

انتخاب ژن و طبقه‌بندی سلول‌های سرطانی بر پایه داده‌های ریزآرایه با استفاده از الگوریتم

ترکیبی BLDA و BPSO

مهسا جروفی^۱، موسی شمسی^۲، حمیدرضا صابرکاری^۳، محمدحسین صدیقی^۴ و علی مومن‌نژاد^۵

۱- کارشناس ارشد مهندسی برق، دانشکده مهندسی برق-دانشگاه صنعتی سهند-تبریز-ایران

m_joroughi@sut.ac.ir

۲- دانشیار گروه برق، دانشکده مهندسی برق-دانشگاه صنعتی سهند-تبریز-ایران

shamsi@sut.ac.ir

۳- کارشناس ارشد مهندسی برق، دانشکده مهندسی برق-دانشگاه صنعتی سهند-تبریز-ایران

h_saberkari@sut.ac.ir

۴- استاد گروه برق، دانشکده مهندسی برق-دانشگاه صنعتی سهند-تبریز-ایران

sedaaghi@sut.ac.ir

۵- کارشناس ارشد مهندسی برق، دانشکده مهندسی برق-دانشگاه صنعتی سهند-تبریز-ایران

a_mommenzhad@sut.ac.ir

چکیده: داده‌های ریزآرایه در تشخیص و طبقه‌بندی انواع بافت‌های سرطانی نقش بسزایی دارند. در پژوهش‌های سرطان همیشه تعداد نسبتاً کم نمونه‌ها در ریزآرایه باعث ایجاد مشکلاتی در طراحی طبقه‌بندها شده است. بنابراین، داده‌های ریزآرایه قبل از طبقه‌بندی از طریق تکنیک‌های انتخاب ژن پیش‌پردازش و ژن‌های فاقد اطلاعات آن‌ها دور ریخته می‌شود. اساساً یک روش انتخاب ژن مناسب می‌تواند به‌طور مؤثر کارایی دسته‌بندی بیماری‌ها (سرطان) را بهبود بخشد. در این مقاله، روش جدیدی بر پایه مدل ترکیبی بهینه‌سازی ازدحام ذرات باینری (BPSO) و آنالیز تفکیک‌کننده خطی بیس (BLDA) برای طبقه‌بندی داده‌های ریزآرایه با ابعاد بالا ارائه شده است. ابتدا موقعیت هر ذره به‌صورت بردار باینری و به‌صورت تصادفی نمایش داده می‌شود؛ به‌طوری‌که هر بیت نشان‌دهنده یک ژن است. بیت صفر نشان‌دهنده این است که ویژگی (ژن) متناظر با آن انتخاب نشده و بیت یک نشان‌دهنده این است که ژن متناظر با آن انتخاب شده است. لذا موقعیت هر ذره بیانگر یک مجموعه ژن بوده و میزان تناسب هر ذره توسط الگوریتم طبقه‌بندی آنالیز تفکیک‌کننده خطی بیس برای ارزیابی کیفیت مجموعه ژن انتخاب شده توسط آن ذره محاسبه می‌شود. الگوریتم پیشنهادی بر روی چهار مجموعه از پایگاه داده سرطان اعمال و نتایج آن با سایر روش‌های موجود مقایسه شده است. نتایج پیاده‌سازی نشان می‌دهد که الگوریتم پیشنهادی از صحت و اعتبار بالایی در مقایسه با سایر روش‌های موجود برخوردار بوده، قادر است مجموعه کوچکی از ژن‌های حاوی اطلاعات را به‌گونه‌ای انتخاب کند که صحت طبقه‌بندی افزایش یابد.

واژه‌های کلیدی: آنالیز تفکیک‌کننده خطی بیس، انتخاب ویژگی، بهینه‌سازی ازدحام ذرات باینری، بیان ژن، ریزآرایه، طبقه‌بندی

فناوری ریزآرایه^۱ در سال ۱۹۹۶ متولد و با عناوین

۱- مقدمه

آرایه‌های DNA^۲، تراشه‌های ژنی، تراشه‌های DNA و

تراشه‌های زیستی نامگذاری شده است. فناوری ریزآرایه

یکی از آخرین پیشرفت‌ها در زمینه زیست‌شناسی ملکولی

است که اجازه نظارت بر بیان هزاران ژن را به‌صورت

همزمان تنها در یک آزمایش هیبریداسیون می‌دهد. علاوه بر

^۱ تاریخ ارسال مقاله: ۱۳۹۱/۱۲/۲۵

تاریخ پذیرش مقاله: ۱۳۹۲/۱۲/۱۲

نام نویسنده مسؤول: حمیدرضا صابرکاری

نشانی نویسنده مسؤول: ایران - تبریز - دانشگاه صنعتی سهند -

دانشکده برق

در ریزآرایه cDNA نمونه‌های نشاندار شده با هم مخلوط می‌شوند و سپس با مولکول‌های DNA سطح اسلاید هیبرید می‌شوند. هر چه جفت بازهای بیشتری با هم مکمل شوند، پیوند هیدروژنی قویتری تشکیل می‌شود. بعد از هیبرید شدن نمونه‌های نشاندار شده با مولکول‌های DNA سطح اسلاید، اسلاید شسته می‌شود. بر اثر شستشو پیوندهای ضعیفتر از بین می‌روند و پیوندهای قویتر باقی می‌مانند. میزان شدت و قدرت سیگنال نهایی وابسته به میزان نمونه‌هایی است که با توالی‌های روی سطح، اتصال قوی برقرار کرده‌اند.

داده‌های ریزآرایه به صورت ماتریسی از هزاران ستون و چند صد سطر هستند که هر سطر نشان‌دهنده یک نمونه و هر ستون نیز نشان‌دهنده یک ژن است. ابعاد بالای ویژگی‌ها و تعداد نسبتاً کم نمونه‌ها باعث ایجاد مشکلاتی در آنالیز داده‌های ریزآرایه شده است. این مشکلات عبارتند از [۴ و ۵]:

- افزایش هزینه محاسباتی و پیچیدگی طبقه‌بندها؛
 - کاهش توانایی تعمیم طبقه‌بندها و کاهش اعتبار آن‌ها در پیش‌بینی نمونه‌های جدید؛
 - به علت بالا بودن تعداد ویژگی‌ها نسبت به نمونه‌ها، احتمال آنکه ژن‌های نامربوط خود را در هنگام یافتن ژن‌های با بیان مختلف و در ساختن مدل‌های پیش‌بینی‌کننده نشان دهند، بسیار زیاد است و
 - تفسیر ژن‌های مسبب بیماری مشکل است؛ زیرا از دیدگاه بیولوژیکی تنها مجموعه کوچکی از ژن‌ها مربوط به بیماری هستند. در نتیجه، داده‌های مربوط به اکثریت ژن‌ها در واقع نقش یک پس‌زمینه نویزی را دارند که می‌تواند اثر آن زیرمجموعه کوچک را محو کند. بنابراین، تمرکز بر روی مجموعه کوچکتری از داده‌های بیان ژن، باعث تفسیر بهتر نقش ژن‌های حاوی اطلاعات می‌شوند.
- از این رو، اولین قدم مهم در آنالیز داده‌های ریزآرایه کاهش تعداد ژن‌ها یا به عبارتی، انتخاب ژن‌های متمایزکننده به منظور طبقه‌بندی است. در این مقاله یک الگوریتم بهینه بر

پتانسیل علمی این فناوری در مطالعه بنیادین بیان ژن؛ یعنی تنظیم و تعاملات ژن‌ها، کاربردهای مهمی در پژوهش‌های دارویی و کلینیکی دارد. برای مثال، با مقایسه بیان ژن در سلول‌های سالم و ناسالم، ریزآرایه می‌تواند در شناسایی ژن‌های ناسالم برای داروهای درمانی یا ارزیابی تاثیر آن‌ها استفاده شود [۱].

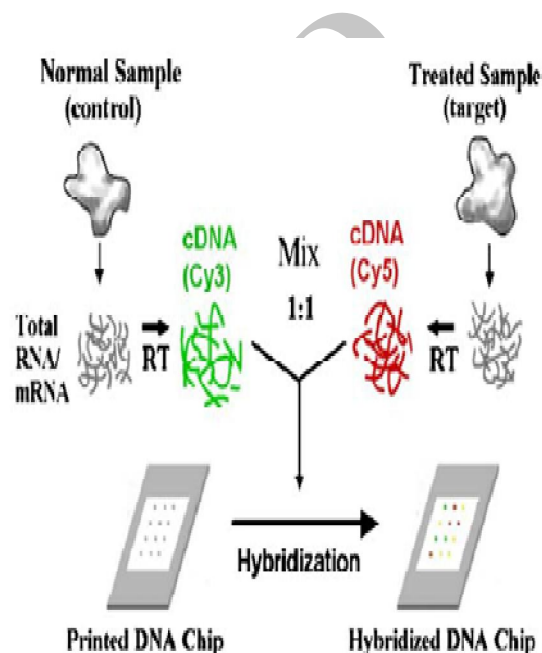
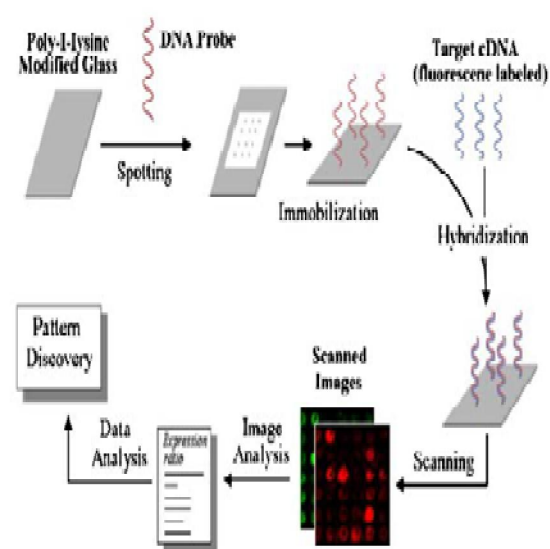
ریزآرایه دارای هزاران نقطه^۳ بوده، هریک از این نقاط حاوی دنباله‌های مختلف شناخته شده DNA، به نام نشانگر^۴ هستند. این نقاط روی یک اسلاید شیشه‌ای توسط یک arrayer رباتیک چاپ می‌شوند. دو نوع ریزآرایه بیشترین کاربرد را دارند: آرایه‌های بر پایه DNA مکمل^۵ و آرایه الیگونوکلوئید که به اختصار الیگو نامیده می‌شود [۱]. در حالت کلی، دو تفاوت عمده بین ریزآرایه cDNA و ریزآرایه الیگونوکلوئید وجود دارد: اول اینکه در ریزآرایه cDNA طول قطعه DNA بیشتر از طول قطعه DNA در ریزآرایه الیگونوکلوئید است؛ دوم اینکه در آزمایش‌های ریزآرایه cDNA دو نمونه RNA، نمونه کنترل و نمونه تجربی، با دو فلورسنت مختلف (Cy3 و Cy5) علامت گذاری می‌شوند [۱-۳]. هدف از ریزآرایه cDNA مقایسه بیان ژن در دو نمونه مختلف است. شکل ۱ مراحل به دست آوردن داده‌های ریزآرایه را نشان می‌دهد. با توجه به این شکل می‌توان مراحل به دست آوردن ریزآرایه DNA را به صورت زیر بیان کرد:

- نمونه‌گیری از نمونه‌های مختلف (برای مثال نمونه‌های سرطانی و سالم)؛
- جداسازی mRNA از نمونه‌ها و انجام رونویسی معکوس و تهیه cDNA؛
- نشاندار کردن نمونه‌های cDNA با رنگ‌های فلورسنت سبز و قرمز (Cy3 و Cy5)؛
- ریختن نمونه‌ها بر روی سطح اسلاید که از قبل توسط توالی‌های ژن مورد نظر پوشیده شده است؛
- شستن سطح اسلاید و
- اسکن کردن آرایه هیبرید شده.

مختلف آن توضیح داده می‌شود. در بخش ۴ مدل ترکیبی پیشنهاد شده بر پایه الگوریتم ترکیبی بهینه‌سازی ازدحام ذرات باینری و آنالیز تفکیک‌کننده خطی بیز معرفی و مراحل مختلف آن به تفصیل بیان می‌شود. نتایج پیاده‌سازی بر روی چهار پایگاه داده در بخش ۵ مطرح شده و نهایتاً بخش ۶ شامل نتیجه‌گیری و جمع‌بندی است.

مبنای مدل ترکیبی بهینه‌سازی ازدحام ذرات باینری^۶ و آنالیز تفکیک‌کننده خطی بیز^۷ برای این منظور ارائه شده است. استفاده از این الگوریتم به کاهش نویز موجود در داده‌های ریزآرایه و همچنین، افزایش صحت طبقه‌بندی آن‌ها منجر می‌شود. در ادامه، در بخش ۲ پایگاه‌های داده ریزآرایه استفاده شده معرفی می‌شود. در بخش ۳، فرآیند کلی استخراج ویژگی و طبقه‌بندی داده‌های ریزآرایه مطرح و مراحل

Flow chart of cDNA microarray technique



شکل (۱): مراحل مختلف به دست آوردن داده‌های ریزآرایه [۱]

مربوط به ALL و ۱۴ مورد مربوط به AML) که در فرآیند تست استفاده شده و ۳۸ نمونه سرطانی (۲۷ مورد مربوط به ALL و ۱۱ مورد مربوط به AML) استفاده شده در فرآیند آموزش، تقسیم می‌شوند.

سرطان پستان: این پایگاه داده شامل ۹۷ نمونه از آزمایش‌های ریزآرایه با ۲۴۴۸۱ سطوح بیان ژن است. داده‌ها به دو دسته ۱۹ نمونه کنترل (۱۲ مورد مربوط به نمونه‌های عود کرده^{۱۱} و ۷ مورد مربوط به نمونه‌های عود نکرده^{۱۱}) که در فرآیند تست استفاده شده و ۷۸ نمونه (۳۴ مورد مربوط به نمونه‌های عود کرده و ۴۴ مورد مربوط به

۲- پایگاه‌های داده

در این مقاله از چهار پایگاه داده ریزآرایه استفاده شده که در ادامه به توضیح آن‌ها می‌پردازیم. شایان ذکر است که تمامی نمونه‌ها با به‌کارگیری آرایه‌های الیگو نوکلئوتیدی با چگالی بالا اندازه‌گیری شده‌اند [۶]. داده‌های استفاده شده در این مقاله از مرجع [۷] استخراج شده است.

سرطان خون: این پایگاه داده شامل ۷۲ نمونه از آزمایش‌های ریزآرایه با ۷۱۲۹ سطوح بیان ژن است. مساله اصلی در آن جداسازی دو نوع از داده‌های سرطان خون، ALL^۸ و AML^۹ است. داده‌ها به دو دسته ۳۴ نمونه کنترل (۲۰ مورد

نمونه‌های عود نکرده) استفاده شده در فرآیند آموزش، تقسیم می‌شوند.

سرطان ریه: این پایگاه داده شامل ۱۸۱ نمونه از آزمایش های ریزآرایه با ۱۲۵۳۳ سطوح بیان ژن است. داده‌ها به دو دسته ۱۴۹ نمونه کنترل (۱۵ مورد مربوط به نمونه‌های MPM^۱ و ۱۳۴ مورد مربوط به نمونه‌های ADCA^{۱۳} که در فرآیند تست استفاده شده و ۳۲ نمونه (۱۶ مورد مربوط به نمونه‌های MPM و ۱۶ مورد مربوط به نمونه‌های ADCA استفاده شده در فرآیند آموزش، تقسیم می‌شوند.

سرطان پروستات: این پایگاه داده شامل ۱۳۶ نمونه از آزمایش های ریزآرایه با ۱۲۶۰۰ سطوح بیان ژن است. داده‌ها به دو دسته ۳۴ نمونه کنترل (۲۵ مورد مربوط به نمونه‌های تومور و ۹ مورد مربوط به نمونه‌های بدون تومور) که در فرآیند تست استفاده شده و ۱۰۲ نمونه (۵۲ مورد مربوط به نمونه‌های تومور و ۵۰ مورد مربوط به نمونه‌های نرمال) استفاده شده در فرآیند آموزش، تقسیم می‌شوند.

۳- فرآیند کلی استخراج ویژگی و طبقه‌بندی داده‌های ریزآرایه

یکی از مسائل مهم در تحلیل داده‌های ریزآرایه انتخاب ژن است و آن فرآیندی است که تعداد کمی از ژن‌ها قبل از طبقه‌بندی انتخاب می‌شوند. ابعاد بالا، تعداد نسبتاً کم نمونه‌ها و تغییرپذیری ذاتی در فرآیندهای آزمایشگاهی و بیولوژیکی باعث ایجاد مشکلاتی در آنالیز داده‌های ریزآرایه شده است. از این‌رو، اولین گام مهم در آنالیز داده‌های ریزآرایه کاهش تعداد ژن‌ها یا به عبارتی انتخاب ژن‌های متمایزکننده به‌منظور طبقه‌بندی است. این مرحله انتخاب ژن نامیده می‌شود. شکل ۲ مراحل کلی استخراج ویژگی و طبقه‌بندی داده‌های ریزآرایه را نشان می‌دهد. این مراحل در

حالت کلی به‌صورت زیر هستند که در ادامه با جزئیات کامل مطرح می‌شوند:

- پیش‌پردازش داده‌های بیان ژن؛
- انتخاب یک مجموعه ژن‌های حاوی اطلاعات؛
- طبقه‌بندی داده‌ها و
- ارزیابی و اعتبارسنجی نتایج حاصل شده.

۳-۱- پیش‌پردازش داده‌های بیان ژن

پیش از اعمال الگوریتم‌های استخراج ویژگی و طبقه‌بندی نیاز است که داده مورد استفاده پیش‌پردازش شود. داده‌های مورد استفاده در این مقاله با به‌کارگیری نرم‌افزار Weka فراخوانی شده که دارای فرمت ARFF^{۱۴} هستند. به-علت تغییرات زیاد داده، گسسته‌سازی مقادیر آن برای دستیابی به صحت طبقه‌بندی مناسب ضروری است. گسسته‌سازی فرآیند تبدیل ویژگی‌ها و متغیرهای پیوسته به مقادیر و یا ویژگی‌های گسسته است. معمولاً طی این فرآیند، داده‌ها به k بخش با طول یکسان (بازه‌های یکسان) و یا $k\%$ از کل داده (فرکانس‌های یکسان) گسسته می‌شوند.

۳-۲- انتخاب مجموعه ژن‌های حاوی اطلاعات برای طبقه‌بندی داده‌های ریزآرایه

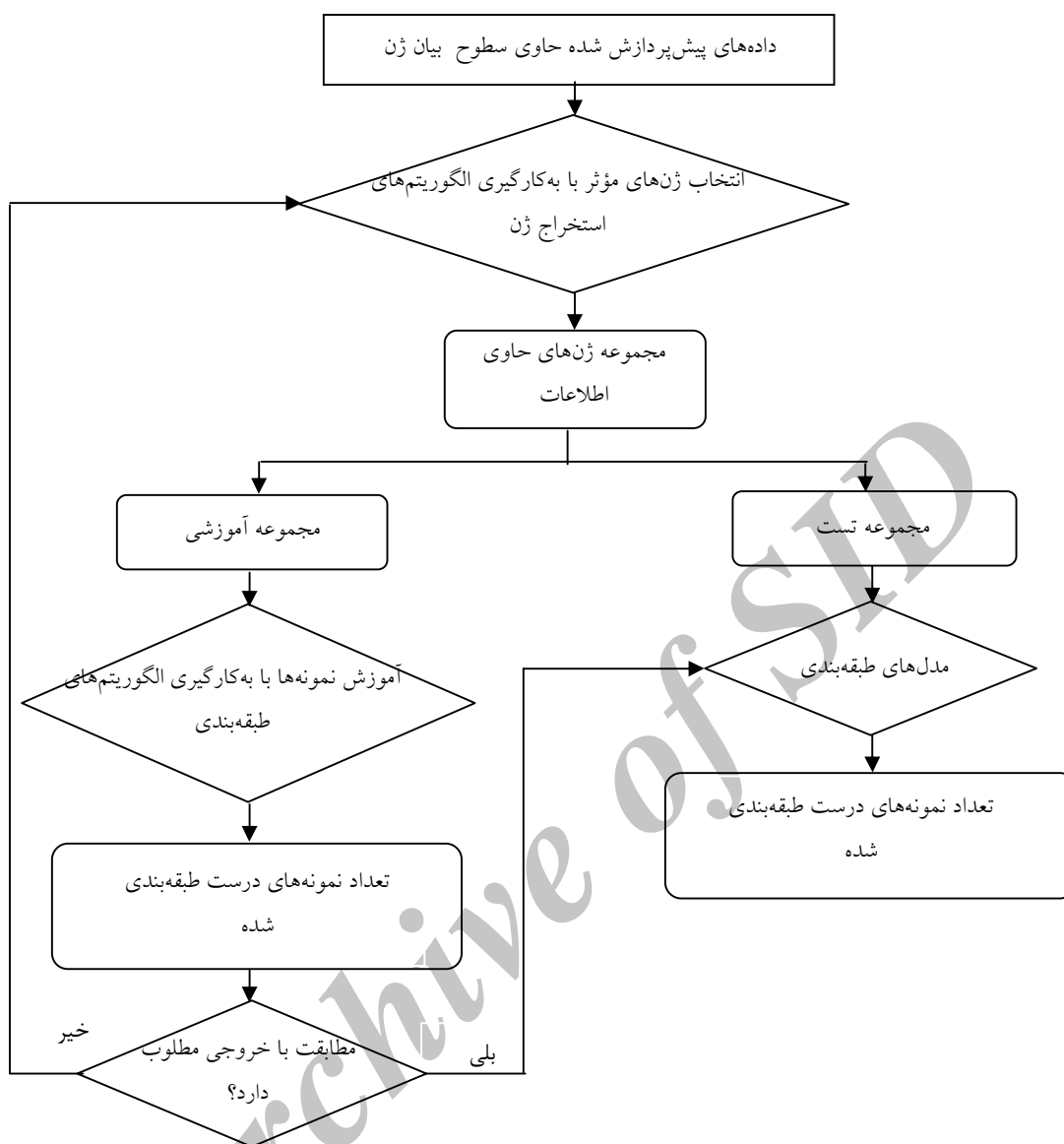
پژوهش‌ها نشان داده که تشخیص دقیق سرطان می‌تواند با طبقه‌بندی داده‌های ریزآرایه عملی شود. با وجود این، مشکل اصلی در تحلیل داده‌های ریزآرایه بعد بالای آن‌هاست که در نتیجه تعداد بسیار زیاد متغیرها (ژن‌ها) در مقابل تعداد کم نمونه‌ها ایجاد می‌شود. اگرچه تعداد بسیار زیادی از ژن در داده‌های ریزآرایه وجود دارند، تنها بخش اندکی از آن‌ها تاثیر بسزایی در صحت طبقه‌بندی می‌گذارند. بسیاری از ژن‌ها عملکرد مشابهی در دو حالت نرمال و

شده است. در [۱۱] الگوریتم‌های ژنتیک موازی با به‌کارگیری عملگرهای تطبیقی گسترش یافته است. همچنین در [۱۲] از مدل ترکیبی الگوریتم ژنتیک و ماشین بردار پشتیبان برای انتخاب یک مجموعه از ژن‌ها و طبقه‌بندی آن‌ها استفاده شده است. در [۱۳] نیز مساله انتخاب و طبقه‌بندی ژن به‌صورت یک مساله بهینه‌سازی چند مرحله‌ای مطرح شده که در آن به‌طور همزمان تعداد ویژگی‌ها (ژن‌ها) و همچنین تعداد نمونه‌های اشتباه طبقه‌بندی شده کاهش داده می‌شود.

نهایتاً در مدل‌های ترکیبی، فرآیند انتخاب یک مجموعه از ژن‌های مؤثر در حین فرآیند آموزش توسط یک طبقه‌بند خاص صورت می‌گیرد. یک نمونه از این روش، استفاده از ماشین بردار پشتیبان به‌همراه حذف ویژگی‌های بازگشتی است. ایده این روش، حذف یک به یک ژن‌ها و بررسی اثر حذف شدن آن‌ها در خطای مورد انتظار است [۱۴]. الگوریتم حذف ویژگی‌های بازگشتی یک روش رتبه‌بندی ویژگی‌ها رو به عقب است؛ به عبارتی آن دسته از ژن‌هایی که در آخرین مرحله حذف می‌شوند، بهترین نتیجه طبقه‌بندی را به‌دست می‌دهند، در حالی که این ژن‌ها ممکن است به-تنهایی همبستگی خوبی با کلاس‌ها نداشته باشند. مدل‌های ترکیبی را می‌توان حالت تعمیم یافته مدل wrapper در نظر گرفت. دو نمونه دیگر از مدل ترکیبی در [۱۵] و [۱۶] اشاره شده است.

غیرنرمال (سرطانی) دارند. همچنین، برخی از ژن‌ها به‌صورت نویز در داده‌ها ظاهر می‌شوند. وجود ژن‌های حاوی نویز در بروز سرطان نقشی نداشته، اثر منفی در صحت طبقه‌بندی می‌گذارند. بنابراین، داده‌های ریزآرایه قبل از طبقه‌بندی از طریق تکنیک‌های انتخاب ژن پیش‌پردازش و ژن‌های فاقد اطلاعات آن‌ها دور ریخته می‌شود. انجام این کار باعث افزایش بازدهی طبقه‌بندی‌کننده و همچنین، کاهش پیچیدگی محاسباتی خواهد شد [۸].

به‌طور کلی، می‌توان گفت سه مدل انتخاب ویژگی (ژن) وجود دارد [۹]. مدل اول مدل فیلتر است که عمل انتخاب ویژگی و طبقه‌بندی را در دو مرحله جداگانه انجام می‌دهد. این مدل ژن‌هایی را به عنوان ژن‌های مؤثر انتخاب می‌کند که دارای توانایی تفکیک‌کنندگی بالایی باشند. این مدل مستقل از طبقه‌بندی یا الگوریتم یادگیری است و از لحاظ محاسبات ساده و سریع است. مدل دوم مدل wrapper است که انتخاب ویژگی و طبقه‌بندی را در یک فرآیند انجام می‌دهد. این مدل در حین فرآیند جست‌وجوی ژن‌های مؤثر، از طبقه‌بند استفاده می‌کند. به عبارتی، مدل wrapper از یک الگوریتم یادگیری برای تست زیر مجموعه ژن انتخاب شده استفاده می‌کند. صحت مدل wrapper نسبت به مدل فیلتر بیشتر است. روش‌های مختلفی برای انتخاب زیرمجموعه‌های مناسب بر مبنای مدل wrapper در مقالات ارائه شده است. در [۱۰] از الگوریتم‌های تکاملی همراه با طبقه‌بند K نزدیکترین همسایگی برای این منظور استفاده



شکل (۲): بلوک دیاگرام مراحل مختلف تحلیل داده‌های ریزآرایه

باقیمانده آن‌ها به‌عنوان نمونه‌های آموزشی در نظر گرفته می‌شود. این روند ۱۰ بار با به‌کارگیری داده‌های آموزشی و تست مختلف انجام گرفته و نتیجه با گرفتن مقدار متوسط در ۱۰ بار تکرار آزمایش به دست می‌آید.

۴- مدل پیشنهاد شده بر پایه الگوریتم

ترکیبی BLDA و BPSO

مدل پیشنهاد شده بر پایه آنالیز همبستگی پیرسون، الگوریتم بهینه‌سازی ازدحام ذرات باینری و آنالیز تفکیک-

۳-۳- ارزیابی و اعتبار سنجی نتایج

آخرین گام در تحلیل داده‌های ریزآرایه ارزیابی نتایج حاصل از اعمال الگوریتم‌های طبقه‌بندی است. در این مقاله ارزیابی نتایج بر اساس روش اعتبارسنجی k -fold انجام گرفته که در آن k اشاره به تعداد دفعات تکرار داشته و مقدار آن برابر ۱۰ در نظر گرفته شده است. از این‌رو اعتبارسنجی 10 -fold به این ترتیب صورت می‌گیرد که ابتدا نمونه‌ها به ۱۰ بخش تقسیم شده و در هر بار اجرای الگوریتم، ۰/۱ کل داده‌ها به‌عنوان نمونه‌های تست و

از ژن‌ها با استفاده از آنالیز همبستگی پیرسون^{۱۵}

به منظور امتیازدهی به هر ژن، دو نشانگر ویژگی ایده‌آل باینری مطابق رابطه (۱) تعریف می‌شود. اولین نشانگر در کلاس A دارای مقدار یک و در کلاس B دارای مقدار صفر است و دومین نشانگر در کلاس A دارای مقدار صفر و در کلاس B دارای مقدار یک است.

$$Ideal_1 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0) \quad (1)$$

$$Ideal_2 = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$$

اگر ژن‌ها مشابه نشانگرها باشند یا به عبارتی فاصله بین ژن‌ها و نشانگرها کوچک باشد، آنگاه این ژن‌ها به‌عنوان ژن‌های حاوی اطلاعات برای طبقه‌بندی انتخاب می‌شوند. در این مقاله برای محاسبه فاصله هر ژن از نشانگرها از معیار آنالیز همبستگی پیرسون استفاده شده که به‌صورت زیر تعریف می‌شود [۸]:

$$PC = \frac{\sum_{i=1}^n (ideal_i - \mu_{ideal})(g_i - \mu_g)}{\sqrt{\sum_{i=1}^n (ideal_i - \mu_{ideal})^2} \sqrt{\sum_{i=1}^n (g_i - \mu_g)^2}} \quad (2)$$

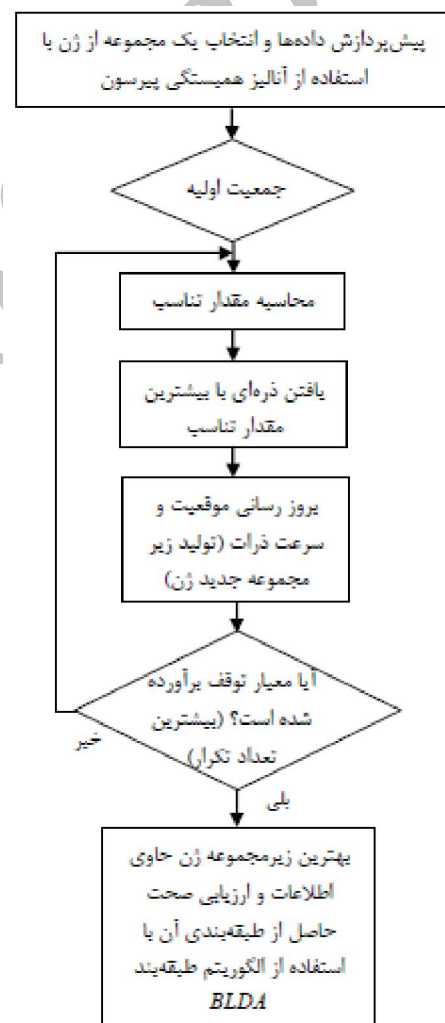
که در آن n تعداد نمونه‌های آموزشی، μ_g میانگین ژن، μ_{ideal} میانگین نشانگر ایده‌آل، g_i i امین مقدار از بردار ژن و $ideal_i$ i امین مقدار باینری از بردار نشانگر ایده‌آل است. به منظور پیش‌پردازش اولیه، معیار PC برای تمامی ژن‌ها محاسبه شده و k ژن که دارای PC کوچکتری هستند به‌عنوان ژن‌های حاوی اطلاعات اولیه انتخاب می‌شوند.

۴-۲- الگوریتم بهینه‌سازی ازدحام ذرات باینری

ایده بهینه‌سازی ازدحام ذرات برای اولین بار توسط کندی و ابرهارت در سال ۱۹۹۵ مطرح شد [۱۷]. PSO یک الگوریتم محاسبه‌ای تکاملی الهام گرفته از طبیعت و براساس تکرار است. منبع الهام این الگوریتم رفتار اجتماعی حیوانات، همانند حرکت دسته جمعی پرندگان و ماهی‌ها بوده است. در این الگوریتم هر پاسخ مساله به‌صورت یک ذره مدل می‌شود. هر ذره برای به‌دست آوردن بهترین پاسخ مساله از تجربه خود و از تجربیات به‌دست آمده از جمعیت استفاده می‌کند. الگوریتم PSO از تعداد مشخصی از ذرات تشکیل می‌شود. برای هر ذره دو مقدار موقعیت و سرعت

کننده خطی بیز است. شکل (۳) بلوک دیاگرام الگوریتم پیشنهادی را نشان می‌دهد. گام‌های اصلی این الگوریتم به صورت زیر هستند که در ادامه با جزئیات کامل مطرح می‌شوند:

- پیش‌پردازش داده‌ها و انتخاب مجموعه‌ای از ژن‌ها با استفاده از آنالیز همبستگی پیرسون؛
- استفاده از الگوریتم ترکیبی BPSO/BLDA برای انتخاب مجموعه ژن‌های حاوی اطلاعات و طبقه‌بندی آن‌ها.



شکل (۳): بلوک دیاگرام الگوریتم پیشنهادی

۴-۱- پیش‌پردازش داده‌ها و انتخاب مجموعه‌ای

است:

$$t = w^T x + n \quad (۳)$$

که در آن t و x به ترتیب بردار هدف و ویژگی، w بردار وزن ($w \in R^D$) و n نویز سفید است. در این صورت تابع همانندی برای وزن‌های w استفاده شده در رگرسیون به صورت زیر بیان می‌شود:

$$p(D | \beta, w) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left(-\frac{\beta}{2} \|X^T w - t\|^2\right) \quad (۴)$$

که در آن $X \in R^{D \times N}$ ماتریس سطری حاوی بردارهای ویژگی، D نشان‌دهنده دو پارامتر $\{X, t\}$ ، β معکوس واریانس و N نشان‌دهنده تعداد نمونه‌ها در مجموعه آموزشی است. به منظور توصیف یک مجموعه بیز باید توزیع پیشین برای بردارهای وزن تعیین شود. این توزیع اطلاعات اولیه‌ای درباره بردار وزن به دست داده و به صورت زیر تعریف می‌شود:

$$p(w | \alpha) = \left(\frac{\alpha_i}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\frac{1}{2} w^T I'(\alpha) w\right) \quad (۵)$$

که در آن α_i نشان‌دهنده معکوس واریانس توزیع اولیه برای بردارهای وزن w_i ، $I'(\alpha)$ یک ماتریس قطری مربعی با ابعاد $D+1$ بوده که D تعداد ویژگی است. با داشتن توزیع پیشین و تابع همانندی می‌توان توزیع پسین را با به کارگیری قانون بیز به صورت زیر به دست آورد:

$$p(w | \beta, \alpha, D) = \frac{p(D | \beta, w) p(w | \alpha)}{\int p(D | \beta, w) p(w | \alpha) dw} \quad (۶)$$

از آنجا که توزیع پیشین و تابع همانندی، گوسی هستند توزیع پسین نیز گوسی خواهد بود. بنابراین، میانگین و کوواریانس این توزیع به صورت زیر است:

$$m = \beta (\beta X X^T + I'(\alpha))^{-1} X t \quad (۷)$$

$$C = (\beta X X^T + I'(\alpha))^{-1}$$

از این رو، توزیع پسین می‌تواند برای محاسبه احتمال توزیع اهداف رگرسیون، \hat{t} ، برای بردار ویژگی جدید \hat{x} استفاده شود. در این صورت، توزیع پیش‌بینی‌کننده به صورت زیر نشان داده می‌شود:

تعریف می‌شود که به ترتیب با یک بردار مکان و یک بردار سرعت مدل می‌شوند. یک حافظه به ذخیره بهترین موقعیت هر ذره در گذشته و یک حافظه نیز به ذخیره بهترین موقعیت پیش آمده در میان همه ذرات اختصاص می‌یابد. با تجربه حاصل از این حافظه‌ها، ذرات تصمیم می‌گیرند که در نوبت بعدی چگونه حرکت کنند. در طول حرکت، هر ذره موقعیت خود را با تغییر سرعت با توجه به بهترین موقعیت خود و بهترین موقعیت پیش آمده در بین همه ذرات تنظیم می‌کند. بنابراین، حرکت هر ذره به سه عامل بستگی دارد: موقعیت فعلی ذره، بهترین موقعیتی که ذره تاکنون داشته است (Pbest) و بهترین موقعیتی که کل مجموعه ذرات تاکنون داشته‌اند (Gbest) [۱۷]. PSO به طور موفقیت آمیزی در زمینه‌های بهینه‌سازی توابع، فرایند آموزش شبکه‌های عصبی، سیستم‌های کنترل فازی و غیره استفاده شده است. کندی و ابرهات در سال ۱۹۹۷ نوع باینری الگوریتم PSO را نیز معرفی کردند که در موارد متغیرهای گسسته باینری استفاده می‌شود. در BPSO موقعیت هر ذره به صورت بردار باینری صفر و یک تعریف می‌شود و سرعت هر ذره به صورت تعداد بیت‌های تغییر یافته در هر تکرار تعبیر می‌شود [۱۸].

۳-۴ الگوریتم آنالیز تفکیک کننده خطی بیز

الگوریتم BLDA یک الگوریتم قابل تنظیم بوده که به منظور جلوگیری از بیش برآزش^{۱۶} در داده‌های با ابعاد بالا به کار می‌رود. با استفاده از این الگوریتم درجه تنظیم می‌تواند به صورت اتوماتیک و به سرعت از طریق داده‌های آموزشی و بدون نیاز به استفاده از اعتبارسنجی تخمین زده شود. این طبقه‌بند برای طبقه‌بندی داده‌های حاوی نویز و همچنین، ویژگی‌هایی که به طور دقیق قابل طبقه‌بندی نیستند استفاده می‌شود [۱۹]. اساس این طبقه‌بند بدین صورت است که در آن رگرسیون در چارچوب بیز صورت می‌گیرد [۲۰]. بدین صورت، هدف‌ها و بردار ویژگی یک رابطه خطی با یکدیگر خواهند داشت. این رابطه به صورت زیر

محاسبه می‌شود. بهترین میزان تناسب هر ذره $pbest$ و بهترین میزان تناسب در میان گروه ذرات $gbest$ نامیده می‌شود. این فرآیند تا زمانی که معیار توقف برآورده نشود تکرار خواهد شد. معیار توقف می‌تواند بیشترین تعداد تکرار و یا میزان تناسب بیشینه تعریف شود [۲۲]. سرعت ذرات در الگوریتم BPSO توسط رابطه زیر به روز رسانی می‌شود:

$$v_{pd}^{new} = w \times v_{pd}^{old} + c_1 \times rand_1 (pbest_{pd} - x_{pd}^{old}) + c_2 \times rand_2 (gbest_d - x_{pd}^{old}) \quad (10)$$

اگر v_{pd}^{new} در بازه $[v_{min}, v_{max}]$ نباشد، آنگاه سرعت ذرات در الگوریتم PSO باینری به صورت زیر تعیین می‌شود:

$$v_{pd}^{new} = \max(\min(v_{max}, v_{pd}^{new}), v_{min}) \quad (11)$$

برای تبدیل بردار سرعت به بردار احتمال از تابع تبدیل سیگموئید به صورت زیر استفاده می‌کنیم:

$$S(v_{pd}^{new}) = \frac{1}{1 + \exp(-v_{pd}^{new})} \quad (12)$$

از این رو، با به کارگیری رابطه (۱۲) موقعیت ذرات نیز به صورت زیر به روز رسانی می‌شود:

$$X_{pd}^{new} = \begin{cases} 1 & ; rand < S(v_{pd}^{new}) \\ 0 & ; otherwise \end{cases} \quad (13)$$

در روابط فوق w وزن اینرسی، d بعد مساله، c_1 و c_2 فاکتورهای شتاب و $rand_1$ ، $rand_2$ و $rand$ اعداد تصادفی در بازه $[0, 1]$ هستند. همچنین v_{pd}^{new} سرعت ذره در مرحله جدید و v_{pd}^{old} سرعت در مرحله قبلی و X_{pd}^{old} موقعیت قبلی ذره است. الگوریتم BPSO به صورت زیر خلاصه می‌شود:

$$p(\hat{t} | \beta, \alpha, \hat{x}, D) = \int p(\hat{t} | \beta, \alpha, w) p(w | \beta, \alpha, D) dw \quad (8)$$

$$\hat{t} = w^T \hat{x}$$

توزیع بیان شده در رابطه (۸) یک توزیع گوسی با میانگین μ و واریانس σ^2 است که برای بردار ورودی جدید \hat{x} ، مقادیر مختلف \hat{t} را تعیین می‌کند. شایان ذکر است که در الگوریتم BLDA هدف‌های رگرسیون برای نمونه‌های کلاس ۱ در $\frac{N}{N_1}$ و برای نمونه‌های کلاس ۲ در $\frac{N}{N_2}$ تنظیم می‌شوند که در آن N تعداد کل نمونه‌های آموزشی، N_1 تعداد نمونه‌های کلاس ۱ و N_2 تعداد نمونه‌های کلاس ۲ است [۲۱]. از این رو، احتمال نمونه‌های کلاس ۱ به صورت زیر به دست می‌آید:

$$p(\hat{y} = 1 | \beta, \alpha, \hat{x}, D) = \frac{p\left(\hat{t} = \frac{N}{N_1} | \beta, \alpha, \hat{x}, D\right)}{p\left(\hat{t} = \frac{N}{N_1} | \beta, \alpha, \hat{x}, D\right) + p\left(\hat{t} = -\frac{N}{N_2} | \beta, \alpha, \hat{x}, D\right)} \quad (9)$$

۴-۴- انتخاب ویژگی و طبقه‌بندی با استفاده از

مدل ترکیبی BPSO/BLDA

در این مقاله از الگوریتم BPSO برای اجرای فرآیند انتخاب ژن و از الگوریتم BLDA به عنوان ارزیابی کننده برای طبقه‌بندی به دست آمده توسط BPSO استفاده شده است. روش کار بدین صورت است که الگوریتم BPSO از ۳۰ ذره و هر ذره نیز از ۷۰ بیت باینری تشکیل شده است. موقعیت هر ذره توسط این ۷۰ بیت باینری تعریف می‌شود. هر بیت نشان‌دهنده یک ژن است؛ به طوری که بیت صفر نشان‌دهنده این است که ویژگی (ژن) متناظر با آن انتخاب نشده و بیت یک نیز نشان‌دهنده این است که ویژگی (ژن) متناظر با آن انتخاب شده است. شکل ۴ نمایشی از ازدحام ذرات را نشان می‌دهد. ذرات بعد از هر بار به روز رسانی ارزیابی می‌شوند. موقعیت هر ذره بیانگر یک مجموعه ژن است؛ لذا میزان تناسب هر ذره توسط طبقه‌بند BLDA برای ارزیابی کیفیت مجموعه ژن انتخاب شده توسط آن ذره

۵- نتایج پیاده‌سازی

کلیه مراحل پیاده‌سازی الگوریتم پیشنهادی بر روی یک کامپیوتر با پردازنده ۳/۴GHz و حافظه RAM برابر ۱GHz انجام گرفته است. مقادیر پارامترهای استفاده شده در الگوریتم BPSO در جدول ۱ نشان داده شده است.

در جدول ۲ نتایج به دست آمده از صحت الگوریتم پیشنهادی با اعمال روش اعتبارسنجی fold-۱۰ و همچنین میانگین صحت در چهار پایگاه داده نشان داده شده است. در این جدول، Acc ($\%$) و Avg (N) به ترتیب نشان‌دهنده صحت در ۱۰ بار اجرای الگوریتم و میانگین تعداد ژن‌های انتخاب شده در هر بار اجرای الگوریتم است. همان‌طور که مشاهده می‌شود، بیشترین صحت طبقه‌بندی در پایگاه داده سرطان خون برابر ۹۲/۳۹ بوده هنگامی که میانگین تعداد ژن‌های انتخاب شده برابر ۴۸ است. کمترین صحت طبقه‌بندی در این پایگاه داده نیز برابر ۸۵/۴۷ است که در این حالت ۶۷ ژن انتخاب شده است. به‌طور مشابه، بیشترین صحت طبقه‌بندی در پایگاه‌های داده ریه، پستان و پروستات به ترتیب ۹۹/۶۶، ۹۴/۶۳ و ۹۶/۶۸ است که میانگین تعداد ژن‌های انتخاب شده در آن‌ها به ترتیب ۳۹، ۴۳ و ۵۰ است. کمترین مقدار صحت طبقه‌بندی در این پایگاه‌های داده نیز به ترتیب برابر ۹۶/۲۶، ۹۰/۷۸ و ۹۴/۰۵ است که در این حالت میانگین تعداد ژن‌های انتخاب شده به ترتیب برابر ۳۸، ۳۸ و ۴۰ است. در شکل ۵ نیز میانگین صحت طبقه‌بندی الگوریتم پیشنهادی در ۱۰ بار اجرای آن در چهار پایگاه داده به‌طور شهودی نشان داده شده است.

الگوریتم BPSO

شروع

تولید ذرات اولیه به صورت تصادفی
تازمانی که معیار توقف برآورده شود، مراحل زیر تکرار می‌شوند:

ارزیابی میزان تناسب ذرات توسط طبقه‌بند $BLDA$
برای $p=1:N$ (تعداد ذرات)

اگر (میزان تناسب x_p) < (میزان تناسب $pbest_p$)
آنگاه

$$pbest_p = x_p$$

پایان

اگر (میزان تناسب یکی از x_p ها) < (میزان

تناسب $gbest$) آنگاه

موقعیت آن ذره = $gbest$

پایان

برای $d=1:D$ (تعداد ابعاد ذرات)

$$v_{pd}^{new} = w \times v_{pd}^{old} + c_1 \times rand_1 (pbest_{pd} - x_{pd}^{old}) + c_2 \times rand_2 (gbest_d - x_{pd}^{old})$$

اگر v_{pd}^{new} در بازه $[v_{min}, v_{max}]$ نباشد آنگاه:

$$v_{pd}^{new} = \max(\min(v_{max}, v_{pd}^{new}), v_{min})$$

استفاده از تابع تبدیل سیگموئید تبدیل بردار سرعت به

بردار احتمال

$$S(v_{pd}^{new}) = \frac{1}{1 + \exp(-v_{pd}^{new})}$$

اگر $(rand < S(v_{pd}^{new}))$ آنگاه $x_{pd}^{new} = 1$ در

غیراینصورت $x_{pd}^{new} = 0$

تکرار مراحل فوق تا برآورده شدن معیار توقف.

پایان

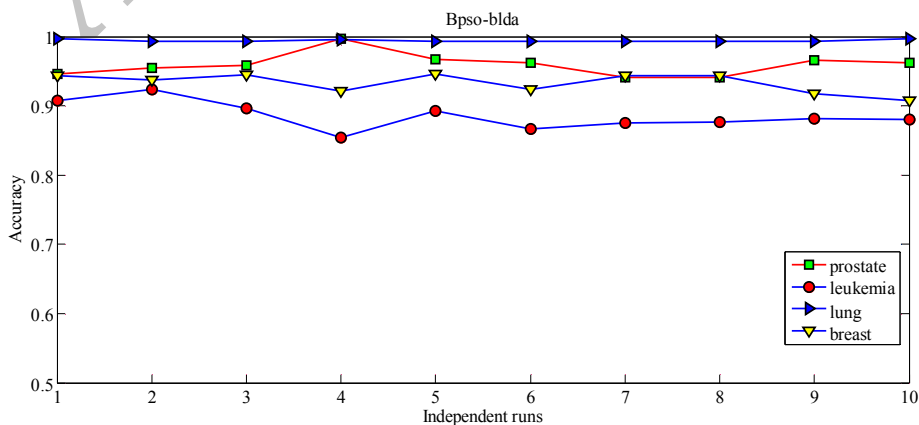
شکل (۴): نمایش ازدحام ذرات در الگوریتم BPSO

جدول (۱): مقادیر پارامترهای به کار رفته در الگوریتم BPSO

مقادیر پارامترهای استفاده شده در الگوریتم PSO باینری	
۳۰	m: تعداد ذرات
۲	C ₁ , C ₂ : فاکتورهای شتاب
۶	V _{max} : بیشینه سرعت ذرات
-۶	V _{min} : کمینه سرعت ذرات
[۰ و ۱]	R ₁ , R ₂ : اعداد تصادفی
۷۰	N: طول هر ذره (تعداد بیت)
۵۰	Max iter: بیشینه تکرار
$w = w_{max} - (((w_{max} - w_{min}) \times iter) / Maxiter)$	w: وزن اینرسی
۰/۹۹۵	W _{max} : بیشینه وزن اینرسی
۰/۵	W _{min} : کمینه وزن اینرسی

جدول (۲): مقادیر کمی صحت به دست آمده از اعمال الگوریتم پیشنهادی در ۱۰ بار اجرای الگوریتم بر روی چهار پایگاه داده

پایگاه‌های داده								تعداد اجرای الگوریتم
سرطان پروستات		سرطان پستان		سرطان ریه		سرطان خون		
Avg	ACC (%)	Avg	ACC (%)	Avg	ACC (%)	Avg	ACC (%)	
۵۴	۹۴/۵۶	۴۳	۹۴/۴۱	۴۰	۹۹/۶۳	۴۲	۹۰/۵۷	۱
۵۲	۹۵/۵۲	۲۴	۹۳/۷۵	۴۰	۹۹/۳۳	۴۸	۹۲/۳۹	۲
۴۵	۹۵/۸۳	۴۱	۹۴/۵۲	۵۰	۹۹/۲۶	۴۸	۸۹/۶۴	۳
۵۰	۹۹/۶۳	۳۴	۹۲/۱۵	۳۹	۹۹/۵۶	۶۷	۸۵/۴۷	۴
۵۰	۹۶/۶۸	۴۳	۹۴/۶۳	۳۸	۹۶/۲۶	۴۷	۸۹/۲۱	۵
۴۰	۹۶/۱۹	۲۹	۹۲/۳۲	۴۰	۹۹/۳۳	۴۵	۸۶/۶۴	۶
۴۰	۹۴/۰۵	۴۸	۹۴/۴۱	۴۰	۹۹/۳۳	۳۹	۸۷/۵۶	۷
۴۰	۹۴/۱۲	۴۲	۹۴/۴۱	۵۳	۹۹/۲۸	۳۷	۸۷/۶۲	۸
۵۷	۹۶/۵۶	۳۴	۹۱/۷۷	۴۰	۹۹/۳۷	۶۵	۸۸/۱۱	۹
۴۵	۹۶/۲۵	۳۸	۹۰/۷۸	۳۹	۹۹/۶۶	۶۲	۸۷/۹۹	۱۰
۴۷	۹۵/۹۳	۳۸	۹۳/۳۱	۴۲	۹۹/۴۳	۵۰	۸۸/۵۳	میانگین



شکل (۵): ترسیم منحنی میانگین صحت طبقه‌بندی الگوریتم پیشنهادی در ۱۰ بار اجرای آن بر روی چهار پایگاه داده

ژنتیک روش‌های جستجوی مبتنی بر جمعیت هستند که با اشتراک گذاری اطلاعات میان اعضای خود و با استفاده از قوانین قطعی و احتمالی، فرآیند جستجو را بهبود می‌بخشند. با این حال BPSO اپراتورهای ژنتیک مانند اپراتورهای تقاطع و جهش را ندارد. البته، مدل اجتماعی تعامل بین ذرات را می‌توان شبیه اپراتور تقاطع در نظر گرفت. برای مثال، پارامترهای rand1 و rand2 (رابطه ۱۰) که بر سرعت ذرات اثر می‌گذارند، شبیه پارامتر جهش در الگوریتم ژنتیک هستند. در حقیقت، تنها تفاوت بین آن‌ها این است که اپراتورهای تقاطع و جهش در الگوریتم ژنتیک احتمالاتی است، در حالی که ذرات جدید در BPSO باید در هر تکرار بدون هیچ احتمالی پردازش شوند. در مقایسه با الگوریتم ژنتیک، مکانیسم به اشتراک گذاری اطلاعات در BPSO به‌طور قابل توجهی متفاوت است. در الگوریتم ژنتیک سیر تکاملی با استفاده از اپراتورهای تقاطع و جهش انجام می‌شود. کروموزوم‌ها اطلاعات را بین یکدیگر به اشتراک گذاشته، کل جمعیت همانند یک گروه به سمت منطقه هدف حرکت می‌کنند. در فضای مساله، این مدل شبیه به جستجوی فقط یک ناحیه است. بنابراین، نقطه ضعف این مدل این است که به راحتی می‌تواند در بهینه‌های محلی گیر کند، اما در BPSO هر ذره در فضای مساله به‌صورت یکنواخت توزیع شده و فقط gbest اطلاعات را برای سایر ذرات فراهم می‌کند. این یک مکانیسم به اشتراک‌گذاری یک‌طرفه است و سیر تکاملی فقط در جهت بهترین راه حل است. عملکرد BPSO تحت تاثیر پارامترهای w و فاکتورهای شتاب $c1$ و $c2$ است. با مقداردهی مناسب این پارامترها (جدول ۱) به راحتی می‌توان به نتایج مطلوب دست یافت. اگر مقادیر این پارامترها خیلی کوچک انتخاب شود، حرکت ذرات نیز خیلی آهسته و زمان‌بر خواهد بود و اگر مقادیر پارامترها بزرگ انتخاب شود، الگوریتم تضعیف شده و مجموعه ویژگی‌های مفید به دست نمی‌آید. بنابراین، تنظیم مناسب پارامترها، الگوریتم BPSO را به راحتی قادر به انتخاب ویژگی‌های مهم می‌کند.

در مورد روش انتخاب طبقه‌بندی نیز می‌توان این‌گونه توضیح داد که اگرچه SVM در طبقه‌بندی تومور در اغلب موارد به خوبی عمل می‌کند، ولی پیچیدگی محاسباتی بالایی

در جدول ۳ مقایسه الگوریتم پیشنهادی با سایر روش‌های موجود در مراجع از نظر صحت طبقه‌بندی در پایگاه داده سرطان خون نشان داده شده است. همان‌طور که مشاهده می‌شود، میزان بهبود صحت طبقه‌بندی در الگوریتم پیشنهادی نسبت به روش‌های PSO+ANN [۲۳]، Nero-Fuzzy [۲۴] و KNN [۲۵] به ترتیب به میزان ۲/۸، ۱/۲ و ۲۱/۹ درصد است. صحت الگوریتم پیشنهادی تنها از الگوریتم طبقه‌بند Bayesian [۲۶] پایین‌تر است. به‌طور مشابه، عملکرد الگوریتم پیشنهادی با سایر روش‌ها در پایگاه داده سرطان ریه مقایسه شده و در جدول ۴ نشان داده شده است. بر اساس این جدول، صحت طبقه‌بندی در الگوریتم پیشنهادی نسبت به روش‌های PSO+SVM [۲۷]، IPso+KNN [۲۸] و Bayesian به ترتیب به میزان ۰/۴، ۲/۹، ۱/۱ و ۱۱/۷ درصد بهبود می‌یابد. صحت طبقه‌بندی الگوریتم پیشنهادی در این پایگاه داده تنها از روش ترکیبی PSO+ Ensemble NN [۲۳] کمتر است. عملکرد الگوریتم پیشنهادی با روش‌های PSO+SVM، GA+SVM [۲۷] و Bayesian در پایگاه داده سرطان پستان نیز مقایسه شده که در جدول ۵ نشان داده شده است. برتری الگوریتم پیشنهادی نسبت به دو روش PSO+SVM و Bayesian به‌وضوح در این جدول مشاهده می‌شود؛ به‌طوری‌که میزان بهبود در الگوریتم پیشنهادی به ترتیب برابر ۹/۴ و ۲۶/۱ درصد است. نهایتاً الگوریتم پیشنهادی با روش‌های IPso+KNN، الگوریتم ترکیبی PSO/GA+SVM [۲۹] و KNN+EA [۳۰] اعمال شده

بر پایگاه داده سرطان پروستات مقایسه و در جدول ۶ نشان داده شده است. برتری الگوریتم پیشنهادی نسبت به همه روش‌های مطرح شده در این جدول مشاهده می‌شود؛ به‌طوری‌که میزان بهبود صحت به ترتیب برابر ۴/۱، ۲/۷ و ۷/۹ درصد است.

دلیل برتری الگوریتم پیشنهادی را می‌توان از دو لحاظ انتخاب ویژگی و روش طبقه‌بندی بررسی نمود. BPSO شباهت زیادی با الگوریتم‌های محاسبه‌ای تکاملی مانند الگوریتم ژنتیک دارد. BPSO بر اساس رفتار اجتماعی در جمعیت‌های بیولوژیکی است. الگوریتم‌های BPSO و

تعداد ۱۶ نمونه مربوط به نمونه‌های MPM و ۱۶ نمونه نیز مربوط به ADCA است و این تفکیک در شکل ۶ (ج) نشان داده شده است. نهایتاً در داده‌های سرطان پروستات نیز تعداد ۵۲ نمونه مربوط به نمونه‌های تومور و ۵۰ نمونه نیز مربوط به نمونه‌های بدون تومور است و این تفکیک نیز در شکل ۶ (د) به‌خوبی انجام گرفته است. در جداول ۷ تا ۱۰ نیز شماره ژن‌های مؤثر به‌دست آمده از اعمال الگوریتم پیشنهادی آورده شده است.

جدول (۳): مقایسه نتایج کمی صحت طبقه‌بندی در الگوریتم

پیشنهادی و سایر روش‌ها در پایگاه داده سرطان خون

الگوریتم انتخاب ژن	طبقه‌بند	صحت طبقه‌بندی (%)
BPSO	BLDA	۸۷/۵
PSO	ANN	۸۶/۱
-	Nero-Fuzzy	۸۷/۵
-	KNN	۷۲/۶
-	Bayesian	۹۱/۲

در یافتن بهترین ترکیب پارامترها دارد [۳۱]. اما طبقه‌بند BLDA به‌راحتی و با صرف زمان کم قابل پیاده‌سازی است. همان‌طور که در جدول‌های ۳ الی ۶ مشاهده می‌شود، ترکیب روش انتخاب ژن BPSO و طبقه‌بند BLDA در اغلب موارد دارای نتیجه بهتری است.

همان‌طور که از شکل‌های ۶ (الف) تا (د) مشاهده می‌شود، الگوریتم پیشنهادی به‌خوبی قادر به جداسازی سطوح بیان ژن‌ها به دو کلاس است. برای مثال در داده سرطان خون (شکل ۶ (الف))، تعداد ۲۷ نمونه مربوط به کلاس ALL و ۱۱ نمونه مربوط به کلاس AML است که این تفکیک به‌خوبی در این شکل نشان داده شده است. به‌طور مشابه، در داده سرطان پستان نیز تعداد ۳۴ نمونه مربوط به نمونه‌هایی است که در آن‌ها سرطان عود کرده و ۴۴ نمونه نیز به نمونه‌هایی مرتبط است که سرطان در آن‌ها عود نکرده است. این تفکیک نیز در شکل ۶ (ب) مشاهده می‌شود. در داده‌های سرطان ریه و پروستات نیز وضعیت مشابهی مشاهده می‌شود؛ به‌طوری‌که در داده‌های سرطان ریه

جدول (۴): مقایسه نتایج کمی صحت طبقه‌بندی در الگوریتم پیشنهادی و سایر روش‌ها در پایگاه داده سرطان ریه

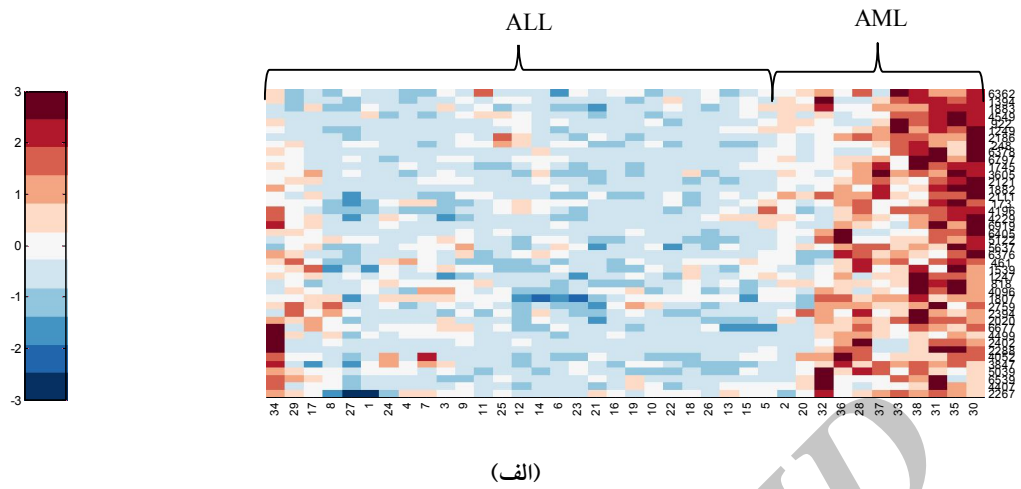
الگوریتم انتخاب ژن	طبقه‌بند	صحت طبقه‌بندی (%)
BPSO	BLDA	۹۹/۵
PSO	SVM	۹۹/۰
IPSO	KNN	۹۶/۵
PSO	Ensemble Neural Network	۱۰۰
PSO	ANN	۹۸/۳
-	Bayesian	۸۹/۴

جدول (۵): مقایسه نتایج کمی صحت طبقه‌بندی در الگوریتم پیشنهادی و سایر روش‌ها در پایگاه داده سرطان پستان

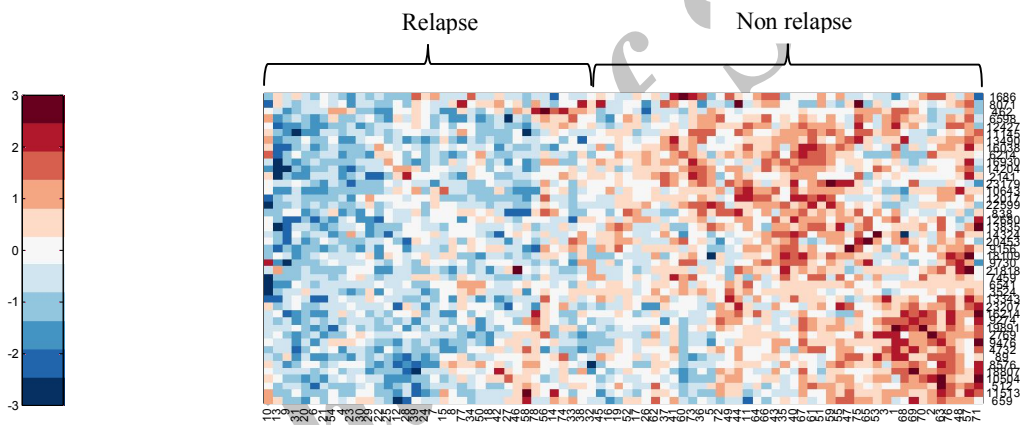
الگوریتم انتخاب ژن	طبقه‌بند	صحت طبقه‌بندی (%)
BPSO	BLDA	۹۳/۵
PSO	SVM	۸۵/۳
GA	SVM	۹۵/۸
-	Bayesian	۷۴/۱

جدول (۶): مقایسه نتایج کمی صحت طبقه‌بندی در الگوریتم پیشنهادی و سایر روش‌ها در پایگاه داده سرطان پروستات

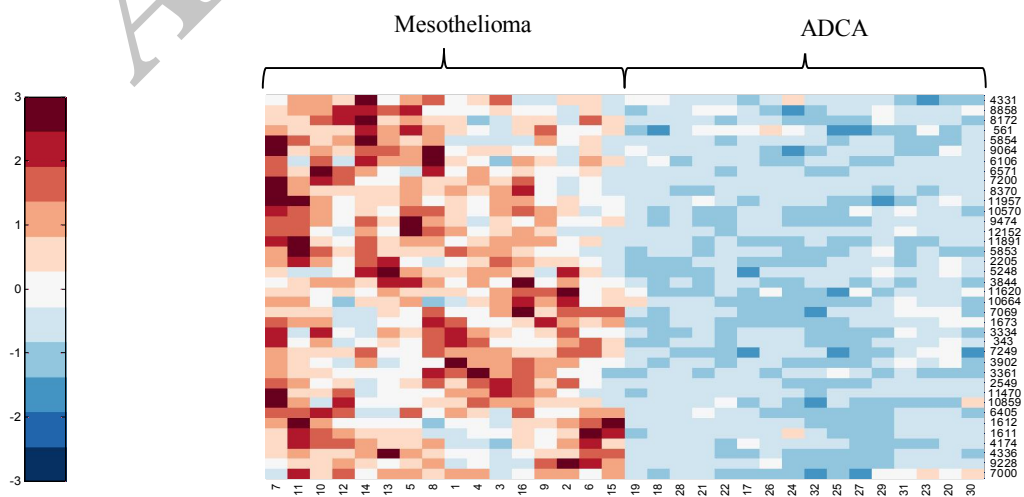
الگوریتم انتخاب ژن	طبقه‌بند	صحت طبقه‌بندی (%)
BPSO	BLDA	۹۵/۹
IPSO	KNN	۹۲/۱
Hybrid PSO/GA	SVM	۹۳/۴
EA	KNN	۸۸/۹



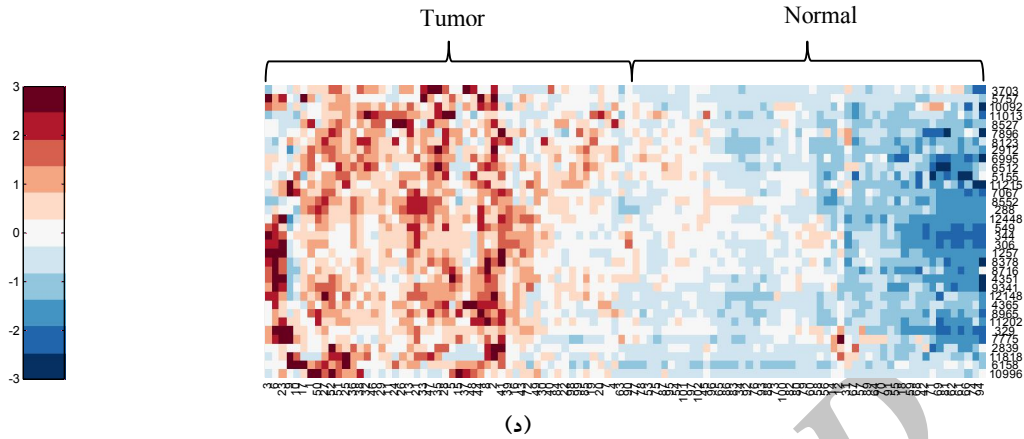
(الف)



(ب)



(ج)



شکل (۶): مجموعه ژن‌های حاوی اطلاعات انتخاب شده با اعمال الگوریتم پیشنهادی در پایگاه‌های داده: (الف) سرطان خون؛ (ب) سرطان پستان؛ (ج) سرطان ریه و (د) سرطان پروستات

جدول (۷): شماره ژن‌های مؤثر در بروز سرطان خون به‌دست آمده از اعمال الگوریتم پیشنهادی

۲۲۶۷	۲۲۸۸	۲۳۹۴	۱۲۴۷	۵۱۲۲	۱۱۷۳	۱۷۴۵	۱۲۴۹	۶۳۶۲
۴۴۰۷	۲۴۰۲	۲۷۵۹	۱۵۳۹	۶۴۰۵	۲۱۱۱	۶۷۹۷	۹۲۲	۵۰۳۹
۶۵۳۹	۴۴۹۹	۱۸۰۷	۴۶۱	۶۹۱۹	۱۸۸۲	۶۳۷۸	۴۵۴۹	
۳۸۴۷	۶۶۷۷	۴۰۹۶	۶۳۷۶	۴۲۲۹	۲۱۲۱	۲۴۸	۱۸۸۳	
۴۰۵۲	۲۰۲۰	۸۱۸	۵۶۳۷	۴۱۹۶	۳۶۰۵	۲۱۸۶	۱۳۹۴	

جدول (۸): شماره ژن‌های مؤثر در بروز سرطان پستان به‌دست آمده از اعمال الگوریتم پیشنهادی

۶۵۹	۸۵۷۶	۱۹۸۹۱	۳۵۲۴	۱۸۱۰۹	۱۲۶۸۰	۲۳۱۷۹	۱۶۰۳۸	۴۶۲
۱۱۵۱۳	۸۹	۹۲۷۴	۶۵۴۱	۹۱۵۶	۸۳۸	۲۱۴۱	۱۳۴۹۰	۸۰۷۱
۵۱۲	۴۷۳۲	۱۶۲۱۴	۷۴۵۹	۲۰۴۵۳	۲۲۵۹۹	۱۴۲۰۴	۱۱۱۴۵	۱۶۸۶
۱۰۵۰۴	۹۴۷۶	۲۳۲۰۷	۲۱۸۱۸	۱۴۳۲۴	۱۲۰۱۷	۱۶۹۳۰	۱۲۴۲۷	
۱۸۸۰۷	۲۷۶۹	۱۳۳۴۳	۹۷۳۰	۱۳۸۳۵	۱۰۶۴۳	۶۲۱۴	۶۵۹۸	

جدول (۹): شماره ژن‌های مؤثر در بروز سرطان ریه به‌دست آمده از اعمال الگوریتم پیشنهادی

۷۰۰۰	۱۶۱۲	۳۳۶۱	۱۶۷۳	۵۲۴۸	۹۴۷۴	۶۵۷۱	۸۱۷۲
۹۲۲۸	۶۴۰۵	۳۹۰۲	۷۰۶۹	۲۲۰۵	۱۰۵۷۰	۶۱۰۶	۸۸۵۸
۴۳۳۶	۱۰۸۵۹	۷۲۴۹	۱۰۶۶۴	۵۸۵۳	۱۱۹۵۷	۹۰۶۴	۴۳۳۱
۴۱۷۴	۱۱۴۷۰	۳۴۳	۱۱۶۲۰	۱۱۸۹۱	۸۳۷۰	۵۸۵۴	
۱۶۱۱	۲۵۴۹	۳۳۳۴	۳۸۴۴	۱۲۱۵۲	۷۲۰۰	۵۶۱	

جدول (۱۰): شماره ژن‌های مؤثر در بروز سرطان پروستات به‌دست آمده از اعمال الگوریتم پیشنهادی

۱۰۹۹۶	۳۲۹	۹۳۴۱	۳۰۶	۸۵۵۲	۶۹۹۵	۱۱۰۱۳
۶۱۵۸	۱۱۲۰۲	۴۳۵۱	۳۴۴	۷۰۶۷	۲۹۱۲	۱۰۰۹۲
۱۱۸۱۸	۸۹۶۵	۸۷۱۶	۵۴۹	۱۱۲۱۵	۸۱۲۳	۵۷۵۷
۲۸۳۹	۴۳۶۵	۸۳۷۸	۱۲۴۴۸	۵۱۵۵	۷۸۹۶	۳۷۰۳
۷۷۷۵	۱۲۱۴۸	۱۲۵۷	۲۸۱	۶۵۱۲	۸۵۲۷	

۶- نتیجه‌گیری

در این مقاله یک الگوریتم جدید و کارآمد بر پایه مدل ترکیبی بهینه‌سازی ازدحام ذرات باینری و الگوریتم آنالیز تفکیک‌کننده خطی بیز برای انتخاب ژن و طبقه‌بندی آن‌ها پیشنهاد و عملکرد آن بر روی چهار پایگاه داده ریزآرایه ارزیابی شد. در ابتدا، کلیه نمونه‌ها به دو دسته نمونه‌های آموزشی و نمونه‌های تست با استفاده از روش اعتبارسنجی fold-10 تقسیم گردید. سپس با به‌کارگیری آنالیز همبستگی پیرسون، یک مجموعه ژن از نمونه‌های آموزشی انتخاب گردید. در مرحله بعد با به‌کارگیری الگوریتم BPSO هر ذره یک مجموعه ژن را انتخاب کرده و سپس میزان تناسب مجموعه ژن انتخاب شده توسط آن ذره با طبقه‌بند BLDA محاسبه گردید. بعد از 50 بار تکرار ذره‌ای که بهترین میزان تناسب (صحت طبقه‌بندی) را داشته است، به‌عنوان مجموعه ژن حاوی اطلاعات در نظر گرفته می‌شود. نتایج پیاده‌سازی نشان داد که الگوریتم پیشنهادی به کاهش بعد داده‌های ریزآرایه منجر می‌شود. همچنین صحت طبقه‌بندی در آن به-علت استفاده از مدل BPSO بهبود می‌یابد؛ علت این امر آن است که در مدل BPSO، همبستگی بین ژن‌ها در نظر گرفته شده و این امر به از بین رفتن افزونگی و همچنین کاهش تعداد ژن‌های فاقد اطلاعات و اضافی منجر می‌شود. با این حال، الگوریتم پیشنهادی در برخی موارد عملکرد ضعیفتری نسبت به سایر الگوریتم‌ها دارد. برای مثال در سرطان ریه، صحت طبقه‌بندی در الگوریتم پیشنهادی برابر 99/5 درصد و در الگوریتم PSO-Ensemble Neural Network برابر 100 درصد است. علت این امر، آن است که استفاده از طبقه‌بندهای گروهی نتیجه بهتری از یک طبقه‌بند دارد. هدف ما در کارهای آینده، ترکیب الگوریتم‌های پیشرفته‌تر با الگوریتم پیشنهادی به منظور افزایش نرخ طبقه‌بندی خواهد بود.

مراجع

- [2] Chu, F., and Wang, L., "Application of support vector machines to cancer classification with microarray data", *International Journal of Neural Systems*, Vol. 15, No. 6, pp. 239-263, 2002.
- [3] Lu, Y., and Han, J., "Cancer Classification Using Gene Expression Data", Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
- [4] Brazma, A., and Vilo, J., *Gene expression data analysis*, European Molecular Biology Laboratory, Outstation Hinxton, 2000.
- [5] Molaiezhadeh, F., Moradi, M. H., "Informative Gene Selection in Microarray Data using Mutual Information and Genetic Algorithm", 13th Conference on Biomedical Engineering, Sharif University of Technology, Tehran, Iran, February 2007.
- [6] Ben-Doe, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z., "Tissue classification with gene expression profiles", *Journal of Computational Biology*, Vol. 7, No. 3-4, pp. 559-583, 2000.
- [7] <http://datam.i2r.a-star.edu.sg/datasets/krbd>.
- [8] Chen, Y., and Zhao, Y., "A novel ensemble of classifiers for microarray data classification", *Applied Soft Computing*, ELSEVIER, No. 8, pp. 1664-1669, 2008.
- [9] Guyon, I. and Elisseeff, A., "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3, No. 3, pp. 1157-1182, 2003.
- [10] Li, L., Weinberg, C. R., Darden, T. A., and Pedersen, L. G., "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method", *Bioinformatics*, Vol. 17, No. 12, pp. 1131-1142, 2001.
- [11] Jourdan, L., *Metheuristics for knowledge discovery: Application to genetic data*, Ph.D. thesis, University of Lille, 2003.
- [12] Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W., and Chen, L., "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines", *FEBS Letter*, Vol. 555, No. 2, pp. 358-362, 2003.
- [13] Reddy, A. R., and Deb, K., *Classification of two-class cancer data reliably using evolutionary algorithms*, Technical Report, KanGAL, 2003.
- [14] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., "Gene selection for cancer classification using support vector machines", *Machine Learning*, Vol. 46, No. 1-3, pp. 389-422, 2002.
- [1] Wee, A., Liew, C., Yah, H., and Yang, M., "Pattern recognition techniques for the emerging field of bioinformatics: A review", *Pattern Recognition*, No. 38, No. 11, pp. 2055 - 2073, 2005.

- algorithm using Bayesian classification approach", *American Journal of Applied Sciences*, Vol. 9, No. 1, pp. 127-131, 2012.
- [27] Alba, E., Garcia-Nieto, J., Jourdan, L., and Ghazali Talbi, El., "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms", *IEEE Congress on Evolutionary Computation (CEC 2007)*, 2007.
- [28] Chuang, L. Y., Chang, H. W., Tu, C. J., and Yang, C. H., "Improved binary PSO for feature selection using gene expression data", *Computational Biology and Chemistry*, Elsevier, Vol. 32, No. 1, pp. 29-38, 2008.
- [29] Li, S., Wu, X., and Tan, M., "Gene selection using hybrid particle swarm optimization and genetic algorithm", *Soft Computing*, Springer, Vol. 12, pp. 1039-1048, 2008.
- [30] Juliusdottir, D., Corne, E., Keedwell, E., and Narayanan, A., "Two-phase EA/KNN for feature selection and classification in cancer microarray datasets", In *CIBCB*, pp. 1-8, 2005.
- [31] T. Li, C.L. Zhang, M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", *Bioinformatics* 20, No. 15, pp. 2429-2437, 2004.
- [15] Saeys, Y., Aeyels Degroev, S., Rouze, D., and Van de peer, Y. P., "Enhancement genetic feature selection through restricted search and Walsh analysis", *IEEE Transactions on Systems, Man and Cybernetics, Part C*, Vol. 34, pp. 398-406, 2004.
- [16] Goh, L., Song, Q., and Kasabov, N., "A novel feature selection method to improve classification of gene expression data", In *Proceedings of the Second Asia-Pacific Conference on Bioinformatics*, pages 161-166, Australian Computer Society, Darlinghurst, Australia, 2004.
- [17] Kennedy, J., and Eberhat, R., "Particle Swarm Optimization", In *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 4, pp. 1942-1948, 1995.
- [18] Kennedy, J., and Eberhat, R., "A Discrete Binary Version of the Particle Swarm Algorithm", In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 5, pp. 4104-4109, 1997.
- [19] Hoffmann, U., Vesin, J. M., Ebrahimi, T., and Diserens, K., "An efficient P300-based brain computer interface for disabled subjects", *Journal of Neuroscience Methods*, Vol. 2, No. 2, pp. 1-5, 2007.
- [20] Hoffmann, U., *Bayesian machine learning applied in a brain computer for disabled users*, Ph.D. thesis, 2007.
- [21] Hoffmann, U., "Bayesian feature selection applied in a P300 brain-computer interface", Vol. 2, No. 3, pp. 1-5, 2007.
- [22] Chuang, L., Chang, H., Tu, C., and Yang, C., "Improved binary PSO for feature selection using gene expression data", *Computational Biology and Chemistry*, ELSEVIER, Vol. 1, No. 32, pp. 29-38, 2008.
- [23] Chen, Y., and Zhao, Y., "A novel ensemble of classifiers for microarray data classification", *Applied Soft Computing*, Elsevier, Vol. 8, No. 4, pp. 1664-1669, 2008.
- [24] Wang, Z. Y., Palade, V., and Xu, Y., "Nero-fuzzy ensemble approach for microarray cancer gene expression data analysis", In *Proceeding of the International Symposium on Evolving Fuzzy Systems*, pp. 241-246, 2006.
- [25] Jirapech,-Umpai, T., and Aitken, S., "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes", *Bioinformatics*, Vol. 6, No. 3, pp. 168-174, 2005.
- [26] Sharma, A., and Paliwal, K., "A gene selection

¹Microarray²Deoxyribonucleic Acid³Spot⁴Prob⁵Complementary DNA (cDNA)⁶Binary Particle Swarm Optimization(BPSO)⁷Bayesian Linear Discriminant Analysis (BLDA)⁸Acute Lymphoblastic Leukemia⁹Acute Myeloid Leukemia¹⁰Relapse¹¹Non-Relapse¹²Malignant Pleural Mesothelioma¹³Adenocarcinoma¹⁴Attribute- Relation File Format¹⁵Person Correlation (PC)¹⁶Over-fitting