نشریه علمی ترویجی محاسبات نرم
سال ششم/ شمارهٔ دوم/ پاییز و زمستان ۱۳۹۶/ صفحهٔ ۴۸ـ۵۹

دریافت مقاله: ۱۳۹۵/۷/۱۷
پذیرش مقاله: ۱۳۹۶/۱۱/۲۸

ارائهٔ الگوریتم ترکیبی پالایشی- پوششی انتخاب ویژگی و کاربرد آن در کاهش بعد داده‌های بیان ژن

زهرا روزبهانی[۱]، جلال رضایی‌نور[۲]، منصوره یاری ایلی[۳]، راضیه قیاسی[۴]

[۱] دانشجوی دکترای مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران
z.roozbahani@stu.qom.ac.ir
[۲] دانشیار و عضو هیئت‌علمی گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران
J.rezaee@qom.ac.ir
[۳] دانشجوی دکترای مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران
m.yari@stu.qom.ac.ir
[۴] دانشجوی دکترای مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم،قم، ایران
raziehghiasi@gmail.com

چکیده: امروزه بالا رفتن حجم داده‌ها و تعداد ویژگی‌ها در مجموعهٔ داده، باعث کاهش دقت الگوریتم یادگیری و پیچیدگی محاسباتی شده است. روش‌های کاهش بعد، نوعی از روش انتخاب مشخصه هستند که به دو صورت پالایشی و پوششی انجام می‌شود. دقت روش‌های پوششی نسبت به روش‌های پالایشی بیشتر است اما در مقابل، روش‌های پالایشی سریع‌تر عمل می‌کنند و پیچیدگی‌های محاسباتی کمتری دارند. با در نظر گرفتن مزایا و معایب الگوریتم‌های پالایشی و پوششی، در این پژوهش یک روش ترکیبی جدید ارائه شده است. در این روش، ابتدا کل مشخصه‌های موجود در مجموعهٔ داده در نظر گرفته می‌شوند سپس با ترکیب الگوریتم‌های پالایشی انتخاب مشخصه و ارزش‌گذاری نتایج آن به روش پوششی، زیرمجموعه‌ای بهینه از مشخصه‌ها انتخاب می‌شوند. با توجه به اینکه بسیاری از بیماری‌ها و مسائل زیست‌سیستمی نظیر سرطان، به کمک بررسی دادهٔ ریزآرایه قابل شناسایی و تشخیص هستند و با توجه به اینکه تعداد مشخصه‌ها در این مجموعه داده‌ها بسیار زیاد است، روش ارائه‌شده در این پژوهش روی دادهٔ ریزآرایهٔ مربوط به سه نوع سرطان مورد ارزیابی قرار گرفته است. این روش، در مقایسه با روش‌های مشابه، به دقت بیشتری در دسته‌بندی و شناسایی عوامل مؤثر در سرطان، به‌خصوص سرطان خون دست یافته است.

واژه‌های کلیدی: کاهش بعد، انتخاب مشخصه، الگوریتم پالایشی- پوششی، شبکهٔ عصبی پرسپترون چندلایه، ریزآرایه.

۱. مقدمه

پیشرفت‌های به‌وجودآمده در جمع‌آوری داده و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از اطلاعات روبه‌رو شویم. بسترهای داده‌ای که ابعاد بالایی دارند به‌رغم فرصت‌هایی که به وجود می‌آورند، چالش‌های محاسباتی زیادی را ایجاد می‌کنند. یکی از مشکلات داده‌های با ابعاد بالا این است که در بیشتر مواقع، تمام ویژگی‌ها برای یافتن دانشی که در داده‌ها نهفته است مهم و حیاتی نیست. برای انتخاب یک زیرمجموعهٔ مناسب از مجموعه ویژگی‌ها می‌توان از روش‌های مبتنی بر کاهش بعد[1] و روش‌های مبتنی بر انتخاب ویژگی[2] استفاده کرد. روش‌های مبتنی بر کاهش بعد با ترکیب مقادیر ویژگی‌های موجود، تعداد کمتری ویژگی به وجود می‌آورند به‌نحوی که این ویژگی‌های جدید بتوانند تمام یا بخش اعظمی از اطلاعات موجود را توصیف کنند. از جمله این روش‌ها PCA است. مشکل بزرگی که در به‌کارگیری این روش‌ها وجود دارد، این است که نمی‌توان هیچ‌یک از ویژگی‌ها را از فضای ورودی حذف کرد. فودر [۱] نگاهی اجمالی به همهٔ روش‌های کاهش بعد مبتنی بر استخراجِ مشخصه داشته است. در مقابل روش‌های استخراج مشخصه، روش‌های مبتنی بر انتخاب مشخصه سعی می‌کنند با انتخاب زیرمجموعه‌ای مناسب از مشخصه‌های اولیه، ابعاد داده را کاهش دهند. روش انتخاب ویژگی به دو دستهٔ پالایشی[3] و پوششی[4] تقسیم می‌شوند.

روش‌های پالایشی از یک معیار برای درجه‌بندی و انتخاب ویژگی‌های کلیدی در دسته‌بند استفاده می‌کنند.برای مثال روش ضریب همبستگی پیرسون[5] و بهرهٔ سیگنال به نویز[6] [۲ و ۳] دو نمونه از روش‌های پالایشی هستند. روش‌های پوششی، فرایند انتخاب ویژگی را با بررسی ویژگی‌های انتخاب‌شده در دسته‌بندی ارزیابی می‌کنند. اما به‌تازگی روش‌های ترکیبی انتخاب ویژگی مورد استقبال بسیاری از محققان قرار گرفته است [۴]. این رویکرد ترکیبی از دو روش پوششی و پالایشی بوده و مزایای هر دو روش را دارد. در روش ترکیبی، ابتدا با روش پالایشی از مجموعه کل ویژگی‌ها، مجموعه ویژگی‌های کاندید انتخاب می‌شود و سپس این مجموعه ویژگی با رویکرد پوششی تصفیه و پالایش می‌شود [۵]. در پژوهش حاضر نیز یک روش ترکیبی پوششی-پالایشی ارائه شده است. مدل ارائه‌شده روی دادهٔ ریزآرایه بیان ژن[7] مربوط به سه نوع سرطان مورد بررسی قرار گرفته است. تکنیک‌های دسته‌بندی برای تحلیل و تفسیر ریزآرایهٔ دادهٔ بیان ژن کاربرد گسترده‌ای دارند. اطلاعات موجود در الگوی ریزآرایهٔ بیان ژن، می‌تواند علائم بسیار ارزشمندی از مشکلات بیولوژیکی، نظیر سرطان را نشان دهد [۶]. چالش اصلی در دسته‌بندی دادهٔ بیان ژن، ابعاد بالای مجموعه داده‌هاست. اغلب از یک نمونهٔ کوچک مورد مطالعه، تعداد بسیار زیادی ویژگی قابل استخراج است. ایدهٔ اصلی در این رویکرد، یافتن مرتبط‌ترین و مهم‌ترین زیرمجموعهٔ ژن برای دسته‌بندی بهتر داده‌هاست که مزایای روش‌های پالایشی یعنی سرعت محاسبات بالا و روش‌های پوششی، دقت محاسباتی بالا را توأمان به همراه دارد. این روش روی سه مجموعه داده مربوط به سرطان اعمال شده و نتایج آن با روش‌های پیشنهادشدهٔ اخیر مقایسه شده است.

در ادامه، ساختار مقاله به این صورت سازمان‌دهی شده است. در بخش دو خلاصه‌ای از پیشینهٔ تحقیق و کارهای مرتبط آورده شده است. بخش سه به مبانی نظری روش‌های انتخاب ویژگی‌ها می‌پردازد. در بخش چهار پیاده‌سازی مدل پیشنهادی ارائه شده و در بخش پنج به ارائهٔ نتایج تجربی و ارزیابی مدل پرداخته شده است. در نهایت در بخش شش نتیجه‌گیری مقاله آمده است.

۲. پیشینهٔ پژوهش

از دهه ۱۹۷۰ میلادی تاکنون، پژوهش‌های بسیاری روی روش‌های انتخاب ویژگی صورت گرفته است [۷-۱۰]. در این

1. dimension reduction based method
2. Feature extraction based method
3. Filter feature selection
4. Wrapper feature selection
5. Pearson Correlation Coefficient
6. Signal to noise ratio
7. microarray gene expression data

بخش به مرور تحقیقات انجام‌شده در زمینهٔ انتخاب ویژگی در داده‌های با ابعاد بالا پرداخته می‌شود. سگن [۱۱] در پژوهش خود، یک روش پوششی برای انتخاب ویژگی مطرح کرده استکه از یک تابع ارزیابی استفاده می‌کند. این تابع مجموع یک معیار اختلاف آماری و یک معیار پیچیدگی ویژگی را محاسبه کرده و آن را مینیمم می‌کند. این الگوریتم اولین ویژگی که بهتر بتواند دسته‌ها را از هم تمییز دهد، می‌یابد؛ سپس ویژگی‌هایی را می‌یابد که در ترکیب با ویژگی‌های نظری انتخاب‌شده، قابلیت تفکیک‌پذیری دسته‌ها را افزایش دهند. این فرایند زمانی متوقف می‌شود که به حداقل معیار بازنمایی مورد انتظار برسد. ونگ [۱۲] نیز در پژوهش خود یک روش پوششی ارائه کرده است. با توجه به اینکه روش‌های پوششی زمان اجرایی بالایی صرف می‌کنند او از یک دسته‌بند نزدیک‌ترین همسایگی[1] استفاده کرده است.

کیرا و همکاران [۱۳] در روش Relief از یک راه‌حل آماری برای انتخاب ویژگی استفاده می‌کنند. Relief یک روش مبتنی بر وزن است. لیکسین [۱۴] در تحقیق خود یک روش انتخاب ویژگی بر اساس الگوریتم ژنتیک و Relief پیشنهاد داده است. او در این روش از معیار درجه‌بندی[2] برای کاهش فضای جست‌وجو استفاده کرده است. سن سامیلاراسو و همکاران [۳] نیز یک روش پالایشی جدید برای انتخاب ویژگی معرفی کرده‌اند. نویسندگان ابتدا به‌منظور کاهش فضای جست‌وجو از روش خوشه‌بندی فازی شهودی[3] برای خوشه‌بندی ویژگی‌ها بر اساس درجهٔ عضویت و درجهٔ عدم قطعیت بین ویژگی‌ها استفاده کرده‌اند؛ سپس از یک الگوریتم ژنتیک که متفاوت از الگوریتم سنتی است، برای حذف خوشه‌های غیرمرتبط برای فرایند دسته‌بندی استفاده شده است.

دارا و بانکا [۱۵] یک روش پالایشی به نام روش بهینه‌سازی ازدحام ذرات PSO دودویی مبتنی بر تئوری زبر[4] را برای انتخاب ویژگی در داده‌های با ابعاد بزرگ ارائه کرده‌اند. آن‌ها از الگوریتم PSO دودویی برای حذف ویژگی‌های اضافی و گسسته‌سازی داده‌ها استفاده کرده‌اند سپس ورودی حاصل از این مرحله برای کاهش بعد به تئوری زبر داده می‌شود. دسته‌بندهایی همچون BLR، LibSVM، رگرسیون لجستیک، MLP، DS، j48 و KNN برای دسته‌بندی و ارزیابی نتایج استفاده شده است.

محمد و همکاران [۱۶] یک روش پالایشی به نام روش بهینه‌سازی ازدحام ذرات PSO بهبودیافته را برای انتخاب ویژگی در داده‌های ریزآرایهٔ سرطان ارائه کرده‌اند. هسو و همکاران [۴] یک روش انتخاب ویژگی ترکیبی (پالایشی-پوششی) ارائه کرده‌اند. در روش آن‌ها ابتدا ویژگی‌های مشترک شناسایی‌شده از دو روش f-Score و بهرهٔ اطلاعاتی[5] به‌عنوان ویژگی‌های کاندید انتخاب می‌گردد (پالایش) سپس مجموعه ویژگی‌های انتخابی با استفاده از روش پوششی تصفیه و پالایش می‌شود.

تاکنون مطالعات زیادی در زمینهٔ ارائهٔ روش‌های جدید انتخاب ویژگی به‌منظورمدیریت داده‌های با ابعاد بالا، نظیر داده‌های ریزآرایهٔ بیان ژن، انجام شده است. برای نمونه اینزا و همکاران [۱۷] ابتدا مقایسه‌ای بین روش‌های انتخاب ویژگی پالایشی و پوششی انجام داده‌اند؛ آن‌ها در مطالعهٔ خود معیارهای جدید پالایشی[6] را برای امتیازدهی به ویژگی‌ها پیشنهاد داده‌اند سپس با استفاده از داده‌های بیان ژن مربوط به سرطان به سنجش عملکرد روش‌های مختلف پرداخته‌اند. آپولونی و همکاران [۱۸] نیز دو الگوریتم نوین انتخاب ویژگی به روش پالایشی-پوششی به نام‌های BDE-X_{Rank} و BDE-X_{Rankf} برای کاهش بعد ریزآرایه با ابعاد بالا ارائه داده‌اند. در این مدل ابتدا ویژگی‌ها به کمک مقدار بهرهٔ اطلاعاتی امتیازدهی می‌شوند سپس در مرحلهٔ دوم یک روش پوششی بر اساس تخمین دیفرانسیل دودویی (BDE) به جست‌وجوی بهترین ویژگی‌ها می‌پردازند. به‌علاوه ویژگی کلیدی روش آن‌ها استفاده از تولید راه‌حل اولیهٔ BDE با تعداد ویژگی‌های کم است. برخی راه‌حل‌ها تنها ویژگی‌های

1. K-nearest neighbor
2. Ranking measure
3. Intuitive fuzzy clustering method
4. Rough set
5. Information Gain
6. Filter metric

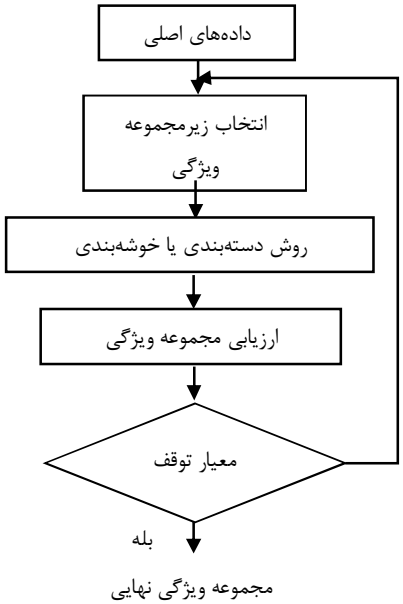
مرتبط هستند، برخی به‌منظور ایجاد تنوع در جمعیت اولیه با ویژگی‌های تصادفی آغاز می‌شوند. در BDE-X_{Rankf} یک تابع تناسب جدید که در آن مقدار امتیاز راه‌حل تحت‌تأثیر فراوانی ویژگی‌ها در جمعیت فعلی است، در الگوریتم مشارکت داده شده است. در پایان به‌منظور ارزیابی روش‌ها از چهار الگوریتم یادگیری ماشین (NB، SVM، C4.5 و KNN) استفاده شده است.

همچنین کاندناس و همکاران [۱۹] به پیاده‌سازی یک روش انتخاب ویژگی ترکیبی مبتنی بر تئوری فازی روی داده‌های با کیفیت پایین، مربوط به انواع سرطان پرداخته است. روش پیشنهادی آن‌ها شامل سه مرحله است: ۱. فرایند نرمال‌سازی و گسسته‌سازی مجموعه ویژگی‌ها، و پیش‌انتخاب ویژگی‌ها با استفاده از فرایند گسسته‌سازی (پالایش). ۲.فرایند رتبه‌بندی ویژگی‌های پیش انتخاب‌شده با استفاده از درخت تصمیم فازی. ۳. انتخاب بهترین ویژگی‌ها به روش پوششی با استفاده از جنگل تصادفی فازی گروهی[1] مبتنی بر اعتبارسنجی ضربدری[2]. نویسندگان در گام اول از روش نرمال‌سازی min-max برای نرمال‌سازی ویژگی‌ها و روش گسسته‌سازی دومرحله‌ای استفاده کرده‌اند. در روش گسسته‌سازی در مرحلهٔ اول با استفاده از درخت تصمیم فازی به تولید مجموعه ویژگی اولیه پرداخته شده و سپس در مرحلهٔ دوم با استفاده از الگوریتم ژنتیک به پالایش مجموعه ویژگی انتخابی پرداخته شده است. در این روش در گام دوم، اهمیت ویژگی‌ها بر اساس عمق آن‌ها در درخت تصمیم فازی تعیین می‌شود؛ یعنی ویژگی‌ای که در بالای درخت ظاهر شده، اهمیت بیشتری از ویژگی‌های پایین درخت دارد. به‌طور کلی نتیجهٔ این تکنیک‌ها بهبود عملکرد و دقت پیش‌بینی مدل، کارایی و مقیاس‌پذیری در فرایند کلاس‌بندی، افزایش سرعت فرایند یادگیری و قابلیت درک و تفسیر بهتر از مدل بوده است.

۳. مبانی نظری

۱.۳. روش‌های پوششی انتخاب ویژگی

عملکرد روش پوششی به این صورت است که در هر تکرار از الگوریتم، دسته‌ای از ویژگی‌ها برای یادگیری انتخاب می‌شوند و در نهایت آن دسته از ویژگی‌ها که مدل با استفاده از آن‌ها به دقت بیشتری برسد، انتخاب می‌شوند [۱۸]. شکل (۱) نحوهٔ عملکرد روش پوششی را نشان می‌دهد.



شکل (۱): نحوهٔ عملکرد روش پوششی

روش‌های پوششی قابلیت آن را دارند که عملکرد پیش‌بینی خوب و دقت دسته‌بندی بالا را به همراه صرفه‌جویی در زمان ارائه دهند اما در مقابل چند نقطه‌ضعف دارند. احتمال بروز بیش‌برازش[3] در آن‌ها بیش از روش‌های پالایشی است و به این دلیل که ارزیابی هر ترکیبی از ویژگی‌ها مستلزم انجام کامل فاز آموزش در الگوریتم یادگیری است، پیچیدگی محاسباتی بالایی دارند. نقطه‌ضعف دیگر روش‌های پوششی، فقدان عمومیت[4] آن‌هاست و استفاده از آن‌ها محدود به مدل‌های رگرسیون خاص است [۱۹].

۲.۳. روش‌های پالایشی انتخاب ویژگی

برای حل مشکلات پوششی، روش پالایشی ارائه شده است [۲۰]. در این روش، قبل از شروع فاز یادگیری، مهم‌ترین ویژگی‌های مرتبط برای یادگیری انتخاب می‌شوند و بقیهٔ

1. Fuzzy Random Forest ensemble
2. Cross Validation
3. Overfitting
4. generalization

ویژگی‌ها در نظر گرفته نمی‌شوند. این روش مستقل از الگوریتم دسته‌بند عمل می‌کند. در شکل (۲) نحوهٔ عملکرد روش‌های پالایشی به‌صورت نمادین نمایش داده شده است. در این روش ویژگی‌های توصیفی که بیشترین همبستگی با دستهٔ هدف را دارند انتخاب می‌شوند. معمول‌ترین معیار برای انتخاب ویژگی‌ها، ضریب همبستگی است [۲۱]. با توجه به سادگی محاسبات در این روش‌ها در حل مسائل با ابعاد بالا، بسیار سریع‌تر از روش‌های پوششی عمل می‌کنند.

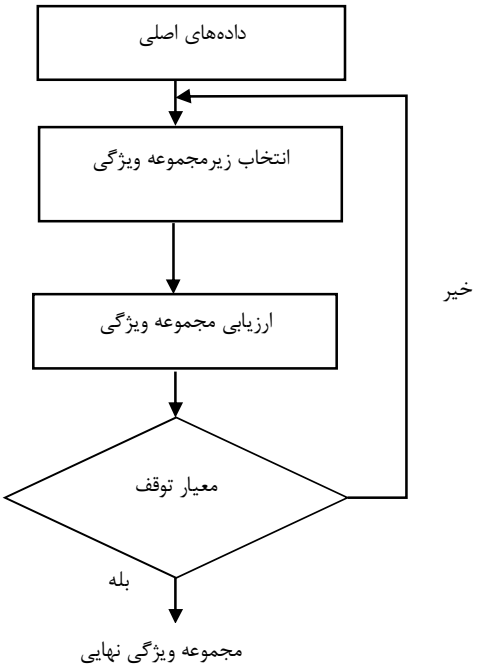
در اکثر روش‌های پالایشی، هر ویژگی به‌صورت جداگانه مورد ارزیابی قرار می‌گیرد و از وابستگی آن با سایر ویژگی‌ها صرف‌نظر می‌شود. به این روش‌ها، روش‌های پالایشی تک‌متغیره گفته می‌شود. این امر ممکن است کارایی دسته‌بند را نسبت به روش‌های پوششی کاهش دهد.

از طرف دیگر یک اشکال عمدهٔ روش‌های پالایشی نادیده گرفتن ارتباط ویژگی‌های انتخاب‌شده با دسته‌بند، طی اجرای الگوریتم است. روش‌های پارامتری مانند t-test و روش تحلیل واریانس و روش‌های ناپارامتری همچون آزمون ویلکاکسون و مجموع مربعات بین- درون دسته‌ای (BSS/WSS) از جمله پرکاربردترین روش‌های پالایشی تک‌متغیره‌اند [۲۲]. با توجه به محدودیت‌های خاص روش‌های تک‌متغیره، محققان روش‌هایی را ارائه کرده‌اند که سعی می‌کنند همبستگی بین ویژگی‌ها را نیز مورد ارزیابی قرار دهند. به این روش‌ها، روش‌های پالایشی چندمتغیره گفته می‌شود. روش‌هایی مانند انتخابویژگی مبتنی بر همبستگی (CFS) [۲۳] روش‌های کمترین افزونگی- بیشترین ارتباط (MRMR) [۲۴] و روش Relief[۲۵] روش‌های پالایشی چندمتغیره‌اند.

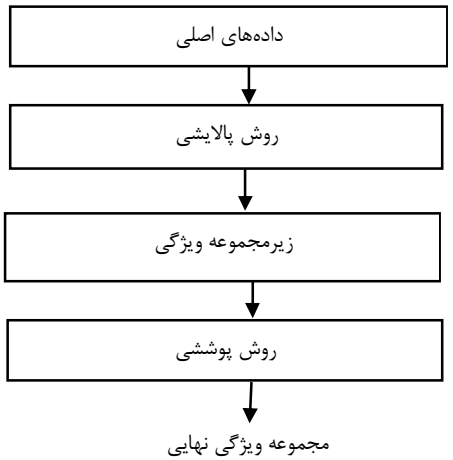
۳.۳. روش‌های ترکیبی انتخاب ویژگی

روش ترکیبی انتخاب ویژگی ترکیبی از دو روش پوششی و پالایشی است. در این روش، ابتدا به کمک روش‌های پالایشی مجموعه ویژگی کاندید انتخاب و سپس این مجموعه ویژگی با رویکرد پوششی تصفیه و پالایش می‌شود (شکل ۳).

در تحقیقات [۱۴ـ۱۵ و ۱۸] یک روش انتخاب ویژگی ترکیبی ارائه شده است.



شکل(۲): نحوهٔ عملکرد روش پالایشی



شکل (۳): روش ترکیبی (پالایشی- پوششی)

۴. مواد و روش‌ها

در این پژوهش، یک روش پوششی-پالایشی انتخاب ویژگی ارائه شده است. این روش در مرحلهٔ اول از الگوریتم‌های مختلف پالایشی انتخاب ویژگی استفاده می‌کند که عبارت‌اند

از: الگوریتم Relief، الگوریتم مبتنی بر نسبت بهره، الگوریتم مبتنی بر بهرهٔ اطلاعاتی، الگوریتم مبتنی بر ضریب عدم اطمینان متقارن و الگوریتم CFS. شمای الگوریتم در شکل (۴) نشان داده شده است.

الگوریتم FWFS شامل مراحل زیر است:

۱. یک مقدار ثابت مانند n به‌عنوان بازهٔ بررسی در نظر گرفته می‌شود. مقدار n باید از تعداد کل ویژگی‌ها کوچک‌تر باشد.

۲. ویژگی‌های موجود در مجموعهٔ داده به‌طور جداگانه و با استفاده از هریک از الگوریتم‌های پالایشی معرفی‌شده امتیازدهی می‌شوند. ویژگی‌های امتیازدهی‌شده در مجموعه $S_i, (i = 1, \dots, 5)$ به‌صورت نزولی قرار می‌گیرند (S_i مجموعه تمام ویژگی‌های مرتب‌شده به‌صورت نزولی توسط الگوریتم انتخاب ویژگی iام است).

۳. N ویژگی اول هریک از مجموعه $S_i, (i = 1, \dots, 5)$ درون زیرمجموعه $sub\_S_i, (i = 1, \dots, 5)$ قرار می‌گیرد. این n ویژگی از مجموعه $S_i, (i = 1, \dots, 5)$ حذف می‌شوند.

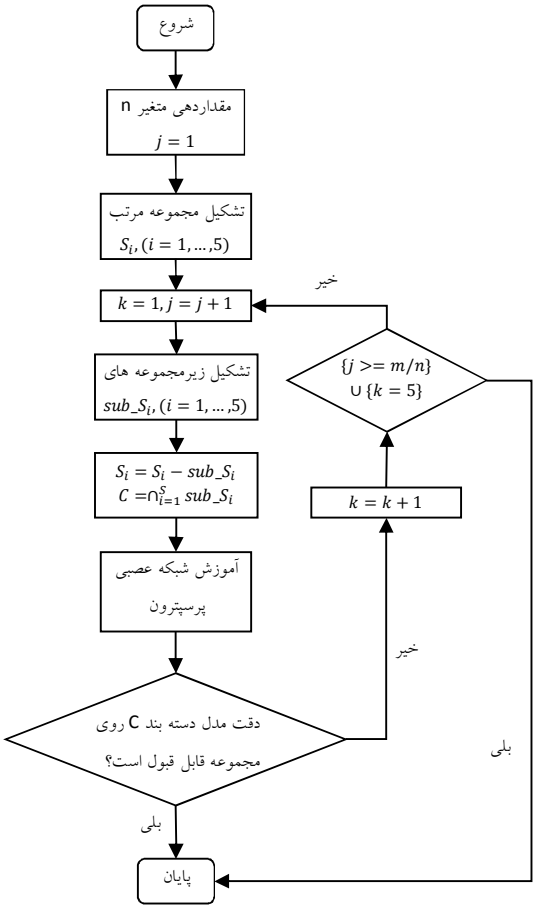
۴. ویژگی‌های مشترک $sub\_S_i, (i = 1, \dots, 5)$ انتخاب شده و در مجموعه C(common Gen) قرار می‌گیرند.

۵. شبکهٔ عصبی پرسپترون چندلایه برای دسته‌بندی داده‌ها روی ویژگی‌های مجموعه C طراحی و آموزش داده می‌شود.

۶. دقت مدل طراحی‌شده با استفاده از روش LOOCV محاسبه می‌شود.

۷. اگر دقت مدل قابل قبول باشد مجموعهٔ C به‌عنوان زیرمجموعهٔ ویژگی‌های انتخاب‌شده معرفی می‌شوند و الگوریتم به پایان می‌رسد؛ در غیر این صورت الگوریتم از گام سوم تا هفتم تکرار می‌شود. اگر تعداد کل ویژگی‌ها m باشد تعداد تکرار الگوریتم حداکثر m/n خواهد بود.

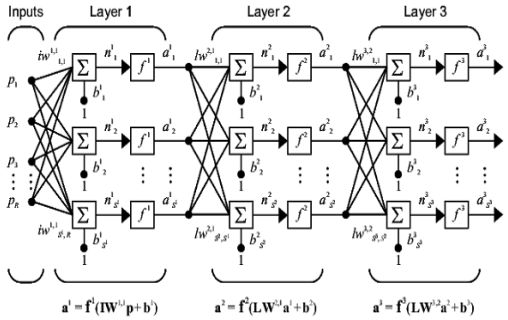
۸. یکی دیگر از شرایط خاتمه، بررسی افزایش دقت پیش‌بینی در اجرای دوبارهٔ الگوریتم است. چنانچه دقت الگوریتم در چند اجرای پیاپی افزایش چشمگیری نداشته باشد، الگوریتم متوقف می‌شود. بنابراین زیرمجموعهٔ نهایی انتخاب‌شده از ویژگی‌ها، زیرمجموعه‌ای است که با اضافه شدن ویژگی به آن، دقت پیش‌بینی افزایش اندک و ناچیزی داشته باشد. در پیاده‌سازی این الگوریتم مقدار k نشان‌دهندهٔ تعداد دفعاتی است که در آن دقت مدل افزایش قابل توجهی نداشته است. در پیاده‌سازی این الگوریتم بر ریزآرایه، مقدار ۵ برای متغیر k در نظر گرفته شده است.



شکل (۴): شمای روش پیشنهادی

در گام اول و دوم این الگوریتم، از روش‌های پالایشی انتخاب ویژگی استفاده شده است. گام‌های سه به بعد، مبتنی بر روش‌های پوششی هستند. همان طور که اشاره شد، در روش‌های پوششی یک الگوریتم دسته‌بند روی مجموعه داده با توجه به ویژگی‌های انتخاب‌شده اعمال می‌شود و این عمل تا وقتی که دقت مدل قابل قبول نباشد، ادامه می‌یابد. مرحلهٔ پنجم شامل طراحی مدل دسته‌بندی مبتنی بر شبکه‌های عصبی پرسپترون چندلایه است. شبکهٔ پرسپترون چندلایه دارای

ویژگی‌هایی است که آن‌ها را در کاربردهایی مانند شناسایی الگو، پیش‌بینی و سایر مسائل یادگیری، ممتاز می‌کند. این نوع شبکه‌ها در حل مسائل بسیار سریع و قابل اطمینان عمل می‌کنند [۲۶]. شکل (۵) یک نمونه از شبکهٔ عصبی چندلایه را نشان می‌دهد.



شکل(۵): یک نمونه شبکهٔ عصبی مصنوعی سه‌لایه با اطلاعات تفکیک‌شده

۵. یافته‌های پژوهش

۱.۵. دادهٔ ریزآرایهٔ مورد مطالعه

روش ارائه‌شده در این پژوهش بر داده‌های سه ریزآرایهٔ (داده‌های بیان ژن) مربوط به سرطان، مورد ارزیابی قرار گرفته که عبارت‌اند از: ریزآرایهٔ سرطان خون، ریزآرایهٔ لنفوما، ریزآرایهٔ DLBCL. جدول (۱) خلاصه‌ای از اطلاعات این مجموعه داده را نشان می‌دهد. این مجموعه داده از طریق منابع [۲۷ و ۲۸] قابل دسترس هستند. ویژگی‌ها در این مجموعه داده، ژن‌ها هستند که برای هریک از نمونه‌ها در این مجموعه داده مقدار بیان هر ژن آمده است. در ادامه، این سه مجموعه داده به‌طور مختصر شرح داده می‌شود. سرطان خون یکی از مهم‌ترین سرطان‌هایی است که جامعهٔ بشری با آن روبه‌روست. معمولاً نشانهٔ مشخصی برای سرطان خون وجود ندارد و زمانی که نشانه‌ها نمایان می‌شوند، بسیار مبهم و پیچیده‌اند. طراحی یک دسته‌بند می‌تواند به تشخیص بموقع این نوع سرطان خون کمک کند و روش‌های درمانی خاص را پیشنهاد دهد.

جدول (۱): خلاصه اطلاعات مجموعه داده‌های ریزآرایه

مجموعه داده ریزآرایه	تعداد ژن	تعداد نمونه‌های دسته c_1	تعداد نمونه‌های دسته c_2	تعداد کل نمونه‌ها	نسبت نمونه به بعد
سرطان خون	۷۱۲۹	۴۷	۲۵	۷۲	۰/۰۱
سرطان لنفوما	۴۰۲۶	۲۳	۲۲	۴۵	۰/۰۱۱
DLBCL	۷۰۲۹	۵۸	۱۹	۷۷	۰/۰۱

مجموعهٔ دادهٔ سرطان خون برای نخستین بار در پژوهش گلوب و همکاران تهیه شد [۲۹]. ریزآرایهٔ سرطان خون حاوی ۷۱۲۹ ژن از ۷۲ بیمار مختلف است. ۴۷ مورد از این بیماران مبتلابه سرطان خون حاد و پیشرفتهٔ لیمفوبلاستی[1] و ۲۵ مورد دیگر مبتلا به سرطان خون حاد مغز استخوان[2] هستند. سرطان لنفوما در غدد یا گره‌های سیستم لنفاوی بدن بروز می‌کند. ریزآرایهٔ مورد بررسی شامل ۴۵ نمونه از این بیماری است که تعداد ۲۲ نفر از آن‌ها به سرطان لنفوما نوع Germinal Center B-like group(GCL) و تعداد ۲۳ نفر از آن‌ها به بیماری نوع Activated Blike group(ACL) مبتلا هستند. تعداد کل ژن‌ها ۴۰۲۶ نمونه است. ریزآرایهٔ انتشار سلول‌های بتا[3] نیز یک نوع از سرطان لنفوماست. این مجموعه دارای ۵۸ نمونه DLBCL و ۱۹ نمونه غدهٔ لنفاوی کیسه‌ای[4] بوده و تعداد ژن‌های توصیف‌کنندهٔ داده‌های این ریزآرایه ۷۰۲۹ مورد است.

۲.۵. روش ارزیابی مدل

تعمیم‌پذیری یکی از مهم‌ترین ویژگی‌های یک مدل محسوب می‌شود. هرچند برآورد تعمیم‌پذیری یک مدل به‌صورت تئوری در غالب موارد بسیار پیچیده و دشوار است، برآورد تقریبی آن با روش‌های شبیه‌سازی به‌سادگی قابل دستیابی است. از پرکاربردترین روش‌های ارزیابی تعمیم‌پذیری مدل، روش اعتبارسنجی k-دسته‌ای است. در این روش یک مجموعه داده با n نمونه به k زیرمجموعه تقسیم می‌شود که

1. Acute lymphoblastic leukemia(ALL)
2. Acute myeloid leukemia(AML)
3. Diffuse large B-cell lymphoma(DLBCL)
4. follicular lymphoma(FL)

اندازهٔ هریک n/k است. در روش k-دسته‌ای، مدل k بار آموزش دیده و ارزیابی می‌شود و در هر مرحله n-k نمونه برای آموزش و k نمونه جهت آزمون در نظر گرفته می‌شود. دقت نهایی مدل میانگین حاصل از k مرحله است.

روش دیگر برای ارزیابی عملکرد مدل، روش اعتبارسنجی LOOCV است. در این روش مدل n بار اجرا می‌شود و در هر مرحله n-1 نمونه برای آموزش و یک نمونه باقی‌مانده برای آزمون در نظر گرفته می‌شود. نتیجهٔ نهایی، میانگین دقت حاصل از n بار تکرار این عملیات است.

برای ارزیابی مدل دسته‌بندی، از دو معیار دقت و صحت استفاده شده است: دقت[1]، نسبت کل نمونه‌های درست پیشگویی‌شده به کل نمونه‌هاست. صحت[2]، بیانگر نسبت نمونه‌های درست پیشگویی‌شدهٔ یک دسته به کل نمونه‌های همان دسته است. روابط (۷) و (۸) این دو معیار را به‌طور رسمی تعریف می‌کنند.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۷)$$

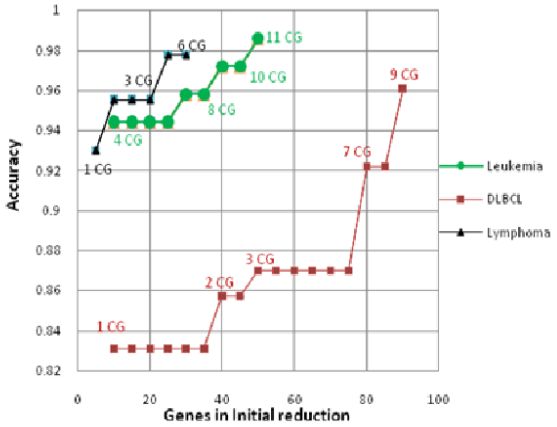
$$Precision = \frac{TP}{TP + FP} \qquad (۸)$$

که در آن TP و TN به‌ترتیب نمونه‌های متعلق به دسته positive و negative هستند که دسته‌بند آن‌ها را به‌درستی پیشگویی کرده است و FP و FN به‌ترتیب تعداد نمونه‌های متعلق به دستهٔ positive و negative می‌باشند که دسته‌بند نتوانسته است آن‌ها را به‌درستی پیشگویی کند.

۳.۵. نتایج تجربی

الگوریتم FWFS روی سه مجموعهٔ داده بیان ژن مربوط به سرطان خون، سرطان لنفوما و DLBCL پیاده‌سازی شده و مورد ارزیابی قرار گرفته است. نمودار (۱) نتایج حاصل از اعمال این الگوریتم را روی هریک از این سه مجموعهٔ داده نشان می‌دهد.

با توجه به این نمودار، دقت دسته‌بندی ریزآرایهٔ لوکمیا در مرحلهٔ اول ۹۴درصد است. در این مرحله، الگوریتم با ۴ ژن مشترک در میان ۵ ژن با بالاترین امتیاز، پیاده‌سازی شده است. دقت مدل در تکرار دوم الگوریتم، روی هشت ژن مشترک از میان ۳۵ ژن با بالاترین امتیاز به ۹۶درصد می‌رسد. این ۳۵ ژن دارای بیشترین امتیاز در میان ژن‌های رتبه‌بندی‌شده توسط هریک از الگوریتم‌های پالایشی انتخاب ژن بوده‌اند. در چهارمین تکرار الگوریتم، دقت مدل بیشتر از ۹۸درصد شده است. در این مرحله، ۱۱ ژن از میان ۵۰ ژن مشترک با بیشترین امتیاز انتخاب شده‌اند. دقت الگوریتم در تکرارهای بعدی افزایش چشمگیری نداشته است، بنابراین الگوریتم پس از چهار بار تکرار متوقف می‌شود.



نمودار (۱): دقت دسته‌بندی به روش LOOCV

الگوریتم FWFS برای مجموعه داده‌های ریزآرایهٔ لنفوما و DLBCL نیز به همین صورت پیاده‌سازی شده است. در ریزآرایهٔ لنفوما، الگوریتم در تکرار سوم متوقف شده و دقت حاصل از مدل با استفاده از شش ژن مشترک از میان ۳۰ ژن با بیشترین امتیاز، نزدیک به ۹۸درصد است. برای داده‌های DLBCL الگوریتم FWFS در تکرار پنجم با انتخاب ۹ ژن مشترک در میان ۹۵ ژن با بیشترین امتیاز متوقف شده و دقت حاصل بیشتر از ۹۶درصد است.

در جدول (۲) شناسه و تعداد ژن‌های انتخاب‌شده پس از اتمام الگوریتم، به‌تفکیک برای هریک از سه مجموعهٔ داده مشاهده می‌شود. با توجه به این جدول پس از پیاده‌سازی مدل، در ریزآرایهٔ سرطان خون از میان ۷۱۲۹ ژن، ۱۱ ژن و در ریزآرایهٔ لنفوما از میان ۴۰۲۶ ژن، ۶ ژن و در ریزآرایهٔ انتشار

1. Accuracy
2. Precision

سلول‌های بتا از میان ۷۰۲۹ ژن، تعداد ۹ ژن انتخاب شده است.

جدول (۲): اطلاعات ژن‌های انتخاب‌شده

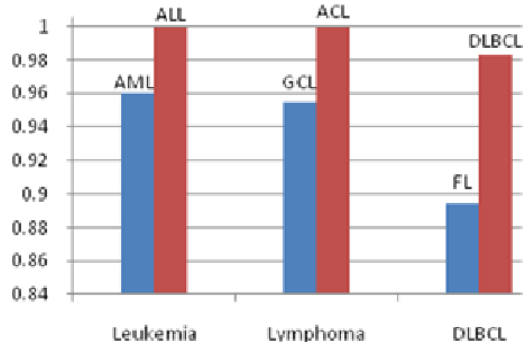
مجموعه داده ریزآرایه	تعداد اولیه ژن‌ها	شناسه ژن‌های انتخاب‌شده توسط الگوریتم FWFS	تعداد ژن‌های انتخاب‌شده
سرطان خون	۷۱۲۹	M84525_at, M23197_at,X95735_at,M27891_at ,U46499-at,L09209_s_at, M31523_at,M92287_at,61587_at, X59417_at,M55150_at	۱۱
سرطان لنفوما	۴۰۲۶	GENE3332X,GENE3330X,GENE 3261X,GENE3335X,GENE3314X, GENE3256X	۶
DLBCL	۷۰۲۹	Z11793_at, L17131_rnal, X02152_at, M63138_at,HG2279-HT2375_at, Z21966_at, HG1980-HT2023_at, X67951_at,D13633_at	۹

در گام پنجم از الگوریتم، از شبکهٔ عصبی پرسپترون چندلایه برای دسته‌بندی داده‌ها استفاده می‌شود. تعداد نورون‌های لایهٔ ورودی برابر با تعداد زیرمجموعهٔ ژن‌های انتخاب شده است. داده‌های بیان ژن بررسی‌شده در این پژوهش دارای دو دستهٔ مختلف‌اند؛ بنابراین لایهٔ خروجی شبکه دارای یک نورون است.

شبکه‌های عصبی به تعداد نورون‌های لایهٔ مخفی خود بسیار حساس‌اند. تعداد نورون‌های کم باعث عدم تطابق و تعداد نورون زیاد باعث بیش برازش می‌شود. لذا انتخاب مناسب تعداد نورون‌ها در هر لایهٔ پنهان بسیار مهم است؛ با وجود این، راه‌حلی عملی برای دست‌یابی به تعداد نورون‌ها و تعداد لایه‌های پنهان برای تعیین معماری شبکه وجود ندارد. رویکرد متعارف برای رسیدن به تعداد مناسب لایه‌ها و نورون‌ها، مورد آزمایش قرار دادن مقادیر مختلف برای تعداد نورون‌ها و تعداد لایه‌های پنهان است.

در لایهٔ پنهان شبکهٔ عصبی طراحی‌شده، از توابع فعال‌سازی سیگموئیدی و تانژانت هذلولی استفاده شده است. به‌هنگام‌سازی مقادیر وزن و بایاس در هر تکرار از روال پس‌انتشار خطا بر اساس روابطی خاص صورت می‌گیرد که به‌اصطلاح، الگوریتم آموزش نامیده می‌شود. الگوریتم‌های آموزش مختلفی برای شبکه‌های پس‌انتشار وجود دارد. در روش FWFS الگوریتم‌های آموزشی دسته‌ای گرادیان نزولی[1]، لونبرگ مارکوارت[2] و الگوریتم پس‌انتشار ارتجاعی[3] بررسی شده‌اند.

دقت دسته‌بند شبکهٔ عصبی پرسپترون چندلایه به روش LOOCV در جدول (۳) نشان داده شده است. این نتایج مربوط به اجرای الگوریتم در آخرین تکرار است. در طراحی شبکهٔ عصبی، الگوریتم‌های آموزش مختلف ارزیابی شده‌اند که نتایج حاصل از آن‌ها در این جدول مشاهده می‌شود. با توجه به جدول (۳) شبکهٔ عصبی با الگوریتم آموزش RB دارای بالاترین دقت است که به‌صورت پررنگ در جدول مشخص است. باتوجه به اینکه الگوریتم آموزش RB بیشترین دقت را کسب کرده است، اندازهٔ صحت این شبکهٔ عصبی بر سه مجموعهٔ ریزآرایه مورد بررسی محاسبه شده است. نمودار (۲) میزان صحت مدل را نشان می‌دهد. شایان ذکر است روش FWFS با استفاده از نرم‌افزار داده‌کاوی WEKA و جعبه ابزار شبکهٔ عصبی نرم‌افزار MATLAB7.10.0 پیاده‌سازی شده است.



نمودار (۲): صحت دسته‌بند MLP NN با استفاده از الگوریتم آموزش RP

۵،۴. مقایسه با سایر روش‌ها

نتایج تجربی روی سه مجموعهٔ داده نشان می‌دهد که روش پیشنهادی در برخی موارد با ویژگی‌های کمتر به دقتی برابر یا بالاتر نسبت به سایر روش‌هایی که به‌تازگی پیشنهاد شده‌اند، دست یافته است. برای مثال آپولونی در [۱۸] با روش ترکیبی

1. Batch gradient descent
2. Levenbery-Marqwardt(LM)
3. Resilient Backpropagation(RB)

پوششی-پالایشی با دسته‌بند SVM با انتخاب ویژگی‌های کمتر نسبت به روش پیشنهادی ما به دقت کمتری (۹۴/۱درصد) در داده‌های سرطان خون دست یافته است. همچنین هسو [۴] ابتدا با انتخاب ویژگی‌های مشترک شناسایی‌شده با استفاده از دو روش F-Score و بهرهٔ اطلاعاتی و سپس پوشش مجموعه ویژگی‌های انتخابی با استفاده از مدل دسته‌بند SVM به دقت ۹۹/۶۱ (برابر با دقت روش پیشنهادی مقاله) دست یافته است. این در حالی است که تعداد مشخصه‌های انتخابی (۷۰ مشخصه) نسبت به روش پیشنهادی این مقاله بسیار زیاد است.

جدول (۳): دقت دسته‌بند شبکهٔ عصبی پرسپترون چندلایه به روش LOOCV

الگوریتم آموزش	سرطان خون		سرطان لمفوما			DLBCL		
	ALL	AML	Overall	ACL	GCL	Overall	DLBCL	FL	Overall
LM[1]	۹۷/۸۷	۹۲	۹۵/۸۳	۱۰۰	۸۶/۳۶	۹۳/۳۳	۹۱/۳۸	۸۹/۴۷	۹۰/۹۰
LM[2]	۹۷/۸۷	۹۲	۹۵/۸۳	۱۰۰	۸۶/۳۶	۹۳/۳۳	۹۱/۳۸	۸۹/۴۷	۹۰/۹۰
LM[3]	۹۷/۸۷	۹۲	۹۵/۸۳	۱۰۰	۸۶/۳۶	۹۳/۳۳	۹۱/۳۸	۸۹/۴۷	۹۰/۹۰
LM[4]	۹۷/۸۷	۹۲	۹۵/۸۳	۱۰۰	۸۶/۳۶	۹۳/۳۳	۹۱/۳۸	۸۹/۴۷	۹۰/۹۰

در مواردی نیز نتایج تجربی اعمال روش پیشنهادی روی برخی مجموعه داده‌ها نسبت به سایر روش‌ها عملکرد کمتری داشته است. برای مثال آپولونی [۱۸] با انتخاب دو مشخصه به دقت ۱۰۰درصد در داده‌های لنفوما دست یافته است، درحالی‌که دقت مدل ما نزدیک به ۹۸درصد است. اما همان طور که اشاره شد، روش پیشنهادی بر مجموعه دادهٔ سرطان خون، دقت بیشتری نسبت به روش آپولونی داشته است.

جدول (۴): مقایسهٔ روش پیشنهادی مقاله با روش‌های ترکیبی موجود

داده	ویژگی کاهش‌یافته	تعداد	دقت	روش
سرطان خون	۷۰		۹۸/۶۱	هسو و همکاران [۴]
	۱		۹۴/۱	آپولونی و همکاران [۱۸]
	۱۱		**۹۸/۶۱**	روش پیشنهادی مقاله
سرطان لمفوما	**۲**		**۱۰۰**	آپولونی و همکاران [۱۸]
	۶		۹۷/۷۷	روش پیشنهادی مقاله

جدول (۵): مقایسهٔ روش پیشنهادی مقاله با روش‌های پوششی موجود

داده	الگوریتم مختلف	ویژگی در گام‌های	تعداد میانگین	دقت	روش
سرطان خون		۳/۱		۹۰/۷	SFS
		۷/۶		۹۷/۱	$IWSS^2$
		۷/۱		**۹۵/۵**	$IWSS^3$
		۴/۴		۹۶/۳	$IWSS_r^2$
		۳/۸		۹۸/۳	$IWSS_r^3$
DLBCL		۳/۵		۸۳/۳	SFS
		۹/۹		۹۳/۶	$IWSS^2$
		۵/۳		۹۲/۵	$IWSS^3$
		۵/۳		۸۹/۵	$IWSS_r^2$
		۳/۸		۹۸/۳	$IWSS_r^3$

جدول (۶): مقایسهٔ روش پیشنهادی مقاله با روش‌های ترکیبی موجود

داده	ویژگی کاهش‌یافته	تعداد	دقت	روش
سرطان خون	۷		۱۰۰	دارا و بانکا [۱۵]
	۳/۵		۱۰۰	محمد و همکاران۱۶]
	۱۱		۹۸/۶۱	روش پیشنهادی
سرطان لمفوما	**۳**		**۹۷/۹۳**	دارا و بانکا [۱۵]
	۶		۹۷/۷۷	روش پیشنهادی

در ادامه، روش پیشنهادی در این مقاله با سایر روش‌های پالایشی و پوششی و ترکیبی، مقایسه شده است. جدول (۴) نتایج حاصل از مقایسه روش ترکیبی پیشنهادی با سایر روش‌های ترکیبی را نشان می‌دهد. جدول (۵) مربوط به نتایج به‌دست‌آمده از مدل‌های پوششی روی مجموعه داده مشابه با این پژوهش را

1. Levenbery-Marqwardt(LM)
2. Levenbery-Marqwardt(LM)
3. Levenbery-Marqwardt(LM)
4. Levenbery-Marqwardt(LM)

نشان می‌دهد. ایگووانگ و همکاران [۱۲] روش‌های پوششی مختلف را مطالعه کرده‌اند. آن‌ها در بهترین حالت با میانگین ۳/۸ ویژگی به دقت ۹۸/۳درصد در داده‌های سرطان خون دست یافته‌اند. این در حالی است که روش پیشنهادی این مقاله به دقت ۹۸/۶۱ با ۱۱ ویژگی، دست پیدا کرده است. دقت حاصل از پیاده‌سازی مدل ایگووانگو همکاران روی مجموعه داده‌های سرطان لنفوما به ۹۸/۳ می‌رسد که در مقایسه با دقت مدل پیشنهادی، عملکرد بهتری داشته است.

در جدول (۶) نیز دقت حاصل از روش انتخاب مشخصهٔ پالایشی، که روی مجموعه دادهٔ سرطان خون و لنفوما اعمال شده، نشان داده شده است. با توجه به بررسی‌های انجام‌شده، برخی روش‌های ارائه‌شده توسط سایر محققان نتایجی مشابه و یا بهتر از روش ارائه‌شده در این پژوهش را داشته‌اند، اما متأسفانه در اکثر پژوهش‌ها به جزئیات مدل و به‌خصوص تعداد تکرار الگوریتم در روش‌های پوششی اشاره‌ای نشده است و در برخی پژوهش‌ها زمان اجرای الگوریتم به‌عنوان معیار ارزیابی، ارائه شده است. تنها معیار مشترک برای مقایسهٔ روش‌های ارائه‌شده در پژوهش‌ها، دقت حاصل از مدل است. به‌طور کلی می‌توان گفت رویکرد ترکیبی حاصل ترکیب دو روش پوششی و پالایشی بوده و مزایای هر دو روش را دارد.

۶. نتیجه‌گیری

انتخاب ویژگی یکی از موضوعات مهم در زمینهٔ یادگیری ماشین و به‌خصوص دسته‌بندی است. یک الگوریتم انتخاب ویژگی مناسب، به کشف بهتر دانش و تفسیر آن کمک می‌کند. در این مقاله یک روش انتخاب ویژگی نوین، با ترکیب روش‌های پالایشی و پوششی معرفی شد. روش پیشنهادی مزایای هر دو رویکرد پالایشی و پوششی را به‌طور همزمان به همراه دارد. همان طور که پیش‌تر اشاره شد، نتایج روش‌های پوششی بسیار دقیق‌تر از روش‌های پالایشی است؛ اما به‌علت بالا بودن پیچیدگی محاسباتی، این روش‌ها در برخورد با داده‌هایی با ابعاد بالا، پژوهشگران علاقه‌مند به استفاده از روش‌های پالایشی هستند. یکی از معایب روش‌های پالایشی نیز نادیده گرفتن نتایج حاصل از دسته‌بند در انتخاب ویژگی‌هاست. روش ترکیبی پیشنهادی در این مقاله از مزیت رویکردهای مبتنی بر روش‌های پالایشی بهره‌مند است و همچنین نتایج مدل دسته‌بند در فرایند انتخاب ویژگی دخیل است.

در ادامه به‌طور خلاصه به بیان نتایج حاصل از اعمال این روش اشاره می‌شود. دقت مدل دسته‌بند بر مجموعهٔ دادهٔ ریزآرایهٔ لوکمیا با انتخاب ۱۱ ژن، بیشتر از ۹۸درصد بوده است. در ریزآرایهٔ لنفوما، الگوریتم در تکرار سوم متوقف شده و دقت حاصل از مدل با استفاده از شش ژن مشترک از میان ۳۰ ژن با بیشترین امتیاز، نزدیک به ۹۸درصد است. برای مجموعه داده‌های DLBCL الگوریتم FWFS در تکرار پنجم با انتخاب ۹ ژن مشترک در میان ۹۵ ژن با بیشترین امتیاز متوقف شده و دقت حاصل بیشتر از ۹۶درصد است. همچنین مقایسهٔ روش پیشنهادی مقاله با سایر روش‌های موجود حاکی از آن است که روش پیشنهادی در داده‌های سرطان خون به دقت بهتری نسبت به سایر روش‌ها دست یافته است. روش پیشنهادی این پژوهش، برای هر مسئلهٔ دسته‌بند با ابعاد بالای مجموعهٔ داده نظیر داده‌های بیان ژن، مناسب است. یکی از محدودیت‌های روش ارائه‌شده تعیین بازهٔ بررسی در گام اول است که باید به روش سعی و خطا محاسبه شود. همچنین یکی دیگر از محدودیت‌ها انتخاب بهینهٔ الگوریتم‌های پالایشی از میان الگوریتم‌های موجود است. در پژوهش‌های آتی، می‌توان با جایگزینی دیگر الگوریتم‌های پالایشی و پوششی انتخاب ویژگی در الگوریتم ارائه‌شده نتایج به‌دست‌آمده را ارزیابی کرد. همچنین با تخمین مقدار بازهٔ بررسی ویژگی‌ها، سرعت اجرای الگوریتم را بهبود بخشید.

مراجع

[1] Fodor, I.K., "*A survey of dimension reduction techniques*", Technical Report, Lawrence Livermore National Laboratory, 2002.

[2] Mishra, D. and Sahu, B., "*Feature selection for cancer classification: a signal-to-noise ratio approach*", International Journal of Scientific and Engineering Research, Vol. 2, No. 4, pp.1-7, 2011.

[3] Senthamilarasu, S. and Hemalatha, M., "*A genetic algorithm based intuitionistic fuzzification technique for attribute selection*/", Indian J. Sci. Technol. Vol. 6, No. 4, pp. 4336–4346, 2013.

[4] Hsu, H.H., Hsieh, C.W. and Lu, M.D. "*Hybrid feature selection by combining filters and wrappers*", Expert Systems with Applications, Vol. 38, No. 7, pp. 8144–

8150, 2011.

[5] Rezaeenoor J., Yari Eili M., Hadavandi, E. and Roozbahani, M.H., "*Developing a new hybrid intelligence approach for prediction online news popularity*", International Journal of Information Science and Management, Vol. 16, No. 1, pp.71-87, 2018.

[6] Rakkeitwinai, S., Lursinsap, C., Aporntewan, Ch. and Mutirangura, A. "*New feature selection for gene expression classification based on degree of class overlap in principle dimensions*", Computers in Biology and Medicine, Vol. 64, pp.1-7, 2015

[۷] نبی‌لو، مریم، دانشپور، نگین، «**ارائهٔ یک الگوریتم خوشه‌بندی برای داده‌های دسته‌ای با ترکیب معیارها**»، مجلهٔ محاسبات نرم، دورهٔ پنجم، شمارهٔ ۱، ۱۴ـ۲۵، ۱۳۹۵.

[۸] وثیقی ذاکر، اکرم، جلیلی، سعید، «**پیش‌بینی ژن‌های بیماری با استفاده از دسته‌بند تک‌کلاسی ماشین بردار پشتیبان**»، مجلهٔ محاسبات نرم، دورهٔ چهارم، شمارهٔ ۱، ۷۴ـ۸۳، ۱۳۹۴.

[9] Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor, B., "*Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks*", Bioinformatics, Vol. 22, No. 14, pp.184–190, 2006.

[10] Nanni, L. and Lumini, A., "*Wavelet selection for disease classification by DNA microarray data*", Expert Systems with Applications, Vol. 38, pp. 990–995, 2011.

[11] Segen, J., "*Feature selection and constructive inference*", In: Proceedings of Seventh International Conference on Pattern Recognition, pp.1344–1346, 1984.

[12] Wang, A., Ning, A., Chen, G. and Li, L., Alterovitz, G., "*Accelerating wrapper-based feature selection with K-nearest-neighbor*", Knowledge-Based Systems, Vol. 83, pp.81-91, 2015.

[13] Kira, K. and Rendell, L.A., "*The feature selection problem: Traditional methods and a new algorithm*", Proceedings of Ninth National Conference on Artificial Intelligence, pp.129–134, 1992.

[14] Zhang, L. X., Wang, J. X., Zhao, Y. N. and Yang, Z. H., "*A novel hybrid feature selection algorithm: using relief estimation for Ga-wrapper search*", international Conference on Machine Learning and Cybernetics, Vol. 1, pp. 380–384,2003.

[15] Dara, S., Banka, H. and Annavarapu, C. S. R. "*A Rough Based Hybrid Binary PSO Algorithm for Flat Feature Selection and Classification in Gene Expression Data*", Annals of Data Science, Vol. 4, No. 3, pp. 341-360, 2017.

[16] Mohamad, M. S., Omatu, S., Deris, S. and Yoshioka, M.,"*Particle swarm optimization for gene selection in classifying cancer classes*", Artificial Life and Robotics, Vol. 14, No. 1, pp.16-19, 2009.

[17] Inza, I.A., Larran aga, P., Blanco, R. and Cerrolaza, A.J., "*Filter versus wrapper gene selection approaches in DNA microarray domains*", Artificial Intelligence in Medicine, Vol. 31, pp. 91-103, 2004.

[18] Apollonia, J., Leguizam´ona, G. and Alba, E., "*Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiment*", Applied Soft Computing, Vol. 38, pp. 922-932,2016.

[19] Cadenas, J.M., Garrido, M.C. and Martínez, R., "Feature subset selection Filter–Wrapper based on low quality data", Expert Systems with Applications, Vol. 40, PP. 6241–6252, 2013.

[18] Frenay, B., Doquire, G. and verleysen M, "*Is mutual information adequate for feature selection in regression?*", Neural Networks, Vol. 48, pp. 1-7, 2013.

[19] Blum, A. L., and Langley, P., "*Selection of relevant features and examples in machine learning.Artificial Intelligence*", Vol. 97, No.1, pp. 245–271, 1997.

[20] Guyon, I. and Elisseeff, A., "*An introduction to variable and feature selection*", *The Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, 2003.

[21] Doquire, G. and Verleysen, M., "*Feature selection with missing data using mutual information estimators*", Neurocomputing, Vol. 90, pp. 3-11, 2012.

[22]Jafari, P. and Azuaje, F., "*An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors*", BMC Medical Informatics and Decision Making, Vol. 6, No.1, 2006.

[23] Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P., "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments", FEBS Letter, Vol. 573, No. 1-3, pp. 83–92, 2004.

[24] Peng, H., Long, F. and Ding, Ch., "*Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, 2005.

[25] Kononenko, I., "Estimating attributes: analysis and extensions of RELIEF", European Conference on Machine Learning, pp.171–182, 1994.

[26] Rezaeenoor, J., Yari Eili, M., Roozbahani, Z. and Ebrahimi, M., "*prediction of protein thermostability by an efficient neural network approach*", Journal of health management and informatics, Vol. 3, No. 4, pp. 102-110, 2016.

[27] datam.i2r.a-star.edu.sg/datasets/krbd.

[28] www.upo.es/eps/aguilar/datasets.html

[29] Golub, TR., Slonim, DK., Tamayo, P., Huard, C., Gaasenbeek, M. and et al, Molecular classification of cancer Class discovery and class prediction by gene expression monitoring, Bloom eld CD, Lander ES, pp. 531–537, 1999.