)=\frac{1}{\sum_{j=1}^{C}e^{y_i}}\begin{bmatrix}e^{y_1}\\e^{y_2}\\\dots\\e^{yC}\end{bmatrix}=\begin{bmatrix}p(y_i=1|x_i)\\p(y_i=2|x_i)\\\dots\\p(y_i=C|x_i)\end{bmatrix} \qquad (۳)$$

متغیر θ پارامتر نگاشت برای رده j است (b_j+W_jx). رویکرد پیشنهادی در این پژوهش مجازات کردن طبقه‌بندی نادرست در تابع هزینه cross-entropy بر اساس هزینه‌های تعیین‌شده در ماتریس هزینه (۷) برای به حداکثر رساندن نزدیکی پیش‌بینی به رده واقعی را

مدنظر قرار می‌دهد. مقدار کل هزینه هر دسته با N نمونه با استفاده از معادله ٤ محاسبه می‌شود:

$$\mathcal{L}(O,y) = -\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(O_j,y_j) \qquad (٤)$$

که در آن مقدار cross-entropy میانگین مقادیر زیان برای کل N طبقه‌بندی است. مقدار زیان برای هر پیش‌بینی توسط معادله ٥ محاسبه می‌شود:

$$\mathcal{L}(O_i,y_i) = -\sum_{i=1}^{C}\left(y_{o,c}\log p(y_i = 1|x_i;\theta_i)\right) \qquad (٥)$$

در این رابطه، $y_{o,c}$ یک شاخص دودویی (۰ یا ۱) است که به پیش‌بینی صحیح مشاهده برای نمونه o اشاره دارد. مقدار $y_{o,c}$ برای رده اشتباه پیش‌بینی‌شده ۱ و برای رده واقعی ۰ است. احتمال طبقه‌بندی اشتباه با در نظر گرفتن هزینه مربوط به رده تغییر می‌یابد (معادله ٦):

$$p(y_i = 1|x_i) = \frac{\gamma_{i,j}\exp(O_i)}{\sum_{i=1}^{C}\exp(O_i)} \qquad (٦)$$

بر اساس معادله ٥، ضرب هزینه مربوط به رده‌های اقلیت، مقدار احتمال جدید را به‌شدت کاهش می‌دهد و بنابراین، منجر به افزایش مقدار زیان طبقه‌بندی در رابطه ٦ می‌شود. به‌این‌ترتیب، رده‌های اقلیت بیشتر از رده‌های اکثریت بر روی تابع هزینه تأثیر می‌گذارند. الگوریتم ۲ شبه‌کد تابع هزینه cross-entropy حساس به هزینه (CSCE[10]) را نشان می‌دهد که برای مدل CNN حساس به هزینه طراحی شده است.

Algorithm 2: Cost-sensitive cross-entropy (CSCE)

Input: cost matrix (γ), Actual values (y_A), Predicted values (y_p)
Output: Loss value $\mathcal{L}$
1: Begin
2: $\mathcal{L} \leftarrow 0$
3: For each i $\in N$
4: $loss_i = y_{Ai} + \log(y_{pi} \times \gamma_{ij})$
5: $\mathcal{L} \leftarrow \mathcal{L} + loss_i$
6: Return $\mathcal{L}$/N
7: End

۴ – نتایج ارزیابی

در این بخش، عملکرد مدل CNN حساس به هزینه پیشنهادشده با مدل‌های دیگر مقایسه می‌شود. این مدل‌ها شامل نسخه‌های غیرحساس برای CNN است. کتابخانه Keras و Tensorflow به‌عنوان Backend برای اجرای مدل‌های DL مورداستفاده قرار گرفتند. تمام مدل‌ها با ۱۰۰ دوره آموزش‌دیده بودند. راهبرد توقف اولیه برای جلوگیری از مشکل بیش‌برازش مورداستفاده قرار گرفت که در آن زمانی که مقدار خطا بر روی‌داده‌های اعتبارسنجی برای چندین دوره تغییر نکرده باشد، فرایند آموزش متوقف می‌شود. از تابع Adam به‌عنوان بهینه‌ساز برای شبکه‌های عصبی استفاده می‌شود. در تمام آزمایش‌ها، ۸۰٪ داده‌ها به‌عنوان مجموعه آموزشی، ۱۰٪ به‌عنوان مجموعه اعتبارسنجی و ۱۰٪ به‌عنوان مجموعه آزمایشی استفاده شد.

۴-۱-معیارهای ارزیابی: برای ارزیابی روش پیشنهاد، از چهار معیار دقت (رابطه ۷)، فراخوانی (رابطه ۸)، صحت (رابطه ۹)، و F1-Score (رابطه ۱۰) استفاده می‌شود.

$$\text{Accuracy} = \frac{TP+TN}{FP+FN+TP+TN} \qquad (٧)$$

در مسائل واقعی که در آن داده‌ها اغلب نامتوازن هستند، دقت نمی‌تواند معیار مناسبی برای ارزیابی کارایی الگوریتم‌های طبقه‌بندی باشد، به این دلیل که در رابطه دقت، ارزش نمونه‌ها برای دسته‌های مختلف یکسان در نظر گرفته می‌شوند. فراخوانی[11] و صحت[12] دو معیار دیگر برای ارزیابی عملکرد دسته‌بند هستند و به ترتیب در رابطه‌های ۸ و ۹ نشان داده‌شده‌اند.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (٨)$$

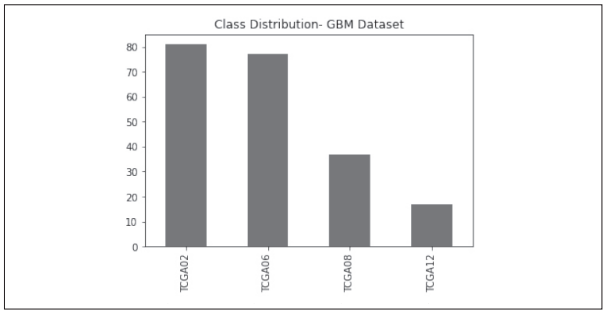
$$\text{Precision} = \frac{TP}{TP+FP} \qquad (٩)$$

در بعضی موارد ممکن است تمرکز بر روی حداکثر کردن فراخوانی یا صحت باشد. بااین‌حال، در مواردی که پیدا کردن یک ترکیب مطلوب بین صحت و فراخوانی مدنظر باشد، دو معیار می‌توانند با استفاده از معیار F1 Score ترکیب شوند. رابطه ۱۰ ترکیبی از دو رابطه فراخوانی و صحت را نشان می‌دهد.

$$\text{F1 Score} = \frac{2*Recall*Percision}{Recall+Percision} \qquad (١٠)$$

10- Cross-Entropy loss function (CSCE)

11- Recall
12- Percision



شکل ۴: توزیع نمونه‌ها برای مجموعه داده GBM

۴-۲- مجموعه داده‌ها: در این پژوهش، مجموعه داده زیرگروه سرطان بیان ژن RNA-Seq برای GBM (گلیوبلاستوما چندحالت[۱۳]) برای آزمایش‌ها استفاده شد. مجموعه داده GBM دارای ۲۱۲ نمونه سرطانی با حدود ۱۲ هزار ژن است که در چهار گروه ژنی دسته‌بندی می‌شوند. در مجموعه داده GBM دو رده TCGA02 و TCGA06 دارای توزیع یکسانی هستند (حدود ۸۰ نمونه) و دو رده دیگر دارای توزیع اقلیتی هستند.

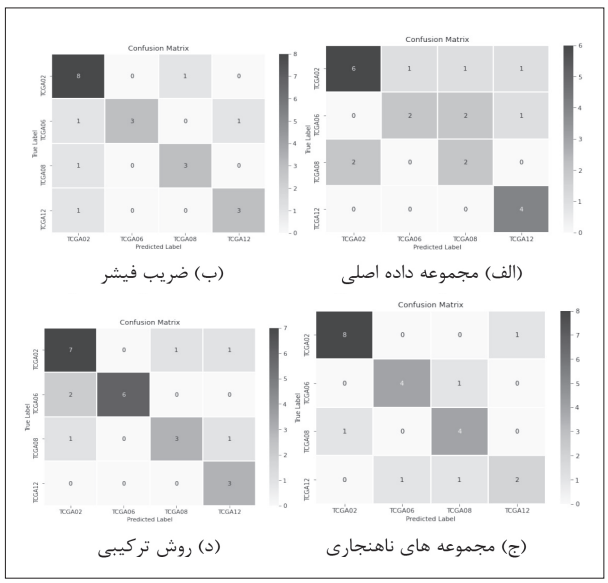
۴-۳- نتایج و بحث: در این بخش عملکرد روش پیشنهادی بر روی مجموعه داده GBM و همچنین نسخه متوازن شده مجموعه داده GBM بررسی می‌شود. جدول ۲ عملکرد روش پیشنهادی بر روی مجموعه داده GBM بر اساس سه معیار فراخوانی، صحت، و F1 Score را نشان می‌دهد. روش فیشر با ژن‌های اصلی مقایسه شده‌اند. همان‌گونه که در جدول مشاهده میشود، سه روش فیشر، ناهنجاری و ترکیبی با ژن‌های اصلی مقایسه شده‌اند. همان‌گونه که در جدول مشاهده می‌شود، روش پیشنهادی بر روی ژن‌های انتخاب شده توسط تکنیک ناهنجاری بهترین عملکرد را داشته است. همچنین شکل ۵ ماتریس‌های آشفتگی برای روش پیشنهادی بر روی داده اصلی را نشان می‌دهد.

شکل ٦ میانگین عملکرد مدل‌های بیان ژن را نشان می‌دهد. بدترین عملکرد برای ژن‌های منتخب با تکنیک نسبت فیشر بود. در مقابل، بالاترین عملکرد مربوط به کاربرد روش پیشنهادی بر روی مجموعه داده انتخابی با استفاده از تکنیک ترکیبی بود.

13- Glioblastoma Multiforme

جدول ۲: عملکرد روش پیشنهادی برای مجموعه داده GBM

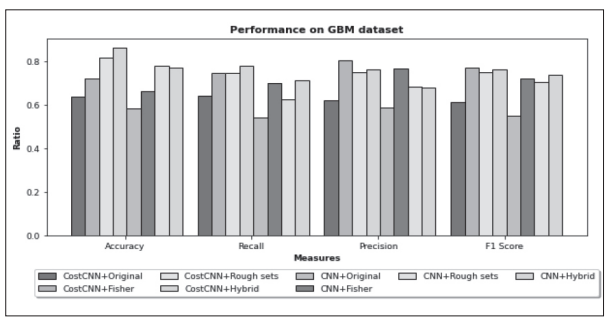
رده	انتخاب ژن	فراخوانی	صحت	F1-score
TCGA02	مجموعه اصلی	۰,۶۷	۰,۷۵	۰,۷۱
	فیشر	۰,۸۸	۰,۷۲	۰,۷۹۲
	مجموعه ناهنجاری‌های مجاور	۰,۸۸	۰,۸۸	۰,۸۸
	ترکیبی	۰,۷۷	۰,۷	۰,۷۴
TCGA06	مجموعه اصلی	۰,۴	۰,۶۷	۰,۵
	فیشر	۰,۶	۱,۰	۰,۸
	مجموعه ناهنجاری‌های مجاور	۰,۸	۰,۸	۰,۸
	ترکیبی	۰,۷۵	۱,۰	۰,۸۵۷
TCGA08	مجموعه اصلی	۰,۵	۰,۴	۰,۴۴
	فیشر	۰,۷۵	۰,۷۵	۰,۷۵
	مجموعه ناهنجاری‌های مجاور	۰,۸	۰,۶۶	۰,۷۵۴
	ترکیبی	۰,۶	۰,۷۵	۰,۶۶
TCGA12	مجموعه اصلی	۱,۰	۰,۶۷	۰,۸
	فیشر	۰,۷۵	۰,۷۵	۰,۷۵
	مجموعه ناهنجاری‌های مجاور	۰,۵	۰,۶۶	۰,۵۶۸
	ترکیبی	۱,۰	۰,۶	۰,۸



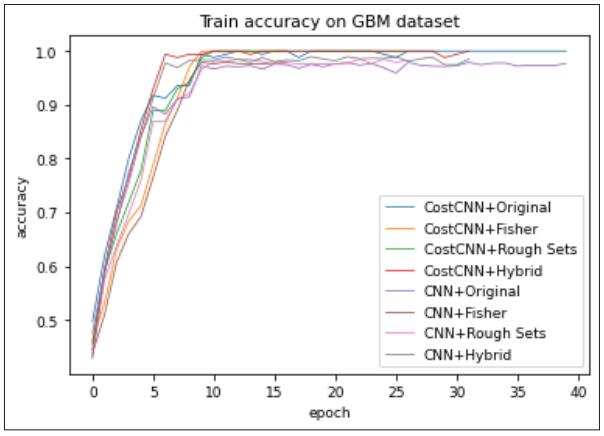
(الف) مجموعه داده اصلی (ب) ضریب فیشر

(ج) مجموعه های ناهنجاری (د) روش ترکیبی

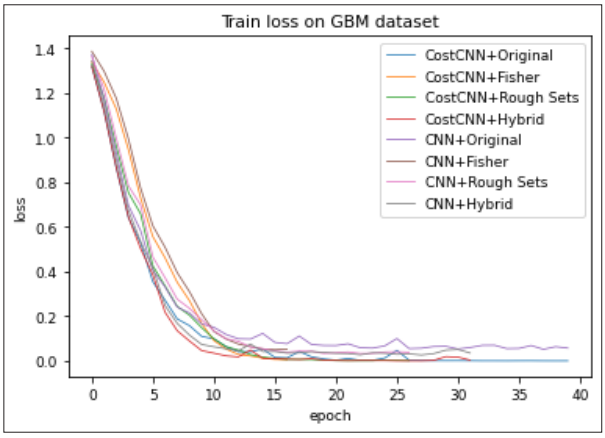
شکل ۵: ماتریس آشفتگی برای مدل با مجموعه ژن‌های انتخاب‌شده متفاوت برای مجموعه داده GBM اصلی



شکل ۶: عملکرد مدل‌های یادگیری ژرف برای دسته‌بندی



شکل ۷: دقت آموزش



شکل ۸: خطای آموزش

شکل‌های ۷ و ۸ تأثیر تعداد دوره‌ها را بر دقت آموزشی و از دست دادن مدل پیشنهادی در مقایسه با روش مشابه بر روی مجموعه داده‌های GBM با تکنیک‌های مختلف انتخاب ژن نشان می‌دهند. مشاهده می‌شود که مدل‌ها در تکرار سی‌ام به حداکثر دقت رسیده‌اند. روش پیشنهادی از دقت بالاتری نسبت به روش مشابه برخوردار بود و با عدد دوره کمتر همگرا شد.

۵- نتیجه‌گیری

یادگیری عمیق (DL) یک تکنیک پرکاربرد در ناحیه بیان ژن است. بااین‌حال، این الگوریتم‌ها در مورد داده‌های با ابعاد بالا و همچنین داده‌های نامتعادل با چالش‌های زیادی روبرو هستند. تعداد زیاد ویژگی‌ها، پیچیدگی مدل‌های یادگیری عمیق را افزایش می‌دهد و همچنین عدم تعادل بین رده‌های سرطان، عملکرد مدل طبقه‌بندی را کاهش می‌دهد. برای رسیدگی به این چالش‌ها، یک رویکرد CNN پیشنهاد شد که با تکنیک انتخاب ژن ادغام شد. یک راهبرد حساس به هزینه نیز برای مقابله با داده‌های نامتعادل استفاده شد. در این راهبرد، طبقه‌بندی‌های اشتباه مختلف دارای هزینه‌های مشخصی هستند که در هنگام محاسبه میزان خطا اعمال می‌شود و برای بهینه‌سازی پارامترهای شبکه استفاده می‌شود. در روش پیشنهادی، ماتریس هزینه با استفاده از یک تابع فرمول‌بندی شده تعیین می‌شود. برای ارزیابی روش پیشنهادی از مجموعه داده استفاده شد. معیارهای دقت، فراخوانی، صحت و F1-Score برای مقایسه روش پیشنهادی با مدل‌های مشابه استفاده شد. اجرای حساس به هزینه CNN با نسخه‌های غیر حساس مقایسه شد. نتایج نشان داد که روش پیشنهادی توانایی تشخیص بالاتری برای طبقات اقلیت دارد. به‌طور متوسط، مدل پیشنهادی عملکرد تشخیص سرطان زیرگروه را حدود ۳٪ افزایش داده است.

۶- کارهای آتی

در آینده، برای بهینه‌سازی هزینه‌های طبقه‌بندی نادرست در داده‌های آموزشی، روش‌های دیگری همانند یادگیری ژرف حساس به هزینه تکاملی برای تشخیص سرطان می‌تواند ارائه شود. در این روش، هزینه‌های دسته‌بندی در طول فرآیند آموزش توسط رویکردهای فراابتکاری بهینه می‌شوند. در این مقاله، تابع Cross-entropy برای عدم تعادل رده بهبود داده شد، با این‌حال، سایر توابع هزینه همانند SVM Hinge Loss و MSE می‌توانند با استفاده از راهبرد حساس به هزینه بهبود یابند.

۷- منابع

[1] Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H. and Khan, M.M., 2019. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. IEEE Access, 7, pp.22086-22095.

[2] Zhu, S., Wang, D., Yu, K., Li, T. and Gong, Y., 2008. Feature selection for gene expression using model-based entropy. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(1), pp.25-36.

[3] Liao, C., Li, S. and Luo, Z., 2006, November. Gene selection using wilcoxon rank sum test and support vector machine for cancer classification. In International Conference on Computational and Information Science (pp. 57-66). Springer, Berlin, Heidelberg.
[4] L. Li, C. Weinberg, T. Darden, and L. G. Pedersen, ``Gene selection for sample classi_cation based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method,'' Bioinformatics, vol. 17, no. 12, pp. 1131_1142, Dec. 2001.
[5] J. Khan et al., ``Classi_cation and diagnostic prediction of cancers using gene expression pro_ling and arti_cial neural networks,'' Nature Med., vol. 7, no. 6, pp. 673_679, 2001.
[6] I. Guyon, J.Weston, S. Barnhill, and V. Vapnik, ``Gene selection for cancer classification using support vector machines,'' Mach. Learn., vol. 46, nos. 1_3, pp. 389_422, 2002.
[7] Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z. and Feng, D.D., 2016. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC bioinformatics, 17(17), pp.243-256.
[8] Y.-A. Chung, H.-T. Lin, and S.-W. Yang, "Cost-aware pre-training for multiclass cost-sensitive deep learning," arXiv preprint arXiv:1511.09337, 2015.
[9] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in International Joint Conference on Neural Networks. IEEE, 2016, pp. 4368–4374.
[10] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3573–3587, 2017.
[11] A. Telikani and A. H. Gandomi, "Cost-sensitive stacked auto-encoders for intrusion detection in the internet of things," Internet of Things, p. 100122, 2019.