ارایه یک الگوریتم جدید ترکیبی انتخاب زیرمجموعه ویژگی جهت تحلیل داده‌های طیف‌جرمی لیزری سرطان تخمدان

حسین منتظری کردی[1]، محمدحسین میران‌بیگی[2*]، محمدحسن مرادی[3]

۱- دانشجوی دکتری تخصصی مهندسی پزشکی، دانشگاه تربیت مدرس تهران

۲- استادیار گروه مهندسی پزشکی، دانشگاه تربیت مدرس تهران

۳- دانشیار گروه مهندسی پزشکی، دانشگاه صنعتی امیرکبیر تهران

تاریخ دریافت نسخه اصلاح شده: ۸۶/۷/۱۹ **تاریخ پذیرش مقاله: ۸۶/۱۱/۲۴**

چکیده

مقدمه: یکی از مشکلات اساسی در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص بموقع آن می‌باشد. واکنشهای شیمیایی درون یک عضو زنده می‌تواند بصورت الگوهایی پروتیینی در مایعاتی نظیر خون، خلط، و ادرار انعکاس داده شود. طیف‌سنج‌جرمی لیزری یک ابزار مفید جهت استخراج اطلاعات پروتیینی از نمونه‌های وابسته به موجودات زنده می‌باشد. انتخاب بهینه زیرمجموعه ویژگی دربین نقاط داده طیف‌جرمی از چالشهای مهم این حوزه محسوب می‌شود.

مواد و روشها: در این تحقیق، داده‌های پروفایل پروتیینی خونابه بیماران مبتلا به سرطان تخمدان در دوگروه مجزا مورد تحلیل قرارگرفت. با استفاده از یک مدل ریاضی، اغتشاشات خطزمینه و نویزالکتریکی در مرحله پیش‌پردازش حذف گردید و سپس، تمام سیگنالهای طیف‌جرمی نرمالیزه شدند. روش پیشنهادی، ترکیبی از تست آماری و اندازه فاصله باچاتاریا می‌باشد که با استفاده از معیار پیش‌بینی‌خطای‌نهایی، حداقل ویژگیهای لازم برای انتخاب بهترین زیرمجموعه از بین ۱۵۱۵۴ نقطه داده موجود را با حفظ ارزش اطلاعاتی و قدرت تفکیک‌پذیری برمی‌گزیند که این زیرمجموعه جهت آشکارسازی نشانگرهای‌حیاتی بکارمی‌رود.

نتایج: با استفاده از روش ارزیابی متقابل K چرخشی، نمونه‌های موجود در گروههای مورد مطالعه به دو دسته یادگیری و آزمون تقسیم شدند. با اعمال حداقل آستانه نقاط دارای اختلاف معنی‌دار انتخاب شدند و سپس، بهترین زیر مجموعه از بین نقاط باقیمانده با شرط دارا بودن حداکثر محتوای اطلاعاتی انتخاب گردید. ابعاد داده ورودی از ۱۵۱۵۴ به تعداد ۸۰ نقطه ویژگی کاهش داده شد. در این زیرمجموعه، تعداد ۱۶ و ۶ نشانگرحیاتی بترتیب برای زیرمجموعه‌های داده I و II انتخاب شد که در قیاس با روشهای دیگر دارای دقت تشخیص ۱۰۰٪، قطعیت ۱۰۰٪، و حساسیت ۱۰۰٪ می‌باشد.

بحث و نتیجه‌گیری: تشخیص بیماری در علم پزشکی نمونه‌ای از تفکیک الگو در علوم مهندسی می‌باشد. در این مقاله، یک روش فیلتری انتخاب زیرمجموعه ویژگی معرفی گردید که با ترکیب روشی آماری و معیارفاصله مبتنی بر سنجش ارزش اطلاعاتی، ویژگیهای مناسب را در بین فضای ورودی برمی‌گزیند. نتایج بدست آمده بر این نکته تاکید دارد که بکارگیری روشهای ترکیبی در استخراج و انتخاب ویژگی از فضاهای با ابعاد بالا، علاوه بر حفظ محتوای اطلاعاتی فضای اولیه، بخوبی می‌تواند کلاسهای الگو را از هم تفکیک نماید. (مجله فیزیک پزشکی ایران، دوره ۴، شماره ۱۴و ۱۵، بهار و تابستان ۸۶ :۹۶-۸۳)

واژگان کلیدی: پروتیین شناسی، سرطان تخمدان، طیف‌جرمی‌لیزری، الگوریتم انتخاب زیر مجموعه ویژگی، نشانگر حیاتی

* **نویسنده مسؤول: محمدحسین میران بیگی**

آدرس: گروه مهندسی پزشکی، بخش برق، دانشکده فنی، دانشگاه تربیت مدرس تهران miranbmh@modares.ac.ir

تلفن: ۸۸۰۱۱۰۰۱- (۲۱) ۹۸+

۱- مقدمه

یکی از مشکلات اساسی و حل نشده در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص بموقع و زودرس آن می‌باشد. تلاش جهت رفع این مشکل منجر به ارایه برخی از روشهای تشخیصی گردیده‌است که عیب عمده آنها عدم دقت کافی با توجه به هزینه بالا می‌باشد. با توجه به اطلاعات حوزه علم ژنتیک، واکنشهای شیمیایی درون یک عضو زنده می‌تواند بصورت الگوهایی پروتیینی در مایعاتی نظیر خون، خلط، و ادرار انعکاس داده شود [۱]. واژه پروتیین‌شناسی[1] به علومی اطلاق می‌شود که امکان مقایسه کمی و کیفی و همچنین تمایز قایل شدن بین پروتیینها تحت شرایط مختلف پاتولوژیک برای فهم فرایندهای بیولوژیک، نظیر شناسایی عوامل وابسته به بیماری، را فراهم نماید.

حوزه فعالیت پروتیین‌شناسی بطورعام به هرنوع تکنولوژی یا تکنیک پردازش‌اطلاعاتی مربوط می‌شود که بتواند داده‌های‌پروتیینی با مقیاس بالا تولید نموده و یا آن را مورد تحلیل قراردهد [۲]. یکی از ابزارهای مورد استفاده جهت استخراج اطلاعات پروتیینی از نمونه‌های وابسته به موجودات زنده، طیف‌سنج‌جرمی لیزری[2] می‌باشد. تحلیل محتوای اطلاعاتی طیف‌جرمی یک روش سریع و ارزان در تشخیص بیماری بدون ایجاد هرگونه عوارض جانبی می‌باشد که می‌تواند امکان بالقوه غربالگری سرطان را فراهم سازد. سرطان تخمدان یکی از عوامل شایع مرگ و میر زنان محسوب می‌شود که تشخیص بموقع این بیماری به نشانگرهای حیاتی[3] جدید با قدرت تشخیص بالا نیاز دارد [۳]. استفاده از تکنیک طیف‌سنجی‌جرمی لیزری بهمراه تحلیل صحیح سطوح تفسیر پروتیینها می‌تواند گامی موثر در راه حل این مشکل و کشف نشانگرهای حیاتی به شمار آید.

با استفاده روبه رشد طیف‌سنجی‌جرمی‌لیزری برای تهیه پروفایلهای‌پروتیینی، برخی از چالشهای مهم در ارتباط با تحلیل این داده‌ها افزایش یافته‌است. از جمله آنها می‌توان به مواردی نظیر: حذف نویز خط زمینه و الکتریکی، نرمالیزه کردن طیف‌جرم، تعیین و تنظیم پیکها، و انتخاب متغیرهای مهم در بین مجموعه داده اشاره نمود[٤].

موریس و همکارانش دریکی از تحقیقات خود[۵] متذکر شده‌اند که مرحله انتخاب ویژگی درتحلیل سیگنال طیف‌جرمی از اهمیت ویژه‌ای برخوردار است. دینگ و همکارانش درتحقیقی دیگر[٦] نشان داده‌اند که در استخراج نشانگرهای‌حیاتی از داده‌های با ابعاد بالا، اتخاذ یک استراتژی انتخاب بهترین زیرمجموعه ویژگی نقش اساسی دارد. ژونگ و همکارانش درتحقیقشان[۷] و همچنین چِن و همکارانش درتحقیق خود[۸] از تستهای آماری جهت انتخاب اولیه نقاط ویژگی استفاده نموده‌اند. سوراس و همکارانش درتحقیقی دیگر[۹] با استفاده از تست آماری ویلکاکسون و انتخاب آستانه 10^{-6} به انتخاب نقاط ویژگی معنی‌دار پرداختند.

آدام و همکارانش در مطالعه خود [۱۰] با استفاده از روش درخت تصمیم، داده طیف‌جرم لیزری سرطان پروستات را مورد تحلیل قرار دادند. کو و همکارانش در ادامه [۱۱] روش درخت تصمیم مبتنی بر تقویت[4] را روی مجموعه داده[۱۰] پیاده سازی کردند. پتریکون و همکارانش درتحقیقات خود[۲و۱۲] با بکارگیری الگوریتم ژنتیک و شبکه عصبی خودسازمانده در جهت کشف نشانگرحیاتی از سیگنال طیف‌جرمی‌لیزری سرطان پروستات و تخمدان تلاش نمودند. یونگ و همکارانش[۱۳] با توسعه یک روش ترکیبی انتخاب پیک مبتنی بر ماشین‌حامی‌بردار[5]، داده طیف جرمی لیزری را تحلیل کردند. عیب عمده این مطالعات، علاوه بر عدم اتخاذ یک روش مناسب جهت حذف اغتشاش خط‌زمینه و نویزالکتریکی، کنارهم قراردادن

1- Proteomics
2- Laser Mass Spectrometry
3- Biomarkers

4- Boosting
5- Support Vector Machine (SVM)

مجموعه‌ای از ویژگیها با رتبه منفرد بالا بوده‌است. یک روش انتخاب ویژگی مناسب ضمن حفظ معنی‌دار بودن نقاط ویژگی بایستی حافظ محتوای اطلاعاتی بردار ویژگی نیز باشد که در روشهای قبلی فقط حصول نتایج طبقه‌بندی بالا مدنظر بوده‌است.

در این مقاله، روشی ترکیبی ارایه شده است که با استفاده از یک تست آماری در مرحله اول بعد داده طیف جرمی را کاهش می‌دهد و سپس با حداکثر نمودن معیار مبتنی بر فاصله بین نقاط‌ویژگی، بهترین زیرمجموعه از آنها را استخراج می‌نماید. این روش ضمن حفظ معنی‌دار بودن نقاط ویژگی، محتوای‌اطلاعاتی بردار ویژگی را نیز محفوظ نگه می‌دارد که مبین قدرت تمایز بین گروههای سالم و سرطانی می‌باشد. با روش ارایه شده، تعداد۸۰ ویژگی‌از بین ۱۵۱۵٤ نقطه برگزیده شد که از این زیرمجموعه تعداد ۱٦ و ٦ پروتیین به عنوان نشانگرحیاتی انتخاب‌گردید. همچنین، عملکرد ویژگیهای‌منتخب با استفاده از چند روش‌طبقه‌بندی مناسب برای داده‌های‌طیف‌جرمی مورد بررسی قرار گرفت. با استفاده از روش ارزیابی‌متقابل K چرخشی، مقادیر دقت تشخیص۱۰۰٪، قطعیت۱۰۰٪، و حساسیت۱۰۰٪ در بین مجموعه آزمایش حاصل گردید.

۲- طیف‌سنجی جرمی لیزری

تاریخ پزشکی همواره شاهد تحولات شگرفی بوده‌است که اختراعات مهم در عرصه تکنولوژی پدیدآورده‌است. تکنیک طیف‌سنجی جرمی در علم شیمی قدمتی طولانی دارد و بیشتر با عنوان طیف‌سنج جرم کروماتوگراف گاز شناخته می‌شود. از این تکنیک جهت تحلیل ترکیبات گازی هوای بازدم به منظور کشف عوامل بیماری استفاده شده‌است [١٤]. ولی استفاده از این تکنولوژی برای تحلیل نمونه‌های بیولوژیک نظیر خون بدلیل محدودیتهای تکنیکی در آماده‌سازی نمونه جهت اعمال به طیف‌سنج جرم تا سالهای اخیر غیر ممکن بوده‌است.

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 اختراع طیف‌سنجهای جرمی لیزری مشکل یونیزاسیون نمونه مرتفع شده‌است.

در طیف‌سنج جرمی لیزری، نمونه بیولوژیک برروی یک صفحه یا ماتریس جذب کننده انرژی قرار می‌گیرد. پرتوهای لیزر به این 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 تاثیر یک ولتاژ بالا در یک لوله خلاء به سمت آشکارساز حرکت می‌کنند که 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 نشاندهنده غلظت یا فراوانی یک مولکول است [۱۵]. شکل۱ نمایی از یک طیف‌سنج جرمی لیزری را جهت تهیه طیف‌جرم نشان می‌دهد.

۳- مواد و پیش‌پردازش

در این مقاله از داده‌های پروفایل پروتیینی خونابه[1] که توسط دستگاه طیف سنجی جرمی با تکنیک جذب- یونیزاسیون لیزری سطحی ارتقا یافته زمان پروازی[2] تهیه شده است، به عنوان داده ورودی برای الگوریتمهای پردازشی استفاده شده‌است. سپس، با استفاده از این داده‌ها غربالگری بیماران مبتلا به سرطان تخمدان و افراد تحت کنترل سالم بمنظور دستیابی به نشانگرهای حیاتی متمایز کننده دو گروه انجام شده‌است.

1- Serum
2- Surface Enhanced Laser Desorption-Ionization- Time of Flight

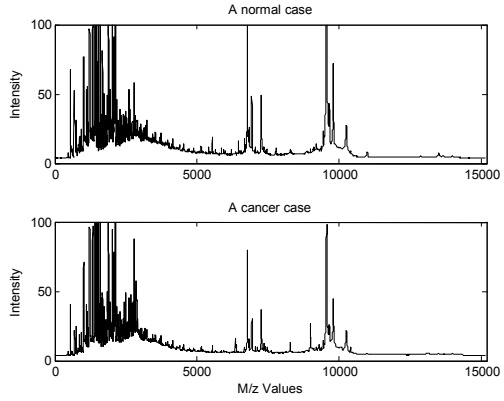
۳-۱- داده‌های طیف سنجی‌جرمی

در این تحقیق، جهت تعیین الگوهای پروتیینی از خونابه که باعث تمایز و تشخیص نمونه‌های مبتلا به سرطان‌تخمدان از افراد تحت‌کنترل‌سالم می‌شود، مجموعه‌داده‌های طیف‌جرمی حاصل از SELDI-TOF مورد استفاده قرار گرفت. این مجموعه داده‌ها از انستیتو ملی سرطان آمریکا اخذ گردید که نحوه توزیع نمونه‌های هر مجموعه به تفکیک گروههای سالم و سرطانی در جدول۱ آورده شده‌است.

جدول۱- نحوه توزیع نمونه‌ها در مجموعه داده‌های تحت مطالعه

مجموعه داده	تعداد نمونه های سالم	تعداد نمونه های سرطانی	تعداد نمونه های با تومور خوشخیم	تعداد ویژگیها در هر طیف جرم
مجموعه I	۱۰۰	۱۰۰	۱٦	۱۵۱۵٤
مجموعه II	۹۱	۱٦۲	۰	۱۵۱۵٤

در این مجموعه داده‌ها، هر منحنی طیف جرم دارای ۱۵۱۵٤ نقطه روی محور نسبت جرم به بار (M/z) می باشد که متناظر با آن یک نقطه روی محور شدت سیگنال وجود دارد. شکل۲ دو نمونه از منحنی طیف جرمی را برای داده‌های تحت مطالعه مجموعهII در جدول۱ نشان می‌دهد که بترتیب مربوط به یک فرد سالم، و بیمار سرطان تخمدان با اثبات بیماری به روش بیوپسی می باشد.



شکل ۲- نمونه‌هایی از طیف‌جرمی متعلق به دو فرد سالم و بیمار از مجموعه داده سرطان تخمدان

۳-۲- مدل ریاضی طیف‌جرم

از دیدگاه مدلسازی، سیگنال طیف‌جرمی می‌تواند با یک مدل ترکیبی نشان داده شود [۵ و۱۶] که اجزای این مدل می‌بایست در برگیرنده عوامل شیمیایی و الکتریکی یک طیف‌سنج جرم باشد. فرض می‌شود که n طیف نمونه‌برداری‌شده در بازه زمانی T از زمان پرواز در فواصل زمانی $t_j, j=1,\ldots,T$ مشاهده‌شده‌است. مدل زیر را می‌توان برای سیگنال طیف جرمی در نظرگرفت [۵، ۱۷]:

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

که در این مدل، $y_i(t_j)$ شدت سیگنال یا فراوانی یک مولکول برای هر جرم مجزا را در زمان t_j نشان می‌دهد. $B_i(t_j)$ به خطای سیستم ناشی از ماده شیمیایی بکاررفته در دستگاه جهت ثابت نگه‌داشتن نمونه بیولوژیک اشاره‌دارد که به عنوان نویز خط زمینه شناخته می‌شود. $S_i(t_j)$ سیگنال اصلی ناشی از مولکولهای پروتیینی موجود در نمونه بیولوژیک می‌باشد که در عامل مقیاس N_i ضرب شده‌است. ε_{ij} معرف اغتشاش الکتریکی دستگاه طیف‌سنج جرمی است که توزیع گوسی برای آن فرض می‌شود.

۳-۳- حذف اغتشاشات خط‌زمینه و الکتریکی

با توجه به مدل ارایه‌شده، جهت کاهش اثرات مخرب ناشی از خط‌زمینه و نویزالکتریکی در مرحله انتخاب ویژگی بایستی نسبت به حذف آنها اقدام نمود. با توجه به کارآیی تبدیل موجک گسسته در پردازش سیگنالها با تغییرات زیاد، از این ابزار جهت حذف اغتشاش از طیف‌جرم استفاده شد[۱۸]. سیگنال طیف جرمی مدل شده در (۱) به عنوان یک تابع گسسته در نظر گرفته شد که با گرفتن تبدیل موجک گسسته از آن اغتشاش خط زمینه که دارای تغییرات آهسته می‌باشد، در ضرایب تقریبات[1] ظاهر می‌شود[۱۹] و نویز الکتریکی نیز در ضرایب جزییات[2] انعکاس می‌یابد [۲۰]. با استفاده از روش توصیف شده در [۲۱]، ضرایب تقریبات جهت تنظیم خط زمینه مورد پردازش قرار گرفتند. از دابیشیز مرتبه ٤ به عنوان تابع ویولت مادر مناسب استفاده گردید. نویز الکتریکی که بصورت عاملی با تغییرات تیز فرض می‌شود، با استفاده از روش بیان شده در [۲۲] به روش آستانه‌دهی نرم حذف گردید. مقدار آستانه مناسب برای حذف نویز الکتریکی بر مبنای آمارگان مرتبه بالا[3] محاسبه شد که این مقدار با روش ذکر شده در[۲۳] بدست آمد.

۳-۴- نرمالیزه کردن طیف

پس از حذف اثرات ناشی از اغتشاشهای خط زمینه و منابع الکتریکی، سیگنال بمنظور حذف اثر ضریب مقیاس با استفاده از رابطه (۲) نرمالیزه گردید [۲٤]:

$$NS = \frac{S - \min(S)}{\max(S) - \min(S)} \qquad (۲)$$

در این رابطه، NS به مقدار شدت طیف نرمالیزه شده و S به مقدار آن قبل از نرمالیزاسیون اشاره دارد. در این رابطه عبارات min(S) و max(S) بترتیب مقادیر حداقل و حداکثر شدت درهرنقطه از سیگنال را نشان می دهند. درشکل۳ یک نمونه از طیف جرمی قبل از عمل پیش‌پردازش مشاهده می‌شود. در همین شکل، اغتشاش خط زمینه تخمین زده شده به همراه نویز الکتریکی نیز دیده می‌شود. سیگنال طیف جرمی بدون اغتشاش و همچنین سیگنال نرمالیزه شده نیز در ادامه در همین شکل نشان‌داده‌شده‌است.

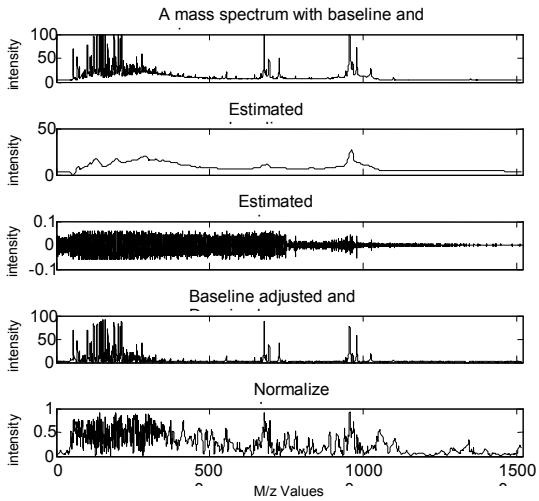
1- Approximations

2- Details

3- Higher Order Statistics

٤- روش انتخاب ویژگی

روشهای انتخاب ویژگی بطور کلی به دو دسته فیلتری و مبتنی بر یادگیری[1] تقسیم می شوند. در روشهای فیلتری، انتخاب کننده ویژگی مستقل از الگوریتم یادگیری ویژه مورد استفاده در طبقه بندی می باشد و از آن به عنوان یک فیلتر جهت رد یا قبول نقاط ویژگی نامناسب بهره گرفته شده است. بعبارت دیگر، در روشهای مبتنی بر یادگیری، انتخاب کننده ویژگی به عنوان یک معیار وابسته به الگوریتم یادگیری معین در انتخاب ویژگیهای مناسب شرکت می کند[۲۵]. در این مقاله، یک روش ترکیبی جهت انتخاب ویژگیهای مناسب ارایه شده‌است که با استفاده از یک معیار فاصله در فضای فیلتر شده توسط یک روش آماری به جستجوی زیرمجموعه‌ای از ویژگیها با شرط حداکثر تمایز بین گروههای سالم و سرطانی می‌پردازد.



شکل ۳- یک نمونه از طیف‌جرمی قبل و بعد از حذف اغتشاشات خط‌زمینه و نویزالکتریکی به همراه طیف نرمالیزه شده

۴-۱- روش آماری

در این مقاله، از روش آمارگان T (TS) به عنوان یک شیوه فیلتری جهت حذف ویژگیهای نامناسب استفاده گردید. آمارگان T برای زمانی که دو گروه تحت مطالعه دارای اندازه‌های نا برابر از لحاظ تعداد نمونه‌های عضو هر کلاس می‌باشد، با استفاده از رابطه زیر بیان می‌گردد [۲۶]:

$$t_i = \frac{\mu_{i1} - \mu_{i2}}{\sigma_{i1,2}}, \quad i = 1,\ldots,M \qquad (۳)$$

در رابطه بالا، μ_{ij} بیانگر مقدار میانگین برای ویژگی iام از کلاس j می‌باشد و پارامتر $\sigma_{i1,2}$ نیز به مقدار واریانس ویژگی iام بین دو کلاس اشاره دارد که از رابطه زیر قابل محاسبه است[۲۵]:

$$\sigma_{i1,2} = \sqrt{\frac{(N_1 - 1){\sigma_{i1}}^2 + (N_2 - 1){\sigma_{i2}}^2}{N_1 + N_2 - 2}} \times \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \qquad (٤)$$

در رابطه بالا، N_i تعداد نمونه‌های هر کلاس بوده و σ_{ij} نیز واریانس متغیر iام از کلاس j می‌باشد. با استفاده از مقدار آمارگان محاسبه شده و روش آستانه‌دهی، ویژگیهایی که عدد آمارگان آنها از مقدار آستانه بیشتر باشد، نگه‌داشته‌شده و مابقی حذف می‌شوند. مقدار آستانه با توجه به درنظر گرفتن توزیع گوسی و یک بازه اطمینان[2] قابل محاسبه می‌باشد[۲۶].

۴-۲- روش حداکثرتمایز

با توجه به اعمال آمارگان T (TS) به هر عضو از مجموعه داده، فرض می‌شود که داده ورودی D با تعداد N نمونه که هر یک از آنها دارای M ویژگی می‌باشد، موجود است. هر عضو از D با $X = \{x_i, i = 1,\ldots,M\}$ نشان‌داده‌می‌شود، و هدف انتخاب زیر مجموعه‌ای از آن می‌باشد بطوریکه منجر به تمایز بهینه بین کلاسهای

1- Wrapper

2- Confidence Interval

موجود گردد. پس هدف یافتن زیر مجموعه‌ای m عضوی از فضای $\Re^M$ است که باعث تفکیک بهینه کلاسهای موجود گردد.

روش فیلتری آمارگان T موجب کاهش بعد اولیه فضای داده ورودی می‌گردد، ولی این شیوه در انتخاب بهترین زوج ویژگیها هیچ نقشی بعهده ندارد. بعبارت دیگر، آمارگان T بهترین ویژگیهای منفرد را از فضای موجود برمی‌گزیند. از آنجاییکه نشان‌داده‌شده‌است، ترکیبی از بهترین ویژگیهای منفرد همواره نمی‌تواند منجر به ایجاد یک زیر مجموعه ویژگی بهتر گردد[٦]. لذا این نکته موجب می‌شود که در فضای زیر مجموعه ویژگی با افزونگی اطلاعات مواجه شویم. از اینرو، ارایه یک روش جستجو مبتنی بر اندازه‌گیری تاثیر انتخاب متغیر جدید در محتوای اطلاعاتی فضای کاهش یافته می‌تواند موجب کاهش افزونگی گردد.

هدف‌از ارایه‌روش حداکثرتمایز[1] (MD) پیداکردن یک زیرمجموعه با تعداد m عضو از فضای $\Re^M$ می‌باشد که در کنارهم منجر به ایجاد حداکثر فاصله بین کلاسهای موجود گردد. برای این مقصود، یک معیار انتخاب ویژگی با تشکیل یک تابع تمایز غیرخطی یا تربیعی[2] معرفی نموده‌ایم که فاصله باچاتاریا بین نقاط ویژگی را اندازه‌گیری می‌کند. اگر ویژگیها دارای توزیع گوسی با مقادیر Σ_i و μ_i فرض شوند که پارامترها بترتیب معرف کوواریانس درون کلاسی و میانگین کلاس بوده و زیرنویس $i=1,2$ به تعداد کلاسها اشاره‌دارد. فاصله باچاتاریا[3] به عنوان یک اندازه از توان تمایز بین دوکلاس بصورت زیر تعریف می‌شود [٢٧]:

$$J(S;c)=\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\log\left(\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}\right) \quad (٥)$$

در رابطه بالا، S به زیر مجموعه ویژگی اشاره دارد و c معرف کلاسهای موجود می‌باشد. هدف انتخاب ابعاد، m، از زیرمجموعه، S، است بطوریکه شرط زیر حداکثر گردد:

$$\max\left(J(S;c)\right), \qquad S=\{x_i, i=1,\ldots,m\} \quad (٦)$$

با روش توصیف شده در بالا بدیهی است که با افزودن تعداد نقاط ویژگی به زیرمجموعه، مقدار فاصله باچاتاریا افزایش می‌یابد و لذا همگرایی روش جهت رسیدن به نقطه بهینه m کند می‌شود. در اینگونه روشهای انتخاب ویژگی، از شیوه‌های مرحله‌ای[4] یا زیربهینه[5] استفاده می‌شود که با بررسی یک معیار توقف در هر مرحله نسبت به همگرا شدن روش تصمیم‌گیری می‌شود. در این مقاله، از معیارخطای پیش‌بینی نهایی آکِیک[6] مبتنی بر حداکثر فاصله جهت بررسی شرط همگرایی روش حداکثر تمایز استفاده شده است [٢٨]. فرض شود که متغیر k به یک ویژگی از زیرفضای با بعد کاهش یافته اشاره نماید، آنگاه تابع خطا بصورت زیر تعریف می‌شود:

$$FPE(k)=\frac{1}{2k}\sum_{i=1}^{k}\left(\tilde{J}(S_i;c)-FPE(i-1)\right)^2 \quad (٧)$$

در رابطه بالا، $\tilde{J}(S_i;c)$ یک تابع تبدیل یافته از تابع حداکثر تمایز در رابطه (٥) می‌باشد که جهت همگرایی سریعتر روش، بصورت نمایی زیر در نظر گرفته می‌شود:

1- Maximum Discrimination
2- Quadratic Discriminate Function
3- Bhattacharyya
4- Stepwise
5- Suboptimal
6- Akaik's final prediction error

$$\tilde{J}(S_i;c)=1-e^{(-J(S_i;c)/i)} \qquad (۸)$$

در روابط بالا، منظور از S_i زیرمجموعه ویژگیها با تعداد i متغیر انتخاب شده می‌باشد. تابع ارایه شده در (۷)، یک تابع تجمعی می‌باشد که با افزایش تعداد ویژگیها در زیر فضای جستجو مقدار خطای پیش‌بینی به سمت ۱ میل می‌کند و می‌توان با قراردادن یک شرط توقف روی مشتق این تابع مقدار مناسب m را انتخاب نمود. با توجه به روابط گفته شده، الگوریتم جستجو در این مقاله بصورت زیر می‌باشد:

االگوریتم ۱:

مرحله ۱: با استفاده از آمارگان T (TS) و انتخاب یک مقدار آستانه مناسب تعداد ویژگیها را از M به d کاهش می‌دهیم.

مرحله ۲: در زیرفضای کاهش یافته d بعدی، اولین ویژگی مناسب را با بررسی شرط حداکثر تمایز (MD) برای کل متغیرهای زیرفضا انتخاب می‌نماییم.

مرحله ۳: مقدار خطای پیش‌بینی نهایی را مقدار دهی اولیه می‌کنیم. از آنجاییکه در مرحله قبل اولین ویژگی انتخاب گردید، لذا مقدار $FPE(0)=0$ انتخاب می‌شود.

مرحله ٤: ویژگی مناسب بعدی با یافتن اندیس متغیر حداکثرکننده رابطه (٦) انتخاب می‌شود. از آنجاییکه نقاط ویژگی بطور اولیه توسط آمارگان T فیلتر شده‌اند، این شرط به معنای انتخاب متغیرهایی با حداکثرتمایز برمبنای آمارگان T می‌باشد که در اینجا روش حداکثر تمایز – آمارگان T (MDTS) نامیده می‌شود.

مرحله ۵: از رابطه (۷) مقدار خطای پیش‌بینی نهایی محاسبه می‌شود و سپس، شرط زیر جهت ارزیابی همگرایی و توقف جستجو مورد بررسی قرار می‌گیرد:

$$\left(FPE(k)-FPE(k-1)\right)\le\varepsilon \qquad (۹)$$

مرحله ٦: در صورت تحقق رابطه (۹) به جستجو خاتمه داده و مقدار k بدست آمده به عنوان بعد نهایی زیرفضا پذیرفته می‌شود، در غیر اینصورت به مرحله (٤) رفته و جستجو را تا برقراری کامل شرط همگرایی ادامه می‌دهیم.

دررابطه(۹)، مقدار ε توسط کاربر تعریف می‌شود. با توجه به نتایج اکثر روشهای‌آماری می‌توان این مقدار را دربازه $\varepsilon<0.05$ انتخاب نمود. اگر این مقدار را بزرگ انتخاب نماییم، همگرایی روش سریعتر می‌شود و امکان از دست دادن اطلاعات مفید وجود دارد. با کوچک انتخاب نمودن این مقدار، همگرایی را کند کرده‌ایم ولی اطلاعات بیشتری را جهت تفکیک نهایی کلاسها حفظ نموده‌ایم. درهرصورت انتخاب مناسب مقدار ε مصالحه‌ای بین زمان همگرایی و از دست‌دادن اطلاعات مفید می‌باشد.

۵- نتایج

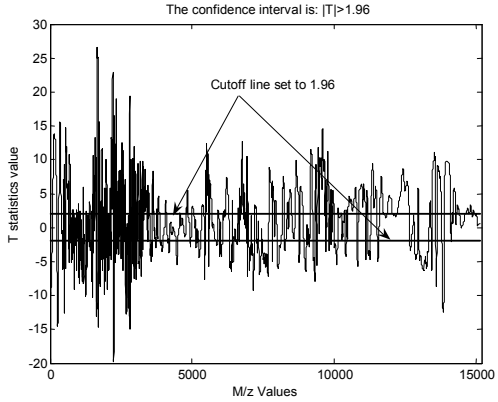
جهت نشان دادن کارآیی روش پیشنهادی(MDTS)، داده طیف سنجی‌جرمی لیزری بیماران سرطان تخمدان مورد تحلیل قرار گرفت. پیش‌پردازش اعمال شده به هر یک از نمونه‌های طیف‌جرمی مطابق با روش بیان شده صورت پذیرفت. هر طیف بطور جداگانه با استفاده از تبدیل موجک گسسته با تابع مادر دابیشیز مرتبه ٤ به روش آستانه‌دهی نرم جهت حذف اغتشاشهای خط زمینه و نویز الکتریکی مورد پردازش قرارگرفت. سپس هر طیف با روش ذکر شده نسبت به تغییرات مقدار شدت نرمالیزه گردید.

در این مطالعه، داده طیف‌سنجی‌جرمی متشکل از دو کلاس به نامهای گروه سالم یا کنترل و گروه سرطانی می‌باشد. جهت بررسی عملکرد روش انتخاب زیرمجموعه ویژگی پیشنهاد شده، بررسی قدرت تمایز نشانگرهای حیاتی منتخب بین دو کلاس موجود، و همچنین حل مشکل تکرارپذیری درون گروهی، داده‌ها به دو مجموعه یادگیری و آزمایش تقسیم شدند. در تقسیم داده‌های هر کلاس به زیرمجموعه‌های یادگیری و آزمون، از روش ارزیابی متقابلK

چرخشی[1] استفاده شد. با انتخاب K=۱۰ داده های هر مجموعه بطور تصادفی با تکرار ۱۰ به دو گروه یادگیری و آزمون تقسیم شدند.

۵-۱- اعمال فیلتر آمارگان T

با توجه به کارآیی آمارگان T در تعیین نقاط مهم در بین دو کلاس متفاوت [۷و۸]، جهت انتخاب حداکثر نقاط دارای اطلاعات مفید بازه اطمینان ۹۵٪ درصد درنظرگرفته‌شد که مقدار آستانه برای آن برابر ۱.۹٦ بدست می‌آید. شکل٤ بیانگر آمارگان محاسبه شده دربین مجموعه یادگیری می‌باشد که با اعمال آستانه موردنظر تعداد نمونه‌های ویژگی از ۱۵۱۵٤ نقطه به ۱۰٤۸۱ نقطه کاهش خواهد یافت.



شکل ٤- نمودار آمارگانT محاسبه شده در بین مجموعه یادگیری (مقدار آستانه ۱.۹٦)

۵-۲- تعیین تعداد ویژگیها

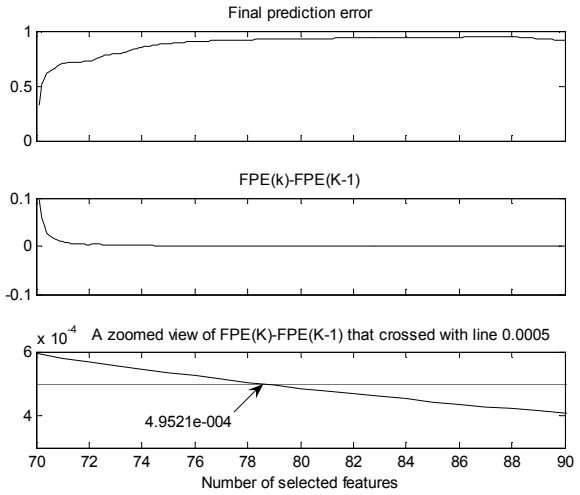
در گام بعدی، باتوجه به مراحل ذکرشده در الگوریتم ۱ و بکارگیری روش حداکثر تمایز نسبت به انتخاب ویژگیهای مناسب از بین تعداد نقاط باقیمانده اقدام گردید. با انجام آزمایشات مختلف روی مجموعه یادگیری که هربار بطور تصادفی از مجموعه کل نمونه‌برداری گردید، مقدار مناسب برای آستانه رابطه (۹) برابر ۰.۰۰۰۵ انتخاب شد. شکل۵ نمودار مقدار خطای نهایی را برای ۲۰۰ نمونه از ۱۰٤۸۱ نقطه با معیار حداکثر تمایز نشان می‌دهد و شرط توقف محاسبه شده از روی رابطه (۹) نیز در همین شکل دیده می‌شود. همچنین، در این شکل تلاقی مقدار تفاضل خطا با خط ۰.۰۰۰۵ مشاهده می‌شود که در نقطه۸۰ از خطای نهایی، برابر با نقطه ۷۹ از تفاضل خطا، شرط توقف صادق گردید. با استفاده از روش حداکثرتمایز- آمارگان T (MDTS) که در الگوریتم۱ توصیف شده‌است، تعداد ویژگیها از ۱۰٤۸۱ نقطه به ۸۰ نقطه کاهش‌داده‌شد.

۵-۳- انتخاب نشانگرهای حیاتی

با استفاده از زیرمجموعه ویژگی بدست‌آمده که در مجموعه دادهI و II بترتیب دارای ابعاد۸۰ × ۲۰۰ و۸۰ × ۲۵۳ می‌باشد و روش ارزیابی‌متقابلK چرخشی با مقدار ۱۰ = K، مجموعه های یادگیری و آزمون تشکیل‌داده‌شد. از طبقه‌بند تحلیل‌تمایزخطی[2] جهت نشان‌دادن کارآیی‌عملکرد روش پیشنهادی درمقایسه با روش آمارگان T استفاده‌شد. با توجه به مجموعه یادگیری، در هر بار آزمایش تعداد ۳۰ ویژگی با روشهای پیشنهادی و آمارگان T استخراج شد و سپس با محاسبه تابع هیستوگرام در بین ویژگیهای انتخاب‌شده، نشانگرهای منتخب با تکرار بیشتر از ۱ جهت بررسی عملکرد با طبقه‌بند در بردار نهایی ویژگی قرارداده‌شد.

1- K fold cross validation

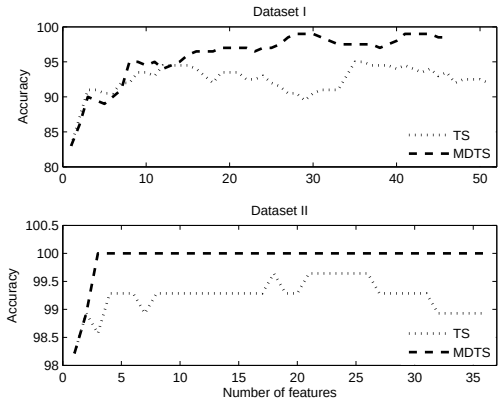
2- Linear discriminate analysis



شکل ٥- نمودار مقدار خطای نهایی به همراه مشتق آن، و یک نمای بزرگ شده از شکل وسط در تقاطع با خط ۰.۰۰۰۵

شکل٦ نتایج حاصل از طبقه‌بندی با استفاده از نشانگرهای انتخاب‌شده توسط دو روش در بین مجموعه آزمون از داده‌های موجود را نشان‌می‌دهد. همانطور که در شکل مشاهده می‌شود، روش پیشنهادی دارای بهبود قابل توجهی از نظر درصد تشخیص نسبت به روش آماری می‌باشد. این بهبود در مجموعه داده I که در شرایط آزمایشگاهی نامناسب تولید شده‌است و دارای واریانس تغییرات زیاد می‌باشد، مشهودتر است.

از بردار ویژگی نهایی با استفاده از روش تحلیل تمایز خطی بترتیب تعداد ١٦ و ٦ نشانگرحیاتی برای مجموعه‌داده I و II انتخاب شد که جهت بررسی کارآیی و قدرت تمایز این نشانگرها، از معیارهای دقت تشخیص، قطعیت، و حساسیت استفاده شد که معیارهای فوق بطور معمول در بررسی نتایج کارهای غربالگری[1] بکارمی‌رود.



شکل ٦- بررسی عملکرد روش پیشنهادی با روش آمارگان T از نظر درصد طبقه‌بندی

٦- بحث

در جدول۲ مقایسه نتایج حاصل از روش پیشنهادی با نشانگرهای گزارش‌شده در مراجع[۱۲و۲۹] برروی مجموعه داده I ارایه‌شده‌است. مقادیر حاکی از بهبود قابل توجه نتایج روش پیشنهادی در قیاس با روش مبتنی بر شبکه عصبی و الگوریتم‌ژنتیک، و روش توسعه یافته آماری دارد. در جدول۳ نتایج روش پیشنهادی با نشانگرهای گزارش شده در مراجع [۳۰و۳۱] برروی مجموعه‌داده II مقایسه‌شده‌است. در این دو مرجع با استفاده از روشهای تحلیل ممیز حداقل مربعات جزیی متعامد و انتخاب متغیر بیز مبتنی بر تبدیل موجک، نشانگرهای حیاتی انتخاب گردید که این نشانگرها در قیاس با روش پیشنهادی در این تحقیق با توجه به تعداد ویژگی بکاررفته، از نتایج ضعیف تری برخورداراست.

1- Screening task

جدول ۲- مقایسه نتایج حاصل از روش پیشنهادی با مراجع ۱۲ و ۲۹ بااستفاده از طبقه‌بند LDA باتوجه به تعداد نشانگرهای‌حیاتی و قدرت تفکیک‌پذیری در مجموعه‌داده I

روش انتخاب ویژگی	تعداد ویژگی	دقت	قطعیت	حساسیت
MDTS	۱۶	٪۱۰۰	٪۱۰۰	٪۱۰۰
مرجع ۱۲	۵	٪۷۱	٪۶۶	٪۷۶
مرجع ۲۹	۱۸	٪۹۲.۵	٪۸۸	٪۹۷

جدول ۳- مقایسه نتایج حاصل از روش پیشنهادی با مراجع ۳۰ و ۳۱ بااستفاده از طبقه‌بند LDA در مجموعه‌داده II برحسب تعداد نشانگرهای‌حیاتی و قدرت جداپذیری کلاسهای الگو

روش انتخاب ویژگی	تعداد ویژگی	دقت	قطعیت	حساسیت
MDTS	۶	٪۱۰۰	٪۱۰۰	٪۱۰۰
مرجع ۳۰	۱۰	٪۹۹.۶۸	٪۹۹	٪۱۰۰
مرجع ۳۱	۹	٪۹۸.۵۷	٪۹۶	٪۱۰۰

باتوجه به نتایج ارایه‌شده، روش پیشنهادی در قیاس با روشهای دیگر بدلیل انتخاب بهترین زیرمجموعه در بین متغیرهایی که از نظر آماری دارای اختلاف معنی‌دار می‌باشند، منجر به حصول جوابهای بهتر شده‌است. استفاده از معیار سنجش محتوای اطلاعاتی مبتنی برفاصله باچاتاریا به همراه ملاک انتخاب مرتبه مدل براساس خطای پیش بینی نهایی، عامل اصلی بهبود نتایج می‌باشد.

در این مقاله نشان داده‌شده‌است که با استفاده از روش ترکیبی و اعمال پیش‌پردازش مناسب روی داده خام ورودی می‌توان کاهش‌بعد قابل ملاحظه‌ای انجام داد، بدون آنکه محتوای اطلاعاتی داده ورودی از نظر قدرت تفکیک‌پذیری در بین گروههای سالم و سرطانی از دست برود.

مزیت و کارآیی روش پیشنهادی نسبت به روشهای دیگر زمانی ملموس‌تر خواهدبود که درتحلیل مجموعه‌داده‌های ضعیف از نظر نحوه تولید و توزیع نقاط‌ویژگی نیز به جوابهای قابل‌قبول می‌رسد. این‌نتیجه‌مهم، امیدواری به دستیابی نشانگرهای‌جدید با قدرت تمایزبالا و قابلیت‌تکرارپذیری درون‌گروهی و بین گروهی را افزایش‌می‌دهد که یکی‌از مشکلات‌اساسی در حوزه تحلیل محتوای پروتیینی سیگنال طیف‌جرمی‌لیزری می‌باشد [۳۲].

۷- نتیجه‌گیری

تشخیص بیماری در علم پزشکی نمونه‌ای از تفکیک الگو در علوم مهندسی می‌باشد. بدون تردید، یکی از عوامل موفقیت در زمینه تشخیص الگو، استخراج و انتخاب نقاط ویژگی مناسب از داده‌های خام ورودی است. از سویی دیگر، پیش پردازش مناسب داده ورودی قبل از عمل انتخاب زیر مجموعه نقاط ویژگی، نقشی کلیدی در رسیدن به موفقیت ایفا می‌نماید.

در این مقاله، یک روش فیلتری انتخاب زیرمجموعه ویژگی معرفی گردید که با ترکیب روشی آماری و معیارفاصله مبتنی بر سنجش ارزش اطلاعاتی، ویژگیهای مناسب را در بین فضای ورودی برمی‌گزیند. با استفاده از معیار خطای پیش‌بینی نهایی و روش پیشنهاد شده (MDTS)، ضریب کاهش‌بعد ۹۴۷:۱ و ۲۵۲۶:۱ بترتیب در بین مجموعه‌داده‌های I و II حاصل شد که ابعاد داده ورودی را از ۱۵۱۵۴ نقطه به تعداد ۱۶ و ۶ نقطه کاهش داد. با استفاده از نشانگرهای حیاتی منتخب و روش ارزیابی متقابل K چرخشی، مقادیردقت‌تشخیص ٪۱۰۰، قطعیت ٪۱۰۰ و حساسیت ٪۱۰۰ در بین مجموعه آزمون بدست آمد.

در بیشتر کارهای غربالگری و همچنین طبقه‌بندی کلاسهای الگو، یک مجموعه از ویژگیهای خوب می‌تواند منجر به حصول تفکیک‌پذیری بالا گردد. از سویی‌دیگر، داده‌های مرتبط با تشخیص بیماری دارای ابعاد بالا می‌باشد که انتخاب اولیه این نقاط با استفاده از روشهای

آستانه‌دهی به دلیل در دست نبودن یک معیار مناسب برای تعیین آستانه، خالی از اشکال نخواهدبود.

نتایج بدست آمده بر این نکته تاکید دارد که بکارگیری روشهای ترکیبی در استخراج و انتخاب ویژگی از فضاهای با ابعاد بالا، علاوه بر حفظ محتوای اطلاعاتی فضای اولیه، بخوبی می‌تواند کلاسهای الگو را از هم تفکیک نماید. هرچند که اعمال یک مقدار حداقل آستانه (P-value < ۰.۰۵) جهت انتخاب اولیه نقاط دارای اختلاف معنی‌دار در بین گروههای کلاس الگو ضروری بنظر می‌رسد ولی بایستی متذکر شد که استفاده از معیارهای مناسب در الگوریتمهای ترکیبی عامل اصلی رسیدن به موفقیت و از دست ندادن اطلاعات معنادار می‌باشد.

منابع

1. Srinivas P. R, Srivastava S, Hanash S, Wright G. L. Proteomics in early detection of cancer. Clin Chem 2001; 47(10): 1901-1911.
2. Petricoin E.F.III, 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. JNCI 2002; 94(20): 1576-1578.
3. Jemal A, Thomas A, Murray T, Thun M. Cancer statistics. CA Can J Clin 2002; 52: 23-47.
4. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinf 2003; 19(13): 1636-1643.
5. 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. Bioinf 2005; 21(9): 1764-1775.
6. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. CSB 2003.
7. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Gen Res 2001; 11: 1878-1887.
8. Chen G, Gharib T.G, Huang C.C, Thomas D.G, Shedden K.A, Taylor J.M.G, Kardia S.L.R, Misek D.E, Giordano T.J, Iannettoni M.D, Orringer M.B, Hanash S.M, Beer D.G. Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. Clin Can Res 2002; 8: 2298-2305.
9. Sorace J.M, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinf 2003; 4: 24-34.
10. Adam B.L, Qu Y, Davis J.W, Ward M.D, Clements M.A, Cazares L.H, Semmes O.J, Schellhammer P.F, Yasui Y, Feng Z, Wright G.L. Serum protein fingerprint-ing coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Can Res 2002; 62: 3609-3614.

11. Qu Y, Adam B.L, Yasui Y, Ward M.D, Cazares L.H, Schellhammer P.F, Feng Z, Semmes O.J, Wright G.L. Boosted decision tree analysis of SELDI mass spectral serum profiles discriminates prostate cancer from non-cancer patients. Clin Chem 2002; 48(10): 1835-1843.

12. Petricoin E.F.III, 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," Lan 2002; 359: 572-577.

13. Oh J.H, Gao J, Nandi A, Gurnani P, Knowles L, Schorge J, Rosenblatt K.P. Diagnosis of early relapse in ovarian cancer using serum proteomic profiling. Gen Inf 2005; 16(2): 195-204.

14. Phillips M, Gleeson K. Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study. Lan 1999; 353: 1930-1933.

15. Finehout E.J, Lee K.H. An introduction to mass spectrometry applications in biological research. BMBE 2004; 32(2): 93-100.

16. Hilario M, Kalousis A, Pellegrini C, Muller M. Processing and classification of protein mass spectra. Mass Spec Riv 2006; 25: 409-449.

17. Boratyn G.M, Merchant M.L, Klein J.B. Utilization of Human Expert Techniques for Detection of Low-Abundant Peaks in High-Resolution Mass Spectra. 28th IEEE EMBS 2006; NY USA.

18. Donoho D.L, Johnstone I.M. Threshold selection for wavelet shrinkage of noisy data. 16th IEEE EMBS 1994; 24a - 25a.

19. Liu B.F, Sera Y, Matsubara N, Otsuka K, Terabe S. Signal denoising and baseline correction by discrete wavelet trandform for microchip capillary electrophoresis. Eleph 2003; 24: 3260-3265.

20. Ojanen J, Miettinen T, Heikkonen J, Rissanen J. Robust denoising of electrophoresis and mass spectrometry signals with minimum description length principle. FEBS 2004; 570: 107-113.

21. Ruckstuhl A.F, Jacobson M.P, Field R.W, Dodd J.A. Baseline subtraction using robust local regression estimation. Qua Spec Rad Tran 2001; 68: 179-193.

22. Donoho D.L. De-Noising by Soft-Thresholding. IEEE Trans. On Information Theory 1995; 41(3): 613-627.

23. Ravier P, Amblard P.O. Wavelet packets and de-noising based on higher-order-statistics for transient detection. SP 2001; 81: 1909-1926.

24. Petricoin E.F.III, Liotta L.A. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancers. Cur Opn Biotech 2004; 24: 24-30.

25. Frosini G, Lazzerini B, Marcelloni F. A modified fuzzy C-means algorithm for feature selection. 19th NAFIPS 2000; 148-152.

26. Hollander M, Wolfe D.A. Nonparametric Statistical Methods. 2nd Edition, Wiley 1999, 121-125.

27. S. Theodoridis, K. Koutroumbas, Pattern Recognition, 2nd Edition, Academic Press 2003, 174-183.

28. Zhang P, Shaman P. Assessing Prediction Error in Autoregressive Models. Tran AMS 1995; 347(2): 627-637.

29. 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): 14666-14671.

30. Whelehan O.P, Earll M.E, Johansson E, Toft M, Eriksson L. Detection of ovarian cancer using chemometric analysis of proteomic profiles. Chem Intl lab sys 2006; 84: 82-87.

31. Vannucci M, Sha N, Brown P.J. NIR and mass spectra classification: baysian methods for wavelet-based feature selection. Chem Intl lab sys 2005; 77: 139-148.

32. Baggerly K.A, Morris J.S, Edmonson S.R, Coombes K.R. Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. JNCI 2005; 97(4): 307-309.