انتخاب ویژگی و طبقه‌بندی سلول‌های سرطانی بر پایه داده‌های ریزآرایه با استفاده از الگوریتم جستجوی فاخته چندهدفه

خدیجه کمری، فرزان رشیدی و عبدالله خلیلی

چکیده: داده‌های ریزآرایه نقش مؤثری در طبقه‌بندی و تشخیص انواع بافت‌های سرطانی ایفا می‌کنند. با این حال در پژوهش‌های مرتبط با سرطان، تعداد نسبتاً کم نمونه‌ها در مقایسه با تعداد بسیار زیاد ژن‌ها، باعث ایجاد مشکلاتی از قبیل کاهش کارایی طبقه‌بندها، افزایش هزینه‌های محاسباتی و پیچیدگی در طبقه‌بندی سلول‌های سرطانی خواهد شد. یک راهکار مناسب جهت افزایش کارایی طبقه‌بندها، حذف ژن‌های نامربوط و انتخاب نمونه‌های مناسب برای آموزش طبقه‌بندها است. در این مقاله یک مدل ترکیبی بر پایه الگوریتم بهینه‌سازی جستجوی فاخته چندهدفه و خوشه‌بندی فازی برای طبقه‌بندی داده‌های ریزآرایه پیشنهاد شده است. در این مطالعه از نسخه دودویی الگوریتم جستجوی فاخته چندهدفه به منظور انتخاب ویژگی‌های مرتبط با بیماری و از نسخه پیوسته آن برای انتخاب تعداد نمونه‌های مناسب برای آموزش طبقه‌بندها استفاده شده است. به منظور تسریع در فرایند بهینه‌سازی و جلوگیری از گیرافتادن الگوریتم در بهینه‌های محلی، راهکارهای ابتکاری جدیدی نیز به الگوریتم اضافه شده‌اند. برای بررسی عملکرد مدل پیشنهادی، شبیه‌سازی‌های متعددی بر روی شش مجموعه داده سرطانی انجام گرفته و نتایج آن با دیگر مقالات مقایسه شده است. نتایج به دست آمده نشان می‌دهند در بسیاری از موارد مدل پیشنهادی قادر است در مقایسه با سایر روش‌ها، با انتخاب مجموعه کوچک‌تری از ژن‌های متمایز، منجر به افزایش کارایی طبقه‌بندها شود.

کلیدواژه: **انتخاب ویژگی، انتخاب نمونه، داده‌کاوی، ریزآرایه، الگوریتم جستجوی فاخته چندهدفه. خوشه‌بندی فازی.**

۱- مقدمه

اطلاعات ژنتیکی افراد می‌تواند نقش مؤثری در تشخیص و طبقه‌بندی انواع بیماری‌ها از جمله سرطان داشته باشد. یکی از دقیق‌ترین روش‌ها در این زمینه، بررسی مقادیر بیان ژنی در افراد مختلف توسط فناوری ریزآرایه[1] است. داده‌های ریزآرایه به صورت ماتریسی از هزاران ستون و حداکثر چند صد سطر هستند که هر سطر نشان‌دهنده یک نمونه و هر ستون نیز بیانگر یک ژن (ویژگی) است. حجم بسیار زیاد ژن‌ها و تعداد نسبتاً کم نمونه‌ها می‌تواند باعث ایجاد مشکلاتی مانند افزایش هزینه‌های محاسباتی، کاهش توانایی در تعمیم طبقه‌بندها و همچنین کاهش کارایی آنها در پیش‌بینی نمونه‌های جدید ریزآرایه شود [۱]. در عین حال به علت بالابودن تعداد ویژگی‌ها، احتمال آن که ژن‌های غیر مرتبط در ساخت مدل‌های پیش‌بینی‌کننده تأثیرگذار باشند بسیار زیاد است [۲]. این امر ممکن است در پاره‌ای از موارد در تفسیر ژن‌های مسبب بیماری مشکلاتی را ایجاد نماید. چرا که از دیدگاه بیولوژیکی تنها مجموعه کوچکی از ژن‌ها مربوط به بیماری هستند و مابقی ژن‌ها در واقع نقش یک پس‌زمینه نویزی را ایفا می‌کنند که بعضاً ممکن است اثر آن مجموعه کوچک را نیز محو نمایند [۳]. به هر حال افزایش هرچه بیشتر تعداد ژن‌ها در مجموعه داده‌های ریزآرایه، باعث کاهش نرخ صحت طبقه‌بندی و همچنین گیج‌شدن طبقه‌بند خواهد شد. از این رو اولین قدم در آنالیز داده‌های ریزآرایه، استخراج ژن‌های مرتبط و متمایز از بین حجم انبوهی از داده‌های بیان ژن است چرا که تمرکز بر روی مجموعه کوچک‌تری از ژن‌ها باعث تفسیر بهتر نقش ژن‌های حاوی اطلاعات می‌شود [۴]. به همین دلیل امروزه تلاش پژوهشگران بر این است که با ارائه راهکارهای جدیدی، ضمن حذف داده‌های نویزی، نامرتبط و اضافی، علاوه بر کاهش ابعاد ریزآرایه، صحت طبقه‌بندی را نیز به طور مطلوبی افزایش دهند.

امروزه با افزایش ابعاد داده‌ها و همچنین با توجه به این که مسأله انتخاب ویژگی و طبقه‌بندی داده‌های ریزآرایه‌ها جزء مسایل بهینه‌سازی غیر چندجمله‌ای سخت[2] است، توجه محققان به سمت استفاده از الگوریتم‌های بهینه‌سازی فراابتکاری جلب شده است [۵] و [۶]. با این حال انتخاب یک الگوریتم بهینه‌سازی مناسب می‌تواند نقش مؤثری در انتخاب ژن‌های متمایز و افزایش کارایی طبقه‌بندها داشته باشد.

تا کنون روش‌ها و الگوریتم‌های بهینه‌سازی متعددی جهت انتخاب ژن‌های مؤثر بر بیماری‌های سرطان به کار گرفته شده‌اند. با این حال در اکثر پژوهش‌های مرتبط با این موضوع، از نسخه تک‌هدفه این الگوریتم‌ها استفاده شده است. به عبارت دیگر، هدف اصلی اکثر این پژوهش‌ها افزایش صحت طبقه‌بندی بوده و اهدافی مانند کاهش تعداد ویژگی‌ها یا توجه به مسأله افزونگی بین آنها در اولویت‌های بعدی قرار گرفته‌اند [۷] و [۸]. این در حالی است که با استفاده از الگوریتم‌های بهینه‌سازی چندهدفه می‌توان علاوه بر کاهش تعداد ویژگی‌ها و توجه به مسأله افزونگی بین آنها، کارایی طبقه‌بندها را تا حد بسیار زیادی افزایش داد [۷] تا [۹]. هدف این مقاله ارائه چارچوب جدیدی جهت انتخاب ژن‌های متمایز در بروز بیماری و طبقه‌بندی آنها است. بدین منظور جهت انتخاب ژن‌های متمایز و مؤثر از الگوریتم جستجوی فاخته چندهدفه دودویی استفاده شده است. همچنین از آنجایی که کارایی الگوریتم‌های طبقه‌بندی به شدت وابسته به کیفیت داده‌های آموزشی است، یکی دیگر از اهداف این پژوهش، انتخاب نمونه‌های مناسب برای آموزش طبقه‌بندها است. برای دست‌یابی به این هدف نیز از ترکیب روش خوشه‌بندی فازی و نسخه پیوسته الگوریتم جستجوی فاخته چندهدفه بهره گرفته شده است. توابع هدفی که توسط

این مقاله در تاریخ ۱۱ تیر ماه ۱۳۹۷ دریافت و در تاریخ ۶ شهریور ماه ۱۳۹۸ بازنگری شد.

خدیجه کمری، دانشکده فنی و مهندسی، دانشگاه هرمزگان، بندرعباس، ایران، (email: pzi.baran@gmail.com).

فرزان رشیدی (نویسنده مسئول)، دانشکده فنی و مهندسی، دانشگاه هرمزگان، بندرعباس، ایران، (email: rashidi@hormozgan.ac.ir).

عبدالله خلیلی، دانشکده فنی و مهندسی، دانشگاه هرمزگان، بندرعباس، ایران، (email: khalili@hormozgan.ac.ir).

1. Microarray

2. NP-Hard

نسخه دودویی الگوریتم جستجوی فاخته بهینه می‌شوند عبارت هستند از بهره اطلاعاتی و شاخص F. همچنین توابع هدف مورد استفاده در نسخه پیوسته الگوریتم نیز فشرده‌سازی سراسری[1] و جداسازی فازی[2] خوشه‌ها هستند.

برای ارزیابی عملکرد ساختار پیشنهادی، روش مورد نظر بر روی سه طبقه‌بند پرکاربرد شامل ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایه (KNN)، نایو بیز (NB) و ترکیب این سه طبقه‌بند بر اساس رأی اکثریت (MV)، پیاده‌سازی و کارایی آنها جهت تشخیص و دسته‌بندی شش مجموعه داده سرطانی مورد بررسی قرار داده شده است. همچنین مقایسه‌ای بین نتایج حاصل از این روش با روش‌های ارائه‌شده در سایر مقالات انجام گرفته است. نتایج به دست آمده بیانگر آن است که روش پیشنهادی قادر است با انتخاب مجموعه کوچک‌تری از ژن‌های متمایز، منجر به افزایش کارایی طبقه‌بندها شود. دلیل استفاده از طبقه‌بندهای فوق، کاربرد بسیار زیاد آنها در مسایل مرتبط با داده‌کاوی و انتخاب ویژگی است.

مقایسه کارایی الگوریتم جستجوی فاخته با دیگر الگوریتم‌های فراابتکاری مبتنی بر هوش جمعی نشان می‌دهد در اغلب مسایل بهینه‌سازی، الگوریتم جستجوی فاخته می‌تواند علاوه بر داشتن سرعت همگرایی مناسب، دقت بهتری در دست‌یابی به جواب بهینه سراسری داشته باشد. به همین دلیل از این الگوریتم در حوزه‌های مختلف بهینه‌سازی مانند پردازش تصویر، یادگیری ماشین و دیگر حوزه‌های مهندسی استفاده شده است. مرجع [۱۰] الگوریتم جستجوی فاخته را با جزئیات مورد بررسی قرار داده و کاربردهای آن را در حوزه‌های مختلف بهینه‌سازی بیان نموده است. بر اساس نتیجه‌گیری انجام‌شده در مرجع فوق، الگوریتم مذکور علی‌رغم داشتن مزایایی همچون سرعت همگرایی بالا و دقت مناسب در دستیابی به جواب بهینه سراسری، بعضاً در مسایل با ابعاد بالا در بهینه‌های محلی گیر خواهد کرد که دلیل آن ثابت‌بودن پارامترهای کنترلی الگوریتم طی فرایند بهینه‌سازی است. نوآوری‌های این مقاله عبارتند از:

۱) ارائه یک مدل ترکیبی در قالب مسایل بهینه‌سازی چندهدفه برای انتخاب ویژگی‌های متمایز از داده‌های ریزآرایه.
۲) ارائه نسخه بهبودیافته الگوریتم جستجوی فاخته تک‌هدفه به منظور تسریع در فرایند بهینه‌سازی و جلوگیری از گیرافتادن آن در بهینه‌های محلی در مسایل با ابعاد بالا.
۳) ارائه نسخه دودویی الگوریتم جستجوی فاخته با الهام از نسخه دودویی الگوریتم بهینه‌سازی ازدحام ذرات.
۴) انتخاب ژن‌های متمایز با استفاده از نسخه دودویی الگوریتم جستجوی فاخته چندهدفه.
۵) انتخاب تعداد نمونه‌های مناسب برای آموزش طبقه‌بندها با استفاده از نسخه پیوسته الگوریتم جستجوی فاخته چندهدفه.
۶) بررسی تأثیر انتخاب ویژگی و انتخاب نمونه بر صحت طبقه‌بندی داده‌های ریزآرایه.

۲- مروری بر کارهای انجام‌شده

در مبحث داده‌کاوی، با هدف کاهش تعداد ویژگی‌ها، از بین بردن ویژگی‌های غیر مرتبط و حذف داده‌های نویز، از الگوریتم‌های انتخاب ویژگی استفاده می‌شود. روش‌های انتخاب ویژگی را می‌توان به طور کلی به سه دسته فیلتری[3]، بسته‌بندی[4] و ادغامی[5] تقسیم‌بندی کرد [۴]. در روش‌های فیلتری، قدرت هر ژن در تفکیک نمونه‌ها بر اساس یک شاخص آماری محاسبه گردیده و سپس ژن‌هایی که بر اساس معیار محاسبه‌شده، قدرت تفکیک بهتری دارند به عنوان مجموعه ژن‌های مؤثر در بروز بیماری انتخاب می‌شوند. از جمله این روش‌ها می‌توان به انتخاب ژن با استفاده از معیار فیشر [۱۱]، معیار مجموع رتبه ویلکوکسون [۱۲]، تحلیل مؤلفه اصلی (PCA)[6] [۱۳]، روش بهره اطلاعاتی[7] [۱۴] و فیلتر مبتنی بر همبستگی سریع (FCBF)[8] [۱۵] اشاره کرد. یکی دیگر از روش‌های مبتنی بر فیلتر، روش ReliefF است که برای حل مشکل داده‌های نویزی و چندکلاسه مورد استفاده قرار می‌گیرد. در این الگوریتم از یک وزن برای نشان‌دادن میزان ارتباط استفاده شده، سپس به صورت تصادفی یک نمونه از بین نمونه‌های موجود انتخاب گردیده و وزن‌ها بر اساس میزان تفاوت بین نمونه انتخابی و نمونه موجود در کلاس مشترک و نمونه موجود در یک کلاس دیگر محاسبه می‌شوند [۱۶].

مرجع [۱۷] یک روش فیلتری با عنوان MASSIVE[9] پیشنهاد کرده که بر اساس یک معیار تئوری اطلاعات به نام DISR[10] عمل انتخاب ویژگی داده‌های ریزآرایه را انجام می‌دهد. در [۱۸] نیز بر اساس روش فیلتری چندوظیفه‌ای، الگوریتمی پیشنهاد داده‌اند که با استفاده از آن می‌توان در مجموعه داده‌های با ابعاد بالا، عمل انتخاب ویژگی‌های متمایز را انجام داد. مرجع [۱۹] نیز یک روش فیلتری با عنوان MWMR را پیشنهاد داده که با استفاده از آن می‌توان در مجموعه داده‌های ریزآرایه، زیرمجموعه بهینه‌ای از ژن‌ها که دارای بیشینه وزن و کمینه همبستگی هستند انتخاب نمود. در این الگوریتم وزن هر ژن نشان‌دهنده اهمیت آن ویژگی است. از مزایای الگوریتم‌های مبتنی بر روش فیلتر، سادگی در پیاده‌سازی و حجم کم محاسبات است. از محدودیت‌های آنها نیز می‌توان به عدم توجه به افزونگی و ارتباط بین ژن‌ها اشاره کرد [۲۰].

در روش‌های بسته‌بندی از یک الگوریتم طبقه‌بند برای انتخاب مجموعه ژن‌های مؤثر در بروز بیماری استفاده می‌شود. در این روش‌ها به کمک یک مکانیزم جستجو، در هر مرحله یک زیرمجموعه ژن انتخاب شده و کیفیت آن بر اساس کارایی طبقه‌بند مورد ارزیابی قرار می‌گیرد. زیرمجموعه‌ای از ژن‌ها که بالاترین کارایی را در طبقه‌بند ایجاد نمایند به عنوان مجموعه ژن‌های متمایز انتخاب می‌شوند [۲۱]. با توجه به گسترده‌بودن فضای جستجو، در این روش غالباً از الگوریتم‌های بهینه‌سازی هوشمند مانند الگوریتم ژنتیک [۵] و [۲۲]، الگوریتم تبرید فلزات [۲۳]، الگوریتم ازدحام ذرات دودویی [۲۴] و [۲۵]، الگوریتم تکامل تفاضلی [۲۶]، الگوریتم IGIS [۲۷]، ترکیب الگوریتم ReliefF با الگوریتم جستجوی گرانشی [۲۸]، الگوریتم بهینه‌سازی مبتنی بر جغرافیای زیستی [۲۹] و [۳۰] و الگوریتم بهینه‌سازی خفاش [۳۱] استفاده می‌شود. از محدودیت‌های این الگوریتم‌ها، همگرایی زودرس و گیرافتادن در بهینه‌های محلی و از مزایای آنها نیز در نظر گرفته شدن تعامل بین ژن‌ها

1. Compactness
2. Fuzzy Separation
3. Filter
4. Wrapper
5. Embedded
6. Principal Component Analysis
7. Information Gain Based Method
8. Fast Correlation-Based Filter
9. Matrix of Average Sub-Subset Information for Variable Elimination
10. Double Input Symmetrical Relevance

و توجه به مسأله افزونگی است [۲۸] و [۳۱].

از دیگر روش‌های بسته‌بندی مورد استفاده در انتخاب ژن‌های ریزآرایه می‌توان به روش انتخاب ویژگی متوالی SFS[1] اشاره کرد. در این روش ابتدا ژن‌های برتر هر بلوک با توجه به عملکردشان در طبقه‌بند مشخص شده و سپس برای دستیابی به بهترین زیرمجموعه، ژن‌ها با هم مقایسه می‌گردند [۳۲].

دسته دیگری از روش‌های انتخاب ژن، روش‌های ادغامی هستند که هدف این روش‌ها، استفاده از مزایای هر دو روش فیلتری و بسته‌بندی است. به عبارت دیگر در عین این که پیچیدگی محاسباتی آن نسبت به روش‌های بسته‌بندی کمتر است، تعامل بین ژن‌ها را نیز در نظر می‌گیرد [۳۳]. از الگوریتم SVMRFE[2]، IWSS[3] و Approximate Markov Blanket می‌توان به عنوان نمونه‌هایی از انواع روش‌های ادغامی نام برد [۳۴].

هرچند تا کنون روش‌ها و الگوریتم‌های متعددی جهت انتخاب ژن‌های متمایز و مؤثر در تشخیص بیماری‌های سرطانی معرفی شده است اما همان طور که اشاره شد، برخی از این روش‌ها دارای محدودیت‌هایی از قبیل عدم توجه به افزونگی ژن‌ها یا افزایش هزینه‌های محاسباتی هستند. وجود افزونگی در مجموعه ژن‌های انتخابی، علاوه بر افزایش حجم محاسبات، منجر به کاهش کارایی طبقه‌بندها نیز خواهد شد، چرا که انتخاب ژن‌های افزونه، اطلاعات بیشتری را فراهم نمی‌کنند. به علاوه ممکن است این مجموعه ژن‌ها جزء ژن‌های دارای پس‌زمینه نویز باشند [۳۵]. یک راهکار برای کاهش افزونگی در مجموعه ژن‌های انتخابی، استفاده از روش خوشه‌بندی است. به طور نمونه، [۳۶] قبل از انتخاب ویژگی (انتخاب ژن)، از روش خوشه‌بندی فازی برای حذف ژن‌های مشابه استفاده کرده است. اما در [۳۷] پس از انتخاب ژن، با اعمال روش خوشه‌بندی سلسله‌مراتبی سعی در حذف ژن‌هایی نموده است که الگوهای بیانی مشابه دارند. در [۳۸] نیز از روش مارکوف بلانکت برای تشخیص و کاهش افزونگی ژن‌ها استفاده شده است. مرجع [۳۹] روشی به نام MRMR[4] را ارائه نموده که هدف آن انتخاب مجموعه ژنی است که به طور هم‌زمان کمترین میزان افزونگی و بیشترین میزان ارتباط با کلاس مورد بررسی را داشته باشد. در [۴۰] نیز ابتدا با استفاده از روش تحلیل مؤلفه اصلی، ژن‌های مؤثر و متمایز استخراج گردیده و سپس با استفاده از طبقه‌بند SVM اقدام به کاهش تعداد ژن‌ها شده است.

۳- الگوریتم جستجوی فاخته

با توجه به این که در این پژوهش از نسخه پیوسته و دودویی الگوریتم جستجوی فاخته چندهدفه جهت کاهش تعداد ویژگی‌ها و انتخاب نمونه‌های مناسب برای آموزش طبقه‌بندها استفاده شده است، در ادامه این الگوریتم مورد بررسی قرار داده شده است.

۳-۱ الگوریتم جستجوی فاخته پیوسته

الگوریتم جستجوی فاخته یک الگوریتم بهینه‌سازی مبتنی بر جمعیت است که اولین بار در سال ۲۰۰۹ توسط یانگ و دِب معرفی گردید. این الگوریتم که از فرایند تخم‌گذاری و زادآوری فاخته‌ها الهام گرفته، از سه قاعده زیر پیروی می‌کند [۴۱]:

۱) هر پرنده در هر زمان تنها یک تخم می‌گذارد و آن را در لانه‌ای قرار می‌دهد که به صورت تصادفی انتخاب شده است (هر لانه یک راه‌حل را در خود نگه می‌دارد).

۲) لانه‌هایی که تخم‌های (راه‌حل‌های) با کیفیت بهتری دارند به نسل بعد منتقل می‌شوند.

۳) تعداد لانه‌های در دسترس در طول اجرای الگوریتم ثابت بوده و پرنده میزبان به احتمال p_a تخم مهمان را شناسایی می‌کند. در این وضعیت پرنده میزبان می‌تواند تخم مهمان را دور بریزد یا لانه را به طور کامل به مکان جدیدی منتقل کند. در الگوریتم ارائه‌شده توسط یانگ برای ساده‌سازی، p_a درصد از N لانه موجود با لانه‌های جدید جایگزین می‌شود (با راه‌حل‌های تصادفی جدید در مکان‌های جدید). همانند سایر الگوریتم‌های بهینه‌سازی مبتنی بر جمعیت، الگوریتم جستجوی فاخته با یک جمعیت اولیه از فاخته‌ها کار خود را شروع می‌کند. این جمعیت از فاخته‌ها تعدادی تخم دارند که آنها را در لانه تعدادی پرنده میزبان می‌گذارند. تعدادی از این تخم‌ها که شباهت بیشتری به تخم‌های پرنده میزبان دارند، شانس بیشتری برای رشد و تبدیل‌شدن به فاخته بالغ خواهند داشت و سایر تخم‌ها توسط پرنده میزبان شناسایی شده و از بین می‌روند. فاخته‌ها برای بیشینه‌کردن شانس نجات تخم‌های خود به دنبال بهترین منطقه می‌گردند. تعداد تخم‌های رشدیافته در یک منطقه به منزله مناسب‌بودن لانه‌های آن منطقه است. هرچه تخم‌های بیشتری در یک منطقه رشد کرده و نجات یابند به همان اندازه تمایل بیشتری به آن منطقه اختصاص می‌یابد. بنابراین موقعیتی که در آن بیشترین تعداد تخم‌ها نجات یابند، پارامتری خواهد بود که الگوریتم جستجوی فاخته قصد بهینه‌سازی آن را دارد. با در نظر گرفتن تعداد تخمی که هر فاخته خواهد گذاشت و همچنین فاصله فاخته‌ها از منطقه بهینه فعلی برای سکونت، تعدادی شعاع تخم‌گذاری تعیین می‌شود. سپس فاخته‌ها شروع به تخم‌گذاری تصادفی در لانه‌هایی داخل شعاع تخم‌گذاری خود می‌کنند. پس از چند تکرار، تمام جمعیت فاخته‌ها در یک نقطه که بیشترین شباهت را به تخم‌های پرندگان میزبان و همچنین بیشترین منابع غذایی دارند جمع می‌شوند. این محل بیشترین سود کلی را خواهد داشت و در آن کمترین تعداد تخم‌ها از بین خواهد رفت.

در الگوریتم جستجوی فاخته، هر تخمی که در لانه‌ای گذاشته می‌شود در واقع نماینده یک راه‌حل است. راه‌حل‌های جدید نیز به کمک پرواز لِوی تولید می‌شوند. پرواز لِوی که رفتار موجوداتی مانند فاخته‌ها، نوعی مرغابی و میمون‌های عنکبوتی را هنگام جستجوی غذا مدل‌سازی می‌کند یک قدم‌زدن تصادفی است که در آن طول گام‌ها از توزیع لِوی پیروی می‌کند [۴۱]. مطالعات نشان داده است که پرواز لِوی می‌تواند کارایی فرایند جستجو را در شرایط عدم اطمینان، بیشینه نماید. در الگوریتم جستجوی فاخته، جواب‌های جدید با استفاده از پرواز لِوی طبق رابطه زیر تولید می‌شوند [۴۱]

$$X_i(t+1) = X_i(t) + \alpha \times Levy(\lambda) \times (X_i(t) - X_{best}(t)) \quad (1)$$

در رابطه فوق، $X_i(t)$ بیانگر موقعیت لانه iام در مرحله تکرار t، $X_{best}(t)$ موقعیت بهترین لانه (بهترین جواب) در مرحله تکرار tام و $Levy(\lambda)$ توزیع پرواز لِوی است. همچنین α طول گام بوده که مقدار آن می‌بایست متناسب با مقیاس مسأله و فضای جواب تعیین شود. طبق

1. Successive Feature Selection
2. Suport Vector Machin Recursive Feature Elimination
3. Incremental Wrapper Subset Selection
4. Max Relevance-Min Redundancy

رابطه فوق هرچه تخم‌های فاخته به تخم‌های پرنده میزبان شباهت بیشتری داشته باشد، شانس عدم شناسایی و احتمال بقای آنها بالاتر است. برای افزایش سرعت جستجوی محلی، بعضی از راه‌حل‌های جدید توسط توزیع لِوی در اطراف بهترین راه‌حلی که در هر مرحله به دست می‌آید ($X_{best}(t)$) انتخاب می‌شوند. همچنین جهت پیمایش فضای جستجو، تعدادی دیگر از راه‌حل‌ها نیز به طور تصادفی تولید می‌گردند. توزیع پرواز لِوی نیز طبق رابطه زیر محاسبه می‌شود [۴۱]

$$Levy \sim u = r^{-\lambda} \quad , \ (1 < \lambda \le 3) \qquad (۲)$$

در رابطه فوق r یک عدد تصادفی در بازه صفر تا یک است.

شبه‌کد نسخه اصلی الگوریتم جستجوی فاخته به شکل زیر است [۴۱]:

۱) تعداد N_p لانه $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ را به صورت تصادفی تولید کن به گونه‌ای که هر یک از X_i ها برداری d بعدی است و d بعد فضای جواب (تعداد متغیرهای تصمیم‌گیری) است.

۲) به صورت تصادفی و با استفاده از پرواز لِوی یک جواب جدید مانند X_j تولید کن.

۳) مقدار تابع هدف f را به ازای عضو جدید X_j محاسبه کن $f(X_j)$.

۴) یکی از لانه‌ها مانند لانه iام را به صورت تصادفی انتخاب و مقدار تابع هدف را به ازای آن محاسبه کن $f(X_i)$.

۵) اگر $f(X_j)$ بهتر از $f(X_i)$ باشد، آن گاه X_j را با X_i عوض کن.

۶) بدترین لانه‌ها را دور بریز (p_a درصد از بدترین لانه‌ها) و به صورت تصادفی در مکان جدید دوباره آنها را بساز.

۷) مقدار تابع هدف را به ازای تمامی جواب‌ها (تخم‌ها) محاسبه کن.

۸) بهترین جواب این مرحله را محاسبه کن $(X_{best}(t))$.

۹) بهترین جواب به دست آمده تا این مرحله را ذخیره کن $(gbest)$.

۱۰) تا زمانی که شرط خاتمه الگوریتم حاصل نشده است به مرحله ۲ برگرد.

مهم‌ترین عاملی که کارایی و دقت یک الگوریتم بهینه‌سازی هوشمند را کنترل می‌کند، ایجاد مصالحه بین دو مؤلفه اکتشاف[1] و استخراج[2] است [۴۲]. مؤلفه اکتشاف به الگوریتم این امکان را می‌دهد که بتواند جهت دستیابی به پاسخ‌های جدید، فضای جواب را بدون گیرافتادن در بهینه‌های محلی جستجو نماید. به عبارت دیگر مفهوم اکتشاف به معنای توانایی الگوریتم در جستجوی مناطق مختلف فضای جواب، جهت یافتن پاسخ‌های جدید است. در حالی که مؤلفه استخراج باعث می‌شود الگوریتم بتواند مکان‌های بهینه را به صورت محلی و متمرکز برای یافتن بهترین جواب، جستجو نماید. به بیان دیگر، استخراج به معنی قابلیت متمرکزکردن جستجو در محدوده مطلوب است تا جواب مورد نظر موشکافی شود. بنابراین برای رسیدن به جواب بهینه سراسری، می‌بایست مصالحه‌ای بین دو مؤلفه اکتشاف و استخراج صورت پذیرد.

در نسخه استاندارد الگوریتم جستجوی فاخته، تعداد p_a درصد از بدترین لانه‌ها از بین رفته و به جای آن لانه‌های جدیدی به طور تصادفی ایجاد می‌شوند. این عمل که همان مفهوم اکتشاف است در حقیقت منجر به افزایش توانایی الگوریتم در جستجوی فضاهای جدید شده و می‌تواند از گیرافتادن آن در بهینه‌های محلی جلوگیری نماید. اما این امر خود ماهیت تصادفی‌بودن الگوریتم را افزایش خواهد داد.

یک استراتژی مناسب برای ایجاد تعادل بین دو مؤلفه اکتشاف و استخراج این است که در اولین تکرارهای الگوریتم، تأثیر مفهوم اکتشاف بیشتر از استخراج باشد و با گذشت زمان و تکرار الگوریتم، از تأثیر اکتشاف کاسته شده و اهمیت بیشتری به مفهوم استخراج داده شود. به این معنی که در تکرارهای اولیه، الگوریتم یک جستجوی سراسری در فضای جواب انجام داده و در تکرارهای آخر، ناحیه‌های یافته‌شده را با دقت بیشتری جستجو کند. با توجه به این که در نسخه استاندارد الگوریتم جستجوی فاخته، عملگری جهت ایجاد تعادل بین دو مؤلفه استخراج و اکتشاف وجود ندارد، در مسایل بهینه‌سازی با ابعاد بالا، کارایی الگوریتم می‌تواند با چالش‌هایی از جمله کندشدن سرعت همگرایی یا بعضاً دورشدن از جواب بهینه سراسری مواجه شود. در این مقاله برای افزایش سرعت همگرایی و بهبود کیفیت جواب‌ها، در مرحله ۶ به جای ایجاد تصادفی لانه‌های جدید، از رابطه زیر که الهام‌گرفته از الگوریتم تکامل تفاضلی است استفاده شده است [۲۶]

$$X_i(t) = X_r + \beta \times (gbest - X_r) \qquad (۳)$$

در رابطه فوق، X_r موقعیت جدید لانه تصادفی و $gbest$ نیز موقعیت بهترین لانه تا مرحله تکرار t است. پارامتر β نیز سرعت همگرایی الگوریتم را کنترل می‌کند. در این مقاله برای ایجاد تعادل بین خاصیت اکتشاف و استخراج از رابطه زیر برای محاسبه پارامتر β استفاده شده است

$$\beta(t) = \beta_{\max} - (\beta_{\max} - \beta_{\min}) \frac{t}{MaxGen} \qquad (۴)$$

در رابطه فوق $MaxGen$ حداکثر تعداد تکرار الگوریتم و $\beta_{\min}$ و $\beta_{\max}$ نیز بیانگر محدوده تغییرات β است.

طبق (۴) مشاهده می‌شود جهت ایجاد تعادل بین دو مفهوم اکتشاف و استخراج در الگوریتم، پارامتر β به صورت خطی از یک مقدار بیشینه به یک مقدار کمینه کاهش می‌یابد به طوری که در مراحل اولیه الگوریتم سعی می‌کند در یک زمان معقول، فضای جواب را پیمایش کرده و در تکرارهای پایانی یک جستجوی محلی در همسایگی بهترین جواب انجام دهد. با تکرار مراحل فوق به تدریج لانه‌ها به سمت نقاط بهینه حرکت می‌کنند و در پایان اجرای الگوریتم تمامی N لانه در اطراف نقطه بهینه جمع می‌شوند. بنابراین اگر تعداد لانه‌ها به اندازه کافی از تعداد نقاط بهینه بیشتر باشد در بیشتر موارد، الگوریتم به سمت جواب بهینه همگرا خواهد شد. با انجام اصلاحات فوق در نسخه اصلی الگوریتم جستجوی فاخته، می‌توان به یک نسخه بهبودیافته از الگوریتم دست یافت که در آن مسأله ثابت‌بودن پارامترهای کنترلی الگوریتم مرتفع شده و امکان ایجاد تعادل بین دو مؤلفه اکتشاف و استخراج نیز وجود دارد. همان طور که در بخش شبیه‌سازی آمده است این موضوع می‌تواند علاوه بر افزایش همگرایی، منجر به کاهش احتمال گیرافتادن الگوریتم در بهینه‌های محلی شود.

۳-۲ الگوریتم جستجوی فاخته دودویی

در حل مسأله انتخاب ویژگی با استفاده از الگوریتم‌های بهینه‌سازی هوشمند، هر جواب مسأله به صورت رشته‌ای از صفرها و یک‌ها تعریف می‌شود. طول رشته نیز برابر تعداد کل ویژگی‌ها است. مقدار صفر یا یک برای هر بیت رشته به ترتیب بیانگر انتخاب یا عدم انتخاب ویژگی متناظر با آن بیت رشته است.

الگوریتم جستجوی فاخته معرفی‌شده در بخش قبل، برای حل مسایل بهینه‌سازی پیوسته به کار می‌رود و برای حل مسأله انتخاب ویژگی که

1. Exploration
2. Exploitation

جزء مسایل بهینه‌سازی دودویی است باید از نسخه گسسته آن استفاده کرد. تا کنون نسخه‌های مختلفی از الگوریتم جستجوی فاخته گسسته ارائه شده است. با توجه به توانایی قابل توجه الگوریتم بهینه‌سازی ازدحام ذرات دودویی در حل مسایل مختلف، در این مقاله با الهام از نسخه دودویی الگوریتم مذکور، از (۵) و (۶) برای تبدیل نسخه پیوسته الگوریتم جستجوی فاخته به نسخه دودویی استفاده شده است [۲۴]

$$\mathrm{sgn}(x_{ij}(t)) = \frac{1}{1 - e^{-xij(t)}} \quad (۵)$$

$$x_{ij}(t) = \begin{cases} 1 & \text{if} \quad \mathrm{sgn}(x_{ij}(t)) \geq rand \\ 0 & \text{otherwise} \end{cases} \quad (۶)$$

در رابطه فوق، $rand$ یک عدد تصادفی در بازه صفر و یک و x_{ij} نیز بیانگر jامین بیت از لانه iام است. سایر مراحل الگوریتم جستجوی فاخته دودویی، همانند نسخه پیوسته بهبود یافته است که جزئیات آن در بخش قبل ذکر گردید.

۳-۳ الگوریتم جستجوی فاخته چندهدفه

الگوریتم جستجوی فاخته‌ای که در بخش‌های قبل معرفی شد، الگوریتمی است که برای حل مسایل بهینه‌سازی تک‌هدفه کاربرد دارد اما مسأله مورد نظر در این پژوهش، از نوع مسایل بهینه‌سازی چندهدفه است. در مسایل بهینه‌سازی چندهدفه، پیداکردن یک جواب منحصربه‌فرد که بتواند به طور هم‌زمان تمامی توابع هدف را بهینه نماید، معمولاً امکان‌پذیر نیست. با این حال می‌توان به مجموعه‌ای از جواب‌ها دست یافت که بهترین تعامل را بین اهداف برقرار کنند. به این مجموعه جواب‌ها، جواب‌های پارتو یا جبهه پارتو گفته می‌شود. جبهه پارتو در حقیقت همان جواب‌های بهینه‌ای است که توسط هیچ جواب دیگری از مجموعه جواب‌ها مغلوب نمی‌شوند. تعریف ریاضی مغلوب‌بودن یک جواب به صورت زیر است [۴۳]:

در یک مسأله بهینه‌سازی mهدفه $f_i, i = 1,2,\ldots,m$ می‌گوییم جواب x_1 جواب x_2 را مغلوب می‌کند اگر و تنها اگر دو شرط زیر برقرار باشد:

۱) به ازای تمامی f_iها، $f_i(x_1)$ بدتر از $f_i(x_2)$ نباشد.

۲) حداقل به ازای یکی از f_iها، $f_i(x_1)$ بهتر از $f_i(x_2)$ باشد.

اگر یکی از دو شرط فوق برقرار نباشد می‌گوییم جواب x_1 جواب x_2 را مغلوب نمی‌کند. شایان ذکر است اگر جواب x_1 بر جواب x_2 غلبه نکند لزوماً به معنی آن نیست که x_2 بر x_1 غلبه می‌کند.

با توجه به تعریف فوق، جواب $x^* \in X$ را بهینه به مفهوم پارتو می‌گوییم اگر به ازای تمامی توابع هدف، هیچ یک از جواب‌های $x \in X$ نتوانند x^* را مغلوب نمایند. یعنی به ازای تمامی xها، $f_i(x^*)$ بدتر از $f_i(x)$ نباشد و حداقل به ازای یکی از f_iها، $f_i(x^*)$ بهتر از $f_i(x)$ باشد. در ادامه با استفاده از توضیحات ذکرشده در بالا، الگوریتم جستجو فاخته تک‌هدفه معرفی‌شده در بخش قبل به گونه‌ای تعمیم داده شده که قابل استفاده برای حل مسایل بهینه‌سازی چندهدفه نیز باشد [۴۴]. ابتدا مفهومی به نام لیست آرشیو را معرفی می‌کنیم. لیست آرشیو، لیستی است که جواب‌های بهینه به مفهوم پارتو در آن قرار می‌گیرند. یعنی جواب‌هایی که اولاً مغلوب هیچ یک از اعضا نمی‌شوند و ثانیاً بهبود در یک تابع هدف، منجر به بدترشدن حداقل یک تابع هدف دیگر می‌شود.

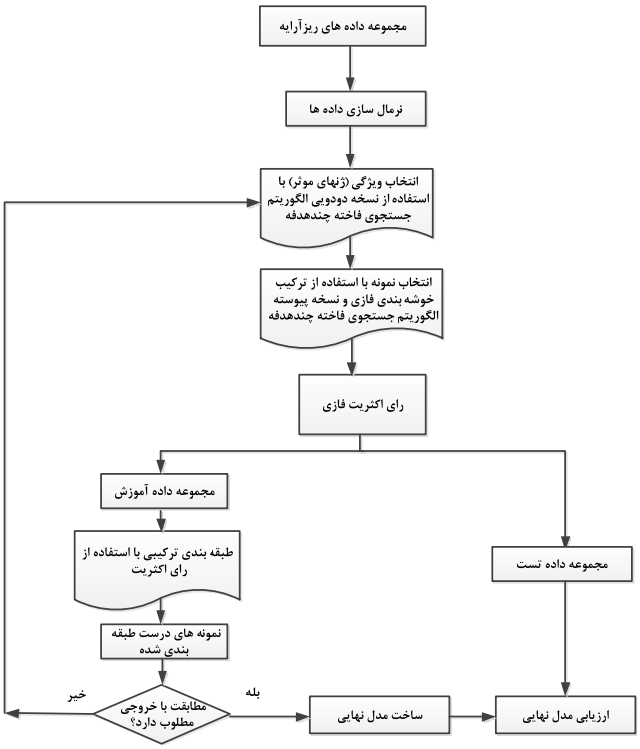
در شروع الگوریتم این لیست خالی بوده و هیچ جوابی در آن قرار ندارد. بعد از اجرای اولین تکرار، تمامی جواب‌ها با هم مقایسه شده و جواب‌هایی که توسط هیچ یک از اعضای جمعیت مغلوب نشده‌اند به این لیست منتقل می‌شوند. در تکرارهای بعدی و با تولید جمعیت جدید، ابتدا تمامی اعضای جدید با هم مقایسه شده و اعضای نامغلوب به لیست آرشیو اضافه می‌شوند. با توجه به این که اعضای جدید منتقل‌شده به لیست آرشیو ممکن است توسط یک یا چند عضو قبلی لیست مغلوب شوند یا بالعکس تعدادی از اعضای قبلی لیست آرشیو، مغلوب این اعضای جدید گردند، می‌بایست در هر تکرار و بعد از اضافه‌شدن اعضای جدید به لیست، لیست آرشیو مجدداً پالایش شده و اعضای مغلوب حذف شوند. این فرایند تا زمانی که شرط خاتمه الگوریتم برآورده نشود تکرار خواهد شد.

موضوع دیگری که می‌بایست مورد توجه قرار گیرد آن است که در الگوریتم‌های بهینه‌سازی تک‌هدفه، بهترین پاسخ که همان پاسخ بهینه است معمولاً منحصربه‌فرد است. بنابراین پارامتر $X_{best}(t)$ در (۱) و $gbest$ در (۳) به ترتیب بهترین جواب الگوریتم در مرحله t و بهترین جواب تا آن مرحله خواهند بود. اما در مسایل بهینه‌سازی چندهدفه همان طور که قبلاً ذکر شد هیچ یک از جواب‌های پارتو بر همدیگر برتری ندارند. سؤالی که مطرح می‌شود این است که برای چندهدفه‌کردن الگوریتم جستجوی فاخته، کدام یک از اعضای موجود در لیست آرشیو می‌بایست جایگزین $X_{best}(t)$ و $gbest$ شوند. در این مقاله برای حفظ تنوع و کیفیت جواب‌ها در هر دو نسخه پیوسته و دودویی الگوریتم، برای انتخاب $X_{best}(t)$ و $gbest$ از مفهوم فاصله ازدحامی استفاده شده است [۴۵]. فاصله ازدحامی مشخص می‌کند مقدار فاصله هر عضو موجود در لیست آرشیو با دو عضو مجاور خود (که آنها نیز در لیست آرشیو هستند) چقدر است. هرچه فاصله ازدحامی مربوط به یک عضو بزرگ‌تر باشد به معنی آن است که در محدوده بزرگ‌تری از آن عضو، جواب بهینه پارتو وجود ندارد. بنابراین آن محدوده جزء محدوده‌های بکر محسوب شده و اعضای بیشتری باید به سمت این محدوده همگرا شوند تا تراکم اعضای پارتو در فضای جواب دارای توزیع یکنواخت باشد.

نکته دیگری که باید مورد توجه قرار گیرد آن است که در مسایل بهینه‌سازی چندهدفه چون حداقل دو تابع هدف وجود دارد، دیگر نمی‌توان در خصوص بهتربودن یک جواب به صورت قطعی نظر داد. برای رفع این مشکل مرحله ۵ شبه‌کد الگوریتم جستجوی فاخته را این گونه اصلاح می‌کنیم که اگر X_j، X_i را مغلوب نماید، جایگزین آن خواهد شد ولی اگر X_i، X_j را مغلوب نماید همان X_i به عنوان عضو بهتر به نسل بعد منتقل می‌شود. اگر هیچ یک از دو راه‌حل X_i و X_j بر همدیگر غلبه نکنند، چون فقط یکی از آنها می‌بایست به نسل بعد منتقل شود با احتمال مساوی یکی از آنها را به صورت تصادفی به نسل بعد منتقل می‌کنیم. با در نظر گرفتن موارد ذکرشده در این بخش، نسخه تک‌هدفه الگوریتم جستجوی خفاش بهبودیافته به نسخه چندهدفه تبدیل می‌شود.

٤- ساختار روش پیشنهادی

شکل ۱ ساختار روش پیشنهادی جهت استخراج ویژگی و طبقه‌بندی داده‌های ریزآرایه را نشان می‌دهد. همان طور که مشاهده می‌شود این ساختار شامل دو فاز انتخاب ویژگی (ژن) و انتخاب نمونه‌های آموزش است. در هر دو فاز از الگوریتم جستجوی فاخته بهبودیافته چندهدفه استفاده شده است. با این تفاوت که در مرحله انتخاب ویژگی، از نسخه دودویی الگوریتم بهره گرفته شده، در حالی که انتخاب نمونه‌های آموزش به کمک نسخه پیوسته الگوریتم انجام گرفته است. در ادامه به اختصار بخش‌های مختلف این ساختار بررسی شده است.



شکل ۱: ساختار روش پیشنهادی.

٤-١ نرمال‌سازی داده‌ها

با توجه به این که واحد اندازه‌گیری ژن می‌تواند بر روی تحلیل داده‌ها اثرگذار باشد و از طرفی بیان یک ویژگی در واحدهای کوچک‌تر، آن ویژگی را دارای وزن بیشتری در تحلیل خواهد کرد، در این مقاله ابتدا قبل از انجام هر پردازشی، داده‌های ریزآرایه با استفاده از (۷) نرمال‌سازی شده‌اند [۱۴]

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \qquad (۷)$$

در رابطه فوق، X و X_{norm} به ترتیب مقدار اصلی و مقدار نرمال‌سازی شده ویژگی (ژن) است. همچنین $\min(X)$ و $\max(X)$ نیز بیانگر حداقل و حداکثر مقدار ویژگی X است.

٤-٢ انتخاب ویژگی با استفاده از نسخه بهبودیافته الگوریتم جستجوی فاخته چندهدفه

با توجه به این که یکی از اهداف الگوریتم‌های انتخاب ویژگی، انتخاب زیرمجموعه‌ای از ژن‌ها است که در عین متمایزبودن دارای حداقل افزونگی نیز باشند، در این مقاله از شاخص بهره اطلاعاتی و مفهوم آنتروپی برای انتخاب ژن‌های مؤثر و متمایز استفاده شده است. بدین منظور پس از نرمال‌سازی داده‌ها، در گام اول با استفاده از (۸)، بهره اطلاعاتی مربوط به هر ژن محاسبه می‌شود [۱۴]

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \qquad (۸)$$

که در رابطه فوق، $IG(S,A)$ بیانگر بهره اطلاعاتی نمونه S برای ویژگی A و $H(S)$ مقدار آنتروپی نمونه‌های S است. $Values(A)$ نیز مجموعه همه ویژگی‌های A است. همچنین S_v زیرمجموعه‌ای از S است که در آن A دارای مقدار v باشد. در رابطه فوق، عبارت اول مقدار آنتروپی ویژگی‌ها و عبارت دوم مقدار آنتروپی مورد انتظار بعد از انتخاب ویژگی‌ها است. اگر ویژگی هدف دارای c مقدار مختلف باشد، آن گاه آنتروپی S نسبت به این دسته‌بندی c گانه به صورت زیر تعریف می‌شود [۱۴]

$$H(S) = -\sum_{i=1}^{c} p_i \times \log_2^{p_i} \qquad (۹)$$

که در آن p_i نسبتی از S است که به کلاس i تعلق دارد. مقادیر بهره اطلاعاتی به دست آمده برای تمامی ویژگی‌ها بر اساس کمترین و بیشترین مقدار بهره اطلاعاتی نرمال‌سازی می‌شوند.

پس از آن می‌بایست یک حد آستانه مناسب برای نگهداری ویژگی‌ها مشخص گردد. در این مقاله جهت تنظیم مقدار آستانه از نسخه دودویی الگوریتم جستجوی فاخته دوهدفه که جزئیات آن در بخش‌های قبل ذکر گردید، استفاده شده است. اهداف مورد نظر که توسط این الگوریتم بهینه می‌شوند عبارتند از: تعداد ویژگی‌ها (تعداد ژن‌ها) که می‌بایست کمینه شود و شاخص F که باید بیشینه گردد. برای تعیین تعداد ویژگی‌ها از آستانه بهره اطلاعاتی و برای محاسبه شاخص F نیز از طبقه‌بند SVM استفاده شده است. با توجه به این که خروجی الگوریتم‌های بهینه‌سازی چندهدفه، مجموعه‌ای از جواب‌های بهینه پارتو است، جهت انتخاب عضو بهینه از بین اعضای پارتو، از دو معیار صحت و سپس فراخوان بهره گرفته‌ایم. اگر بر اساس این دو معیار، تنها یک عضو بهینه وجود داشته باشد، همان عضو انتخاب خواهد شد ولی اگر تعداد اعضای پارتو بیش از یک عضو باشد، جواب نهایی با توجه به معیار فاصله ازدحامی انتخاب می‌شود.

در حل مسأله انتخاب ویژگی با استفاده از الگوریتم‌های بهینه‌سازی هوشمند، روش‌های مختلفی جهت کدگذاری وجود دارد که به تناسب ماهیت مسأله و الگوریتمی که به کار گرفته می‌شود، می‌بایست روش مناسبی انتخاب گردد. در این مقاله هر یک از اعضای جمعیت الگوریتم فاخته به صورت یک بردار دودویی به طول M در نظر گرفته شده‌اند که M تعداد کل ویژگی‌ها است. هر بیت از این بردار نیز نشان‌دهنده یک ژن است. اگر یک ویژگی انتخاب شده باشد، بیت متناظر با آن ویژگی یک و در غیر این صورت صفر خواهد بود. پس از تولید جمعیت اولیه به صورت تصادفی، هر یک از اعضای جمعیت توسط الگوریتم طبقه‌بندی و توابع هدف محاسبه و طی فرایند بهینه‌سازی توسط الگوریتم به روز رسانی خواهند شد. اعضای جمعیت پس از به روز رسانی ارزیابی می‌شوند. این فرایند تا زمانی که شرط خاتمه الگوریتم برآورده نشود تکرار خواهد شد. در اینجا، شرط خاتمه، تکرار الگوریتم به تعداد از پیش تعیین شده در نظر گرفته شده است.

٤-٣ انتخاب نمونه‌ها

یکی دیگر از چالش‌های مطرح در ریزآرایه‌ها، کم‌بودن تعداد نمونه‌ها است. این عامل باعث می‌شود الگوریتم‌های طبقه‌بندی در مرحله آموزش، با تعداد نمونه‌های کافی آموزش نبینند و در نتیجه در مرحله تشخیص، نتوانند نمونه‌های جدید را به خوبی از هم متمایز کنند [۴۶]. در این مقاله برای انتخاب نمونه‌های مناسب جهت آموزش طبقه‌بندها، از ترکیب خوشه‌بندی فازی و نسخه پیوسته الگوریتم جستجوی فاخته چندهدفه استفاده شده است. توابع هدفی که توسط این الگوریتم بهینه می‌شوند عبارتند از شاخص فشرده‌سازی سراسری خوشه‌ها (π) که می‌بایست کمینه شود و شاخص جداسازی فازی (sep) که می‌بایست بیشینه گردد. این دو شاخص توسط روابط ریاضی زیر بیان می‌شوند [۴۷]

$$\pi = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} u_{ik}^{m} D(s_i, x_k)}{\sum_{k=1}^{n} u_{ik}} \quad (۱۰)$$

$$sep = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} \mu_{ij}^{m} D(s_i, s_j) \quad (۱۱)$$

در روابط فوق، $D(s_i, x_k)$ فاصله بین داده x_k از مرکز خوشه s_i، u_{ik} میزان تعلق داده kام (یعنی x_k) به خوشه iام، μ_{ij}^{m} میزان تعلق s_j به s_i و m نیز یک عدد حقیقی بزرگ‌تر از یک است که در اکثر پژوهش‌ها مقدار آن ۲ در نظر گرفته می‌شود. همچنین n تعداد کل داده‌ها و c نیز تعداد خوشه‌ها است.

در این مقاله با توجه به پیوسته‌بودن فضای جستجو، برای حل مسأله خوشه‌بندی به کمک الگوریتم جستجوی فاخته، از روش کدگذاری مبتنی بر مراکز با داده‌های اعشاری استفاده شده است. از طرفی چون تعداد خوشه‌ها برابر با تعداد کلاس‌های ریزآرایه است و تعداد کلاس‌ها نیز از قبل مشخص و ثابت می‌باشد، برای کد گذاری از روش کدگذاری با طول ثابت استفاده شده است. در کدگذاری با طول ثابت، تمام راه‌حل‌های خوشه‌بندی، دارای تعداد از پیش تعیین شده K سرخوشه هستند. از این رو برای هر مجموعه داده ریزآرایه، اندازه تمام اعضای جمعیت در طول فرایند بهینه‌سازی یکسان و به اندازه $D\times K$ خواهد بود که D بیانگر ابعاد یا تعداد ویژگی‌های مجموعه داده است. در الگوریتم جستجوی فاخته نیز همانند سایر الگوریتم‌های بهینه‌سازی هوشمند، ابتدا جمعیت اولیه به صورت تصادفی ایجاد می‌شود. در این الگوریتم هر عضو جمعیت که یک تخم نامیده می‌شود بیانگر یک راه‌حل برای مسأله است. هر راه‌حل نیز شامل تعدادی متغیر تصمیم‌گیری است که در اینجا به دلیل استفاده از روش رمزگذاری مبتنی بر مرکز، متغیرهای تصمیم‌گیری شامل ویژگی‌های سرخوشه‌ها خواهد بود. به عبارت دیگر، هر تخم به عنوان یک راه‌حل، مختصات K سرخوشه را در خود دارد. پس از تولید جمعیت اولیه، اعضای جمعیت می‌بایست توسط توابع هدف مورد ارزیابی قرار گیرند. توابع هدف مورد استفاده در این مقاله همان توابع هدف معرفی‌شده در (۱۰) و (۱۱) هستند. برای ارزیابی الگوریتم خوشه‌بندی از دو شاخص Silhouette و ARI[1] استفاده شده است. محدوده تغییرات هر دو شاخص عددی بین [-۱،۱] است. در شاخص Silhouette هرچه مقدار شاخص به یک نزدیک‌تر باشد، خوشه فشرده‌تر و دورتر از سایر خوشه‌ها است. منفی‌بودن این شاخص نیز به این معنا است که داده‌ها به اشیای خوشه دیگری غیر از خوشه‌ای که به آن تعلق دارند نزدیک‌تر هستند. شاخص ARI نیز تعیین‌کننده میزان تطابق خوشه‌های واقعی با خوشه‌های تخمینی است. برای این شاخص، عدد یک نشان‌دهنده هم‌پوشانی کامل خوشه‌های واقعی با خوشه‌های تخمینی است. همچنین مقادیر ۰ و ۱-، عملکرد بسیار ضعیف روش خوشه‌بندی را نشان می‌دهد. روابط ریاضی و جزئیات بیشتری از این دو شاخص در [۴۸] آورده شده است.

٤-٤ رأی اکثریت فازی و ساخت مدل نهایی

همان طور که در بخش‌های قبل نیز اشاره شد با توجه به ماهیت داده‌های ریزآرایه، در این پژوهش از نسخه بهبودیافته الگوریتم جستجوی فاخته چندهدفه برای یافتن مراکز خوشه‌ها و از تکنیک رأی اکثریت فازی برای تقسیم‌بندی داده‌های آموزش و تست استفاده شده است. پس از آن که به کمک الگوریتم جستجوی فاخته چندهدفه، مراکز خوشه‌ها تعیین شدند با استفاده از معیار شباهت فازی برای هر نمونه یک درجه عضویت محاسبه می‌شود که از آن می‌توان برای تعیین وزن هر نمونه در خوشه فازی استفاده کرد. وزن نمونه x_k در خوشه s_i از رابطه زیر به دست می‌آید [۴۸]

$$w_i(x_k) = \frac{u_{ik}}{\sum_{j=1}^{n} u_{ij}} \quad (۱۲)$$

در نهایت به ازای هر خوشه فازی مانند خوشه iام، یک مجموعه وزن W_i برای نمونه‌های آموزشی به دست می‌آید به طوری که [۴۸]

$$W_i = \{w_i(x_k) \,|\, \forall x_k \in X\} \quad (۱۳)$$

چون درجه عضویت یک نمونه به یک خوشه فازی بیان‌کننده میزان شباهت آن نمونه به نمونه‌های موجود در خوشه مورد نظر است، بنابراین می‌توان از این درجه عضویت برای وزن‌دهی به تصمیم طبقه‌بند متناظر با آن خوشه استفاده کرد. انتخاب تعداد نمونه‌های آموزش و تست وابسته به دو پارامتر σ و γ است. پارامتر σ نشان‌دهنده آستانه درجه عضویت است. بر این اساس مجموعه داده‌هایی برای آموزش طبقه‌بند انتخاب می‌شوند که درجه عضویت آنها بیشتر از این مقدار باشد. پارامتر γ نیز بیانگر آستانه رأی اکثریت فازی است. این پارامتر تعیین می‌کند که حداقل چه تعداد از راه‌حل‌های ارائه‌شده باید در رأی‌گیری فازی با هم توافق داشته باشند [۳۹]. اگر پارامترهای σ و γ افزایش یابند، تعداد داده‌های مجموعه آموزش کاهش خواهند یافت. این موضوع به این معنی است که تعداد بیشتری از راه‌حل‌های ارائه‌شده با یکدیگر هم‌خوانی دارند. کاهش پارامترهای σ و γ نیز به منزله افزایش حجم داده‌های آموزش است. این امر نشان‌دهنده آن است که تعداد کمتری از نمونه‌ها، به نامناسب‌بودن بعضی از داده‌ها در میان خود توافق دارند و به همین دلیل این داده‌ها در مجموعه آموزش قرار می‌گیرند. برای ایجاد مصالحه بین داده‌های آموزش و تست، مقادیر این دو پارامتر معمولاً در محدوده ۰/۵ تا ۰/۷ در نظر گرفته می‌شوند [۴۹] که در مقاله مقادیر هر دو پارامتر برابر ۰/۶۵ انتخاب شده است.

پس از آموزش طبقه‌بندها جهت پیش‌بینی یک نمونه جدید، ابتدا درجه عضویت و کلاس آن نمونه به ترتیب در خوشه‌ها و طبقه‌بندهای متناظر با خوشه‌ها محاسبه می‌شود. سپس با استفاده از قاعده رأی اکثریت فازی، تصمیم طبقه‌بندها با هم ترکیب و نتیجه نهایی تعیین می‌شود به طوری که [۴۹]

$$V(x_k) = \arg\max \sum_{i=1}^{K} u_{ik} . I(y == h_i(x_k)) \quad (۱۴)$$

در رابطه فوق، $V(x_k)$ رأی اکثریت گروه به ازای نمونه x_k و $h_i(x_k)$ تصمیم طبقه‌بند متناظر با خوشه iام به ازای x_k است. مقدار $I(*)$ نیز در صورتی که شرط * درست باشد، برابر یک و در غیر این صورت برابر صفر خواهد بود.

در این مقاله از چهار طبقه‌بند ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایه (KNN)، نایو بیز (NB) و طبقه‌بند ترکیبی استفاده شده است. جزئیات بیشتری از طبقه‌بندهای فوق در [۵۰] آورده شده است. برای ارزیابی کارایی طبقه‌بندهای مورد استفاده در این پژوهش نیز از شاخص صحت و F-measure بهره گرفته‌ایم.

1. Adjusted Rand Index

جدول ۱: مشخصات مجموعه داده‌ها [۵۱].

Dataset	No. of genes	No. of classes	Class Lables	No. of samples
Colon	۲۰۰۰	۲	Tumor	۲۲
			Normal	۴۰
Leukemia	۷۱۲۹	۲	ALL	۴۷
			AML	۲۵
Ovarian	۱۵۱۵۴	۲	Cancer	۱۶۲
			Normal	۹۱
Breast	۲۴۴۸۱	۲	Relapse	۴۶
			Non-Relapse	۵۱
ALL-AML-۳	۷۱۲۹	۳	B-Cell	۳۸
			T-Cell	۹
			AML	۲۵
SRBCT	۲۳۰۸	۴	EWS	۲۹
			BL	۱۱
			NBL	۱۸
			RMS	۲۵

٤-۵ اعتبارسنجی مدل

در مبحث یادگیری ماشین برای سنجش کارایی یک مدل، غالباً از روش اعتبارسنجی متقابل $k-\text{fold}$ استفاده می‌شود. در این روش معمولاً داده‌ها به دو قسمت تقسیم می‌شوند، یک قسمت داده‌های آموزش و دیگری داده‌های تست. از داده‌های آموزش برای ایجاد مدل و از داده‌های تست برای بررسی کارایی آن استفاده می‌شود. اصول حاکم بر این روش بدین صورت است که ابتدا داده‌ها به k زیرمجموعه افراز می‌شوند. از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و $k-۱$ تای دیگر برای آموزش به کار می‌روند. این فرایند k بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی به کار می‌روند. در نهایت میانگین نتیجه این k بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود. در این پژوهش از روش اعتبارسنجی ۳-fold استفاده شده است.

۵- شبیه‌سازی و تحلیل نتایج

برای ارزیابی عملکرد الگوریتم‌های بهینه‌سازی هوشمند، شاخص‌های عمومی مختلفی وجود دارند که مهم‌ترین آنها شاخص بهترین جواب، سرعت همگرایی و نرخ موفقیت است. نرخ موفقیت طبق تعریف عبارت است از نسبت تعداد دفعاتی که الگوریتم به یک مقدار از پیش تعیین شده همگرا می‌شود به تعداد کل اجراها. به عنوان مثال اگر از ۱۰ بار اجرای یک الگوریتم، ۸ بار آن، الگوریتم به مقدار بهینه از پیش تعیین شده همگرا شود می‌گوییم نرخ موفقیت آن الگوریتم برابر ۸۰٪ است.

عملکرد الگوریتم جستجوی فاخته بهبودیافته با نسخه اصلی آن روی ۵ تابع استاندارد Rosenbrock، Rastrigin، Schwefel، Griewangk و Ackley مورد ارزیابی قرار گرفت. این توابع جزء توابع استانداردی هستند که برای ارزیابی کارایی الگوریتم‌های بهینه‌سازی هوشمند مورد استفاده قرار می‌گیرند. میانگین نرخ موفقیت نسخه اصلی الگوریتم در مقایسه با نسخه بهبودیافته آن به ازای بیست بار اجرا به ترتیب ۸۸٪ در مقابل ۹۶٪ به دست آمد. همچنین به طور میانگین در ۹۲٪ اجراها، بهترین جواب حاصل از نسخه بهبودیافته الگوریتم، بهینه‌تر از پاسخ حاصل از نسخه اصلی الگوریتم بود. سرعت همگرایی الگوریتم را می‌توان به نوعی متناظر با تعداد دفعات ارزیابی تابع هدف برای رسیدن به یک مقدار از پیش تعیین شده دانست. هرچه تعداد دفعات ارزیابی تابع هدف کمتر باشد، سرعت همگرایی الگوریتم نیز بیشتر خواهد بود. نتایج ارزیابی‌ها نشان دادند در ۹۶٪ اجراها، سرعت همگرایی نسخه بهبودیافته الگوریتم، بیشتر از نسخه استاندارد آن است که این خود نشان‌دهنده کارایی راهکار ارائه‌شده جهت ایجاد مصالحه بین دو مؤلفه اکتشاف و استخراج در نسخه بهبودیافته الگوریتم جستجوی فاخته است.

در ادامه نتایج حاصل از روش پیشنهادی بر روی شش مجموعه از داده‌های ریزآرایه که در دیگر مقالات به کار گرفته شده‌اند، ارائه و با سایر کارهای انجام‌شده مقایسه گردیده است. مشخصات این ۶ مجموعه داده در جدول ۱ آورده شده است [۵۱].

پارامترهای الگوریتم جستجوی فاخته بهبودیافته به شرح زیر انتخاب شده‌اند

$$N_p = ۵۰$$
$$MaxGen = ۱۰۰$$
$$\alpha = ۰/۲$$
$$\beta_{\min} = ۰/۶$$
$$\beta_{\max} = ۱/۵$$

که N_p تعداد اعضای جمعیت و $MaxGen$ حداکثر تعداد تکرار یا همان شرط خاتمه الگوریتم است. ظرفیت لیست آرشیو نیز ۵۰۰ انتخاب شده است.

در این مقاله نتایج حاصل از شبیه‌سازی‌ها به ازای چهار حالت زیر مورد مطالعه قرار گرفته است:

الف) بدون انتخاب نمونه و انتخاب ویژگی (Non-Pre): در این حالت تمامی داده‌های هر مجموعه ریزآرایه برای ایجاد طبقه‌بندها مورد استفاده قرار داده شده‌اند و به جز نرمال‌سازی داده‌ها، هیچ پردازش دیگری از جمله انتخاب ویژگی یا انتخاب نمونه بر روی داده‌ها صورت نگرفته است.

ب) انتخاب ویژگی (FS): در این حالت ابتدا داده‌های هر یک از ریزآرایه‌ها به عنوان ورودی الگوریتم جستجوی فاخته چندهدفه دودویی در نظر گرفته شده‌اند. خروجی این الگوریتم برای هر مجموعه داده، انتخاب تعدادی ویژگی از بین کل ویژگی‌ها است. سپس داده‌ها بر اساس ویژگی‌های انتخاب‌شده به طبقه‌بندها اعمال می‌شوند.

ج) انتخاب نمونه (IS): در این حالت ابتدا هر یک از مجموعه داده‌های ریزآرایه به عنوان ورودی الگوریتم جستجوی فاخته چندهدفه پیوسته در نظر گرفته می‌شوند. سپس بر اساس رأی اکثریت فازی تعدادی از نمونه‌ها که می‌توانند باعث بهبود در عملکرد طبقه‌بندها در مرحله آموزش شوند، انتخاب می‌گردند. بعد از انتخاب نمونه‌های آموزش و تست، این نمونه‌ها به طبقه‌بندها اعمال می‌گردند.

د) انتخاب ویژگی و انتخاب نمونه (FSIS): در این حالت ابتدا داده‌های ریزآرایه به الگوریتمی که برای انتخاب ویژگی پیشنهاد شده است وارد می‌شوند و در مرحله بعد داده‌ها با توجه به ویژگی‌های انتخاب‌شده در مرحله انتخاب ویژگی، وارد الگوریتمی می‌شوند که برای انتخاب نمونه‌های آموزشی و تست پیشنهاد گردیده است. بعد از گذراندن این مراحل، داده‌های انتخاب‌شده جهت آموزش با ویژگی‌های انتخاب‌شده به طبقه‌بندهای مورد نظر اعمال می‌گردند.

برای بررسی نتایج حاصل از چهار حالت فوق، از دو شاخص صحت و

جدول ۲: کارایی طبقه‌بندها بر روی مجموعه داده‌های ریزآرایه.

Datasets	No. of features	No. of samples	Accuracy%				F-measure%			
			SVM	KNN	NB	MV	SVM	KNN	NB	MV
Colon										
Non-Pre	۲۰۰۰	۶۲	۲٫۰۸±۷۸٫۸۴	۲٫۴۱±۷۷٫۲۳	۲٫۵۳±۷۶٫۶۶	۱٫۸۶±۸۳٫۴۲	۱٫۹۴±۶۵٫۴۳	۲٫۰۷±۶۴٫۹۶	۱٫۵۴±۶۴٫۸۴	۱٫۱۳±۷۸٫۹۱
FS	۰٫۶۹±۴٫۴۰	۶۲	۰٫۹۸±۸۹٫۵۳	۰٫۷۶±۹۲٫۸۶	۰٫۹۳±۹۳٫۱۲	۰٫۴۲±۹۶٫۵۶	۰٫۶۷±۸۴٫۱۵	۰٫۵۸±۸۶٫۷۲	۰٫۷۱±۸۸٫۶۴	۰٫۳۵±۹۰٫۲۹
IS	۲۰۰۰	۳٫۰۱±۳۷٫۲۰	۱٫۰۶±۸۵٫۹۱	۱٫۱۸±۸۵٫۸۲	۱٫۱۳±۸۶٫۰۹	۰٫۸۳±۹۲٫۵۹	۱٫۲۳±۷۲٫۱۶	۱٫۰۳±۸۱٫۴۷	۱٫۰۴±۸۵٫۹۷	۰٫۸۷±۸۹٫۸۴
FSIS	۰٫۴۸±۴٫۳۰	۲٫۱۳±۴۰٫۹۰	۰٫۶۲±۹۱٫۳۴	۰٫۴۹±۹۳٫۳۲	۰٫۷۴±۹۶٫۰۳	۰٫۱۶±۹۹٫۲۴	۰٫۵۳±۸۴٫۷۵	۰٫۲۴±۸۵٫۱۵	۰٫۴۹±۸۸٫۱۲	۰٫۳۸±۹۰٫۰۷
Leukemia										
Non-Pre	۷۱۲۹	۷۲	۱٫۹۳±۹۴٫۲۴	۲٫۰۳±۹۰٫۱۱	۱٫۸۲±۹۳٫۴۷	۰٫۵۶±۹۷٫۱۵	۱٫۲۶±۸۳٫۱۳	۱٫۵۷±۷۸٫۲۳	۱٫۵۹±۸۰٫۳۲	۰٫۸۷±۸۵٫۱۴
FS	۰٫۷۰±۴٫۵۰	۷۲	۰٫۸۴±۹۶٫۵۸	۰٫۶۳±۹۵٫۰۴	۰٫۵۳±۹۷٫۳۶	۰٫۳۱±۹۹٫۳۷	۰٫۷۳±۸۳٫۶۴	۰٫۸۴±۸۲٫۳۵	۰٫۷۳±۸۲٫۷۴	۰٫۲۸±۹۳٫۶۴
IS	۷۱۲۹	۱٫۷۶±۵۰٫۳۰	۰٫۳۸±۹۴٫۹۳	۰٫۴۳±۹۱٫۱۶	۰٫۳۷±۹۵٫۱۸	۰٫۲۶±۹۷٫۱۳	۰٫۶۱±۸۲٫۷۸	۰٫۳۹±۷۸٫۷۳	۰٫۶۱±۸۰٫۲۶	۰٫۵۸±۸۹٫۳۲
FSIS	۰٫۵۱±۴٫۳۰	۱٫۵۲±۶۰٫۱۰	۰٫۰۹±۹۹٫۰۷	۰٫۲۱±۹۵٫۸۸	۰٫۱۷±۹۸٫۱۲	۰٫۱۴±۹۹٫۸۷	۰٫۲۱±۸۸٫۱۵	۰٫۱۶±۸۱٫۵۷	۰٫۰۹±۸۷٫۳۴	۰٫۲۵±۹۳٫۱۲
Ovarian										
Non-Pre	۱۵۱۵۴	۲۵۳	۱٫۸۷±۸۶٫۲۳	۱٫۷۶±۸۸٫۸۴	۱٫۶۴±۸۱٫۶۷	۱٫۴۳±۸۹٫۰۸	۱٫۳۸±۷۸٫۳۹	۱٫۵۲±۷۸٫۶۳	۱٫۷۳±۷۱٫۲۵	۱٫۳۲±۷۸٫۸۱
FS	۰٫۸۲±۴٫۳۰	۲۵۳	۰٫۶۸±۹۳٫۶۷	۱٫۱۴±۹۳٫۰۶	۰٫۹۴±۹۲٫۱۲	۰٫۳۴±۹۶٫۳۴	۰٫۵۴±۸۳٫۴۲	۱٫۱۲±۸۲٫۹۸	۰٫۸۸±۸۲٫۳۸	۰٫۴۷±۸۴٫۵۶
IS	۱۵۱۵۴	۳٫۹۴±۱۶۹	۱٫۱۷±۹۰٫۱۶	۱٫۶۸±۸۸٫۷۶	۰٫۹۷±۸۳٫۱۷	۰٫۹۱±۹۴٫۰۲	۱٫۰۸±۷۸٫۷۵	۱٫۲۳±۷۷٫۹۶	۰٫۵۶±۷۳٫۵۲	۰٫۶۳±۸۲٫۲۸
FSIS	۰٫۷۸±۳٫۱	۱٫۲۵±۱۸۳٫۷۰	۰٫۳۱±۹۶٫۸۹	۰٫۳۷±۹۵٫۱۴	۰٫۲۸±۹۴٫۱۶	۰٫۱۲±۹۸٫۵۳	۰٫۲۴±۸۷٫۳۶	۰٫۴۱±۸۵٫۷۲	۰٫۳۷±۸۳٫۹۵	۰٫۱۷±۸۹٫۲۳
Breast										
Non-Pre	۲۴۴۸۱	۹۷	۱٫۶۴±۸۵٫۱۱	۲٫۱۶±۷۹٫۳۶	۱٫۹۳±۷۶٫۲۳	۰٫۹۸±۸۶٫۱۸	۱٫۱۸±۷۱٫۱۴	۱٫۹۷±۶۵٫۸۶	۲٫۱۶±۶۱٫۵۷	۰٫۳۵±۷۱٫۹۸
FS	۰٫۶۹±۱۳٫۴	۹۷	۰٫۲۷±۸۸٫۰۶	۰٫۳۶±۸۶٫۹۱	۰٫۶۳±۸۴٫۷۸	۰٫۲۲±۹۰٫۷۱	۰٫۴۸±۷۲٫۱۵	۰٫۷۶±۷۰٫۳۳	۰٫۹۳±۶۸٫۴۹	۰٫۷۳±۸۰٫۴۶
IS	۲۴۴۸۱	۰٫۷۰±۷۱٫۵۰	۱٫۱۴±۸۶٫۱۴	۱٫۴۷±۸۰٫۹۵	۱٫۱۸±۸۰٫۳۸	۰٫۷۵±۸۷٫۱۴	۱٫۶۷±۷۱٫۴۳	۱٫۰۶±۶۶٫۱۵	۱٫۵۸±۶۰٫۷۲	۰٫۹۳±۷۲٫۳۱
FSIS	۰٫۸۴±۱۳٫۶	۰٫۸۲±۷۸٫۷۰	۰٫۱۲±۸۸٫۲۴	۰٫۱۹±۸۷٫۸۸	۰٫۱۶±۸۶٫۴۲	۰٫۴۸±۹۳٫۱۶	۰٫۳۶±۷۴٫۸۱	۰٫۵۴±۷۴٫۲۳	۰٫۱۱±۷۲٫۳۹	۰٫۱۷±۸۲٫۷۳
ALL-AML-۳										
Non-Pre	۷۱۲۹	۷۲	۲٫۳۱±۹۱٫۱۴	۱٫۷۹±۹۲٫۶۸	۲٫۱۶±۹۰٫۶۹	۱٫۱۸±۹۳٫۶۴	۱٫۸۳±۷۹٫۰۵	۱٫۱۳±۷۹٫۵۳	۱٫۶۸±۷۸٫۴۲	۱٫۱۲±۷۹٫۸۸
FS	۰٫۸۲±۶٫۳۰	۷۲	۰٫۸۷±۹۴٫۹۲	۰٫۹۶±۹۳٫۴۷	۰٫۶۷±۹۵٫۱۴	۰٫۳۱±۹۷٫۸۴	۰٫۵۳±۸۸٫۴۵	۱٫۱۵±۸۷٫۸۲	۰٫۹۴±۸۹٫۰۸	۰٫۱۴±۹۱٫۰۵
IS	۷۱۲۹	۰٫۸۴±۵۳٫۴	۱٫۳۶±۹۱٫۸۸	۱٫۴۲±۹۳٫۱۴	۱٫۶۷±۹۰٫۹۳	۰٫۹۷±۹۴٫۱۶	۱٫۲۸±۷۹٫۶۳	۱٫۲۳±۷۹٫۷۷	۱٫۰۶±۷۸٫۸۴	۰٫۵۸±۸۰٫۱۳
FSIS	۰٫۴۸±۶٫۳۰	۰٫۸۲±۶۴٫۷۰	۰٫۷۱±۹۶٫۱۷	۰٫۸۸±۹۵٫۲۷	۰٫۱۸±۹۵٫۵۴	۰٫۳۵±۹۸٫۸۹	۰٫۳۱±۹۰٫۱۶	۰٫۲۶±۸۹٫۵۱	۰٫۲۷±۸۹٫۵۳	۰٫۲۷±۹۱٫۶۷
SRBCT										
Non-Pre	۲۳۰۸	۸۳	۱٫۷۴±۸۷٫۶۶	۱٫۶۱±۹۰٫۴۲	۲٫۰۸±۸۸٫۰۷	۱٫۲۷±۹۱٫۲۲	۱٫۲۱±۷۱٫۳۴	۱٫۱۴±۷۲٫۰۵	۱٫۵۴±۷۱٫۵۳	۱٫۰۲±۷۲٫۸۴
FS	۰٫۶۹±۵٫۴۰	۸۳	۰٫۳۲±۹۴٫۱۷	۰٫۴۲±۹۳٫۶۶	۰٫۲۷±۹۵٫۴۳	۰٫۱۸±۹۸٫۲۴	۰٫۴۱±۷۸٫۳۴	۰٫۶۱±۷۷٫۸۸	۰٫۳۸±۷۸٫۵۶	۰٫۲۳±۷۹٫۴۷
IS	۲۳۰۸	۰٫۹۴±۶۶٫۷۰	۱٫۰۷±۸۸٫۶۱	۰٫۷۷±۹۱٫۶۳	۱٫۱۴±۹۰٫۱۳	۰٫۷۸±۹۲٫۵۶	۰٫۸۵±۷۲٫۰۵	۰٫۳۷±۷۲٫۷۴	۱٫۰۵±۷۲٫۱۸	۰٫۳۸±۷۶٫۴۵
FSIS	۰٫۵۱±۵٫۴۰	۱٫۰۵±۷۹٫۷۰	۰٫۱۴±۹۵٫۶۱	۰٫۱۸±۹۴٫۱۹	۰٫۱۳±۹۶٫۲۷	۰٫۰۷±۹۸٫۶۳	۰٫۵۳±۷۸٫۸۴	۰٫۴۳±۷۸٫۷۵	۰٫۲۷±۷۹٫۱۱	۰٫۲۴±۸۴٫۰۳

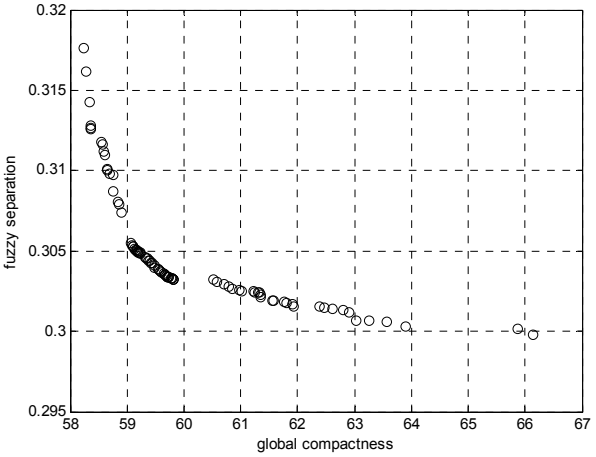
F-measure و برای اعتبارسنجی نتایج نیز از روش اعتبارسنجی ۳–fold استفاده شده است.

با توجه به این که الگوریتم‌های بهینه‌سازی هوشمند دارای ماهیت تصادفی هستند، جهت بررسی مقاوم‌بودن آنها و همچنین افزایش قابلیت اطمینان نتایج، در تمامی شبیه‌سازی‌ها از میانگین حاصل از ۱۰ بار اجرای مستقل برای ارزیابی نتایج استفاده شده است.

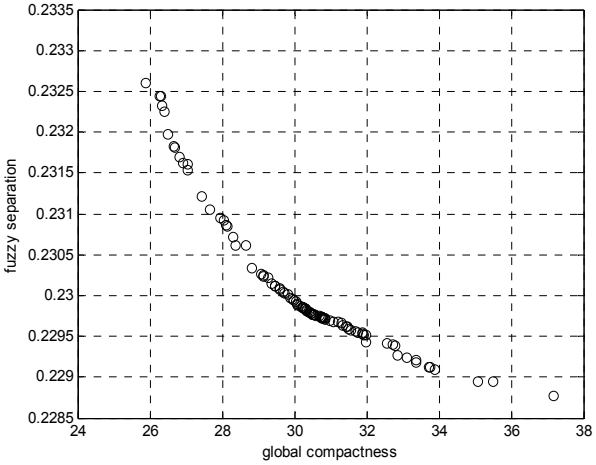
نتایج حاصل از چهار حالت فوق برای سه طبقه‌بند پایه و طبقه‌بند ترکیبی در جدول ۲ آورده شده است. با توجه به نتایج این جدول مشاهده می‌شود برای هر شش مجموعه داده ریزآرایه، از نظر هر دو شاخص صحت و F-measure، کارایی کلیه طبقه‌بندها در دو حالت FS و FSIS نسبت به حالت Non-Pre بهبود پیدا کرده است. دلیل این امر را می‌توان استفاده از الگوریتم جستجوی فاخته اصلاح‌شده دوهدفه در انتخاب ویژگی‌های مؤثر بر تشخیص بیماری دانست. به عبارت دیگر الگوریتم جستجوی فاخته توانسته است از بین تمام ویژگی‌ها، ویژگی‌های نویزی یا غیر مرتبط را حذف کند و ویژگی‌هایی که بیشترین تأثیر را بر تشخیص بیماری دارند، شناسایی و انتخاب نماید.

در مورد داده‌های ریزآرایه، با توجه به کم‌بودن تعداد نمونه‌ها نسبت به تعداد ویژگی‌ها، حساسیت در انتخاب بهترین داده‌ها جهت آموزش طبقه‌بندها نیز بیشتر می‌شود چرا که اگر طبقه‌بندها با داده‌های ضعیف آموزش ببینند، کارایی الگوریتم نیز تحت تأثیر قرار خواهد گرفت. همان طور که از جدول ۲ مشاهده می‌شود کارایی طبقه‌بندها در حالت IS نیز نسبت به حالت Non-Pre بهبود یافته است.

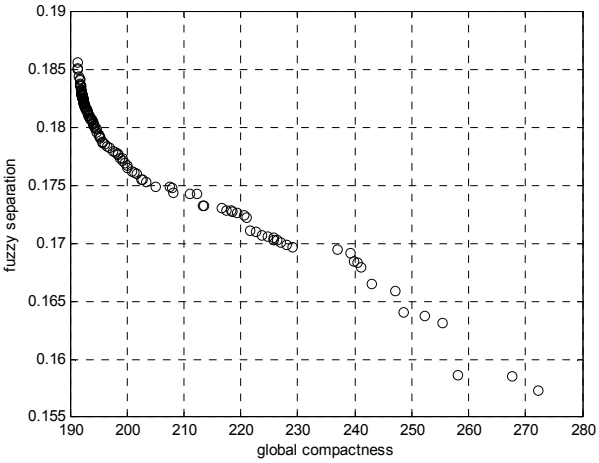
نتیجه دیگری که از جدول ۲ قابل استخراج است این است که از بین دو حالت FS و IS، کارایی طبقه‌بندها در حالت FS به مراتب بهتر از حالت IS است. به طور نمونه، کمترین و بیشترین درصد بهبودی حاصل‌شده برای شاخص صحت در حالت IS نسبت به حالت Non-Pre به ترتیب برابر ۰٫۹۸٪ برای مجموعه داده Leukemia و ۸٫۷۳٪ برای مجموعه داده Colon است. این در حالی است که کمترین و بیشترین درصد بهبودی حاصل‌شده برای حالت FS نسبت به حالت پایه به ترتیب برابر ۳٫۲۲٪ (مجموعه داده Leukemia) و ۱۳٫۱۴٪ (مجموعه داده Colon) می‌باشد.



شکل ۲: جواب‌های بهینه پارتو در حالت FSIS در مجموعه داده Colon.



شکل ۳: جواب‌های بهینه پارتو در حالت FSIS در مجموعه داده Leukemia.



شکل ۴: جواب‌های بهینه پارتو در حالت FSIS در مجموعه داده Ovarian.

به عبارت دیگر در مورد ریزآرایه‌ها، از بین دو پردازش انتخاب ویژگی و انتخاب نمونه‌های آموزش، تأثیر انتخاب ویژگی بر بهبود کارایی طبقه‌بندها، بیشتر از انتخاب نمونه‌های آموزش است. این نتیجه دقیقاً منطبق بر نتیجه‌گیری ارائه‌شده در [۵۲] است.

همچنین بر اساس نتایج ذکرشده در جدول ۲ مشاهده می‌شود برای هر شش مجموعه ریزآرایه، در هر چهار حالت مورد مطالعه، طبقه‌بند ترکیبی توانسته در مقایسه با طبقه‌بندهای پایه به طور مناسبی هر دو شاخص صحت و F-measure را بهبود بخشد. به طور نمونه در حالت FSIS بیشترین مقادیر به دست آمده برای شاخص‌های صحت و F-measure

جدول ۳: میانگین صحت حاصل از الگوریتم‌های طبقه‌بندی به ازای تمامی مجموعه داده‌ها (٪).

	Non-pre	FS	IS	FSIS
SVM	۱٫۹۲±۸۷٫۲۰	۰٫۶۶±۹۲٫۸۲	۱٫۰۳±۸۹٫۶۰	۰٫۳۳±۹۴٫۵۵
KNN	۱٫۱۹۶±۸۶٫۴۴	۰٫۷۱۱±۹۲٫۵۰	۱٫۱۵±۸۸٫۵۷	۰٫۳۸±۹۳٫۶۱
NB	۲٫۰۲±۸۴٫۴۶	۰٫۶۶±۹۲٫۹۹	۱٫۰۷±۸۷٫۶۴	۰٫۲۷±۹۴٫۴۲
MV	۱٫۲۱±۸۹٫۹۴	۰٫۲۹±۹۶٫۵۸	۰٫۷۵±۹۲٫۸۶	۰٫۲۲±۹۷٫۹۲

جدول ۴: میانگین F-MEASURE حاصل از الگوریتم‌های طبقه‌بندی به ازای تمامی مجموعه داده‌ها (٪).

	Non-pre	FS	IS	FSIS
SVM	۱٫۴۶±۷۴٫۷۴	۰٫۵۶±۸۱٫۶۹	۱٫۱۲±۷۶٫۱۳	۰٫۳۷±۸۴٫۰۱
KNN	۱٫۵۶±۷۳٫۲۱	۰٫۸۴±۸۱٫۳۴	۰٫۸۹±۷۶٫۱۴	۰٫۳۴±۸۲٫۴۸
NB	۱٫۷۰±۷۱٫۳۲	۰٫۷۷±۸۱٫۶۵	۰٫۹۸±۷۵٫۲۴	۰٫۲۶±۸۳٫۴۰
MV	۰٫۹۷±۷۷٫۹۲	۰٫۳۶±۸۶٫۵۷	۰٫۶۵±۸۱٫۷۲	۰٫۲۵±۸۸٫۴۷

به ترتیب ۹۹٫۸۷ و ۹۳٫۱۲ درصد است که مربوط به مجموعه داده Leukemia می‌باشد. همچنین کمترین مقادیر این دو شاخص نیز به ترتیب ۹۳٫۱۶ و ۸۲٫۷۳ درصد می‌باشد که مربوط به مجموعه داده Breast است. در مورد سایر مجموعه داده‌ها نیز این دو شاخص در محدوده مقادیر فوق هستند. نتیجه دیگری که از جدول ۲ به دست می‌آید این است که علی‌رغم آن که کارایی طبقه‌بندها در حالت FSIS نسبت به حالت Non-Pre بهبود قابل ملاحظه‌ای یافته است، اما تفاوت محسوسی بین مقادیر حاصل از دو شاخص صحت و F-measure وجود دارد، به گونه‌ای که برای برخی از مجموعه داده‌ها این تفاوت به بالای ۱۰ درصد نیز می‌رسد. یکی از دلایل این موضوع می‌تواند عدم توازن بین داده‌های ریزآرایه باشد. به طور معمول عدم توازن بین مجموعه داده‌های ریزآرایه باعث تمایل الگوریتم‌های طبقه‌بندی به سمت کلاس اکثریت می‌شود [۵۳].

به طور مشخص از نظر تعداد ویژگی، الگوریتم پیشنهادی توانسته است در مقایسه با حالت Non-Pre، با انتخاب مجموعه بسیار کوچک‌تری از ژن‌های حاوی اطلاعات، کارایی طبقه‌بندها را از نظر شاخص‌های صحت و F-measure بهبود بخشد. به طور نمونه برای مجموعه داده Colon در عین این که در حالت Non-Pre نسبت به حالت FSIS، شاخص‌های صحت و F-measure به ترتیب از ۸۳٫۴۲ و ۷۸٫۹۱ درصد به ۹۹٫۲۴ و ۹۰٫۰۷ درصد افزایش یافته‌اند، تعداد ژن‌ها نیز از ۲۰۰۰ ژن به حدود ۴٫۳ ژن کاهش پیدا کرده است. به عبارت دیگر مدل پیشنهادی توانسته است با انتخاب مجموعه کوچک‌تری از ژن‌های متمایز، منجر به افزایش کارایی طبقه‌بندها شود. این موضوع در مورد سایر مجموعه داده‌های ریزآرایه نیز صدق می‌کند. شکل‌های ۲ تا ۴ مجموعه جواب‌های بهینه پارتوی مربوط به ریزآرایه‌های Colon، Leukemia و Ovarian را به ازای یکی از ده بار اجرای مستقل در حالت FSIS نشان می‌دهد. با توجه به این شکل‌ها مشاهده می‌شود الگوریتم پیشنهادی به خوبی قادر به طبقه‌بندی سطوح بیان ژن‌ها به دو کلاس است.

در جداول ۳ و ۴ میانگین نتایج حاصل از دو شاخص صحت و F-measure بر روی طبقه‌بند پایه و طبقه‌بند ترکیبی برای شش مجموعه داده ریزآرایه نشان داده شده است. در خصوص طبقه‌بندهای پایه همان طور که از این جداول قابل مشاهده است، در حالت Non-Pre که هیچ پردازشی بر روی مجموعه داده‌های ریزآرایه انجام نشده است، طبقه‌بند SVM دارای کارایی بهتری نسبت به دو طبقه‌بند KNN و NB است.

جدول ۵: نتایج آماری حاصل از روش پیشنهادی بر اساس شاخص صحت.

Dataset	Worst%	Average%	Best%	SD
Colon	۹۸٫۸۵	۹۹٫۲۴	۱۰۰	۰٫۱۶
Leukemia	۹۹٫۵۷	۹۹٫۸۷	۱۰۰	۰٫۱۴
Ovarian	۹۸٫۱۳	۹۸٫۵۳	۹۹٫۲۹	۰٫۱۸
Breast	۹۱٫۹۸	۹۳٫۱۶	۹۴٫۲۷	۰٫۴۸
ALL-AML-۳	۹۷٫۸۸	۹۸٫۸۹	۹۹٫۰۷	۰٫۳۵
SRBCT	۹۸٫۴۸	۹۸٫۶۳	۹۹٫۵۷	۰٫۱۱

جدول ۶: نتایج آماری حاصل از روش پیشنهادی بر اساس شاخص F.

Dataset	Worst%	Average%	Best%	SD
Colon	۸۹٫۲۲	۹۰٫۰۷	۹۰٫۶۳	۰٫۳۸
Leukemia	۹۲٫۴۸	۹۳٫۱۲	۹۳٫۳۶	۰٫۲۵
Ovarian	۸۸٫۹۰	۸۹٫۲۳	۸۹٫۴۱	۰٫۱۷
Breast	۸۲٫۱۴	۸۲٫۷۳	۸۳٫۱۳	۰٫۲۱
ALL-AML-۳	۹۱٫۱۵	۹۱٫۶۷	۹۲٫۲۵	۰٫۲۷
SRBCT	۸۳٫۷۷	۸۴٫۰۳	۸۴٫۸۱	۰٫۲۴

جدول ۷: مقایسه کارایی طبقه‌بندها بر روی مجموعه داده‌های ریزآرایه بر اساس شاخص صحت.

Datasets	FSIS-MMoCS	FCBF	BPSO	PSO-DT	MBEGA	CFS-TCBPSO-1NN	CFS-iBPSO-NB	IGIS	ReliefF-BGSA
Colon	۹۹٫۲۴	۸۵٫۴۸	۷۴٫۱۹	۹۰٫۳۲	۸۶٫۶۶	۹۸٫۴۳	۹۴٫۸۹	۷۷٫۴۷	۹۳٫۱۲
Leukemia	۹۹٫۸۷	۱۰۰	۱۰۰	۹۵٫۸۳	۹۵٫۸۹	۱۰۰	۱۰۰	۸۹٫۸۷	۱۰۰
Ovarian	۹۸٫۵۳	۹۹٫۹۱	۹۴٫۰۷	۹۷٫۲۳	۹۹٫۷۱	۱۰۰	۱۰۰	۹۹٫۴۹	۹۸٫۹۹
Breast	۹۳٫۱۶	۵۷٫۷۳	۵۴٫۶۳	۶۷٫۰۱	۸۰٫۷۴	۹۱٫۷	۹۲٫۷۵	–	۹۴٫۵۶
ALL-ML-۳	۹۸٫۸۹	۹۵٫۸۳	۹۴٫۰۷	۹۵٫۸۳	۹۶٫۶۴	۹۹٫۶۷	۱۰۰	۸۷٫۷۱	۹۳٫۶۷
SRBCT	۹۸٫۶۳	۹۸٫۹۴	۹۹٫۰۰	۹۲٫۴۹	۹۹٫۲۳	۱۰۰	۱۰۰	۹۰٫۰۰	۹۵٫۶۷
Average	۹۸٫۰۵	۸۹٫۶۵	۸۵٫۹۹	۸۹٫۷۹	۹۳٫۱۵	۹۸٫۳۰	۹۷٫۹۴	۸۸٫۹۱	۹۶٫۰۰

جدول ۸: مقایسه کارایی طبقه‌بندها بر روی مجموعه داده‌های ریزآرایه بر اساس تعداد ژن‌های انتخاب‌شده.

Datasets	FSIS-MMoCS	FCBF	BPSO	PSO-DT	MBEGA	CFS-TCBPSO-1NN	CFS-iBPSO-NB	IGIS	ReliefF-BGSA
Colon	۴٫۳	۱۴	۴۷۸٫۱	۶۴۳٫۳	۲۴٫۵	۹	۴٫۲	۵٫۳	۸٫۶
Leukemia	۴٫۳	۵۱	۵۷۲٫۴	۱۴۶۸	۹٫۰	۳٫۱	۴٫۳	۵٫۰	۷٫۶
Ovarian	۳٫۱	۳۰	۵۰۷۴٫۷	۳۵۹۴٫۲	۱۵٫۸	۱۰	۳٫۳	۲٫۸	۵٫۴
Breast	۱۳٫۶	۹۲	۱۹۳۰	۱۰۴۶۵	۱۴٫۵	۳۰٫۲	۳۲٫۷	–	۵۶٫۷
ALL-ML-۳	۶٫۳	۵۳	۳۳۷۹	۱۲۹۴٫۱	۲۰٫۱	۱۰٫۲	۶٫۰	۷٫۵	۷٫۸
SRBCT	۵٫۴	۸۲	۷۹۴	۸۷۴	۶۰٫۷	۳۹٫۱	۳۴٫۱	۹٫۲	۳۹٫۲۳
Average	۶٫۱۷	۵۳٫۶۷	۲۰۳۸٫۰۳	۳۰۵۶٫۴۳	۲۴٫۱	۱۶٫۹۳	۱۴٫۱	۵٫۹۶	۲۰٫۸۹

این بهبود کارایی و دقت بالاتر بیانگر این موضوع است که طبقه‌بند SVM حساسیت کمتری نسبت به تعداد زیاد ویژگی‌ها دارد. پس از SVM، طبقه‌بند KNN بالاترین کارایی را از نظر شاخص صحت و F-measure از خود نشان داده است.

در خصوص تأثیر انتخاب ویژگی و انتخاب نمونه‌های آموزش بر کارایی طبقه‌بندها، بر اساس نتایج ارائه‌شده در این جداول، مشاهده می‌شود به ازای هر شش مجموعه داده، مجدداً طبقه‌بند SVM دارای کارایی بهتری نسبت به دو طبقه‌بند دیگر است. با توجه به میانگین نتایج ارائه‌شده در جداول فوق مشاهده می‌شود در مقایسه با حالت Non-Pre، به صورت نسبی، کارایی طبقه‌بندهای SVM، KNN، NB و MV در حالت FSIS برای شاخص صحت به ترتیب ۸٫۴۲، ۸٫۲۹، ۱۱٫۷۸ و ۸٫۸۶ درصد و برای شاخص F-measure نیز ۱۲٫۳۹، ۱۲٫۶۷، ۱۶٫۹۴ و ۱۳٫۵۳ درصد بهبود یافته است. به بیانی دیگر با انتخاب ژن‌های مؤثر و متمایز، همچنین انتخاب نمونه‌های مناسب می‌توان علاوه بر کاهش هزینه‌های محاسباتی، کارایی الگوریتم‌های طبقه‌بندی را تا حد زیادی بهبود بخشید.

جدول ۵ به ترتیب بدترین جواب، بهترین جواب، میانگین جواب‌ها و انحراف معیار مربوط به شاخص صحت را به ازای ۱۰ بار اجرای الگوریتم در حالت FSIS نشان می‌دهد. با توجه به نتایج این جدول مشاهده می‌شود الگوریتم پیشنهادی توانسته است در دو مجموعه داده Colon و Leukemia به بهترین شاخص صحت ۱۰۰٪ برسد. مقادیر آماری مربوط به شاخص F-measure نیز در جدول ۶ آورده شده است.

برای ارزیابی کارایی روش پیشنهادی، مقایسه‌ای بین این روش با روش‌های ارائه‌شده در مقالات دیگر انجام شده است. در جدول ۷ میانگین نتایج حاصل از ۱۰ بار اجرای مستقل الگوریتم به ازای شاخص صحت آورده شده است. جدول ۸ نیز میانگین تعداد ویژگی‌های به دست آمده را نشان می‌دهد. همان طور که مشاهده می‌شود از نظر شاخص صحت، الگوریتم CFS-TCBPSO-1NN با میانگین ۹۸٫۳۰٪ توانسته است بهترین کارایی را از خود نشان دهد. همچنین الگوریتم پیشنهادی نیز با اختلاف بسیار ناچیزی (با میانگین ۹۸٫۰۵٪) پس از الگوریتم CFS-TCBPSO-1NN در رتبه دوم قرار گیرد. این در حالی است که از نظر میانگین تعداد ویژگی‌ها، الگوریتم پیشنهادی در رتبه اول و الگوریتم CFS-TCBPSO-1NN در رتبه پنجم قرار دارد. همان طور که قبلاً ذکر شد مسأله انتخاب ویژگی یک مسأله بهینه‌سازی چندهدفه است که در آن می‌بایست علاوه بر کمینه‌کردن تعداد ویژگی‌ها، شاخص‌های دیگری مانند شاخص صحت نیز بیشینه گردد. اما در اکثر پژوهش‌های مرتبط با این موضوع (از جمله در روش CFS-TCBPSO-1NN) از نسخه تک‌هدفه الگوریتم‌های بهینه‌سازی هوشمند استفاده شده است. به این معنی که با ترکیب وزنی توابع هدف، مسأله به یک مسأله بهینه‌سازی تک‌هدفه تقلیل داده می‌شود. در حالت کلی در مقایسه با الگوریتم‌های بهینه‌سازی چندهدفه، الگوریتم‌های بهینه‌سازی تک‌هدفه نمی‌توانند در هنگام بهینه‌سازی هم‌زمان چند تابع هدف کارایی مطلوبی داشته باشند. چرا که انتخاب مناسب وزن‌های متناظر با هر تابع، مستلزم داشتن اطلاعات

محیطی گسترده در مورد مسأله است که معمولاً از ابتدا در دسترس نمی‌باشند [۴۳]. به عبارت دیگر تبدیل یک مسأله بهینه‌سازی چندهدفه به یک مسأله بهینه‌سازی تک‌هدفه با استفاده از ترکیب وزنی اهداف، خود یک مسأله بهینه‌سازی است که در آن می‌بایست بسته به نوع و بزرگی توابع هدف، وزن‌های بهینه متناظر با هر هدف محاسبه شوند [۴۳]. شایان ذکر است چون در [۲۷] نتیجه اعمال روش IGIS بر روی مجموعه داده Breast داده نشده است، برای ارزیابی بهتر نتایج، در کلیه روش‌ها از اثر این مجموعه داده صرف نظر شده است.

به طور خلاصه دلیل برتری روش پیشنهادی را می‌توان از دو جنبه مورد بررسی قرار داد. یکی الگوریتم بهینه‌سازی و دیگری الگوریتم طبقه‌بندی. الگوریتم بهینه‌سازی فاخته جهت جستجو در فضای جواب و تولید گام‌های تصادفی از پرواز لِوی بهره می‌گیرد در حالی که در سایر الگوریتم‌های بهینه‌سازی این گام‌های تصادفی توسط توزیع گاوسی تولید می‌شود. توزیع لِوی دارای میانگین و واریانس بی‌نهایت است و این ویژگی باعث می‌شود هم‌پوشانی بین زیربخش‌ها در پردازش موازی این الگوریتم نسبت به سایر الگوریتم‌های مبتنی بر توزیع گاوسی حداقل شده و همگرایی الگوریتم نیز سریع‌تر گردد. به علاوه رویکرد نخبه‌گرایانه الگوریتم با نگه‌داشتن بهترین جواب‌ها و جایگزین‌کردن بدترین آنها با جواب‌هایی بهتر، همچنین استخراج محلی با برداشتن گام‌هایی به سوی بهترین جواب منجر به ایجاد مصالحه بین دو مفهوم استخراج و اکتشاف خواهد شد. به عبارت دیگر در سایر الگوریتم‌های بهینه‌سازی هوشمند با عملگرهایی مواجه هستیم که بعضاً فقط یک هدف خاص را بر عهده دارند ولی در الگوریتم جستجوی فاخته عملگرهای تعریف‌شده به صورت هم‌زمان چندین هدف را تحقق می‌بخشند.

به طور نمونه، خوشه‌بندی در این الگوریتم به فاخته‌ها کمک می‌کند به سرعت ناحیه جواب را به چندین بخش تقسیم کرده و نواحی مستعد جواب بهینه سراسری را به صورت تخمینی مشخص کنند. سپس تمامی فاخته‌ها به سمت این نواحی مهاجرت کرده و ناحیه‌های یافته‌شده را با دقت بیشتری جستجو می‌کنند. این امر موجب همگرایی بسیار سریع‌تر الگوریتم فاخته می‌شود. سؤالی که ممکن است مطرح شود این است که آیا این فرایند می‌تواند منجر به همگرایی زودرس الگوریتم شود؟ جواب این سؤال در فرایند تخم‌گذاری فاخته‌ها و راهکار ابتکاری ارائه‌شده در این مقاله نهفته است. برخلاف سایر الگوریتم‌ها، راهکار ابتکاری ارائه‌شده، علاوه بر توزیع تخم‌ها در اطراف بهینه فعلی منجر به ایجاد تعادل بین دو مؤلفه استخراج و اکتشاف شده و از گیرافتادن الگوریتم در بهینه‌های محلی جلوگیری می‌کند. از طرفی ویژگی توزیع لوی نیز به الگوریتم کمک می‌کند بتواند نواحی بیشتری از فضای مسأله را جستجو نماید.

در مورد روش انتخاب الگوریتم طبقه‌بندی نیز می‌توان این گونه توضیح داد که چون در الگوریتم‌های طبقه‌بند استاندارد، توزیع کلاس‌ها متوازن در نظر گرفته می‌شود، این دسته از الگوریتم‌ها در مواجهه با مجموعه داده‌های نامتوازن عملکرد خیلی مناسبی از خود ارائه نمی‌دهند چرا که الگوریتم‌های معمول طبقه‌بند به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل خواهند شد. این موضوع باعث افزایش خطا در شناسایی نمونه‌های اقلیت می‌شود. یکی از روش‌های حل نامتوازن‌بودن داده‌ها روش‌های مبتنی بر ترکیب طبقه‌بندها است. با توجه به این که هر طبقه‌بند پایه بر اساس پارامترهای کنترلی خود پاسخ متفاوتی به مسأله مورد نظر خواهد داد، می‌توان انتظار داشت با ترکیب طبقه‌بندهای پایه بر اساس رأی اکثریت، کارایی طبقه‌بند ترکیبی نیز افزایش یابد.

٦- نتیجه‌گیری

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 باعث ایجاد مشکلاتی در آنالیز داده‌های ریزآرایه خواهد شد. از طرفی کم‌بودن تعداد نمونه‌ها، طبقه‌بندها را برای ارائه یک مدل مناسب جهت پیش‌بینی نمونه‌های آزمایش تضعیف می‌کند. در این مقاله برای انتخاب ویژگی و انتخاب نمونه‌های آموزش از الگوریتم جستجوی فاخته اصلاح‌شده چندهدفه، جهت دستیابی به بهترین راه‌حل‌های موجود استفاده گردید. روش پیشنهادی شامل دو مرحله انتخاب ویژگی و انتخاب نمونه‌های مناسب جهت آموزش بود. در مرحله انتخاب ویژگی از نسخه دودویی الگوریتم جستجوی فاخته اصلاح‌شده چندهدفه برای تنظیم آستانه بهره اطلاعاتی استفاده شد. همچنین در مرحله انتخاب نمونه‌های آموزش نیز یک روش خوشه‌بندی فازی مبتنی بر نسخه پیوسته الگوریتم فاخته اصلاح‌شده چندهدفه با استفاده از کدگذاری مراکز خوشه‌ای پیشنهاد گردید. نتایج بررسی‌های انجام‌شده بر روی شش مجموعه داده ریزآرایه نشان دادند الگوریتم پیشنهادی می‌تواند با در نظر گرفتن همبستگی بین ژن‌ها و همچنین افزونگی بین آنها، ژن‌های اضافی و فاقد اطلاعات را حذف کند و منجر به افزایش کارایی طبقه‌بندها گردد. با این حال مقایسه بین نتایج حاصل از روش ارائه‌شده در این پژوهش با دیگر روش‌ها بیانگر این واقعیت است که الگوریتم پیشنهادی در برخی موارد عملکرد ضعیف‌تری نسبت به سایر الگوریتم‌ها از خود نشان می‌دهد. بر این اساس می‌توان چنین نتیجه‌گیری کرد که نمی‌توان الگوریتم و راهکار مشخصی را معرفی کرد که بتواند علاوه بر کاهش تعداد ویژگی‌ها، بهبود منجر به افزایش کارایی طبقه‌بندها در مقایسه با سایر روش‌ها باشد.

مراجع

[1] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J. M. Benitez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111-135, 20 Oct. 2014.

[2] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91-107, Mar. 2017.

[3] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Applied Soft Computing*, vol. 62, pp. 203-215, Jan. 2018.

[4] G. Ditzler, R. Polikar, and G. Rosen, "A sequential learning approach for scaling up filter-based feature subset selection," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2530-2544, Jun 2017.

[5] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets," *Expert Systems with Applications*, vol. 121, pp. 233-243, May 2019.

[6] G. I. Sayed, A. E. Hassanien, and A. T. Azar, "Feature selection via a novel chaotic crow search algorithm," *Neural Computing and Applications*, vol. 31, pp. 171-188, 2019.

[7] B. Cao, J. Zhao, P. Yang, P. Yang, X. Liu, J. Qi, *et al.*, "Multiobjective feature selection for microarray data via distributed parallel algorithms," *Future Generation Computer Systems*, vol. 100, no. 2, pp. 952-981, Nov. 2019.

[8] A. K. Das, S. K. Pati, and A. Ghosh, "Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm," *Knowledge and Information Systems*, vol. 62, no. 2, pp. 1-33, Feb. 2019.

[9] X. Li and M. Yin, "Multiobjective binary biogeography based optimization for feature selection using gene expression data," *IEEE Trans. on NanoBioscience*, vol. 12, no. 4, pp. 343-353, Dec. 2013.

on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764, Nov. 2011.
[33] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *J. of Machine Learning Research*, vol. 13, pp. 1393-1434, May 2012.
[34] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans. on Nanobioscience*, vol. 9, pp. 31-37, Oct. 2009.
[35] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: a mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371-6385, Oct. 2014.
[36] H. H. Hsu, C. W. Hsieh, and M. D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8144-8150, Jul. 2011.
[37] Y. Ye, Q. Wu, J. Z. Huang, M. K. Ng, and X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognition*, vol. 46, no. 3, pp. 769-787, Mar. 2013.
[38] M. K. Ebrahimpour and M. Eftekhari, "Ensemble of feature selection methods: a hesitant fuzzy sets approach," *Applied Soft Computing*, vol. 50, no. C, pp. 300-312, Jan. 2017.
[39] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Jun. 2005.
[40] A. Bai and A. Pradhan, "An efficient approach for feature extraction and classification of microarray cancer data," *International J. of Computational Intelligence Studies*, vol. 3, no. 4, pp. 339-355, Jan. 2014.
[41] X. S. Yang and S. Deb, "Cuckoo search: recent advances and applications," *Neural Computing and Applications*, vol. 24, no. 1, pp. 169-174, Jan. 2014.
[42] L. Lin and M. Gen, "Auto-tuning strategy for evolutionary algorithms: balancing between exploration and exploitation," *Soft Computing*, vol. 13, no. 2, pp. 157-168, Jan. 2009.
[43] H. Rashidi and J. Khorshidi, "Exergoeconomic analysis and optimization of a solar based multigeneration system using multiobjective differential evolution algorithm," *J. of Cleaner Production*, vol. 170, pp. 978-990, Jan. 2018.
[44] X. S. Yang and S. Deb, "Multiobjective cuckoo search for design optimization," *Computers & Operations Research*, vol. 40, no. 6, pp. 1616-1624, Jun. 2013.
[45] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, Aug. 2002.
[46] J. A. Olvera-Lopez, J. A. Carrasco-Ochoa, J. F. Martinez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 133-143, Aug. 2010.
[47] Y. Chen, J. Bi, and J. Z. Wang, "MILES: multiple-instance learning via embedded instance selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931-1947, Oct. 2006.
[48] L. Vendramin, R. J. Campello, and E. R. Hruschka, "Relative clustering validity criteria: a comparative overview," *Statistical Analysis and Data Mining: the ASA Data Science J.*, vol. 3, no. 4, pp. 209-235, Aug.. 2010.
[49] R. J. Campello and E. R. Hruschka, "A fuzzy extension of the silhouette width criterion for cluster analysis," *Fuzzy Sets and Systems*, vol. 157, no. 21, pp. 2858-2875, Nov. 2006.
[50] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, Oct. 2007.
[51] Z. Zhu, Y. S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, no. 11, pp. 3236-3248, Nov. 2007.
[52] A. A. Aburomman and M. B. I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Applied Soft Computing*, vol. 38, pp. 360-372, Jan. 2016.
[53] D. H. Mazumder and R. Veilumuthu, "An enhanced feature selection filter for classification of microarray cancer data," *ETRI J.*, vol. 41, no. 3, pp. 358-370, Jun. 2019.

خدیجه کمری مدرک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر به ترتیب در سال‌های ۱۳۹۲ و ۱۳۹۶ از دانشگاه غیرانتفاعی جهاد دانشگاهی اصفهان و دانشگاه هرمزگان دریافت کرد. زمینه‌های تحقیقاتی مورد علاقه‌ی او یادگیری ماشین، بازشناسی الگو، الگوریتم‌های تکاملی و رایانش نرم است.

[10] A. Joshi, O. Kulkarni, G. Kakandikar, and V. Nandedkar, "Cuckoo search optimization-a review," in *Materials Today: Proc.*, vol. 4, pp. 7262-7269, Jan. 2017.
[11] N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Trans. on Neural Networks*, vol. 13, no. 1, pp. 143-159, Jan. 2002.
[12] B. K. Singh, K. Verma, and A. Thoke, "Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images," *Expert Systems with Applications*, vol. 66, pp. 114-123, Dec. 2016.
[13] H. Uguz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024-1032, Oct. 2011.
[14] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing & Management*, vol. 42, no. 1, pp. 155-165, Jan. 2006.
[15] L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," in *Proc. of the 20th Int. Conf. on Machine Learning, ICML'03*, pp. 856-863, Washington DC, USA, ???. 2003.
[16] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, pp. 23-69, Oct. 2003.
[17] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 261-274, Jun. 2008.
[18] L. Lan and S. Vucetic, "Improving accuracy of microarray classification by a simple multi-task feature selection filter," *International J. of Data Mining and Bioinformatics*, vol. 5, no. 2, pp. 189-208, Jan. 2011.
[19] J. Wang, L. Wu, J. Kong, Y. Li, and B. Zhang, "Maximum weight and minimum redundancy: a novel framework for feature subset selection," *Pattern Recognition*, vol. 46, no. 6, pp. 1616-1627, Jun. 2013.
[20] N. Garcia-Pedrajas and J. Perez-Rodriguez, "Multi-selection of instances: a straightforward way to improve evolutionary instance selection," *Applied Soft Computing*, vol. 12, no. 11, pp. 3590-3602, Nov. 2012.
[21] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 1, pp. 483-519, Mar. 2013.
[22] D. Keco, A. Subasi, and J. Kevric, "Cloud computing-based parallel genetic algorithm for gene selection in cancer classification," *Neural Computing and Applications*, vol. 30, no. 5, pp. 1601-1610, Sept. 2018.
[23] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302-312, Oct. 2017.
[24] K. H. Chen, K. J. Wang, K. M. Wang, and M. A. Angelia, "Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data," *Applied Soft Computing*, vol. 24, pp. 773-780, Nov. 2014.
[25] L. Y. Chuang, C. S. Yang, K. C. Wu, and C. H. Yang, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13367-13377, Sept. 2011.
[26] A. Ghosh, A. Datta, and S. Ghosh, "Self-adaptive differential evolution for feature selection in hyperspectral image data," *Applied Soft Computing*, vol. 13, no. 4, pp. 1969-1977, Apr. 2013.
[27] S. Nakariyakul, "A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification," *PloS ONE*, vol. 14, no. 2, Article No.: e0212333, 15 Feb. 2019.
[28] X. Han, D. Li, P. Liu, and L. Wang, "Feature selection by recursive binary gravitational search algorithm optimization for cancer classification," *Soft Computing*, vol. 24, pp. 1-19, 2020.
[29] B. Liu, M. Tian, C. Zhang, and X. Li, "Discrete biogeography based optimization for feature selection in molecular signatures," *Molecular Informatics*, vol. 34, no. 4, pp. 197-215, Apr. 2015.
[30] S. Yazdani, J. Shanbehzadeh, and E. Aminian, "Feature subset selection using constrained binary/integer biogeography-based optimization," *ISA Trans.*, vol. 52, no. 3, pp. 383-390, Mar. 2013.
[31] D. Rodrigues, L. A. Pereira, R. Y. Nakamura, K. A. Costa, X. S. Yang, A. N. Souza, *et al.*, "A wrapper approach for feature selection based on bat algorithm and optimum-path forest," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2250-2258, May 2014.
[32] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans.*

فرزان رشیدی تحصیلات خود را در مقاطع کارشناسی، کارشناسی ارشد و دکتری برق به‌ترتیب در سال‌های ۱۳۷۹، ۱۳۸۱ و ۱۳۹۴ از دانشگاه شیراز، دانشگاه تهران و دانشگاه صنعتی شیراز به پایان رسانده است و هم‌اکنون عضو هیأت علمی دانشگاه هرمزگان می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: الگوریتم‌های تکاملی و رایانش نرم، یادگیری ماشین، انرژی‌های نو و کاربرد سیستم‌های خبره در مهندسی برق.

عبدالله خلیلی تحصیلات خود را در مقطع کارشناسی، کارشناسی ارشد و دکتری مهندسی کامپیوتر به‌ترتیب در سال‌های ۱۳۸۸، ۱۳۹۰ و ۱۳۹۵ از دانشگاه اصفهان، دانشگاه علم و صنعت و دانشگاه شیراز به پایان رسانیده است و از سال ۱۳۹۵ عضو هیأت علمی دانشگاه هرمزگان می‌باشد. زمینه های مورد علاقه ایشان، امنیت سیستم‌های کامپیوتری و صنعتی، یادگیری عمیق و داده‌کاوی می‌باشد.