مدل‌سازی و طراحی الگوریتم تشخیص و غربالگری بیماری مبتنی بر دسته‌بندی ماشین بردار پشتیبان خطی با قابلیت انتخاب ویژگی ترکیبی

تارا غفوری، نگین معنوی‌زاده*

گروه مهندسی برق و الکترونیک، آزمایشگاه ادوات نانوساختار الکترونیکی، دانشکده مهندسی برق، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران.

دریافت: ۱۴۰۱/۰۴/۳۰ ویرایش: ۱۴۰۱/۰۵/۰۶ پذیرش: ۱۴۰۱/۰۶/۲۵ آنلاین: ۱۴۰۱/۰۷/۰۱

چکیده

زمینه و هدف: در مطالعه حاضر، یک رویکرد انتخاب ویژگی ترکیبی از روش‌های فیلتر و بسته‌بندی، با هدف تشخیص وضعیت بیماری و بقای بیمار، برای تعدادی از مجموعه دادگان علوم زیستی با تعداد متفاوت نمونه، ویژگی و کلاس پیاده‌سازی می‌شود؛ بنابراین، این راهبرد از مزایای هر دو روش، شامل سرعت عملکرد، تعمیم‌پذیری و دقت بالا بهره می‌برد.

روش بررسی: الگوریتم‌های انتخاب ویژگی در چارچوب بازشناسی آماری الگو در نرم‌افزار Matlab R2021a طی فروردین و اردیبهشت ۱۴۰۱ مدل‌سازی شده‌اند. ابتدا ویژگی‌ها بر پایه اطلاعات متقابل بهنجار شده رتبه‌بندی می‌شوند و یک زیرمجموعه ویژگی بهینه با بالاترین دقت دسته‌بند انتخاب می‌شود. پس از خوشه‌بندی مجموعه داده به‌روش Mini Batch K-means و استخراج ویژگی‌های رتبه‌بندی‌شده، الگوریتم‌های شمول و خروج ویژگی به مجموعه دادگان اعمال می‌شوند.

یافته‌ها: رویکردهای انتخاب ویژگی پیشنهادی برای مجموعه دادگان زیست‌شناسی مولکولی، ویروس هپاتیت C و باکتری E.coli، امتیاز صحت و فراخوانی بالای ۹۸٪ را نتیجه می‌دهند، که به معنای حضور تعداد بسیار کم موارد مثبت کاذب و منفی کاذب در دسته‌بندی با ماشین بردار پشتیبان خطی است. برای مجموعه داده ویروس هپاتیت C، با انتخاب ۹ ویژگی مرتبط از ۱۳ ویژگی موجود با روش خروج ویژگی، دقت دسته‌بندی ۹۸/۹۲٪ و امتیاز F1 ۹۹/۰۲٪ به‌دست می‌آید. رویکرد شمول ویژگی نیز با یک اختلاف جزیی، دقت ۹۸/۷۸٪ را نتیجه می‌دهد.

نتیجه‌گیری: نتایج حاصل نشان‌دهنده توانمندی رویکردهای انتخاب ویژگی به‌کار رفته برای مجموعه دادگان علوم زیستی با ابعاد بالای ویژگی همچون مجموعه داده بیان پروتیین می‌باشد. قابلیت تعمیم‌پذیری به سایر دسته‌بندها و تعیین خودکار تعداد ویژگی‌های بهینه در طول فرآیند انتخاب ویژگی، این رویکردها را در بسیاری از کاربردهای داده‌کاوی برای علوم زیستی انعطاف‌پذیر می‌سازد.

کلمات کلیدی: انتخاب ویژگی ترکیبی، پایگاه داده زیستی، دسته‌بند ماشین بردار پشتیبان خطی، خوشه‌بندی k-میانگین مینی‌بَچ، اطلاعات متقابل بهنجارشده.

* نویسنده مسئول: تهران، دانشگاه صنعتی خواجه نصیرالدین طوسی، دانشکده مهندسی برق، گروه مهندسی برق و الکترونیک.

تلفن: ۰۲۱-۸۴۰۶۲۲۰۷

E-mail: manavizadeh@kntu.ac.ir

مقدمه

یکی از مهم‌ترین چالش‌ها در بازشناسی آماری الگو، مسئله نفرین ابعاد (Curse of dimensionality) است که متداول‌ترین راهکار پیشنهاد شده تاکنون برای رفع آن، کاهش ابعاد با دو رویکرد تبدیل ویژگی و انتخاب ویژگی است. اهداف اصلی کاهش ابعاد با حفظ کمترین تعداد ویژگی، بهبود دقت دسته‌بندی، کاهش زمان محاسبات و جلوگیری از بیش‌برازش می‌باشد.[1] تبدیل‌های ویژگی براساس دو دیدگاه عمل می‌کنند: در رویکرد مبتنی بر داده‌ها (فیلتر)، هدف از اعمال تبدیل و کاهش بُعد، ارایه توصیف متفاوتی از داده‌ها در فضای جدید است. تحلیل مولفه‌های اصلی (PCA)، شناخته‌شده‌ترین روش تبدیل ویژگی با معیار بیان داده است. تابع هدف در این روش، کاهش

وابستگی آماری بین ویژگی‌ها یا افزایش پراکندگی داده‌ها در فضای جدید است، در حالی‌که، رویکرد وابسته به کلاس (بسته‌بندی)، داده‌ها را با هدف کاهش خطای طبقه‌بندی یا افزایش فاصله بین نماینده کلاس‌ها توام با کاهش پراکندگی داده‌های درون کلاسی در فضای جدید ویژگی با ابعاد پایین‌تر نگاشت می‌کند.[۲] به‌طور معمول، تحلیل تشخیصی خطی (LDA) به‌عنوان تابع نگاشت برای چنین تبدیلاتی به‌کار می‌رود.

درصورتی‌که روند استخراج ویژگی پرهزینه و زمان‌بر باشد یا ویژگی‌ها ماهیت عددی نداشته و نیاز به تصمیم‌گیری بر مبنای داده‌ها در فضای ویژگی اصلی باشد، انتخاب زیرمجموعه ویژگی به تبدیل ویژگی برتری می‌یابد. در فضای جستجو، زیرمجموعه ویژگی‌هایی انتخاب می‌شوند که تابع هدف را بهینه سازند.[۳]

به‌طور کلی، توابع هدف مبتنی بر روش‌های فیلتر (استقلال ویژگی‌ها، معیارهای تئوری اطلاعاتی و غیره) و یا روش‌های بسته‌بندی (دقت طبقه‌بندی) و یا ترکیبی از این دو رویکرد تعیین شده و پس از اعتبارسنجی، زیرمجموعه ویژگی نهایی انتخاب می‌شود.

از جمله معیارهای فیلتر مرسوم، بیشینه‌سازی فاصله بین میانگین کلاس‌ها، بیشینه‌سازی وابستگی بین ویژگی و برچسب کلاس توام با بیشینه‌سازی استقلال ویژگی‌ها از یکدیگر (مقادیر ویژه LDA) می‌باشد. تاکنون، معیارهای گوناگونی برای بیان تابع هدف با رویکرد همبستگی ویژگی‌ها معرفی شده‌است.[۴و۵و۶] معیار دیگر برای تعیین تابع هدف به‌منظور انتخاب زیرمجموعه ویژگی، سنجش اطلاعات متقابل میان مجموعه ویژگی انتخابی و برچسب کلاس است.[۷]

روش‌های فیلتر در مقایسه با روش‌های بسته‌بندی از اجرای سریع‌تر و تعمیم‌پذیری بالاتر برخوردار بوده، با این وجود، در صورت جمع‌آوری مجموعه دادگان مناسب، این رویکرد به انتخاب کل مجموعه ویژگی و یا حذف تعداد اندکی از آن در بهینه‌سازی تابع هدف گرایش دارد. همچنین رویکرد مستقل از کلاس اطلاعاتی در مورد ارتباط بین ویژگی و کلاس هدف به‌دست نمی‌دهد. در مقابل، دقت روش بسته‌بندی بالاتر بوده و به‌دلیل شمول اعتبارسنجی، قابلیت تعمیم‌پذیری نیز دارد، با این وجود، این قابلیت برای طبقه‌بند به‌کار رفته اختصاصیت داشته و بنابراین در مقایسه با روش فیلتر از نقطه قوت آن کاسته می‌شود. همچنین، از آن‌جایی که مرحله آموزش در انتخاب ویژگی دخالت دارد، فرآیند اجرای آن کُند است.[۸و۹]

روش‌های متداول جستجو نیز شامل الگوریتم‌های نمایی، ترتیبی و تصادفی می‌باشد.[۱۰و۱۱و۱۲] یک رویکرد هیبریدی فیلتر-بسته‌بندی با در نظر گرفتن به‌ترتیب، اطلاعات متقابل و حذف بازگشتی ویژگی (RFE) ارایه شده و کارآیی دسته‌بند جنگل تصادفی مورد استفاده با شاخص‌های ارزیابی دقت، صحّت، فراخوانی و امتیاز F1 به‌دست آمده است.[۱۳] در یک بررسی مشابه، به‌منظور جست‌وجوی ویژگی‌هایی که اطلاعات متقابل مجموعه ویژگی پیشنهادی را با برچسب کلاس افزایش و با سایر ویژگی‌ها کاهش می‌دهند، از الگوریتم ژنتیک استفاده شده است.[۱۴] مقدار اطلاعات متقابل دوتایی (PMI) ویژگی‌ها بین صفر و یک بهنجار شده تا از بایاس شدن به ویژگی‌های چند ارزشی جلوگیری شود.[۱۵] همچنین، یک رویکرد انتخاب ویژگی مبتنی بر کاوش زیرگراف چگال به کمک یک الگوریتم تخمین و معیار اطلاعات متقابل بهنجار شده میان زوج ویژگی‌ها با روش جست‌وجوی شناور رو به جلو-رو به عقب معرفی شده است.[۱۶]

کاربرد انتخاب ویژگی در پردازش کلان‌داده‌ها و به‌ویژه مجموعه دادگان ژنومی بوده که در بازشناسی آماری الگو، بخش عمده این پردازش را یادگیری با نظارت و بدون نظارت داده‌ها تشکیل می‌دهد.[۱۷]

سه گام اصلی در روش‌های یادگیری بدون ناظر غیرپارامتری موسوم به خوشه‌بندی، تعریف یک معیار شباهت یا عدم شباهت بین نمونه‌ها (تابع فاصله)، بیان یک تابع هزینه برای خوشه‌بندی و ارایه یک الگوریتم تکراری برای بهینه‌سازی تابع هزینه با استفاده از تابع فاصله است.[۱۸] معیارهای اعتبارسنجی خوشه نیز غالبا بر پایه فاصله داده‌های درون هر خوشه از میانگین آن و همزمان فاصله خوشه‌ها از یکدیگر استوار است.[۱۹و۲۰و۲۱]

انتخاب نقطه اولیه و به‌دنبال آن حرکت از خوشه‌ای به خوشه دیگر در جهت بهینه‌سازی تابع هزینه در همگرایی مسئله حایز اهمیت است. از پرکاربردترین روش‌های خوشه‌بندی الگوریتم‌های مسطح و سلسله مراتبی بوده که در رویکرد اول، خوشه‌ها به‌طور کامل از یکدیگر مستقل هستند (خوشه‌بندی k-میانگین و تعمیم‌یافته آن با قابلیت خودسازماندهی داده‌ها و انتخاب خودکار تعداد خوشه‌ها (ISODATA) و در رویکرد دوم، خوشه‌ها در یک ساختار شبه درختی در هر تکرار، به خوشه‌های کوچکتر شکسته شده و یا از

ادغام آن‌ها، خوشه بزرگتر تشکیل می‌شوند (خوشه‌بندی تقسیمی و تجمیعی).

خوشه‌بندی سلسله مراتبی اطلاعات دقیق‌تری تولید کرده و برای تحلیل داده‌های جزیی پیشنهاد می‌شود، با این وجود، به‌دلیل پیچیدگی محاسباتی بالا، برای مجموعه دادگان با اندازه بزرگ مناسب نبوده و روش‌های مسطح کاربرد بیشتری دارند.[۲۲]

در این کار، برای نخستین بار، یک روش انتخاب ویژگی ترکیبی فیلتر–بسته‌بندی شامل اطلاعات متقابل بهنجارشده normalized mutual information (NMI) و کارآیی دسته‌بند ماشین بردار پشتیبان خطی Linear support vector machine (LSVM) پیشنهاد شده و برای مجموعه دادگان علوم زیستی از پایگاه داده یادگیری ماشین UC Irvine Machine Learning Repository (UCI) با ابعاد بالای ویژگی با رویکرد بیماریابی پیاده‌سازی می‌شود.[۲۳] بر این اساس، ویژگی‌های هر مجموعه داده رتبه‌بندی شده و بهترین زیرمجموعه ویژگی انتخاب می‌شود. الگوریتم k-میانگین مینی‌بَچ برای خوشه‌بندی دادگان در سطح ویژگی به‌منظور کاهش نرخ تغییر نقطه مرکزی هر خوشه و در نتیجه، همگرایی سریع‌تر محاسبات به‌کار رفته و با روش زیرنمونه‌برداری تصادفی اعتبارسنجی خوشه‌بند مورد استفاده بررسی می‌شود.

به‌طور کلی اهداف این بررسی عبارتند از: ارایه یک روش ترکیبی فیلتر–بسته‌بندی برای انتخاب ویژگی مناسب برای دادگان بزرگ، بهبود دقت طبقه‌بندی و سرعت انتخاب ویژگی به‌طور هم‌زمان در مقایسه با روش‌های مشابه برای انتخاب ویژگی ترکیبی، حذف ویژگی‌های دارای افزونگی، و در نتیجه افزایش کیفیت خوشه‌بندی.

روش بررسی

با افزایش اندازه مجموعه داده، زمان محاسبه افزایش می‌یابد، زیرا لازم است تمام داده‌ها در حافظه اصلی موجود باشند. همچنین، درصورت بزرگ بودن مجموعه داده، نرخ همگرایی روش k–میانگین مرسوم افت قابل‌توجهی خواهد داشت. از این‌رو، برای مجموعه دادگان بزرگ، خوشه‌بندی k–میانگین به دسته‌های تصادفی کوچک داده انتخاب مناسب‌تری است.

به‌منظور تسهیل ذخیره‌سازی در حافظه، بَچ‌های تصادفی کوچک داده با اندازه ثابت در نظر گرفته می‌شود. در هر تکرار، خوشه با دربرگرفتن نمونه‌های تصادفی جدید از مجموعه داده بروزرسانی می‌شود. با فرض مجموعه داده $D=\{x_1, x_2, x_3, \ldots, x_p\}$، که $x_i \in \mathbb{R}^{m\times n}$ است، x_i نشان‌دهنده نمونه‌های ثبت شده در یک بردار حقیقی n بُعدی با تعداد m می‌باشد. مطابق تعریف تابع هزینه، کمینه‌سازی مجموع مربعات خطا (J_{MSE})، مبتنی بر معیار فاصله اقلیدسی در رابطه (۱) و با در نظر گرفتن یک مجموعه S از مراکز خوشه $s \in \mathbb{R}^n$ منتخب از نمونه‌های ثبت شده مجموعه داده D، خوشه‌بندی k–میانگین انجام می‌شود.

$$J_{MSE} = Min\sum_{x\in D}\|f(S,x) - x\|^2 \quad (۱)$$

که در این رابطه، f(S,x) نزدیک‌ترین مرکز خوشه $s \in S$ به نمونه ثبت شده x را نتیجه داده و تعداد خوشه‌ها $k=|S|$ می‌باشد. در این پیاده‌سازی، تعداد خوشه‌ها برابر با تعداد کلاس‌ها در نظر گرفته شده است. خروجی الگوریتم k–میانگین، ماتریس برچسب کلاس منتسب به هر نمونه در سطح ویژگی و k تا مرکز نهایی تعیین شده است.

اطلاعات متقابل معیاری برای ارزیابی اطلاعات به‌دست آمده از کلاسترها در مورد ارتباط و افزونگی ویژگی و کلاس هدف است. در صورتی‌که خوشه‌بندی تصادفی انجام شود و هیچ دانشی در مورد کلاس‌ها فراهم نکند، ارزش این معیار برابر صفر و درصورت تشکیل کلاس‌ها به شکل مطلوب، بیشینه مقدار را خواهد داشت. روابط (۲–۵) این مفهوم را به‌صورت ریاضی بیان می‌کنند.

$$H(S) = -\sum_j P(s_j) log P(s_j) \quad (۲)$$

$$H(\Omega|S) = \sum_j P(s_j)\left(-\sum_k P(d_k|s_j)\log P(d_k|s_j)\right) \quad (۳)$$

$$MI(\Omega,S) = H(\Omega) - H(\Omega|S) = \sum_j\sum_k P(s_j \cap d_k) log \frac{P(s_j\cap d_k)}{P(s_j)P(d_k)} \quad (٤)$$

که در این روابط H بیانگر آنتروپی شانون (Shannon Entropy)، S مجموعه کلاس‌ها، Ω مجموعه خوشه‌ها، $P(s_j)$ احتمال حضور نمونه در کلاس s_j، $P(d_k)$ احتمال حضور نمونه در خوشه d_k، و $P(s_j \cap d_k)$ احتمال حضور نمونه در ناحیه اشتراک s_j و d_k می‌باشد. در واقع، $M1(\Omega,S)$ بیانگر فاصله تابع توزیع توام $P(s_j \cap d_k)$ با حاصل‌ضرب $P(s_j)P(d_k)$ است. درصورتی‌که Ω و s از هم مستقل آماری باشند، تابع

توزیع توام برابر با حاصل ضرب احتمالات بوده و بنابراین فاصله بین این دو مقدار صفر می‌شود. با بهنجارش این رابطه و محدود کردن ارزش آن بین صفر و یک به‌دست خواهیم آورد:

$$NMI(\Omega,S)=\frac{MI(\Omega,S)}{[H(\Omega)+H(S)]/2} \quad (۵)$$

بهنجارش نشان می‌دهد که تعداد کمتر خوشه‌ها مطلوب است، زیرا آنتروپی معمولا با افزایش تعداد خوشه افزایش می‌یابد.

روش‌های انتخاب ویژگی نقش مهمی در کارکرد بهینه مدل‌های یادگیری دارند. روش ترکیبی به‌کار رفته در این کار برای انتخاب ویژگی‌های بهینه از دو تابع معیار و دو فاز متناظر مجزا تشکیل شده است، ابتدا تابع رتبه‌بندی ویژگی مبتنی بر روش فیلتر به‌کار رفته و سپس ویژگی‌های نامزد با یک روش بسته‌بندی که به خروجی الگوریتم یادگیری بستگی دارد، انتخاب می‌شوند. در این کار، از ماشین بردار پشتیبان خطی به‌عنوان طبقه‌بند استفاده شده، با این وجود، انتظار می‌رود این رویکرد به انواع دسته‌بندها قابلیت تعمیم داشته باشد. هم‌چنین این روش آسیب‌پذیری بسیار کمی در برابر بیش‌برازش داشته، محاسبات با سرعت بالایی انجام شده و بدون نیاز به ورود تعداد ویژگی بهینه در الگوریتم یادگیری توسط کاربر، بهترین زیرمجموعه ویژگی به‌دست می‌آید.

فاز اول، رتبه‌بندی ویژگی: رویکرد اصلی، خوشه‌بندی ویژگی‌ها با هدف توسعه یک الگوریتم انتخاب بوده که زمان محاسبات کمتری در مقایسه با الگوریتم‌های مرسوم داشته باشد و از این‌رو، روش k–میانگین مینی‌بَچ به‌کار می‌رود.[۲۴]

مینی‌بَچ‌ها زیرمجموعه‌های داده ورودی هستند که به‌طور تصادفی در هر تکرار نمونه‌برداری می‌شوند. پس از تشکیل مینی‌بَچ‌ها با تابع هدف یکسان، نمونه‌های داخل هر یک به نزدیک‌ترین مرکز خوشه منتسب شده و در گام بعد، مراکز بَچ‌های کوچک داده با متوسط گرفتن از نمونه‌های منتسب به آن مراکز بروزرسانی می‌شوند.

این مراحل تا رسیدن به همگرایی یا تعداد از پیش تعیین شده تکرارها انجام می‌شود و تاثیر این پیاده‌سازی، کاهش نرخ تغییر برای هر مرکز است. کیفیت خوشه معیاری برای یافتن ارتباط ویژگی با کلاس هدف بوده و هرچه کیفیت خوشه افزایش یابد، ویژگی مرتبط‌تر و با درجه اهمیت بیشتری به‌شمار می‌آید. معیار رتبه‌بندی، NMI، با مقدار واحد نشان‌دهنده بهترین دسته‌بندی با استفاده از ویژگی‌های نامزد است. این معیار برای همه ویژگی‌ها جداگانه ارزیابی شده و با مقایسه آن‌ها یک لیست رتبه‌بندی براساس ارتباط هر ویژگی نامزد و برچسب کلاس به‌دست می‌آید.

فاز دوم، انتخاب زیرمجموعه ویژگی بهینه: متغیرهای ویژگی دارای افزونگی تمایل به تولید نتایج نامتوازن داشته و از این‌رو، از لیست رتبه‌بندی به‌دست آمده از فاز اول (روش فیلتر) حذف می‌شوند. برای این منظور، از دو رویکرد پیشنهادی در یک مطالعه برای انتخاب زیرمجموعه ویژگی بهینه استفاده شده و نتایج به‌دست آمده از آن‌ها مقایسه می‌شوند.[۲۵] با این تفاوت که در فاز دوم (روش بسته‌بندی) در مطالعه مذکور، جنگل تصادفی و در این پژوهش، ماشین بردار پشتیبان خطی برای دسته‌بندی داده‌ها به‌کار رفته است.

نتیجه به‌کارگیری SVM، یک اپتیمم سراسری خواهد بود، این دسته‌بند تعمیم‌پذیری خوبی دارد، بیش‌برازش رخ نمی‌دهد و به داده‌های پرت (outliers) نیز حساس نیست، زمان مورد نیاز برای حل مسئله، چند جمله‌ای بوده و پاسخ آن پایدار و قابل تکرار است؛ جواب مسئله فقط وابسته به تعداد محدودی از نمونه‌های مجموعه داده (بردارهای پشتیبان) بوده و از این رو مسئله با یک روش خلوت (sparse) حل می‌شود با محاسبه تابع بهینه‌سازی به روش دوگان لاگرانژی (Lagrangean duality)، عملکرد SVM مستقل از بُعد داده‌ها خواهد بود با کنترل پارامتر هزینه بین خطای دسته‌بندی و حاشیه بیشینه در LSVM می‌توان کارآیی دسته‌بند را تا حد امکان بهبود داد.

شمول ویژگی: ویژگی‌های رتبه‌بندی‌شده از فاز اول به‌ترتیب نزولی رتبه، یک به یک به زیرمجموعه ویژگی انتخابی که در ابتدا تهی مفروض است، اضافه می‌شوند و در هر تکرار، پس از افزودن ویژگی با بالاترین رتبه، شاخص‌های ارزیابی محاسبه می‌گردند. این روند تا زمانی‌که ویژگی با پایین‌ترین رتبه وارد زیرمجموعه ویژگی‌ها شود، ادامه می‌یابد. در صورتی‌که افزودن ویژگی در هر مرتبه بهبود دقت دسته‌بندی را نتیجه دهد، آن ویژگی در زیرمجموعه ویژگی‌های انتخابی گنجانده شده و در غیراینصورت حذف می‌شود (الگوریتم ۱).

الگوریتم ۱: شبه کد رتبه‌بندی مبتنی بر شمول ویژگی با رویکرد دست‌یابی به زیرمجموعه ویژگی بهینه.

Input: The total set of ranked features, $T=\{t_0, t_1, t_2, \ldots, t_n\}$, where n=total number of features, t_0=the highest ranked feature, t_n=the least ranked feature

Output: The selected set of features
Begin
1. Lst=T [0] and prev = 0
2. [m,n]=size (data); Loop process
3. for k=0 to n-1, do (train the model based on LSVM classifier and store the accuracy on acc)
4. data=data [Lst];
5. [$label_{train}$ $data_{train}$ $label_{test}$ $data_{test}$]=RandomSubsampling (data, 100, m, 1); svmparms='-s 0 -t 0 -c 100 ';
6. model = svmtrain (label ($label_{train}$), $data_{train}$, svmparms);
7. [$label_{assigned}$, acc, decision_value]=svmpredict (label($label_{test}$), $data_{test}$, model);
8. if acc>prev then
9. if (k≠n-1) then
10. Add T[k+1] into the Lst
11. prev=acc
12. else
13. Print Lst
14. end if
15. else
16. Remove T [k] object from the Lst
17. if (k≠n-1) then
18. Add T [k+1] to the Lst
19. else
20. Print Lst
21. end if
22. end if
23. end for
24. return Lst End

خروج ویژگی: در این رویکرد، ابتدا زیرمجموعه ویژگی انتخابی شامل همه ویژگی‌های موجود در مجموعه داده فرض می‌شود. ویژگی‌ها از پایین‌ترین رتبه (مستخرج از فاز محاسبه اطلاعات متقابل بهنجار شده) به‌ترتیب صعودی، یک به یک از لیست کل مجموعه ویژگی‌ها حذف شده و در هر گام، دقت دسته‌بندی برای لیست متناظر محاسبه می‌شود. این فرآیند تا زمانی‌که لیست تهی شود، ادامه می‌یابد. در پایان، بالاترین کارآیی دسته‌بندی در میان همه تکرارها به‌عنوان خروجی این الگوریتم لحاظ شده و مجموعه ویژگی متناظر انتخاب می‌شود (الگوریتم ۲).

به‌عبارتی، ورودی هر دو الگوریتم اشاره شده، مجموعه ویژگی‌های رتبه‌بندی‌شده است که با محاسبه NMI در فاز اول به‌دست آمده و در خروجی، زیرمجموعه ویژگی‌های انتخاب‌شده نمایش داده می‌شوند. این دو رویکرد، محدودیتی برای کاربر در اعمال تعداد ویژگی‌های لازم به‌عنوان پارامتر ورودی ایجاد نمی‌کنند.

به‌منظور سنجش توانایی این روش‌ها در انتخاب ویژگی‌های مرتبط، در رویکرد سوم همه ویژگی‌های موجود در مجموعه داده برای دسته‌بندی و مقایسه معیارهای ارزیابی در نظر گرفته می‌شوند.

الگوریتم ۲: شبه کد رتبه‌بندی مبتنی بر خروج ویژگی با رویکرد دستیابی به زیرمجموعه ویژگی بهینه.

Input: The total set of ranked features, $T=\{t_0, t_1, t_2, \ldots, t_n\}$, where n=total number of features, t_0=the least ranked feature, t_n =the highest ranked feature
Output: The resultant subset for every eliminated feature from the feature list
Begin
1. Lst=T and prev=0
2. [m,n]=size (data); Loop process
3. for k=0 to n-1, do (train the model based on LSVM classifier and store the accuracy on acc)
4. data=data [Lst];
5. [$label_{train}$ $data_{train}$ $label_{test}$ $data_{test}$]=RandomSubsampling (data, 100, m, 1);
6. svmparms=-s 0 -t 0 -c 100
7. model=svmtrain (label ($label_{train}$), $data_{train}$, svmparms)
8. [$label_{assigned}$, acc, decision_value]=svmpredict (label($label_{test}$), $data_{test}$, model);
9. if acc>prev then
10. prev=acc//to store the greatest accuracy
11. feature=k//to store the no. of feature eliminated
12. end if
13. Delete Lst [0] //deleting the least ranked feature
14. end for
15. return End

الگوریتم‌های انتخاب ویژگی برای دو رویکرد اشاره شده در نرم‌افزار Matlab R2021a برای ۱۵ مجموعه داده UCI در شاخه علوم زیستی با تعداد متفاوت ویژگی، نمونه و کلاس با هدف شناسایی بیماری و یا بقای بیمار مطابق جدول ۱ پیاده‌سازی شده‌اند.

در مرحله پیش‌پردازش، ویژگی‌های توصیفی در مجموعه دادگان دارای چنین ویژگی‌هایی به کُدهای عددی معادل کاراکترهای اَسکی American standard code for information interchange (ASCII) تبدیل شده‌اند. همچنین، به نمونه‌های با مقادیر گمشده از مجموعه داده ارزش صفر نسبت داده شده است.

به‌منظور اعتبارسنجی و تعیین مقدار مناسب پارامتر هزینه بین خطای دسته‌بندی و بیشینه حاشیه دسته‌بند ماشین بردار پشتیبان خطی از روش زیرنمونه‌برداری تصادفی استفاده می‌شود. به‌علاوه، در هر آزمایش، ۱۵–۱۰ مرتبه شبیه‌سازی با نمونه‌های آموزشی تصادفی جدید تکرار شده و میانگین دقت دسته‌بندی محاسبه می‌گردد. دسته‌بند LSVM در این پژوهش با استفاده از جعبه ابزار Libsvm، با تنظیم پارامتر هزینه (C) برابر ۱۰۰ و انتخاب تابع هسته خطی پیاده‌سازی شده است.[۲۶]

یافته‌ها

مطابق روش‌های ذکر شده برای انتخاب ویژگی در بخش ۳، هرچه مقدار NMI بزرگ‌تر باشد، رتبه تخصیص‌یافته به ویژگی متناظر بالاتر است. در پیاده‌سازی کد NMI از رابطه آنتروپی شانون (شرطی) حاکم میان توابع توزیع احتمال متغیرهای تصادفی $P(d_k)$ و $P(s_j)$ مطابق روابط (۵-۲) استفاده شده است.

شکل‌های ۱ (الف) تا ۱۵ (الف)، مقادیر رتبه حاصل از اطلاعات متقابل بهنجار شده را به‌ازای تمام ویژگی‌های موجود در مجموعه دادگان بررسی شده نمایش می‌دهند. تغییرات مقادیر NMI و دقت محاسبه شده به‌ازای هر ویژگی افزوده شده (در رویکرد شمول ویژگی) یا حذف شده از زیرمجموعه ویژگی‌های انتخابی (در رویکرد خروج ویژگی) برای مجموعه دادگان، می‌تواند بازتاب تغییر الگوی ویژگی‌های انتخابی در طول آموزش داده‌ها باشد.

در شکل‌های ۱ (ب) تا ۱۵ (ب) نیز دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به زیرمجموعه ویژگی‌های انتخابی در رویکرد شمول ویژگی برای مجموعه دادگان بررسی شده نمایش داده شده است. همچنین شکل‌های ۱ (ج) تا ۱۵ (ج) دقت محاسبه شده به‌ازای هر ویژگی حذف شده با پایین‌ترین رتبه از زیرمجموعه ویژگی‌های انتخابی در رویکرد خروج ویژگی برای مجموعه دادگان بررسی شده را نشان می‌دهند.

بررسی‌ها نشان می‌دهند در مجموعه دادگان با بیشترین اطلاعات متقابل بین ویژگی و کلاس هدف (مقادیر بالای NMI در نمودارهای رتبه‌بندی ویژگی)، با تقریب خوبی هر دو الگوریتم پیشنهادی انتخاب ویژگی با اتخاذ کمترین تعداد ویژگی، از دقت بالاتری در مقایسه با زمانی که پیش‌پردازشی روی ویژگی‌ها اعمال نشده، برخوردار است. مزیت این دو رویکرد نسبت به روش‌های مرسوم انتخاب ویژگی، قابلیت انتخاب بهترین زیرمجموعه ویژگی به‌صورت خودکار و بدون نیاز به کاربر برای تعیین تعداد ویژگی‌های انتخابی است. معیارهای ارزیابی شامل دقت دسته‌بندی و امتیاز F1 که تابعی از صحت و پوشش است، برای هر دو رویکرد انتخاب ویژگی و مجموعه کل ویژگی‌ها در هر مجموعه داده مطابق روابط (۸-۶) محاسبه می‌شوند.[۲۷]

صحت، بیانگر این است چه تعداد از نمونه‌های انتخابی درست هستند، و پوشش بدین مفهوم است که چه تعداد از نمونه‌های صحیح موجود انتخاب شده‌اند. امتیاز F1 نیز نشان‌دهنده میانگین وزن‌دار صحت و پوشش هر کلاس است. در این روابط، TrPos بیانگر این است که مدل به‌درستی نمونه‌های متعلق به یک کلاس را پیش‌بینی کرده و برچسب می‌زند.

FlPos بیانگر این است که مدل به غلط نمونه‌های غیرمتعلق به یک کلاس را منتسب به کلاس پیش‌بینی کرده و برچسب می‌زند. FlNeg بیانگر این است که مدل به غلط نمونه‌های متعلق به کلاس را غیرمنتسب به کلاس پیش‌بینی کرده و برچسب می‌زند. TrNeg بیانگر این است که مدل به درستی نمونه‌های غیرمتعلق به کلاس را پیش‌بینی کرده و برچسب می‌زند.

$$Accuracy = \frac{\sum(TrPos+TrNeg)}{\sum total\ population} = \frac{\sum(TrPos+TrNeg)}{\sum(TrPos+TrNeg+FlPos+FlNeg)} \quad (۶)$$

$$\begin{cases} Precision\ score = Pr = \frac{\sum TrPos}{\sum(TrPos+FlPos)} \\ Recall\ score = Rc = \frac{\sum TrPos}{\sum(TrPos+FlNeg)} \end{cases} \quad (۷)$$

$$F1\ score = 2\left(\frac{Pr \times Rc}{Pr+Rc}\right) \quad (۸)$$

بحث

نتایج مستخرج از هر سه رویکرد انتخاب ویژگی از لحاظ معیارهای مختلف ارزیابی محاسبه و در جدول ۱ مقایسه شده‌اند. در یک بررسی کلی مشاهده می‌شود در مجموعه دادگان با ابعاد بالاتر ویژگی، نظیر داده‌های سرطان ریه، سرطان دهانه رحم، پوست‌شناسی و بیان پروتیین، رویکرد شمول ویژگی به بهبود شاخص‌های ارزیابی کمک بیشتری کرده است.

مطابق جدول ۱، هر دو رویکرد پیشنهادی برای مجموعه دادگان سرطان ریه، سرطان پستان و اکوکاردیوگرام، با انتخاب کمترین تعداد ویژگی ممکن، دقت بسیار بالاتری در مقایسه با حضور همه ابعاد ویژگی نشان داده است. این در حالی است که دقت بالا در مجموعه دادگان التهاب حاد، پارکینسون، اوتیسم کودکان و بیان پروتیین با تمایل به حفظ یا حذف همه ویژگی‌ها در مجموعه داده به‌دست می‌آید.

جدول ۱: معیارهای ارزیابی روش انتخاب ویژگی مورد استفاده در این پژوهش مبتنی بر دسته‌بند LSVM.

مجموعه داده بالینی	رویکرد انتخاب ویژگی	تعداد ویژگی گنجانده شده (+) یا تعداد ویژگی خارج شده (-)	دقت دسته‌بندی (درصد)	امتیاز F1 (درصد)
	شمول ویژگی	+۱۲	۸۳/۳۳۳	۹۷/۹۷۳
لنفوگرافی	خروج ویژگی	-۹	۷٤/۵۱۰	۹۱/۲۱٦
	حضور کل ویژگی‌ها	۱۸	۷٤/۵۰۹	۹۱/۲۱٦
	شمول ویژگی	+۱۲	٦۱/٦٦۷	۸٤/۳۷۵
سرطان ریه	خروج ویژگی	-۳۲	٦۲/۵۰۰	۸۱/۲۵۰
	حضور کل ویژگی‌ها	۵٦	۵۰/۰۰۰	۸۱/۲۵۰
	شمول ویژگی	+٤	۹٦/۹۲۳	۹۸/٤۲٦
سرطان پستان	خروج ویژگی	-۵	۹۳/۸۲۳	۹٤/۹۹۳
	حضور کل ویژگی‌ها	۱۰	۷۱/۵٤۳	۷۹/٦۸۵
	شمول ویژگی	+۱۰	۹۵/۸۵۵	۹٦/۸۵۳
سرطان دهانه رحم	خروج ویژگی	-۲۰	۹٤/۷۲۳	۹۳/۵۸۹
	حضور کل ویژگی‌ها	۳٦	۹۳/۲۱۳	۹٤/۲۸۹
	شمول ویژگی	+۳	۹۸/۷۸۳	۹۸/۲۱۱
ویروس هپاتیت C	خروج ویژگی	-٤	۹۸/۹۲٤	۹۹/۰۲٤
	حضور کل ویژگی‌ها	۱۳	۹۸/۰٤۷	۹۸/۵۳۷
	شمول ویژگی	+٦	۹۸/٦٤۵	۹۹/۷۰۲
عفونت باکتریایی E.coli	خروج ویژگی	-٤	۹۸/٦۲۱	۹۹/٤۰۵
	حضور کل ویژگی‌ها	۸	۹۹/۰۹۱	۹۹/۷۰۲
التهاب حاد مثانه و نارسایی مزمن کلیه	شمول ویژگی	+٦	۹۹/٤۲۳	۹۹/۵۷۱
	خروج ویژگی	-۳	۹۹/۸۵٤	۱۰۰/۰۰۰
	حضور کل ویژگی‌ها	٦	۹۹/٤۲۳	۹۹/۵۷۱
	شمول ویژگی	+٤	۸۲/٦۱۹	۹۰/۹٦۹
نارسایی قلبی	خروج ویژگی	-٤	۷۷/۰۲۰	۸۵/۲۸٤
	حضور کل ویژگی‌ها	۱۳	۸۳/۵۸۲	۹۲/٦٤۲
	شمول ویژگی	+۳	۹۰/۱۲۵	۹۹/۲٤۲
اکوکاردیوگرام	خروج ویژگی	-٦	۸۹/۳۹٤	۹٤/٦۹۷
	حضور کل ویژگی‌ها	۱۲	۷۵/۰۰۰	۹٦/۹٦۹
	شمول ویژگی	+۲۳	۹۹/٦۱۲	۹۹/۷۱۱
پارکینسون	خروج ویژگی	-۲۱	۹۹/۲۱۷	۹۹/۵۹٤
	حضور کل ویژگی‌ها	۲۸	۹۸/۸۰۹	۹۹/۰۳۸
	شمول ویژگی	+۱٦	۹٦/۵۸٦	۹۹/٤۵٤
پوست‌شناسی	خروج ویژگی	-۱٤	۹۷/۵۲۱	۹۸/۹۰۷
	حضور کل ویژگی‌ها	۳٤	۹۳/۹۳۹	۹۸/۳٦۱
	شمول ویژگی	+۲۰	۹۹/۹۷۳	۹۹/۸۹۱
اختلال طیف اوتیستیک کودکان	خروج ویژگی	-۱۹	۹۹/۹۵۲	۹۹/۸۳۳
	حضور کل ویژگی‌ها	۲۰	۹۹/۹۷۳	۹۹/۸۹۱
آسیب شبکیه چشم در اثر دیابت (رتینوپاتی دیابتی)	شمول ویژگی	+۵	۷۲/٦۰۲	۷٦/٤۵۵

جدول ۱: معیارهای ارزیابی روش انتخاب ویژگی مورد استفاده در این پژوهش مبتنی بر دسته‌بند LSVM. (ادامه)

مجموعه داده بالینی	رویکرد انتخاب ویژگی	تعداد ویژگی گنجانده شده (+) یا تعداد ویژگی خارج شده (-)	دقت دسته‌بندی (درصد)	امتیاز F1 (درصد)
رفتار دنباله ژن پیش‌برنده در بیولوژی مولکولی	خروج ویژگی	−۱۰	۷۲/۲۱۷	۷۵/۵۸۶
	حضور کل ویژگی‌ها	۱۹	۷۳/۳۹۶	۷۸/۰۱۹
	شمول ویژگی	+۷	۹۲/۷۷۸	۹۸/۱۱۳
	خروج ویژگی	−۳۶	۹٤/٤٤٤	۹۹/۰۱۲
	حضور کل ویژگی‌ها	۵۸	۹۷/۲۲۲	۹۹/۰۵٦
	شمول ویژگی	+۷۱	۹۹/۹۵٦	۱۰۰/۰۰۰
بیان پروتیین	خروج ویژگی	−۷۳	۹۹/۸۲۱	۱۰۰/۰۰۰
	حضور کل ویژگی‌ها	۸۱	۹۹/٦٤۳	۹۹/۹۰۷

علت این رخداد این است که بیشتر ویژگی‌ها در این مجموعه دادگان از نوع Boolean (بله/خیر) و یا توصیفی (مشخصه رفتاری، مصوت‌ها، نام بیمار، کشور محل سکونت و توالی پروتیین) بوده که عملکرد الگوریتم یادگیری برای این‌گونه داده‌ها وابستگی زیادی به شیوه کدگذاری پیدا می‌کند. به‌عبارتی، برای چنین مجموعه دادگانی استقلال ویژگی‌های توصیفی (مطابق نمودارهای رتبه‌بندی NMI) دقت دسته‌بندی نزدیک ٪۱۰۰ را نتیجه داده که در این شرایطِ تعریف شده معتبر است. بنابراین پیاده‌سازی روش پیشنهادی انتخابِ ویژگی مبتنی بر LSVM برای مجموعه دادگان عددی قابلیت اطمینان دارد. همچنین، اگرچه حفظ تمام ویژگی‌های موجود در مجموعه دادگان با بعُد پایین، مانند داده باکتری E.coli با ۸ ویژگی، بالاترین دقت دسته‌بندی برابر با ۹۹/۰۹٪ و امتیاز F1 برابر با ۹۹/۷۰٪ را نتیجه داده، رویکرد شمول ویژگی به دقت قابل قبول ۹۸/٦۵٪ بدون تغییر امتیاز F1 دست یافته است.

از طرفی، در بیشتر مجموعه دادگان مورد بررسی، امتیاز F1 نرخ بالاتری در مقایسه با دقت دسته‌بند LSVM نشان داده که بیانگر این است که در دسته‌بندی نمونه‌های متعلق به کلاس هدف نسبت به دسته‌بندی نمونه‌های غیرمتعلق به آن کلاس، خطای کمتری رخ داده است (حساسیت اندکی بیشتر از اختصاصیت).

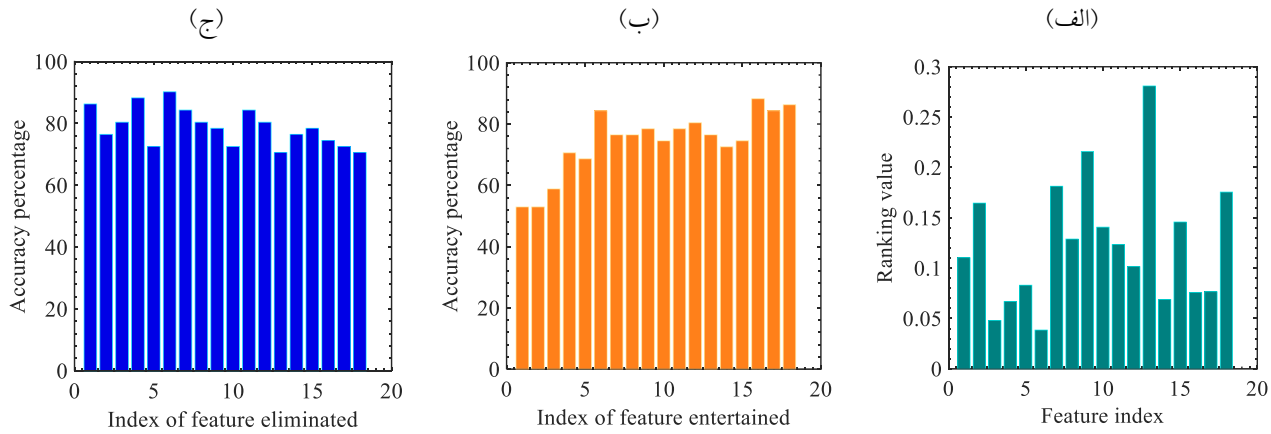
در این راستا، هر دو رویکرد انتخاب ویژگی پیشنهادی برای مجموعه دادگان بیولوژی مولکولی، ویروس هپاتیت C و باکتری E.coli، امتیاز صحت و فراخوانی بالای ۹۸٪ را نتیجه داده‌اند، که به معنای حضور تعداد بسیار کم موارد مثبت کاذب و منفی کاذب در دسته‌بندی با LSVM است (FlPos=FlNeg≈0).

در میان مجموعه دادگان عددی قابل کاربرد برای دسته‌بند LSVM، بهترین نتیجه پیاده‌سازی روش‌های انتخاب ویژگی به‌کار رفته، برای مجموعه داده ویروس هپاتیت C با انتخاب ۹ ویژگی مرتبط از بین ۱۳ ویژگی موجود با روش خروج ویژگی به‌دست آمده، که دقت دسته‌بندی ۹۸/۹۲٪ و امتیاز F1 ۹۹/۰۲٪ را نتیجه داده است. رویکرد شمول ویژگی نیز با یک اختلاف جزیی، دقت مطلوب ۹۸/۷۸٪ را نتیجه داده است.

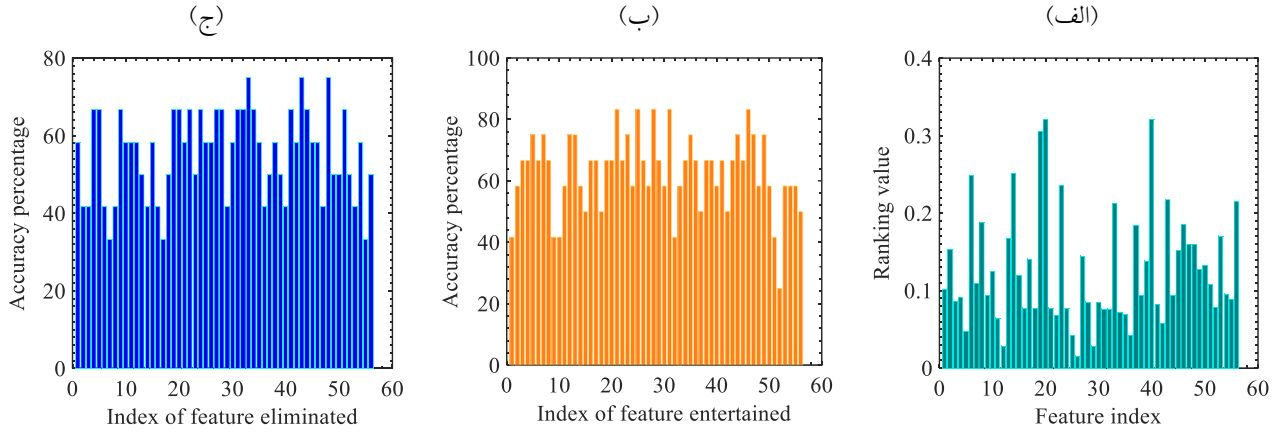
به‌منظور ارزیابی عملکرد انتخاب ویژگی پیشنهادی، نتایج حاصل با روش‌های انتخاب ویژگی گزارش‌شده در ادبیات برای سه مجموعه داده پُرکاربرد سرطان پستان، لنفوگرافی و پارکینسون در جدول ۲ مقایسه شده‌اند.

الگوریتم‌های انتخاب ویژگی در همه بررسی‌ها به ایجاد دسته‌بندی بهتر با لیست کردن ویژگی‌های ارجح کمک می‌کنند، که به نوبه خود، در کاهش بار محاسباتی نیز نقش دارند. با مقایسه نتایج به‌دست آمده از رویکرد انتخاب ویژگی پیشنهادی و سایر روش‌ها به عملکرد مطلوب انتخاب ویژگی در این پژوهش پی می‌بریم

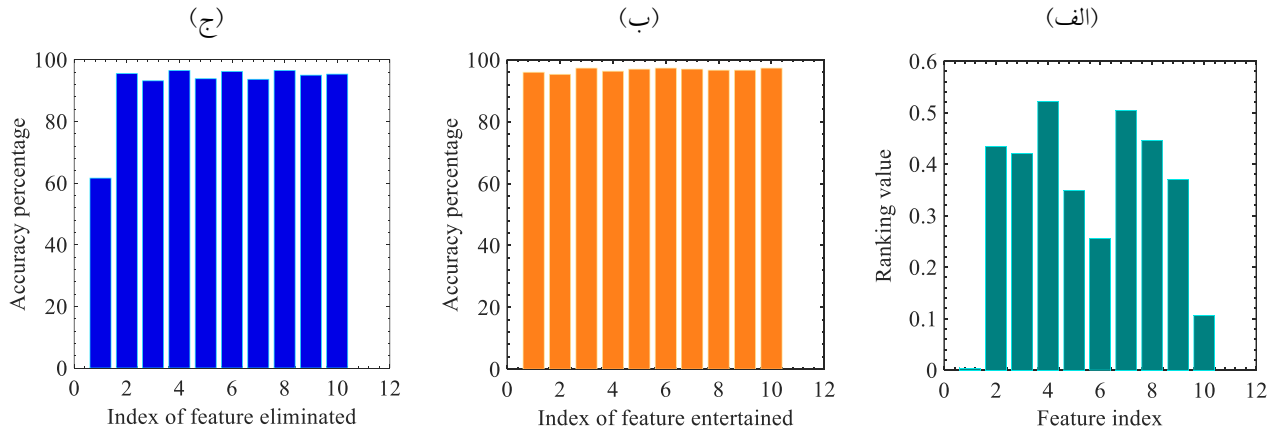
برای مجموعه دادگان سرطان پستان و پارکینسون، به‌دلیل توزیع غیرمتوازن نمونه‌ها در کلاس‌ها، مقادیر دقت و سایر شاخص‌های ارزیابی منتج از رویکردهای گوناگون تفاوت بیشتری دارد.[۳۶و۳۷]



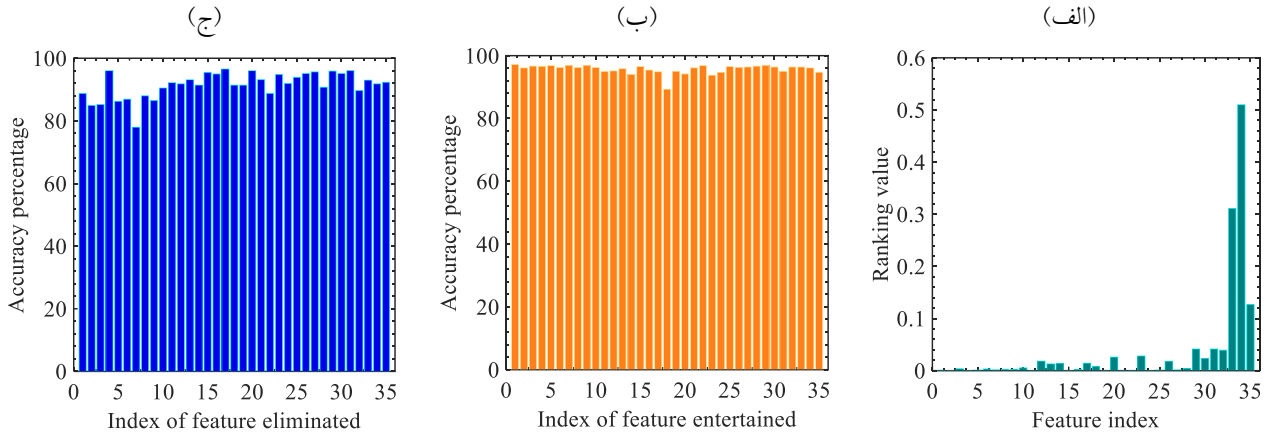
شکل ۱: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجار شده به‌ازای همه ویژگی‌های موجود در مجموعه داده لنفوگرافی. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



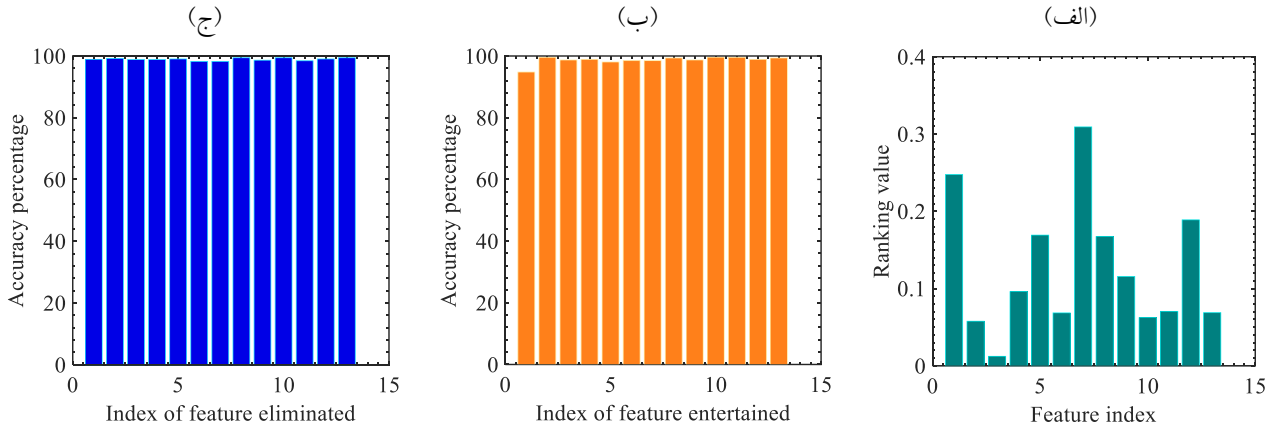
شکل ۲: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجار شده به‌ازای همه ویژگی‌های موجود در مجموعه داده سرطان ریه. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



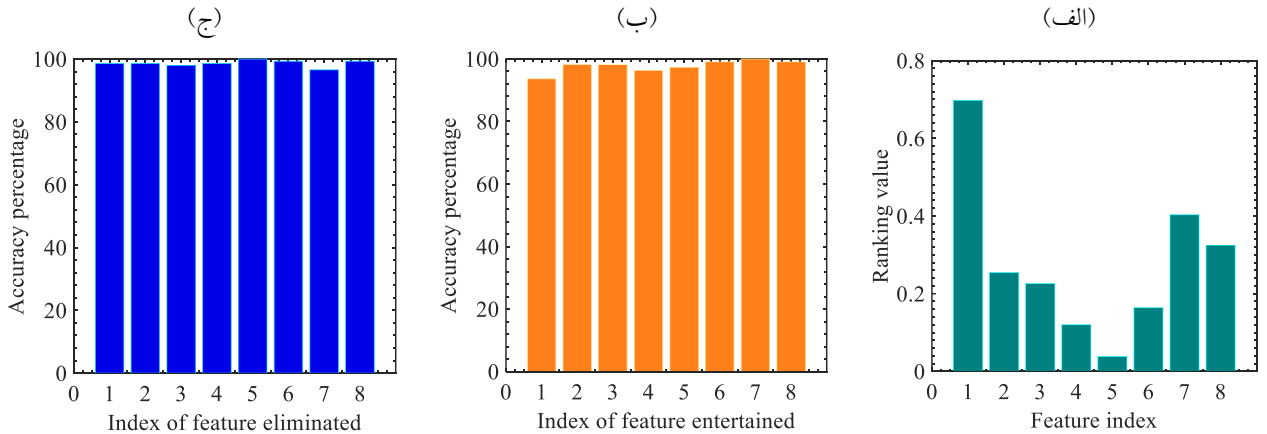
شکل ۳- (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجار شده به‌ازای همه ویژگی‌های موجود در مجموعه داده سرطان پستان. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در رویکرد شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



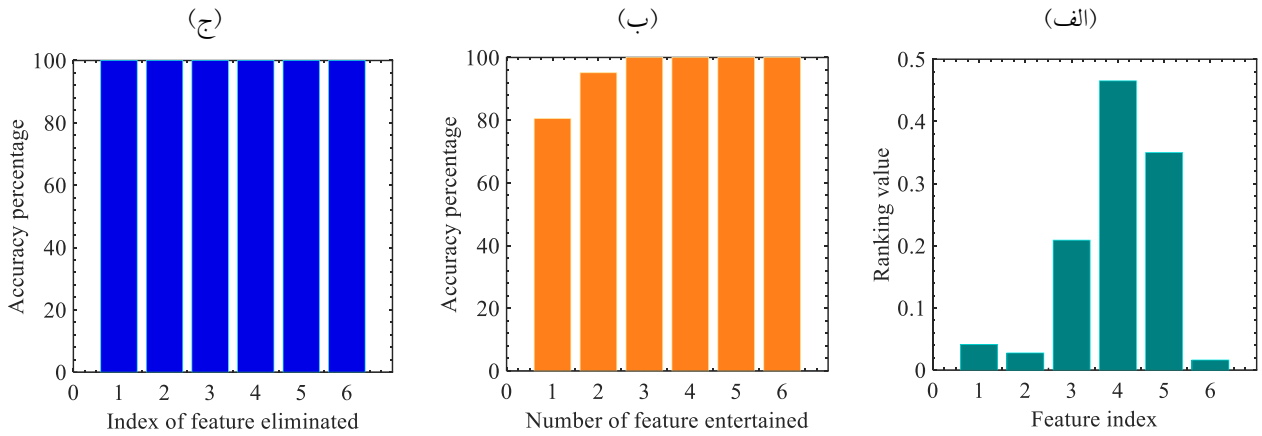
شکل ٤: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده سرطان دهانه رحم. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



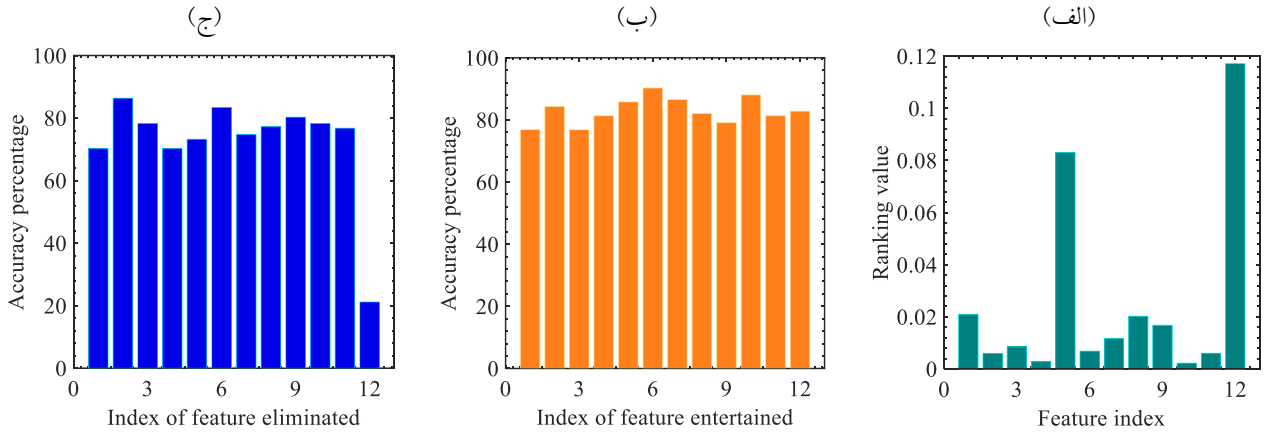
شکل ٥: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده ویروس هپاتیت C. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



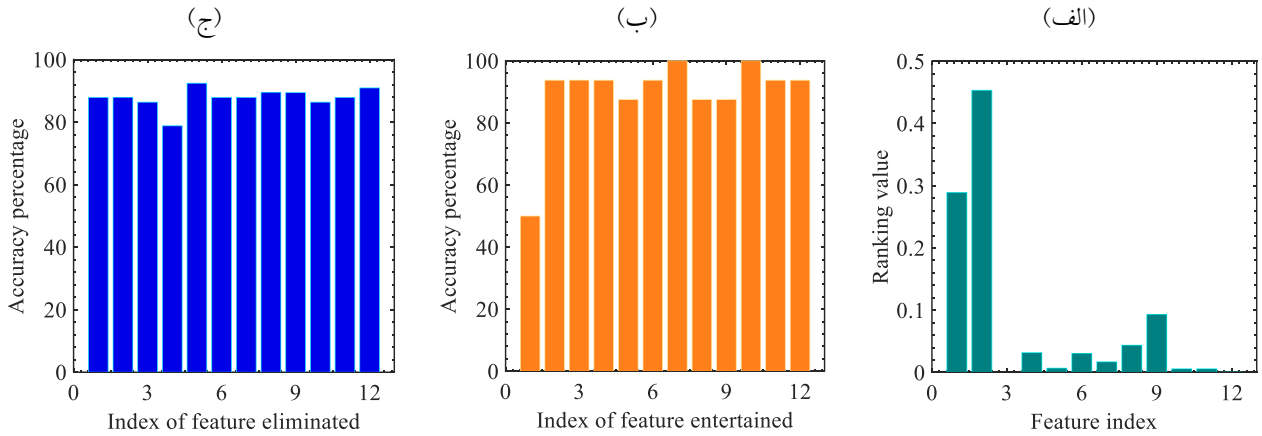
شکل ٦: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده عفونت باکتریایی E.coli. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



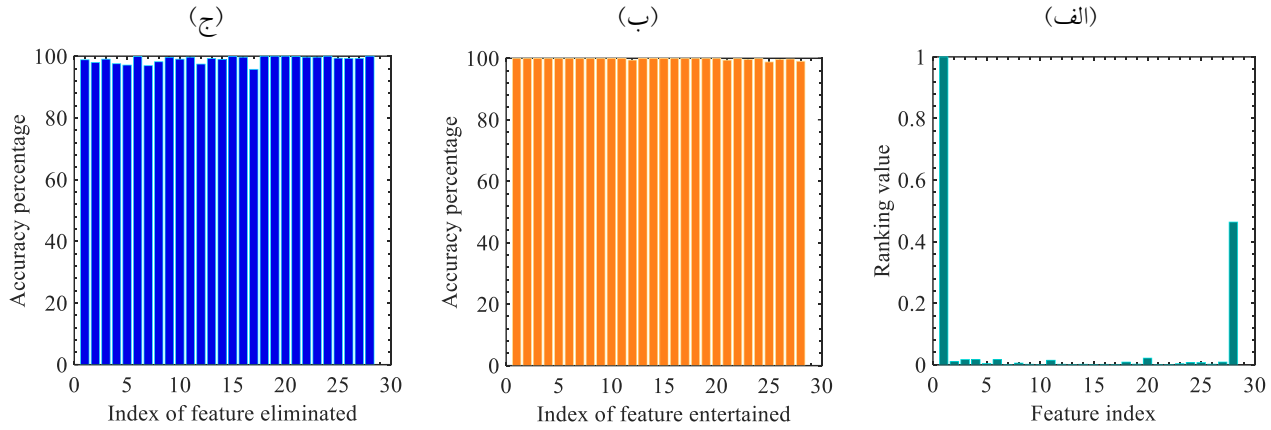
شکل ۷: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده التهاب حاد مثانه و نارسایی مزمن کلیه. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



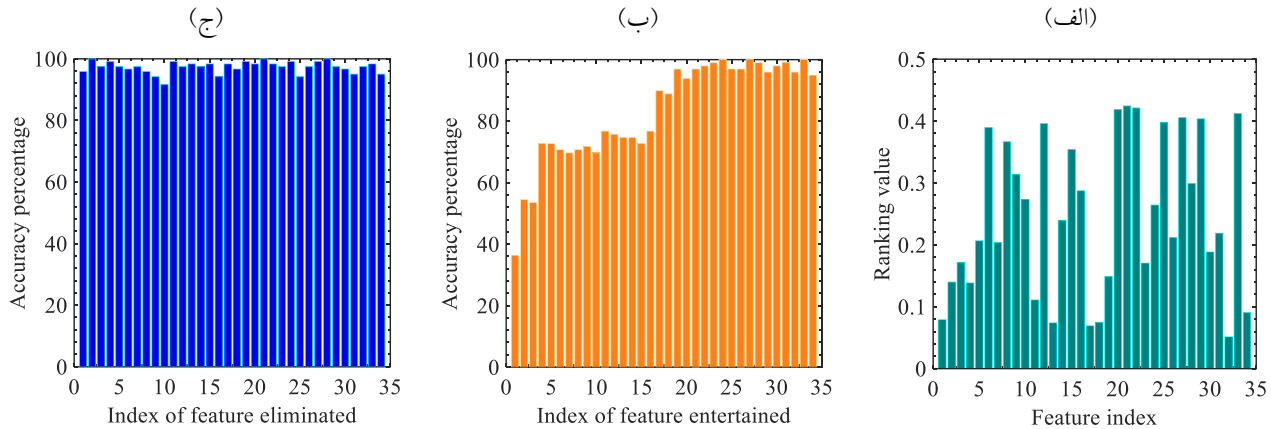
شکل ۸: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده نارسایی قلبی. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



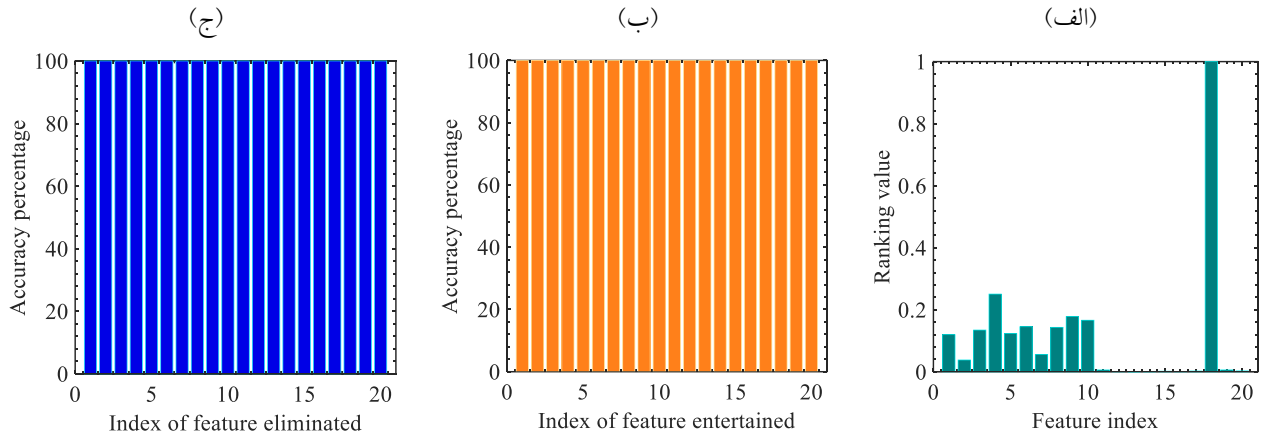
شکل ۹: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده اکوکاردیوگرام. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



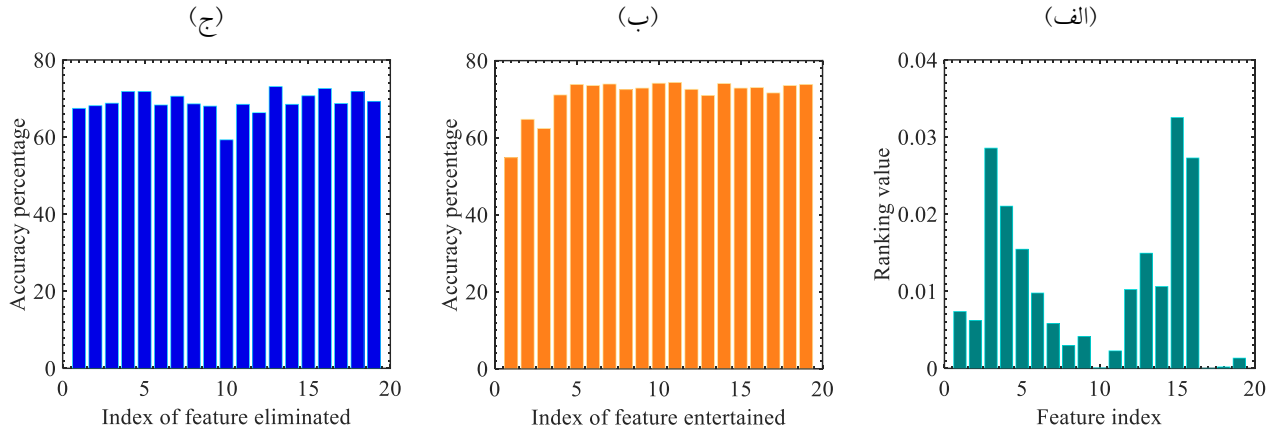
شکل ۱۰: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده گفتاری پارکینسون. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



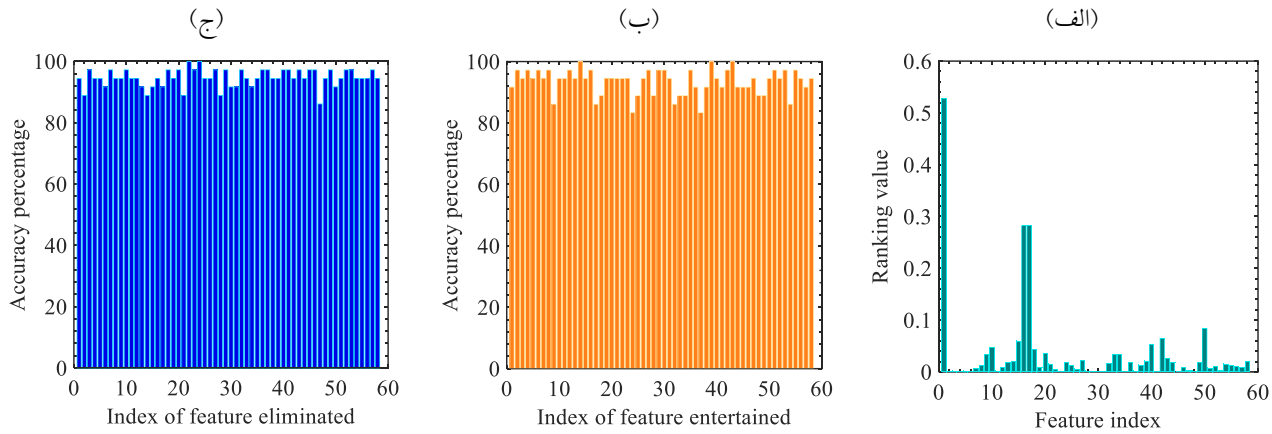
شکل ۱۱: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده پوست‌شناسی. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



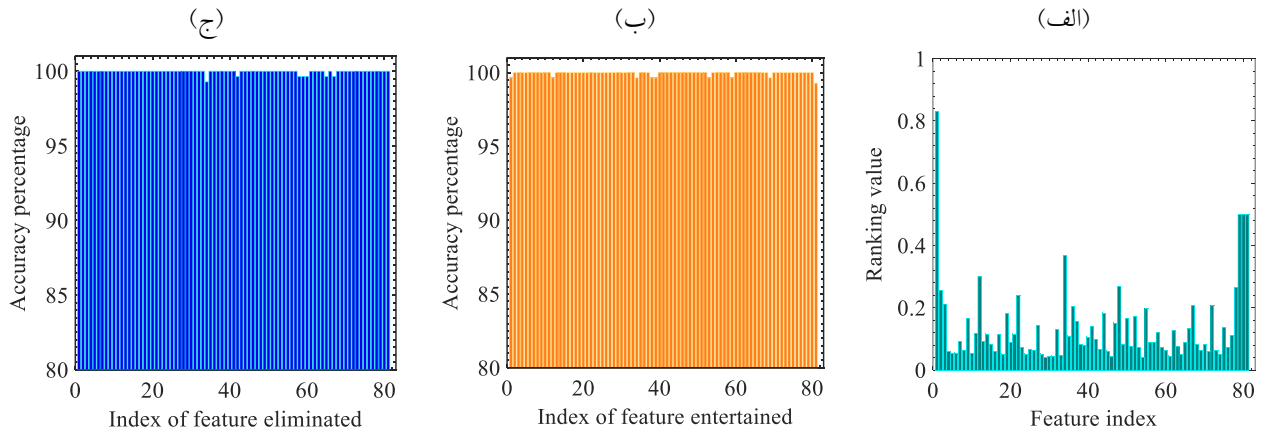
شکل ۱۲: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده اختلال طیف اوتیستیک کودکان. (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



شکل ۱۳: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده رتینوپاتی دیابتی. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.



شکل ۱٤: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجار شده به‌ازای همه ویژگی‌های موجود در مجموعه داده بیولوژی مولکولی (دنباله ژن پروموتر). (ب) دقت محاسبه شده به‌ازای هر ویژگی اضافه شده به مجموعه ویژگی‌های انتخابی در رویکرد شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با پایین‌ترین رتبه در روش خروج ویژگی.



شکل ۱۵: (الف) رتبه‌بندی حاصل از اطلاعات متقابل بهنجارشده به‌ازای همه ویژگی‌های موجود در مجموعه داده بیان پروتئین موش. (ب) دقت محاسبه شده برای هر ویژگی اضافه‌شده به مجموعه ویژگی‌های انتخابی در روش شمول ویژگی. (ج) دقت به‌دست آمده از حذف ویژگی با کمترین رتبه در روش خروج ویژگی.

جدول ۲: مقایسه معیارهای ارزیابی برای روش انتخاب ویژگی پیشنهادی در این پژوهش مبتنی بر دسته‌بند LSVM و روش‌های گزارش‌شده در ادبیات

مجموعه داده	رویکرد انتخاب ویژگی	تعداد ویژگی انتخاب شده	دقت دسته‌بندی (درصد)	امتیاز F1 (درصد)
	۲۸	٤	۷۲/۲۰۰	۷٦/٤۹۰
	۲۹	۹	۸۷/۸٤۲	۹۰/۱۲۹
سرطان پستان	۳۰	۱	۷٦/۰۰۰	۷۰/۰۰۰
	بررسی حاضر (شمول ویژگی)	٤	۹٦/۹۲۳	۹۸/٤۲٦
	بررسی حاضر (خروج ویژگی)	۵	۹۳/۸۲۳	۹٤/۹۹۳
	۳۱	۱۰	۷۲/۸۰۰	۸۸/۳۲۰
	۳۲	۲	۷٤/۱۲۰	---
لنفوگرافی	۲۵	۲	۷٦/٦٦٦	۸۰/۰۰۰
	بررسی حاضر (شمول ویژگی)	۱۲	۸۳/۳۳۳	۹۷/۹۷۳
	بررسی حاضر (خروج ویژگی)	۹	۷٤/۵۱۰	۹۱/۲۱٦
	۳۳	۱۸	۹۵/۱۰۰	۹۵/٦۲۱
	۳٤	۱۳	۹٤/٦۳۰	۹۲/٦۷٦
پارکینسون	۳۵	۱٤	۸۰/۲۱۰	۸۲/۰۰۰
	بررسی حاضر (شمول ویژگی)	۲۳	۹۹/٦۱۲	۹۹/۷۱۱
	بررسی حاضر (خروج ویژگی)	۷	۹۹/۲۱۷	۹۹/۵۹٤

با این وجود، در میان روش‌های موجود، الگوریتم پیشنهادی خروج ویژگی برای تشخیص بیماری پارکینسون با حفظ کمترین تعداد ویژگی به عملکرد بهینه دست یافته است، زیرا کدگذاری اَسکی تعریف شده برای دادگان توصیفی بیشینه استقلال ویژگی‌ها را نتیجه داده است. از سوی دیگر، هر دو روش پیشنهادی از تفسیرپذیری و پایداری اولویت‌های ویژگی برخوردارند، به طوری‌که همبستگی بالای ویژگی‌های انتخابی و کلاس هدف اغلب نتایج عملکردی بهینه یکسانی را تولید کرده، که حاکی از قدرت تکرارپذیری الگوریتم انتخاب ویژگی است، در حالی‌که ناپایداری روش‌های مرسوم اطمینان ویژگی‌های انتخابی را کاهش می‌دهد.[۳۸]

به‌طور خلاصه، هدف از کاهش ابعاد و انتخاب ویژگی در این پژوهش، بهبود دقت دسته‌بندی و دیگر شاخص‌های ارزیابی به کمک رویکردهای شمول ویژگی (افزودن ویژگی‌های ارتقادهنده دقت دسته‌بند) و خروج ویژگی (حذف ویژگی‌های دارای افزونگی) می‌باشد. ایده اصلی، ارایه روشی برای انتخاب ویژگی با ترکیبی از روش‌های فیلتر و بسته‌بندی است. در روش فیلتر مستقل از کلاس، پس از خوشه‌بندی داده‌ها با بَچ‌های کوچک تصادفی، از اطلاعات متقابل بین یک ویژگی و کلاس هدف استفاده شده و پس از رتبه‌بندی ویژگی‌ها بر حسب مقادیر به‌دست آمده از اطلاعات متقابل بهنجارشده، در رویکرد بسته‌بندی، دقت دسته‌بندی به‌عنوان معیاری برای انتخاب ویژگی ارزیابی شده است

اعتبارسنجی به‌روش زیرنمونه‌برداری تصادفی برای دسته‌بندی نمونه‌های آموزش و آزمایش منتج با استفاده از ماشین بردار پشتیبان خطی انجام شده است.

نتایج به‌دست آمده نشان‌دهنده قدرت رویکردهای مورد استفاده در بهبود دقت دسته‌بندی و دیگر شاخص‌های ارزیابی به‌ویژه در مجموعه دادگان با ابعاد بسیار بالای ویژگی مانند مجموعه داده توصیف مولکول‌های پروتیینی دارای زنجیره‌های اسید آمینه می‌باشند. از مهمترین کاربرد الگوریتم‌های پیشنهادی بازشناسی الگو در تشخیص پزشکی می‌توان به اختلالاتی اشاره کرد که پایه بیولوژیکی دارند، اما تشخیص آن‌ها منحصرا رفتاری است (به‌عنوان مثال، تشخیص اوتیسم در سنین پایین یا اختلالات روان پزشکی).[۳۹]

در این راستا، الگوریتم‌های پیشنهادی رویکرد امیدوارکننده‌ای برای کشف نشانگر زیستی بر پایه روش‌های بازشناسی الگو پیاده‌سازی شده در داده‌های پزشکی ارایه می‌دهند. قابلیت تعمیم‌پذیری به سایر دسته‌بندها و تعیین خودکار تعداد ویژگی‌های منتخب در طول فرآیند انتخاب ویژگی، این رویکرد را در بسیاری از کاربردهای پردازش تصویر، بیوانفورماتیک، داده‌کاوی و یادگیری ماشین انعطاف‌پذیر می‌سازد.

References

1. Chizi B, Maimon O. Dimension reduction and feature selection. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. 1st Ed. Springer, Boston, MA; 2010. p. 83-100.
2. Hammami M, Bechikh S, Hung C-C, Said LB. A multi-objective hybrid filter-wrapper evolutionary approach for feature selection. Memet Comput 2018;11:193-208.
3. Jie C, Jiawei L, Shulin W, Sheng Y. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70-9.
4. Schober P, Boer C, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesth Analg 2018;126(5):1763-8.
5. Xu H, Deng Y. Dependent evidence combination based on Shearman coefficient and Pearson coefficient. IEEE Access 2017;6:11634-40.
6. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 2020;21(6):1-13.
7. Gao W, Hu L, Zhang P. Class-specific mutual information variation for feature selection. Pattern Recognit 2018;79:328-39.
8. Ghosh M, Guha R, Sarkar R, Abraham A. A wrapper-filter feature selection technique based on ant colony optimization. Neural Comput Appl 2020;32(12):7839-57.
9. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. Artif Intell Rev 2019;53:907-48.
10. Venkatesh B, Anuradha J. A review of feature selection and its methods. Cybern Inf technol 2019;19(1):1-26.
11. Aggrawal R, Pal S. Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science 2020;1(344):1-16.
12. Shancheng J, Kwai-Sang C, Long W, Gang Q, Kwok L T. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. Expert Syst Appl 2017;82:216-30.
13. Venkatesh B, Anuradha J. A hybrid feature selection approach for handling a high-dimensional data. In: Saini HS, Sayal R, Govardhan A, Buyya R, editors. Innovations in computer science and engineering. 1st Ed. Springer, Singapore; 2019. p. 365-73.
14. Siddiqi UF, Sait SM, Kaynak O. Genetic algorithm for the mutual information-based feature selection in univariate time series data. IEEE Access 2020;8:9597-609.
15. Lim H, Kim DW. Pairwise dependence-based unsupervised feature selection. Pattern Recognit 2021;111:107663.
16. Bhadra T, Bandyopadhyay S. Supervised feature selection using integration of densest subgraph finding with floating forward–backward search. Inf Sci 2021;566:1-18.
17. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature selection methods and genomic big data: a systematic review. J Big Data 2019;6(79):1-24.
18. Lin YS, Jiang JY, Lee SJ. A similarity measure for text classification and clustering. IEEE Trans Knowl Data Eng 2014;26(7):1575-90.
19. Mittal H, Saraswat M. A new fuzzy cluster validity index for hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids. IEEE Trans Fuzzy Syst 2020;29(11):3249-58.
20. Xu Q, Zhang Q, Liu J, Luo B. Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Syst Appl 2020;151:113367.
21. Akhanli SE, Hennig C. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Stat Comput 2020;30:1523-44.
22. Peterson AD, Ghosh AP, Maitra R. Merging k-means with hierarchical clustering for identifying general-shaped groups. Stat 2018;7(1):e172.
23. Frank A. UCI Machine Learning Repository [Internet]. 2010. Available from: http://archive.ics.uci.edu/ml
24. Peng K, Leung VCM, Huang Q. Clustering approach based on mini batch k-means for intrusion detection system over big data. IEEE Access 2018;6:11897-906.
25. Thejas GS, Joshi SR, Iyengar SS, Sunitha NR, Badrinath P. Mini-batch normalized mutual information: A hybrid feature selection method. IEEE Access 2019;7:116875-85.
26. Fu Z, Robles-Kelly A, Zhou J. Mixing linear SVMs for nonlinear classification. IEEE Trans Neural Netw 2010;21(12):1963-75.
27. Thejas GS, Garg R, Iyengar SS, Sunitha NR, Badrinath P, Chennupati S. Metric and accuracy ranked feature inclusion: Hybrids of filter and wrapper feature selection approaches. IEEE Access 2021;9:128687-701.
28. Wu M, Zhong X, Peng Q, Xu M, Huang S, Yuan J, Ma J, Tan T. Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. *Eur J Radiol* 2019;114:175-84.
29. Zhou HF, Zhang JW, Zhou YQ, Guo XJ, Ma YM. A feature selection algorithm of decision tree based on feature weight. *Expert Syst Appl* 2021;164:113842.
30. Memon MH, Li JP, Haq AU, Memon MH, Zhou W. Breast cancer detection in the IOT health environment using modified recursive feature selection. *Wirel Commun Mob Comput* 2019;2019:5176705.
31. Zawba HM, Hassanien AE, Emary E, Yamany W, PARV B. Hybrid flower pollination algorithm with rough sets for feature selection. *11th Int Comput Eng Conf (ICENCO)*, Cario, Egypt; 2015:278-83.
32. Zhang Y, Yang A, Xiong C, Wang T, Zhang Z. Feature selection using data envelopment analysis. *Knowl Based Syst* 2014;64:70-80.
33. Qasim HM, Ata O, Ansari MA, Alomary MN, Alghamdi S, Almehmadi M. Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem. *Medicina* 2021;57:1217.
34. Chen H-L, Wang G, Ma C, Cai ZN, Liu WB, Wang SJ. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing* 2016;184:131-44.
35. Nahar N, Ara F, Neloy MAI, Biswas A, Hossain MS, Andersson K. Feature selection based machine learning to improve prediction of Parkinson disease. *In:* Mahmud M, Kaiser MS, Vassanelli S, Dai Q, Zhong N, editors. *Brain Informatics. BI 2021*. Lecture Notes in Computer Science, Springer, Cham. 2021;12960:496-508.
36. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform* 2019;90:103089.

37. Polat K. Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets. *Neural Comput Appl* 2018;30:987-1013.
38. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. *J. King Saud Univ. - Comput. Inf. Sci.* 2022;34(4):1060-73.
39. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* 2015:57;328-49.

TUMJ Tehran University Medical Journal, October 2022; Vol. 80, No. 7: 546-562 Original Article

Modeling and design of a diagnostic and screening algorithm based on hybrid feature selection-enabled linear support vector machine classification

Tara Ghafouri Ph.D. Student
Negin Manavizadeh, Ph.D.*

Department of Electrical and Electronic Engineering, Nanostructured-Electronic Devices Laboratory, Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran.

*Corresponding author: Department of Electrical and Electronic Engineering, Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran.
Tel: +98-21-84062207
E-mail: manavizadeh@kntu.ac.ir

Abstract

Received: 21 Jul. 2022 Revised: 28 Jul. 2022 Accepted: 16 Sep. 2022 Available online: 23 Sep. 2022

Background: In the current study, a hybrid feature selection approach involving filter and wrapper methods is applied to some bioscience databases with various records, attributes and classes; hence, this strategy enjoys the advantages of both methods such as fast execution, generality, and accuracy. The purpose is diagnosing of the disease status and estimating of the patient survival.

Methods: Feature selection algorithms have been modeled in Matlab R2021a during April and May 2022 in the framework of statistical pattern recognition. First, the features are ranked based on normalized mutual information, as a metric of relevance and redundancy of features, and accordingly, an optimum feature subset with the highest accuracy of classification is selected. Two feature selection algorithms, *i.e.*, inclusion of features enhancing the classification accuracy and exclusion of irrelevant features are applied to the interest datasets, subsequent to the mini-batch *k*-means clustering of records.

Results: At the end of the execution of both feature selection methods, evaluation metrics including accuracy, precision, recall, and F1 score are measured and compared. Both proposed feature selection approaches for the molecular biology, hepatitis C virus (HCV), and *E. coli* bacteria datasets result in the precision and recall scores more than 98 percent, meaning that there are few false positives and false negatives in the linear support vector machine (LSVM) classification. Regarding the HCV dataset, selection of nine relevant features among the thirteen present ones using the feature exclusion method yields the classification accuracy and F1 score of 98.92 percent and 99.02 percent, respectively. The feature inclusion approach also results in an accuracy of 98.78 percent with a slight discrepancy.

Conclusion: The results reveal superior strength of the feature selection methods used here for life science datasets with higher-order features such as protein/gene expression database. The potentials to generalize to other classifiers and automatically specify the optimal number of features during the feature selection procedure make these approaches flexible in many data mining applications for the life sciences.

Keywords: hybrid feature selection, life science datasets, linear support vector machine (LSVM), mini-batch k-means clustering, normalized mutual information.