

الگوریتم ترکیبی داده های طیف جرمی لیزری سرطان تخمدان

ارایه یک الگوریتم جدید ترکیبی انتخاب زیرمجموعه ویژگی جهت تحلیل داده های طیف جرمی لیزری سرطان تخمدان

حسین منتظری کردی^۱، محمدحسین میران بیگی^{۲*}، محمدحسن مرادی^۳

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 پذیرش مقاله: ۸۶/۱۱/۲۴

تاریخ دریافت نسخه اصلاح شده: ۸۶/۷/۱۹

چکیده

مقدمه: یکی از مشکلات اساسی در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص بموقع آن می باشد. واکنشهای شیمیایی درون یک عضو زنده می تواند بصورت الگوهای پروتئینی در مایعاتی نظیر خون، خلط، و ادرار انعکاس داده شود. طیفسنج جرمی لیزری یک ابزار مفید جهت استخراج اطلاعات پروتئینی از نمونه های وابسته به موجودات زنده می باشد. انتخاب بهینه زیرمجموعه ویژگی در بین نقاط داده طیف جرمی از چالشهای مهم این حوزه محسوب می شود.

مواد و روشها: در این تحقیق، داده های پروفایل پروتئینی خونابه بیماران مبتلا به سرطان تخمدان در دو گروه مجزا مورد تحلیل قرار گرفت. با استفاده از یک مدل ریاضی، اغتشاشات خطرزمینه و نویز الکترونیکی در مرحله پیش پردازش حذف گردید و سپس، تمام سیگنالهای طیف جرمی نرمالیزه شدند. روش پیشنهادی، ترکیبی از تست آماری و اندازه فاصله باچاتاریا می باشد که با استفاده از معیار پیش بینی خطای نهایی، حداقل ویژگیهای لازم برای انتخاب بهترین زیرمجموعه از بین ۱۵۱۵۴ نقطه داده موجود را با حفظ ارزش اطلاعاتی و قدرت تفکیک پذیری برمیگزیند که این زیرمجموعه جهت آشکارسازی نشانگرهای حیاتی بکار می رود.

نتایج: با استفاده از روش ارزیابی متقابل K چرخشی، نمونه های موجود در گروه های مورد مطالعه به دو دسته یادگیری و آزمون تقسیم شدند. با اعمال حداقل آستانه نقاط دارای اختلاف معنی دار انتخاب شدند و سپس، بهترین زیرمجموعه 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 نشانگر حیاتی بترتیب برای زیرمجموعه های داده I و II انتخاب شد که در قیاس با روشهای دیگر دارای دقت تشخیص ۱۰۰٪، قطعیت ۱۰۰٪، و حساسیت ۱۰۰٪ می باشد.

بحث و نتیجه گیری: تشخیص بیماری در علم پزشکی نمونه ای از تفکیک الگو در علوم مهندسی می باشد. در این مقاله، یک روش فیلتری انتخاب زیرمجموعه ویژگی معرفی گردید که با ترکیب روشی آماری و معیار فاصله 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 می تواند کلاسهای الگو را از هم تفکیک نماید. (مجله فیزیک پزشکی ایران، دوره ۴، شماره ۱۴ و ۱۵، بهار و تابستان ۸۶: ۹۶-۸۳)

واژگان کلیدی: پروتئین شناسی، سرطان تخمدان، طیف جرمی لیزری، الگوریتم انتخاب زیرمجموعه ویژگی، نشانگر حیاتی

* نویسنده مسؤول: محمدحسین میران بیگی

آدرس: گروه مهندسی پزشکی، بخش برق، دانشکده فنی، دانشگاه

تربیت مدرس تهران
miranbmi@modares.ac.ir

تلفن: ۰۲۱-۸۸۰۱۱۰۰۱ (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 می باشد. تلاش جهت رفع این مشکل منجر به ارایه برخی از روشهای تشخیصی گردیده است که عیب عمده آنها عدم دقت کافی با توجه به هزینه بالا می باشد. با توجه به اطلاعات حوزه علم ژنتیک، واکنشهای شیمیایی درون یک عضو زنده می تواند بصورت الگوهای پروتئینی در مایعاتی نظیر خون، خلط، و ادرار انعکاس داده شود [۱]. واژه پروتئین شناسی^۱ به علمی اطلاق می شود که امکان مقایسه کمی و کیفی و همچنین تمایز قایل شدن بین پروتئینها تحت شرایط مختلف پاتولوژیک برای فهم فرایندهای بیولوژیک، نظیر شناسایی عوامل وابسته به بیماری، را فراهم نماید.

حوزه فعالیت پروتئین شناسی بطورعام به هر نوع تکنولوژی یا تکنیک پردازش اطلاعاتی مربوط می شود که بتواند داده های پروتئینی با مقیاس بالا تولید نموده و یا آن را مورد تحلیل قرار دهد [۲]. یکی از ابزارهای مورد استفاده جهت استخراج اطلاعات پروتئینی از نمونه های وابسته به موجودات زنده، طیف سنج جرمی لیزری^۲ می باشد. تحلیل محتوای اطلاعاتی طیف جرمی یک روش سریع و ارزان در تشخیص بیماری بدون ایجاد هرگونه عوارض جانبی می باشد که می تواند امکان بالقوه غربالگری سرطان را فراهم سازد. سرطان تخمدان یکی از عوامل شایع مرگ و میر زنان محسوب می شود که تشخیص بموقع این بیماری به نشانگرهای حیاتی^۳ جدید با قدرت تشخیص بالا نیاز دارد [۳]. استفاده از تکنیک طیف سنج جرمی لیزری بهمراه تحلیل صحیح سطوح تفسیر پروتئینها می تواند گامی موثر در راه حل این مشکل و کشف نشانگرهای حیاتی به شمار آید.

با استفاده روبه رشد طیف سنجی جرمی لیزری برای تهیه پروفایلهای پروتئینی، برخی از چالشهای مهم در ارتباط با تحلیل این داده ها افزایش یافته است. از جمله آنها می توان به مواردی نظیر: حذف نویز خط زمینه و الکتریکی، نرمالیزه کردن طیف جرم، تعیین و تنظیم پیکها، و انتخاب متغیرهای مهم در بین مجموعه داده اشاره نمود [۴].

موریس و همکارانش در یکی از تحقیقات خود [۵] متذکر شده اند که مرحله انتخاب ویژگی در تحلیل سیگنال طیف جرمی از اهمیت ویژه ای برخوردار است. دینگ و همکارانش در تحقیقی دیگر [۶] نشان داده اند که در استخراج نشانگرهای حیاتی از داده های با ابعاد بالا، اتخاذ یک استراتژی انتخاب بهترین زیرمجموعه ویژگی نقش اساسی دارد. ژونگ و همکارانش در تحقیقشان [۷] و همچنین چن و همکارانش در تحقیق خود [۸] از تستهای آماری جهت انتخاب اولیه نقاط ویژگی استفاده نموده اند. سوراس و همکارانش در تحقیقی دیگر [۹] با استفاده از تست آماری ویلکاکسون و انتخاب آستانه 10^{-6} به انتخاب نقاط ویژگی معنی دار پرداختند.

آدام و همکارانش در مطالعه خود [۱۰] با استفاده از روش درخت تصمیم، داده طیف جرم لیزری سرطان پروستات را مورد تحلیل قرار دادند. کو و همکارانش در ادامه [۱۱] روش درخت تصمیم مبتنی بر تقویت^۴ را روی مجموعه داده [۱۰] پیاده سازی کردند. پتريکون و همکارانش در تحقیقات خود [۲ و ۱۲] با بکارگیری الگوریتم ژنتیک و شبکه عصبی خودسازمانده در جهت کشف نشانگرهای حیاتی از سیگنال طیف جرمی لیزری سرطان پروستات و تخمدان تلاش نمودند. یونگ و همکارانش [۱۳] با توسعه یک روش ترکیبی انتخاب پیک مبتنی بر ماشین حامی بردار^۵، داده طیف جرمی لیزری را تحلیل کردند. عیب عمده این مطالعات، علاوه بر عدم اتخاذ یک روش مناسب جهت حذف اغتشاش خط زمینه و نویز الکتریکی، کنارهم قراردادن

4- Boosting
5- Support Vector Machine (SVM)

1- Proteomics
2- Laser Mass Spectrometry
3- Biomarkers

الگوریتم ترکیبی داده های طیف جرمی لیزری سرطان تخمدان

در یک طیفسنج جرمی ابتدا نمونه در مرحله آماده سازی جهت یونیزه کردن مولکولها قرار می گیرد. در طیفسنجهای کروماتوگراف، مرحله یونیزاسیون با حرارت دادن محلولی حاوی نمونه صورت می پذیرد که برای نمونه های بیولوژیک بدلیل ناپایداری در برابر حرارت این امکان وجود ندارد. با اختراع طیفسنجهای جرمی لیزری مشکل یونیزاسیون نمونه مرتفع شده است.

در طیفسنج جرمی لیزری، نمونه بیولوژیک بر روی یک صفحه یا ماتریس جذب کننده انرژی قرار می گیرد. پرتوهای لیزر به این صفحه تابانده شده و انرژی آن توسط ماتریس جذب می شود. سپس، قسمتی از انرژی جذب شده به ماده بیولوژیک بازپس داده می شود که موجب یونیزه شدن مولکولهای ماده می گردد. این یونها تحت تاثیر یک ولتاژ بالا در یک لوله خلاء به سمت آشکارساز حرکت می کنند که زمان رسیدن مولکولها به آشکارساز برحسب جرم آنها متفاوت می باشد. خروجی آشکارساز یک طیف می باشد که محور افقی بیانگر نسبت جرم به بار و محور عمودی نیز نشاندهنده غلظت یا فراوانی یک مولکول است [۱۵]. شکل ۱ نمایی از یک طیفسنج جرمی لیزری را جهت تهیه طیف جرم نشان می دهد.

۳- مواد و پیش پردازش

در این مقاله از داده های پروفایل پروتئینی خونابه^۱ که توسط دستگاه طیفسنجی جرمی با تکنیک جذب- یونیزاسیون لیزری سطحی ارتقا یافته زمان پروازی^۲ تهیه شده است، به عنوان داده ورودی برای الگوریتمهای پردازشی استفاده شده است. سپس، با استفاده از این داده ها غربالگری بیماران مبتلا به سرطان تخمدان و افراد تحت کنترل سالم بمنظور دستیابی به نشانگرهای حیاتی متمایز کننده دو گروه انجام شده است.

مجموعه ای از ویژگیها با رتبه منفرد بالا بوده است. یک روش انتخاب ویژگی مناسب ضمن حفظ معنی دار بودن نقاط ویژگی بایستی حافظ محتوای اطلاعاتی بردار ویژگی نیز باشد که در روشهای قبلی فقط حصول نتایج طبقه بندی بالا مدنظر بوده است.

در این مقاله، روشی ترکیبی ارایه شده است که با استفاده از یک تست آماری در مرحله اول بعد داده طیف جرمی را کاهش می دهد و سپس با حداکثر نمودن معیار مبتنی بر فاصله بین نقاط ویژگی، بهترین زیرمجموعه از آنها را استخراج می نماید. این روش ضمن حفظ معنی دار بودن نقاط ویژگی، محتوای اطلاعاتی بردار ویژگی را نیز محفوظ نگه می دارد که مبین قدرت تمایز بین گروههای سالم و سرطانی می باشد. با روش ارایه شده، تعداد ۸۰ ویژگی از بین ۱۵۱۵۴ نقطه برگزیده شد که از این زیرمجموعه تعداد ۱۶ و ۶ پروتئین به عنوان نشانگر حیاتی انتخاب گردید. همچنین، عملکرد ویژگیهای منتخب با استفاده از چند روش طبقه بندی مناسب برای داده های طیف جرمی مورد بررسی قرار گرفت. با استفاده از روش ارزیابی متقابل K چرخشی، مقادیر دقت تشخیص ۱۰۰٪، قطعیت ۱۰۰٪، و حساسیت ۱۰۰٪ در بین مجموعه آزمایش حاصل گردید.

۲- طیفسنجی جرمی لیزری

تاریخ پزشکی همواره شاهد تحولات شگرفی بوده است که اختراعات مهم در عرصه تکنولوژی پدید آورده است. تکنیک طیفسنجی جرمی در علم شیمی قدمتی طولانی دارد و بیشتر با عنوان طیفسنج جرم کروماتوگراف گاز شناخته می شود. از این تکنیک جهت تحلیل ترکیبات گازی هوای بازدم به منظور کشف عوامل بیماری استفاده شده است [۱۴]. ولی استفاده از این تکنولوژی برای تحلیل نمونه های بیولوژیک نظیر خون بدلیل محدودیتهای تکنیکی در آماده سازی نمونه جهت اعمال به طیفسنج جرم تا سالهای اخیر غیر ممکن بوده است.

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



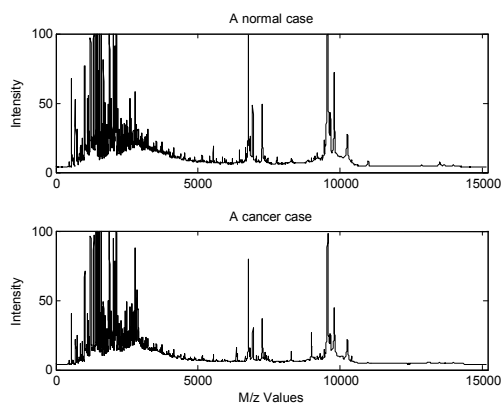
۱-۳- داده‌های طیف سنجی جرمی

در این تحقیق، جهت تعیین الگوهای پروتئینی از خونابه که باعث تمایز و تشخیص نمونه‌های مبتلا به سرطان تخمدان از افراد تحت کنترل سالم می‌شود، مجموعه داده‌های طیف جرمی 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 جدول ۱ آورده شده است.

جدول ۱- نحوه توزیع نمونه‌ها در مجموعه داده‌های تحت مطالعه

مجموعه داده	تعداد نمونه‌های سالم	تعداد نمونه‌های سرطانی	تعداد نمونه‌های با تومور خوشخیم	تعداد ویژگی‌ها در هر طیف جرم
I	۱۰۰	۱۰۰	۱۶	۱۵۱۵۴
II	۹۱	۱۶۲	۰	۱۵۱۵۴

در این مجموعه داده‌ها، هر منحنی طیف جرم دارای ۱۵۱۵۴ نقطه روی محور نسبت جرم به بار (M/Z) می‌باشد که متناظر با آن یک نقطه روی محور شدت سیگنال وجود دارد. شکل ۲ دو نمونه از منحنی طیف جرمی را برای داده‌های تحت مطالعه مجموعه II در جدول ۱ نشان می‌دهد که بترتیب مربوط به یک فرد سالم، و بیمار سرطان تخمدان با اثبات بیماری به روش بیوپسی می‌باشد.



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

۲-۳- مدل ریاضی طیف جرم

از دیدگاه مدلسازی، سیگنال طیف جرمی می تواند با یک مدل ترکیبی نشان داده شود [۵ و ۱۶] که اجزای این مدل می بایست در برگیرنده عوامل شیمیایی و الکتریکی یک طیف سنج جرم باشد. فرض می شود که n طیف نمونه برداری شده در بازه زمانی T از زمان پرواز در فواصل زمانی $t_j, j=1, \dots, T$ مشاهده شده است. مدل زیر را می توان برای سیگنال طیف جرمی در نظر گرفت [۵، ۱۷]:

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \varepsilon_{ij} \quad (1)$$

که در این مدل، $y_i(t_j)$ شدت سیگنال یا فراوانی یک مولکول برای هر جرم مجزا را در زمان t_j نشان می دهد. $B_i(t_j)$ به خطای سیستم ناشی از ماده شیمیایی بکاررفته در دستگاه جهت ثابت نگه داشتن نمونه بیولوژیک اشاره دارد که به عنوان نویز خط زمینه شناخته می شود. $S_i(t_j)$ سیگنال اصلی ناشی از مولکولهای پروتئینی موجود در نمونه بیولوژیک می باشد که در عامل مقیاس N_i ضرب شده است. ε_{ij} معرف اغتشاش الکتریکی دستگاه طیف سنج جرمی است که توزیع گوسی برای آن فرض می شود.

۳-۳- حذف اغتشاشات خط زمینه و الکتریکی

با توجه به مدل ارائه شده، جهت کاهش اثرات مخرب ناشی از خط زمینه و نویز الکتریکی در مرحله انتخاب ویژگی بایستی نسبت به حذف آنها اقدام نمود. با توجه به کارایی تبدیل موجک گسسته در پردازش سیگنالها با تغییرات زیاد، از این ابزار جهت حذف اغتشاش از طیف جرم 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 می باشد، در ضرایب تقریبات^۱ ظاهر می شود [۱۹] و نویز الکتریکی نیز

در ضرایب جزئیات^۲ انعکاس می یابد [۲۰]. با استفاده از روش توصیف شده در [۲۱]، ضرایب تقریبات جهت تنظیم خط زمینه مورد پردازش قرار گرفتند. از دابیشیز مرتبه ۴ به عنوان تابع ویولت مادر مناسب استفاده گردید. نویز الکتریکی که بصورت عاملی با تغییرات تیز فرض می شود، با استفاده از روش بیان شده در [۲۲] به روش آستانه دهی نرم حذف گردید. مقدار آستانه مناسب برای حذف نویز الکتریکی بر مبنای آمارگان مرتبه بالا^۳ محاسبه شد که این مقدار با روش ذکر شده در [۲۳] بدست آمد.

۳-۴- نرمالیزه کردن طیف

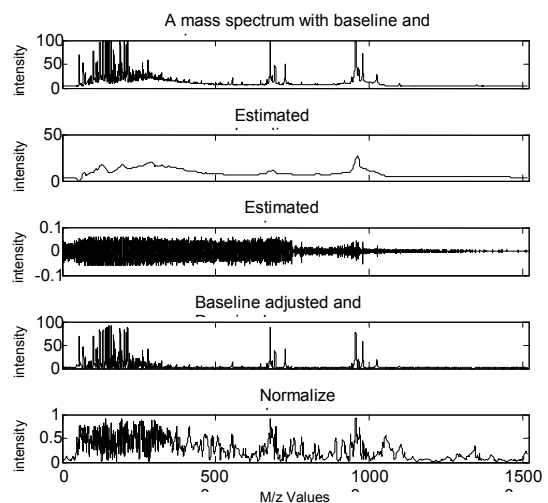
پس از حذف اثرات ناشی از اغتشاشهای خط زمینه و منابع الکتریکی، سیگنال بمنظور حذف اثر ضریب مقیاس با استفاده از رابطه (۲) نرمالیزه گردید [۲۴]:

$$NS = \frac{S - \min(S)}{\max(S) - \min(S)} \quad (2)$$

در این رابطه، NS به مقدار شدت طیف نرمالیزه شده و S به مقدار آن قبل از نرمالیزاسیون اشاره دارد. در این رابطه عبارات $\max(S)$ و $\min(S)$ بترتیب مقادیر حداقل و حداکثر شدت در هر نقطه از سیگنال را نشان می دهند. در شکل ۳ یک نمونه از طیف جرمی قبل از عمل پیش پردازش مشاهده می شود. در همین شکل، اغتشاش خط زمینه تخمین زده شده به همراه نویز الکتریکی نیز دیده می شود. سیگنال طیف جرمی بدون اغتشاش و همچنین سیگنال نرمالیزه شده نیز در ادامه در همین شکل نشان داده شده است.

۴- روش انتخاب ویژگی

روشهای انتخاب ویژگی بطور کلی به دو دسته فیلتری و مبتنی بر یادگیری^۱ 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 ارائه شده است 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 می پردازد.



شکل ۳- یک نمونه از طیف جرمی قبل و بعد از حذف اغتشاشات خطرزمینه و نویز الکتریکی 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]:

$$t_i = \frac{\mu_{i1} - \mu_{i2}}{\sigma_{i1,2}}, \quad i = 1, \dots, M \quad (3)$$

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

$$\sigma_{i1,2} = \sqrt{\frac{(N_1 - 1)\sigma_{i1}^2 + (N_2 - 1)\sigma_{i2}^2}{N_1 + N_2 - 2}} \times \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \quad (4)$$

در رابطه بالا، N_i تعداد نمونه‌های هر کلاس بوده و σ_{ij} نیز واریانس متغیر i ام از کلاس j می‌باشد. با استفاده از مقدار آمارگان محاسبه شده و روش آستانه‌دهی، ویژگیهایی که عدد آمارگان آنها از مقدار آستانه بیشتر باشد، نگه‌داشته شده و مابقی حذف می‌شوند. مقدار آستانه با توجه به در نظر گرفتن توزیع گوسی و یک بازه اطمینان^۲ قابل محاسبه می‌باشد [۲۶].

۴-۲- روش حداکثر تمایز

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 $X = \{x_i, i = 1, \dots, M\}$ نشان داده می‌شود، و هدف انتخاب زیر مجموعه‌ای از آن می‌باشد بطوریکه منجر به تمایز بهینه بین کلاسهای

$$J(S; c) = \frac{1}{8} (\mu_1 - \mu_2)^T (\Sigma_1^{-1} + \Sigma_2^{-1}) (\mu_1 - \mu_2) + \frac{1}{2} \log \left(\frac{|\Sigma_1 + \Sigma_2|}{2(|\Sigma_1| |\Sigma_2|)^{\frac{1}{2}}} \right) \quad (5)$$

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

$$\max(J(S; c)), \quad S = \{x_i, i = 1, \dots, m\} \quad (6)$$

با روش توصیف شده در بالا بدیهی است که با افزودن تعداد نقاط ویژگی به زیرمجموعه، مقدار فاصله باچاتاریا افزایش می یابد و لذا همگرایی روش جهت رسیدن به نقطه بهینه m کند می شود. در اینگونه روشهای انتخاب ویژگی، از شیوه های مرحله ای^۴ یا زیربهینه^۵ استفاده می شود که با بررسی یک معیار توقف در هر مرحله نسبت به همگرا شدن روش تصمیم گیری می شود. در این مقاله، از معیارخطای پیش بینی نهایی آیک^۶ مبتنی بر حداکثر فاصله جهت بررسی شرط همگرایی روش حداکثر تمایز استفاده شده است [۲۸]. فرض شود که متغیر k به یک ویژگی از زیرفضای با بعد کاهش یافته اشاره نماید، آنگاه تابع خطا بصورت زیر تعریف می شود:

$$FPE(k) = \frac{1}{2k} \sum_{i=1}^k (\tilde{J}(S_i; c) - FPE(i-1))^2 \quad (7)$$

در رابطه بالا، $\tilde{J}(S_i; c)$ یک تابع تبدیل یافته از تابع حداکثر تمایز در رابطه (۵) می باشد که جهت همگرایی سریعتر روش، بصورت نمایی زیر در نظر گرفته می شود:

موجود گردد. پس هدف یافتن زیر مجموعه ای m عضوی از فضای \mathbb{R}^M است که باعث تفکیک بهینه کلاسهای موجود گردد. روش فیلتری آمارگان T موجب کاهش بعد اولیه فضای داده ورودی می گردد، ولی این شیوه در انتخاب بهترین زوج ویژگیها هیچ نقشی بعهد ندارد. عبارت دیگر، آمارگان T بهترین ویژگیهای منفرد را از فضای موجود برمیگزیند. از آنجاییکه نشان داده شده است، ترکیبی از بهترین ویژگیهای منفرد همواره نمی تواند منجر به ایجاد یک زیر مجموعه ویژگی بهتر گردد [۶]. لذا این نکته موجب می شود که در فضای زیر مجموعه ویژگی با افزودن اطلاعات مواجه شویم. از اینرو، ارایه یک روش جستجو مبتنی بر اندازه گیری تاثیر انتخاب متغیر جدید در محتوای اطلاعاتی فضای کاهش یافته می تواند موجب کاهش افزودن گردد.

هدف از ارایه روش حداکثر تمایز^۱ (MD) پیدا کردن یک زیرمجموعه با تعداد m عضو از فضای \mathbb{R}^M می باشد 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^۲ معرفی نموده ایم که فاصله باچاتاریا بین نقاط ویژگی را اندازه گیری می کند. 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 کلاسهها اشاره دارد. فاصله باچاتاریا^۳ به عنوان یک اندازه از توان تمایز بین دو کلاس بصورت زیر تعریف می شود [۲۷]:

4- Stepwise
5- Suboptimal
6- Akaik's final prediction error

1- Maximum Discrimination
2- Quadratic Discriminate Function
3- Bhattacharyya

$$\tilde{J}(S_i; c) = 1 - e^{(-J(S_i; c)/i)} \quad (8)$$

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:

الگوریتم 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 ۴: ویژگی مناسب بعدی با یافتن اندیس متغیر حداکثرکننده رابطه (۶) انتخاب می‌شود. از آنجاییکه نقاط ویژگی بطور اولیه توسط آمارگان T فیلتر شده‌اند، این شرط به معنای انتخاب متغیرهایی با حداکثر تمایز بر مبنای 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:

$$(FPE(k) - FPE(k-1)) \leq \varepsilon \quad (9)$$

مرحله ۶: 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 (۴) رفته و جستجو را تا برقراری کامل شرط همگرایی ادامه می‌دهیم.

در رابطه (۹)، مقدار ε توسط کاربر تعریف می‌شود. با توجه به نتایج اکثر روشهای آماری می‌توان این مقدار را در بازه $0.05 < \varepsilon$ 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. در هر صورت انتخاب مناسب مقدار ε مصالحه‌ای بین زمان همگرایی و از دست دادن اطلاعات مفید می‌باشد.

۵- نتایج

جهت نشان دادن کارایی روش پیشنهادی (MDTS)، داده طیف سنجی جرمی لیزری بیماران سرطان تخمدان مورد تحلیل قرار گرفت. پیش‌پردازش اعمال شده به هر یک از نمونه‌های طیف جرمی مطابق با روش بیان شده صورت پذیرفت. هر طیف بطور جداگانه با استفاده از تبدیل موجک گسسته با تابع مادر دابیشیز مرتبه ۴ به روش آستانه‌دهی نرم جهت حذف اغتشاشهای خط زمینه و نویز الکتریکی مورد پردازش قرار گرفت. سپس هر طیف با روش ذکر شده نسبت به تغییرات مقدار شدت نرمالیزه گردید.

در این مطالعه، داده طیف سنجی جرمی 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 متقابل K

الگوریتم ترکیبی داده های طیف جرمی لیزری سرطان تخمدان

آستانه رابطه (۹) برابر 0.0005 انتخاب شد. شکل ۵ نمودار مقدار خطای نهایی را برای 200 نمونه از 10481 نقطه با معیار حداکثر تمایز نشان می دهد و شرط توقف محاسبه شده از روی رابطه (۹) نیز در همین شکل دیده می شود. همچنین، در این شکل تلاقی مقدار تفاضل خطا با خط 0.0005 مشاهده می شود که در نقطه 80 از خطای نهایی، برابر با نقطه 79 از تفاضل خطا، شرط توقف صادق گردید. با استفاده از روش حداکثر تمایز-آمارگان T (MDTS) که در الگوریتم ۱ توصیف شده است، تعداد ویژگیها از 10481 نقطه به 80 نقطه کاهش داده شد.

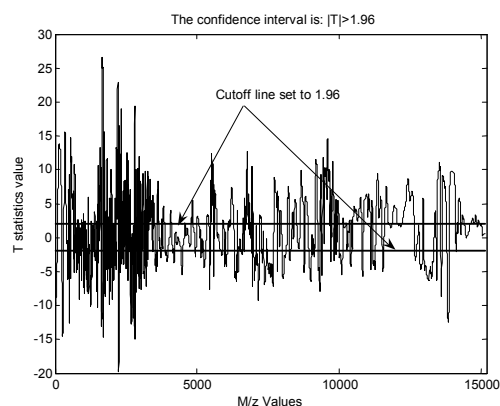
۵-۳- انتخاب نشانگرهای حیاتی

با استفاده از زیرمجموعه ویژگی بدست آمده که در مجموعه داده I و II بترتیب دارای ابعاد 200×80 و 253×80 می باشد و روش ارزیابی متقابل K چرخشی با مقدار $K = 10$ ، مجموعه های یادگیری و آزمون تشکیل داده شد. از طبقه بندی تحلیل تمایز خطی^۲ جهت نشان دادن کارایی عملکرد روش پیشنهادی در مقایسه با روش آمارگان T استفاده شد. با توجه به مجموعه یادگیری، در هر بار آزمایش تعداد 30 ویژگی با روشهای پیشنهادی و آمارگان T استخراج شد و سپس با محاسبه تابع هیستوگرام در بین ویژگیهای انتخاب شده، نشانگرهای منتخب با تکرار بیشتر از ۱ جهت بررسی عملکرد با طبقه بندی در بردار نهایی ویژگی قرار داده شد.

چرخشی^۱ استفاده شد. با انتخاب $K=10$ داده های هر مجموعه بطور تصادفی با تکرار 10 به دو گروه یادگیری و آزمون تقسیم شدند.

۵-۱- اعمال فیلتر آمارگان T

با توجه به کارایی آمارگان T در تعیین نقاط مهم در بین دو کلاس متفاوت [۷ و ۸]، جهت انتخاب حداکثر نقاط دارای اطلاعات مفید بازه اطمینان 95% درصد در نظر گرفته شد که مقدار آستانه برای آن برابر 1.96 بدست می آید. شکل ۴ بیانگر آمارگان محاسبه شده در بین مجموعه یادگیری می باشد که با اعمال آستانه مورد نظر تعداد نمونه های ویژگی از 15154 نقطه به 10481 نقطه کاهش خواهد یافت.



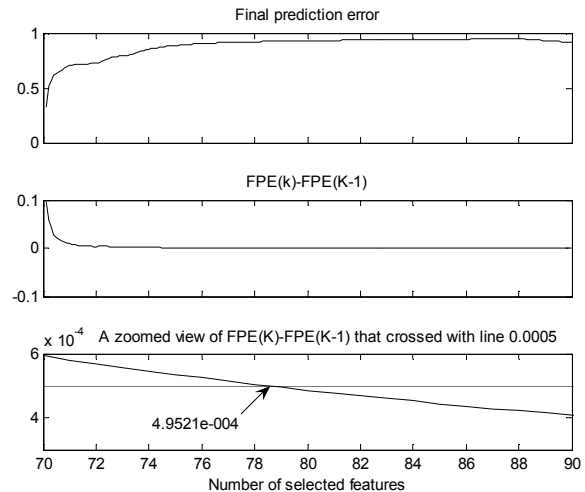
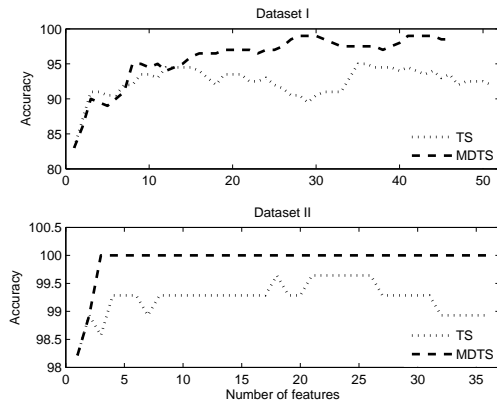
شکل ۴- نمودار آمارگان T محاسبه شده در بین مجموعه یادگیری (مقدار آستانه 1.96)

۵-۲- تعیین تعداد ویژگیها

در گام بعدی، با توجه به مراحل ذکر شده 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 هر بار بطور تصادفی از مجموعه کل نمونه برداری گردید، مقدار مناسب برای

2- Linear discriminate analysis

1- K fold cross validation



شکل ۶- بررسی عملکرد روش پیشنهادی با روش آمارگان T از نظر درصد طبقه‌بندی

۶- بحث

در جدول ۲ مقایسه نتایج حاصل از روش پیشنهادی با نشانگرهای گزارش شده در مراجع [۲۹ و ۱۲] بر روی مجموعه داده I ارایه شده است. مقادیر حاکی از بهبود قابل توجه نتایج روش پیشنهادی در قیاس با روش مبتنی بر شبکه عصبی و الگوریتم ژنتیک، و روش توسعه یافته آماری دارد. در جدول ۳ نتایج روش پیشنهادی با نشانگرهای گزارش شده در مراجع [۳۰ و ۳۱] بر روی مجموعه داده II مقایسه شده است. در این دو مرجع با استفاده از روشهای تحلیل ممیز حداقل مربعات جزئی 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 برخوردار است.

شکل ۵- نمودار مقدار خطای نهایی به همراه مشتق آن، و یک نمای بزرگ شده از شکل وسط در تقاطع با خط 0.0005 .
شکل ۶ نتایج حاصل از طبقه‌بندی با استفاده از نشانگرهای انتخاب شده توسط دو روش در بین مجموعه آزمون از داده‌های موجود را نشان می‌دهد. 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 شده است و دارای واریانس تغییرات زیاد می‌باشد، مشهودتر است.
از بردار ویژگی نهایی با استفاده از روش تحلیل تمایز خطی بترتیب تعداد ۱۶ و ۶ نشانگر حیاتی برای مجموعه داده I و II انتخاب شد که جهت بررسی کارایی و قدرت تمایز این نشانگرها، از معیارهای دقت تشخیص، قطعیت، و حساسیت استفاده شد که معیارهای فوق بطور معمول در بررسی نتایج کارهای غربالگری^۱ بکار می‌رود.

1- Screening task

الگوریتم ترکیبی داده های طیف جرمی لیزری سرطان تخمدان

نحوه تولید و توزیع نقاط ویژگی نیز به جوابهای قابل قبول می‌رسد. این نتیجه مهم، امیدواری به دستیابی نشانگرهای جدید با قدرت تمایز بالا و قابلیت تکرارپذیری درون گروهی و بین گروهی را افزایش می‌دهد که یکی از مشکلات اساسی در حوزه تحلیل محتوای پروتئینی سیگنال طیف جرمی لیزری می‌باشد [۳۲].

۷- نتیجه گیری

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 کاهش بعد ۱:۹۴۷ و ۱:۲۵۲۶ بترتیب در بین مجموعه داده‌های 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، مقادیر دقت تشخیص 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 سویی دیگر، داده‌های مرتبط با تشخیص بیماری دارای ابعاد بالا می‌باشد که انتخاب اولیه این نقاط با استفاده از روشهای

جدول ۲- مقایسه نتایج حاصل از روش پیشنهادی با مراجع ۱۲ و ۲۹ با استفاده از طبقه‌بند LDA باتوجه به تعداد نشانگرهای حیاتی و قدرت تفکیک‌پذیری در مجموعه داده I

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

جدول ۳- مقایسه نتایج حاصل از روش پیشنهادی با مراجع ۳۰ و ۳۱ با استفاده از طبقه‌بند LDA در مجموعه داده II برحسب تعداد نشانگرهای حیاتی و قدرت جداپذیری کلاسهای الگو

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

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

در این مقاله نشان داده شده است که با استفاده از روش ترکیبی و اعمال پیش پردازش مناسب روی داده خام ورودی می‌توان کاهش بعد 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 کارایی روش پیشنهادی نسبت به روشهای دیگر زمانی ملموس تر خواهد بود که در تحلیل مجموعه داده‌های ضعیف از نظر

آستانه ($P\text{-value} < 0.05$) 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.

آستانه‌دهی به دلیل در دست نبودن یک معیار مناسب برای تعیین آستانه، خالی از اشکال نخواهد بود. نتایج بدست آمده بر این نکته تاکید دارد که بکارگیری روش‌های ترکیبی در استخراج و انتخاب ویژگی از فضاها 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

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, Velasco 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 transform 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, Lazzarini 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: bayesian 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.