مقاله پژوهشی

لیزر پزشکی؛ ۱۳۸۷، دوره ۵، شماره ۲: صفحات ۲۵-۱۶.

استخراج الگوی پروتئینی از داده‌طیف‌جرمی‌لیزری جهت تشخیص سرطان پستان با استفاده از الگوریتم داده‌کاوی

حسین منتظری کردی[۱]
محمدحسین میران‌بیگی[۲]
محمدحسن مرادی[۳]
معصومه نجفی[۴]

[۱] دانشجوی دکتری تخصصی مهندسی پزشکی، دانشگاه تربیت مدرس
[۲] استادیار مهندسی پزشکی، دانشگاه تربیت مدرس
[۳] دانشیار مهندسی پزشکی، دانشگاه صنعتی امیرکبیر
[۴] مرکز بیماریهای پستان، جهاد دانشگاهی علوم پزشکی تهران

نویسنده مسئول: محمدحسین میران‌بیگی، گروه مهندسی پزشکی، بخش برق، دانشکده فنی، دانشگاه تربیت مدرس، تهران، ایران- تلفن: ۸۲۸۸۳۳۷۰-(۰۲۱)۹۸+
پست الکترونیکی: miranbmh@modares.ac.ir

خلاصه

زمینه و هدف: یکی از مشکلات اساسی در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص زودرس آن می‌باشد. سرطان پستان یکی از بیماری‌های شایع در بین زنان می‌باشد که تشخیص در مراحل اولیه می‌تواند تأثیر بسزایی در میزان مرگ و میر زنان داشته باشد. در حال حاضر، نشانگرهای تومور مناسب برای تشخیص زودرس این بیماری وجود ندارد. واکنش‌های شیمیایی درون یک عضو زنده می‌تواند بصورت الگوهایی پروتئینی در مایعاتی نظیر خون، خلط و ادرار انعکاس داده شود. طیف‌سنج جرمی جذب- یونیزاسیون لیزری سطحی ارتقاء یافته زمان پروازی یک ابزار مناسب جهت تهیه پروفایل‌های پروتئینی از نمونه‌های بیولوژیک می‌باشد. ارایه یک روش داده‌کاوی جهت انتخاب نشانگرهای حیاتی تفکیک کننده گروه‌های سالم از سرطانی، جزء چالش‌های مهم در تحلیل الگوهای پروتئینی محسوب می‌شود.

روش بررسی: در این تحقیق، داده‌های پروفایل پروتئینی خونابه بیماران مبتلا به سرطان پستان مورد تحلیل قرارگرفت. با ارایه یک مدل ریاضی و استفاده از تبدیل موجک گسسته، اغتشاشات خط زمینه و نویز الکتریکی در مرحله پیش‌پردازش حذف گردید و سپس، تمام سیگنال‌های طیف جرمی نرمالیزه شدند. در این مقاله، یک الگوریتم داده کاوی ترکیبی مبتنی بر سه معیار آزمون آماری، اندازه تفکیک‌پذیری کلاس و امتیازدهی نقاط، معرفی شده است. با روش پیشنهاد شده، بهترین زیرمجموعه پروتئین‌ها از بین ۱۳۴۸۸ نقطه موجود با حفظ ارزش اطلاعاتی و قدرت تفکیک‌پذیری انتخاب شد و برای تعیین نشانگرهای حیاتی استفاده گردید. با استفاده از روش ارزیابی متقابل K چرخشی، نمونه‌های موجود در مجموعه داده به دو دسته یادگیری و آزمون، بطور تصادفی تقسیم شدند. حداقل آستانه برای آمارگان T مقدار ۱/۹۶ انتخاب شد. الگوریتم داده‌کاوی به نقاط باقیمانده از مرحله آستانه‌دهی اعمال شد و بهترین زیرمجموعه ویژگی‌ها شامل نشانگرهای حیاتی با قدرت تمایز بالا انتخاب گردید.

یافته ها: با استفاده از روش تحلیل تمایز خطی، تعداد ۱۹ پروتئین بعنوان نشانگر حیاتی برگزیده شد که توانست نمونه‌های سالم و سرطانی را با دقت تشخیص ۱۰۰٪، حساسیت ۱۰۰٪ و قطعیت ۱۰۰٪ از هم تمیز دهد.

بحث و نتیجه‌گیری: با تولید اطلاعات کامل از نمونه‌های بیولوژیک می‌توان از آنها در تشخیص بیماری‌های با عوامل تشخیصی ضعیف نظیر سرطان استفاده نمود. تشخیص بیماری نمونه‌ای از تفکیک الگو می‌باشد. در این مقاله، یک الگوریتم داده کاوی جهت انتخاب بهترین زیرمجموعه از پروتئین‌ها معرفی گردید. روش پیشنهادی نشان داد که با کاهش تعداد نشانگرهای حیاتی منتخب، که از مزیت‌های این روش می‌باشد، قدرت تفکیک‌پذیری از سطح مناسبی برخوردار است. نتایج بدست آمده تأکید دارد که انتخاب مناسب زیرمجموعه پروتئین‌های شاخص تأثیر بسزایی در تعیین نشانگرهای حیاتی جهت تشخیص صحیح بیماری دارد.

واژه‌های کلیدی: سرطان پستان، پروتیین‌شناسی، طیف‌جرم، الگوریتم داده‌کاوی، نشانگرحیاتی، تشخیص الگو

مقدمه

یکی از مشکلات اساسی و حل نشده در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص بموقع و زودرس آن می‌باشد. متأسفانه، سرطان پستان یکی از بیماری‌های شایع در بین زنان می‌باشد که تشخیص در مراحل اولیه می‌تواند تأثیر بسزایی در کاهش میزان مرگ و میر زنان داشته باشد [۱]. سالیانه در آمریکا بیش‌از ۲۰۰۰۰۰ مورد جدید از بیماری سرطان پستان گزارش می‌شود که از این تعداد بیش از ۹۹/۳٪ مربوط به زنان می‌باشد. حدود ۳۱٪ از سرطان‌های تشخیص داده شده برای جمعیت زنان در آمریکا به سرطان پستان اختصاص دارد که حایز بالاترین نرخ مشاهده در این گروه می‌باشد. میزان مرگ و میر ناشی از سرطان پستان در گروه زنان تقریبا ۱۵٪ است که بعد از سرطان ریه دومین رتبه از نظر عاملیت مرگ به این بیماری تعلق دارد [۲-۱].

با توجه به اطلاعات حوزه علم ژنتیک، واکنش‌های شیمیایی درون یک عضو زنده می‌تواند بصورت الگوهایی پروتئینی در مایعاتی نظیر خون، خلط و ادرار انعکاس داده شود [۳]. واژه پروتئین‌شناسی[1] به علومی اطلاق می‌شود که امکان مقایسه کمی و کیفی و همچنین تمایز قائل شدن بین پروتئینها تحت شرایط مختلف پاتولوژیک را فراهم نماید. حوزه فعالیت پروتئین‌شناسی بطور عام به هر نوع تکنولوژی یا

[1] Proteomics

تکنیک پردازش اطلاعاتی مربوط می‌شود که بتواند داده‌های پروتئینی را با مقیاس بالا تولید نموده و یا آنها را مورد تحلیل قراردهد [۴].

یکی از تکنیک‌های جدید مورد استفاده جهت استخراج اطلاعات پروتئینی از نمونه‌های وابسته به موجودات زنده، طیف‌سنج جرمی جذب- یونیزاسیون لیزری سطحی ارتقاء یافته زمان پروازی[2] (SELDI-TOF) می‌باشد. تحلیل محتوای اطلاعاتی طیف جرمی یک روش سریع و بطور نسبی کم هزینه در تشخیص بیماری بدون ایجاد هرگونه عوارض جانبی می‌باشد، که می‌تواند امکان بالقوه غربالگری سرطان را فراهم سازد. در بسیاری از کارهای غربالگری، داده ورودی بوسیله ویژگی‌های زیادی توصیف شده‌است که تنها تعداد کمی از آنها برای پیش‌بینی عامل بیماری یا برچسب کلاس مناسب هستند. از اینرو، استفاده از تکنیک‌های استخراج یا انتخاب ویژگی از اهمیت ویژه‌ای در تحلیل و تفسیر این نوع داده‌ها برخوردار می‌باشد.

استفاده از پروفایل پروتئینی تهیه شده توسط SELDI-TOF برای کشف نشانگرهای حیاتی جدید با قدرت تمایز بالا بین گروه‌های سالم و سرطانی در حال توسعه می‌باشد[۵]. پتریکون و همکارانش [۶] با استفاده از یک بسته نرم افزاری داده‌کاوی مبتنی بر شبکه عصبی و الگوریتم ژنتیک، تحقیقاتی برای کشف نشانگرهای حیاتی از داده‌های پروتئینی سرطان تخمدان انجام دادند. هو و همکارانش [۷] از الگوریتم‌های داده‌شناسی‌حیاتی[3] مبتنی بر شبکه عصبی و تحلیل تمایز[4] برای استخراج الگوهای پروتئینی بهره گرفتند. شین و همکارش [۸] یک روش انتخاب ویژگی مبتنی بر خوشه‌بندی را توسعه دادند و از آن برای تحلیل داده‌های پروتئینی استفاده نمودند. ژانگ و همکارانش [۹] جهت استخراج نشانگرهای‌حیاتی از یک روش مبتنی‌بر حذف متوالی ویژگیها بر پایه الگوریتم یادگیری ماشین حامی‌بردار (R-SVM) بهره گرفتند که از این شیوه در تحلیل داده‌های طیف‌جرمی سرطان پستان استفاده کردند.

یکی از معایب عمده اکثر تحقیقات در زمینه تحلیل داده طیف‌جرمی عدم اتخاذ یک روش مناسب در پیش‌پردازش این نوع از داده‌ها برای حذف اغتشاشات ذاتی طیف بوده است. بائولینگ وو و همکارانش [۱۰] در مقاله خود به پیش‌پردازش صحیح داده بعنوان یک مرحله حساس در تحلیل طیف‌جرم اشاره نمودند و انتخاب متغیر را یکی از چالش‌های دیگر این حوزه برشمردند. از دیگر معایب عمده می‌توان به استفاده از روش‌های آستانه‌دهی جهت انتخاب اولیه پروتئین‌ها اشاره نمود که در بیشتر موارد با انتخاب یک آستانه بالا تعداد زیادی از متغیرها بدون بررسی عملکردشان حذف می‌شوند.

موریس و همکارانش [۱۱] نیز بر استخراج و انتخاب ویژگی‌های مناسب از داده طیف‌جرمی با یک روش معنی‌دار علمی تأکید داشته‌اند.

در حال حاضر، برخی از نشانگرهای تومور، مانند CA15.3، در تشخیص سرطان پستان مورد استفاده قرار می‌گیرند که دارای حساسیت (۲۳٪) و قطعیت (۶۹٪) می‌باشند [۱۲]. با نشانگرهای موجود، تشخیص بیماری در مراحل بالای پاتولوژیک امکان‌پذیر می‌باشد. تحلیل پروفایل پروتئینی خونابه می‌تواند امیدی در راستای دستیابی به نشانگرهای حیاتی جهت تشخیص زودرس بیماری سرطان باشد [۱۳].

در این مطالعه، مرحله پیش‌پردازش بعنوان یکی از چالش‌های اصلی تحلیل داده‌های طیف‌جرمی مورد بررسی قرار گرفته است. نشان داده می‌شود که با استفاده از قابلیت تبدیل موجک، می‌توان اغتشاشات داده را بخوبی حذف نمود. در این تحقیق، یک روش ترکیبی داده‌کاوی مبتنی بر سه معیار: آزمون معنی‌دار بودن آماری، اندازه تفکیک‌پذیری کلاس و توصیف‌گر شکل معرفی شده‌است و روش پیشنهادی در تحلیل داده طیف‌جرمی لیزری SELDI-TOF در تشخیص سرطان پستان مورد استفاده قرار گرفته است.

روش بررسی

در این تحقیق، از داده‌های طیف‌جرمی لیزری نمونه‌های خونابه بعنوان الگوهای پروتئینی ورودی برای الگوریتم پیشنهادی استفاده شده است. طیف‌جرمی حاوی اطلاعاتی نظیر جرم پروتئین‌های موجود در یک ترکیب، به همراه فراوانی یا غلظت آنها می‌باشد که از آن بعنوان سیگنال یا پروفایل پروتئینی یاد می‌شود. در این تحقیق، از طیف جرمی‌لیزری جهت انتخاب نشانگرهای‌حیاتی متمایز کننده نمونه‌های گروه‌های سالم و سرطانی با قدرت تفکیک‌پذیری بالا استفاده شده است.

۱- طیف‌سنج جرمی لیزری

در یک طیف‌سنج جرمی ابتدا نمونه در مرحله آماده‌سازی جهت یونیزه کردن مولکول‌ها قرار می‌گیرد. در طیف‌سنج‌های کروماتوگراف، مرحله یونیزاسیون با حرارت دادن محلولی حاوی نمونه صورت می‌پذیرد که برای نمونه‌های بیولوژیک بدلیل ناپایداری در برابر حرارت این امکان وجود ندارد. با ابداع و توسعه طیف‌سنج‌های جرمی لیزری مشکل یونیزاسیون نمونه مرتفع گردیده است [۱۴].

در طیف‌سنج جرمی لیزری، نمونه بیولوژیک بر روی یک صفحه یا ماتریس جذب کننده انرژی قرار می‌گیرد. پرتوهای لیزر به صفحه تابانده شده و انرژی آن توسط ماتریس جذب می‌شود. سپس، قسمتی از انرژی جذب شده به ماده بیولوژیک بازپس داده می‌شود که موجب یونیزه شدن مولکول‌های ماده می‌گردد. این یون‌ها تحت تأثیر یک ولتاژ بالا در یک لوله خلاء به سمت آشکارساز حرکت می‌کنند که

[2] Surface-Enhanced Laser Desorption-Ionization Time-of-Flight Mass Spectrometry
[3] Bioinformatics
[4] Discriminant Analysis

زمان رسیدن مولکول‌ها به آشکارساز برحسب جرم آنها متفاوت می‌باشد. خروجی آشکارساز یک طیف می‌باشد که محور افقی بیانگر نسبت جرم به بار و محور عمودی نیز نشان دهنده غلظت یا فراوانی یک مولکول است [۱۴]. شکل شماره ۱ نمایی از یک طیف‌سنج جرمی لیزری را با جزئیات مراحل مختلف تهیه طیف‌جرم نشان می‌دهد.

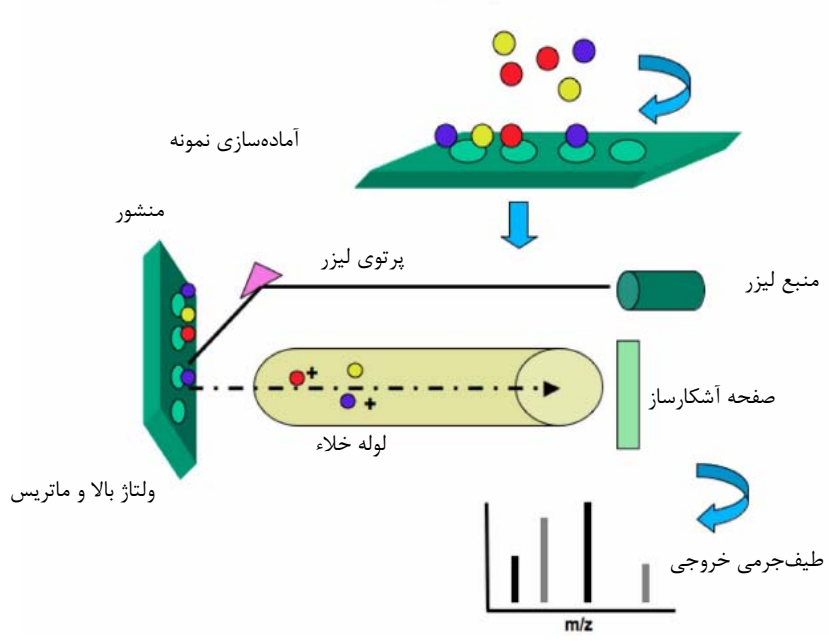
۲- توصیف داده

در تحقیق حاضر، مجموعه داده طیف جرمی لیزری سرطان پستان حاصل از دستگاه طیف‌سنج جرمی سطحی ارتقاء یافته جذب یونیزاسیون لیزری زمان پروازی[5] (SELDI-TOF) مورد استفاده قرار گرفته است. مجموعه داده شامل۱۶۷ نمونه طیف جرمی لیزری می‌باشد که از نمونه خونابه افراد مبتلا به بیماری سرطان پستان (گروه سرطان) و اشخاص سالم بدون عارضه (گروه کنترل یا سالم) تهیه شده‌است. این ۱۶۷ نمونه طیف‌جرمی به تفکیک شامل ۵۵ نمونه سرطانی با HER2 مثبت، ۳۵ نمونه با ER\PR مثبت، ۶۴ نمونه سالم و ۱۳ نمونه از یک فرد سالم می‌باشد. نمونه‌های خون از یک مطالعه تحت کنترل وابسته به مرکز سرطان دانا فاربر دانشگاه هاروارد جمع‌آوری شده‌است که داده‌های طیف جرمی سرطان پستان از مسئول نگهدارنده این داده‌ها اخذ گردیده است [۱۵].

هر طیف بطور جداگانه دارای ۱۳۴۸۸ نقطه داده می‌باشد که در محدوده جرمی بین ۰ تا ۲۰ کیلو دالتون قرار دارند. این طیف‌ها دارای دو بعد می‌باشند که در یک بعد جرم پروتئین‌ها بر حسب دالتون و در بعد دیگر میزان فراوانی مربوط به هر جرم لیست شده‌است. شکل شماره ۲، دو نمونه از منحنی طیف جرمی را برای داده‌های تحت مطالعه نشان می‌دهد که به ترتیب مربوط به یک فرد سالم و بیمار دارای سرطان پستان می‌باشد. محور افقی نشان‌دهنده مقادیر نسبت جرم به بار می‌باشد و محور عمودی بیانگر فراوانی هر پروتئین بر حسب آشکارسازی توسط تحلیل‌گر جرم دستگاه طیف‌سنج می‌باشد.

۳- مدل ریاضی طیف‌جرمی

قبل از انجام هرگونه عملیات پردازشی روی سیگنال طیف جرمی، ابتدا یک مدل ریاضی برای آن ارایه می‌شود که در این مدل اثرات شیمیایی و الکتریکی دستگاه طیف‌سنج در نظر گرفته شده است.



شکل۱- نمایی از مراحل مختلف دستگاه طیف‌سنج جرمی لیزری جهت تهیه پروفایل پروتئینی از نمونه بیولوژیک

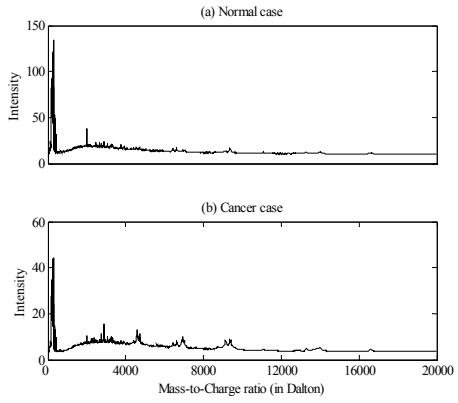
فرض می‌شود که n طیف نمونه برداری‌شده در بازه زمانی T از زمان پرواز در فواصل زمانی $t_j, j=1,\ldots,T$ مشاهده شده است. مدل زیر را می‌توان برای سیگنال طیف جرمی در نظر گرفت [۱۶]:

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \varepsilon_{ij} \qquad (۱)$$

در این مدل، $y_i(t_j)$ شدت سیگنال برای هر جرم مجزا؛ $B_i(t_j)$ خطای سیستم ناشی از ماده شیمیایی بکار رفته که به آن خط زمینه نیز اطلاق می شود؛ $S_i(t_j)$ سیگنال اصلی ناشی از مولکول‌های پرونئینی موجود در نمونه بیولوژیک؛ N_i عامل مقیاس ضربی، ε_{ij}

[5] Surface Enhanced Laser Desorption-Ionization Time-of-Flight

معرف اغتشاش الکتریکی دستگاه طیف سنج جرمی با توزیع گوسی می باشد.



شکل۲- نمونه‌هایی از طیف‌جرمی متعلق به دو فرد سالم و بیمار از مجموعه داده سرطان پستان: (a) کنترل، (b) سرطانی

۴- پیش‌پردازش

داده سطری بدست آمده از طیف‌سنج جرمی SELDI-TOF جهت استفاده مستقیم در مرحله استخراج ویژگی مناسب نمی‌باشد. هدف از مرحله پیش‌پردازش ارتقای کیفیت داده به نحوی است که منجر به کاهش خطاهای سیستماتیک در مراحل بعدی گردد. با توجه به مدل ارائه شده در رابطه (۱)، امکان تفکیک اغتشاش خط‌زمینه و نویز الکتریکی از یکدیگر وجود ندارد. بنابراین نیاز به نگاشت طیف‌جرمی در فضایی که جدا کردن این دو اغتشاش امکان‌پذیر باشد، ضروری خواهد بود.

تبدیل موجک گسسته[6] یکی از نگاشت‌های مفید در بسط سیگنال به فضای پایه متعامد می‌باشد. در فضای توابع با انرژی محدود، $L^2(\Re)$، یک موجک پایه متعامد با شیفت‌دادن و تغییر مقیاس تابع موجک مادر Ψ بصورت $\Psi_{j,k}(x)=2^{-j/2}\Psi(2^{-j}x-k)$ بدست می‌آید که در این رابطه j و k مقادیر صحیح می‌باشند. یک تابع $f(x)$ بطول N می‌تواند توسط تابع مقیاس $\Phi(x)$، که تابع موجک $\Psi(x)$ از روی آن ساخته می‌شود، و همچنین تابع موجک مادر بصورت زیر بسط داده شود[۱۷]:

$$f(x)=\langle f,\Phi\rangle\Phi(x)+\sum_{j=0}^{J}\sum_{k=0}^{2^j-1}\langle f,\Psi_{j,k}\rangle\Psi_{j,k}(x) \qquad (۲)$$

در تحلیل موجک، یک سیگنال یا بردار (نظیر منحنی طیف جرم) به مجموعه‌ای از تقریبات[7] و جزییات[8] که شکل ساده‌تری از تابع می‌باشند، با مقیاس بندی و شیفت دادن تجزیه می‌شوند. در تبدیل موجک چند وضوحی[9] فضای نگاشت مربوط به هر تقریب، می‌تواند به دو زیر فضای تقریب و جزئیات دیگر تبدیل شود [۱۸].

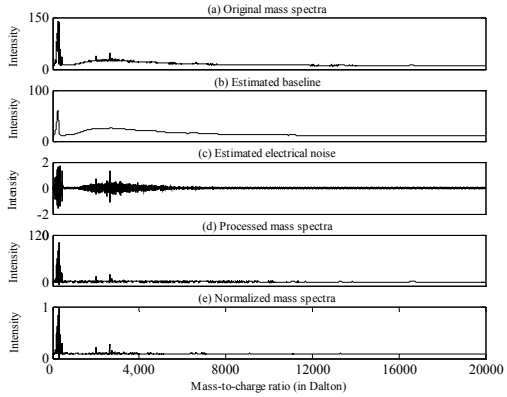
در این تحقیق، از تبدیل موجک جهت حذف توام خط زمینه و نویز الکتریکی استفاده شده است. وقتی‌که تبدیل ویولت به رابطه (۱) اعمال شود آنگاه سیگنال مشاهده‌شده به دو قسمت تقریبات و جزئیات تجزیه می‌شود که اغتشاش خط زمینه در ضرایب تقریبات و نویز الکتریکی در ضرایب جزئیات ظاهر می‌شوند [۱۹و۲۰]. در این مطالعه، تکنیک تخمین هوشمند خط‌زمینه به ضرایب تقریبات برای حذف خط‌زمینه اعمال گردید [۲۱]. جهت حذف نویز الکتریکی، روش آستانه‌دهی نرم مورد استفاده قرارگرفت [۲۲] و مقدار مناسب آستانه با استفاده از آمارگان مرتبه بالا[10] محاسبه گردید [۲۳].

پس از حذف اثرات ناشی از اغتشاش‌های خط زمینه و منابع الکتریکی، سیگنال بمنظور حذف اثر ضریب مقیاس با استفاده از رابطه (۳) نرمالیزه گردید:

$$NS=\frac{S-\min(S)}{\max(S)-\min(S)} \qquad (۳)$$

در این رابطه، NS به مقدار شدت طیف نرمالیزه شده و S به مقدار آن قبل از نرمالیزاسیون اشاره دارد. همچنین عبارات min(S) و max(S) به ترتیب مقادیر حداقل و حداکثر شدت در هر نقطه از سیگنال را نشان می‌دهند. در شکل شماره ۳ نمونه‌ای از سیگنال طیف‌جرمی قبل و بعد از پیش پردازش مشاهده می‌شود. برای پیش‌پردازش هر طیف‌جرمی از ویولت‌دابیشیز مرتبه ۴ استفاده گردید. همچنین، طیف جرمی نرمالیزه شده نیز در شکل قابل رؤیت است.

روش ترکیبی داده‌کاوی در مواردی که تعداد نقاط ویژگی از تعداد نمونه‌ها خیلی بیشتر می‌باشد، استخراج یا انتخاب ویژگی حایز اهمیت خواهد بود. زیرا استفاده از تمام ویژگی‌ها غیر عملی بوده و موجب کاهش کارآیی مدل می‌گردد [۲۴]. از اینرو، هر نوع الگوریتم یا روش



شکل۳- نمونه‌ای از یک طیف‌جرمی قبل و بعد از پیش‌پردازش (a) سیگنال اصلی، (b) خط‌زمینه تخمین زده شده، (c) نویز الکتریکی تخمین زده شده، (d) سیگنال پردازش شده، (e) طیف‌جرمی نرمالیزه شده

[6] Discrete wavelet transform
[7] Approximations
[8] Details
[9] Multiresolution
[10] Higher-order statistics

توسعه یافته‌ای که توانایی استخراج الگوهای با قدرت تفکیک‌پذیری بالا و حداقل تعداد ویژگی لازم از این فضای اطلاعاتی ابعاد بالا را داشته باشد، یک رهیافت داده‌کاوی نامیده می‌شود [۲۵و۲۶]. در این تحقیق، یک روش انتخاب زیرمجموعه ویژگی مبتنی بر رهیافت فیلتری جهت انتخاب نشانگرهای حیاتی از داده‌های طیف‌جرمی با ابعاد بالا توسعه داده شده است. روش پیشنهادی قادر است تا بهترین گروه از ویژگی‌ها مشتمل بر نشانگرهای تفکیک کننده گروه‌های سالم از سرطانی را انتخاب نماید.

۵- آزمون معناداربودن آماری

در بیشتر کارهای تشخیصی بطور عام تمام نقاط داده ثبت شده بیانگر اختلاف معنی‌دار بین گروه‌های تحت مطالعه نمی‌باشند. بنابراین، اتخاذ یک روش مناسب در انتخاب اولیه نقاط معنی‌دار از نقطه نظر فیزیولوژیک حایز اهمیت می‌باشد. در تکنیک‌های داده کاوی، از این مرحله بعنوان استخراج ویژگی یاد می‌شود. برای این منظور، آزمون‌های آماری جهت تعیین سطوح معنی‌دار بودن اختلاف بین نقاط ویژگی، گزینه‌ای مناسب به شمار می‌آیند. در این مقاله، آزمون آمارگان T برای انتخاب نقاط معنی‌دار مورد استفاده قرارگرفت.

فرض می‌شود که ماتریس داده ورودی $D_{N\times M}$ با N نمونه و تعداد M ویژگی وجود دارد که هر عضو از این مجموعه بصورت $X=\{x_i, i=1,\ldots,M\}$ نشان داده می‌شود. هدف استخراج ویژگی پیداکردن یک زیرفضا با m ویژگی، $\Re^m$، از فضای مشاهدات با M ویژگی، $\Re^M$، خواهد بود که نقاط این زیرفضا از نظر آماری دارای اختلاف معنی‌دار باشند. فرض می‌شود که دو متغیر x و y با توزیع نرمال وجود دارند آنگاه آمارگان T با استفاده از رابطه (۴) محاسبه می‌شود:

$$TS=\frac{\mu_x-\mu_y}{\sqrt{\frac{{\sigma_x}^2}{N_x}+\frac{{\sigma_y}^2}{N_y}}} \qquad (۴)$$

در رابطه بالا N,σ,μ بترتیب بیانگر میانگین، واریانس و تعداد نمونه‌های جوامع آماری x و y می‌باشند. با استفاده از مقدار آمارگان محاسبه شده و روش‌آستانه‌دهی، ویژگی‌هایی که عدد آمارگان آنها از مقدار آستانه بیشتر باشد، نگه داشته شده و مابقی حذف می‌شوند. مقدار آستانه با توجه به در نظر گرفتن توزیع‌گوسی و یک بازه اطمینان[11] قابل‌محاسبه می‌باشد [۲۷].

۶- اندازه تفکیک‌پذیری کلاس

روش فیلتری آمارگان T موجب کاهش بعد اولیه فضای داده ورودی می‌گردد، ولی این شیوه در انتخاب بهترین زوج ویژگی‌ها هیچ نقشی به عهده ندارد. نشان داده شده است که ترکیبی از بهترین ویژگی‌های منفرد همواره نمی‌تواند منجر به ایجاد یک زیر مجموعه ویژگی بهتر گردد [۲۸ و۲۹]. برای انتخاب بهترین زیرمجموعه ویژگی، می‌توانیم از اندازه‌های جداپذیری کلاس الگو استفاده کنیم.

اکنون، ماتریس داده‌ورودی، $D_{N\times m}$، می‌باشد که تعداد نقاط ویژگی برابر m است. می‌خواهیم تعداد d ویژگی تشکیل دهنده زیر فضای $\Re^d$ را انتخاب کنیم بطوری‌که این زیر فضا حداکثر فاصله را بین گروههای کلاس الگو از نظر تفکیک پذیری ایجاد نماید. در این مقاله، از فاصله باچاتاریا[12] بعنوان یک معیار جهت اندازه‌گیری جداپذیری کلاس‌های الگو با استفاده از ویژگی‌های منتخب، استفاده شده است. فاصله باچاتاریا برای متغیرهای با توزیع نرمال در دو کلاس الگو از رابطه (۵) قابل محاسبه می‌باشد [۳۰]:

$$b_{ij}=\frac{1}{8}(\mu_i-\mu_j)^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\mu_i-\mu_j)+\frac{1}{2}\log\left(\frac{\left|\frac{\Sigma_i+\Sigma_j}{2}\right|}{\left(|\Sigma_i||\Sigma_j|\right)^{\frac{1}{2}}}\right) \qquad (۵)$$

در رابطه بالا، Σ و μ بترتیب معرف واریانس درون کلاسی و میانگین کلاس می‌باشند. همچنین، عملگرهای $(\cdot)^T$، $(\cdot)^{-1}$ و $|\cdot|$ بترتیب به ترانهاده، معکوس ماتریس و دترمینان اشاره دارد. اگر زیرفضای با تعداد ویژگی d بصورت S_d و تعداد کلاس‌های الگو با c نمایش داده شود آنگاه معیار انتخاب ویژگی می‌تواند بصورت زیر بیان شود:

$$\max\left(J_b(S_d|c)\right),\qquad J_b(S_d|c)=\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}b_{ij} \qquad (۶)$$

جهت انتخاب بهترین زیر مجموعه ویژگی، S_d، دارای d بعد از فضای فیلتر شده با ابعاد m تعداد حالت‌های ممکن برابر ترکیب $\sum_{i=1}^{d}\binom{m}{i}$ خواهد بود. در این حالت جستجوی کل فضا کاری سخت می‌باشد. در اینجا می‌توان از روش‌های زیر بهینه مبتنی بر جستجوهای متوالی استفاده نمود که با انتخاب یک مقدار مناسب برای d، اندازه مناسب برای زیرفضای جستجو تعیین می‌شود [۳۱].

۷- توصیف‌گر شکل

جهت تحلیل داده‌های طیف‌جرمی، برخی محققان ابتدا پیک‌های موجود در طیف را استخراج نموده و سپس با رتبه بندی آنها اقدام به

[11] Confidence Interval

[12] Bhattacharyya

انتخاب نشانگر حیاتی می‌نمایند و بعضی دیگر نیز از تمام نقاط طیف بعنوان ویژگی‌های متمایز کننده گروه‌های سالم و سرطانی استفاده می‌کنند [۳۴-۳۲]. با توجه به این مهم، که اکثر دستگاه‌های طیف‌سنج جرمی با استفاده از نرم‌افزار تهیه شده توسط شرکت سازنده، پیک‌های وابسته به پروتئین‌ها را استخراج می‌نمایند لذا محور نسبت جرم به بار در داده‌های تهیه شده بصورت یکنواخت نمونه‌برداری نمی‌شود. بنابراین، جهت حفظ ارزش اطلاعاتی تمامی نقاط و همچنین در نظر گرفتن نقاط دارای شکل پیک، استفاده از روش‌های امتیازدهی نقاط ضروری می‌باشد.

در این مقاله، یک روش امتیازدهی برای نقاط طیف‌جرمی پیشنهاد داده‌شده‌است که به هر نقطه از نسبت جرم به بار متناسب با شکل تغییرات شدت آن نقطه یک وزن اختصاص می‌دهد. این شیوه امتیازدهی یا اختصاص وزن را بعنوان عملگر *توصیف‌گر شکل*[13] می‌نامیم. فرض می‌شود که بردار میانگین ماتریس داده ورودی با $\overline{\mathbf{d}}$ نشان داده شود:

$$\overline{\mathbf{d}} = \frac{1}{N}\sum_{i=1}^{N} d_{ij}, \qquad D_{N\times M} = \left[d_{ij}\right]_{N\times M} \tag{۷}$$

جهت امتیازدهی هر نقطه، یک پنجره بطول W با مرکزیت آن نقطه در نظر گرفته می‌شود. اگر نقطه مرکزی با $\overline{d}_j$ نشان داده شود آنگاه جهت محاسبه شکل تغییرات نقطه مرکزی نسبت به نقاط همسایه در پنجره از یک تابع علامت بصورت زیر استفاده می‌شود:

$$\mathrm{SDO}_j = \frac{1}{2\mathrm{W}} \sum_{i=j-\frac{\mathrm{w}}{2}}^{j+\frac{\mathrm{w}}{2}} sign(\overline{d}_j - \overline{d}_i - s) + \frac{1}{2} + \alpha \tag{۸}$$

این عملگر به هر نقطه که در پنجره نسبت به نقاط همسایه دارای شدت بیشتری باشد، امتیاز ۱ اختصاص می‌دهد و به نقاطی که در حداقل مقدار نسبت به نقاط مجاور قرار دارند، وزن صفر نسبت می‌دهد. پارامترهای s و α بترتیب بیانگر مقدار حساسیت و عامل تنظیم می‌باشند که حساسیت موجب تعریف یک فاصله بین شدت نقاط جهت انتخاب مناسب پیک‌ها می‌شود و عامل تنظیم باعث کاهش اثرات نویز می‌گردد. طول پنجره در این عملگر یک عدد صحیح زوج می‌باشد. استفاده از طیف میانگین اثر تغییر مکان پروتئین‌ها بر روی محور جرم به بار ناشی از خطای طیف‌سنج جرمی را کاهش می‌دهد.

۸- الگوریتم ترکیبی داده‌کاوی

در این مقاله، یک الگوریتم ترکیبی داده کاوی پیشنهاد شده است که زیرمجموعه‌ای از ویژگی‌ها را با تلفیق یک آزمون آماری، اندازه جداپذیری کلاس و امتیازدهی نقاط انتخاب می‌کند. روش جستجو مبتنی بر انتخاب افزایشی مستقیم می‌باشد که d ویژگی از فضای M بعدی برمی‌گزیند. مقدار مناسب برای d می‌تواند با محاسبه حداقل خطای طبقه بندی به روش ارزیابی متقابل بطور تجربی تعیین شود. روش انتخاب بهترین زیرمجموعه ویژگی مشتمل بر ۵ مرحله می‌باشد که در زیر توصیف شده‌است:

گام اول: مقدار آمارگان T برای ماتریس داده ورودی D از رابطه (۴) محاسبه می‌شود و سپس مقادیر جرم به بار با آمارگان بالاتر از یک حداقل آستانه انتخاب می‌شوند:

$$m_i = \arg\max\left\{\left|TS\right| \geq threshold\right\} \tag{۹}$$

گام دوم: اولین اندیس ویژگی، $d = ۱$، طوری انتخاب می‌شود که دارای بالاترین مقدار آمارگان در بین m_i های منتخب از گام اول می‌باشد.

$$m_1 = \arg\max\left\{\left|TS(m_i)\right|\right\} \tag{۱۰}$$

گام سوم: مقدار تابع توصیف‌گر شکل، SOD، برای m_i های منتخب از گام اول از ماتریس داده ورودی D با استفاده از رابطه (۸) حساب می‌شود.

گام چهارم: اندیس ویژگیهای بعدی، $d \geq ۲$، طوری انتخاب می‌شود که معیار زیر را حداکثر نماید:

$$m_i = \arg\max\left\{J_b\left(S_d\middle|c\right)\times SOD\right\}, \quad i \geq 2 \tag{۱۱}$$

گام پنجم: گام چهارم تا رسیدن به مقدار مطلوب d که بطور تجربی تعیین می‌شود، ادامه خواهد یافت.

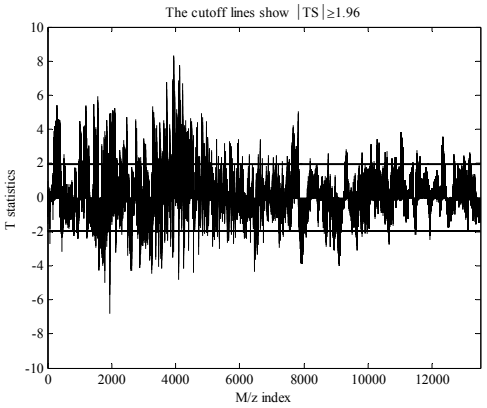
یافته ها و بحث

جهت نشان دادن کارآیی روش پیشنهادی در تعیین نشانگر حیاتی، داده‌های طیف‌سنجی‌جرمی بیماری سرطان پستان مورد تحلیل قرار گرفت. پیش‌پردازش جهت حذف خط زمینه و نویز الکتریکی روی تمام طیف‌ها مطابق با روش توصیف شده در بخش۲ انجام شد. به‌منظور جداسازی، از داده ورودی یک مجموعه یادگیری و آزمون بطور تصادفی برای گروه‌های سالم و سرطانی انتخاب شد. با توجه به تعداد کم نمونه‌های مجموعه داده در مقابل تعداد زیاد نقاط ویژگی، از روش ارزیابی متقابل۱۰چرخشی جهت جلوگیری از ایجاد هرگونه بایاس و خطا در هنگام انتخاب ویژگی و طبقه‌بندی استفاده شد.

در اولین گام، مقدار آمارگان T برای مجموعه یادگیری محاسبه گردید. از آنجایی‌که هنوز روش خاصی جهت انتخاب مقدار دقیق

[13] Shape descriptor operator

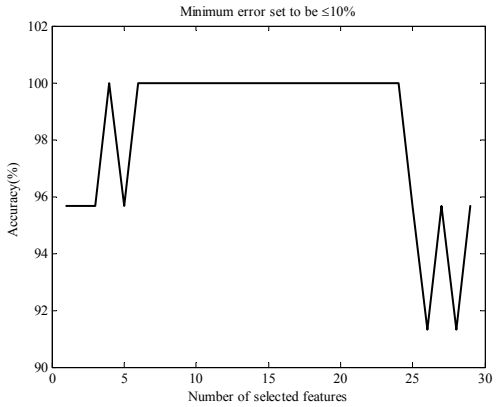
آستانه با توجه به تعداد متغیرها و نوع داده ارایه نشده‌است، بنابراین در بیشتر کارها انتخاب مقدار آستانه سلیقه‌ای بوده‌است [۳۵و۳۶]. در این مطالعه، از بازه اطمینان ۹۵٪ با حداقل مقدار آستانه ۱/۹۶ (P-value≤۰/۰۵) استفاده‌شد. شکل شماره ۴ مقدار آمارگان محاسبه شده برای مجموعه یادگیری را نشان می‌دهد که با اعمال آستانه مورد نظر تعداد ۲۳۸۰ نقطه جرم به بار از بین ۱۳۴۸۸ نقطه ویژگی باقی خواهد ماند.



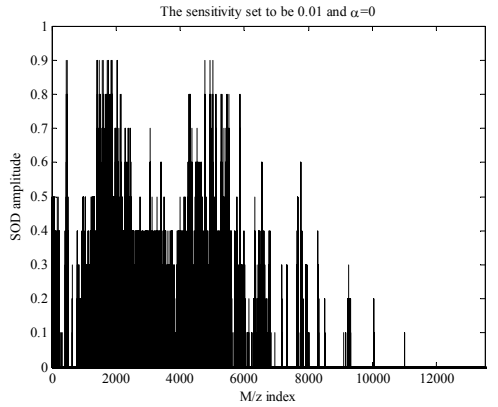
شکل۴- مقدار آمارگان T محاسبه شده در مجموعه یادگیری با خط متقاطع در مقدار آستانه ۱/۹۶

تابع توصیف‌گر شکل موجب می‌شود که نقاط طیف جرمی صرف نظر از مقدار شدت برحسب شکل تغییرات در بردار ویژگی نهایی تأثیرگذار باشند. یکی از معایب طیف‌سنجی جرمی وجود تغییرات غیر فیزیولوژیک در مقدار غلظت نقاط جرم به بار می‌باشد. همچنین، این تابع بدلیل استفاده از طیف میانگین موجب کاهش اثر جابجایی در محل نسبت جرم به بار ناشی از دستگاه طیف‌سنج می‌گردد. شکل شماره ۵ مقدار این تابع را برای کل مجموعه داده نشان می‌دهد. در محاسبه وزن از مقدار حساسیت ۰/۰۱ و بایاس ۰ استفاده شده است. این شکل نشان می‌دهد که نقاط با شدت‌های غیر یکسان از وزن برابر برحسب شکل تغییرات برخوردار می‌شوند.

جهت تعیین مقدار مناسب برای تعداد نهایی متغیرها در بردار ویژگی، d، یک مجموعه یادگیری بطور تصادفی در بین گروه‌های سالم و سرطانی انتخاب شد. با استفاده از روش ارزیابی متقابل K نگهدار[14] با تکرار ۱۰۰ و در نظر گرفتن حداقل خطای طبقه‌بندی ۱۰٪، جهت تعیین مقدار مناسب d اقدام شد. در اجراهای متوالی، بطور تجربی مقدار نهایی ۳۰ ویژگی انتخاب گردید. در شکل شماره ۵ مقدار میانگین درصد تشخیص در ۱۰۰ تکرار برای تعداد ویژگی‌های منتخب نشان داده شده است. در این شکل، هدف تأکید برروی کافی



شکل ۵- درصد تشخیص بر حسب تعداد ویژگی‌های منتخب برای بردار ویژگی نهایی با تعداد ۳۰ ویژگی با استفاده از طبقه‌بند LDA



شکل۶- مقدار تابع وزن توصیف‌گر شکل برای کل مجموعه داده با حساسیت ۰/۰۱ و پارامتر تنظیم صفر

بودن تعداد ۳۰ ویژگی جهت انتخاب زیر مجموعه نهایی بوده است و تغییرات در نرخ تشخیص به اهمیت ترتیب ویژگی‌ها اشاره‌ای ندارد.

با روش ارزیابی متقابل۱۰ چرخشی، تعداد ۱۹ پروتئین بعنوان ویژگی‌های متمایز کننده گروه سالم از نمونه‌های سرطانی بوسیله طبقه‌بندی تحلیل تمایز خطی(LDA)[15] انتخاب گردید برای بررسی قدرت تفکیک‌پذیری الگوی پروتیینی منتخب علاوه‌بر دقت تشخیص، معیارهای حساسیت و قطعیت نیز محاسبه شد، جدول شماره ۱ نتایج حاصل از طبقه‌بندی نمونه‌ها را با استفاده از طبقه‌بندی‌های تحلیل تمایز خطی و ماشین حامی بردار (SVM) فهرست نموده است.

جهت مقایسه عملکرد روش پیشنهادی با دیگر روش‌های پیاده‌شده بر روی همین مجموعه داده از نظر تأثیر انتخاب زیر مجموعه ویژگی بر روی معیارهای تشخیصی، الگوریتم داده کاوی با نتایج حاصل از کار ژانگ و همکارانش [۹] قیاس شده است. در مرجع [۹]

[14] K-hold-cross-validation

[15] Linear Discriminant Analysis

تعداد ۹۸ پروتئین از ۱۳۴۸۸پروتئین موجود در طیف جرمی با یک روش خوشه‌بندی بعنوان ویژگی‌های شاخص انتخاب شده بود. با روش توصیف شده در [۹] و تعداد ۲۶ اجرای متوالی بهترین نرخ تشخیص با روش ارزیابی متقابل ۱۰ چرخشی ۸۲/۴ ٪ با تعداد ۸ پروتئین بوده‌است. با الگوریتم داده‌کاوی پیشنهادی در این تحقیق، تعداد ۹۵ پروتئین از ۱۳۴۸۸ نقطه طیف‌جرمی انتخاب شده بود. همچنین، با روشی مشابه بطور حذفی متوالی، نتایج حاصل از طبقه‌بندی نمونه‌های داده براساس حذف نشانگرها از زیر فضای ویژگی مورد بررسی قرار گرفته است. جدول شماره ۲ عملکرد الگوریتم ترکیبی را برحسب تعداد ویژگی‌ها در هر مرحله از اجرا نشان می‌دهد. در این بررسی نیز از طبقه‌بندی ماشین حامی بردار استفاده شده است و معیارهای تشخیص با روش ارزیابی متقابل ۱۰ چرخشی محاسبه شده است. نتایج این جدول تأثیر انتخاب خوب پروتئین‌های شاخص را بر عملکرد تشخیص نشان می‌دهد که تشخیص کامل با حداقل تعداد ۱۹ پروتئین میسر شده است. همچنین با کاهش تعداد نشانگرهای حیاتی هنوز معیارهای تشخیصی از اعتبار خوبی برخوردار می‌باشند که نشان‌دهنده کارآیی روش پیشنهادی در تحلیل پروفایل‌های پروتئینی می‌باشد. در این جدول، تعداد ویژگی‌ها با علامت NF و مراحل حذفی متوالی با اجرا نشان داده شده است.

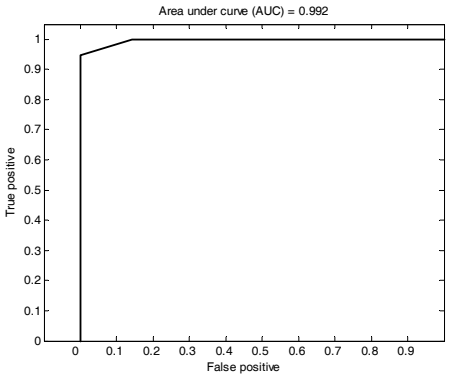
جدول ۱- عملکرد طبقه‌بندی نمونه‌ها با استفاده از طبقه‌بندهای LDA و SVM بر اساس ۱۹ پروتئین منتخب با روش داده کاوی ترکیبی پیشنهادی

نوع طبقه‌بندی		معیارهای تشخیصی (٪)		
		دقت	حساسیت	قطعیت
	LDA	۱۰۰	۱۰۰	۱۰۰
	SVM	۱۰۰	۱۰۰	۱۰۰

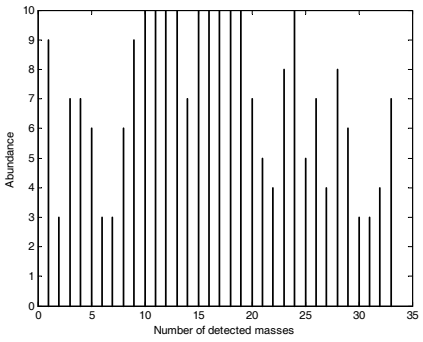
جدول ۲- بررسی عملکرد تأثیر روش پیشنهادی در انتخاب پروتئین‌های شاخص از طیف‌جرمی بر اساس تعداد نشانگرهای حیاتی و معیارهای تشخیصی استاندارد

نوع طبقه‌بندی ماشین حامی بردار (SVM)	اجرا	NF	معیارهای تشخیصی (٪)		
			دقت	حساسیت	قطعیت
	۱	۹۵	۸۷/۴	۸۳/۳۴	۹۰
	۲	۵۸	۹۸/۲۶	۹۶/۶۷	۹۹/۳
	۳	۴۷	۹۹/۵۷	۹۹/۳	۱۰۰
	۴	۳۹	۹۹/۵۷	۹۸/۹	۱۰۰
	۵	۳۸	۹۹/۵۷	۹۸/۹	۱۰۰
	۶	۳۶	۱۰۰	۱۰۰	۱۰۰
	۷	۱۹	۱۰۰	۱۰۰	۱۰۰
	۸	۹	۹۸/۷	۹۶/۶۷	۱۰۰

از آنجایی‌که نتایج حاصل از هر روش تشخیصی وابسته به پارامترهای مرتبط با آن می‌باشد، لذا جهت بررسی کارآیی روش پیشنهادی از منحنی مشخصه عملکرد[16] (ROC) و سطح زیر آن[17] (AUC) استفاده شده است. شکل شماره ۷ منحنی ROC را برای نشانگرهای حیاتی انتخاب شده با استفاده از روش پیشنهادی براساس الگوریتم یادگیری SVM نشان می‌دهد. مقدار بالای AUC بیانگر عملکرد خوب این روش می‌باشد که در جدول شماره ۲ نتایج تشخیصی آن با روش R-SVM مقایسه شده‌است.



شکل۷- منحنی ROC برای الگوریتم یادگیری ماشین حامی بردار در با استفاده از نشانگرهای حیاتی منتخب



شکل۸- نمودار هیستوگرام نشانگرهای حیاتی استخراج شده از پروفایل پروتئینی بعنوان معیاری از تکرارپذیری

یکی دیگر از معیارهای سنجش توانایی یک روش در استخراج الگوهای تفکیک‌پذیر از پروفایل پروتئینی جنبه تکرارپذیری آن می‌باشد. تکرارپذیری می‌تواند بصورت استخراج الگوهای یکسان در اجراهای متوالی روش بر روی مجموعه داده با انتخاب تصادفی نمونه‌های یادگیری و آزمون تعریف شود [۳۷]. بدین منظور با استفاده از روش ارزیابی متقابل ۱۰ چرخشی، تعداد ۳۰ نشانگر حیاتی در هر بار اجرای متوالی استخراج گردید. نمودار هیستوگرام جرم‌های آشکار شده بعنوان نرخ تکرارپذیری روش معرفی شده در شکل شماره ۸

[16] Receiver operating charactristics
[17] Area under curve

نشان داده شده است. در این نمودار، ۳۳ نشانگر حیاتی با تکرار بیش از ۲ بار بعنوان الگوهای پروتئینی معتبر ترسیم شده‌اند که حاکی از دارا بودن نرخ تکرارپذیری قابل قبول می‌باشد.

نتیجه‌گیری

با تولید اطلاعات کامل از نمونه‌های بیولوژیک می‌توان از آنها در تشخیص بیماری‌های با عوامل تشخیصی ضعیف نظیر سرطان استفاده نمود. طیف‌سنجی‌جرمی امیدی در راستای دستیابی به نشانگرهای مولکولی با قدرت تفکیک بالا برای تشخیص سرطان در مراحل اولیه می‌باشد. یکی از مشکلات حوزه پروتئین‌شناسی، تحلیل صحیح این داده‌ها می‌باشد که با تعداد زیاد متغیر در برابر نمونه‌های کم مواجه می‌باشیم. الگوریتم‌های داده کاوی در این موارد نقشی کلیدی در انتخاب نشانگرهای حیاتی ایفا می‌نمایند.

در این مقاله، یک الگوریتم داده کاوی ترکیبی پیشنهاد شد که توانست تعداد ۱۹ پروتئین شاخص را از بین ۱۳۴۸۸ نقطه ویژگی استخراج نماید. یکی از عوامل موفقیت در راستای تحلیل داده طیف جرمی ارایه یک روش پیش‌پردازش هوشمند در کنار انتخاب ویژگی بوده است. در تحلیل داده‌های طیف‌جرمی سرطان پستان با استفاده از ۱۹ نشانگر حیاتی، دقت تشخیص ۱۰۰٪، حساسیت ۱۰۰٪ و قطعیت ۱۰۰٪ حاصل گردید که در قیاس با روشی مشابه از عملکرد بسیار خوبی برخوردار بوده است. اگرچه این مقایسه براساس تعداد ویژگی‌های منتخب و تأثیر آن در معیارهای تشخیص بوده است ولی بیانگر اهمیت بالای نقش انتخاب زیر مجموعه ویژگی در تعیین نشانگرهای حیاتی با تفکیک‌پذیری کامل در بین نمونه‌های سالم و سرطانی می‌باشد.

منابع

1. Jemal A, Thomas T, Murray M Thun. Cancer statistics. CA Cancer J. Clin 2002; 52: 23-47.
2. Jemal RC, Tiwari T Murray. Cancer statistics. CA Cancer J. Clin 2004; 54: 8-29.
3. Srinivas PR, Srivastava S, Hanash S, Wright G.L. Proteomics in early detection of cancer. Clinical Chemistry 2001; 47(10): 1901-11.
4. Petricoin III E.F, Ornstein D.K, Paweletz C.P, Ardekani A.M, Hackett P.S, Hitt B.A, Velassco A, Trucco C, Wiegand L, Wood K, Simone C.B, Levine P.J, Linehan W.M, Emmert-Buck M.R, Steinberg S.M, Kohn E.C, Liotta L.A. Serum proteomic patterns for detection of prostate cancer. Journal of National Cancer Institute 2002; 94 (20): 1576-8.
5. Bertucci D, Birnbaum A, Goncalves. Proteomics of breast cancer. Molecular & Cellular Proteomics 2006; 5: 10.
6. Petricoin III E.F, Ardekani A.M, Hitt B.A, Levine P.J, Fusaro V.A, Steinberg S.M, Mills G.B, Simone C, Fishman D.A, Kohn E.C, Liotta L.A. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 2002; 359: 572-7.
7. Hu Y, Zhang S, Yu J, Liu J, Zheng S. SELDI-TOF-MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. The Breast 2005; 14: 250-5.
8. Shin B, Sheu M, Joseph M.K, Markey. Guilt-by-association feature selection: Identifying biomarkers from proteomic profiles. Journal of Biomedical Informatics 2007. (doi:10.1016/j.jbi.2007.04.003).
9. Zhang X, Lu X, Shi X, Xu X.Q, Leung H.C.E, Harris L.N, Iglehart J.D, Miron A, Liu J.S, Wong W.H. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics 2006 (doi:10.1186/1471-2105-7-197).
10. Wu T, Abbott D, Fishman W, McMurray G, Mor K, Stone D. Ward K, Williams H, Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003; 19 (13): 1636-43.
11. Morris J.S, Coombes K.R, Koomen J, Baggerly K.A, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 2005; 21(9): 1764-75.
12. Li J, Zhang Z, Rosenzwieg J, Wang Y.Y, Chang D.W. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clinical Chemistry 2002; 48(8): 1296-304.
13. Becker S, Cazares L.H, Watson P, Lynch H, Semmes O.J, Drake R.R, Laronga C. Surface-Enhanced Laser Desorp- tion- Ionization Time-of-Filght (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. Annals of Surgical Oncology 2004; 11(10): 907-14.
14. Finehout E.J, Lee K.H. An introduction to mass spectrometry applications in biological research. Biochemistry and Molecular Biology Education 2004; 32(2): 93-100.
15. Shi Q, Harris L.N, Lu X, Li X, Hwang J, Gentleman R, Iglehart J.D, Miron A. Declining plasma fibrinogen alpha fragment identifies HER2-positive breast cancer patients and reverts to normal levels after surgery. Journal of Proteome research 2006; 5: 2947-55.
16. Malyarenko D.I, Cooke W.E, Adam B.L, Malik G, Chen H, Tracy E.R, Trosset M.W, Sasinowski M, Semmes O.J, Manos D.M. Enhancement of sensitivity and resolution of SELDI-TOF mass spectrometric

records for serum peptides using time-series analysis techniques. Clinical Chemistry 2005; 51(1): 65-74.

17. Qu Y, Adam B.L, Thornquist M, Potter J.D, Thompson M.L, Yasui Y, Davis J, Schellhammer P.F, Cazares L, Clements M.A, Wright G.L, Feng Z. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. Biometrics 2003; 59: 143-51.
18. Mallat. A wavelet tour of signal processing. Academic press 1998.
19. Liu B.F, Sera Y, Matsubara N, Otsuka K, Terabe S. Signal denoising and baseline correction by discrete wavelet transform microchip capillary electrophoresis. Electrophoresis 2003; 24: 3260-5.
20. Hu Y, Jiang T, Shen T, Li W, Wang W, Hu J. background elimination method based on wavelet transform for Raman spectra. Chemometrics and Intelligent Laboratory Systems 2007; 85: 94-101.
21. Ruckstuhl F, Jacobson M.P, Field R.W, Dodd J.A. Baseline subtraction using robust local regression estimation. Quantitative Spectroscopy & Radiative Transfer 2001; 68: 179-93.
22. Donoho D.L. De-Noising by Soft-Thresholding. IEEE Trans. On Information Theory 1995; 41(3).
23. Ravier P, Amblard P.O. Wavelet packets and de-noising based on higher-order-statistics for transient detection. Signal processing 2001; 81: 1909-26.
24. Sima, Dougherty E.R. What should be expected from feature selection in small-sample settings. Bioinformatics 2006; 22(19): 2430-6.
25. Adam B.L, Vlahou A, Semmes O.J, Jr G.L, Wright. Proteomic approaches to biomarker discovery in prostate and bladder cancers. Proteomics 2001; 1: 1264-70.
26. Li L, Tang H, Wu Z, Gong J, Gruigl M, Zou J, Tockman M, Clark R.A. Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine 2004; 32: 71-83.
27. Hollander M, Wolfe D.A. Nonparametric Statistical Methods, 2nd Edition, Wiley 1999.
28. Jain A.K, Duin R.P.W, Mao J.Statistical pattern recognition: a review. IEEE Trans. On Pattern Analysis and Machine Intelligence 2000; 22(1): 4-37.
29. Peng H,Long F,Ding C. Feature selection based on mutual information: criteria of Max-dependency, Max-relevance, and Min-redundancy. IEEE Trans. On Pattern Analysis and Machine Intelligence 2005; 27(8): 1226-38.
30. Theodoridis S, Koutroumbas K. Pattern Recognition, 2nd Edition, Academic Press 2003.
31. Webb A.R. Statistical pattern recognition, John Wiley & Sons, May 2003.
32. Ressom H.W, Varghese R.S, Abdel-Hamid M, Eissa S.A.L, Saha D, Goldman L, Petricoin E.F, Conrads T.P, Veenstra T.D, Loffredo C.A, Goldman R. Analysis of mass spectral serum profiles for biomarker selection. Bioinformatics 2005; 21(21): 4039-5.
33. Yu J.S, Ongarello S, Fiedler R, Chen X.W, Toffolo G, Cobelli C, Trajanoski Z. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 2005; 21(10): 2200-9.
34. Li L, Tang H, Wu Z, Gong J, Gruigl M, Zou J, Tockman J, Clark R.A. Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine 2004; 32: 71-83.
35. Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach J. S. Detection of cancers-specific markers amid massive mass spectral data. PNAS 2003; 100(25): 14666-71.
36. Sorace J.M, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003; 4(24).
37. Baggerly K.A, Morris J.S, Coombes K.R. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different expriments. Bioinformatics 2004; 20(5): 777-85.