الگوریتمهای یادگیری ماشین در تحلیل داده‌های توالی- RNA (RNA-Seq) بیماران سرطانی

فاطمه قلی پور[1]، علی محمدیان مصمم [2*]، بهنام آقاجان[3]

1 دانشگاه زنجان، دانشکده علوم، ، دانشجوی کارشناسی ارشد علم داده، *fghfgholipour7732@gmail.com*

2 دانشگاه زنجان، دانشکده علوم، عضو هیات علمی گروه آمار، *a.m.mosammam@znu.ac.ir*

3 دانشگاه زنجان ، دانشکده ریاضی، دانشجوی دکتری ریاضی، *behnamaghajan.71@yahoo.com*

* نویسنده مسئول

چکیده

با توسعه فناوری‌های نوین در عرصه پزشکی، حجم عظیمی از داده‌های مرتبط با سرطان گردآوری شده و در اختیار جامعه پژوهشی قرار گرفته است. استفاده از هوش مصنوعی و یادگیری ماشین در حوزه‌های مختلف پزشکی، به‌ویژه در پیش‌بینی، غربالگری و تشخیص سرطان، با هدف بهبود دقت و کارایی انجام می‌شود. با توجه به پیچیدگی و حجم بالای داده‌های ژنتیکی، یادگیری ماشین قادر است الگوهای پنهان را شناسایی کرده و دقت پیش‌بینی را نسبت به روش‌های سنتی افزایش دهد. در این مقاله با استفاده از الگوریتم‌های یادگیری ماشین به طبقه‌بندی داده های توالی نسل بعدی(*NGS*) بویژه داده‌های *(RNA-Seq)* می پردازیم. داده های مورد مطالعه در این مقاله سرطان دهانه رحم با بیان‌های 714 میکرو *RNA (miRNA)* از نمونه‌های انسانی را اندازه‌گیری کرده است. در این مجموعه داده، 29 نمونه توموری و 29 نمونه غیرتوموری از دهانه رحم وجود دارد و این دو گروه به عنوان دو کلاس جداگانه در نظر گرفته می‌شوند.

کلمات کلیدی: الگوریتم‌های یادگیری ماشین، بیان ژن، *RNA-Seq*، *SVM*

1- مقدمه

سرطان بیماری‌ای است که در آن تعدادی از سلول‌های بدن بدون کنترل رشد کرده و به نواحی دیگر بدن سرایت می‌کنند. با توسعه فناوری‌های نوین در عرصه پزشکی، حجم عظیمی از داده‌های مرتبط با سرطان گردآوری شده و در اختیار جامعه پژوهشی قرار گرفته است. با این وجود، پیش‌بینی دقیق نتیجه یک بیماری همچنان یکی از وظایف چالش‌برانگیز برای پزشکان محسوب می‌شود*[1]*.

با توجه به اینکه سرطان سالانه بیش از 12 میلیون مورد جدید را در سراسر جهان شامل می‌شود و پیش‌بینی می‌شود تا سال 2050 این تعداد به 27 میلیون تشخیص سالانه برسد، این بیماری یک چالش بزرگ بهداشت و سلامت عمومی را مطرح می‌کند. در حالی که طی دهه گذشته پیشرفت‌های قابل‌توجهی در درک زیست‌شناسی سرطان به دست آمده است، پیشرفت‌های قابل توجهی در زمینه‌های تشخیص زودهنگام، آزمایش‌های غربالگری، یا درمان‌های هدفمند و با سمیت کمتر برای سرطان مشاهده نشده است. نشانگرهای زیستی حساس و خاص نقش حیاتی در تشخیص زودهنگام و پایش سرطان ایفا می‌کنند. شناسایی نشانگرهای مولکولی جدید تومور به منظور کاهش مرگ و میر و عوارض ناشی از سرطان یک اولویت کلیدی است*[2]* .

فناوری ***NGS*** به عنوان یک روش نوآورانه برای تعیین توالی ***DNA*** و ***RNA*** و شناسایی واریانت‌ها و جهش‌ها معرفی شده است. این فناوری، امکان تعیین توالی موازی گسترده‌ای از توالی‌های گوناگون *DNA* یا*RNA* ، و حتی کل ژنوم، را در مدت زمانی نسبتاً کوتاه فراهم می‌کند. این پیشرفت، بعد از تعیین توالی **سنگر**، به عنوان یک گام انقلابی در حوزه تعیین توالی شناخته می‌شود. *NGS* شامل چند مرحله کلیدی در فرآیند تعیین توالی است. به عنوان مثال، فرآیند *NGS* برای *DNA* شامل مراحل تکه‌تکه کردن*DNA* ، آماده‌سازی کتابخانه، تعیین توالی گسترده به صورت موازی، تحلیل داده‌های بیوانفورماتیکی و تفسیر واریانت‌ها و جهش‌ها است*[3]*. تکه‌تکه کردن *DNA*برای خرد کردن *DNA* هدف به قطعات کوتاه‌تر، معمولاً به طول 100 تا 300 جفت باز *(bp)* به کار می‌رود. این فرآیند با استفاده از روش‌های مختلفی امکان‌پذیر است، از جمله روش‌های مکانیکی، هضم آنزیمی*[4]* و روش‌های دیگر. به عنوان نمونه، می‌توان با استفاده از فراصوت *DNA* را به قطعات کوتاه‌تر شکافت. سپس، قطعات کوتاه مربوط به توالی‌های هدف *DNA* با استفاده از پروب‌های مکمل خاص از مجموعه استخراج می‌شوند. روش جایگزین دیگر شامل واکنش زنجیره‌ای پلیمری *(PCR)* است که در آن از جفت‌های متعددی از پرایمرها برای تکثیر بخش‌های خاصی از *DNA* هدف به وسیله *PCR* استفاده می‌شود. محصولات حاصل از *PCR* به عنوان قطعات کوتاهی از *DNA* هدفمند عمل می‌کنند. سپس این قطعات *DNA* برای آماده‌سازی کتابخانه به کار گرفته می‌شوند*[5]*.

در این راستا، روش‌های یادگیری ماشین به ابزارهای پرکاربردی برای محققان پزشکی تبدیل شده‌اند. این تکنیک‌ها توانایی کشف و تحلیل الگوها و روابط پیچیده میان داده‌ها را دارند و می‌توانند به طور مؤثر نتایج احتمالی انواع مختلف سرطان را پیش‌بینی کنند. یکی از وظایف اصلی ، طبقه‌بندی مبتنی بر بیان ژنی می‌باشد، در زمینه طبقه‌بندی مبتنی بر بیان ژن، الگوریتم‌های مختلفی با استفاده از داده‌های میکروآرایه توسعه یافته و تطبیق داده

شده‌اند.*RNA-Seq*، که از تکنولوژی‌های پیشرفته توالی‌یابی نسل جدید *(NGS)* استفاده می‌کند، که نسبت به میکروآرایه‌ها مزایای عمده‌ای مانند تولید داده‌های با نویزی کمترو شناسایی رونویسی‌ها و ایزوفرم‌های جدید را دارد. این مزایا می‌توانند عملکرد الگوریتم‌های طبقه‌بندی را بهبود بخشند.*[6]*

2- روش‌شناسی

داده‌های مورد مطالعه در این بخش شامل سرطان دهانه رحم با بیان‌های 714 میکرو *RNA (miRNA)* از نمونه‌های انسانی است. در این داده‌ها، 29 نمونه توموری و 29 نمونه غیرتوموری از دهانه رحم وجود دارد و این دو گروه به عنوان دو کلاس جداگانه در نظر گرفته می‌شوند. فریم داده‌ای که از آن صحبت می‌شود، شامل 58 نمونه انسانی است که در هر نمونه بیان 714 *miRNA* اندازه‌گیری شده است. این داده‌ها در قالب یک جدول یا ماتریس سازماندهی شده‌اند که هر ردیف آن یک نمونه و هر ستون آن یک *miRNA* خاص را نشان می‌دهد*[7]* .

ماشین‌های بردار پشتیبان *(SVM)* یک تکنیک یادگیری ماشین نظارت شده است که بر مبنای نظریه یادگیری آماری توسعه یافته است. این روش به عنوان یک دسته‌بند باینری عمل می‌کند اصطلاح "ماشین" به این دلیل استفاده می‌شود که الگوریتم *SVM* یک خروجی باینری تولید می‌کند. اما قادر است مسائل با چندین کلاس مختلف و همچنین وظایف پیش‌بینی عددی را نیز پوشش دهد. افزون بر این، *SVM*می‌تواند با داده‌هایی که به صورت خطی قابل تفکیک نیستند نیز کار کند. مشابه به شبکه‌های عصبی، *SVM*از نگاشت غیرخطی بردارهای ورودی به یک فضای ویژگی با ابعاد بالا استفاده می‌کند که در آن می‌توان داده‌ها را از طریق ابرصفحه‌های جداکننده بهینه به‌صورت خطی تفکیک کرد*[8]* .

یکی از وظایف مهم در استفاده از داده‌های بیان ژن، شناسایی یک زیرمجموعه کوچک از ژن‌ها برای ساخت طبقه‌بندهای تشخیصی به ویژه برای بیماری‌های سرطان است. پراکندگی بیش از حد یکی دیگر از مشکلاتی است که نیازمند مدل‌سازی دقیق رابطه میانگین و واریانس داده‌های *RNA-Seq* است. بنابراین از طبقه بندی های مبتنی بر *Voom*استفاده میکنیم*[9]* .

توالی‌یابی، برخلاف داده‌های ریزآرایه‌ای که پیوسته هستند، داده‌هایی گسسته و صحیح (غیرمنفی) تولید می‌کند، به‌طوری‌که تعداد خوانش‌ها در مناطق خاص را نشان می‌دهد. در این راستا، ویتن (2011) رویکرد تحلیل تفکیک خطی پواسون را برای داده‌های *RNA-Seq* پیشنهاد داده است .با این حال، مدل پواسون ممکن است در شرایط پراکندگی بیش از حد یا زمانی که تکرارهای بیولوژیکی در دسترس‌اند، به اندازه توزیع دوجمله‌ای منفی مناسب نباشد. به این معنا که در توزیع دوجمله‌ای منفی، واریانس ممکن است بیشتر یا مساوی با میانگین باشد. هرچند مدل‌سازی با توزیع دوجمله‌ای منفی پیچیده‌تر است زیرا شامل پارامتر پراکندگی می‌شود که نیاز به تخمین دارد [10,11].

در این مطالعه، برای تحلیل داده‌های *miRNA* مربوط به دهانه رحم از نرم افزار *R* و کتابخانه ***MLSeq*** استفاده کردیم. مراحل انجام کار شامل انتخاب *miRNA* هایی با بیشترین پراکندگی و تقسیم داده‌ها به دو مجموعه آموزش و آزمون بود. سپس، عملیات پیش‌پردازش داده‌های آموزشی با استفاده از چهار روش مختلف نرمال‌سازی انجام شد:

1. ***deseq-vst:*** نرمال‌سازی با روش نسبت میانه در ***DESeq*** انجام شد و سپس تبدیل تثبیت واریانس به داده‌های نرمال‌شده اعمال گردید.
2. ***deseq-rlog:*** نرمال‌سازی با روش نسبت میانه در ***DESeq*** انجام شد و سپس تبدیل لگاریتمی تنظیم شده *(Regularized Logarithmic Transformation)* به داده‌های نرمال‌شده اعمال شد.
3. ***deseq-logcpm:*** نرمال‌سازی با روش نسبت میانه در ***DESeq*** انجام شد و سپس لگاریتم تعداد بر میلیون *(Log of Counts-Per-Million)* به داده‌های نرمال‌شده اعمال شد.
4. ***tmm-logcpm:*** نرمال‌سازی با روش میانگین بریده‌شده مقادیر ***M*** *(Trimmed Mean of M values)* انجام شد و سپس لگاریتم تعداد بر میلیون *(Log of Counts-Per-Million)* به داده‌های نرمال‌شده اعمال شد.

پس از پیش‌پردازش داده‌ها، الگوریتم‌های مختلف یادگیری ماشین برای مدل‌سازی به‌کار گرفته شدند. جزئیات این الگوریتم‌ها و روش‌های پیش‌پردازش به شرح زیر است:

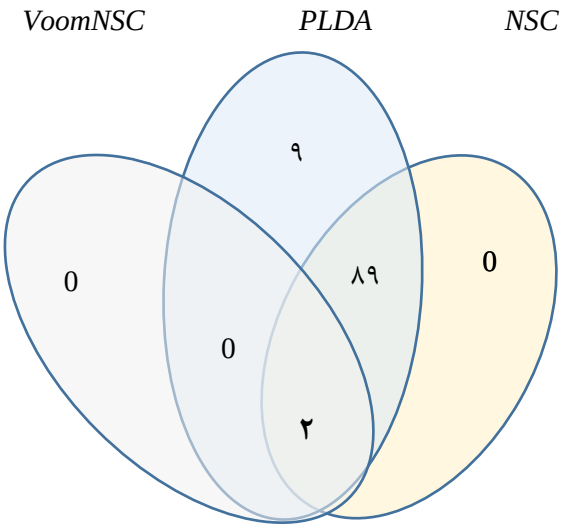
- ***SVM:*** مدل *SVM* با استفاده از هسته شعاعی *("svmRadial")* و پیش‌پردازش *"deseq-vst"* پیاده‌سازی شد.
- ***Poisson:*** مدل پواسون با استفاده از روش تفکیک خطی پواسون *("PLDA")* و پیش‌پردازش *"deseq"* پیاده‌سازی شد.
- ***Negative Binomial:*** مدل دوجمله‌ای منفی با استفاده از روش تفکیک خطی دوجمله‌ای منفی *("NBLDA")* و پیش‌پردازش *"deseq"* پیاده‌سازی شد.
- ***VOOMDLDA:*** مدل *VOOM* با استفاده از روش تفکیک خطی *VOOM ("voomDLDA")* و پیش‌پردازش *"deseq"* پیاده‌سازی شد.
- ***VOOM-NGS:*** مدل *VOOM-NGS* با استفاده از روش دسته‌بندی نزدیک‌ترین همسایه *VOOM ("voomNSC")* و پیش‌پردازش *"deseq"* پیاده‌سازی شد.

نتایج این مدل‌سازی‌ها به طور خلاصه در جدول زیر ارائه شده است:

جدول: نتایج طبقه بندی برای داده های دهانه رحم با استفاده از الگوریتم های مختلف

طبقه بندی	دقت	حساسیت
SVM	۰.۹۴۴	
PLDA	۰.۸۸۹	۱.۰۰
NBLDA	۰.۸۳۳	
NSC	۰.۸۸۹	۰.۹۱۰
voomNSC	۰.۷۲۲	۰.۰۲۰
voomDLDA	۰.۸۸۹	

صدها تا هزاران ویژگی ژنی می‌توانند برای بررسی ارتباط با یک بیماری یا شرایط خاص مورد توالی‌یابی قرار گیرند. اما معمولاً فقط یک زیرمجموعه کوچک از این ویژگی‌ها به طور معناداری میان گروه‌های مختلف متفاوت بیان می‌شود و می‌تواند به تفکیک و شناسایی گروه‌ها کمک کند. بنابراین، شناسایی ویژگی‌های با بیان دیفرانسیل *(DE)*یکی از اهداف اصلی در مطالعات *RNA-Seq* است. ما با استفاده ازطبقه کننده هایی پراکنده می توانیم این ویژگی‌ها را تشخیص دهیم که به صورت شکل زیر می باشد:



شکل*1*: ویژگی های انتخاب شده از طبقه بندی های پراکنده

۳- نتیجه‌گیری

در این مقاله با استفاده از الگوریتم‌های یادگیری ماشین به طبقه‌بندی داده های توالی نسل بعدی(*NGS*) بویژه داده‌های *(RNA-Seq)* می پردازیم. داده های مورد مطالعه در این مقاله سرطان دهانه رحم با بیان‌های 714 میکرو *RNA (miRNA)* از نمونه‌های انسانی را اندازه‌گیری کرده است. در این مجموعه داده، 29 نمونه توموری و 29 نمونه غیرتوموری از دهانه رحم وجود دارد و این دو گروه به عنوان دو کلاس جداگانه در نظر گرفته می‌شوند. علاوه بر این در این مقاله، عملکرد مدل‌ها را ارزیابی و مقایسه کردیم که نشان داد ***SVM***بالاترین دقت طبقه‌بندی را دارد. به همین ترتیب، ***voomNSC***کمترین میزان پراکندگی را در مقایسه با سایر طبقه‌بندی‌ها نشان می‌دهد. برخی از ویژگی‌ها بین طبقه‌بندی‌های پراکنده مشترک هستند. به عنوان مثال، ***voomNSC***، ***PLDA***، ***PLDA2***و ***NSC***معمولاً دو ویژگی را به‌عنوان نشانگرهای زیستی ممکن کشف می‌کنند. برای نمونه، ***voomNSC***تنها ۲٪ از تمامی ویژگی‌ها را انتخاب کرده است. اگرچه نشانگرهایی که توسط سایر طبقه‌بندی‌ها شناسایی شده‌اند نیز ممکن است حاوی اطلاعات ارزشمندی باشند، این دو ویژگی به‌ویژه از اهمیت بالایی برخوردار هستند و باید به دقت در نظر گرفته شوند.

مراجع

.1 *Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.*

.2 *Heneghan HM, Miller N, Kerin MJ. MiRNAs as biomarkers and therapeutic targets in cancer. Current opinion in pharmacology. 2010;10(5):543-50.*

.3 *Qin D. Next-generation sequencing and its clinical application. Cancer biology & medicine. 2019;16(1):4.*

.4 *Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PloS one. 2011;6(11):e28240.*

.5 *Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD, et al. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Analytical and bioanalytical chemistry. 2009;393:171-5.*

.6 *Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Duru IP, Unver T, et al. MLSeq package: Machine Learning Interface to RNA-Seq Data.*

.7 *Witten D, Tibshirani R, Gu SG, Fire A, Lui W-O. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC biology. 2010;8:1-14.*

.8 *Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. Taipei, Taiwan; 2003.*

.9 *Zararsiz G, Goksuluk D, Klaus B, Korkmaz S, Eldem V, Karabulut E, et al. voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ. 2017;5:e3890.*

.10 *Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. BMC bioinformatics. 2016;17:1-10.*

.11 *Witten DM. Classification and clustering of sequencing data using a Poisson model. 20.*11