یک روش انتخاب ویژگی ترکیبی برای داده‌های با بعد بالا مبتنی بر خرد جمعی

امیررضا روحی و حسین نظام‌آبادی‌پور

چکیده: امروزه با ظهور و گسترش داده‌های بعد بالا، روند انتخاب ویژگی نقش بسیار مهمی را در زمینه یادگیری ماشینی و به خصوص مسایل طبقه‌بندی داده، بازی می‌کند. کار بر روی داده‌های با بعد بالا از جمله داده‌های میکروآرایه‌ای با مشکلاتی همچون وجود ویژگی‌های نامرتبط و افزونه بسیار روبه‌رو است که باعث کاهش نرخ صحت طبقه‌بند، افزایش هزینه محاسباتی و معضل "نفرین بعد" می‌شود. در این مقاله به ارائه یک روش ترکیبی با استفاده از رویکردهای خرد جمعی برای انتخاب ویژگی در داده‌های با بعد بالا پرداخته می‌شود. در روش پیشنهادی، ابتدا در مرحله اول از یک روش فیلتری برای کاهش بعد داده استفاده می‌شود، سپس در مرحله دوم، دو الگوریتم روزآمد پیچشی با استفاده از رویکرد خرد جمعی بر روی ویژگی‌های کاهش‌یافته اعمال شده و نتیجه تجمیع می‌گردد. روش پیشنهادی بر روی ۸ پایگاه داده میکروآرایه‌ای مورد ارزیابی قرار گرفته و مقایسه نتایج با چندین روش روزآمد و شناخته‌شده در حوزه انتخاب ویژگی، کارایی روش پیشنهادی را تأیید می‌کند.

***کلیدواژه:* انتخاب ویژگی، داده‌های با بعد بالا، روش‌های ترکیبی، روش‌های فراابتکاری، روش‌های فیلتری، روش‌های خرد جمعی.**

۱- مقدمه

در دنیای امروز، افزایش بیش از پیش ابعاد داده به یک مسئله اساسی در مبحث داده‌کاوی تبدیل شده است. در گذشته، داده‌های کلاسیک که دارای حداکثر چند ده ویژگی بوده‌اند نتایج قابل قبولی در صحت طبقه‌بند داشته‌اند. امروزه با داده‌هایی همانند داده‌های بزرگ و بعد بالا روبه‌رو هستیم که تعداد ویژگی آنها به شدت افزایش پیدا کرده است به طوری که شامل هزاران ویژگی هستند. این افزایش بعد، هزینه محاسباتی سیستم را افزایش داده و منجر به کاهش نرخ صحت طبقه‌بند می‌شود [۱]. در این وضعیت می‌توان گفت که انتخاب ویژگی از اساسی‌ترین و مهم‌ترین مباحث در یادگیری ماشینی به حساب می‌آید.

امروزه محققان در بسیاری از مباحث با داده‌های با بعد بالا و داده‌های بزرگ روبه‌رو هستند [۲]. در دو دهه اخیر و با افزایش ابعاد داده‌ها در علم پزشکی، داده‌هایی به نام داده‌های میکروآرایه‌ای به وجود آمده‌اند. داده‌های میکروآرایه‌ای داده‌هایی هستند که از نمونه‌های بافت و سلول با توجه به تفاوت‌های موجود در ژن، استخراج می‌شوند. این نوع داده‌ها می‌توانند برای تشخیص بیماری و انواع تومور در علم پزشکی مفید باشند. داده‌های مذکور دارای تعداد بسیار زیادی ویژگی (بین ۲۰۰۰ تا ۶۰۰۰۰ ویژگی) و تعداد کمی نمونه هستند [۳] و [۴].

این مقاله در تاریخ ۱۰ دی ماه ۱۳۹۵ دریافت و در تاریخ ۳ شهریور ماه ۱۳۹۶ بازنگری شد.

امیررضا روحی، بخش مهندسی برق، دانشگاه شهید باهنر کرمان، کرمان، (email: amirreza.rouhi1@gmail.com).

حسین نظام‌آبادی‌پور، بخش مهندسی برق، دانشگاه شهید باهنر کرمان، کرمان، (email: nezam@uk.ac.ir).

تا به امروز، روش‌های انتخاب ویژگی مختلفی معرفی شده که به طور کلی می‌توان آنها را به سه گروه عمده دسته‌بندی کرد: ۱) روش‌های فیلتری، ۲) روش‌های پیچشی و ۳) روش‌های ترکیبی[۱].

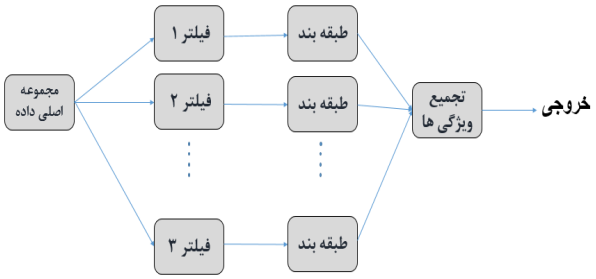
روش‌های فیلتری[۲] [۵] مستقل از الگوریتم یادگیری عمل می‌کنند. در این روش‌ها، خصوصیات ذاتی داده برای رتبه‌بندی مورد استفاده قرار گرفته و ویژگی‌های با بالاترین رتبه انتخاب می‌شوند [۶] و [۷]. این روش‌ها دارای سرعت بالایی هستند و در نتیجه به کارگیری این روش‌ها در داده‌های با بعد بالا مناسب می‌باشد. از جمله روش‌های فیلتری می‌توان به بهره اطلاعاتی[۳] (IG) [۸]، امتیاز فیشر[۴] (F-Score) [۹]، ریلیف- اف[۵] [۱۰]، روش فیلتری بر اساس همبستگی سریع[۶] (FCBF) [۱۱]، انتخاب ویژگی مبتنی بر همبستگی[۷] (CFS) [۱۲] و اثر متقابل[۸] [۱۳] اشاره کرد.

روش‌های پیچشی[۹] به روش‌هایی گفته می‌شود که میزان خطای طبقه‌بند ملاک سنجش کیفیت زیرمجموعه‌های ویژگی است و در نتیجه، این روش‌ها از دقت بسیار بالایی برخوردار هستند. به دلیل بررسی میزان شایستگی هر زیرمجموعه توسط طبقه‌بند، این روش‌ها سرعت پایین و پیچیدگی محاسباتی بالایی دارند [۱۴] که همین امر باعث ایجاد محدودیت در به کارگیری آنها در داده‌های با بعد بالا می‌شود.

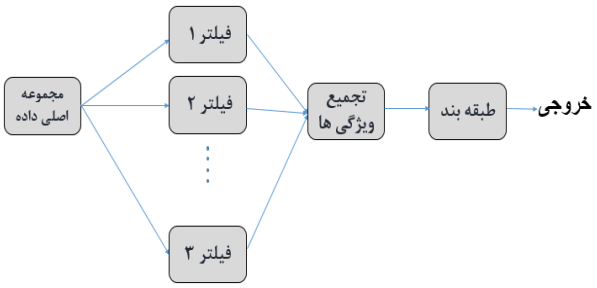
از آنجایی که هر کدام از روش‌های فیلتری یا پیچشی، یافتن بهترین جواب را تضمین نکرده و هر کدام دارای مزایا و نقص‌هایی هستند، می‌توان آنها را به صورت مکمل در کنار هم استفاده کرد که به این روش، روش ترکیبی می‌گویند. به عبارت دیگر، روش‌های ترکیبی از دو طبقه تشکیل شده‌اند. در طبقه اول، روش‌های فیلتری به کاهش ابعاد داده می‌پردازند و پس از آن توسط روش‌های پیچشی، بهترین زیرمجموعه‌های ویژگی، انتخاب خواهد شد. در نتیجه، احتمال حذف‌شدن ویژگی‌های مطلوب در روش‌های ترکیبی نسبت به روش‌های فیلتری کمتر است. استفاده از روش‌های ترکیبی می‌تواند یکی از بهترین گزینه‌ها برای انتخاب ویژگی در داده‌های با بعد بالا باشد [۱۵] زیرا حذف ویژگی‌های نامرتبط و افزونه بدون کاهش سرعت و افزایش نه چندان زیاد پیچیدگی محاسباتی انجام می‌شود.

امروزه با افزایش بعد داده‌ها و همچنین بالارفتن پیچیدگی محاسباتی برای رسیدن به جوابی مطلوب با کمترین خطای ممکن، توجه محققان به

1. Hybrid Methods
2. Filter Methods
3. Information Gain
4. Fisher Score
5. ReliefF
6. A Fast Correlation-Based Filter Solution
7. Correlation Based Feature Selection
8. Interact
9. Wrapper Methods



شکل ۱: نمونه ۱ روش خرد جمعی [۲۴].



شکل ۲: نمونه ۲ روش خرد جمعی [۲۴].

سمت الگوریتم‌های فراابتکاری[1] جلب شده است. این روش‌ها، راه حل مناسبی برای حل مسایل پیچیده و زمان بر می‌باشند به طوری که راه حل‌های نه صد در صد بهینه، بلکه نزدیک به جواب بهینه را با هزینه محاسباتی مطلوب جستجو می‌کنند. این الگوریتم‌ها با الهام از فرایندهای فیزیکی و بیولوژیکی در طبیعت به دست آمده‌اند و غالب آنها به صورت جمعیتی عمل می‌کنند [۱۶] و به دلیل آن که نرخ صحت طبقه‌بندی را به عنوان تابع ارزیابی مجموعه ویژگی‌های انتخاب‌شده قرار می‌دهند، عموماً در قالب روش‌های پیچشی به کار گرفته می‌شوند. در مسایل بعد بالا به دلیل پیچیدگی زمانی حاصل از اعمال این روش‌ها بر روی داده‌های بعد بالا، امکان عملی برای استفاده آنها به تنهایی در این قبیل داده‌ها مقرون به صرفه نیست بلکه می‌توان از آنها در قالب روش‌های ترکیبی بهره برد تا ابتدا بُعد داده به صورت چشم‌گیری کاهش یابد و سپس روش‌های پیچشی بر روی آنها اعمال شوند.

از جمله الگوریتم‌های فراابتکاری می‌توان به الگوریتم وراثتی[2] با الهام از علم وراثت [۱۷]، الگوریتم بهینه‌سازی جمعیت ذرات[3] با شبیه‌سازی رفتار جمعی پرندگان [۱۸]، الگوریتم جستجوی گرانشی[4] با الهام از جرم و نیروی جاذبه [۱۶] و الگوریتم جمعیت مورچگان[5] با تقلید از رفتار مورچه‌ها مورچه‌ها در جستجوی غذا [۱۳] اشاره کرد.

این مقاله به ارائه یک روش انتخاب ویژگی ترکیبی با به کارگیری دو الگوریتم فراابتکاری "الگوریتم جستجوی گرانشی بهبودیافته" (IBGSA)[6] و "الگوریتم فراابتکاری مورچگان باینری پیشرفته"[7] به صورت خرد جمعی می‌پردازد. هدف این روش انتخاب زیرمجموعه‌ای با ویژگی‌های برجسته از داده‌های با بعد بالا و میکروآرایه‌ای ارائه شده است.

ادامه این مقاله این گونه سازماندهی شده که در بخش دوم مروری بر کارها و روش‌های ارائه‌شده در انتخاب ویژگی در داده‌های با بعد بالا می‌شود. در بخش سوم به معرفی تعدادی از روش‌های فیلتری و همچنین دو الگوریتم فراابتکاری به کار رفته در مقاله پرداخته می‌شود. در بخش چهارم روش پیشنهادی برای حل مسایل انتخاب ویژگی بر روی داده‌ها با بعد بالا معرفی می‌گردد. در بخش پنجم کارایی روش پیشنهادی روی چند مجموعه داده استاندارد میکروآرایه‌ای بررسی می‌شود و سپس نتایج مقایسه با سایر روش‌ها گزارش می‌گردد. در نهایت در بخش ششم، مقاله جمع‌بندی می‌شود.

۲- مروری بر کارهای انجام‌شده و بیان چالش‌ها

تا کنون روش‌های متعددی مبتنی بر روش‌های فیلتری برای انتخاب ویژگی در داده‌های بعد بالا و میکروآرایه‌ای ارائه شده است. مرجع [۱۹] در سال ۲۰۰۸ یک روش فیلتری به نام MASSIVE برای انتخاب ویژگی روی داده‌های میکروآرایه‌ای بر اساس یک معیار تئوری اطلاعات به نام DIRS پیشنهاد کرد. در سال ۲۰۱۱ نیز نویسندگان در [۲] به معرفی یک روش انتخاب ویژگی بر روی داده‌های با بعد بالا بر اساس روش فیلتری چندوظیفه‌ای[8] پرداخته‌اند. در سال ۲۰۱۳، [۲۰] به ارائه یک روش فیلتری برای رویارویی با داده‌های میکروآرایه‌ای پرداخته است. در این روش بر اساس بیشینه وزن و کمینه افزونگی، زیرمجموعه ویژگی بهینه انتخاب می‌شود که وزن هر ویژگی، نشان‌دهنده اهمیت آن ویژگی و افزونگی، نشان‌دهنده همبستگی میان ویژگی‌ها است.

از جمله روش‌های پیچشی اعمال‌شده بر روی داده‌های میکروآرایه‌ای می‌توان به روش انتخاب ویژگی متوالی SFS که در [۲۱] پیشنهاد شده اشاره کرد. در این روش ابتدا ویژگی‌های برتر هر بلوک با توجه به عملکردشان در طبقه‌بند مشخص می‌شوند، سپس برای دستیابی به بهترین زیرمجموعه ویژگی با هم مقایسه می‌گردند.

در [۲۲] یک روش ترکیبی برای انتخاب ویژگی در داده‌های با بعد بالا ارائه شده است. در این روش که از سه مرحله تشکیل شده، ابتدا توسط دو روش فیلتری امتیاز فیشر و بهره اطلاعاتی، ویژگی‌های نامرتبط و افزونه حذف می‌شوند که باعث کاهش بعد داده‌ها می‌شود. سپس نتایج حاصل از از این دو روش توسط عملگرهای AND و XOR، ترکیب شده و دو زیرمجموعه ویژگی مجزا تولید می‌کند. در مرحله آخر، روش پیچشی با یک الگوریتم یادگیری به انتخاب ویژگی‌های مطلوب می‌پردازد. در [۲۳] روشی به نام R-m-GA ارائه شده که از ترکیب سه روش ReliefF، Mrmr و الگوریتم وراثتی بهره می‌برد.

داده‌ها ممکن است از نظر تعداد ویژگی‌ها و نمونه‌ها خیلی بزرگ بوده و همچنین از نظر نویز، افزونگی و غیر خطی بودن دارای مشکل باشند. اخیراً روش‌های انتخاب ویژگی به نام روش‌های خرد جمعی[9] مورد توجه محققان قرار گرفته که در این روش‌ها به جای به کار بردن یک روش انتخاب ویژگی خاص و پذیرش نتایج آن به عنوان زیرمجموعه ویژگی نهایی از ترکیب نتایج چندین روش متفاوت استفاده می‌شود. به طور کلی این رویکرد از دو مرحله تشکیل شده است: در مرحله اول تعدادی از انتخاب‌گرهای ویژگی در نظر گرفته می‌شوند. در مرحله دوم خروجی‌های حاصل‌شده از مرحله اول به صورت مجزا و با روش‌های مختلفی با هم ترکیب شده و ویژگی‌های منتخب نهایی را در خروجی ایجاد می‌کنند. در شکل‌های ۱ و ۲ دیاگرام دو رویکرد انتخاب ویژگی به روش خرد جمعی نشان داده شده است [۲۴].

1. Meta-Heuristic
2. Genetic Algorithm
3. Particle Swarm Optimization
4. Gravitational Search Algorithm
5. Ant Colony Optimization
6. Improved Binary Gravitational Search Algorithm
7. Advanced Binary ACO
8. Multi-Task
9. Ensemble Methods

مرجع [۲۵] در سال ۲۰۱۶ به ارائه یک روش خرد جمعی مبتنی بر خوشه‌بندی برای داده‌های با بعد بالا پرداخته است. در [۲۶] نویسندگان به ارائه یک روش ترکیبی پرداخته‌اند که در آن از روش‌های خرد جمعی نیز بهره برده‌اند. در این روش در مرحله اول نتایج حاصل از سه روش "ریلیف- اف"، "بهره اطلاعات" و روش فیلتری بر اساس "همبستگی سریع" توسط عمگر AND منطقی به صورت خرد جمعی با هم ترکیب می‌شوند و در مرحله بعد، الگوریتم فراابتکاری $ABACO_H$ بر روی ویژگی‌های کاهش‌یافته در مرحله اول اعمال می‌شود.

نویسندگان در [۲۷] به ارائه یک روش مبتنی بر روش فراابتکاری الگوریتم جمعیت مورچگان (ACO) به نام MGSACO پرداخته‌اند که از تلفیق روش بهینه‌سازی مورچگان و روش فیلتری حاصل شده است.

تحقیقات مختلف بر روی داده‌های با بعد بالا نشان داده که با وجود صدها و هزاران ویژگی موجود در این داده‌ها، تعداد زیادی از این ویژگی‌ها حاوی اطلاعات مفید نیستند زیرا نسبت به برچسب خروجی، نامرتبط یا افزونه هستند [۲۸] و [۲۹]. به هر حال افزایش هرچه بیشتر تعداد ویژگی‌ها در این داده‌ها، کاهش نرخ صحت طبقه‌بندی و همچنین گیج‌شدن طبقه‌بند را به همراه دارد که از معضلات نفرین بعد محسوب می‌شود. در این میان، انتخاب ویژگی بیشترین ضرورت را برای طبقه‌بندی دارد زیرا باعث حذف ویژگی‌های اضافی و نامرتبط می‌شود [۳۰]. امروزه تلاش محققان برای یافتن روش‌های جدیدی است که در داده‌های بزرگ و بعد بالا بتوانند با حذف داده‌های نویزی، نامرتبط و اضافی علاوه بر کاهش بعد، نرخ صحت طبقه‌بندی مطلوبی نیز حاصل کنند [۳۱] و [۳۲].

در میان روش‌های موجود، روش‌های پیچشی نرخ صحت طبقه‌بندی مطلوبی دارند ولی به دلیل سرعت کار پایین و پیچیدگی محاسباتی بالا، نمی‌توان به خوبی و به تنهایی از آنها در داده‌های بعد بالا بهره برد. از طرفی، روش‌های فیلتری سرعت بسیار بالاتری نسبت به روش‌های پیچشی دارند اما دارای نرخ صحت طبقه‌بندی کمتری نسبت به روش‌های پیچشی هستند. امروزه توجه محققان معطوف به ترکیب این روش‌ها برای رسیدن به روشی است که علاه بر سرعت و دقت بالا، کاهش بعد را نیز به خوبی انجام دهد. حال چگونگی ترکیب این روش‌ها و رسیدن به روشی که نسبت به سایر روش‌های ارائه‌شده، نتایج بهتری حاصل کند یکی از مباحث و چالش‌های موجود می‌باشد.

۳- مفاهیم پایه

در این قسمت به مرور چندین روش فیلتری و همچنین دو الگوریتم فراابتکاری جستجوی گرانشی بهبودیافته و مورچگان باینری پیشرفته پرداخته می‌شود.

۳-۱ روش بهره اطلاعات

این روش، یک روش فیلتری تک‌متغیره است و به محاسبه میزان اطلاعات موجود در یک ویژگی می‌پردازد. بهره اطلاعات بر پایه مفهوم آنتروپی در تئوری اطلاعات است و به عبارت دیگر، بهره اطلاعات ویژگی xi در مجموعه نمونه‌های Sx به صورت ذیل نوشته می‌شود [۳۳]

$$IG(Sx,xi)=H(Sx)-\sum_{v=values(xi)}^{\frac{|Sxi=v|}{|Sx|}}H(Sxi=v) \qquad (۱)$$

که $values(xi)$ مجموعه مقادیری است که ویژگی xi می‌تواند داشته باشد و آنتروپی مجموعه S به صورت زیر تعریف می‌شود

$$H(S)=-(p_-)\log_۲(p_+)-(p_-)\log ۲(p_-) \qquad (۲)$$

که در آن $p+$ بیانگر نسبت تعداد نمونه‌های مثبت به کل نمونه‌ها و $p-$ بیانگر نسبت تعداد نمونه‌های منفی به کل نمونه‌هاست. هرچه یک ویژگی، مقدار IG بزرگ‌تری داشته باشد در انتخاب ویژگی از اعتبار بیشتری برخوردار است.

۳-۲ روش فیلتری بر اساس همبستگی سریع

روش فیلتری بر اساس همبستگی سریع (FCBF)، یکی از روش‌های فیلتری چندمتغیره است که در سال ۲۰۰۳ توسط [۱۱] برای انتخاب ویژگی در داده‌های با بعد بالا طراحی شده است. این الگوریتم دو معیار زیر را در نظر گرفته و به اندازه‌گیری این دو می‌پردازد:

۱) همبستگی ویژگی با برچسب داده

۲) همبستگی ویژگی با ویژگی

FCBF با انتخاب مجموعه‌ای از ویژگی‌هایی که همبستگی زیادی با برچسب داده‌ها بر اساس معیار عدم قطعیت متقارن دارند، شروع به انتخاب ویژگی می‌کند. معیار عدم قطعیت متقارن، معیاری است که بر پایه نسبت بین بهره اطلاعات و آنتروپی بین دو ویژگی تعریف می‌شود. این روش ابتدا ویژگی‌های افزونه را حذف کرده و سپس ویژگی‌های مرتبط با برچسب داده را نگه می‌دارد [۱۱].

۳-۳ روش ریلیف و ریلیف- اف

ریلیف[1] [۳۴] یک الگوریتم انتخاب ویژگی بر اساس روش فیلتری تک‌متغیره است و به جستجوی ویژگی‌هایی می‌پردازد که به طور آماری مرتبط با برچسب داده‌ها باشند. این الگوریتم تنها برای مسایل دوگروهی (دو کلاس خروجی) تعریف شده است. بر اساس این روش، ویژگی‌ای مطلوب‌تر است که اختلاف بیشتری را میان نمونه‌های گروه‌های مختلف ایجاد نموده و مقادیر یکسانی را برای نمونه‌های گروه مشابه داشته باشد [۳۴].

از مشکلات الگوریتم ریلیف آن است که قابلیت رسیدگی به داده‌های غیر کامل و نویزی را ندارد و همچنین این الگوریتم نمی‌تواند به بررسی مسایل چندگروهی (چندکلاسی) بپردازد. از این رو در الگوریتم ریلیف- اف [۱۰] که تعمیم‌یافته الگوریتم ریلیف است این مشکلات برطرف شده و توانایی رسیدگی به داده‌های غیر کامل و همچنین بررسی مسایل چندگروهی اضافه شده است.

۳-۴ روش امتیاز فیشر

ایده الگوریتم انتخاب ویژگی امتیاز فیشر، پیداکردن زیرمجموعه‌ای از ویژگی‌ها به گونه‌ای است که فاصله بین نقاط داده در گروه‌های متفاوت تا حد امکان زیاد و فاصله بین نقاط داده در یک گروه تا حد امکان کم باشد [۹].

ویژگی Xi از یک مجموعه داده m گروهی را در نظر بگیرید. اگر مجموعه نمونه‌ها از ویژگی iام در گروه kام X_i^k باشد و بدانیم $\left|X_i^k\right|=n_k$ که $k=۱,۲,\ldots,m$ می‌باشد و همچنین $\bar{X}_i^k$ و $\overline{X_i}$ به ترتیب میانگین ویژگی‌ها در X_i^k و X_i باشند، آن گاه امتیاز فیشر این ویژگی به صورت زیر تعریف می‌شود [۳۵]

1. Relief

$$F(X_i)=\frac{\sum_{k=1}^{m} n_k(\bar{X}_i^k-\bar{X}_i)^2}{\sum_{k=1}^{m}\sum_{x\in X_i^k}(x-\bar{X}_i^k)^2} \quad (۳)$$

که صورت کسر نشان‌دهنده تمیزدهندگی بین دو گروه و مخرج کسر نشان‌دهنده میزان پراکندگی در هر گروه است. هرچه امتیاز فیشر یک ویژگی بیشتر باشد آن ویژگی قابلیت تمیزدهندگی بیشتری خواهد داشت. بعد از محاسبه امتیاز فیشر هر ویژگی، تعدادی از ویژگی‌ها با امتیاز بیشتر را بر اساس یک حد آستانه از پیش تعریف شده به عنوان زیرمجموعه ویژگی نهایی انتخاب می‌شوند.

۳-۵ الگوریتم جستجوی گرانشی بهبودیافته

الگوریتم جستجوی گرانشی نمونه‌ای از الگوریتم‌های فراابتکاری است که در سال ۲۰۰۹ با الهام از جرم و نیروی جاذبه، توسط [۱۶] ارائه شده است. بر طبق قانون نیوتن هر ذره در جهان به ذره‌های دیگر نیرویی وارد می‌کند که این نیرو متناسب با حاصل‌ضرب جرم‌های آنهاست و با مجذور فاصله بین آنها نسبت عکس دارد.

در سال ۲۰۱۰ نسخه باینری این الگوریتم به نام BGSA در [۳۶] ارائه گردید. سپس در سال ۲۰۱۴ نیز نسخه بهبودیافته این الگوریتم به نام IBGSA توسط [۳۷] برای جلوگیری از ایجاد رکود و گیرافتادن در بهینه‌های محلی برای حل مسایل انتخاب ویژگی ارائه شد. در ادامه به توضیح الگوریتم فراابتکاری IBGSA پرداخته شده است.

در یک سیستم که دارای s عامل است، موقعیت iامین عامل به صورت (۴) بیان می‌شود

$$X_i=(x_i^1,\ldots,x_i^d,\ldots,x_i^n) \quad , \quad i=1,2,\ldots,s \quad (۴)$$

که x_i^d موقعیت بعد d از عامل i را نشان می‌دهد و n نیز نشان‌دهنده بعد فضای جستجو است. جرم هر عامل نیز پس از محاسبه مقدار شایستگی جمعیت فعلی با استفاده از رابطه زیر محاسبه می‌شود

$$M_i(t)=\frac{fit_i(t)-worst(t)}{\sum_{j=1}^{S} fit_j(t)-worst(t)} \quad (۵)$$

$$worst(t)=\max fit_j(t) \quad , \quad j\in\{1,2,\ldots,s\} \quad (۶)$$

که در آن $M_i(t)$ و fit_i به ترتیب جرم و مقدار شایستگی عامل iام در زمان tام می‌باشد. برایند نیروهای وارد به عامل iام از سوی مجموعه عامل‌های سنگین‌تر، شتاب و سرعت هر عامل از روابط زیر به دست می‌آیند

$$F_i^d(t)=\sum_{\substack{j\in kbest\\ j\neq i}} randjG(t)\frac{Mj(t)Mi(t)}{Rij(t)+\varepsilon}(x_j^d(t)-x_i^d(t)) \quad (۷)$$

$$a_i^d(t)=\frac{F_i^d(t)}{M_i(t)}=\sum_{\substack{j\in kbest\\ j\neq i}} rand_j G(t)\frac{Mj(t)}{Rij(t)+\varepsilon}(x_j^d(t)-x_i^d(t)) \quad (۸)$$

$$v_i^d(t+1)=rand_i\times v_i^d(t)+a_i^d(t) \quad (۹)$$

در روابط فوق $rand_i$ و $rand_j$ دو عدد تصادفی با تابع توزیع یکنواخت در فاصله $[0,1]$ هستند و ε یک مقدار کوچک می‌باشد. $R_{ij}(t)$ فاصله همینگ دو عامل i و j می‌باشد و به صورت (۱۰) تعریف می‌شود

$$R_{ij}(t)=\frac{1}{n}\sum^{n}\left|x_j^d(t)-x_i^d(t)\right| \quad (۱۰)$$

$Kbest$ شامل مجموعه K عامل برتر جمعیت با شایستگی بیشتر است و تابعی از زمان می‌باشد که در ابتدا با مقدار $K0$ شروع و با زمان کاهش می‌یابد. ثابت گرانشی G نیز تابعی از زمان بوده و با مقدار G_0 مقداردهی شده و با زمان کاهش می‌یابد

$$G=G_0(1-\exp(-\alpha\frac{t}{T})) \quad (۱۱)$$

مکان عامل‌ها با یک احتمال بر طبق (۱۲) که به تابع انتقال معروف است تغییر می‌کند

$$Tfn(v_i^d(t))=A+(1-A)\times\left|\tanh v_i^d(t)\right| \quad (۱۲)$$

در رابطه فوق A از (۱۳) به دست می‌آید

$$A=k_1(1-\exp\frac{Fc}{k_2}) \quad (۱۳)$$

که در رابطه فوق k_1 پارامتر ثابت، k_2 ثابت زمانی که بر اساس مسأله تعریف می‌شود و Fc نیز شمارنده شکست است. شکست زمانی اتفاق می‌افتد که بهترین جواب "دیده‌شده تا کنون" بعد از یک تکرار تغییر نکند [۳۶]. در نهایت عامل‌ها با توجه به (۱۴) حرکت می‌کنند

$$x_i^d(t+1)=\begin{cases}\text{complement } x_i^d(t) & \text{if rand}()<Tfn(v_i^d(t+1))\\ x_i^d(t) & \text{else}\end{cases} \quad (۱۴)$$

مکان عامل تنها در صورتی تغییر می‌کند که مکان جدید، شایستگی بهتر یا حداقل مساوی شایستگی قبل ایجاد کند. به بیان ریاضی این خاصیت را می‌توان به صورت زیر بیان نمود

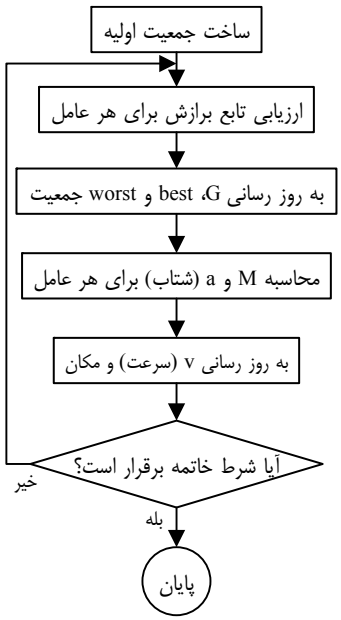
$$X_i(t+1)=\begin{cases}X_i(t+1) & \text{if } fit(X_i(t+1))<fit(X_i(t))\\ X_i(t) & \text{otherwise}\end{cases} \quad (۱۵)$$

پس از به روز رسانی مکان‌ها، الگوریتم تا رسیدن به شرط توقف تکرار می‌شود. برای توقف تعداد سیکل‌های الگوریتم، ملاک‌های مختلفی در نظر گرفته می‌شود. روش به کار رفته در این مقاله توقف پس از تعداد سیکل‌های معین است. در شکل ۳ روندنمای الگوریتم جستجوی گرانشی بهبودیافته نشان داده شده است.

۳-۶ الگوریتم بهینه‌ساز جمعیت مورچگان

الگوریتم بهینه‌ساز جمعیت مورچگان در سال ۱۹۹۱ با الهام از رفتار مورچه‌ها در جستجوی غذا توسط [۱۳] ارائه شد. مورچه‌ها قادرند با وجود کور و کم‌هوش‌بودن، با ارتباط با یکدیگر و انتقال اطلاعات مسیر از طریق رد پای فرمونی کوتاه‌ترین مسیر رفت و برگشت از خانه تا غذا را پیدا کنند.

در سال‌های اخیر الگوریتم بهینه‌سازی مورچگان باینری (BACO) بر مبنای الگوریتم مورچگان ارائه شده است. مشکل اصلی الگوریتم باینری مورچگان در انتخاب ویژگی این است که هر مورچه در گره i تنها قادر به تصمیم‌گیری درباره ویژگی بعدی می‌باشد و اگر از این ویژگی بگذرد و این ویژگی را انتخاب نکند قادر به بررسی حضور این ویژگی در گره‌های بعدی نیست. همچنین در این روش امکان ارائه راه حل جامع برای بینایی وجود ندارد. در سال ۲۰۱۳ نسخه‌ای از الگوریتم مورچگان باینری پیشرفته به نام $ABACO_H$ توسط [۳۸] ارائه شد که در این روش از دو الگوریتم مورچگان باینری و گسسته به صورت ترکیبی استفاده شده است.



شکل ۳: روندنمای الگوریتم IBGSA.

در انتخاب ویژگی توسط این الگوریتم، مسأله باید به صورت گراف تعریف شود که ویژگی‌ها به جای گره‌های گراف قرار می‌گیرند. مکان مورچه‌ها در ابتدا به صورت تصادفی روی گراف انتخاب می‌شود و سپس از طریق رابطه

$$P^{k}_{i_x,j_y}(t)= \begin{cases} \dfrac{\tau_{i_x,j_y}\eta_{i_x,j_y}}{\sum_{j}\tau_{i_x,j.}\eta_{i_x,j.}+\tau_{i_x,j_1}\eta_{i_x,j_1}} & , \ j\in \text{addmissible nodes} \\ 0 & , \ \text{otherwise} \end{cases} \quad (16)$$

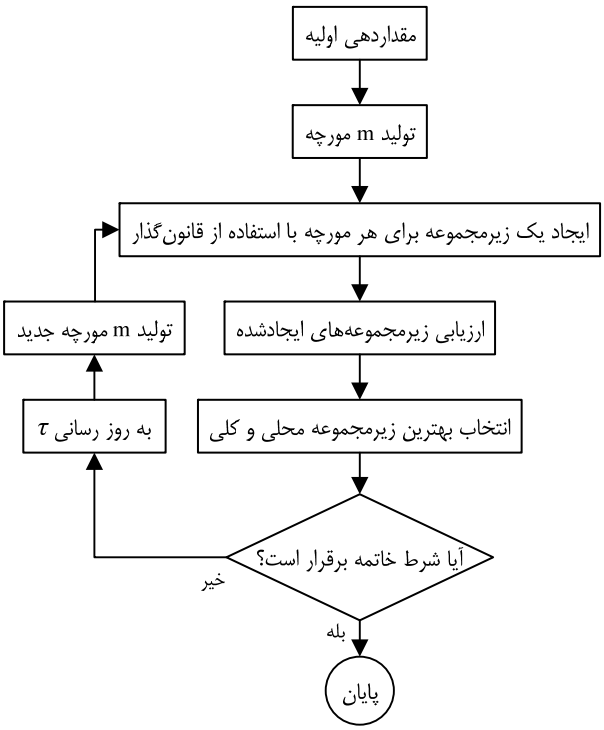
احتمال انتخاب بیت با مقدار $y\in\{0,1\}$ در گره بعدی (گره jام) برای مورچه kام که در زمان t در موقعیت $x\in\{0,1\}$ در گره iام قرار داشته است مشخص می‌شود. همچنین $\tau_{i_0,j.}$، $\tau_{i_1,j.}$، τ_{i_0,j_1} و $\tau_{i_0,j.}$ به ترتیب بیانگر رد پای فرومونی بین مسیرهای متصل‌کننده گره‌های iام و jام به ترتیب روی یال‌های (۰ به ۰)، (۱ به ۰)، (۰ به ۱) و (۱ به ۱) است [۳۸] و Δij هزینه جابه‌جایی از گره iام به گره jام می‌باشد. α و β پارامترهای مسئله‌اند که میزان اهمیت رد پا در مقابل بینایی را کنترل می‌کنند. τ_{ij} نیز رد پای فرمونی یال موجود بین گره i و j می‌باشد. به روز کردن رد پای شاخه‌ها از روابط زیر انجام می‌پذیرد [۳۸]

$$\tau_{ij}(new)=(1-\rho)\tau_{ij}(old)+\Delta\tau^{k}_{ij} \quad (17)$$

$$\Delta\tau^{k}_{ij}=\sum_{k=1}^{m}\Delta\tau \quad (18)$$

$$\Delta\tau^{k}_{ij}=\begin{cases}\dfrac{Q}{F_k} & , \ \text{if the } K-\text{th } ant \ traverse \ arc(i,j) \\ 0 & , \ \text{otherwise}\end{cases} \quad (19)$$

در روابط فوق، ρ ضریب تبخیر رد پا می‌باشد که از انباشته‌شدن بیش از حد رد پا جلوگیری می‌کند و $\Delta\tau^{k}_{ij}$ و $\Delta\tau_{ij}$ به ترتیب مقدار رد پایی است که مورچه kام و رد پایی که تمام مورچه‌ها بر یال (i,j) اضافه می‌کنند. m تعداد مورچه‌ها است و F_k هزینه طی‌کردن مسیری که مورچه kام از آن گذشته است و شاخه (i,j) در آن مسیر قرار دارد. T_k نیز تور مورچه kام یا مسیر حرکت این مورچه است. اگر مورچه‌های زیادی از



شکل ۴: روندنمای الگوریتم $ABACO_H$. (به انتهای روندنما، پایان اضافه شد)

مسیری عبور کنند، مقدار رد پای آن مسیر بیشتر شده و بالعکس رد پای مسیرهایی که مورچه‌های کمتری از آن عبور کرده‌اند به مرور زمان تبخیر می‌شود [۳۸].

شایان ذکر است که مورچه صرفاً به یکی از زیرگره‌های صفر یا یک سفر می‌کند که انتخاب زیرگره یک به منزله انتخاب آن ویژگی و انتخاب زیرگره صفر به منزله عدم انتخاب آن ویژگی است. شرط توقف نیز در این روش می‌تواند معیارهایی مانند توقف پس از همگراشدن مورچه‌ها به یک جواب یا توقف پس از تعداد سیکل‌های معین باشد. در الگوریتم به کار رفته در این مقاله از معیار دوم استفاده شده است. شکل ۴ روندنمای الگوریتم مورچگان باینری پیشرفته را نشان می‌دهد.

الگوریتم $ABACO_H$ به مورچه اجازه می‌دهد تا بین تمام ویژگی‌های موجود کاوش کند که یک مزیت بسیار مهم در الگوریتم مورچگان محسوب می‌شود. در این روش، برخلاف روش‌های مورچگان ارائه‌شده در گذشته که مشاهده ویژگی به منزله انتخاب آن است، مورچه می‌تواند ویژگی‌های مشاهده‌شده را انتخاب یا رد کند. همچنین در این روش جدید، مورچه‌ها قادر به دیدن تمام ویژگی‌هایی هستند که قبلاً از آنها بازدید نکرده‌اند که این مزیت در روند انتخاب ویژگی نتایج بهتری را در پی دارد.

٤- روش پیشنهادی

روش‌های ترکیبی به دلیل بهره‌مندی از هر دو روش فیلتری و پیچشی، سرعت بالاتر و همچنین پیچیدگی محاسباتی کمتری نسبت به روش‌های پیچشی دارند و در عین حال از نرخ صحت طبقه‌بندی مطلوب‌تری نسبت به روش‌های فیلتری بهره می‌برند و از این رو گزینه مناسبی برای انتخاب ویژگی در داده‌های با بعد بالا می‌باشند. روش‌های خرد جمعی نیز به دلیل اعمال نظرات چندین روش مختلف به طور مستقل و در نهایت ترکیب آنها، مورد توجه بسیاری از محققان قرار گرفته‌اند.

همان طور که ذکر شد در روش ارائه‌شده در [۲۶]، در مرحله اول، نتایج حاصل از چندین روش فیلتری متفاوت با هم ترکیب شده و سپس یک روش پیچشی فراابتکاری واحد بر روی ویژگی‌های کاهش‌یافته در مرحله

جدول ۱: پایگاه داده به کار رفته در روش پیشنهادی.

No	Dataset	#Features	#Classes	#Samples
۱	Breast Cancer	۲۴۴۸۱	۲	۹۷
۲	Colon	۲۰۰۰	۲	۶۲
۳	SRBCT	۲۳۰۸	۴	۸۳
۴	Leukemia (ALL-AML)	۷۱۲۹	۲	۷۲
۵	Prostate	۱۰۵۰۹	۲	۱۰۲
۶	Lung	۱۲۶۰۰	۵	۲۰۳
۷	Lung_Cancer	۱۲۵۳۳	۲	۱۸۱
۸	Ovarian	۱۵۱۵۴	۲	۲۵۳

اول اعمال می‌شود. می‌توان گفت در این روش به دلیل ترکیب نظرات چند الگوریتم فیلتری، تأثیر روش‌های فیلتری روی داده‌ها مؤثرتر و بیشتر از روش پیچشی به کار رفته در این روش است.

در این مقاله به ارائه روشی می‌پردازیم که با ترکیب نظرات چندین روش پیچشی به صورت خرد جمعی، تأثیر روش‌های پیچشی را روی نتایج حاصل افزایش دهد. به این صورت که در مرحله اول یک روش فیلتری مؤثر بر روی داده‌ها اعمال می‌شود تا بعد داده‌ها را به میزان زیادی کاهش دهد. سپس چندین روش پیچشی متفاوت بر روی ویژگی‌های انتخاب‌شده در مرحله اول به صورت مستقل اعمال می‌شوند. در مرحله بعد، نتایج حاصل از هر کدام از روش‌های پیچشی با یکدیگر ترکیب و ویژگی‌های منتخب نهایی را تشکیل می‌دهند. بلوک دیاگرام چهارچوب پیشنهادی در شکل ۵ نشان داده شده است.

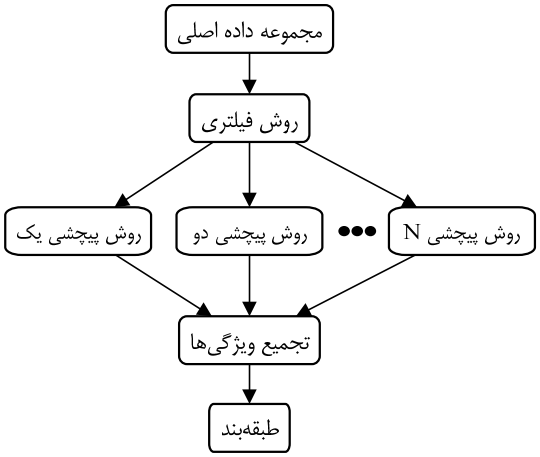
بر خلاف روش ارائه‌شده در [۲۶] که در آن، روش خرد جمعی بر روی روش‌های فیلتری اعمال شده است، روش پیشنهادی در این مقاله از روش خرد جمعی برای ترکیب نتایج الگوریتم‌های فراابتکاری بهره برده است.

در روش پیشنهادی از روش‌های فراابتکاری ترکیب‌شده با یک طبقه‌بند به عنوان روش‌های پیچشی استفاده شده است. هر کدام از الگوریتم‌های فراابتکاری در هر تکرار، زیرمجموعه‌ای بهینه از ویژگی‌ها را بر اساس تابع ارزیابی خود پیدا کرده و تا فرارسیدن شرط توقف، تکرار می‌شوند و در نهایت، بهترین زیرمجموعه ویژگی به دست آمده توسط الگوریتم فراابتکاری ارائه می‌شود و مجموعه نهایی ویژگی از تجمیع ویژگی‌های روش‌های پیچشی متفاوت به دست خواهد آمد.

در روش پیشنهادی از الگوریتم‌های فراابتکاری $ABACO_H$ و IBGSA که در [۲۶] و [۳۵] کارایی مطلوب خود را بر روی داده‌های بعد بالا نشان داده‌اند، استفاده شده است. جهت انتخاب روش فیلتری مناسب، چهار روش فیلتری "بهره اطلاعات (IG)"، "ریلیف- اف (ReliefF)"، "همبستگی سریع (FCBF) " و "امتیاز فیشر" مورد مقایسه قرار گرفته‌اند. همچنین برای انتخاب روش ترکیب (تجمیع) نتایج حاصل از روش‌های فراابتکاری، مقایسه‌ای بین نتایج دو عملگر AND و OR صورت گرفته است.

۵- آزمایش‌ها و نتایج

در این بخش، کارایی روش پیشنهادی روی ۸ مجموعه داده با بعد بالا و میکروآرایه‌ای بررسی گردیده و سپس نتایج با چندین روش انتخاب ویژگی روی داده‌های میکروآرایه‌ای، مقایسه و نتایج گزارش می‌شوند.



شکل ۵: بلوک دیاگرام چهارچوب پیشنهادشده. (از انتهای شکل، فلش حذف شد)

۵-۱ پایگاه داده

جهت انجام آزمایش‌ها از ۸ مجموعه داده میکروآرایه‌ای که در جدول ۱ ذکر شده‌اند استفاده شده است. مجموعه داده‌ها در پایگاه‌های [۳۹] و [۴۰] در دسترس هستند.

۵-۲ معیارهای ارزیابی عملکرد

برای مقایسه عملکرد الگوریتم پیشنهادی با سایر الگوریتم‌ها از معیار نرخ طبقه‌بندی صحیح، حساسیت، تشخیص، ضریب همبستگی متیو، میانگین هندسی و پارامتر کاهش ویژگی استفاده شده است.

نرخ طبقه‌بندی صحیح[1]، معادل نسبت تعداد نمونه‌های آزمون که به درستی طبقه‌بندی شده‌اند به تعداد کل نمونه‌های آزمون است و با استفاده از (۲۰) به دست می‌آید

$$Acc = \frac{a}{b} \qquad (۲۰)$$

که در آن a تعداد نمونه‌های آزمون طبقه‌بندی شده صحیح و b تعداد کل نمونه‌های آزمون است. هرچه نرخ صحت طبقه‌بندی بزرگ‌تر باشد به این معنی است که زیرمجموعه ویژگی‌های انتخاب‌شده نقش بیشتری در طبقه‌بندی صحیح داده داشته‌اند و بنابراین ویژگی‌های مناسب‌تری به شمار خواهند آمد.

دو معیار حساسیت و تشخیص نیز جهت ارزیابی کارایی طبقه‌بندی‌های دودویی طراحی شده‌اند. به طور کلی اگر کلاس‌های یک مجموعه داده دوکلاسی را به صورت کلاس‌های مثبت و منفی در نظر بگیریم و TP تعداد نمونه‌های آزمونی که به درستی در کلاس مثبت طبقه‌بندی شده‌اند، FP تعداد نمونه‌های آزمونی که به اشتباه در کلاس مثبت طبقه‌بندی شده‌اند، TN تعداد نمونه‌های آزمونی که به درستی در کلاس منفی طبقه‌بندی شده‌اند و FN تعداد نمونه‌های آزمونی که به اشتباه در کلاس منفی طبقه‌بندی شده‌اند، فرض شود در این صورت حساسیت، تشخیص، میانگین هندسی و ضریب همبستگی متیو به ترتیب به صورت زیر بیان می‌شوند

$$Sensitivity(SN) = \frac{TP}{TP + FN} \qquad (۲۱)$$

1. Classification Accuracy

$$Specificity(SP)=\frac{TN}{TN+FP} \quad (۲۲)$$

$$Gmean=\sqrt{Sensitivity\times Specificity} \quad (۲۳)$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)()}} \quad (۲۴)$$

پارامتر کاهش ویژگی [1](Fr) نیز نسبت تعداد ویژگی‌های انتخاب‌شده توسط الگوریتم به تعداد کل ویژگی‌ها می‌باشد و با رابطه زیر به دست می‌آید

$$Fr=\frac{p-q}{p} \quad (۲۵)$$

در این رابطه p تعداد کل ویژگی‌ها و q تعداد ویژگی‌های انتخاب‌شده است. طبق این رابطه، هرچه مقدار Fr به ۱ نزدیک‌تر باشد کاهش تعداد ویژگی‌ها بیشتر بوده و مطلوب‌تر است. در نظر داشته باشید که هرچه تعداد ویژگی‌های انتخاب‌شده کمتر باشد، پیچیدگی محاسباتی کمتر خواهد بود.

از آنجایی که تعداد ویژگی‌های انتخاب‌شده و پارامتر کاهش ویژگی به تنهایی نمی‌توانند معیار برتری یا ضعف یک روش باشند باید از معیارهایی همانند نرخ صحت طبقه‌بند نیز در کنار آن بهره برد. در این مقاله برای در نظر گرفتن اثر دو معیار نرخ صحت طبقه‌بندی و پارامتر کاهش ویژگی به طور هم‌زمان، از میانگین هندسی بین این دو معیار استفاده شده که به صورت زیر بیان می‌شود

$$GM=\sqrt{ACC\times Fr} \quad (۲۶)$$

۵-۳ تنظیم پارامترها

در الگوریتم IBGSA تعداد عامل‌های جمعیت برابر ۵۰ و دو ثابت k_1 و k_2 در پارامتر A به ترتیب برابر ۱ و ۵۰ در نظر گرفته شده است. در الگوریتم $ABACO_H$ نیز مقدار اعضای جمعیت ۵۰، ضریب تبخیر (ρ) برابر با ۰/۰۴۹، مقدار کمینه و بیشینه رد پا نیز به ترتیب برابر با ۰/۱ و ۶ لحاظ شده‌اند. مقدار رد پای اولیه نیز (τ) نیز ۰/۱ در نظر گرفته شده است.

تعداد تکرارها در هر یک از دو روش $ABACO_H$ و IBGSA مقدار ۵۰ و تابع شایستگی نیز نرخ طبقه‌بندی صحیح حاصل از طبقه‌بند k همسایه نزدیک‌تر ($k=1$) در نظر گرفته شده است.

۵-۴ نتایج

آزمایش‌ها به دو بخش تقسیم می‌شوند: در قسمت اول با اعمال چندین روش فیلتری در چارچوب پیشنهادشده به ارائه نتایج حاصل از آنها پرداخته می‌شود. پس از بررسی نتایج حاصل، روشی مطلوب در چارچوب پیشنهادی جهت انتخاب ویژگی در داده‌های میکروآرایه‌ای و بعد بالا ارائه می‌شود. در قسمت دوم، روش پیشنهادی با چندین روش روزآمد مقایسه و کارایی روش پیشنهادی مورد ارزیابی قرار می‌گیرد.

همان طور که ذکر شد در چهارچوب روش پیشنهادی از دو روش فراابتکاری IBGSA و $ABACO_H$ به همراه تابع شایستگی ناشی از یک طبقه‌بند به عنوان روش‌های پیچشی استفاده می‌شود. جهت انتخاب روش فیلتری و عملگر ترکیبی مناسب، روش‌های فیلتری "بهره اطلاعات"، "ریلیف-اف"، "همبستگی سریع" و "امتیاز فیشر" و همچنین دو عملگر منطقی AND و OR در چارچوب پیشنهادی مورد ارزیابی و مقایسه قرار گرفته‌اند. لازم به ذکر است که عملگر AND، زمانی یک ویژگی را انتخاب می‌کند که این ویژگی توسط هر دو الگوریتم فراابتکاری به عنوان ویژگی مطلوب انتخاب شده باشد در حالی که در عملگر OR، یک ویژگی انتخاب می‌شود در صورتی که آن ویژگی حداقل توسط یکی از الگوریتم‌ها انتخاب شده باشد.

جدول ۲ نتایج حاصل از روش پیشنهادی را بر مبنای چهار روش فیلتری مختلف و ترکیب نتایج بر اساس دو عملگر AND و OR بر روی ۵ مجموعه داده میکروآرایه نشان می‌دهد. در این جدول از روش‌هایی که از عملگر AND جهت ترکیب بهره برده‌اند به اختصار با نام HEMA[2] به همراه نام الگوریتم فیلتری مورد بررسی استفاده شده است. همچنین در روش‌هایی که از عملگر OR جهت ترکیب خرد جمعی استفاده شده است به اختصار با نام HEMO[3] به همراه نام روش فیلتری به کار رفته در آزمایش استفاده شده است.

در این آزمایش به بررسی هفت معیار شامل نرخ صحت طبقه‌بند (ACC)، تعداد ویژگی‌های انتخاب‌شده (Fr)، حساسیت (SN)، تشخیص (SP)، میانگین هندسی حساسیت و تشخیص، ضریب همبستگی متیو (MCC) و میانگین هندسی GM (میان نرخ صحت طبقه‌بند و پارامتر کاهش بعد) پرداخته شده است. جهت ارزیابی، داده‌ها به دو گروه داده‌های آموزش و داده‌های آزمون به ترتیب با نسبت ۲/۳ و ۱/۳ تقسیم شده‌اند. تمام نتایج توسط طبقه‌بند KNN ($k=1$) و پس از ۵ بار تکرار متوالی به دست آمده‌اند. همچنین قابل ذکر است که روش‌های فیلتری به کار رفته به جز روش FCBF به صورت رتبه‌بندی عمل می‌کنند. به این صورت که این روش‌ها پس از اعمال بر روی داده، به هر ویژگی رتبه‌ای اختصاص داده و در حقیقت تمام ویژگی‌های موجود در داده را رتبه‌بندی می‌کنند. در نتیجه، این روش‌ها جهت انتخاب ویژگی‌های مطلوب، نیاز به یک حد آستانه دارند که در این مقاله، حد آستانه برای همه این روش‌ها، ۰/۰۰۴ در نظر گرفته شده است.

با توجه به نتایج ذکرشده در جدول ۲ برای داده Colon با ۲۰۰۰ ویژگی و ۶۲ نمونه، روش HEMO-FCBF توانسته بیشترین مقادیر را در چهار معیار نرخ صحت طبقه‌بند، ضریب همبستگی متیو (MCC)، تشخیص و میانگین هندسی GM به دست آورد. بیشترین مقدار حساسیت (SN) و میانگین هندسی Gmean این داده توسط روش HMEBA-IG به دست آمده است.

برای داده Leukmia که داده‌ای با ۷۱۲۹ ویژگی و ۷۲ نمونه می‌باشد روش HEMO-FScore توانسته بیشترین نرخ صحت طبقه‌بند، حساسیت، میانگین هندسی GM و ضریب همبستگی متیو را کسب کند. بیشترین مقدار تشخیص و میانگین هندسی GM توسط روش HEMO-FCBF به دست آمده است.

داده Prostate با ۱۰۵۰۹ ویژگی و ۱۰۲ نمونه یکی از سخت‌ترین مجموعه داده‌های میکروآرایه‌ای است. روش HEMO-FCBF در این داده توانسته بیشترین نرخ صحت طبقه‌بند، ضریب همبستگی متیو، تشخیص، میانگین هندسی Gmean و میانگین هندسی GM را به دست آورد و بیشترین مقدار حساسیت در این داده توسط روش HEMO-ReliefF حاصل شده است.

در داده Lung_Cancer با ۱۲۵۳۳ ویژگی و ۱۸۱ نمونه، روش HEMO-FCBF توانسته بهترین نتایج را برای معیارهای صحت طبقه‌بند،

1. Feature Reduction

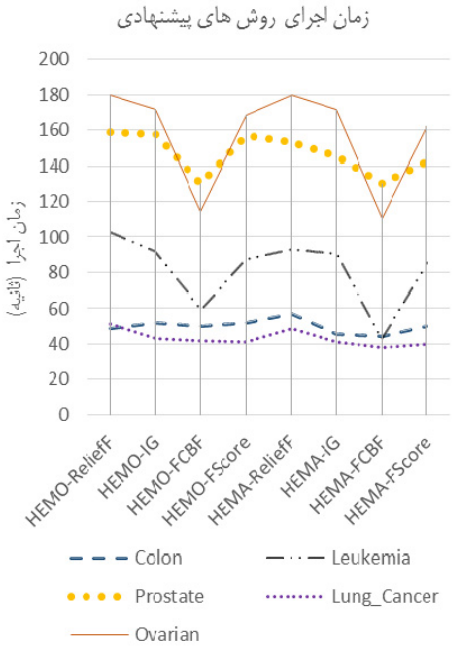
2. Hybrid Ensemble Method with and Operation

3. Hybrid Ensemble Method with or Operation

جدول ۲: نتایج روش‌های پیشنهادی بر روی ۵ پایگاه داده میکروآرایه‌ای با روش ارزیابی ۲/۳ داده آموزش و ۱/۳ داده آزمون.

			Colon	Leukemia	Prostate	Lung_Cancer	Ovarian	Average
AND	Proposed Method with ReliefF (HEMA-ReliefF)	ACC	۰٫۷۷۱۴	۰٫۹۱۶۷	۰٫۸۸۲۴	۰٫۸۵۱۲	۰٫۹۹۵۹	۰٫۸۸۳۵۲
		FS	۲٫۸	۷٫۲	۱۵٫۴	۱۵	۱۳٫۶	۱۰٫۸
		SN	۰٫۸۶۴۵	۰٫۹۳۹۸	۰٫۹۰۱۷	۰٫۸۶۲	۰٫۹۹۶۲	۰٫۹۱۲۸٤
		SP	۰٫۶۲۴۸	۰٫۹۴۴۸	۰٫۸۹۲	۰٫۸۲۱۷	۰٫۹۹۳۸	۰٫۸۵۵٤۲
		GMEAN	۰٫۷۳۴۹۴۲	۰٫۹۴۲۲۹۷	۰٫۸۹۶۸۳۷	۰٫۸۴۱۶۰۹	۰٫۹۹۴۹۹۹	۰٫۸۸۳٦٦٤
		MCC	۰٫۶۸۲۱	۰٫۹۰۲۸	۰٫۸۰۱۲	۰٫۶۳۸	۰٫۹۸۹۹	۰٫۸۰۲۸
		GM	۰٫۸۷۸	۰٫۹۵۷	۰٫۹۳۹	۰٫۹۲۲	۰٫۹۹۷	۰٫۹۳۸
	Proposed Method with IG (HEMA-IG)	ACC	۰٫۸۰۹	۰٫۹۵۰	۰٫۸۹۴	۰٫۸۹۴	۰٫۹۹۸	۰٫۹۰۹
		FS	۳٫۶۰۰	۷٫۶۰۰	۱۱٫۴۰۰	۱۶٫۲۰۰	۴٫۶۰۰	۸٫٦۸۰
		SN	۰٫۹۰۲	۰٫۹۳۶	۰٫۹۴۰	۰٫۹۴۰	۰٫۹۹۶	۰٫۹٤۳
		SP	۰٫۷۸۱	۰٫۹۱۰	۰٫۹۱۵	۰٫۹۰۲	۰٫۹۹۹	۰٫۹۰۱
		GMEAN	۰٫۸۴۰	۰٫۹۲۳	۰٫۹۲۸	۰٫۹۲۱	۰٫۹۹۷	۰٫۹۲۲
		MCC	۰٫۷۶۱	۰٫۹۰۹	۰٫۸۲۱	۰٫۸۲۰	۰٫۹۹۰	۰٫۸٦۰
		GM	۰٫۸۹۹	۰٫۹۷۴	۰٫۹۴۵	۰٫۹۴۵	۰٫۹۹۸	۰٫۹۵۲
	Proposed Method with F-Score (HEMA-FScore)	ACC	۰٫۷۵۲	۰٫۹۸۳	۰٫۸۵۹	۰٫۸۶۶	۰٫۹۸۶	۰٫۸۸۹
		FS	۲٫۶۰۰	۶٫۲۰۰	۹٫۶۰۰	۱۵٫۰۰۰	۱۴٫۰۰۰	۹٫٤۸۰
		SN	۰٫۸۰۲	۰٫۹۷۳	۰٫۹۱۲	۰٫۹۱۹	۱٫۰۰۰	۰٫۹۲۱
		SP	۰٫۷۸۰	۰٫۹۶۰	۰٫۸۴۹	۰٫۸۵۰	۰٫۹۶۰	۰٫۸۸۰
		GMEAN	۰٫۷۹۱	۰٫۹۶۶	۰٫۸۸۰	۰٫۸۸۴	۰٫۹۸۰	۰٫۹۰۰
		MCC	۰٫۶۵۱	۰٫۹۲۴	۰٫۸۰۷	۰٫۷۹۱	۰٫۹۴۱	۰٫۸۲۳
		GM	۰٫۸۶۷	۰٫۹۹۱	۰٫۹۲۶	۰٫۹۳۰	۰٫۹۹۲	۰٫۹٤۱
	Proposed Method with FCBF (HEMA-FCBF)	ACC	۰٫۷۱۲	۰٫۹۵۹	۰٫۸۵۳	۰٫۸۶۱	۰٫۹۹۳	۰٫۸۷٦
		FS	۴٫۰۰۰	۴٫۰۰۰	۱۲٫۸۰۰	۱۴٫۱۰۰	۸٫۴۰۰	۸٫٦٦۰
		SN	۰٫۷۰۸	۰٫۹۷۱	۰٫۸۹۸	۰٫۸۵۱	۱٫۰۰۰	۰٫۸۸٦
		SP	۰٫۶۹۱	۰٫۹۴۲	۰٫۸۹۷	۰٫۸۹۱	۱٫۰۰۰	۰٫۸۸٤
		GMEAN	۰٫۶۹۹	۰٫۹۵۶	۰٫۸۹۷	۰٫۸۷۱	۱٫۰۰۰	۰٫۸۸۵
		MCC	۰٫۶۱۱	۰٫۹۲۴	۰٫۷۸۱	۰٫۷۵۲	۰٫۹۸۵	۰٫۸۱۰
		GM	۰٫۸۴۳	۰٫۹۷۹	۰٫۹۲۳	۰٫۹۲۷	۰٫۹۹۶	۰٫۹۳۳
OR	Proposed Method with ReliefF (HEMO-ReliefF)	ACC	۰٫۷۹۵	۰٫۹۴۶	۰٫۹۲۱	۰٫۹۰۸	۰٫۹۹۳	۰٫۹۱۳
		FS	۴٫۴۰۰	۲۰٫۲۰۰	۳۲٫۸۰۰	۳۴٫۱۰۰	۴۳٫۸۰۰	۲۷٫۰٦۰
		SN	۰٫۸۳۶	۰٫۹۶۷	۰٫۹۴۶	۰٫۹۰۸	۰٫۹۸۹	۰٫۹۲۹
		SP	۰٫۷۹۲	۰٫۹۲۰	۰٫۹۱۲	۰٫۸۲۱	۰٫۹۹۸	۰٫۸۸۹
		GMEAN	۰٫۸۱۳	۰٫۹۴۳	۰٫۹۲۹	۰٫۸۶۴	۰٫۹۹۴	۰٫۹۰۹
		MCC	۰٫۶۰۱	۰٫۹۱۵	۰٫۸۹۳	۰٫۷۹۸	۰٫۹۸۴	۰٫۸۳۸
		GM	۰٫۸۹۱	۰٫۹۷۱	۰٫۹۵۸	۰٫۹۵۲	۰٫۹۹۵	۰٫۹۵۳
	Proposed Method with IG (HEMO-IG)	ACC	۰٫۸۰۵	۰٫۹۳۸	۰٫۸۷۴	۰٫۹۲۶	۰٫۹۹۸	۰٫۹۰۸
		FS	۴٫۰۰۰	۲۰٫۶۰۰	۳۰٫۰۰۰	۳۲٫۸۰۰	۱۳٫۲۰۰	۲۰٫۱۲۰
		SN	۰٫۸۴۲	۰٫۹۴۱	۰٫۸۶۳	۰٫۹۲۲	۰٫۹۹۶	۰٫۹۱۳
		SP	۰٫۷۴۹	۰٫۹۲۸	۰٫۸۷۲	۰٫۸۴۲	۱٫۰۰۰	۰٫۸۷۸
		GMEAN	۰٫۷۹۴	۰٫۹۳۴	۰٫۸۶۷	۰٫۸۸۱	۰٫۹۹۸	۰٫۸۹۵
		MCC	۰٫۶۶۹	۰٫۹۲۱	۰٫۸۳۲	۰٫۷۸۲	۰٫۹۹۵	۰٫۸٤۰
		GM	۰٫۸۹۶	۰٫۹۶۷	۰٫۹۳۳	۰٫۹۶۱	۰٫۹۹۸	۰٫۹۵۱
	Proposed Method with F-Score (HEMO-FScore)	ACC	۰٫۸۰۰	۰٫۹۸۳	۰٫۸۴۴	۰٫۹۳۵	۰٫۹۹۸	۰٫۹۱۲
		FS	۳٫۴۰۰	۲۳٫۱۰۰	۲۹٫۶۰۰	۳۲٫۲۰۰	۴۵٫۰۰۰	۲٦٫٦٦۰
		SN	۰٫۸۳۸	۰٫۹۸۰	۰٫۹۰۹	۰٫۹۴۶	۰٫۹۹۶	۰٫۹۳٤
		SP	۰٫۷۸۳	۰٫۹۳۵	۰٫۸۱۲	۰٫۹۰۱	۱٫۰۰۰	۰٫۸۷٤
		GMEAN	۰٫۸۱۰	۰٫۹۵۷	۰٫۸۵۹	۰٫۹۲۳	۰٫۹۹۸	۰٫۹۰۳
		MCC	۰٫۶۵۱	۰٫۹۳۱	۰٫۷۵۸	۰٫۸۹۲	۰٫۹۹۵	۰٫۸٤۵

	GM	۰/۸۹۴	۰/۹۹۰	۰/۹۱۷	۰/۹۶۶	۰/۹۹۷	۰/۹۵۳
Proposed Method with FCBF (HEMO-FCBF)	ACC	۰/۸۱۴	۰/۹۴۱	۰/۹۲۹	۰/۹۴۳	۱/۰۰۰	۰/۹۲۵
	FS	۷/۰۰۰	۵/۰۰۰	۲۶/۴۰۰	۱۸/۷۵۰	۱۷/۲۰۰	۱۴/۸۷۰
	SN	۰/۸۵۵	۰/۹۷۵	۰/۹۴۶	۰/۹۵۲	۱/۰۰۰	۰/۹۴۶
	SP	۰/۷۸۵	۰/۹۶۵	۰/۹۳۱	۰/۹۰۴	۱/۰۰۰	۰/۹۱۵
	GMEAN	۰/۸۲۰	۰/۹۷۰	۰/۹۳۸	۰/۹۲۸	۱/۰۰۰	۰/۹۳۰
	MCC	۰/۷۷۲	۰/۹۱۲	۰/۹۰۲	۰/۸۹۱	۱/۰۰۰	۰/۸۹۵
	GM	۰/۹۰۱	۰/۹۷۰	۰/۹۶۳	۰/۹۷۰	۰/۹۹۹	۰/۹۶۱



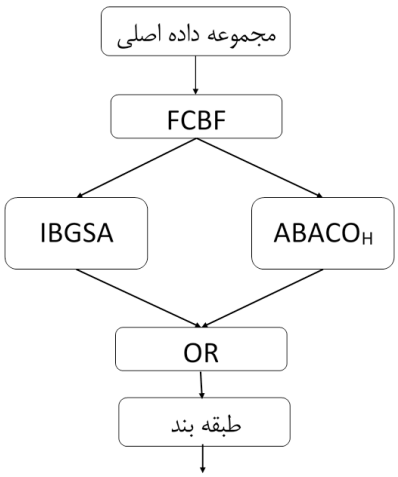
شکل ۶: نمودار زمان اجرای روش‌های پیشنهادی.

حساسیت، تشخیص، میانگین هندسی Gmean و میانگین GM کسب کند. بیشترین مقدار معیار همبستگی متیو در این داده توسط روش HEMO-FScore به دست آمده است.

برای داده Ovarian که با ۱۵۱۵۴ ویژگی و ۲۵۳ نمونه، بیشترین تعداد ویژگی را در میان داده‌های مورد بررسی به خود اختصاص داده، روش HEMO-FCBF توانسته در معیارهای صحت طبقه‌بند، حساسیت، تشخیص، میانگین هندسی Gmean و ضریب همبستگی متیو (MCC) مقدار ۱/۰۰ را کسب کند. میانگین هندسی، GM، این روش روی این داده مقدار ۰/۹۹۹ است که در مقایسه با سایر روش‌ها بر روی این داده بهترین نتیجه محسوب می‌شود. بررسی میانگین مقادیر به دست آمده از ۵ معیار ذکرشده در جدول ۲ نیز نشان می‌دهد که روش HEMO-FCBF توانسته در ۵ مجموعه داده مورد بررسی به برتری مطلوبی دست پیدا کند.

شکل ۶ نمودار زمان اجرای روش‌های پیشنهادی در جدول ۲ را بر روی ۵ مجموعه داده نشان می‌دهد. همان طور که مشاهده می‌شود دو روش HEMA-FCBF و HEMO-FCBF در زمان اجرا توانسته‌اند بهترین عملکرد را از خود نشان دهند. اگرچه روش HEMA-FCBF از لحاظ زمان اجرا توانسته نتیجه بهتری را کسب کند، با ارزیابی نتایج به دست آمده در جدول ۲ می‌توان به این نکته دست یافت که روش HEMO-FCBF کارایی و عملکرد بهتری در مواجهه با داده‌های میکروآرایه‌ای و بعد بالا داشته است.

با توجه به برتری روش HEMO-FCBF نسبت به سایر روش‌های مورد بررسی، بلوک دیاگرام پیشنهادی به صورت شکل ۷ می‌باشد.

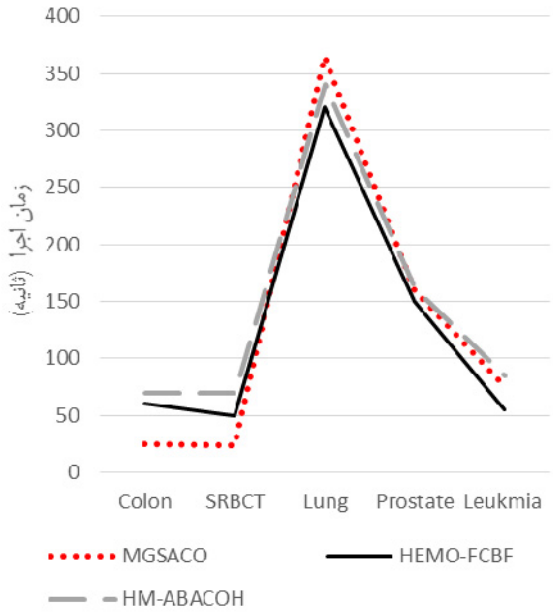


شکل ۷: بلوک دیاگرام روش پیشنهادشده.

جدول ۳: نتایج روش پیشنهادی با روش‌های ارائه‌شده در [۲۶] و [۲۷] بر اساس نرخ طبقه‌بندی صحیح با روش ارزیابی ۲/۳ داده آموزش و ۱/۳ داده آزمون. اعداد درون پرانتز پارامتر کاهش بعد را نشان می‌دهند.

Datasets	[۲۷] MGSACO	[۲۶] HM-ABACOH	HEMO-FCBF
Colon	۰/۸۰ (۰/۹۹۰)	۰/۸۰ (۰/۹۹۶)	**۰/۸۱۴۳ (۰/۹۹۷)**
SRBCT	۰/۸۴۱ (۰/۹۹۱)	۰/۹۳۸۹ (۰/۹۹۱)	**۰/۹۶۵۰ (۰/۹۹۲)**
Leukemia	۰/۹۲۳۱ (۰/۹۹۷)	**۰/۹۴۱۶** (۰/۹۹۸)	۰/۹۴۱۱ **(۰/۹۹۹)**
Prostate	۰/۸۶۶۲ **(۰/۹۹۸)**	۰/۸۷۳ **(۰/۹۹۸)**	**۰/۹۲۹۴** (۰/۹۹۷)
Lung	۰/۸۰ **(۰/۹۹۸)**	۰/۹۰۷۹ (۰/۹۹۷)	**۰/۹۴۲۶ (۰/۹۹۸)**
Average	۰/۸۶۷۹ (۰/۹۹۴۸)	۰/۸۹۲۳ (۰/۹۹۶)	**۰/۹۱۸۵ (۰/۹۹۶۶)**

در جدول ۳ میانگین نرخ صحت طبقه‌بند و کاهش ویژگی در روش پیشنهادی HEMO-FCBF (توسط طبقه‌بند KNN) با روش‌های ارائه‌شده در [۲۶] و [۲۷] بر روی پنج بار اجرای مستقل الگوریتم‌های مذکور نشان داده شده است. روش ارزیابی داده‌ها در این آزمایش نیز تقسیم داده‌ها به دو گروه داده‌های آموزش و آزمون به ترتیب با نسبت‌های ۲/۳ و ۱/۳ می‌باشد. قابل ذکر است که بیشترین نرخ صحت طبقه‌بندی در [۲۷] مربوط به طبقه‌بند بیز ساده است که نتایج آن در جدول ۳ نشان داده شده است. در این جدول، روش پیشنهادی در [۲۶] به اختصار با نام HM-ABACOH نشان داده شده و اعداد ذکرشده داخل پرانتز پارامتر کاهش ویژگی توسط هر روش را بیان می‌کنند.



شکل ۸: زمان اجرای ۳ روش بر روی ۵ دیتاست میکروآرایه‌ای با توجه به جدول ۳.

همان طور که مشاهده می‌شود روش پیشنهادی HEMO-FCBF در چهار پایگاه داده Colon، SRBCT، Prostate و Lung به نتایج بهتری در نرخ طبقه‌بندی صحیح نسبت به دو روش دیگر دست یافته است. همچنین نرخ کاهش ویژگی روش HMEBO-FCBF در چهار دیتاست Colon، SRBCT، Leukemia و Lung نسبت به دو روش دیگر به برتری دست یافته است.

در شکل ۸ زمان اجرای روش پیشنهادی و دو روش ذکرشده در جدول ۳ آمده است. با توجه به نمودار زمان اجرا، روش HEMO-FCBF در ۳ داده Lung، Prostate و Leukmia زمان اجرای کمتری نسبت به دو روش دیگر داشته که با توجه به نتایج حاصل از این اجراها که در جدول ۳ ذکر شده نشان می‌دهد که روش HEMO-FCBF می‌تواند نتایج قابل قبولی بر روی داده‌های بعد بالا و میکروآرایه‌ای داشته باشد.

در جدول ۴ روش پیشنهادی HEMO-FCBF با روش APCES که در سال ۲۰۱۶ در [۲۵] ارائه شده، مقایسه گردیده است. در این جدول، میانگین ۴ معیار حساسیت، تشخیص، نرخ صحت طبقه‌بند و میانگین هندسی Gmean مورد بررسی قرار گرفته‌اند. لازم به ذکر است که به جهت یکسان‌بودن شرایط آزمایش‌ها با نتایج روش مذکور در [۲۵]، مقادیر ذکرشده در جدول ۴ با روش ارزیابی ۵-CV[1] و روی ۱۰ بار اجرای مستقل الگوریتم، حاصل شده است.

با توجه به جدول ۴ معیار حساسیت روش پیشنهادی HEMO-FCBF در ۴ داده Colon، ALL-AML، Lung_Cancer و Ovarian نتایج بهتری نسبت به روش APCES کسب نموده است. در معیار تشخیص و میانگین هندسی Gmean روش پیشنهادی توانسته در تمامی دیتاست‌ها به برتری دست یابد. در معیار نرخ صحت طبقه‌بند نیز روش پیشنهادی HEMO-FCBF توانسته جز در دو پایگاه داده Breast Cancer و Colon در ۳ پایگاه داده دیگر به برتری دست یابد.

۵-۵ تحلیل نتایج و بحث

همان طور که ذکر شد در روش‌های خرد جمعی می‌توان از مزیت‌های چندین روش انتخاب ویژگی بهره برد. همچنین با افزایش میزان تأثیر روش‌های فراابتکاری به کار برده شده در الگوریتم می‌توان به نتایج مطلوب‌تری دست یافت. در این مقاله که هدف اصلی آن افزایش میزان تأثیر روش‌های فراابتکاری در روش‌های ترکیبی می‌باشد به ارائه یک روش ترکیبی با بهره‌گیری از ۲ روش فراابتکاری IBGSA و $ABACO_H$ در قالب روش‌های خرد جمعی پرداخته شده است. با توجه به نتایج ذکرشده از چندین روش پیشنهادی در جدول ۲ روش‌هایی همچون HEMO-FCBF، HEMA-IG و HEMO-FScore نتایج مطلوبی را بر روی ۵ پایگاه داده میکروآرایه‌ای کسب کرده‌اند. در این میان، روش HEMO-FCBF عملکرد بهتری نسبت به سایر روش‌های مورد بررسی کسب کرده است. از لحاظ زمانی نیز گرچه این روش رتبه دوم را از لحاظ پایین‌بودن زمان اجرا به خود اختصاص داده اما با در نظر گرفتن معیارهایی چون نرخ صحت طبقه‌بند، تشخیص و حساسیت، این روش را می‌توان به عنوان روش برتر در نظر گرفت.

در مقایسه روش HEMO-FCBF با دیگر روش‌های به روز در انتخاب ویژگی در داده‌های میکروآرایه‌ای، نتایج این روش ابتدا با روش‌های MGSACO و $HM\text{-}ABACO_H$ بر اساس سه معیار نرخ صحت طبقه‌بند، نرخ کاهش ویژگی و از لحاظ زمان اجرای الگوریتم مقایسه شده است. با توجه به نتایج ذکرشده در جدول ۳ روش پیشنهادی HEMO-FCBF در دو معیار نرخ صحت طبقه‌بند و پارامتر کاهش ویژگی در ۴ پایگاه داده از ۵ پایگاه مورد بررسی قرار گرفته نسبت به دو روش دیگر به برتری دست یافته است. با توجه به نمودار زمان اجرای الگوریتم‌های ذکرشده در جدول ۳ روش پیشنهادی زمان اجرایی قابل قبولی نسبت به دو روش دیگر کسب کرده است.

از جمله دیگر روش‌های روزآمد در زمینه انتخاب ویژگی در داده‌های بعد بالا روش APCES می‌باشد. در جدول ۴ نتایج حاصل از اعمال روش APCES بر اساس ۴ معیار نرخ صحت طبقه‌بند، حساسیت، تشخیص و میانگین هندسی Gmean بر روی ۵ پایگاه داده میکروآرایه‌ای (که نتایج آن در [۲۶] ذکر شده است) با روش HEMO-FCBF مورد بررسی قرار گرفته است. با توجه به نتایج مذکور در جدول ۴ روش پیشنهادی توانسته در دو معیار تشخیص و میانگین هندسی Gmean در هر ۵ پایگاه داده نتایج بهتری نسبت به روش APCES کسب کند.

در معیار حساسیت روش پیشنهادی توانسته به جز در پایگاه داده Breast Cancer در ۴ پایگاه داده دیگر به برتری دست پیدا کند. در معیار نرخ صحت طبقه‌بند روش APCES در دو پایگاه داده Breast Cancer و Colon توانسته نتایج بهتری نسبت به روش HEMO-FCBF کسب کند در حالی که در سه پایگاه داده دیگر ALL-AML، Lung_Cancer و Ovarian روش HEMO-FCBF به نتایج بهتری نسبت به روش APCES دست پیدا کرده است.

همان طور که مشاهده می‌شود میانگین در هر چهار معیار حساسیت، تشخیص، میانگین هندسی Gmean و نرخ صحت طبقه‌بندی نیز برتری روش HEMO-FCBF را نشان می‌دهند.

با توجه به رویکرد ارائه‌شده در این مقاله و بررسی نتایج حاصل از آن می‌توان نتیجه گرفت که استفاده از روش‌های فراابتکاری به صورت خرد جمعی و سپس به کارگیری آن در روش‌های ترکیبی به منظور کاهش پیچیدگی‌های محاسباتی و زمانی، می‌تواند نتایج مطلوبی در روند انتخاب ویژگی بر روی داده‌های با بعد بالا داشته باشد. دلیل عملکرد بهتر روش پیشنهادی نسبت به سایر روش‌های مورد بررسی، استفاده از نظرات چندین روش فراابتکاری به صورت خرد جمعی پس از کاهش چشم‌گیر بعد داده توسط روش فیلتری می‌باشد.

1. 5-Fold Cross-Validation

جدول ۴: نتایج دو روش HEMO-FCBF و روش APCES [۲۶] (آزمایش‌ها با روش ارزیابی CV−۵ انجام گرفته‌اند).

Dataset	SN		SP		ACC		Gmean	
	[۲۶] APCES	HEMO-FCBF	[۲۶] APCES	HEMO-FCBF	[۲۶] APCES	HEMO-FCBF	[۲۶] APCES	HEMO-FCBF
Breast Cancer	۰٫۸۲۷	۰٫۷۸۶	۰٫۶۴	**۰٫۷۰۵۲**	۰٫۷۳۸	۰٫۷۰۲	۰٫۷۱۸	**۰٫۷۳٤۷۸**
Colon	۰٫۸۵	**۰٫۸٦۲**	۰٫۸۲۲	**۰٫۸۳**	۰٫۸۴	۰٫۸۳۷۲	۰٫۸۲۶	**۰٫۸٤۵۸٤۹**
ALL-AML	۰٫۹۵۸	**۰٫۹۸۱۸**	۰٫۸۲۲	**۰٫۹۹۰٤**	۰٫۹۷۵	**۰٫۹۹۲۵**	۰٫۹۷	**۰٫۹۹۰۸۵۸**
Lung_Cancer	۰٫۹۴	**۰٫۹۹۳۵**	۰٫۹۸۴	**۱**	۰٫۹۴۲	**۰٫۹۸٤**	۰٫۹۵۵	**۰٫۹۹٦۷٤۵**
Ovarian	۰٫۹۹۲	**۱**	۰٫۹۸۸	**۱**	۰٫۹۹	**۱**	۰٫۹۹۱	**۱**
Average	۰٫۹۱۳	**۰٫۹۲۵**	۰٫۸۵۱	**۰٫۹۰۵**	۰٫۸۹۷	**۰٫۹۰۳**	۰٫۸۹۲	**۰٫۹۱٤**

به عبارت دیگر در این مقاله سعی شده که در مرحله اول با کاهش بعد بسیار در داده‌های با بعد بالا و حذف ویژگی‌های اضافی و نامرتبط، زمینه را برای حضور روش‌های پیچشی مهیا کرده و سپس در قسمت دوم با اعمال دو روش فراابتکاری، از مزیت‌های به کارگیری روش‌های پیچشی به صورت خرد جمعی استفاده شود. افزایش تأثیر و تنوع در روش‌های فراابتکاری می‌تواند در روند انتخاب ویژگی در داده‌های بعد بالا نتایج بسیار مطلوبی در پی داشته باشد و باعث افزایش قدرت الگوریتم گردد.

٦- جمع‌بندی

با بالارفتن ابعاد داده، تأثیر اساسی مبحث انتخاب ویژگی در یادگیری ماشینی و داده‌کاوی نمایان می‌شود. امروزه می‌توان انتخاب ویژگی را یکی از اساسی‌ترین مراحل در یادگیری ماشینی دانست. در رویارویی با داده‌های بزرگ و داده‌های با بعد بالا علاوه بر معضل نفرین بعد که گیج‌شدن طبقه‌بند را به همراه دارد، همواره ویژگی‌های افزونه و نامرتبط یکی از مهم‌ترین عوامل در کاهش دقت طبقه‌بندی هستند که حذف آنها کمک شایانی به افزایش دقت طبقه‌بند می‌کند. از این رو با افزایش بعد در این داده‌ها، انتخاب ویژگی را می‌توان یک روند غیر قابل حذف در مواجهه با این داده‌ها دانست. حال با افزایش روز به روز حجم این داده‌ها، روش‌های قدیمی انتخاب ویژگی دیگر توانایی رویارویی با این داده‌ها را نخواهند داشت و نمی‌توانند مانند داده‌های کلاسیک بر روی داده‌های با بعد بالا نیز مؤثر باشند.

در روش‌های خرد جمعی، نتایج حاصل از چندین روش متفاوت با هم ترکیب شده و پاسخ هر یک از روش‌ها بر روی نتایج نهایی تأثیرگذار می‌باشد. در نتیجه این روش‌ها می‌توانند در انتخاب ویژگی‌های مطلوب و همچنین حذف ویژگی‌های اضافی و نامرتبط بسیار تأثیرگذار باشند.

روش‌های پیچشی از نرخ صحت طبقه‌بندی بهتری نسبت به روش‌های فیلتری بهره‌مند هستند ولی به دلیل پیچیدگی محاسباتی بالا، راه حل مناسبی برای کاهش بعد در داده‌های با بعد بالا نمی‌باشند. از این روش‌ها می‌توان در قالب روش‌های ترکیبی بهره برد به این صورت که ابتدا توسط روش‌های فیلتری به کاهش بعد داده‌ها پرداخته می‌شود و سپس توسط روش‌های پیچشی ویژگی‌های مطلوب با دقت بالاتری انتخاب خواهند شد. در این میان استفاده از روش‌های خرد جمعی در قالب روش‌های ترکیبی نیز می‌تواند رویکرد مناسبی در رویارویی با داده‌های با بعد بالا باشد.

در این مقاله با ارائه روشی مبتنی بر روش‌های ترکیبی و با بهره‌گیری از روش‌های خرد جمعی به انتخاب ویژگی در داده‌های با بعد بالا پرداخته شده است. هدف اصلی در این مقاله افزایش میزان تأثیر و تنوع روش‌های فراابتکاری در روند انتخاب ویژگی در داده‌های با بعد بالا است. برای بررسی کارایی روش ارائه‌شده از ۸ پایگاه داده میکروآرایه‌ای که همگی مجموعه داده‌هایی با بعد بالا هستند استفاده شده است. ارزیابی کارایی روش پیشنهادی با دیگر روش‌ها بر اساس نرخ صحت طبقه‌بند و نرخ کاهش ویژگی صورت گرفته که نتایج به دست آمده نشان از برتری روش پیشنهادی نسبت به دیگر روش‌های ارائه‌شده دارد.

مراجع

[1] M. M. Kabir, M. Shahjahan, and K. Murase, "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, no. 17, pp. 2914-2928, Oct. 2011.

[2] 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, 2011.

[3] S. Rakkeitwinai, C. Lursinsap, C. Aporntewan, and A. Mutirangura, "New feature selection for gene expression classification based on degree of class overlap in principal dimensions," *Computers in Biology and Medicine*, vol. 64, pp. 292-298, Sept. 2015.

[4] Z. Zhao and H. Liu, "Searching for interacting features," in *Proc. of the 20th Inte. Joint Conf. on Artifical intelligence*, *IJCAI'07*, pp. 1156-1161, Hyderabad, India, 6-12 Jan. 2007.

[5] A. J. Ferreira and M. A. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition*, vol. 45, no. 9, pp. 3048-3060, Sept. 2012.

[6] I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh, "Stable feature selection for clinical prediction: exploiting ICD tree structure using tree-lasso," *J. of Biomedical Informatics*, vol. 53, pp. 277-290, Feb. 2015.

[7] M. Liu and D. Zhang, "Feature selection with effective distance," *Neurocomputing*, vol. 215, pp. 100-109, Nov. 2016.

[8] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," in *Proc. of the 21st Australasian Computer Science Conference ACSC'98,* pp. 181-191, Perth, Australia, 4-6 Feb. 1998.

[9] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," arXiv preprint arXiv: 1202.3725, 2012.

[10] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," in *Proc. European Conf. on Machine Learning*, pp. 171-182, 1994.

[11] 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.

[12] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning*, The University of Waikato, Ph.D. Thesis, 1999.

[13] A. Colorni, M. Dorigo, and V. Maniezzo, "Distributed optimization by ant colonies," in *Proc. of the 1st European Conf. on Artificial Life*, pp. 134-142, Paris, France, 1991.

[14] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Engineering Applications of Artificial Intelligence*, vol. 32, pp. 112-123, Jun. 2014.

[15] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognition*, vol. 45, no. 1, pp. 531-539, Jan. 2012.

[16] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "GSA: a gravitational search algorithm," *Information Sciences*, vol. 179, no. 13, pp. 2232-2248, 13 Jun. 2009.

[17] K. S. Tang, K. F. Man, S. Kwong, and Q. He, "Genetic algorithms and their applications," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 22-37, Nov. 1996.
[18] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*, Ed: Springer, pp. 760-766, 2011.
[19] 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.
[20] 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.
[21] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 754-764, May/Jun. 2012.
[22] 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.
[23] S. Shreem, S. Abdullah, M. Nazri, and M. Alzaqebah, "Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection," *J. of Theoretical and Applied Information Technology*, vol. 46, no. 2, pp. 1034-1039, Dec. 2012.
[24] 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.
[25] J. Meng, H. Hao, and Y. Luan, "Classifier ensemble selection based on affinity propagation clustering," *J. of Biomedical Informatics*, vol. 60, pp. 234-242, Apr. 2016.
[26] A. Rouhi and H. Nezamabadi-pour, "A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm," in *Proc. 1st Conf. on Swarm Intelligence and Evolutionary Computation, CSIEC'16,* pp. 70-75, Bam, Iran, 9-11 Mar 2016.
[27] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization," *Neurocomputing*, vol. 168, pp. 1024-1036, 30 Nov. 2015.
[28] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "Data classification using an ensemble of filters," *Neurocomputing*, vol. 135, pp. 13-20, 5 Jul. 2014.
[29] S. M. Vieira, J. M. Sousa, and U. Kaymak, "Fuzzy criteria for feature selection," *Fuzzy Sets and Systems*, vol. 189, no. 1, pp. 1-18, 16 Feb. 2012.
[30] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing*, vol. 105, no. 7, pp. 3-11, 1 Apr. 2013.
[31] A. B. Brahim and M. Limam, "Robust ensemble feature selection for high dimensional data sets," in *Proc. Int. Conf. on High Performance Computing and Simulation, HPCS'13,* pp. 151-157, Helsinki, Finland, 1-5 Jul. 2013.
[32] L. Y. Chuang, C. H. Yang, K. C. Wu, and C. H. Yang, "A hybrid feature selection method for DNA microarray data," *Computers in Biology and Medicine*, vol. 41, no. 4, pp. 228-237, Apr. 2011.
[33] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
[34] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Proc, of the 10th National Conf. on Artificial Intelligence*, *AAAI'92,* vol. 2, pp. 129-134, San Jose, CA, USA, 12-16 Jul. 1992.
[35] N. Taheri and H. Nezamabadi-pour, "A hybrid feature selection method for high-dimensional data," in *Proc. 4th Int. eConf. on Computer and Knowledge Engineering, ICCKE'14*, pp. 141-145, Mashhad, Iran, 29-30 Oct. 2014.
[36] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "BGSA: binary gravitational search algorithm," *Natural Computing*, vol. 9, no. 3, pp. 727-745, Sept. 2010.
[37] E. Rashedi and H. Nezamabadi-pour, "Feature subset selection using improved binary gravitational search algorithm," *J. of Intelligent and Fuzzy Systems*, vol. 26, no. 3, pp. 1211-1221, 2014.
[38] S. Kashef and H. Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection," *Neurocomputing*, vol. 147, pp. 271-279, 5 Jan. 2015.
[39] A. Statnikov, C. F. Aliferis, and I. Tsamardinos, *Gems: Gene Expression Model Selector*, Available: http://www.gems-system.org, 2005.
[40] Datasets, Available on http://datam.i2r.a-star.edu.sg/datasets/krbd.

امیررضا روحی در سال ۱۳۹۵ مدرک کارشناسی مهندسی برق مخابرات خود را از دانشگاه شهید باهنر کرمان دریافت نمود. نام‌برده در سال ۱۳۹۵ به آزمایشگاه پردازش داده هوشمند (IDPL) دانشگاه شهید باهنر کرمان پیوست. او هم‌اکنون دانشجوی دوره دکتری در دانشگاه پلی‌تکنیک میلان ایتالیا است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: بیوانفورماتیک، بازشناسی الگو، و انتخاب ویژگی.

حسین نظام‌آبادی‌پور دوره کارشناسی خود را در مهندسی برق- الکترونیک در دانشگاه شهید باهنر کرمان در سال ۱۳۷۷ به پایان رساند. پس از آن، مدارک کارشناسی ارشد و دکتری خود را نیز در مهندسی برق-الکترونیک از دانشگاه تربیت مدرس به ترتیب در سال‌های ۱۳۷۹ و ۱۳۸۳ دریافت کرد. وی هم‌اکنون استاد بخش مهندسی برق دانشگاه شهید باهنر کرمان است. زمینه‌های پژوهشی مورد علاقه‌ی او پردازش تصویر، بازشناسی الگو، کاربرد رایانش نرم در پردازش تصویر و روشهای بهینه‌سازی ابتکاری است..