دوفصلنامه محاسبات و سامانه های توزیع شده
سال پنجم، شماره اول، شماره پیاپی ۹، صفحه ۱۵ تا ۲۱، سال ۱۴۰۱

شبکه‌های عصبی پیچشی حساس به هزینه برای طبقه‌بندی زیرگروه‌های سرطان

راضیه هاشمی عالم، محبوبه شمسی*، مجید آقایی
برق و کامپیوتر، صنعتی قم، قم، ایران.

چکیده

طبقه‌بندی زیرگروه‌های سرطان وظیفه بسیار مهمی برای تشخیص و پیش آگهی سرطان است. 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 پایگاه داده بیان ژن و عدم تعادل داده‌ها بین طبقات مختلف تأثیر مستقیمی بر پیچیدگی و عملکرد مدل‌های طبقه‌بندی زیرگروه سرطان دارد. برای پرداختن به مشکل داده‌های نامتعادل، یک مدل شبکه عصبی کانولوشن (CNN) با استفاده از یک استراتژی حساس به هزینه برای افزایش دقت مدل در شناسایی کلاس‌های اقلیت پیشنهادشده است. از سوی دیگر، از تکنیک ضریب فیشر برای کاهش ژن‌ها در مرحله پیش‌پردازش استفاده می‌شود. در روش حساس به هزینه، ماتریس هزینه بر اساس توزیع کلاس‌ها ایجاد می‌شود و سپس از این ماتریس در مرحله تابع هزینه شبکه CNN برای محاسبه میزان خطا استفاده می‌شود. دو مجموعه از مجموعه داده‌های سرطان برای ارزیابی روش پیشنهادی استفاده می‌شود. نتایج با استفاده از سه معیار دقت، فراخوانی و دقت مقایسه می‌شوند. نتایج نشان می‌دهد که انتخاب ژن‌های مناسب برای طبقه‌بندی به همراه استفاده از یادگیری حساس به هزینه برای این منظور می‌تواند عملکرد روش پیشنهادی نسبت به مدل CNN بدون انتخاب ویژگی و یادگیری حساس به هزینه حدود ۱۱٪، ۱۰٪ و ۱۸٪ به ترتیب برای دقت، فراخوانی و صحت افزایش دهد.

کلمات کلیدی: دسته‌بندی، داده‌های نامتوازن، زیرگروه‌های سرطان، داده‌های بیان ژن، یادگیری عمیق CNN.

A Cost-Sensitive Convolution Neural Network for Cancer Subgroups Classification

Razieh hashemi Alam, Mahbobeh Shamsi*, Majid Aghaee

Electricity and computers, Qom Industrial, Qom, Iran.

Abstract

Classification of cancer subtypes is very important task for the diagnosis and prognosis of cancer. In recent years, deep learning methods have gained considerable popularity for this reason; however, it is difficult to determine the structure of the neural network because the function of the deep network depends largely on its structure. In addition, the high number of genes in the gene expression database and the imbalanced data between different classes have a direct effect on the complexity and performance of cancer subgroup classification models. To address the problem of unbalanced data, a convolution neural network (CNN) model using a cost-sensitive strategy is proposed to increase the model's accuracy in identifying minority classes. On the other hand, the fisher ratio technique is used to reduce genes in the preprocessing stage. In techniques the cost-sensitive method, a cost matrix is created based on the distribution of classes, and then this matrix is used in the CNN network cost function step to calculate the amount of error. Two sets of cancer datasets are used to evaluate the proposed method. The results show that selecting the appropriate genes for classification along with the use of cost-sensitive learning can increase the performance of the proposed method compared to the CNN model without selecting the feature and cost-sensitive learning about 11%, 10% and 18% in terms of three criteria of accuracy, recall and precision, respectively.

Keywords: Classification, Imbalanced Data, Cancer subgroups, Gene expression,Convolution Neural Networks, Cost-sensitive learning

تاریخچه مقاله:

تاریخ ارسال: ۱۴۰۰/۱۲/۱۵
تاریخ اصلاحات: ۱۴۰۱/۰۱/۰۵
تاریخ پذیرش: ۱۴۰۱/۰۳/۲۵
تاریخ انتشار: ۱۴۰۱/۰۵/۱۱

Keywords:

Classification,
Imbalanced Data,
Cancer subgroups,
Gene expression,
Convolution Neural Networks,
Cost-sensitive learning

*ایمیل نویسنده مسئول:
shamsi@qut.ac.ir

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 نویز 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 مسئله‌ی عدم توازن کلاس را در نظر نمی‌گیرند، تمایل دارند که توسط کلاس اکثریت تحت پوشش قرار داده شوند و در مقابل توسط کلاس 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 کلاس‌ها 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 طبق با توزیع داده‌ها تنظیم می‌شوند.

استراتژی یادگیری حساسیت به هزینه یکی از تکنیک‌های مبتنی بر سطح دسته‌بند است که هزینه‌های مرتبط با دسته-بندی 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 کلاس 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 استراتژی برای به حداقل رساندن هزینه کلی یادگیری است که باعث می‌شود یک مدل یادگیری به‌گونه‌ای باشد که روند آموزش نسبت به کلاسهایی که هزینه کمتری دارند حساس‌تر باشد.

هنگام استفاده از یادگیری حساس به هزینه در مدل‌های یادگیری ژرف، فرایند آموزش نسبت به کلاس‌هایی که هزینه بالاتری دارند حساس‌تر است. برخی از تلاش‌های تحقیقاتی هزینه‌های خاص کلاس را در طبقه‌بندی کننده‌های یادگیری ژرف بررسی کرده‌اند. اولین بار یادگیری ژرف حساس به هزینه توسط چونگ[1] و همکاران معرفی شد [۸]، که هزینه‌ها را در تابع خطا[2] مرحله قبل از آموزش DNN و CNN ادغام شد. وانگ[3] و همکاران [۹] با در نظر گرفتن میانگین خطا در هر کلاس ، عملکرد از دست دادن میانگین خطای مربع (MSE[4]) برای DNN را بهبود بخشید. خان و همکاران [۱۰] یک روش ابتکاری را برای اختصاص دادن هزینه به‌طور خودکار به هر کلاس با توجه به توزیع داده‌های کلاس، 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 SAE 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 ژرف حساس به هزینه این است که نیازی به استفاده از ماتریس هزینه دست‌ساز نیست زیرا هزینه‌ها از طریق آمار داده‌ها تعیین می‌شود.

این مقاله هم استراتژی‌های یادگیری حساس به هزینه و هم استراتژی کاهش ویژگی را برای رسیدگی به مشکل عدم تعادل طبقاتی و مشکلات ابعادی در طبقه‌بندی زیرگروه سرطان ادغام می‌کند. تکنیک ضریب فیشر برای حذف ژن‌های نامربوط و غیرمفید در مرحله پیش‌پردازش استفاده می‌شود. علاوه بر این، عملکرد تلفات متقابل آنتروپی با استفاده از استراتژی یادگیری حساس به هزینه با ادغام هزینه‌های مربوط به کلاس هنگام محاسبه ارزش تلفات در طول آموزش CNN، بهبود می‌یابد. در این استراتژی هزینه‌ها بر اساس آمار داده‌های دسته‌های سرطان تعریف می‌شود. این رویکرد باعث می‌شود که مدل CNN نسبت به کلاس‌های سرطان با فرکانس پایین حساس باشد و عملکرد مدل بیان ژن را در این نوع سرطان‌ها افزایش می‌دهد. آزمایش‌های مختلفی روی مجموعه داده GBM انجام می‌شود و نتایج

[1]Chung
[2]Loss function
[3]Wang
[4]Mean Squared Error (MSE)

ازنظر دقت، فراخوانی و صحت 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.

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 ۴ نتایج ارزیابی کارایی را ارائه می‌دهد و سپس نتیجه‌گیری در بخش ۵ استنتاج می‌شود.

۲- کارهای مرتبط

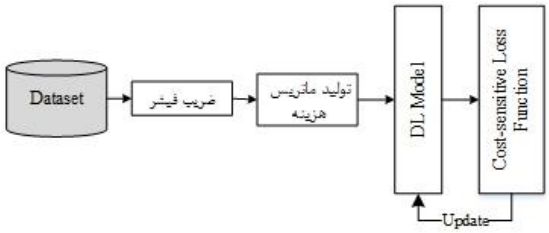
سه دسته برای انتخاب ژن وجود دارد که شامل فیلتر، بسته‌بندی و جاسازی شده است. در [۲] از آزمون t برای غلبه بر مشکل پراکندگی برای انتخاب ژن استفاده شد. لیائو و همکاران [۳] از آزمون مجموع رتبه ویلکاکسون به همراه ماشین بردار پشتیبان برای ارزیابی اهمیت ژن ها استفاده کرد. رویکردهای Wrapper از یک طبقه‌بندی کننده برای ارزیابی عملکرد یک زیرمجموعه ویژگی استفاده می‌کنند.

k-نزدیکترین همسایه [۴]، شبکه عصبی [۵] و ماشین بردار پشتیبان [۶] 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.

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.



(شکل-۱): روش پیشنهادی

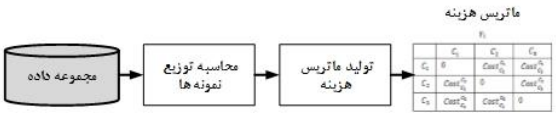
۳-۱- انتخاب ژن: داده‌های بیان ژن به‌طورکلی از هزاران ژن تشکیل‌شده است، درحالی‌که تعداد نمونه‌های موجود اغلب اندک است. در میان هزاران ویژگی در داده‌های بیان ژن، تنها چند ژن درواقع با زیرگروه‌های سرطان همراه هستند درحالی‌که بقیه ممکن است به‌عنوان ویژگی‌های زائد یا عامل اغتشاش در نظر گرفته شوند. بنابراین، 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.

نسبت فیشر نسبت فواصل کلاس با فواصل طبقه‌بندی‌شده است. اگر دو طبق بندی در یک مجموعه داده وجود داشته باشد، هر نمونه می‌تواند به‌عنوان نشان داده شود چون $Y\in\{+1,-1\}$و داده‌های بیان ژن می‌توانند باشند به‌عنوان $x_i=\{x_1^i,...,x_n^i\}$ برای هر ژن میزان انحراف استاندارد $\sigma_i^+(resp.,\sigma_i^-)$ و انحراف میانگین $\mu_i^+(resp.,\mu_i^-)$ محاسبه‌شده و میزان ضریب فیشربر اساس فرمول زیر محاسبه میشود:

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

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

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 می‌شود. برخلاف بسیاری از روش‌های قبلی برای تولید ماتریس هزینه که به‌صورت دستی و از طریق کاربرمتخصص 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 کلاس‌ها تعیین می‌شوند. شکل ۲ فرآیند تولید ماتریس هزینه γ را نشان می‌دهد.



(شکل-۲): فرآیند تولید ماتریس هزینه

در مرحله اول، توزیع هر کلاس 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 کلاس‌های اقلیت در نظر گرفته می‌شود درحالی‌که هزینه طبقه‌بندی پایین‌تری برای کلاس‌های اکثریت تعیین می‌شود. هزینه طبقه‌بندی نادرست کلاس i در کلاس j با استفاده از رابطه ۳-۱ محاسبه می‌شود.

$$\begin{cases}\gamma_{i,j}=\frac{\alpha_i}{\alpha_i+\alpha_j} & i,j=1,2,\dots,C \\ subject\ to\ i\neq j\end{cases} \qquad (۱)$$

در یک ماتریس هزینه، سطر مورب ماتریس به‌عنوان بردار سودمندی شناخته می‌شود. این بردار نشان‌دهنده طبقه‌بندی‌های صحیح را نشان می‌دهد و به صفر تنظیم می‌شود. همچنین، تمام هزینه‌ها غیر منفی هستند، یعنی $\gamma_{i,j}>0$. در این رابطه، α_i و α_j به ترتیب تعداد نمونه‌های کلاس‌های i و j هستند.

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 کانولوشن است که هر یک از آن‌ها دارای یک کانولوشن و به دنبال آن لایه‌های تابع فعال‌ساز Relu و ادغام حداکثری است. اندازه فیلتر برای لایه کانولوشن ۱×۸ و ۱= stride است و هر لایه ادغام حداکثری یک ورودی ۱×۴ را با۲=stride پردازش می‌کند. بعد از هر لایه ReLu ، از نرمال‌سازی دسته‌ای و Dropout با نسبت ۰/۰۵ استفاده می‌شود. بعد از لایه‌های کانولوشن، دولایه اتصال کامل برای طبقه‌بندی ژن استفاده شد.

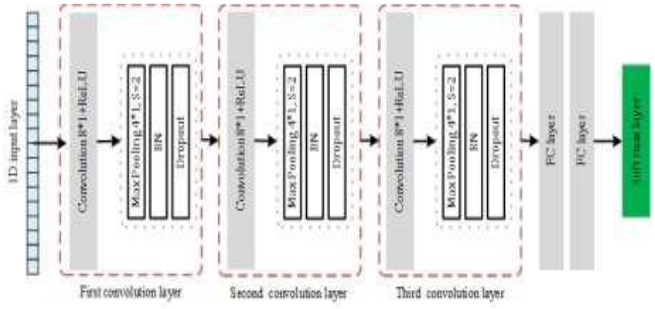
۳-۴- تابع هزینه حساس به هزینه: در این بخش یک تابع هزینه حساس به هزینه پیشنهاد می‌شود که نسبت به طبقه-بندی نادرست کلاس‌های اقلیت حساس‌تر است. در طول آموزش، روش یادگیری پیشنهادی به‌طور مشترک هزینه‌های وابسته به کلاس و پارامترهای شبکه عصبی را بهینه می‌کند. در مقایسه با رویکردهای سطح داده (نمونه‌برداری مجدد)، 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، درروش پیشنهادی هزینه‌های مرتبط با هر کلاس 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 کلاس اکثریت طبقه‌بندی می‌شود بیشتر از زمانی است که نمونه اکثریت به‌اشتباه به‌عنوان کلاس اقلیت طبقه‌بندی می‌شود. همان‌طور که در بخش قبلی اشاره کردیم، کلاس‌های اقلیت و اکثریت تعیین می‌شوند و فقط باید هزینه مربوطه را از ماتریس هزینه پیدا کنیم. برتری الگوریتم ما این است که تعیین نوع کلاس‌ها ازنظر اقلیت یا اکثریت لازم نیست. درواقع، هزینه‌ها فقط بر اساس توزیع کلاس‌ها 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 کلاس‌های 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 استراتژی 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 کلاس‌ها هستند. تابع Softmax احتمال اینکه نمونه i (x_i) متعلق به یک کلاس باشد را محاسبه می‌کند:



(شکل۳): حساس به هزینه CNN معماری

$$f_\theta(x) = \frac{1}{\sum_{j=1}^{C} e^{y_i}} \begin{bmatrix} e^{y_1} \\ e^{y_2} \\ \dots \\ e^{y_C} \end{bmatrix} = \begin{bmatrix} p(y_i = 1|x_i) \\ p(y_i = 2|x_i) \\ \dots \\ p(y_i = C|x_i) \end{bmatrix} \quad (۲)$$

متغیر θ پارامتر نگاشت برای کلاس j است $(b_j + W_j x)$. رویکرد پیشنهادی در این پژوهش مجازات کردن طبقه‌بندی نادرست در تابع هزینه cross-entropy بر اساس هزینه‌های تعیین‌شده در ماتریس هزینه (γ) برای به حداکثر رساندن نزدیکی پیش‌بینی به کلاس واقعی را مدنظر قرار می‌دهد. مقدار کل هزینه هر دسته با N نمونه با استفاده از معادله ۳ محاسبه می‌شود:

$$\mathcal{L}(O, y) = -\frac{1}{N}\sum_{j=1}^{N} \mathcal{L}(O_j, y_j) \quad (۳)$$

که در آن مقدار cross-entropy میانگین مقادیر زیان برای کل N طبقه‌بندی است. مقدار زیان برای هر پیش‌بینی توسط معادله ۴ محاسبه می‌شود:

$$\mathcal{L}(O_i, y_i) = -\sum_{i=1}^{C} \left(y_{o,c} \log p(y_i = 1|x_i; \theta_i)\right) \quad (۴)$$

در این رابطه، $y_{o,c}$ یک شاخص دودویی (۰ یا ۱) که به پیش‌بینی صحیح مشاهده برای نمونه o اشاره دارد. مقدار $y_{o,c}$ برای کلاس اشتباه پیش‌بینی‌شده ۱ و برای کلاس واقعی ۰ است. احتمال طبقه‌بندی اشتباه با در نظر گرفتن هزینه مربوط به کلاس تغییر می‌یابد (معادله ۵):

$$p(y_i = 1|x_i) = \frac{\gamma_{i,j} \exp(O_i)}{\sum_{i=1}^{C} \exp(O_i)} \quad (۵)$$

بر اساس معادله ۳-۶، ضرب هزینه مربوط به کلاس‌های اقلیت، مقدار احتمال جدید را به‌شدت کاهش می‌دهد و بنابراین، منجر به افزایش مقدار زیان طبقه‌بندی در رابطه ۳-۵ می‌شود. به‌این‌ترتیب، کلاس‌های اقلیت بیشتر از کلاس‌های اکثریت بر روی تابع هزینه تأثیر می‌گذارند.

۴- نتایج ارزیابی

در این فصل، 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. استراتژی توقف اولیه برای جلوگیری از مشکل بیش برازش مورداستفاده قرار گرفت که در آن زمانی که مقدار خطا بر روی‌داده‌های اعتبار سنجی 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 ۶)، فراخوانی (رابطه ۷) و صحت (رابطه ۸) استفاده می‌شود.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{FP+FN+TP+TN}} \quad (۶)$$

$$Recall = \frac{TP}{TP+FN} \quad (۷)$$

$$Precision = \frac{TP}{TP+FP} \quad (۸)$$

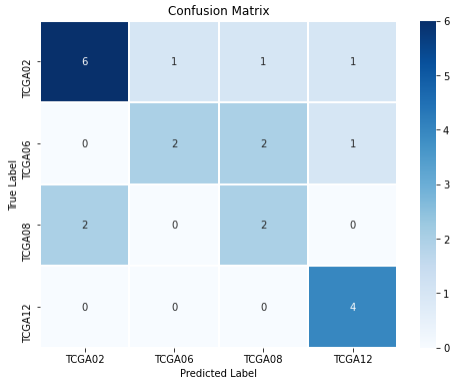
۴-۲- مجموعه داده‌ها: در این پژوهش، مجموعه داده زیرگروه سرطان بیان ژن RNA-Seq برای GBM (گلیوبلاستوما چندحالت[5]) برای آزمایش‌ها استفاده شد. مجموعه داده GBM دارای ۲۱۲ نمونه سرطانی با حدود ۱۲ هزار ژن است که در چهار گروه ژنی دسته‌بندی می‌شوند. در مجموعه داده GBM دو کلاس TCGA02 و TCGA06 دارای توزیع یکسانی هستند (حدود ۸۰ نمونه) و دو کلاس 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 دو معیار فراخوانی و صحت را نشان می‌دهد. روش فیشر با ژن‌های اصلی مقایسه شده‌اند. همان‌گونه که در جدول مشاهده می‌شود، روش پیشنهادی بر روی ژن‌های انتخاب‌شده توسط تکنیک ناهنجاری بهترین عملکرد را داشته است. همچنین شکل ۴ ماتریس‌های آشفتگی برای روش پیشنهادی بر روی داده اصلی را نشان می‌دهد.

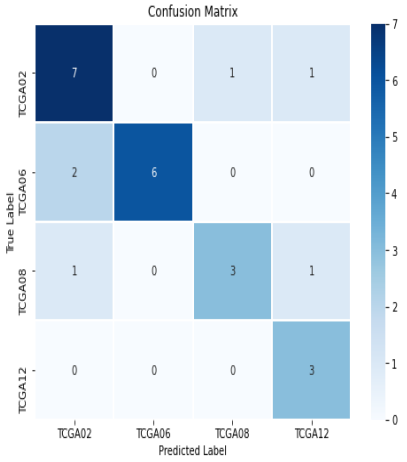
(جدول-۲): عملکرد روش پیشنهادی برای مجموعه داده GBM

کلاس	انتخاب ژن	فراخوانی	صحت
TCGA02	مجموعه اصلی	۰.۶۷	۰.۷۵
	فیشر	۰.۸۸	۰.۷۲
TCGA06	مجموعه اصلی	۰.۴	۰.۶۷
	فیشر	۰.۶	۱.۰
TCGA08	مجموعه اصلی	۰.۵	۰.۴
	فیشر	۰.۷۵	۰.۷۵
TCGA12	مجموعه اصلی	۱.۰	۰.۶۷
	فیشر	۰.۷۵	۰.۷۵

[5]Glioblastoma Multiforme



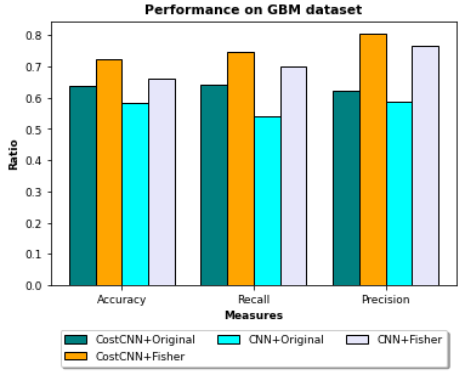
(الف) مجموعه داده اصلی



(ب) ضریب فیشر

(شکل-۴): ماتریس آشفتگی برای مدل با مجموعه ژن‌های انتخاب‌شده متفاوت برای مجموعه داده GBM اصلی

شکل ۵ 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.



(شکل-۵): عملکرد مدل‌های یادگیری ژرف برای دسته‌بندی

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 کلاس‌های 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 استراتژی حساس به هزینه نیز برای مقابله با داده‌های نامتعادل استفاده شد. در این استراتژی، 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، یادآوری، دقت 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.

۵- 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 pretraining 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.

دکتر مجید آقایی، نرم افزار ،دانشکده مهندسی کامپیوتر نرم افزار هیئت علمی دانشگاه صنعتی قم
aghaee@qut.ac.ir

روش ارجاع به مقاله : ر. هاشمی عالم، م. شمسی، م. آقایی. شبکه های عصبی پیچشی حساس به هزینه برای طبقه بندی زیرگروه های سرطان. دوفصلنامه محاسبات و سامانه های توزیع شده سال پنجم، شماره اول، شماره پیاپی ۹، صفحه ۱۵ تا ۲۱، سال ۱۴۰۱.

How to cite: Razieh hashemi alam, mahbobeh shamsi, majid aghaee. A cost-sensitive convolution neural network for cancer subgroups classification. Journal of Distributed Computing and Systems(JDCS), Vol 5, Issue 1, Page 15-21, 2022.

راضیه هاشمی عالم، کارشناسی ارشد نرم افزار،
دانشکده مهندسی کامپیوتر، دانشگاه صنعتی قم
hashemialam.r@qut.ac.ir

دکتر محبوبه شمسی، دکتری مهندسی کامپیوتر نرم افزار (پردازش تصویر)، استادیار، هیئت علمی دانشگاه صنعتی قم
shamsi@qut.ac.ir