بررسی الگوریتم XGBoost روی سرطان

حورا محمودیان اصفهانی

دانشجوی مهندسی کامپیوتر دانشگاه صنعتی اصفهان

Hoora79hoora79@gmail.com

زهرا مستاجران

دانشجوی مهندسی کامپیوتر دانشگاه صنعتی اصفهان

Zahramostajeran.cs@gmail.com

حوری دهش

دانشجوی مهندسی کامپیوتر دانشگاه صنعتی اصفهان

h.dahesh.2000@gmail.com

چکیده

طبقه‌بندی سرطان‌ها می‌تواند به پزشکان در انتخاب راهبردهای مراقبت و درمان بیماران کمک زیادی کند از این رو امروزه به جای روش‌های سنتی از روش‌های رایانه‌ای استفاده می‌شود به طوری که کمک گرفتن از یادگیری ماشین در حل مسئله‌ها یک بخش جدایی‌ناپذیر از جامعه امروزی شده است. یادگیری ماشین شامل الگوریتم‌های متعددی است که از آن‌ها برای مسئله‌های گوناگون استفاده می‌شود که در این مطالعه برای مسئله تشخیص سرطان از الگوریتم XGBoost استفاده شده است. این الگوریتم زیرمجموعه‌ی یادگیری بانظارت است و با استفاده از این الگوریتم مدلی ساخته می‌شود که برای طراحی آن از معماری CRISP بهره گرفته می‌شود. این معماری دارای شش فاز است که با طی کردن پنج فاز اول مدل ما ساخته می‌شود در نتیجه برای اطمینان از این مدل آن را با شش مدل دیگر طبق معیار ارزیابی مساحت زیر منحنی مقایسه می‌کنند سپس این نتیجه حاصل می‌شود که این مدل قادر است به جامعه پزشکان در تشخیص سریع سرطان‌ها کمک کند.

کلمات کلیدی: یادگیری ماشین، یادگیری بانظارت، طبقه‌بندی، الگوریتم‌های تقویتی، XGBoost ، معماری CRISP

۱. مقدمه

سرطان بیماری است که در اثر رشد کنترل نشده و تکثیر غیرعادی سلول‌ها ایجاد می‌شود و ابتلا به آن سالیانه تعداد زیادی را به کام مرگ می‌کشاند. به همین خاطر هم هست که تشخیص و درمان آن یکی از مسائل چالش برانگیز در دهه اخیر بوده است.

امروزه به خاطر پیشرفت علم و فناوری، بسیاری از سلول‌های سرطانی در مراحل اولیه از طریق آزمایش خون و یا تصویربرداری‌های پیشرفته تشخیص داده می‌شوند. اما در این میان هوش مصنوعی می‌تواند نقش مهم‌تری را ایفا کرده و با تفسیر صحیح و علمی تصاویر پزشکی و همچنین آزمایش‌های گرفته شده از بیمار جزئیات بسیاری مهمی را کشف کرده و در اختیار پزشکان قرار دهد.

در ادامه ساختار مقاله بدین صورت است که در بخش ۲ به متن اصلی و در بخش ۳ به نتیجه‌گیری پرداخته می‌شود.

۲. متن اصلی

۲-۱. یادگیری ماشین

از یادگیری ماشین می‌توان به عنوان یکی از زیر شاخه‌های هوش مصنوعی نام برد که در آن به‌جای استفاده‌ی مستقیم از دستور العمل‌های خاص از مطالعه علمی الگوریتم‌ها و مدل‌های آماری کمک گرفته و از تفسیر آنها برای انجام کار مورد نظر استفاده می‌شود. این علم سبب می‌شود رایانه‌ها بدون نیاز به یک برنامه صریح در مورد یک موضوع خاص یاد بگیرند.

امروزه می‌توان نشانه‌های حضور یادگیری ماشین را در زمینه‌های مختلف از جمله مهندسی، زبان‌شناسی، تجارت و پزشکی به راحتی پیدا کرد؛ برای نمونه در دنیای بورس می‌توان به کمک این علم قیمت سهام شرکت‌های مختلف را با تقریب خوبی پیش‌بینی کرد [۱].

یادگیری ماشین می‌تواند به سه روش اصلی انجام پذیرد:

۱. یادگیری بدون‌نظارت

۲. یادگیری بانظارت

۳. یادگیری تقویتی

۲-۱-۱. یادگیری بانظارت

یادگیری بانظارت به دلیل درک شهودی آسان و پیاده‌سازی و اجرای راحت‌تر نسبت به دیگر روش‌های یادگیری به عنوان محبوب‌ترین روش شناخته می‌شود. در این روش ابتدا بسته به نیاز، تعدادی برچسب تعریف خواهد شد سپس تعدادی ورودی در اختیار الگوریتم قرار می‌گیرد. در این مرحله به الگوریتم اجازه داده می‌شود که برای هر ورودی برچسب مرتبط را انتخاب کند و بعد از انتخاب، به آن بازخورد می‌دهد که آیا به درستی انتخاب کرده است یا خیر. به مرور الگوریتم ماهیت رابطه بین ورودی‌ها و برچسب‌ها

را بهتر درک خواهد کرد زمانی که الگوریتم کاملا آموزش دید قادر خواهد بود در مورد نمونه های کاملا جدید که تاکنون برخوردی با آنها نداشته، برچسب متناسب را انتخاب کند [۲].

در این مرحله بسته به گسسته یا پیوسته بودن خروجی الگوریتم یادگیری، با دو دسته مسئله روبه‌رو خواهیم شد. به مسائل با خروجی گسسته مسائل طبقه‌بندی و به دسته‌ی دیگر مسائل رگرسیون گفته می‌شود [۲].

۲-۱-۲. طبقه‌بندی

طبقه‌بندی فرآیند قرار دادن نمونه‌های جدید در طبقات مختلف براساس داده‌های قدیمی است و برای این‌کار به یک مدل طبقه‌بند یا الگوریتم طبقه‌بند نیاز است. درختان تصمیم، نایو بیز، شبکه های عصبی و الگوریتم تقویتی[1] نمونه‌هایی از الگوریتم طبقه‌بند هستند [۲].

۲-۱-۳. الگوریتم تقویتی

در این روش تعدادی سیستم تصمیم‌گیری ضعیف وجود دارد که تنها کمی بهتر از حالت تصادفی رفتار می‌کنند، این سیستم‌ها کنار هم شروع به فعالیت می‌کنند و در هر مرحله بسته به عملکردی که داشته‌اند وزن جدیدی به‌دست می‌آورند. سیستم‌هایی که عملکرد خوبی داشته‌اند وزن بیشتر و آنهایی که عملکرد ضعیفی داشته‌اند وزن کمی خواهند گرفت و در نهایت به کمک هم یک سیستم تصمیم‌گیری قوی ایجاد خواهند کرد. لازم به ذکر است که سه مورد از محبوب‌ترین الگوریتم‌های تقویتی XGBoost[2]، Gradient Boosting، Adaptive Boosting نام دارند [۳].

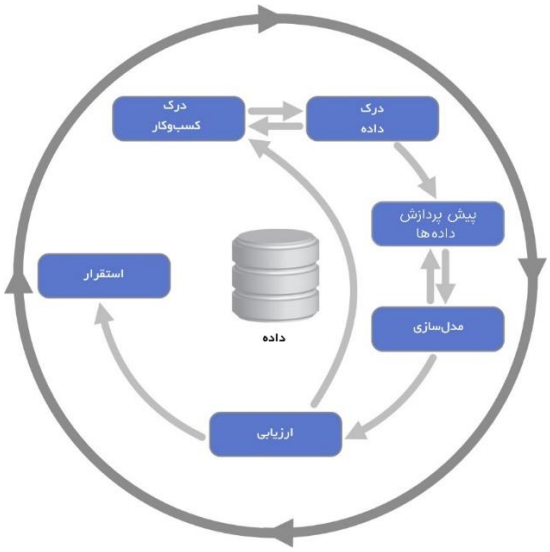
۲-۱-۴. XGBoost

الگوریتم XGBoost یک پیاده‌سازی از تقویت گرادیان درخت تصمیم‌گیری است. این الگوریتم به صورت کتابخانه‌ی نرم افزاری برای زبان‌های مختلف مثل Scala, Pyton, Java, Julia, R در دسترس است. XGBoost در مقایسه با سایر پیاده‌سازی‌های تقویت گرادیان به وضوح سریع‌تر عمل می‌کند [۴].

۲-۲. معماری CRISP

برای طراحی مدل در یادگیری ماشین از معماری CRISP[3] استفاده می‌شود این معماری دارای شش فاز است که پنج فاز اول آن در حوزه‌ی مسئله‌ی یادگیری ماشین است و فاز آخر آن در حوزه‌ی نرم افزار قرار دارد [۵]. شکل ۱ معماری CRISP را نشان می‌دهد.

[1] Boosting
[2] Extreme gradient boosting
[3] Cross-industry standard process



شکل ۱. معماری CRISP [۶]

۲-۲-۱. درک کسب‌وکار

در این مرحله، باید با کسب‌وکاری که قرار است برای آن مدلی ساخته شود آشنایی اولیه صورت گیرد بدین معنی که زوایای مختلف آن کسب‌وکار، محدودیت‌ها، شرایط موجود و اهداف آن مورد بررسی واقع شود [۵].

۲-۲-۲. درک داده‌ها

در ابتدا باید داده‌های مورد نیاز در کسب‌وکار را جمع‌آوری کرد. داده‌های مورد استفاده در این تحقیق از اطلس ژنوم سرطان (TCGA) جمع‌آوری شده است. این مسئله بر چهار نوع سرطان متمرکز می شود که شامل کارسینوم سلول شفاف کلیه (KIRC) با ۵۳۷ نمونه، کارسینوم سلول پاپیلاری کلیه (KIRP) با ۲۹۱ نمونه، کارسینوم سلول سنگفرشی ریه (LUSC) با ۵۰۴ نمونه و کارسینوم سلول سنگفرشی سروگردن (HNSC) با ۵۲۸ نمونه است. لازم به ذکر است که اطلاعات موجود در هر مجموعه داده شامل mRNA، داده‌هایmiRNA-seq، داده‌های متیلاسیون DNA و اطلاعات بالینی است [۷].

۲-۲-۳. پیش‌پردازش داده‌ها

بعد از اینکه داده‌ها جمع‌آوری شدند باید پیش‌پردازش شوند به این علت که داده‌های تمیز برای مدل‌سازی به دست آورده شود. نمونه‌های متعددی برای پیش‌پردازش داده‌ها وجود دارد که رایج‌ترین آن‌ها در ادامه بیان شده است.

۲-۲-۳-۱. پیدا کردن نمونه‌های از دست‌رفته یا گمشده و حذف آن‌ها

اغلب در هر مجموعه داده، گمشدگی وجود دارد؛ به عنوان مثال شخصی به یک سوال نظرسنجی پاسخ نمی‌دهد یا یک سنسور دچار مشکل می شود، این موارد تنها دو مثال از چگونگی گم‌شدن داده‌ها می‌باشند اما گمشدگی همیشه با نشانگرهای خاص مطرح نمی‌شود. گاهی مقادیر منفی می‌توانند داده‌های از دست‌رفته را نشان دهند؛ به عنوان مثال ۹۹۹- به عنوان سن فرد نشان‌دهنده گمشدگی است. اینکه گمشدگی چگونه نشان داده شود به سیستم یا نرم‌افزار مبدأ و هم‌چنین نحوه‌ی مدیریت داده‌ها بستگی دارد [۸].

۲-۲-۳-۲. تبدیل مقادیر

گاهی باید مقادیر ویژگی‌ها را تغییرداد یعنی ممکن است توزیع‌های بسیار اریب وجود داشته باشد و یک تبدیل غیرخطی، این توزیع‌ها را عادی کند که این کار می‌تواند به بسیاری از الگوریتم‌ها کمک نماید [۸].

۲-۲-۳-۳. رمزگذاری متغییرهای طبقه‌بندی شده

بسیاری از الگوریتم‌های یادگیری ماشین در تلاش برای تبدیل داده‌ها به یک نمایش عددی مناسب هستند تا بتوانند روی آن‌ها کار کنند؛ به عنوان مثال به جای اینکه ستونی به نام سرطان با مقدار kirc یا kirp وجود داشته باشد، یک ستون kirc با مقدار صفر یا یک ایجاد شود که به‌ترتیب نشان‌دهنده‌ی kirc نبودن و kirc بودن، است و برای kirp هم نیز به همین‌صورت عمل می‌شود [۸].

برای هر نوع سرطان، بر روی بیمارانی که هر سه نوع داده مولکولی (DNA, mRNA, miRNA) و اطلاعات مرحله پاتولوژیک دارند، تمرکز می‌شود و برای DNA، فقط نمونه‌هایی که بیشترین همبستگی منفی با ژن دارند، حفظ می‌شوند [۷].

در این مطالعه برای مدیریت گمشدگی‌ها، هر ویژگی با بیش از ۲۰ درصد گمشدگی در ویژگی‌های بیولوژیکی حذف می‌شود و درنهایت ۱۲ مجموعه داده اصلی برای تحلیل‌های پایین دستی به دست می‌آید [۷].

۲-۲-۴. مدل‌سازی

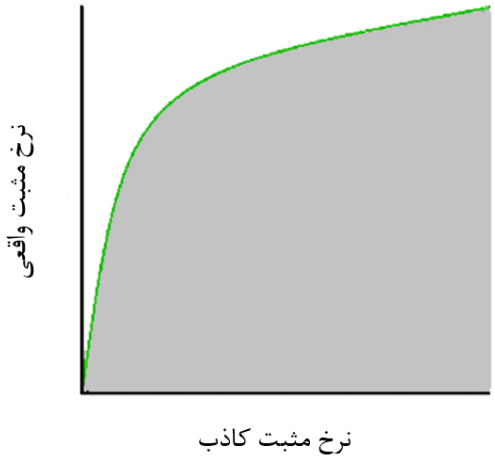
برای اینکه مدل‌سازی انجام شود باید از داده‌های پیش‌پردازش شده که از فاز قبل به دست آمده استفاده شود در نهایت این داده‌ها به دو قسمت آموزش[1] و آزمایش[2] تقسیم بندی می‌گردند [۹]؛ به عنوان مثال ۱۰۰ نمونه سوال در اختیار یک دانشجو قرار می‌گیرد و دانشجو به دلخواه خودش این ۱۰۰ نمونه سوال را به دو قسمت ۹۰ درصدی و ۱۰ درصدی تقسیم‌بندی می‌کند و از آن ۹۰ درصد برای آموزش خودش استفاده می‌کند و در نهایت با آن ۱۰ درصد خود را محک می‌زند مدل XGBoost هم به این طریق ساخته می‌شود.

[1] Train
[2] Test

زمانی که مدل XGBoost ساخته شد باید این مدل بهینه گردد که این امر با کمک الگوریتم‌های بهینه‌سازی ممکن می‌شود. به این منظور الگوریتم‌های بهینه‌سازی متعددی وجود دارد که در این مسئله از الگوریتم جستجوی شبکه‌ای[1] استفاده شده است. این الگوریتم به این صورت کار می‌کند که مقادیر مختلف برای هایپرپارامترهای هر مدل را می‌گیرد و با ترکیب مختلف آن‌ها مدل‌های متعددی ایجاد می‌کند. در نهایت دقیق‌ترین مدل را به عنوان مدل بهینه‌ی نهایی برمی‌گرداند [۱۰].

۲-۲-۵. ارزیابی مدل

برای ارزیابی مدل معیاری‌های متعددی از جمله صحت، دقت، پوشش و مساحت زیر منحنی (AUC) وجود دارد. معیار مورد استفاده برای این مسئله، مساحت زیر منحنی است. شکل ۲ این معیار را نشان می دهد. معیار مورد نظر به این صورت کار می‌کند که هر چه مساحت زیر آن منحنی بیشتر باشد به منزله‌ی آن است که مدل دقیق‌تر است و این بدین معناست که تعداد مثبت واقعی که مدل پیش‌بینی کرده، بیشتر است؛ به عنوان مثال یک فرد سرطان دارد و مدل به درستی تشخیص می‌دهد که آن فرد سرطان داشته است [۱۱].



شکل ۲. معیار مساحت زیر منحنی

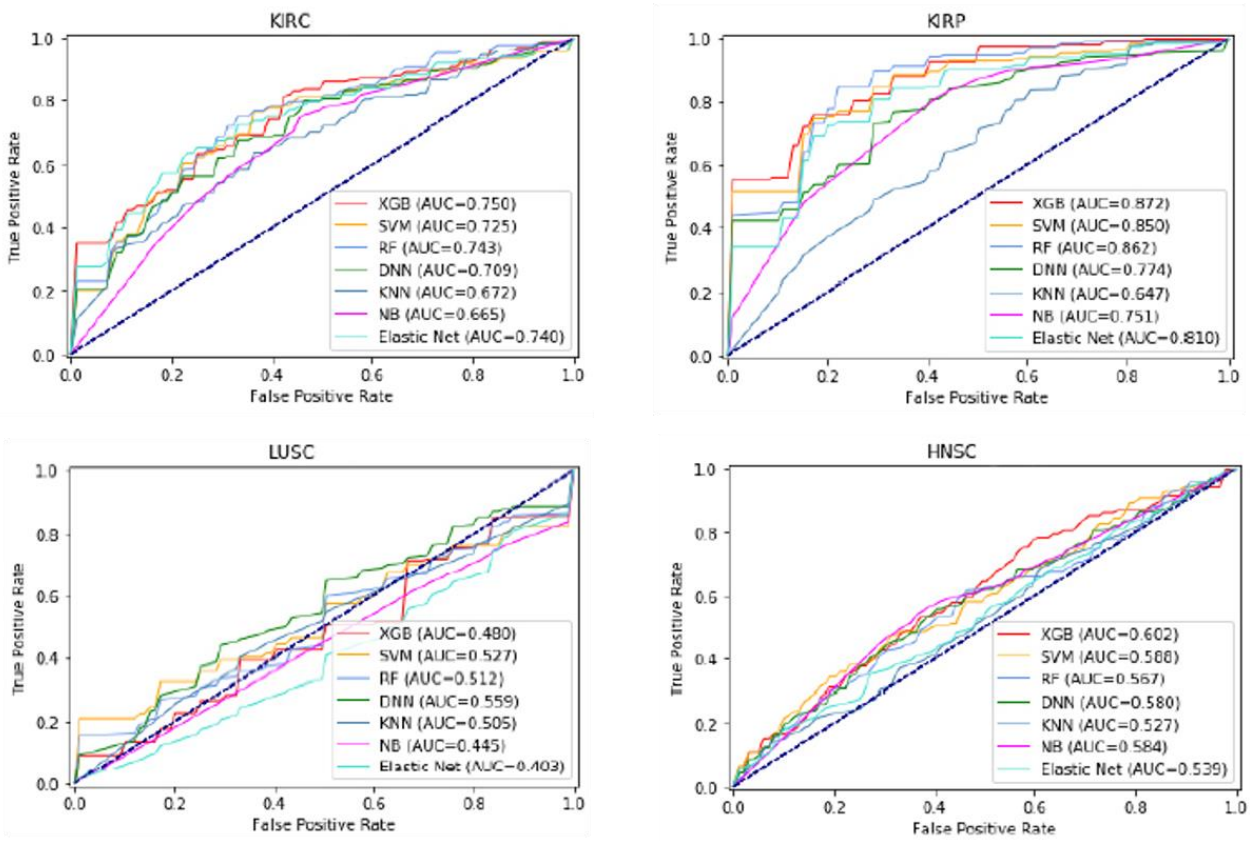
۲-۲-۶. استقرار

فاز آخر این معماری با کمک یک مهندس نرم افزار تکمیل می‌گردد بدین صورت که او با در اختیار داشتن پنج فاز اول اقدام به طراحی و ایجاد نرم افزار می‌کند تا آن را در اختیار پزشک قرار دهد. پزشک با وارد کردن ورودی‌ها، خروجی‌های مدنظر را دریافت می‌کند. لازم به ذکر است که با طراحی نرم افزار، دیگر نیازی نیست که پزشک با کدها درگیر شود [۵].

[1] Grid search

۳. نتیجه‌گیری

در این مطالعه برای دستیابی به بهترین مدل، هفت مدل XGBoost، ماشین بردار پشتیبان (SVM)، جنگل تصادفی (RF)، شبکه عصبی عمیق (DNN)، k _ نزدیکترین همسایه (KNN)، نایو بیز (NB) و Elastic Net طبق معیار AUC با یکدیگر مقایسه و ارزیابی شدند که براساس این ارزیابی نمودارهای هر مدل به صورت شکل ۳ به دست آمدند. طبق نتایج نشان داده شده در شکل ۳، مدل XGBoost در تمامی نمودارها، مساحت زیر منحنی بیشتری داشته است پس این مدل به عنوان مدل نهایی برای مسئله‌ی تشخیص سرطان مورد استفاده قرار خواهد گرفت و به جامعه پزشکان در تشخیص این بیماری کمک شایانی خواهد کرد.



شکل ۳. مقایسه هفت مدل طبق معیار AUC [۷]

مراجع

[۱]- https://blog.faradars.org/introduction-to-machine-learning/

[۲]- http://cafetadris.com/blog/%DB%8C%D8%A7%D8%AF%DA%AF%DB%8C%D8%B1%DB%8C-%D8%A8%D8%A7-%D9%86%D8%A7%D8%B8%D8%B1-supervised-learning/

[۳]- https://de.wikipedia.org/wiki/Boosting

[۴]- https://en.wikipedia.org/wiki/XGBoost

[۵]- https://chistio.ir/%D9%81%D8%B1%D8%A2%DB%8C%D9%86%D8%AF-%DA%A9%D8%B1%DB%8C%D8%B3%D9%BE-crisp-%D9%BE%D8%B1%D9%88%DA%98%D9%87-%D8%AF%D8%A7%D8%AF%D9%87-%DA%A9%D8%A7%D9%88%DB%8C/

[۶]- https://blog.faradars.org/wp-content/uploads/2018/09/CRISPDM_Process_Diagram.png

[۷]- B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, “Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data,” Computers in Biology and Medicine, vol. 121, p. 103761, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103761.

[۸]- https://towardsdatascience.com/pre-processing-and-training-data-d16cc12dfbac

[۹]- https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets

[۱۰]- https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/#:~:text=Grid%20Search%20uses%20a%20different,the%20number%20of%20hyperparameters%20involved.

[۱۱]- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5