25 th International Conference on
Information Technology, Computer and Telecommunication
Event Place: Tbilisi,Georgia
www.itctconf.ir
بیست و پنجمین کنفرانس بین المللی
فناوری اطلاعات، کامپیوتر و مخابرات | گرجستان
25th International Conference on Information Technology,Computer and Telecommunication
PUBLISH IN JOURNALS — INTERNATIONAL CERTIFICATION

یک مرور جامع بر روی روش های انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری

سعید برشنده*

گروه کامپیوتر، دانشکده فنی و مهندسی، موسسه آموزش عالی آفاق، ارومیه، ایران

فرناز صمد زاد

گروه کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

معصومه قاسمی

گروه کامپیوتر، دانشگاه پیام نور، تهران، ایران

ملیکا بزمانی

گروه کامپیوتر، دانشگاه پیام نور، تهران، ایران

چکیده

الگوریتم های یادگیری ماشینی جایگاه ویژه در کاربردهای دنیای واقعی امروزی دارند. این الگوریتم ها قادر هستند محاسبات پیچیده و زمانگیر را در زمان مناسب انجام داده و بسته به نوع مساله پاسخ های بهینه را تولید کنند. این الگوریتم ها در شاخه های مختلفی از علوم از جمله تشخیص بیماری، تخمین هزینه، طبقه بندی، تشخیص نفوذ، پردازش تصاویر و بسیاری دیگر کاربرد دارند. با اینحال، کارایی این الگوریتم ها وابستگی زیادی به داده های آموزشی داده شده به آنها دارد. همچنین، پیشرفت تکنولوژی و ابزارهای جمع آوری اطلاعات منجر به ظهور مجموعه داده های جدید با ابعاد بزرگ شده است که کارایی الگوریتم های یادگیری ماشینی را تحت شعاع قرار می دهد. با افزایش ابعاد مجموعه داده ها، زمان آموزش مدل های یادگیری افزایش یافته و به دلیل وجود ویژگی های غیر مرتبط و زائد در مجموعه داده ها، دقت آنها کاهش می یابد. یکی از تکنیک های پیش پردازش داده ها در الگوریتم های یادگیری ماشینی انتخاب ویژگی است که به واسطه آن ویژگی های غیر مرتبط و زائد از مجموعه داده حذف می شود. با حذف ویژگی های غیر مرتبط، علاوه بر کاهش زمان یادگیری مدل ها، دقت مدل ها نیز افزایش می یابد زیرا که مدل بر روی ویژگی ها مهم و تاثیر گذار تمرکز می کند. در سال های گذشته روش های مختلفی برای انتخاب ویژگی ارائه شده است که از مهمترین آنها به روش های مبتنی بر الگوریتم های فراابتکاری اشاره کرد. این روش ها با در نظر گرفتن مساله مورد نظر و تابع هدف آن ویژگی های تاثیر گذار مجموعه داده را مشخص می کنند. از اینرو، این مقاله جدیدترین الگوریتم های انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری را مورد مطالعه قرار داده و در کنار خلاصه سازی آنها را از جنبه های مختلف مقایسه می کند.

واژگان کلیدی: یادگیری ماشین، انتخاب ویژگی، بهینه سازی، الگوریتم های فراابتکاری

۱- مقدمه

انتخاب ویژگی یکی از روش های مهم در الگوریتم های یادگیری ماشین است که اساس آن انتخاب مجموعه ویژگی های مرتبط و مهم از مجموعه داده‌ها است. در دنیای واقعی، با توجه به حجم بالای داده ها و تنوع ویژگی ها، انتخاب ویژگی های مهم می تواند در کاهش پیچیدگی مدل و افزایش دقت پیش بینی آنها مفید و کارزار باشد. همچنین انتخاب ویژگی می‌تواند با حذف ویژگی های غیر ضروری یا زائد، ازبیش‌برازش جلوگیری کند زیرا در این حالت مدل می‌تواند الگو های واقعی را بهتر شناسایی کند. به عبارتی دیگر می توان چنین بیان کرد که انتخاب ویژگی کارایی الگوریتم های یادگیری ماشین را افزایش داده و منجر به تفسیر پذیری بهتر نتایج می‌شود [۱].

افزایش حجم داده ها در حوزه یادگیری ماشین و داده کاوی چالش های زیادی را به ارمغان آورده است. یکی از اصلی ترین مشکلات موجود در این زمینه، نیاز به منابع محاسباتی بیشتر است؛ زیرا که پردازش و تحلیل داده های بزرگ زمان بر و هزینه بر است. با افزایش حجم داده ها، خطر نویز و وجود اطلاعات غیر مفید نیز افزایش می‌یابد که می‌تواند دقت مدل‌ها را کاهش دهد. علاوه بر این، مدیریت و ذخیره سازی داده های کلان نیازمند زیر ساخت های پیشرفته ای است که ممکن است برای بسیاری از سازمان ها قابل دسترس نباشد. در این راستا، انتخاب ویژگی به عنوان یک راهکار مهم مطرح می‌شود که شامل شناسایی و انتخاب ویژگی های کلیدی از میان مجموعه ای بزرگ از داده‌هاست تا مدل های یادگیری ماشین بتوانند با کارایی بیشتری عمل کنند [۲].

برای مقابله با داده های حجیم امروزی، روش‌های متعددی ارائه شده است که در میان آن ها روش های انتخاب ویژگی از جایگاه اهمیت به ویژه ای برخوردار است. این روش ها شامل تکنیک های مبتنی بر filter، Embedded وWrapper هستند. که هر کدام به نوعی به کاهش ابعاد داده و بالا بردن دقت مدل می‌پردازند. در روش های مبتنی بر filter با استفاده از معیارهای آماری مانند همبستگی یا اطلاعات متقابل، ویژگی هایی که برای محاسبات ضروری نیستندشناسایی و حذف می‌شوند. در روش های Wrapper، با استفاده از یک مدل یادگیری ماشین برای ارزیابی ترکیب های مختلف ویژگی ها، بهترین زیر مجموعه انتخاب می شود. روش های Embedded نیز به طور هم زمان فرآیند انتخاب ویژگی و آموزش مدل را انجام می‌دهند که در نهایت منجر به کارایی بهتر می‌شود و مقدار زمان محاسباتی را کمتر می‌کند [۳]. این رویکرد ها در کنار الگوریتم های فراابتکاری مانند الگوریتم ژنتیک و رویکردهای اکتشافی مانند جستجوی محلی، ابزار های قدرتمندی را برای مدیریت داده های بزرگ و پیچیده فراهم می‌آورند. همانطور که پیش تر بیان گردید انتخاب ویژگی فرآیندی است که در آن ویژگی های غیر ضروری یا کم اهمیت از مجموعه داده ها حذف می‌شوند تا مدل های یادگیری ماشین بهینه تر و کارآمد تری ایجاد شوند. این فرآیند ابعاد داده ها را کاهش داده و منجر به افزایش دقت مدل و کاهش زمان آموزش می شود. همچنین، انتخاب ویژگی از بیش برازش یا کم برازش مدل ها جلوگیری می کند. انتخاب ویژگی همچنین می‌تواند باعث درک بهتر داده ها و شناسایی عوامل کلیدی مؤثر بر نتایج شود. با اینحال، انتخاب ویژگی با چالش هایی نیز مواجه است. یکی از چالش های اساسی آن این است که ممکن است یک سری از ویژگی های مهم که به تنهایی تأثیر آنچنانی ندارند، در ترکیب با دیگر ویژگی ها اطلاعات مفیدی ارائه دهند. در این حالت، انتخاب نادرست ویژگی ها می‌تواند منجر به کاهش عملکرد مدل شود؛ پس نحوه تشخیص ویژگی های مهم، از اهمیت بالایی برخوردار است. از چالش های دیگر می توان به تعیین معیارهای مناسب برای ارزیابی اهمیت ویژگی ها و مدیریت تعاملات پیچیده بین ویژگی ها اشاره کرد. از این رو، انتخاب صحیح و مؤثر ویژگی ها نیازمند تجربه و دانش عمیق در زمینه داده کاوی و یادگیری ماشین است [۴].

در روش های انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری که جزئی از روش های مبتنی بر wrapper هستند، بهترین زیر مجموعه از ویژگی‌ها در یک مجموعه داده توسط تکنیک های خاص این الگوریتم های شناسایی و انتخاب می شوند.. الگوریتم های فراابتکاری مانند الگوریتم ژنتیک و جستجوی ازدحام ذرات توانایی بلقوه ای در جست وجوی فضای بزرگ و پیچیده دارند که می

تواند منجر به یافتن بهترین ترکیب ویژگی ها شود. این الگوریتم ها معمولاً از فرآیندهای طبیعی و رفتارهای جمعی الهام گرفته شده اند. این روش ها معمولاً کیفیت زیر مجموعه های مختلف را با استفاده از یک معیار های خاصی مانند دقت مدل یا معیار های اطلاعاتی ارزیابی می‌کنند. یکی از مزایای اصلی این رویکردها انعطاف پذیری آن ها در مواجهه با مسائلی است که پیچیدگی محاسباتی آنها غیر خطی بوده و روش های سنتی انتخاب ویژگی قادر به حل آنها نمی باشند [۵].

در این پژوهش، برجسته ترین روش های انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری که در سال های اخیر معرفی شده اند مورد مطالعه قرار گرفته و یک مرور کلی بر روی آنها انجام می شود. در این راستا، الگوریتم های فراابتکاری بکار رفته در روش ها، توابع هدف بکار گرفته شده، مساله و مجموعه داده ها، معیارهای ارزیابی استفاده شده به همراه نقاط ضعف و قوت هر کدام مورد بحث قرار می گیرند.

۲- پیش زمینه

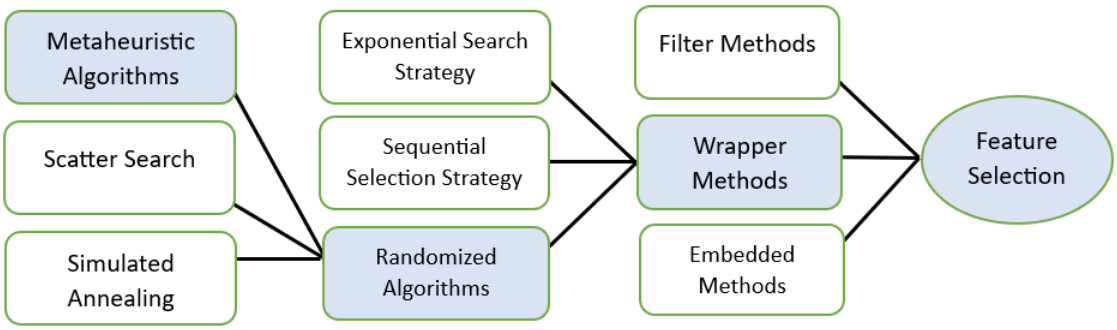
انتخاب ویژگی یک فرآیند حیاتی در پیش پردازش داده ها است که هدف آن استخراج زیر مجموعه ای از ویژگی های مرتبط و ضروری از مجموعه داده ها برای بهبود کارایی مدل های یادگیری ماشین است. این فرآیند با کاهش ابعاد داده ها، پیچیدگی مسئله را کم کرده و فرآیند یادگیری را تسهیل می کند. الگوریتم های فراابتکاری نیز به عنوان ابزارهایی قدرتمند در بهینه سازی فرآیند انتخاب ویژگی ها نقش مهمی ایفا می کنند زیرا به یافتن بهترین یا نزدیک به بهترین مجموعه ویژگی ها کمک می کنند. در این بخش، مفاهیم پایه ای انتخاب ویژگی به طور جامع بررسی می شوند.

۲-۱- انتخاب ویژگی

انتخاب ویژگی به مواجهه با ویژگی هایی که ضرورت و ارتباط لازم با هدف مساله را ندارند پرداخته و بهترین ویژگی ها را از مجموعه داده انتخاب می کند [۶]. انتخاب ویژگی یکی از چالش برانگیز ترین مسائل در یادگیری ماشین است که کاربردهای متنوعی در حوزه های مختلف دارد. از جمله مهمترین این کاربردها می توان به کاربردهای زیست‌پزشکی برای یافتن بهترین ژن از میان ژن‌های کاندید [۷]، متن کاوی برای یافتن کلمات یا عبارات خاص و مشخص [۸]، پردازش تصویر برای انتخاب بهترین محتوای بصری مانند پیکسل‌ها یا رنگ [۹] و غیره اشاره کرد [۱۰].

به طور کلی، روش های انتخاب ویژگی مختلفی برای به یافتن زیرمجموعه‌بهینه از ویژگی ها توسعه یافته اند. این روش ها را می توان بسته به اینکه مجموعه آموزشی دارای برچسب می باشد یا نه در سه دسته نظارت شده، نظارت نشده و نیمه نظارت شده سازمان دهی کرد. روش های انتخاب ویژگی‌نظارت شده نیز به سه دسته روش های مبتنی بر فیلتر، روش های Embedded و روش های مبتنی بر wrapper تقسیم بندی کرد. روش های مبتنی بر فیلتر، انتخاب ویژگیرا از فرآیند یادگیری جدا می کنند تا از هرگونه تداخل الگوریتم یادگیری با الگوریتم انتخاب ویژگی جلوگیری شود. این روش ها معمولاً بر روی ویژگی های کلی مجموعه داده تمرکز می کنند. از سوی دیگر، روش های مبتنی بر wrapper برای ارزیابی بهینگی ویژگی های انتخاب شده، از دقت مدل بکار گرفته شده استفاده می کنند [۱۱]. این روش ها اغلب برای مجموعه داده های با ویژگی های زیاد، پزمانگیر هستند. معمولاً این روش ها شامل الگوریتم های طبقه بندی بوده و با مدل های یادگیرنده تعامل دارند. این روش ها معمولاً نتایج بهتری نسبت به روش مبتنی بر فیلتر ارائه می دهد، اما کندتر بوده و از نظر محاسباتی پرهزینه می باشد زیرا به نتیجه مدل وابسته می باشند. روش های Embedded ترکیبی از دو روش قبلی می با شند. در این روش ها، انتخاب ویژگی در فرآیند آموزش گنجانده شده و نتیجه الگوریتم های یادگیری نیز لحاظ می گردد. این روش ها می توانند از جنبه های مختلفی کارآمدتر باشد، زیرا با جلوگیری از آموزش مجدد پیش بینی‌کننده برای همه زیرمجموعه ها،سریع‌تر به راه‌حل می رسند [۱۲].

در این پژوهش، روش های انتخاب ویژگی که براساس الگوریتم های فراابتکاری می باشند مورد مطالعه قرار گرفته اند. همانطوری که در شکل ۱ نشان داده شده است، این روش ها زیر مجموعه ای از روش های مبتنی بر wrapper می باشند. از اینرو که این روش ها دقت نهایی مدل ها را مد نظر قرار می دهند، زیر مجموعه های مناسب تری را برای مساله مورد نظر انتخاب کرده و به نتایج بهتری دست پیدا می کنند. از اینرو، این روش ها کاربردهای گسترده ای در زمینه های مختلف و الگوریتم های گوناگون دارند.

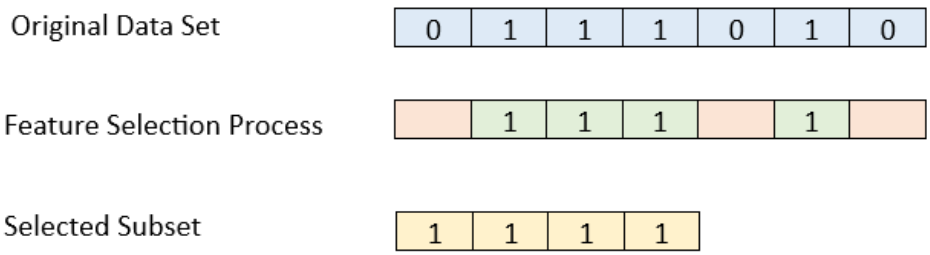


شکل ۱ دسته بندی روش های انتخاب ویژگی

در یک رویکرد ممکن، فرآیند انتخاب ویژگی در روش های مبتنی بر الگوریتم های فراابتکاری به صورت زیر تعریف می شود [۱]:

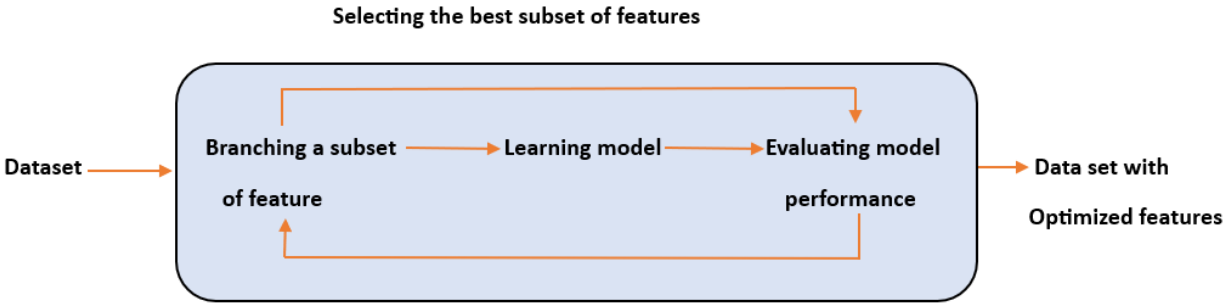
۱. یک بردار به اندازه تعداد ویژگی های مجموعه داده تعریف می شود
۲. خانه های بردار با مقادیر صفر و یک مقداردهی می شوند
۳. ویژگی هایی که مقدار خانه های متناظر آنها در بردار برابر صفر هستند، از مجموعه داده حذف می شوند
٤. ویژگی هایی که مقدار خانه های متناظر آنها در برابر برابر یک هستند، ویژگی های انتخاب شده هستند

مراحل انتخاب ویژگی فوق به صورت بصری در شکل ۲ نشان داده شده است.



شکل ۲ نحوه انتخاب ویژگی در روش های مبتنی بر الگوریتم های فراابتکاری

بعد از اینکه ویژگی های انتخاب شده مجموعه داده مشخص شدند، مدل یادگیری ماشین با مجموعه داده با ویژگی های منتخب آموزش داده می شود. سپس، کارایی مدل آموزش دیده شده با ویژگی های منتخب توسط معیارهای از پیش تعریف شده مانند دقت مدل ارزیابی می شود. در مرحله بعدی، ویژگی های انتخاب شده که توسط بردار مربوطه مشخص می شود، توسط الگوریتم های فراابتکاری تغییر داده می شود. مدل یادگیری مجدداً با ویژگی های انتخاب شده جدید آموزش داده شده و ارزیابی می شود. این مراحل به تعداد مشخصی تکرار شده و در نهایت بهینه ترین زیر مجموعه از ویژگی ها که منجر به بهترین مقدار معیار ارزیابی می شود، به عنوان ویژگی های نهایی انتخاب می شوند. شکل ۳ این فرآیند را به صورت شماتیک ارائه می کند.



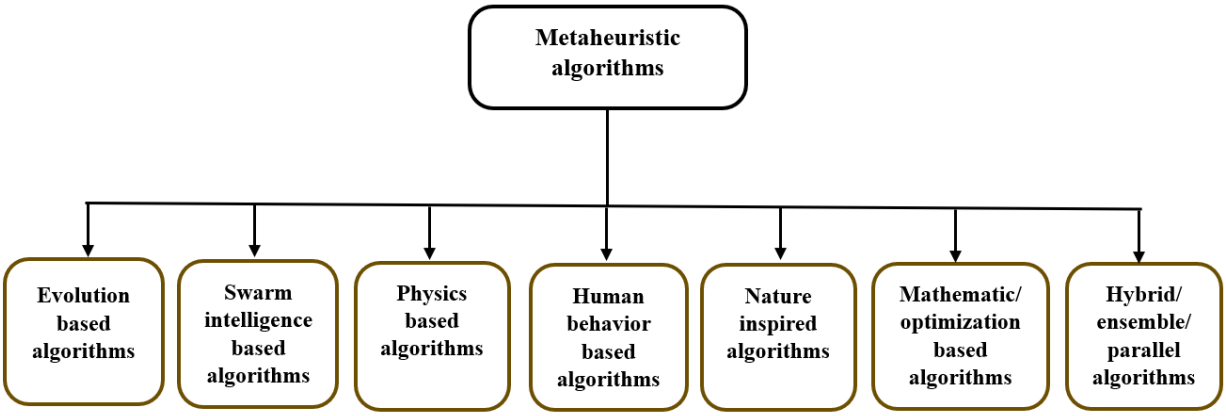
شکل ۳ فرآیند انتخاب ویژگی

۲-۲- الگوریتم‌های فراابتکاری

امروزه، پیچیدگی محاسباتی مسائل دنیای واقعی بطور قابل توجهی بالا می باشد به نحوی که روش های مرسوم قادر به حل آنها در زمان مناسب نمی باشند. به همین منظور، نیاز به توسعه تکنیک های جدید برای کاهش پیچیدگی محاسباتی و زمان اجرای الگوریتم ها بیش از پیش احساس می شود [۱۳]. در نتیجه، دسته جدید از الگوریتم ها به نام الگوریتم های فراابتکاری توسعه یافته اند که قادر هستند مسائل محاسباتی با پیچیدگی نمایی را در مدت زمان معقول با دقت مناسب حل نمایند. رفتار الگوریتم های فراابتکاری، تصادفی است. بدین معنی که بدون نیاز به اطلاعات اضافی در مورد فضای مساله، فرآیند بهینه سازی را با توزیع تصادفی عامل های جستجو در آغاز می‌کنند. سپس، با بروزرسانی این عامل ها در یک تکرار مشخص، مقدار تابع هدف مشخص شده را بهینه می کنند [۱۴]. همانطور که بیان گردید، الگوریتم های فراابتکاری به دنبال به حداقل رساندن یا به حداکثر رساندن ارزش یک تابع به نام تابع هدف هستند. قابلیت های الگوریتم های فراابتکاری در حل مسائل دنیای واقعی متفاوت است زیرا هر الگوریتم از روش های مختلفی برای حل مسائل و بهینه سازی تابع هدف استفاده می کنند. با اینحال، میزان تلاش الگوریتم برای کمینه یا بیشینه کردن مقدار این تابع توسط دو پارامتر تعداد تکرار و تعداد عامل های جستجو مشخص می گردد. هر چقدر تعداد تکرارها و عامل های جستجو بیشتر باشد، الگوریتم به نتایج بهتری دست می یابد ولی زمان اجرای آن نیز افزایش پیدا می کند [۱۵, ۱۶].

اکتشاف و بهره برداری دو عامل اصلی در الگوریتم های فراابتکاری می باشند. الگوریتم توسط قابلیت اکتشاف کل فضای جستجو را برای یافتن مناطق امیدوار کننده جستجو می کند. در حالیکه توسط قابلیت بهره برداری، مناطق امیدوارکننده یافته شده تاکنون را برای یافتن راه حل های بهینه کاوش می کند [۱۷]. الگوریتم هایی که توانایی اکتشاف بالایی دارند به سختی در بهینه های محلی گیر می کنند ولی سرعت همگرایی ضعیفی دارند. در مقابل، الگوریتم هایی که قابلیت بهره برداری خوبی دارند با سرعت بیشتری به سمت بهینه سراسری حرکت می کنند ولی مستعد گیر کردن در بهینه های محلی هستند. برای دستیابی به نتایج رضایت‌بخش، نیاز به برقراری تعادل دقیق بین قابلیت های اکتشاف و بهره برداری است که یکی از چالش های اساسی این الگوریتم ها می باشد [۱۸]. الگوریتم های فراابتکاری را می توان از لحاظ معیارهای گوناگونی دسته بندی کرد. یکی از دسته بندی های رایج این الگوریتم ها در شکل ۴ نشان داده شده است که براساس پدیده هایی می باشند که از آنها الهام گرفته شده اند. الگوریتم های مبتنی بر طبیعت از پدیده های طبیعی، گیاهان، رفتارهای حیوانات و غیره الهام گرفته شده اند. الگوریتم های مبتنی بر فیزیک الهام گرفته شده از قوانین فیزیک از جمله انتقال حرارت، نیروی گرانشی، حرکت ذرات و غیره هستند. الگوریتم Evolutionary الگوریتم های تکاملی هستند که از جمله شاخص ترین آنها می توان به الگوریتم تکامل تفاضلی اشاره کرد. الگوریتم های مبتنی بر ازدحام از رفتارهای جمعی، تعامل بین حیوانات و سایر جوامع هوشمند الهام گرفته شده اند. الگوریتم های مبتنی بر انسان نیز از رفتارهای انسانی، بدن، ارتباطات و استراتژی های انسانی الهام گرفته شده است. در دسته نهایی، الگوریتم های فراابتکاری قرار می گیرند که از معادلات و فرمول های

ریاضی بدست آمده اند. همچنین، در روش هایی نیز این پدیده ها و الگوریتم ها با هم ترکیب می شوند که دسته الگوریتم های فراابتکاری ترکیبی را تشکیل می دهند [۱۶].



شکل ۴ انواع الگوریتم های فراابتکاری

در ادامه مقاله از نمادهای مربوط به الگوریتم ها و استراتژی و یا معادل فارسی آنها استفاده می شود. برای هدایت بهتر خوانندگان، نمادها و فرم کامل آنها در جدول ۱ فراهم شده است. همچنین، در ادامه مجموعه داده های بکار گرفته شده در روش های مورد مطالعه مطرح می شوند. جزئیات کامل این مجموعه داده ها در جدول ۲ ارائه شده است. علاوه بر این، معیارهای ارزیابی استفاده شده در روش های انتخاب ویژگی نیز مورد مقایسه قرار می گیرند. از اینرو، نام معیارها به همراه فرمول ریاضی آنها در جدول ۳ بیان گردیده اند.

جدول ۱ نمادها و اختصارات

نماد	حالت کامل انگلیسی	معادل فارسی
ABC	Artificial Bee Colony	کلونی زنبورعسل مصنوعی
ACO	Ant Colony Optimization	بهینه سازی کلونی مورچه
AdaBoost	Adaptive Boosting	توقیت/ارتقا تطبیقی
ADASYN	Adaptive Synthetic Sampling	نمونه گیری مصنوعی تطبیقی
ALO	Ant Lion Optimizer	بهینه ساز شیر مورچه
AOA	Arithmetic Optimization Algorithm	الگوریتم بهینه سازی حسابی
ASO	Atom Search Optimization	بهینه سازی جستجوی اتم
BA	Bat Algorithm	الگوریتم خفاش
BHA	Black Hole Algorithm	الگوریتم سیاه چاله
BBO	Biogeography-Based Optimization	بهینه سازی مبتنی بر جغرافیای زیستی
CEE	Classification Error Rate	نرخ خطای طبقه بندی
CM	Chaotic Map	نگاشت آشوبناک
CMAES	Covariance Matrix Adaptation Evolution Strategy	استراتژی تکامل انطباق ماتریس کوواریانس
CSA	Crow Search Algorithm	الگوریتم جستجوی کلاغ
DA	Dragonfly Algorithm	الگوریتم سنجاقک
DE	Differential Evolution	تکامل تفاضلی
DT	Decision Tree	درخت تصمیم

EO	Equilibrium Optimizer	بهینه ساز تعادل
FA	Firefly Algorithm	الگوریتم کرم شب تاب

ادامه جدول ۱

نماد	حالت کامل انگلیسی	معادل فارسی
FF	Fitness Function	تابع برازندگی
FPA	Flower Pollination Algorithm	الگوریتم گرده افشانی گل
FR	Random Forest	جنگل تصادفی
FS	Feature Selection	انتخاب ویژگی
FSR	Feature Selection Rate	نرخ انتخاب ویژگی
GA	Genetic Algorithm	الگوریتم ژنتیک
GOA	Grasshopper Optimization Algorithm	الگوریتم بهینه سازی ملخ
GSA	Gravitational Search Algorithm	الگوریتم جستجوی گرانشی
GWO	Grey Wolf Optimizer	بهینه ساز گرگ خاکستری
HC	Hill Climbing strategy	استراتژی تپه نوردی
HGSO	Henry Gas Solubility Optimization	بهینه سازی حلالیت گاز هنری
HHO	Harris Hawks Optimization	بهینه سازی شاهین هاریس
HS	Harmony Search Algorithm	الگوریتم جستجوی هارمونی
ISA	Immune System Algorithm	الگوریتم سیستم ایمنی
KNN	K-Nearest Neighbor	K نزدیک ترین همسایه
LF	Lévy-Flight	پرواز Levy
LOA	Lion Optimization Algorithm	الگوریتم بهینه سازی شیر
MA	Mayfly Algorithm	الگوریتم شاپره
MBO	Monarch Butterfly Optimization	بهینه سازی شاه پروانه
MFO	Moth-Flame Optimization	بهینه سازی شعله-پروانه
MVO	Multi-Verse Optimizer	بهینه ساز چند وضهی
NB	Naïve Bayes	بیزین ساده
NSGA-II	Non-dominated Sorting Genetic Algorithm-II	الگوریتم ژنتیک چند هدفه
PSO	Particle Swarm Optimization	بهینه سازی ازدحام ذرات
QRL	Quasi-Reflective Learning	یادگیری شبه انعکاسی
RGA	Random Grouping Algorithm	الگوریتم گروه بندی تصادفی
RSA	Reptile Search Algorithm	الگوریتم جستجوی خزندگان
SA	Simulated Annealing	تبرید شبیه سازی شده
SBO	Spiral-Based Optimization	بهینه سازی مبتنی بر مارپیچ
SO	Snake Optimizer	بهینه ساز مار
SSA	Sparrow Search Algorithm	الگوریتم جستجوی گنجشک
SSA	Salp Swarm Algorithm	الگوریتم ازدحام سالپ ها
STD	Standard Deviation	انحراف معیار
SVM	Support Vector Machine	ماشین بردار پشتیبان
TF	Transfer Function	تابع انتقال
TSA	Tabu Search Algorithm	الگوریتم جستجوی ممنوعه
UCI	University of California, Irvine	مخزن UCI دانشگاه کالیفرنیا

WOA	Whale Optimization Algorithm	الگوریتم بهینه سازی وال
WSA	Wind-driven Swarm Algorithm	الگوریتم ازدحام باد محور
PNS	Probabilistic Neighborhood Search	جستجوی همسایگی احتمالی

جدول ۲ مجموعه داده ها و جزئیات آنها

#	نام	تعداد نمونه ها	تعداد ویژگی ها	تعداد کلاس ها	زمینه
۱	Diabetes	۷۶۸	۸	۲	Medical
۲	Heart	۳۰۴	۱۳	۲	Medical
۳	Glass	۲۱۴	۱۰	۶	Physical
۴	Dermatology	۳۶۶	۳۴	۶	Medical
۵	Breast cancer	۶۹۹	۹	۲	Biology
۶	Tic-tac-toe	۹۵۸	۹	۲	Game
۷	Exactly۱	۱۰۰۰	۱۳	۲	Biology
۸	Exactly۲	۱۰۰۰	۱۳	۲	Biology
۹	Heart	۲۷۰	۱۳	۲	Biology
۱۰	M-of-n	۱۰۰۰	۱۳	۲	Biology
۱۱	Wine	۱۷۸	۱۳	۳	Chemistry
۱۲	Congress	۴۳۵	۱۶	۲	Politics
۱۳	Vote	۳۰۰	۱۶	۲	Politics
۱۴	Zoo	۱۰۱	۱۶	۶	Artificial
۱۵	Lymphography	۱۴۸	۱۸	۲	Biology
۱۶	Spect	۲۶۷	۲۲	۲	Biology
۱۷	Breast	۵۶۹	۳۰	۲	Biology
۱۸	Ionosphere	۳۵۱	۳۴	۲	Electromagnetic
۱۹	Krvskp	۳۱۹۶	۳۶	۲	Game
۲۰	Waveform	۵۰۰۰	۴۰	۳	Physical
۲۱	Sonar	۲۰۸	۶۰	۲	Biology
۲۲	Penglung	۷۳	۳۲۵	۲	Biology
۲۳	Arrhythmia	۴۵۲	۲۷۹	۱۳	Life
۲۴	Clean۱	۴۷۶	۱۶۸	۲	Physical
۲۵	Colon	۶۲	۲۰۰۰	۲	Biology
۲۶	DNA	۳۱۸۶	۱۸۰	۳	Biology
۲۷	Eeg-eye-state	۱۴,۹۸۰	۱۴	۲	Life
۲۸	Fri_c۰_۱۰۰۰_۱۰	۱۰۰۰	۱۰	۲	Statistical
۲۹	Fri_c۱_۱۰۰۰_۱۰	۱۰۰۰	۱۰	۲	Statistical
۳۰	German	۱۰۰۰	۲۴	۲	Financial
۳۱	Kc۱	۲۱۰۹	۲۱	۲	N/A
۳۲	Leukemia	۷۲	۷۱۲۹	۲	Biology
۳۳	Madelon	۲۶۰۰	۵۰۰	۲	N/A
۳۴	Opt digits	۵۶۲۰	۶۴	۱۰	Computer
۳۵	Page blocks	۵۴۷۳	۱۰	۲	Computer
۳۶	Pen digits	۱۰۹۹۲	۱۶	۲	Handwriting
۳۷	Satellite	۶۴۳۵	۳۶	۶	Physical
۳۸	Segment	۲۳۱۰	۱۹	۷	N/A
۳۹	Semeion	۱۵۹۳	۲۵۶	۱۰	Computer
۴۰	Spam base	۶۴۰۱	۵۷	۲	Computer
۴۱	Wisconsin	۵۶۹	۳۰	۲	Biology

۴۳	Lung	۲۰۳	۳۳۱۲	۵	Biological
۴۴	Churn	۳۱۵۰	۱۶	۲	Telecom
۴۵	Hepatitis	۱۵۵	۱۹	۳	Biological

ادامه جدول ۲

#	نام	تعداد نمونه ها	تعداد ویژگی ها	تعداد کلاس ها	زمینه
۴۶	TOX ۱۷۱	۱۷۱	۵۷۴۸	۴	Biological
۴۷	Lymphoma	۹۶	۴۰۲۶	۹	Biological
۴۸	GLIOMA	۵۰	۴۴۳۴	۴	Biological
۴۹	ALLAML	۷۲	۷۱۲۹	۲	Biological
۵۰	Prostate GE	۱۰۲	۵۹۶۶	۲	Biological
۵۱	CLL_SUB_۱۱۱	۱۱۱	۱۱۳۴۰	۳	Biological
۵۲	nci۹	۶۰	۹۷۱۲	۹	Biological
۵۳	GLI_۸۵	۸۵	۲۲۲۸۳	۲	Biological
۵۴	SMK_CAN_۱۸۷	۱۸۷	۱۹۹۹۳	۲	Biological
۵۵	orlraws۱۰P	۱۰۰	۱۰۳۰۴	۱۰	Face image
۵۶	warpAR۱۰P	۱۳۰	۲۴۰۰	۱۰	Face image
۵۷	warpPIE۱۰P	۲۱۰	۲۴۲۰	۱۰	Face image
۵۸	Yale	۱۶۵	۱۰۲۴	۱۵	Face image
۵۹	Contraceptive	۱۴۷۳	۹	۳	Biological
۶۰	Vowel	۹۹۰	۱۳	۱۱	Biological
۶۱	Australian	۶۹۰	۱۴	۲	Financial
۶۲	SPECT	۱۶۰	۲۲	۲	Biology
۶۳	WDBC	۵۶۹	۳۱	۲	Biology
۶۴	Madelon	۴۴۰۰	۵۰۰	۲	Computer
۶۵	PD speech	۷۵۶	۷۵۴	۲	Medical
۶۶	CANE۹	۱۰۸۰	۸۵۶	۹	Game
۶۷	CNS	۶۰	۷۱۲۹	۲	Biology
۶۸	Iris	۱۵۰	۴	۵	Biology
۶۹	Gastrointestinal	۱۵۲	۶۹۸	۲	Medical
۷۰	Parkinson	۷۵۶	۷۵۴	۵	Medical
۷۱	Chemical Water	۱۷۸	۱۳	۳	Chemistry
۷۲	Chess	۳۱۹۶	۳۶	۲	Text
۷۳	dbworld	۶۴	۲۴۲	۲	Text
۷۴	Lung cancer	۳۲	۵۶	۳	Life
۷۵	SRBCT	۸۳	۲۳۰۸	۴	Microarray
۷۶	Prostate cancer	۱۰۲	۱۰۵۰۹	۲	Microarray
۷۷	Brain Tumor_۱	۹۰	۵۹۲۰	۵	Microarray
۷۸	۱۱_Tumors	۱۷۴	۱۲۵۳۳	۱۱	Microarray

۳- روش شناسی تحقیق

در این بخش از مقاله، برجسته ترین روش هایی که برای حل مساله انتخاب ویژگی از الگوریتم های فراابتکاری بهره گرفته اند تشریح می گردند. نویسندگان در این بخش تلاش کرده اند روش های جدیدی که در مجلات معتبر ارائه شده اند را برای مطالعه انتخاب نمایند. در ادامه، نویسندگان روش های ارائه شده را بررسی کرده و یک مرور جامع از هر کدام را ارائه داده اند. در مرور انجام شده،

الگوریتم یا الگوریتم های فراابتکاری بکار گرفته شده، توابع هدف، مجموعه داده های استفاده شده، نقاط قوت و ضعف، معیار ارزیابی کارایی، مدل یادگیری مورد استفاده و غیره از هر پژوهش استخراج و بیان شده است. برای ایجاد یک پیش زمینه ذهنی برای خوانندگان و ارائه یک مقایسه کلی، روش های مورد مطالعه در این پژوهش در جدول ۴ مقایسه شده اند.

جدول ۳ معیارهای ارزیابی و فرمول ریاضی آنها

#	نام معیار	فرمول ریاضی
۱	Accuracy	$A = \frac{TP + TN}{TP + TN + FP + FN}$
۲	Precision	$P = \frac{TP}{TP + FP}$
۳	Recall	$R = \frac{TP}{TP + FN}$
۴	F۱-Score	$F1 = ۲ \times \frac{Precision \times Recall}{Precision + Recall}$
۵	Chi-Square	$X^۲ = \sum \frac{(O - E)^۲}{E}$
۶	Mutual Information	$MI(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$
۷	Correlation Coefficient	$C = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^۲ \sum(y_i - \bar{y})^۲}}$
۸	Entropy	$E(x) = -\sum p(x) \log p(x)$
۹	Mean Squared Error (MSE)	$MSE = \frac{۱}{n}\sum_{i=۱}^{n}(y_i - \hat{y}_i)^۲$
۱۰	Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{۱}{n}\sum_{i=۱}^{n}(y_i - \hat{y}_i)^۲}$
۱۱	Mean Absolute Error (MAE)	$MAE = \frac{۱}{n}\sum_{i=۱}^{n}\lvert y_i - \hat{y}_i \rvert$
۱۲	Stability	$S = ۱ - \frac{۱}{n}\sum_{i=۱}^{n}\lvert s_i - s_j \rvert$

جدول ۴ مرور و مقایسه روش های انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری

طرح	بهینه ساز(ها)	مجموعه داده ها	مدل یادگیری	معیارهای ارزیابی	رقبا
[۹]	SSA	Breast cancer, Tic-tac-toe, Zoo, Wine, Spect, Sonar, Ionosphere, Heart, Congress, Krvskp, Waveform, Exactly۱, Exactly۲, M-of-N, Vote, Breast, Semeion, Clean ۱, Clean ۲, Lymphography, and Penghung	KNN	Fitness value Accuracy	SSA RGA GSA D-GSA BH-GSA C-GSA AR-GSA SSARM-SCASSA
[۱۰]	HHO	Camel, Jedit, Log٤j, and Xalan	KNN DT LDA	Sensitivity Specificity AUC	ESBHHO EVBHHO GSA GOA WOA BBA GA ALO
[۱۱]	RBEO-LS	Breast cancer, Breast, Congress, Exactly۱, Exactly۲, Heart, Ionosphere, lymphography, M-of-n, Penglung, Sonar, Spect, Tic-tac-toe, Vote, Wine, and Zoo	KNN	Classification Accuracy Number of selected Features Fitness value	WOA-CM GOA-M WFACOFS ECWSA-٤ ABGWO S-SCALO V-SCALO BBO-SBO CISA
[۱۲]	Jaya JS	Diabetes, Contraceptive, Glass, Breast cancer, Vowel Australian, Zoo, SPECT, WDBC, Ionosphere, Sonar, Semeion, Madelon, PD speech, CANE۹, Colon, CNS, and Lung	NB	Best accuracy Best fitness Feature selection ratio Average accuracy	PSO CSA DFO GA GWO GOA
[۱۳]	Beta hill-climbing optimizer	Ic-tac-toe, Breast cancer, Heart, Exactly۱, Exactly۲, M-of-n, Wine, Congress, Vote, Zoo, Spect, Lymphography, Breast, Ionosphere, Krvskp, Waveform, Sonar, clean۱, semeion, Penglung, Colon, and Leukemia	KNN	Classification accuracy CPU time Number of selected features Fitness value	GOA GWO GSA BA SSA HGSA WOA DA GA

PSO

ادامه جدول ۴

طرح	بهینه ساز(ها)	مجموعه داده ها	مدل یادگیری	معیارهای ارزیابی	رقبا
[۱٤]	AOA	Glass, Hepatitis, Lymphography, Primary, Tumor, Soybean, Horse Colic, Ionosphere, Zoo, Musk ۱, Dermatology, SPECT Heart, Libras, Movement, ILPD, LSVT, CADI, TOX ۱۷۱, Leukemia, Lung, and Colon	KNN	Mean fitness values Standard deviation of fitness values Feature selection ratio	HLBDA DA ABC MVO PSO CSA COA CMAES LSHADE FFO AOA
[۱۵]	ChOA	Chess, dbworld, Lymphography, Lung cancer, Yale, SRBCT, Prostate cancer, Leukemia, Brain, Tumor	KNN NB RBF	Number of selected features Fitness value Classification accuracy Computational time	GA BA ACO FA FPA
[۱٦]	EO	Breast cancer, Tic-tac-toe, Wine, Heart, Exactly۱, Exactly۲, M-of-n, Zoo, Vote, Congress, Lymphography, Spect, Breast, Ionosphere, Krvskp, Waveform, Sonar, and glung	KNN	Classification accuracy Number of selected features Pair-wise Friedman test Nemenyi post-hoc test	SSDs-LAHC RTHS AβBSF HSGW RSGW ASGW GA PSO
[۱۷]	ASO	Glass, Hepatitis, Iris, Lymphography, Primary, Tumor, Soybea, Horse Colic, Ionosphere, Zoo, Wine, Breast Cancer, Musk, Arrhythmia, Dermatology, SPECT Hear, Seeds, LSVT, Gastrointestinal, Parkinson, and Leukemia	KNN	Best fitness value Mean fitness value Worst fitness value STD value Average accuracy feature selection ratio	ASO ALO SSA CSA FPA PSO
[۱۸]	WOA	TOX ۱۷۱, Leukemia, Lymphoma, Colon, GLIOMA, ALLAML, Prostate GE, CLL SUB ۱۱۱, nci۹, Lung, GLI ۸۵, SMK CAN ۱۸۷, orlraws۱۰P, warpAR۱۰P, warpPIE۱۰P, and Yale	KNN	fitness value Standard deviation of fitness values Accuracy Precision F-measure	JA MFO GNDO CSA SSA CSO

HDBPSO	Computational cost Wilcoxon test				

ادامه جدول ۴

رقبا	معیارهای ارزیابی	مدل یادگیری	مجموعه داده ها	بهینه ساز(ها)	طرح
PSO GWO MVO WOA SSA RSA SO	Classification accuracy Most optimal feature subset Fitness value Computation time	KNN	Breast cancer, Breast, Churn, Heart, Ionosphere, Krvskp, Sonar, SpectEW, Tic-tac-toe, Vote, Chemical, and Zoo	RSA SO	[۱۹]
SSA ABC PSO BA GWO WOA GOA SFO HHO BSA BASO HGSO	Fitness value Number of selected features Classification accuracy Wilcoxon's rank-sum	KNN SVM RF	Breast cancer, Breast, Congress, Exactly۱, Exactly۲, Heart, Ionosphere, Krvskp, Lymphography, M-of-n, Penglung, Sonar, SpectEW, Tic-tac-toe, Vote, Waveform, Wine, and Zoo	SSA	[۲۰]
PSO ACO ABC CDGAFS	Accuracy Number of selected features Execution time	KNN SVM Ada Boost	Spam Base, Sonar, Arrhythmia, Madelon, Isolet, and Colon	GA	[۲۱]
GA PSO ALO GSA DA SSA WOA GWOPSO ECWSA GOA WOASAT ALO	Accuracy Number of selected features	KNN	Breast cancer, Tic-tac-toe, Exactly۱, Exactly۲, Heart, M-of-n, Wine, Congress, Vote, Zoo, Lymphography, Spect, Breast, Ionosphere, Krvskp, Waveform, Sonar, and Penglung	MA HS	[۲۲]
MBO MBA-SA WOASAT GSA HGSA	Average fitness value Worst fitness value Best fitness values	KNN	Breast cancer, Breast, Congress, Credit, Exactly۱, Exactly۲, Derm۱, Derm۲, Heart, Ionosphere, Krvskp, LED, Lung, Lymphography, M-of-n,	MBO	[۲۳]

Feature Size Accuracy	Mushroom, Penglung, Sonar, Spect, Tic-tac-toe, Vote, Waveform, Wine, WQ, and Zoo

در اولین پژوهش مورد بررسی در این مقاله، روش ارائه شده توسط زیدکوویک و همکاران مورد مطالعه قرار گرفته است که در آن یک روش جدید انتخاب ویژگی براساس الگوریتم ازدحام سالپ ها[1] استفاده شده است [۱۹]. هدف اصلی زیدکوویک و همکاران در این پژوهش بهبود هرچه بیشتر فرآیند انتخاب ویژگی در یادگیری ماشین از طریق بهبود الگوریتم ازدحام سالپ ها بوده است. بنابراین آنها در این پژوهش، الگوریتم ازدحام سالپ های پایه را با افزودن یک مکانیزم اضافی بهبود داده و آن را با الگوریتم فراابتکاری سینوس کسینوس در حوزه هوش ازدحامی ترکیب نمودند. در این مقاله، الگوریتم بر روی ۲۱ مجموعه داده اعمال شده و در آن از طبقه بند k نزدیک ترین همسایه استفاده شده است. سپس، نتایج الگوریتم پیشنهادی با نتایج الگوریتم‌های SSA، RGA و نسخه های مختلف GSA مقایسه شده است. بر اساس یافته‌های تجربی و تحلیل‌های مقایسه‌ای دقیق با دیگر رویکردهای پیشرفته اخیر، الگوریتم پیشنهادی SSARM-SCA به عنوان یک بهینه‌ساز کارآمد عمل می‌کند که به طور قابل توجهی سرعت همگرایی و کیفیت نتایج را نسبت به SSA پایه و دیگر الگوریتم‌های پیشرفته بهبود می‌بخشد. علاوه بر این، نتایج به‌دست‌آمده نشان می‌دهند که روش پیشنهادی موفق شده است دقت طبقه‌بندی بهتری ارائه دهد و از تعداد کمتری از ویژگی‌ها استفاده کند، بنابراین راه‌حل مؤثرتری برای چالش انتخاب ویژگی ارائه می‌دهد.

در پژوهشی دیگر، طاهر و آرمانیک نسخه ارتقا داده شده و باینری از الگوریتم بهینه‌سازی شاهین را ارائه معرفی کرده اند که به‌طور ویژه برای انتخاب ویژگی در پیش‌بینی خطاهای نرم‌افزاری طراحی شده است [۲۰]. نویسندگان با معرفی یک الگوریتم چندجمعیتی باینری کارآمد، به چالش‌های مرتبط با داده‌های نامتعادل ابعاد بالای مجموعه داده‌های خطاهای نرم‌افزاری پرداخته‌اند. این روش شامل تقسیم جمعیت الگوریتم به زیرگروه‌های کوچکتر است که هر یک توسط یک رهبر هدایت می‌شوند تا اکتشاف را بهبود بخشیده و از همگرایی زودرس به سمت بهینه‌های محلی جلوگیری کنند. همچنین، یک فرآیند باینری‌سازی دو مرحله‌ای، با بهره‌گیری از توابع انتقال S-shaped و V-shaped ، الگوریتم پیوسته را برای مسئله انتخاب ویژگی باینری مناسب می‌سازد. تابع هدف بکار برده شده در روش پیشنهادی ترکیبی از معیارهای نرخ خطای طبقه‌بندی و تعداد ویژگی‌های انتخاب شده است که هدف آن به حداقل رساندن هر دو معیار می‌باشد. برای مقابله با داده‌های نامتعادل، از نمونه گیری مصنوعی تطبیقی (ADASYN) استفاده شده است. عملکرد الگوریتم با استفاده از سه معیار: sensitivity، specificity و مساحت زیر منحنی (AUC) در پانزده مجموعه داده واقعی پیش‌بینی خطاهای نرم‌افزاری ارزیابی شده است. از جمله نقاط قوت اصلی الگوریتم پیشنهادی می توان به رویکرد چندجمعیتی آن که قابلیت اکتشاف را افزایش داده و همگرایی زودرس را که در روش‌های تک‌جمعیتی رایج است کاهش می‌دهد. همچنین، استفاده از ADASYN به‌طور مؤثری مشکل عدم تعادل کلاس ها در مجموعه داده‌ها را برطرف کرده و منجر به بهبود دقت پیش‌بینی می‌شود. مقایسه‌های انجام‌شده با سایر الگوریتم‌های پیشرفته، عملکرد برتر این الگوریتم، به ویژه نسخه تابع انتقال V-shape را از نظر مساحت زیر منحنی و تعداد ویژگی‌های انتخاب شده نشان می‌دهد. به‌کارگیری چندین طبقه‌بند ارزیابی قوی‌تری از فرآیند انتخاب ویژگی ارائه می‌دهد. یکی از محدودیت‌های احتمالی این روش، وابستگی آن به تابع هدف مورد استفاده است که در آن خطای طبقه‌بندی و تعداد ویژگی‌ها با استفاده از وزن‌های تعریف‌شده توسط کاربر ترکیب شده اند. زیرا که تعادل بهینه بین این اهداف ممکن است در مجموعه داده‌های مختلف متفاوت باشد. همچنین، انتخابمقادیر بهینه این وزن‌ها می‌تواند بر نتایج تأثیرگذار باشد.. همچنین، رویکرد

[1] Salp Swarm Algorithm (SSA)

چندجمعیتی عملکرد را بهبود می‌بخشد ولی با اینحالممکن است در مقایسه با روش‌های تک‌جمعیتی سربار محاسباتی بیشتر داشته باشد.

علاوه بر این، وادفل و عبدالعزیز یک روش ترکیبی جدید با نام RBEO-LS برای حل مساله انتخاب ویژگی معرفی کرده اند [۲۱]. در روش پیشنهادی آنها، الگوریتم فراابتکاری EO توسط یک استراتژی جستجوی محلی و مکانیزم ReliefF بهبود داده شده است. روش RBEO-LS از دو مرحله تشکیل شده است. در مرحله اول، الگوریتم ReliefF به عنوان یک گام پیش‌پردازشی برای اختصاص وزن به ویژگی‌ها استفاده می‌شود که ارتباط آن‌ها با فرآیند طبقه‌بندی را تخمین می‌زند. در مرحله دوم، الگوریتم EO که باینری شده است به عنوان یک روش جستجوی محلی به کار می‌رود. به عبارتی دیگر، ویژگی‌ها در روش پیشنهادی بر اساس وزن‌های تخصیص داده شده به آنها رتبه‌بندی می‌شوند. همچنین، در روش RBEO-LS الگوریتم EO با یک استراتژی جستجوی محلی ادغام شده است تا عملکرد آن بهبود یابد. در روش پیشنهادی، ضریب همبستگی پیرسون و رتبه بندی ویژگی ها به منظور حذف ویژگی های تکراری و اضافه شدن ویژگی ها مرتبط بکار برده شده است. روش RBEO-LS با هدف دستیابی به دقت بالا با کوچک‌ترین زیرمجموعه از ویژگی‌ها، ویژگی‌های نامرتبط و ویژگی‌های تکراری کههمبستگی بالایی دارند را حذف می‌کند. برای رسیدن به این هدف، RBEO-LS از یک روش دو مرحله‌ای برای تعیین زیرمجموعه بهینه ویژگی‌ها استفاده می‌کند. ابتدا، از ReliefF به عنوان یک مرحله پیش‌پردازش برای تخمین همبستگی بین ویژگی‌ها و کلاس‌ها استفاده می‌کند. بدین صورت که به هر ویژگی یک وزن اختصاص داده می‌شود که اهمیت آن را منعکس می‌کند. در مرحله دوم، الگوریتم EO دودویی انتخاب ویژگی را انجام می دهد. در فاز مقداردهی اولیه روش پیشنهادی، وزن ویژگی‌ها به صورت نزولی مرتب می شود و ویژگی هایی که وزن آنها پایین تر است با احتمال بیشتری انتخاب می شوند.در مقابل، هر چه رتبه یک ویژگی بالاتر باشد، احتمال حذف آن از جمعیت اولیه بیشتر خواهد بود پس از مرحله مقداردهی اولیه، الگوریتم BEO از طریق یک فرآیند تکراری به دنبال کوچک‌ترین زیرمجموعه از ویژگی‌ها می گردد که عملکرد بالاتری را ارائه دهد. با این حال، رتبه‌بندی ReliefF هر ویژگی را به صورت مستقل در نظر می‌گیرد و ارتباطات ممکن بین ویژگی‌ها را که می‌توانند منجر به افزونگی شوند و دقت طبقه‌بندی را تغییر دهند، لحاظ نمی‌کند. برای رفع این مساله، یک استراتژی جستجوی محلی در روش پیشنهادی بکار برده شده که هم همبستگی بین ویژگی‌ها و هم وزن‌های آن‌ها را در نظر بگیرد. به واسطه این رویکرد، راه حل ها در طی فرآیند جستجو به سمت بهترین زیرمجموعه ویژگی‌ها هدایت می شوند. عملکرد الگوریتم پیشنهادی روی شانزده مجموعه داده UCI و ده مجموعه داده زیستی با ابعاد بالا ارزیابی شده است. مجموعه داده‌های UCI شامل تعداد زیادی نمونه با تعداد کم یا متوسطی از ویژگی‌ها هستند، در حالیکه مجموعه داده‌های زیستی تعداد زیادی ویژگی با تعداد کمی نمونه ارائه می‌دهند. در آزمایشات صورت گرفته، نتایج الگوریتم RBEO-LS با شش الگوریتم فراابتکاری مقایسه شده است. این نتایج به توانایی بالای روش RBEO-LSدر حل مسئله انتخاب ویژگی و توانایی آن در ایجاد تعادل بین خطای طبقه‌بندی و تعداد ویژگی‌های انتخاب‌شده اشاره دارد. از این نتایج می‌توان دریافت که روش پیشنهادی RBEO-LS نسبت به روش های مورد مقایسه عملکرد بهتری دارد.

در پژوهشی دیگر، چاودهاری و ساهو سه روش جدید با نام های BJaya-S، BJaya-V و BJaya-JS برای انتخاب ویژگی های از اصلی از مجموعه داده ها ارائه کردند که به ترتیب از توابع انتقال S-shape، توابع انتقال V-shape و شاخص شباهت ژاکارد استفاده می کنند [۲۲]. روش اصلی پیشنهادی (BJaya-JS) کاملاً در فضای باینری عمل کرده و نیازمند تبدیل فضای پیوسته به باینری از طریق توابع انتقال که در الگوریتم‌های مشابه رایج است، نیست. همچنین، روش های پیشنهادی از یک تکنیک جستجوی محلیبه نام تولید کننده راه‌حل جدید (NSG) به منظور بهبود قابلیت های اکتشاف و بهره‌برداری استفاده می کنند. این تکنیک جستجوی همسایگی از سه همسایه به جای یک همسایه برای انجام جستجوی کارآمد در فضای مساله استفاده می‌کند. عملکرد این الگوریتم بر روی ۱۸ مجموعه داده از مخزن UCI ارزیابی و با چندین روش انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری مانند BPSO، BCSA، BDFO، GA، BGWO، و BGOA مقایسه شده است. ارزیابی ها با استفاده از معیارهایی مانند میانگین دقت، مقادیر برازندگی، نسبت ویژگی های انتخاب شده و زمان محاسباتی انجام شده است. همچنین، در این پژوهش از آزمون‌های آماری ویلکاکسون و

فریدمن استفاده شده است. نتایج نشان می‌دهند که BJaya-JS به طور مداوم عملکرد بهتری نسبت به دو روش پیشنهادی دیگر والگوریتم‌های رقیب در مجموعه داده‌های مختلف دارد. این الگوریتم در اکثر موارد بهترین دقت و مقدار برازندگی را به دست آورده، نسبت ویژگی های انتخاب شده کمتری داشته و معمولاً به زمان محاسباتی کمتری نیاز دارد. یکی از نقاط قوت الگوریتم پیشنهادی ماهیت بدون پارامتر آن است که فرآیند بهینه‌سازی را ساده تر کرده و ریسک تنظیم نادرست پارامترها را کاهش می‌دهد. با این حال، ضعف آن این است که عملکرد آن به (عملکرد جستجوی همسایگی وابسته است. همچنین، اگرچه مقادیر بهینه پارامترها از طریق آزمون و خطا تعیین شده است، حساسیت به این پارامترها می‌تواند تعمیم‌پذیری آن به مسائل مختلف را محدود کند. ارزیابی کارایی نیز به طبقه‌بند Naive Bayes محدود شده است.

ال بتار و همکاران نیز در پژوهشی مشابه به بررسی تاثیر مکانیزم تپه نوردی[۲] در حل مسئله انتخاب ویژگی پرداخته اند [۲۳]. مکانیزم تپه نوردی، یک الگوریتم نوین مبتنی بر جستجوی محلی است که قادر به ارائه راه‌حل‌های مؤثر برای انواع مسائل بهینه‌سازی می‌باشد. برای استفاده از این روش در انتخاب ویژگی، الگوریتم باید به گونه‌ای تنظیم شود که با داده‌های باینری سازگار باشد، بنابراین در این پژوهش از توابع انتقال S شکل برای باینری سازی بهره گرفته شده است. همچنین در این پژوهش، تأثیر پارامترهای مکانیزم تپه نوردی بر نرخ همگرایی با در نظر گرفتن معیارهایی همچون دقت، تعداد ویژگی‌های انتخاب شده، مقادیر برازندگی و زمان محاسباتی مورد بررسی قرار گرفته است. فرآیند مکانیزم تپه نوردی با تولید راه‌حل های اولیه آغاز می‌شود که می‌تواند به صورت تصادفی یا ابتکاری باشد. سپس، راه‌حل فعلی توسط سه عملگر بهبود می‌یابد. عملگر اول بخشی از فضای جستجو را برای کاوش انتخاب می‌کند. عملگر دوم فضای انتخاب شده را گسترش داده و ارزیابی ها را انجام می دهد.. در نهایت، عملگر سوم با استفاده از رویکرد Survival of the fittest، بهترین راه‌حل‌ها را حفظ می‌کند. این روند تا رسیدن به حداکثر تعداد تکرارها یا پایان زمان تعیین‌شده ادامه می‌یابد.برای هماهنگی با ساختار باینری مساله انتخاب ویژگی، عملگر جدیدی در روش پیشنهادی با نام عملگر انتقال معرفی شده است که با استفاده از تابع سینوسی، متغیرهای پیوسته را به مقادیر باینری تبدیل می‌کند. نویسندگان برای بررسی کارایی الگوریتم خود از ۲۲ مجموعه داده واقعی استاندارد استفاده کرده اند. نتایج بدست آمده در این پژوهش با نتایج سه الگوریتم جستجوی محلی و ده الگوریتم فرااکتشافی مقایسه شده است. یافته‌ها نشان می‌دهد که بهینه‌ساز باینری مبتنی بر مکانیزم تپه نوردی، در ۱۶ مورد از ۲۲ مجموعه داده، نسبت به سایر الگوریتم‌های جستجوی محلی عملکرد بهتری از نظر دقت طبقه‌بندی داشته و در ۷ مجموعه داده از الگوریتم‌های فرااکتشافی پیشی گرفته است.

در تحقیق انجام شده توسط باچانین و همکاران یک الگوریتم جدید با نام QRLAOA-FS برای حل مساله انتخاب ویژگی معرفی شده که براساس الگوریتم بهینه‌سازی حسابی می باشد [۲۴]. یادگیری شبه‌انعکاسی و الگوریتم جستجوی کرم شب‌تاب. که نسخه بهبودیافته‌ای از الگوریتم بهینه‌سازی حسابی اصلی (AOA) است. در روش پیشنهادی این مقاله، مکانیزم یادگیری شبه‌انعکاسی در الگوریتم بهینه سازی حسابی به منظور افزایش تنوع جمعیت تعبیه شده است. همچنین، از الگوریتم کرم شب‌تاب (FA) برای بهبود قابلیت بهره‌برداری الگوریتم بهینه سازی حسابی استفاده شده است. هدف اصلی این پژوهش، ارائه یک روش مبتنی بر " wrapper برای بهینه سازی دقت طبقه‌بندی از طریق انتخاب زیرمجموعه بهینه‌ای از ویژگی‌ها است. این مطالعه تابع برازش را به گونه‌ای انتخاب کرده که دقت طبقه‌بندی را افزایش دهد و در عین حال تعداد ویژگی‌های انتخاب‌شده را به حداقل برساند. برای نشان دادن تأثیر خطای طبقه‌بندی و اندازه ویژگی‌ها بر تابع برازش، دو ضریب وزنی استفاده شده‌اند. برای تبدیل مقادیر پیوسته به باینری، از توابع انتقال S شکل و V شکل استفاده شده است. همچنین، از طبقه‌بند k نزدیک‌ترین همسایه برای تعیین خطای طبقه‌بندی استفاده شده است زیرا سربار محاسباتی پایینی دارد. علاوه بر این، سرعت همگرایی روش پیشنهادی با بهره گیری از قابلیت‌های بهره‌برداری الگوریتم کرم شب تابافزایش یافته است. این رویکرد ترکیبی همچنین از گیر افتادن الگوریتم در بهینه محلی جلوگیری کرده و عملکرد کلی AOA را بهبود می بخشد. الگوریتم پیشنهادی روی ده تابع آزمایشی بدون محدودیت و سپس روی ۲۱ مجموعه داده

[۲] Hill Climbing

استاندارد از مخازن دانشگاه کالیفرنیا، UCI و دانشگاه ایالتی آریزونا آزمایش و با روش‌های شناخته‌شده مقایسه شده است. همچنین، این روش روی مجموعه داده‌های مرتبط با بیماری کرونا آزمون شده است. در آزمایشات انجام شده برای بررسی عملکرد الگوریتم پیشنهادی، از چهار معیار دقت طبقه‌بندی، نسبت ویژگی های انتخاب شده، مقدار میانگین تابع برازندگی و انحراف معیار مقدار آن استفاده شده است. شایان ذکر است که در آزمایشات، از روش اعتبارسنجی ۱۰-Fold Cross-Validation در تقسیم مجموعه داده ها به مجموعه های آموزشی و ارزیابی بهره گرفته شده است. نتایج آزمایشات نشان دادند که روش پیشنهادی در ۱۳ مجموعه داده به بهترین مقدار میانگین برازندگی، در ۹ مجموعه داده به کمترین مقادیر انحراف معیار و در ۱۱ مجموعه داده به کمترین تعداد ویژگی های انتخاب شده دست یافته است. برای تحلیل عمیق‌تر، آزمون های آماری Friedman Test، Iman-Davenport Test و Holm's Step-Down Procedure بکار گرفته شده اند. نتایج این تحلیل‌های آماری نشان دادند که الگوریتم پیشنهادی در بیشتر موارد از سایر الگوریتم‌های آزمایش‌شده بهتر عمل می‌کند و عملکرد AOA اصلی را به طور قابل‌توجهی ارتقا می‌بخشد.

در کنار این موارد، الناز و الهام پاشایی یک روش جدید انتخاب ویژگی wrapper-based بر پایه الگوریتم بهینه‌سازی شامپانزه برای طبقه‌بندی داده‌های زیست‌پزشکی معرفی کرده اند [۲۵]. این الگوریتم شامپانزه در اصل برای فضاهای پیوسته طراحی شده و موفقیت‌هایی در حل این نوع مسائل داشته است در نتیجه، در این مطالعه دو نسخه باینری از الگوریتم بهینه سازی شامپانزه برای حل مسئله انتخاب ویژگی ارائه شده است. در نسخه اول، دو تابع انتقال S شکل و V شکل به منظور تبدیل نسخه پیوسته این الگوریتم به باینری به کار گرفته شده است. نسخه دوم علاوه بر استفاده از این توابع انتقال، عملگر تقاطع را نیز اضافه می‌کند تا توانایی کاوش الگوریتم بهبود یابد. همچنین در روش پیشنهادی نسخه‌ای از عملگر Intersection استفاده شده است تا قابلیت اکتشاف آن بهبود یابد. برای طبقه‌بندی نیز از مدل های پرکاربرد KNN و NB استفاده شده است. برای ارزیابی کارایی روش پیشنهادی، پنج مجموعه داده زیست‌پزشکی با ابعاد بالا و چند مجموعه داده دیگر از حوزه‌های مختلف بررسی شده است. نتایج بدست آمده سپس با شش الگوریتم wrapper-based انتخاب ویژگی شامل الگوریتم ژنتیک چندهدفه، بهینه‌سازی ازدحام ذرات، الگوریتم خفاش، بهینه‌سازی کلونی مورچه، الگوریتم کرم شب‌تاب و الگوریتم گرده‌افشانی گل و همچنین دو روش استاندارد فیلتری انتخاب ویژگی filter-based مقایسه شده‌اندنتایج آزمایش‌ها نشان می‌دهند که روش پیشنهادی قادر است تعداد کمتری از ویژگی را انتخاب کرده و هم‌زمان دقت طبقه‌بندی را نیز افزایش دهد.

در پژوهشی مشابه، یک رویکرد ترکیبی جدید با نام AIEOU برای انتخاب ویژگی های مهم از مجموعه داده ها معرفی شده است [۲۶]. در این روش پیشنهادی، الگوریتم فراابتکاری EO با مکانیزم Automated Learning ادغام شده تا پارامترهای آن به صورت بهینه مقداردهی شوند. همچنین در این روش، الگوریتم β-Adaptive Ascending بکار گرفته شده تا تعادل موثرتری بین قابلیت های جستجوی روش پیشنهادی حاصل شود. الگوریتم β-Adaptive Ascending پیش‌تر روی توابع بهینه‌سازی مختلف و برای مساله انتخاب ویژگی، همراه با بهینه‌ساز Sailfish به کار رفته است. به طور کلی، برای ترکیب الگوریتم‌های فرااکتشافی دو رویکرد وجود دارد: سطح پایین و سطح بالا. در رویکرد سطح پایین، یک الگوریتم درون الگوریتم دیگر ادغام می‌شود، در حالی که در رویکرد سطح بالا، الگوریتم‌ها به صورت متوالی اجرا می‌شوند. این پژوهش از رویکرد سطح بالا استفاده کرده که مبتنی بر مدل خط لوله‌ای است، به این معنا که خروجی یک الگوریتم به عنوان ورودی الگوریتم بعدی عمل می‌کند. در این پژوهش، برای طبقه بندی داده ها از طبقه بند k نزدیک ترین همسایه که یک الگوریتم یادگیری ماشینی با سربار کم است استفاده شده است. از آنجا که انتخاب ویژگی یک مسئله بهینه‌سازی باینری است، تابع انتقال U شکل برای محدود کردن خروجی به مقادیر صفر و یک به کار رفته است برای بررسی کارایی روش پیشنهادی، نویسندگان روش پیشنهادی خود را بر روی ۱۸ مجموعه داده از زمینه های مختلف اعمال کرده و نتایج بدست آمده را با ۸ روش انتخاب ویژگی مبتنی بر الگوریتم های فراابتکاری مقایسه کرده اند. افزون بر این، این روش روی مجموعه داده‌های Microarray با ابعاد بالا که معمولاً شامل تعداد زیادی ویژگی و تعداد محدودی نمونه هستند، اعمال شده است. این نوع

داده‌ها اغلب با مشکلی به نام curse of dimensionality روبه‌رو هستند. نتایج نشان می‌دهد که AIEOU می‌تواند به عنوان رویکردی مؤثر برای حل مسئله انتخاب ویژگی استفاده شود.

همچنین، تو و عبداله به چالش های انتخاب ویژگی‌های مهم از داده‌های با ابعاد بالا پرداخته‌اند که یک مشکل رایج در داده‌کاوی بوده و منجر به پردازش کندتر، پیچیدگی بالاتر و دقت پیش‌بینی کمتر می‌شود [۲۷]. آنها برای مقابله با این چالش ها یک روش جدید به نام بهینه‌سازی جستجوی اتم آشوبناک (CASO) را برای انتخاب ویژگی های تاثیر گذار از مجموعه داده ها معرفی کرده اند. نویسندگان معتقدند که ادغام نگاشت های آشوبناک با الگوریتم های فراابتکاری می‌تواند نرخ همگرایی و کارایی آن‌ها را بهبود بخشد بنابراین، آنها. دوازده نقشه آشوبناک مختلف را به الگوریتم بهینه‌سازی جستجوی اتم (ASO) اضافه کرده اند. برای باینری سازی الگوریتم جستجوی اتم آشوبناک جدید از یک معادله ساده و برای طبقه بندی داده ها از طبقه بند k نزدیک ترین همسایه استفاده شده است. در نهایت، عملکرد الگوریتم CASO بر روی بیست مجموعه داده مرجع از مخزن یادگیری ماشین UCI ارزیابی شده است. نتایج بدست آمده از روش پیشنهادی با نتایج روش های مشابه که از الگوریتم های ازدحام سالپ ها (SSA)، بهینه‌سازی شیر مورچه (ALO)، الگوریتم جستجوی اتم پایه (ASO)، الگوریتم گرده‌افشانی گل (FPA)، الگوریتم جستجوی کلاغ (CSA)، و بهینه‌سازی ازدحام ذرات استفاده می کنند مقایسه شده است. معیارهای ارزیابی که در آزمایشات بکار برده شده اند شامل بهترین تناسب، بدترین تناسب، میانگین تناسب، انحراف معیار تناسب ، دقت و نسبت ویژگی های انتخاب شده است. نتایج تجربی نشان می‌دهند که CASO، به‌ویژه نسخه‌ای که از نگاشت Logical-Tent ستفاده می‌کند به طور قابل‌توجهی نسبت به سایر الگوریتم‌ها در دقت طبقه‌بندی و تعداد ویژگی‌های انتخابی بهتر عمل می‌کند. آزمون رتبه‌دار ویلکاکسون نیز تفاوت معناداری بین الگوریتم CASO و سایر الگوریتم‌ها در بیشتر مجموعه داده ها نشان می دهد. نویسندگان موفقیت CASO را به توانایی آن در کاوش مؤثرتر فضای جستجو و فرار از بهینه‌های محلی نسبت داده اند که به دلیل معرفی رفتار پویای نگاشت های آشوبناک است. منحنی‌های همگرایی نیز به‌صورت بصری عملکرد برتر CASO را نشان می‌دهند. قوی‌ترین جنبه الگوریتم CASO نرخ بالای همگرایی آن و توانایی اجتناب آن از بهینه‌های محلی در مقایسه با الگوریتم ASO استاندارد است. استفاده از نگاشت های آشوبناک متعدد به یک فرآیند جستجوی مقاوم و قابل تطبیق منجر شده است که به دقت بالاتر و ویژگی‌های کمتری می‌انجامد. با این حال، یک ضعف قابل ذکر این است که نگاشت آشوبناک بهینه باید به‌صورت تجربی تعیین شود و هیچ تحلیل تئوری برای آن وجود ندارد. در حالیکه CASO در دقت و کاهش ویژگی‌ها برتری دارد، هزینه محاسباتی الگوریتم به‌صراحت مورد بحث قرار نگرفته است که می‌تواند محدودیتی برای مجموعه داده‌های بسیار بزرگ باشد.

بر اساس مطالعه انجام شده توسط تو و همکاران، روش جدیدی با نام SBWOA برای انتخاب ویژگی از مجموعه داده‌های با ابعاد بالا معرفی شده است [۲۸]. ایده اصلی این روش، بهبود کارایی الگوریتم بهینه‌سازی نهنگ (WOA) از طریق استفاده از یک استراتژی Spatial Boundary است. این استراتژی تعداد ویژگی‌های انتخاب‌شده توسط هر راه‌حل را در یک محدوده‌ی مشخص تنظیم می‌کند و از همگرایی زودرس جلوگیری کرده و کاوش در فضای ویژگی را بهبود می‌بخشد. این الگوریتم از یک طرح Tournament Selection برای انتخاب بهترین نرخ‌های ابعاد و مکانیزم به‌روزرسانی فضایی برای بهبود فرآیند جستجو استفاده می‌کند. طبقه بند k نزدیک‌ترین همسایه به عنوان مدل یادگیری ماشین برای ارزیابی کیفیت ویژگی‌های انتخاب‌شده در روش پیشنهادی استفاده شده است. نقطه قوت اصلی SBWOA تواناییجستجو در فضاهای با ابعاد بالا است که با استفاده از مکانیزم های Spatial Boundary و Tournament Selection حاصل شده است. این مکانیزم ها به‌طور مؤثری محدودیت‌های WOA استاندارد مانند همگرایی زودرس و گرفتار شدن در بهینه‌های محلی را بر طرف می کنند. یک ضعف این روش پیشنهادی، وابستگی آن به Initial Dimension Rates ایت که باید به‌صورت تجربی تعیین شوند که ممکن است برای همه مجموعه داده‌ها بهینه نباشد. عملکرد روش SBWOA با استفاده از ۱۶ مجموعه داده با ابعاد بالا از دانشگاه ایالتی آریزونا ارزیابی شده و نتایج بدست آمده با هشت روش انتخاب ویژگی که براساس الگوریتم های فراابتکاری می باشند مقایسه شده است. معیارهای ارزیابی بکار گرفته شده در آزمایشات شامل میانگین تناسب، انحراف معیار،

accuracy، Precision، معیار F-measure، تعداد ویژگی های انتخاب شده و زمان محاسباتی است. نتایج نشان می‌دهند که SBWOA به‌طور قابل‌توجهی از WOA سنتی و سایر الگوریتم‌های رقیب برتری دارد. همچنین، روش پیشنهادیتعداد کمتری ویژگی انتخاب می‌کنند در حالیکه دقت طبقه‌بندی بالایی را حفظ می‌کنند. آزمون‌های رتبه‌بندی ویلکاکسون نیز برتری عملکرد SBWOA را به‌طور آماری تأیید می‌کنند.

در تحقیقی که الشورباجی و همکاران انجام داده اند، یک روش جدید با نام RSA-SO برای انتخاب ویژگی معرفی شده است [۲۹]. در روش انتخاب ویژگی معرفی شده در این پژوهش، یک مکانیزم موازی قرار دارد که الگوریتم جستجوی خزندگان را با الگوریتم بهینه ازی مار ترکیب می‌کند. این رویکرد موازی به منظور کاهش خطر گیر افتادن الگوریتم‌ها در بهینه‌های محلی و بهبود تعادل بین قابلیت های اکتشاف و بهره‌برداری در حین جستجوی مجموعه ویژگی‌های بهینه طراحی شده است. این الگوریتم بر روی دوازده مجموعه داده شامل ده مجموعه از مخزن UCI و دو مسئله مهندسی واقعی آزمایش شده و نتایج آن با هفت روش فراابتکاری محبوب مقایسه شده است. معیارهای ارزیابی شامل دقت طبقه‌بندی، تعداد ویژگی‌های انتخاب‌شده، مقادیر تناسب (بهترین، بدترین، میانگین و انحراف معیار) و زمان محاسباتی است. نتایج نشان می‌دهد که RSA-SO عملکرد رقابتی دارد و اغلب دقت بالاتر و تعداد کمتری از ویژگی‌ها را انتخاب می‌کند. عملکرد الگوریتم همچنین در دو مسئله بهینه‌سازی مهندسی با نام های Pressure Vessel Design و Cantilever Beam Design ارزیابی شده و نتایج امیدوارکننده‌ای نشان داده است. آزمون رتبه‌بندی فریدمن برای مقایسه آماری نتایج RSA-SO و سایر الگوریتم‌های فراابتکاریاستفاده شده است. نقطه قوت RSA-SO در اجرای موازی RSA و SO است که تعادل بین جستجوی فضای راه‌حل و تمرکز بر مناطق امیدبخش را برقرار می‌کند. این تعادل منجر به بهبود دقت و کاهش تعداد ویژگی‌ها نسبت به سایر الگوریتم‌های مقایسه شده گردیده است. توانایی الگوریتم برای همگرایی سریع به راه‌حل‌های بهینه در بیشتر مجموعه داده‌های آزمایش شده یک مزیت قابل توجه دیگر است. با این حال، زمان محاسباتی RSA-SO یکی از ضعف های این روش پیشنهادی است. اگرچه در برخی موارد از نظر سرعت عملکرد بهتری دارد، اما در همه مجموعه داده‌ها سریع‌ترین نیست. نویسندگان نیز اذعان داشته اند که پیچیدگی زمانی نیاز به بهبود بیشتری دارد، به ویژه هنگام کار با داده‌های با ابعاد بالا.

علاوه بر این ها، گد و همکاران یک روش جدید با نام الگوریتم بهبودیافته جستجوی گنجشک باینری یا به اختصار IBSSA را برای حل مساله انتخاب ویژگی در طبقه بندی داده ها ارائه داده اند [۳۰]. روش پیشنهادی نویسندگان در این پژوهش عملکرد SSA الگوریتم جستجوی گنجشک پایه در حوزه یادگیری ماشین را به طور چشمگیری افزایش داده است. برای این منظور، نویسندگان در این پژوهش به محدودیت های الگوریتم جستجوی گنجشک پایه که شامل ضعف در اکتشاف و رکود در بهره برداری است پرداخته و دو رویکرد اساسی را برای رفع آنها بکار برده اند. در رویکرد اول، یک استراتژی جستجوی همسایگی و در رویکرد دوم یک الگوریتم جستجوی محلی جدید با نام Random Repositioning of Roaming Agents بکار گرفته شده اند. همچنین، برای تبدیل خروجی های پیوسته الگوریتم به مقادیر باینری متناظر که نشان دهنده ویژگی‌ها هستند از نه تابع انتقال مختلف استفاده شده است شایان ذکر است که در روش پیشنهادی از سه طبقه پرکاربرد KNN، SVM و RF (جنگل تصادفی) بهره گرفته شده است که یکی از نقاط قوت اصلی این پژوهش به شمار می آید. در انتها برای ارزیابی کارایی، روش IBSSA بر روی ۱۸ مجموعه داده استاندارد از مخزن UCI اعمال شده است. نتایج کسب شده توسط روش پیشنهادی با نتایج ۱۲ الگوریتم فراابتکاری باینری دیگر مقایسه شده است. در مقایسات، از معیارهای میانگین دقت طبقه بندی ، میانگین مقدار تابع برازندگی و تعداد ویژگی های انتخاب شده استفاده شده است. همچنین، برای تأیید برتری IBSSA از آزمون رتبه بندی آمرای ویلکاکسون استفاده شده است. یکی از نقاط قوت اصلی این پژوهش، ارزیابی جامع و دقیق روش پیشنهادی بر روی مجموعه داده های مختلف، با استفاده از چندین طبقه بند و توابع انتقال مختلف است که نشان دهنده ارزیابی قوی از عملکرد الگوریتم پیشنهادی است. استفاده از تحلیل‌های آماری مانند آزمون ویلکاکسون نیز اعتبار بیشتری به نتایج تحقیق بخشیده است. با این حال، یک ضعف قابل توجه این است که IBSSA در برخی موارد، به ویژه در مجموعه

داده های با ابعاد بالا و نمونه‌های کم، تعداد بیشتری از ویژگی ها را نسبت به الگوریتم های رقیب انتخاب می کند. این مسئله نشان دهنده نیاز به بهبود بیشتر در استراتژی انتخاب ویژگی الگوریتم است.

در پژوهش انجام شده توسط رستمی و همکاران یک روش نوین مبتنی بر الگوریتم ژنتیک برای انتخاب ویژگی ارائه می‌دهند که با نام CDGAFS شناخته می‌شود و هدف آن حل مشکلات مرتبط با داده‌های با ابعاد بالا در حوزه یادگیری ماشین است [۳۱]. فرآیند این الگوریتم شامل سه گام اساسی است: در مرحله اول، شباهت‌های میان ویژگی‌ها از طریق محاسبه ضریب همبستگی پیرسون ارزیابی می‌شود. در گام بعدی، ویژگی‌های مشابه با بهره‌گیری از یک الگوریتم تشخیص اجتماع به نام Iterative search algorithm for community detection خوشه‌های مرتبط گروه‌بندی می‌شوند. در نهایت، الگوریتم ژنتیک با استفاده از یک مکانیزم community-based repair operation زیرمجموعه‌ای از ویژگی‌ها را انتخاب می‌کند که هم از نظر ارتباطی قوی و هم از لحاظ تنوع مناسب هستند. یکی از مزایای کلیدی این روش، توانایی آن در تعیین خودکار تعداد بهینه ویژگی‌ها و خوشه‌ها است که آن را از سایر روش‌های مرسوم متمایز می‌کند. علاوه بر این، نویسندگان در این پژوهش تحلیل حساسیت پارامترهای کلیدی الگوریتم θ ژنتیک پیچیدگی محاسباتی آن را انجام داده‌اند. روش پیشنهادی قادر است تعداد بهینه ویژگی ها و خوشه ها تعیین کرده کند. همچنین، به واسطه تابع برازندگی چند هدفه بکار برده شده می تواند معیار مرتبط بودن ویژگی ها را تشخیص داده و افزونگی ویژگی ها کاهش دهد. برای طبقه بندی داده ها، روش پیشنهادی سه طبقه بند متفاوت (KNN، SVM و AdaBoost) را بکار گرفته است. یکی از محدودیت‌های روش پیشنهادی این پژوهش، وابستگی آن به پارامترهای θو ω است که باید توسط کاربر تعیین شوند; هر چند تحلیل حساسیت به مقادیر پارامترها ارائه شده است. همچنین، در مواجهه با مجموعه داده‌های بسیار بزرگ، چالش‌های محاسباتی ممکن است ایجاد شوند.. در بخش آزمایشات انجام شده، روشCDGAFS بر روس شش مجموعه داده استاندارد اعمال شده و نتایج آن با الگوریتم های PSO، ACO و ABC مقایسه شده است. معیارهای اصلی ارزیابی شامل دقت طبقه‌بند و تعداد ویژگی‌های انتخابی بوده است. نتایج نیز حاکی از آن است که CDGAFS در تمامی مجموعه داده‌ها و طبقه‌بندها از نظر دقت عملکرد برتری داشته و در اکثر موارد زیرمجموعه‌های کوچک‌تر و بهینه تری از ویژگی‌ها را انتخاب کرده است. علاوه بر این، برتری CDGAFS از طریق آزمون آماری فریدمن به تأیید رسیده است.

در پژوهشی مشابه که توسط باتاچاریاج و همکاران انجام شده است، یک روش جدید با نام MA-HS برای انتخاب ویژگی‌ها ارائه شده است [۳۲]. این روش، دو الگوریتم بهینه سازی Mayfly و جستجوی هارمونی را با هم ترکیب می‌کند. الگوریتم Mayfly یک الگوریتم فرابتکاری مبتنی بر جمعیت است که از رفتار جفت گیری شاپره ها الهام گرفته شده است. الگوریتم Harmony Search (HS) جستجوی هارمونی نیز یک الگوریتم بهینه سازی است که بر اساس اصول موسیقی و هارمونی ها طراحی شده است. هدف اصلی این ترکیب این دو الگوریتم در این پژوهش، استفاده از توانایی های اکتشافی قوی الگوریتم جستجوی هارمونی برای ایجاد زیرمجموعه های متنوع از ویژگی ها و سپس بهبود این زیرمجموعه ها با استفاده از Mayfly برای دست یابی به راه حل های بهینه است .برای تبدیل فضای جستجوی پیوسته به فضای باینری مناسب برای انتخاب ویژگی‌ها، از ت,ابع انتقال S-shaped استفاده شده است. همچنین، در روش پیشنهادی از طبقه بند k نزدیک ترین همسایه استفاده شده است که در آن مقدار k برابر ۵ تنظیم شده است. تابع برازش نیز هم خطای طبقه بندی و هم تعداد ویژگی های انتخاب شده را در نظر می‌گیرد و هدف آن کاهش همزمان این دو معیار است. از جمله مزایای الگوریتم MA-HS می توان به عملکرد برتر آن در مقایسه با روش های موجود، توانایی آن در مدیریت مجموعه داده های با ابعاد مختلف و سرعت همگرایی مناسب اشاره کرد. با این حال، برخی محدودیت ها نیز مانند احتمال همگرایی زودرس وجود دارد. هرچند نویسندگان راهنمایی هایی برای تنظیم پارامترهای آن ارائه کرده اند، با اینحال احتمال وابستگی به مقادیر این پارامترها هنوز وجود داردر بخش آزمایشات الگوریتم MA-HS روی ۱۸ مجموعه داده UCI و ۳ مجموعه داده میکروآرایه با ابعاد بالا آزمایش شده و نتایج آن با ۱۲ روش انتخاب ویژگی مبتنی بر الگوریتم های فرابتکاری دیگر مقایسه شده است. نتایج مقایسه ها

ناشن دادند که الگوریتم MA-HS از نظر دقت طبقه بندی و تعداد ویژگی های انتخاب شده عملکرد بهتری نسبت به سایر روش ها داشته و در بیشتر موارد بهترین رتبه را کسب کرده است. اعتبارسنجی آزمایشات توسط آزمون فریدمن تأیید شده است.

در کنار این موارد، نویسندگان در [۳۳] به بررسی روش‌های بهبود انتخاب ویژگی درحوزه‌های داده‌کاوی و هوش مصنوعی پرداخته و دو روش جدید با نام های MBOLF و MBOICO برای انتخاب ویژگی های تاثیر گذار از مجموعه داده ها ارائه کرده اند. در روش های MBOLF و MBOICO از الگوریتم بهینه سازی شاه پروانه برای انجام فرآیند بهینه سازی استفاده شده است. در الگوریتم MBOLF از توزیع Levy flight random walk و در الگوریتم MBOICO از یک نسخه از عملگر crossover برای افزایش سرعت همگرایی الگوریتم ها استفاده شده اند. تغییرات اعمال‌شده بر روی الگوریتم MBO پایه منجر به بهبود چشمگیر عملکرد آن در مسائل انتخاب ویژگی شده است. یکی از نقاط قوت این پژوهش، ارزیابی های گسترده آن بر روی مجموعه داده‌های متنوع و مقایسه‌ی آن با الگوریتم‌های فراابتکاری معتبر است. با این حال، محدودیت این تحقیق در استفاده از تنها یک طبقه‌بند برای ارزیابی است. به کارگیری چندین طبقه‌بند مختلف می‌تواند ارزیابی جامع‌تر و قابل اعتمادتری از قابلیت تعمیم روش‌های پیشنهادی ارائه کند. همچنین، اگرچه مقاله به بررسی پیچیدگی زمانی الگوریتم‌ها پرداخته است، ارائه یک تحلیل تجربی دقیق‌تر از زمان اجرا می‌تواند نتایج مربوط به سرعت همگرایی را تقویت کند. معیارهای اصلی برای ارزیابی عملکرد شامل دقت طبقه‌بندی، تعداد ویژگی‌های انتخاب شده و سرعت همگرایی الگوریتم‌ها می باشد. همچنین، یک تابع هدف که ترکیبی از خطای طبقه‌بندی و تعداد ویژگی‌های انتخاب‌شده است، به عنوان معیار هدایت فرآیند بهینه‌سازی بکار گرفته شده است. شایان ذکر است که برای طبقه بندی داده ها در هر دو روش از طبقه بند k نزدیک ترین همسایه استفاده شده است. در انتها، روش پیشنهادی بر روی ۲۵ مجموعه داده استاندارد از مخزن UCI مورد ارزیابی قرار گرفته و نتایج آن با الگوریتم مشابه از جمله GA، PSO، ALO و WOASAT مقایسه شده است. یافته‌ها حاکی از آن است که روش MBOICO در مقایسه با سایر روش‌ها، از نظر دقت طبقه‌بندی (با میانگین ۹۳٪) و کاهش تعداد ویژگی‌های انتخاب شده عملکرد بهتری داشته است. الگوریتم MBOLF نیز بهبودهایی در سرعت همگرایی نشان داده اما از نظر دقت کمی ضعیف‌تر از MBOICO عمل کرده است. برای اعتبارسنجی آزمایشات از آزمون t-test استفاده شده است.

.

۴- خلاصه مطالب

این مقاله یک بررسی جامع در مورد الگوریتم‌های انتخاب ویژگی که براساس الگوریتم های فراابتکاری توسعه یافته‌اند ارائه می کند. مقالات بررسی شده در این پژوهش مربوط به سال ها ۲۰۲۰ تا ۲۰۲۳ می باشند که در مجلات معتبر علمی به چاپ رسیده اند. در ابتدا، مقدمه بر مساله انتخاب ویژگی، لزوم و ضرورت آن، کاربردها و مزایای آن، چالش هایی که انتخاب ویژگی با آنها مواجهه است و غیره ارائه شده اند. در فصل دوم مقاله، مساله انتخاب واحد به طور جامعی تعریف گردیده و نحوه کدگذاری آن در روش های مبتنی بر الگوریتم های فراابتکاری فراهم گردیده است. در ادامه، الگوریتم های فراابتکاری تشریح گردیده و انواع دسته بندی آنها ارائه شده است. همچنین، رویکرد حل مساله این الگوریتم ها به همراه چالش های پیش رو مطرح گردیده اند. در کنار این، برای درک صحیح تر خوانندگان، توصیف دقیق و مدل ریاضی مسئله انتخاب ویژگی و لزوم باینری سازی الگوریتم ها ارائه شده است. نحوه حل مساله انتخاب ویژگی توسط الگوریتم های فراابتکاری به همراه شماتیکی از روند فرآیند ان فراهم گردیده است.

در بررسی های انجام شده بر روی روش های مورد مطالعه، کاستی ها و ضعف های الگوریتم ها که شامل نرخ همگرایی پایین آنها، حرکت های تصادفی نابهینه، جستجو در جهت های نامشخص، گیر کردن در بهینه های محلی، همگرایی زودرس، وابستگی به مقادیر پارامترها و غیره می باشد مشخص شده اند. الگوریتم بهینه سازی بکار گرفته شده در هر روش استخراج شده و نوآوری های صورت گرفته بر روی هرکدام در جهت رفع کاستی های ذکر شده مورد بررسی قرار گرفته اند. نحوه باینری کردن الگوریتم های فراابتکاری در هر پژوهش مشخص و ذکر شده است. علاوه بر این، طبقه بند استفاده شده در هر پژوهش بیان شده و مقایسه گردیده است. در ادامه، مساله و مجموعه داده هایی که روش های انتخاب ویژگی بر روی آنها اعمال شده اند برجسته شده و جزئیات کامل آنها ارائه

شده است. همچنین، معیارهای ارزیابی بکار رفته در هر پژوهش به منظور ارزیابی کارایی روش پیشنهادی استخراج و مقایسه شده اند. فرمول های ریاضی این معیارهای ارزیابی نیز در این پژوهش فراهم شده اند. در نهایت، نقاط قوت و ضعف هر روش معرفی شده بیان شده و راهکاریی نیز ارائه شده اند.

مراجع

١. Barshandeh, S., et al., *MPCASMA: A Multi-Population Chaotic-based Hybrid Algorithm for Global Optimization and Its Application in Feature Selection*, in هفتمین کنفرانس بین المللی پژوهش های کاربردی در علوم و مهندسی. ١٤٠٢.

٢. Theng, D. and K.K. Bhoyar, *Feature selection techniques for machine learning: a survey of more than two decades of research.* Knowledge and Information Systems, ٢٠٢٤. ٦٦(٣): p. ١٥٧٥-١٦٣٧.

٣. Barbieri, M.C., B.I. Grisci, and M. Dorn, *Analysis and comparison of feature selection methods towards performance and stability.* Expert Systems with Applications, ٢٠٢٤: p. ١٢٣٦٦٧.

٤. Wang, H., et al., *Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods.* Journal of Big Data, ٢٠٢٤. ١١(١): p. ٤٤.

٥. Maseno, E.M. and Z. Wang, *Hybrid wrapper feature selection method based on genetic algorithm and extreme learning machine for intrusion detection.* Journal of Big Data, ٢٠٢٤. ١١(١): p. ٢٤.

٦. Liu, H. and L. Yu, *Toward integrating feature selection algorithms for classification and clustering.* IEEE Transactions on knowledge and data engineering, ٢٠٠٥. ١٧(٤): p. ٤٩١-٥٠٢.

٧. Ahmed, S., M. Zhang, and L. Peng. *Enhanced feature selection for biomarker discovery in LC-MS data using GP*. in *2013 IEEE congress on evolutionary computation*. ٢٠١٣. IEEE.

٨. Aghdam, M.H., N. Ghasem-Aghaee, and M.E. Basiri, *Text feature selection using ant colony optimization.* Expert systems with applications, ٢٠٠٩. ٣٦(٣): p. ٦٨٤٣-٦٨٥٣.

٩. Ghosh, A., A. Datta, and S. Ghosh, *Self-adaptive differential evolution for feature selection in hyperspectral image data.* Applied Soft Computing, ٢٠١٣. ١٣(٤): p. ١٩٦٩-١٩٧٧.

١٠. Aggarwal, C.C., et al., *Active learning: A survey*, in *Data classification*. ٢٠١٤, Chapman and Hall/CRC. p. ٥٩٩-٦٣٤.

١١. Hassan, A., et al., *A wrapper feature selection approach using Markov blankets.* Pattern Recognition, ٢٠٢٥. ١٥٨: p. ١١١٠٦٩.

١٢. Mandal, A.K., et al., *Feature subset selection for high-dimensional, low sampling size data classification using ensemble feature selection with a wrapper-based search.* IEEE Access, ٢٠٢٤.

١٣. Barshandeh, S., et al., *A Chaotic-integrated Harris Hawks Optimization Algorithm for Solving Numerical Optimization Problems*, in دومین کنفرانس برق، مکانیک ،هوافضا، کامپیوتر و علوم مهندسی. ١٤٠٢.

١٤. Barshandeh, S., R. Dana, and P. Eskandarian, *A learning automata-based hybrid MPA and JS algorithm for numerical optimization problems and its application on data clustering.* Knowledge-Based Systems, ٢٠٢٢. ٢٣٦: p. ١٠٧٦٨٢.

١٥. Sabzalizadeh, R., S. Barshandeh, and S. Gholizadeh, *An Invasive Weed Optimization-based Energy and Resource-efficient Workflow Scheduling Algorithm for the Cloud Environment*, in بیستمین کنفرانس بین المللی فناوری اطلاعات، کامپیوتر و مخابرات. ١٤٠٢.

١٦. Barshandeh, S. and M. Haghzadeh, *A new hybrid chaotic atom search optimization based on tree-seed algorithm and Levy flight for solving optimization problems.* Engineering with Computers, ٢٠٢١. ٣٧(٤): p. ٣٠٧٩-٣١٢٢.

١٧. Barshandeh, S., et al., *A learning-based metaheuristic administered positioning model for 3D IoT networks.* Applied Soft Computing, ٢٠٢٣. ١٣٦: p. ١١٠١١٣.

١٨. Barshandeh, S., F. Piri, and S.R. Sangani, *HMPA: an innovative hybrid multi-population algorithm based on artificial ecosystem-based and Harris Hawks optimization algorithms for engineering problems.* Engineering with computers, ٢٠٢٢. **٣٨**(٢): p. ١٥٨١-١٦٢٥.

١٩. Zivkovic, M., et al., *Novel improved salp swarm algorithm: An application for feature selection.* Sensors, ٢٠٢٢. **٢٢**(٥): p. ١٧١١.

٢٠. Thaher, T. and N. Arman. *Efficient multi-swarm binary harris hawks optimization as a feature selection approach for software fault prediction*. in *2020 11th International conference on information and communication systems (ICICS)*. ٢٠٢٠. IEEE.

٢١. Ouadfel, S. and M. Abd Elaziz, *Efficient high-dimension feature selection based on enhanced equilibrium optimizer.* Expert Systems with Applications, ٢٠٢٢. **١٨٧**: p. ١١٥٨٨٢.

٢٢. Chaudhuri, A. and T.P. Sahu, *Binary Jaya algorithm based on binary similarity measure for feature selection.* Journal of Ambient Intelligence and Humanized Computing, ٢٠٢٢. **١٣**(١٢): p. ٥٦٢٧-٥٦٤٤.

٢٣. Al-Betar, M.A., et al., *Binary β-hill climbing optimizer with S-shape transfer function for feature selection.* Journal of Ambient Intelligence and Humanized Computing, ٢٠٢١. **١٢**(٧): p. ٧٦٣٧-٧٦٦٥.

٢٤. Bacanin, N., et al., *Quasi-reflection learning arithmetic optimization algorithm firefly search for feature selection.* Heliyon, ٢٠٢٣. **٩**(٤).

٢٥. Pashaei, E. and E. Pashaei, *An efficient binary chimp optimization algorithm for feature selection in biomedical data classification.* Neural Computing and Applications, ٢٠٢٢. **٣٤**(٨): p. ٦٤٢٧-٦٤٥١.

٢٦. Ahmed, S., et al., *AIEOU: Automata-based improved equilibrium optimizer with U-shaped transfer function for feature selection.* Knowledge-Based Systems, ٢٠٢١. **٢٢٨**: p. ١٠٧٢٨٣.

٢٧. Too, J. and A.R. Abdullah, *Chaotic atom search optimization for feature selection.* Arabian Journal for Science and Engineering, ٢٠٢٠. **٤٥**(٨): p. ٦٠٦٣-٦٠٧٩.

٢٨. Too, J., M. Mafarja, and S. Mirjalili, *Spatial bound whale optimization algorithm: an efficient high-dimensional feature selection approach.* Neural Computing and Applications, ٢٠٢١. **٣٣**(٢٣): p. ١٦٢٢٩-١٦٢٥٠.

٢٩. Al-Shourbaji, I., et al., *An efficient parallel reptile search algorithm and snake optimizer approach for feature selection.* Mathematics, ٢٠٢٢. **١٠**(١٣): p. ٢٣٥١.

٣٠. Gad, A.G., et al., *An improved binary sparrow search algorithm for feature selection in data classification.* Neural Computing and Applications, ٢٠٢٢. **٣٤**(١٨): p. ١٥٧٠٥-١٥٧٥٢.

٣١. Rostami, M., K. Berahmand, and S. Forouzandeh, *A novel community detection based genetic algorithm for feature selection.* Journal of Big Data, ٢٠٢١. **٨**(١): p. ٢.

٣٢. Bhattacharyya, T., et al., *Mayfly in harmony: A new hybrid meta-heuristic feature selection algorithm.* IEEE Access, ٢٠٢٠. **٨**: p. ١٩٥٩٢٩-١٩٥٩٤٥.

٣٣. Alweshah, M., *Solving feature selection problems by combining mutation and crossover operations with the monarch butterfly optimization algorithm.* Applied Intelligence, ٢٠٢١. **٥١**(٦): p. ٤٠٥٨-٤٠٨١.

25 th International Conference on Information Technology, Computer and Telecommunication
Event Place: Tbilisi,Georgia
www.itctconf.ir
بیست و پنجمین کنفرانس بین المللی
فناوری اطلاعات، کامپیوتر و مخابرات | گرجستان
25th International Conference on Information Technology,Computer and Telecommunication
PUBLISH IN JOURNALS — INTERNATIONAL CERTIFICATION

A Comprehensive Survey on Metaheuristic-based Feature Selection Methods

Saeid Barshandeh*
Department of Computer Science, School of Engineering, Afagh Higher Education Institute, Urmia, Iran
Farnaz Samadzad Azar Keshtiban
Department of computer engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
Masoumeh Ghasemi Kouchmeshki
Department of Computer Engineering, Payame Noor University, Tehran, Iran
Melika Bazmani
Department of Computer Engineering, Payame Noor University, Tehran, Iran

Abstract

Machine Learning (ML) algorithms have a special role in today's real-world applications. These algorithms are capable of solving complex and time-consuming problems in a timely manner. ML algorithms can generate optimal solutions considering the type of problem. These algorithms are used in various branches of science, including disease diagnosis, cost estimation, classification, intrusion detection, image processing, and many others. However, the efficiency of these algorithms is highly dependent on the given training dataset. Also, the advancement of technology and data collection tools has led to the emergence of large-scale datasets that overshadow the efficiency of ML algorithms. As the dimension of the dataset increases, the training time of learning models increases. Additionally, their accuracy decreases due to the presence of irrelevant and redundant features in the dataset. One of the data preprocessing techniques in ML algorithms is feature selection, through which irrelevant and redundant features are omitted from the dataset. By removing irrelevant features, in addition to reducing the learning time of the models, the accuracy of the models also increases because the model focuses on more important and influential features. In recent years, various feature selection methods have been proposed, the most important of which are methods based on metaheuristic algorithms. These methods determine the influential features of the data set by considering the problem and its objective function. Therefore, this paper investigates the latest metaheuristic-based feature selection algorithms, and along with summarizing them, compares them from different aspects.

Keywords: Machine Learning, Feature Selection, Optimization, Metaheuristic Algorithms