تشخیص ژن های مرتبط با میزان تأثیر شیمی درمانی با استفاده از یک الگوریتم تجمیعی

رقیه اسماعیلی نفط چالی[۱]، محمد صنیعی آباده[۲]

[۱]دانشگاه تربیت مدرس-دانشکده مهندسی برق و کامپیوتر، r.esmaeili@modares.ac.ir
[۲]دانشگاه تربیت مدرس-دانشکده مهندسی برق و کامپیوتر، saniee@modares.ac.ir

چکیده

سرطان سلول غیرکوچک ریه، به عنوان رایج ترین نوع سرطان ریه، یکی از عوامل اصلی مرگ و میر در جهان است. این مسئله به علت تشخیص دیرهنگام این بیماری، که اغلب در مراحل پیشرفته صورت می گیرد، می باشد. جراحی به همراه شیمی درمانی کمکی، درمان های پیشنهاد شده برای سرطان سلول غیرکوچک ریه می باشد. این مطالعه برآن است تا به کمک متد پیش پردازش هوشمند، سودمندی/ بیهودگی شیمی درمانی روی بیماران مبتلا به این سرطان را پیش بینی نماید. جهت انتخاب ژن های مرتبط با شیمی درمانی از یک الگوریتم تجمیعی هدفمند انتخاب زن، بهره گرفتیم. دسته بند NB برای دسته بندی نمونه ها به کار گرفته شده است و ارزیابی نتایج، به کمک ۱۰-fold cross validation صورت پذیرفته است.

با استفاده از الگوریتم تجمیعی پیشنهادی، و به کمک ۲ ژن، به دقتی بالاتر از دقت کارهای انجام شده در این حوزه دست یافته ایم. از آنجا که شیمی درمانی فرایندی پرهزینه از بعد زمانی و اقتصادی می باشد، مدل پیشگوی ما می تواند از انجام شیمی درمانی غیر ضروری در موارد بیهوده پیشگیری نماید. هدف اصلی این پژوهش، یافتن ژن های مرتبط با شیمی درمانی و دسته بندی بیماران به کمک آن می باشد. با توجه به نتایج امید بخش به دست آمده، می توان گفت، متد انتخاب ژن هوشمند تجمیعی پیشنهادی، کیفیت دسته بندی را به طور چشم گیری بهبود بخشیده است.

واژه‌های کلیدی

داده توصیف ژنی، SVM-RFE، انتخاب ژن، دسته بندی، شیمی درمانی

۱- مقدمه

سرطان ریه به عنوان یکی از عوامل اصلی مرگ و میر در دنیا بر دو نوع می باشد: سرطان سلول کوچک ریه و سرطان سلول غیر کوچک ریه. NSCLC[۱] از نظر بافت شناسی به سه دسته squamous cell carcinoma، adenocarcinoma و large cell تقسیم می شود. جراحی، شیمی درمانی مبتنی بر سیس پلاتینیوم و رادیوتراپی راهکارهای درمانی پیشنهادی در NSCLC می باشند. از آنجا که شیمی درمانی کمکی، از بازگشت مجدد بیماری و یا پیشرفت آن جلوگیری به عمل می آورد، مطالعات متعددی تلاش کرده اند تا مزایای شیمی درمانی را روی بیماران NSCLC تحلیل و بررسی نمایند. به دلیل نتایج متضاد گزارش شده[۱-۴]، شیمی درمانی کمکی روی بیماران مبتلا به NSCLC همچنان محل مباحثه می باشد[۳]. برخی از بیماران سرطانی، تنها اندکی پس از شروع فرایند شیمی درمانی جان خود را از دست می دهند. درحالی که در مواردی بیمار، مدت طولانی پس از اتمام دوره شیمی درمانی، زنده می ماند. در عین حال مواردی هم مشاهده شده است که بیماران سرطانی بدون دریافت هیچ کمکی از شیمی درمانی برای مدتی نسبتاً طولانی در قید حیات می مانند. این نکته بیانگر آنست که همه ی بیماران سرطانی الزاماً نیازمند شیمی درمانی نیستند. علاوه براین از آنجا که شیمی درمانی حاوی مواد سمی قابل توجهی می باشد، تأثیرات ناخوشایندی نیز روی بیمار خواهد داشت که با توجه به میزان تأثیرگذاری روی حیات وی، ممکن است غیرضروی باشد. از این رو چنان چه بتوان بیماران را از حیث سودمندی/ بیهودگی شیمی درمانی کمکی، تحلیل و دسته بندی نمود، کمک بزرگی برای بیماران و جامعه پزشکان خواهد بود. در این پژوهش با استفاده از رویکرد یادگیری ماشین به تحلیل سودمندی/ بیهودگی شیمی درمانی کمکی، روی بیماران مبتلا به NSCLC می پردازیم. براساس اطلاعات ما، تحلیل رفتار ژن ها در حوزه سرطان شناسی و شیمی درمانی به منظور پیش بینی سودمندی/ بیهودگی شیمی درمانی کمکی، در بیماران مبتلا به NSCLC، به کمک الگوریتم تجمیعی مجهز به تکنیک پیش

[۱] Non-Small Cell Lung Cancer

پردازش هوشمند و با استفاده از داده های توصیف ژنی و رویکرد یادگیری ماشین کاملاً بدیع می باشد. الگوریتم تجمیعی ما شامل انتخاب ژن ترکیبی- افزایشی: آزمون کای[۲]، SVM-RFE[۳] و ماتریس همبستگی، جهت انتخاب ارزشمندترین و مرتبط ترین ژن ها و دسته بندی از نقطه نظر سودمندی/ بیهودگی دریافت شیمی درمانی به کمک دسته بند NB[۴] می باشد.

۲- کارهای مرتبط

برخی از محققین کوشیده اند مزیت شیمی درمانی را در NSCLC به کمک داده های بیمارستانی و توصیف ژنی مورد تحقیق و بررسی قرار دهند[۵-۹]. برجسته ترین پژوهش صورت گرفته در رابطه با بررسی سودمندی/ بیهودگی شیمی درمانی روی بیماران NSCLC توسط چن و همکارانش انجام شده است[۱۰]. در[۱۰] زیر مجموعه ای از ژن های سرطانی از پیش شناخته شده از بین ده ها هزار ژن حاضر در مسئله انتخاب شدند. سپس ۱۰ویژگی برتر به کمک آزمون کای (X^2) از این زیرمجموعه استخراج گردید. در نهایت ترکیب های دو تایی مختلفی از این ۱۰ ویژگی و یک داده بیمارستانی، به عنوان ورودی به دسته بند شبکه های عصبی تزریق گردید. چن و همکارانش به کمک ۳ ویژگی به دقت ۶۵,۷۱٪ دست یافتند. دراین پژوهش برآن هستیم تا با اعمال یک متد انتخاب ویژگی کارآمد بر روی تمام فضای مسئله و داده های توصیف ژنی ، ژن های مرتبط با شیمی درمانی کمکی، را شناسایی نموده و مدل پیشگوی دقیق تری را طراحی نماییم.

۳- روش کار

از آنجا که داده ی توصیف ژنی حاوی ده ها هزار پراب ژن میباشد که تنها تعداد اندکی از آن ها ارزشمند هستند، نیازمند کاهش ابعاد داده هستیم. ابتدا در مرحله انتخاب ژن، ژن های حاوی اطلاعات مفید را انتخاب نموده و در گام بعد، آن ها را به دسته بند تزریق می نماییم.

۳-۱- برچسب زدن نمونه ها

از آنجا که نزدیک به نیمی از بیماران مبتلا در مجموعه داده ی ما حدود ۴۰ ماه در قید حیات بوده اند، آستانه دسته بندی را روی ۴۰ تنظیم می نماییم. ازین رو بیمارانی که شیمی درمانی را انجام داده و بیش از ۴۰ ماه زنده مانده اند و یا بیمارانی که بدون انجام شیمی درمانی کمتر از ۴۰ ماه در قید حیات بوده اند در دسته سودمند قرار می گیرند. در سمت دیگر، بیمارانی که با وجود دریافت شیمی درمانی در کمتر از ۴۰ ماه از دنیا رفته اند و یا بیمارانی که بدون دریافت شیمی درمانی بیش از۴۰ ماه زیسته اند، در دسته بیهوده قرار می گیرند.

۳-۲- انتخاب ژن

بخش مهمی از تحلیل داده های توصیف ژنی شامل فرایند انتخاب ژن می باشد. چرا که می توان به کمک آن ژن های غیر مرتبط را از مجموعه داده حذف نمود. تکنیک مایکروآرایه در تحلیل داده ی توصیف ژنی به دانشمندان امکان بررسی ده ها هزار بیان ژن را در یک آزمایش ساده می دهد. این نوع داده با نام داده عظیم شناخته می شود که دارای ویژگی منحصر به فرد تولید حجم بالای داده با سرعت زیاد می باشد [۱۱]. به علت آنچه که از آن با نام طلسم ابعاد یاد می شود، به کاربستن روشهای آماری و محاسباتی روی این نوع داده بسیار دشوار می باشد. به منظور غلبه بر این دشواری ها روش های متعدد کاهش ابعاد و انتخاب ژن پیشنهاد شده است[۱۲-۱۳]. در مجموع باید گفت، کاهش تعداد ژن ها به منظور یافتن ژن های مرتبط و ارزشمند، فرایندی حیاتی و تأثیرگذار در تحلیل داده ی توصیف ژنی می باشد.

۳-۲-۱- الگوریتم انتخاب ژن تجمیعی افزایشی پیشنهادی،Chi-SVM-RFE به همراه ماتریس همبستگی، انتخاب ژن افزایشی بدون جایگزاری.

استفاده از رویکردهای فیلتر و توکار در انتخاب ویژگی ضعف ها و کاستی هایی به همراه دارد که می تواند روی کیفیت فرایند انتخاب ژن و به تبع آن روی دسته بندی تأثیر نامطلوبی بگذارد. جهت غلبه بر این کمبودها، ما به جای استفاده از یک متد انتخاب ژن، از ترکیبی از دو رویکرد کارا و متفاوت فیلتر و توکار، به صورت متد تجمیعی Chi-SVM-RFE بهره می گیریم.

ابتدا آزمون کای را روی کل ویژگی ها اعمال می نماییم. آزمون کای یک فیلتر تک متغیره، ساده، سریع، به سادگی مقیاس پذیر برای داده هایی با ابعاد بالا و مستقل از دسته بند می باشد[۱۳-۱۴].

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad (1)$$

O_{ij}: تعداد تکرار دیده شده

E_{ij}: تعداد تکرار مورد انتظار

به کمک آزمون کای، ویژگی ها را رتبه بندی می نماییم و زیرمجموعه ای کوچک از ژنها را برای گام بعدی الگوریتم برمی گزینیم.

بعد از محاسبه ی X^2، مجموعه ی ۱۰۰۰ ویژگی با رتبه ی بالاتر را از بین مجموعه رتبه بندی شده انتخاب می کنیم. سپس الگوریتم SVM-RFE را روی این زیرمجموعه ۱۰۰۰ تایی اعمال می نماییم. SVM-RFE توسط گایون و همکارانش جهت انتخاب ژن معرفی گردید [۱۵]. SVM-RFE یک متد توکار مبتنی بر SVM خطی است. در هر دور، ویژگی هایی با وزن پایین تر که کم ترین تأثیر را روی دسته بندی دارند، کنار گذاشته می شوند. این متد وابستگی های بین ویژگی ها را در نظر گرفته و نسبت به متدهای انحصاری، پیچیدگی محاسباتی کمتری دارد. SVM-RFE در هرمرحله ۲۰٪ از ویژگی هایی را که ارزش پایین تری دارند حذف نموده و ۸۰٪ آن ها را برای دور بعد نگه می دارد. حذف کردن ویژگی ها تا جایی ادامه می یابد که فقط ۲ ویژگی از مجموعه اولیه باقی بماند. اگر این دو ویژگی شرط مورد نظر(به دست آمدن دقتی بالاتر از دقت آستانه،۶۵,۷۱٪، که بالاترین دقت گزارش شده با کمترین تعداد ویژگی ممکن(۳ ویژگی) تا به امروز می باشد[۱۰]) را ارضاء نمایند، در مجموعه بهترین درج می شوند. در غیر این صورت در مجموعه بدترین جای می گیرند. در هر دور این دو ویژگی انتخاب شده ی برتر، از مجموعه ۱۰۰۰ تایی خارج شده و کار با باقی ویژگی ها به

[۲] Chi-square test
[۳] Support Vector Machine-Recursive Feature Elimination
[۴] Naïve Bayes

همین ترتیب ادامه می یابد. این روند انتخاب افزایشی بدون جایگزاری چندین مرتبه (بیش از ۱۵۰ دور) تکرار می شود تا جایی که (برای بیش از ۵۰ دور) هیچ افزایشی در مجموعه بهترین روی ندهد. اکنون مجموعه بهترین مشتمل بر ۴ ویژگی/پراب ژن برتر ایجاد شده است. حال از بین این مجموعه بهترین، کوچکترین زیرمجموعه با بهترین دقت ممکن را به کمک ماتریس همبستگی استخراج خواهیم کرد.

۳-۲-۲- ماتریس همبستگی

همبستگی عددی است بین۱- و ۱+ که بیانگر درجه ارتباط بین دو ویژگی می باشد (دو ویژگی را X و Y در نظر بگیرید). همبستگی مثبت بیانگر ارتباط مثبت و همبستگی منفی نشانه ی ارتباط منفی یا معکوس می باشد.

$$Correlation = \sum_{i=1}^{n} \frac{(X_i - X').(Y_i - Y')}{(n-1).S(X).S(Y)} \quad (۲)$$

X و Y: ویژگی ها

X' و Y': میانگین ها

S (X) و S (Y): انحراف استاندارد

ماتریس همبستگی این ۴ ویژگی برتر، در جدول ۱ نشان داده شده است. همانطور که می بینید میزان همبستگی بین این ۴ ویژگی بسیار ضعیف می باشد. با این حال، ما برای انتخاب مفیدترین و مرتبط ترین پراب ژن ها، دو ویژگی، با مقدار کمینه همبستگی را از ماتریس همبستگی بر می گزینیم. این دو ویژگی عبارتند از: TGFA و SEMA۶C. در ادامه دسته بندی بیماران را براساس این دو ژن انجام می دهیم.

۳-۳- دسته بندی

دراین گام نمونه ها را با کمک ژن های انتخاب شده در مرحله قبلی، دسته بندی می نماییم. بر اساس اطلاعات ما، دسته بند NB تاکنون جهت تحلیل ریسک شیمی درمانی کمکی، در بیماران مبتلا به NSCLC مورد استفاده قرار نگرفته است. به طور کلی دسته بند NB یک دسته بند احتمالاتی مبتنی بر نظریه بیز می باشد که دسته بندی را با فرض استقلال بین ویژگی ها انجام می دهد[۱۴]:

$$f_i(x) = \prod_{j=1}^{n} P(x_j/c_i)\, P(c_i) \quad (۳)$$

I: خروجی ها یا دسته های ممکن

بردار ($X_1 \ldots X_n$): بیانگر تعداد ویژگی ها

نتایج دسته بندی در بخش ۴ تحلیل و بررسی خواهد شد.

۳-۴- ارزیابی

برای بررسی کارایی الگوریتم پیشنهادی، از تکنیک ۱۰-fold cross-validation بهره می گیریم. محققین متعددی از این تکنیک به عنوان متد ارزیابی به ویژه روی داده های توصیف ژنی استفاده کرده اند[۱۰و۱۶]. در ۱۰-fold مجموعه داده به صورت خودکار به ده بخش تقسیم می شود. در هر دور، نه بخش برای آموزش و یک بخش برای آزمایش مورد استفاده قرار می گیرد. این عمل ده مرتبه تکرار شده و میانگین نتایج (دقت،specificity و sensitivity) به عنوان معیارهای ارزیابی در نظر گرفته می شوند. اگر گروهی که شیمی درمانی کمکی در آن ها اثر بخش ارزیابی شده است را به عنوان دسته yes و گروهی که شیمی درمانی روی آن ها بیهوده می باشد به عنوان دسته no در نظر بگیریم:

$$accuracy = \frac{true\ yes + true\ no}{all} \quad (۴)$$

$$sensitivity = \frac{true\ yes}{all\ yes} \quad (۵)$$

$$specificity = \frac{true\ no}{false\ yes + true\ no} \quad (۶)$$

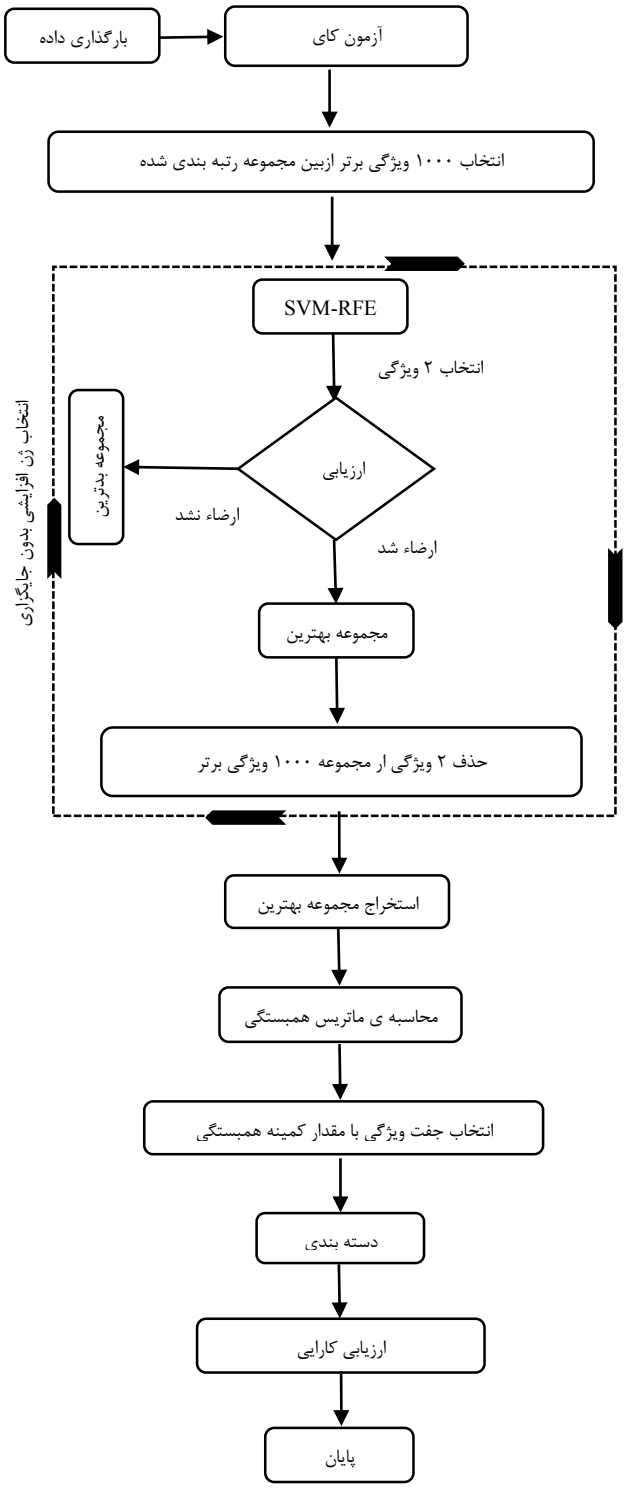
روند نمای الگوریتم پیشنهادی ما در شکل ۱ نشان داده شده است.

۴- دست آوردها و مباحثه

این برای نخستین بار است که یک الگوریتم تجمیعی نوآورانه براساس متد پیش پردازش هوشمند و روشمند، جهت ارزیابی بیماران مبتلا به NSCLC از لحاظ سودمندی/ بیهودگی شیمی درمانی ارائه می شود. در این پژوهش، به کمک رویکرد یادگیری ماشین، ژنهای مرتبط با شیمی درمانی را شناسایی کرده و در نهایت بیماران مناسب برای دریافت شیمی درمانی را پیش بینی نموده ایم. در بخش حاضر به تحلیل نتایج و دست آوردها در زیربخش های زیر می پردازیم: (۱) تحلیل مجموعه داده، (۲) دست آوردهای عددی و محاسباتی، (۳) خصوصیات بایولوژیکی ژن های انتخاب شده، (۴) مقایسه ی نتایج به دست آمده با پژوهشهای مرتبط و در نهایت (۵) مباحثه پیرامون نتایج.

۴-۱- داده

به علت قدرت داده های توصیف ژنی در پیش بینی، پیش آگهی و تشخیص بیماری ها و به ویژه سرطان، شاهد بروز تمایلات فزاینده ای جهت تحلیل داده های توصیف ژنی در این حوزه می باشیم. در این پژوهش جهت ارزیابی سودمندی/ بیهودگی شیمی درمانی روی بیماران مبتلا به NSCLC، داده های توصیف ژنی از [۱۰] استخراج شده است. این مجموعه داده شامل ۴۶۲ نمونه از ۴ انستیتو : دانشگاه مرکزی سرطان میشیگان، مرکز سرطان مافیت، مرکز سرطان یادبود اسلوان-کترینگ وانستیتوی سرطان دانا-فاربر جمع آوری شده است. پلتفرم بیان ژن ها Affymetrix Human Genome U۱۳۳A Array می باشد که توسط Affymetrix فراهم شده است. در شروع فرایند پیش پردازش، نمونه هایی که مقدار مربوط به مدت زمان حیات آن ها از دست رفته است، اطلاعات مربوط به شیمی درمانی آنها ناشناخته بوده است و نمونه های تکراری را از مجموعه داده حذف نمودیم. لازم به ذکر است که ۲ نوع مجموعه داده داریم. مجموعه داده بالینی یا بیمارستانی که برای تقسیم مشاهدات به دو گروه شیمی درمانی- سودمند: گروهی که براساس الگوریتم پیشنهادی ما شیمی درمانی برای آنها پیشنهاد می شود، و شیمی درمانی-بیهوده: که بایستی از شیمی درمانی پرهیز نمایند، مورد استفاده قرار گرفته است. نوع دوم داده، مجموعه داده توصیف ژنی می باشد که برای طراحی و توسعه مدل پیشگوی پیشنهادی، از آن بهره گرفتیم. مجموعه داده ی توصیف ژنی شامل ۲۲۲۸۳ ویژگی/پراب ژن، ۲۸۰ نمونه و ۲ کلاس:کلاس no (شیمی درمانی- بیهوده) حاوی ۱۷۱ بیمار و کلاس yes (شیمی درمانی-سودمند) حاوی ۱۰۹ بیمار، می باشد.



شکل ۱: روندنمای الگوریتم تجمیعی پیشنهادی

۴-۲- نتایج عددی

به کمک تکنیک انتخاب ژن ترکیبی Chi-SVM-RFE و ماتریس همبستگی، مجموعه ی بهترین شامل ۴ پراب ژن برتر تولید می شود. این ۴ پراب ژن عبارتند از TGFA (۲۰۵۰۱۵_s_at)، SEMA۶C (۲۰۸۱۰۰_x_at)، SPO۱۱ (۲۲۲۲۵۹_s_at) و RAB۴۰A (۲۱۷۵۸۹_at). از بین این ۴ ژن وقتی TGFA و SEMA۶C به عنوان ورودی به دسته بند NB داده می شوند بهترین نتیجه حاصل می شود. نتایج ۱۰-fold cross-validation مربوط به مدل پیشنهادی ما به کمک ۲ ژن در جدول ۲ ارائه شده است. نتایج، بیانگر بهبود در دسته بندی ریسک پذیری شیمی درمانی کمکی در بیماران سرطانی می باشد. ما به دقت ۶۸,۹۳٪ در دسته بندی، به کمک ۲ ژن TGFA (۲۰۵۰۱۵_s_at) و SEMA۶C (۲۰۸۱۰۰_x_at) و دسته بند NB دست یافتیم. اگر به دو ویژگی بالا پراب ژن SPO۱۱ (۲۲۲۲۵۹_s_at) را نیز اضافه کنیم، دقت تا ۷۰,۷۱٪ افزایش می یابد. همراه با افزایش تعداد ژن ها، دقت ابتدا افزایش و سپس کاهش می یابد (شکل ۲). در این پژوهش، دسته بندهای مختلفی ازجمله: NB، ماشین بردار پشتیبان (SVM)، k تا نزدیکترین همسایه[۵] (K-NN) و شبکه های عصبی چندلایه(MLP)[۶] مورد آزمایش قرار گرفتند. اگرچه که بهترین Sensitivity توسط K-NN و بهترین Specificity توسط SVM به دست آمده است، بالاترین دقت زمانی حاصل می شود که از دسته بند NB بهره جستیم.

۴-۳- خصوصیات بایولوژیکی ژن های انتخاب شده

ژن TGFA به عنوان یک ژن مرتبط با چندین سرطان از جمله سرطان ریه شناخته شده است. فاکتور رشد تبدیلی (TGF-α) عضوی از خانواده فاکتور رشد اپیدرمال مایتوژن ها می باشد که توسط ژن TGFA رمزنگاری شده است. این ژن با دریافت کننده ی فاکتور رشد اپیدرمال (EGFR) جهت آغاز یک سری فعالیت های بایولوژیکی از جمله تمایز، توسعه و تکثیر سلولی در ارتباط می باشد [۱۷]. همان طور که می دانیم EGFR در NSCLC نقش مؤثری دارد.EGF و TGFA رگولاتورهای بالادستی EGFR و ERBB۲ در NSCLC می باشند [۱۸-۱۹]. ژن انتخاب شده ی بعدی توسط الگوریتم پیشنهادی، SEMA۶C (۲۰۸۱۰۰_x_at) میباشد. این ژن عضوی از خانواده سمافورین ها را رمزنگاری نموده و نقش سازنده ای در بازسازی رشته های عصبی دارد. این خانواده به خاطر تأثیرگذاری در پاسخ های ایمنی و پیشرفت تومورها نیز شناخته می شوند. اگرچه که این خانواده در برخی از سرطان ها حضور پر رنگی دارند ولی SEMA۶C به عنوان یک ژن سرطانی شناخته نمی شود. پژوهش پیش رو، این دو ژن را به عنوان ژن های مرتبط با شیمی درمانی کمکی در NSCLC، پیشنهاد می کند.

۴-۴- مباحثه

با نگاهی به ماتریش همبستگی ۴ ژن برتر در می یابیم که میزان همبستگی بین آنها ناچیز می باشد (جدول ۱). همان طور که در شکل ۳ مشاهد می فرمایید کمترین همبستگی (کوچکترین حباب های آبی تیره) بین دو ژن TGFA (۲۰۵۰۱۵_s_at) و SEMA۶C (۲۰۸۱۰۰_x_at) وجود دارد. به همین دلیل، زمانی که دسته بند با این ۲ ژن آموزش داده می شود بهترین دقت حاصل می شود.

[۵] K- Nearest Neighbor

[۶] Multi-Layer Perceptron

جدول ۱: ماتریس همبستگی ۴ ژن برتر

ویژگی ها	۲۰۵۰۱۵_s_at	۲۰۸۱۰۰_x_at	۲۱۷۵۸۹_at	۲۲۲۲۵۹_s_at
۲۰۵۰۱۵_s_at	۱	-.۰۰۴۶	۰,۰۱۲	۰,۰۵۰
۲۰۸۱۰۰_x_at	-۰,۰۴۶	۱	۰,۱۰۹	۰,۰۸۷
۲۱۷۵۸۹_at	۰,۰۱۲	۰,۱۰۹	۱	۰,۰۱۹
۲۲۲۲۵۹_s_at	۰,۰۵۰	۰,۰۸۷	۰,۰۱۹	۱

جدول ۲: ماتریس درهم ریختگی داده توصیف ژنی- با ۲ ژن

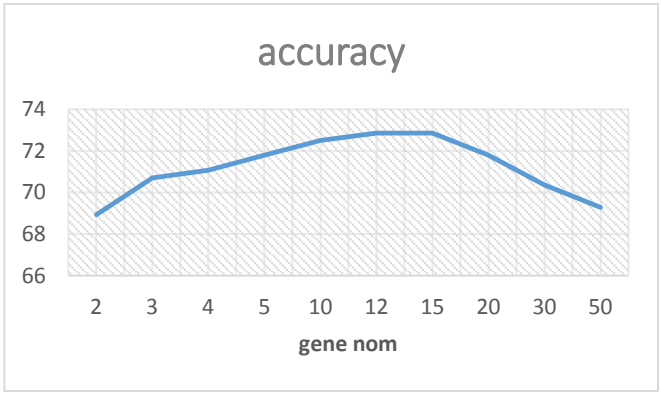
TRUE	ACT ineffectiveness	ACT usefulness	precision
pred. ACT/ ineffectiveness	۱۶۳	۷۹	۶۷,۳۶٪
pred. ACT usefulness	۸	۳۰	۷۸,۹۵٪
recall	۹۵,۳۲٪	۲۷,۵۲٪	
Accuracy			۶۸,۹۳٪

با کمی دقت در ماتریس همبستگی، می توان گفت که میزان همبستگی بین ژن ها بسیار اندک و نزدیک به ۰ است، تا جایی که می توان ادعا کرد که این ویژگی ها مستقل از هم می باشند. این بیانگر کیفیت و قدرت الگوریتم انتخاب ژن پیشنهادی ما است. این جاست که به اهمیت و ضرورت استفاده از یک متد انتخاب ژن مؤثر و کارآمد، که با حذف ویژگی های افزونه و غیرمرتبط، ارزشمندترین آن ها را انتخاب نماید، پی می بریم. در این پژوهش استفاده از متد انتخاب ژن تجمیعی Chi-SVM-RFE به همراه ماتریس همبستگی، به ما در استخراج ژن های حاوی اطلاعات ارزشمند کمک شایانی کرده و تأثیر بسزایی در کیفیت فرایند دسته بندی داشته است.

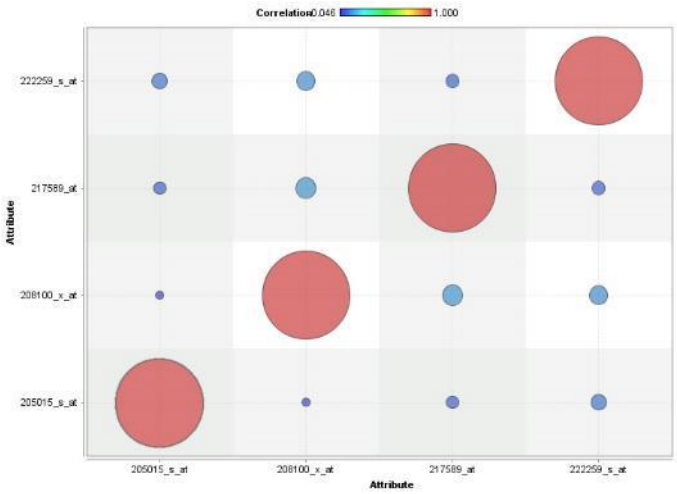
۴-۵- مقایسه ی دست آوردها

بهترین دقت گزارش شده به همراه کمترین تعداد ویژگی جهت بررسی سودمندی / بیهودگی شیمی درمانی در بیماران مبتلا به NSCLC، ٪۶۵,۷۱ می باشد که به کمک ۳ ویژگی و دسته بند MLP، به دست آمده است[۱۰]. در حالی که الگوریتم تجمیعی ما به کمک ۲ ژن، دقت ٪۶۸,۹۳ و با ۳ ژن دقت ٪۷۰,۷۱ را به دست می دهد. ویژگی های برگزیده، محصول الگوریتم انتخاب ژن تجمیعی افزایشی هستند که برکل فضای مسئله اعمال شده است. این درحالیست که [۱۰] تنها روی زیرمجموعه ای از ژن های ازپیش شناخته شده ی سرطانی که از سایر مقالات استخراج شده است کار کرده است و هیچ متد پیش پردازش مشخصی که روی تمام فضای مسئله اعمال شود ارائه نداده و از مقابله با چالش ابعاد بالای داده های ژنی پرهیز نموده است.

الگوریتم خلاقانه ی ما با جستجویی هوشمند روی تمام ویژگیها، با طلسم ابعاد بالای داده های توصیف ژنی مواجهه نموده و تلاش دارد از زاویه ای متفاوت به حل مسئله بپردازد. نتایج مقایسه ای مربوط به ۳ ویژگی به دست آمده توسط الگوریتم پیشنهادی با سایر کارهای انجام شده در این حوزه در جدول ۳ آمده است. همان طور که پیداست مدل پیشگوی ما در پیش بینی ریسک پذیری شیمی درمانی کمکی روی بیماران مبتلا به NSCLC در همه موارد عملکرد بهتری داشته است.



شکل۲: تغییرات ۱۰-fold cross-validation همراه با افزایش تعداد ژنها



شکل ۳: نمودار همبستگی حبابی ۴ ژن. افزایش همبستگی از آبی پر رنگ به سمت قرمز.اندازه و رنگ حباب ها: همبستگی

۵- پیشنهادات

به منظور پیش گیری از هدر رفت منابع از یک سو و انجام شیمی درمانی های هدفمند روی بیماران سرطانی از سوی دیگر، بررسی های تخصصی تر روی رفتار دو ژن SEMA۶C و TGFA پیشنهاد می گردد. علاوه بر این می توان تأثیر گذاری عواملی همچون کشیدن سیگار، میزان موفقیت آمیز بودن عمل جراحی، عفونت ها و خونریزی های احتمالی حین و بعد از عمل را در سودمندی/بیهودگی شیمی درمانی نیز مورد بررسی قرار داد.

جدول ۳: نتایج مقایسه ای - با ۳ ویژگی ورودی

	NB		SVM		K-NN		MLP	
	[۱۰]	ما	[۱۰]	ما	[۱۰]	ما	[۱۰]	ما
۱۰-fold cross validation	N/A	۷۰,۷۱٪	۶۱,۴۲٪	۶۹,۶۴٪	۴۵,۷۱٪	۶۸,۹۳٪	۶۵,۷۱٪	۷۰,۳۶٪
Sensitivity	۱۵,۶۰٪	۳۳,۱۸٪	۴۲,۳۰٪	۲۸,۶۴٪	۹۰,۴۰٪	۴۷,۴۳٪	۴۴,۴۰٪	۳۴٪
Specificity	۹۸,۸۰٪	۹۴,۷۷٪	۷۷,۳۰٪	۹۵,۹۲٪	۱۰,۲۰٪	۸۲,۳۹٪	۸۷,۲۰٪	۹۳,۶۳٪

۶- نتیجه گیری

برای نخستین بار در حوزه سرطان شناسی و تحلیل تأثیر شیمی درمانی کمکی، روی بیماران مبتلا به NSCLC، یک الگوریتم تجمیعی مبتنی برپیش پردازش هدفمند که قادر به غربال بیماران از لحاظ سودمندی/ بیهودگی شیمی درمانی می باشد، ارائه شده است. در این پژوهش انتخاب ژن های مرتبط با شیمی درمانی به کمک الگوریتم تجمیعی انتخاب ژن، که بر روی تمام فضای مسئله اعمال می شود، گامی اساسی در طراحی و پیاده سازی مدل پیشگو می باشد. علاوه بر این در این پژوهش جهت ساخت مدل از دسته بند NB بهره گرفته ایم که تاکنون در این حوزه استفاده نشده است. هدف شیمی درمانی کمک به بیماران واجد شرایط میباشد، ما معتقدیم که مدل پیشگوی مبتنی بر الگوریتم تجمیعی پیشنهادی، برای یافتن بیمار درست، کمک بزرگی خواهد بود.

مراجع

[۱] R. Arriagada, A. Dunant, J.P. Pignon, et al., Long-term results of the international adjuvant lung cancer trial evaluating adjuvant Cisplatin-based chemotherapy in resected lung cancer, J. Clin. Oncol. ۲۸ (۱) (۲۰۱۰) ۳۵–۴۲.

[۲] G.M. Strauss, J.E. Herndon, M.A. Maddaus, et al., Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB ۹۶۳۳ with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups, J. Clin. Oncol. ۲۶ (۳۱) (۲۰۰۸) ۵۰۴۳–۵۰۵۱.

[۳] J.Y. Douillard, Adjuvant chemotherapy for non-small-cell lung cancer: it does not always fade with time, J. Clin. Oncol. ۲۸ (۱) (۲۰۱۰) ۳–۵.

[۴] C.A. Butts, K. Ding, L. Seymour, et al., Randomized phase III trial of vinorelbine plus cisplatin compared with observation in completely resected stage IB and II non-small-cell lung cancer: updated survival analysis of JBR-۱۰, J. Clin. Oncol. ۲۸ (۱) (۲۰۱۰) ۲۹–۳۴.

[۵] H. Tang, G. Xiao, C. Behrens, et al., A ۱۲-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients, Clin. Cancer Res. ۱۹ (۶) (۲۰۱۳) ۱۵۷۷–۱۵۸۶.

[۶] R.K. Van laar, Genomic signatures for predicting survival and adjuvant chemotherapy benefit in patients with non-small-cell lung cancer, BMC Med. Genom. ۵ (۲۰۱۲) ۳۰.

[۷] D.T. Chen, Y.L. Hsu, W.J. Fulp, et al., Prognostic and predictive value of a malignancy-risk gene signature in early-stage non-small Cell lung cancer, J. Natl. Cancer Inst. ۱۰۳ (۲۴) (۲۰۱۱) ۱۸۵۹–۱۸۷۰.

[۸] Y. Xie, J.D. Minna, Non-small-cell lung cancer mRNA expression signature predicting response to adjuvant chemotherapy, J. Clin. Oncol. ۲۸ (۲۹) (۲۰۱۰) ۴۴۰۴–۴۴۰۷.

[۹] Y. - C. Chen et al., Risk classification of cancer survival using ANN with gene expression data from multiple laboratories, Computers in Biology and Medicine ۴۸ (۲۰۱۴) ۱–۷.

[۱۰] Y.-C. Chen et al., Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer, Journal of Biomedical Informatics ۵۶ (۲۰۱۵) ۱–۷.

[۱۱] S. Karimi, M. Farrokhnia, Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique, Chemometrics and Intelligent Laboratory Systems. ۱۳۹ (۲۰۱۴) ۶–۱۴

[۱۲] Y. Saeys, I. Inza, P. L. naga, a review of feature selection techniques in bioinformatics, Bioinformatics. ۲۳ (۲۰۰۷) ۲۵۰۷–۲۵۱۷.

[۱۳] V. Elyasigomari, et al., Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization, Appl. Soft Comput. J. (۲۰۱۵).

[۱۴] J. Novakovic, P. Strbac, D. Bulatovic, Toward Optimal Feature Selection, Yugoslav Journal of Operations Research ۲۱ (۲۰۱۱) ۱۱۹-۱۳۵.

[۱۵] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning. ۴۶ (۲۰۰۲) ۳۸۹–۴۲۲.

[۱۶] E. Lotfi, A. Keshavarz, Gene expression microarray classification using PCA–BEL, Computers in Biology and Medicine ۵۴ (۲۰۱۴) ۱۸۰–۱۸۷.

[۱۷] AA. Badawy, A. El-Hindawi, O. Hammam, M. Moussa, S. Gabal, N. Said, Impact of epidermal growth factor receptor and transforming growth factor-α on hepatitis C virus-induced hepatocarcinogenesis. APMIS ۱۲۳ (۱۰) (۲۰۱۵) ۸۲۳–۳۱.

[۱۸] Hsieh et al., Transcription factor and microRNA-regulated network motifs for cancer and signal transduction networks, BMC Systems Biology (۲۰۱۵) ۹ (Suppl ۱): S۵.

[۱۹] T. Mukohara, S. Kudoh, K. Matsuura, S. Yamauchi, T. Kimura, H. Wanibuchi, et al., Activated Akt expression has significant correlation with EGFR and TGF-alpha expressions in stage I NSCLC, Anticancer research (۲۰۰۴) ۲۴ ۱۱-۱۷.