نوع مقاله: پژوهشی

ارائه یک روش تکاملی جدید مبتنی بر باکتری برای حل مسئله انتخاب ویژگی

محمود پرنده[۱]، دانشجو دکتری، مینا زلفی لیقوان[۲]، استادیار

۱- مهندسی فناوری اطلاعات – مهندسی برق و کامپیوتر – دانشگاه تبریز - تبریز - ایران – parandeh@tabrizu.ac.ir
۲- مهندسی کامپیوتر - دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران – mzolfy@tabrizu.ac.ir

چکیده: ابعاد عظیم داده به یکی از چالش برانگیزترین مسائل این روزها تبدیل شده است. ابعاد داده بالا، که به عنوان "نفرین ابعاد" شناخته می شود، الگوریتم‌های یادگیری ماشین را نیز به چالش می‌کشد. الگوریتم‌های انتخاب ویژگی زیادی برای غلبه بر این مشکل پیشنهاد شده است. وجود داده‌های نامتوازن در برخی از برنامه‌ها، ضمن کاهش کارایی اکثر الگوریتم‌های انتخاب ویژگی، آن‌ها را به سمت کلاس با داده اکثریت سوق می‌دهد. هدف اصلی این مقاله ارائه یک الگوریتم جدید انتخاب ویژگی تکاملی مبتنی بر باکتری است که کارایی خود را در مواجهه با داده‌های نامتوازن حفظ می‌کند. روش پیشنهادی از عملگرهای جهش و تزریق هوشمند استفاده می‌کند که رفتار باکتری‌ها را شبیه‌سازی می‌کند. در این دو عملگر، باکتری‌های برتر بهترین ویژگی‌های برجسته خود را به باکتری‌های دیگر منتقل می‌کنند و عملگر جهش، آن ویژگی‌ها را با کمترین رتبه جایگزین می‌کند. عملگر نخبه‌گرایی با تزریق بهترین ویژگی‌ها به سایر اعضا سعی در بهبود اعضای آرشیو دارد. در مجموعه داده Isolet، روش پیشنهادی در مقایسه با روش‌های مرسوم، به ترتیب، به کاهش انتخاب ویژگی و افزایش دقت ۴۳ و ۷ درصد منجر می‌شود. همچنین، F-measure برای ارزیابی ویژگی‌های انتخاب شده استفاده می‌شود. نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی در انتخاب تعداد ویژگی‌ها با توجه به خطای الگوریتم در مقایسه با روش‌های موجود عملکرد بهتری دارد.

واژه‌های کلیدی: الگوریتم تکاملی، انتخاب ویژگی، جهش هوشمند، نخبه‌گرایی، بهینه‌سازی چندهدفه، الگوریتم باکتری

Introduce a new bacterial-based evolutionary algorithm to solve the feature selection problem

Mahmoud parandeh, phd student[1], Mina zolfy lighvan, associate professor[2]

1- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, parandeh@tabrizu.ac.ir
2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, mzolfy@tabrizu.ac.ir

Abstract: In contemporary times, dealing with enormous data dimensions has emerged as a critical challenge. The "curse of dimensionality" due to high data dimensions poses a continuing challenge to machine learning algorithms. To overcome this problem, many feature selection algorithms have been proposed. However, the presence of imbalanced data in some applications limits the efficiency of most feature selection algorithms and pushes them towards the majority class. Therefore, this study introduces a new bacterial-based evolutionary feature selection algorithm that maintains its efficiency even in the case of imbalanced data. The proposed method employs intelligent mutation and injection operators, which simulate the behavior of bacteria. In these two operators, prior bacteria transfer their best-ranked and prominent features to other bacteria, and the mutation operator replaces those features with the lowest rank. The elitism operator aims to enhance archive members by injecting the best features into other members. The proposed method outperforms conventional methods in terms of feature selection reduction and accuracy enhancement by 43% and 7%, respectively, on the Isolet dataset. The selected features are evaluated using F-measure. The simulation results indicate that the proposed method performs better in selecting the number of features with respect to the algorithm's error compared to state-of-the-art methods.

Keywords: Evolutionary algorithm, feature selection, intelligent mutation, elitism, multi-objective optimization, bacterial algorithm.

تاریخ ارسال مقاله: ۱۴۰۱/۰۷/۰۱
تاریخ اصلاح مقاله: ۱۴۰۲/۰۲/۱۴
تاریخ پذیرش مقاله: ۱۴۰۲/۰۴/۰۶
نام نویسنده مسئول: مینا زلفی لیقوان
نشانی نویسنده مسئول: ایران – تبریز – دانشگاه تبریز – دانشکده مهندسی برق و کامپیوتر – گروه مهندسی کامپیوتر

Digital Object Identifier 10.22034/JASP.

۱- مقدمه

امروزه پیشرفت تکنولوژی و افزایش فضاهای ذخیره‌سازی، باعث پدید آمدن داده‌ها با ابعاد بسیار زیاد شده است [۱]. هر چه ابعاد داده‌های جمع‌آوری شده افزایش یابد، فضای جستجو بزرگتر خواهد شد و چالش‌هایی مانند کاهش تعمیم‌پذیری را به دنبال خواهد داشت [۲-۴]. در اینگونه داده‌ها معمولا تعداد بسیاری ویژگی وجود دارد که اطلاعات مناسبی راجع به کلاس داده ارائه نمی‌کنند. به‌همین دلیل کاهش ابعاد داده به یکی از مسائل مهم در الگوریتم‌های یادگیری ماشین تبدیل شده است. یکی از راه‌کارهای مقابله با این مشکل، استفاده از الگوریتم‌های انتخاب ویژگی است. الگوریتم‌های انتخاب ویژگی، زیر مجموعه‌ای از ویژگی‌های داده را بدون کاهش چشمگیر در دقت الگوریتم یادگیری انتخاب می‌کنند.

از داده‌های ریزآرایه[۱] می‌توان به عنوان مثالی از داده با ویژگی‌های بالا یاد کرد [۵]. در اینگونه داده‌ها، ژن‌ها دارای اطلاعات مهمی هستند که در کاربردهایی مانند تشخیص بیماری [۶] می‌تواند موثر واقع شود. از آنجایی که تعداد نمونه‌های بیمار در مقایسه با نمونه‌های سالم بسیار کمتر است و همچنین داده‌ها، اغلب دارای ابعاد بزرگی هستند، چالش جدیدی برای تشخیص نمونه‌های بیمار به‌وجود می‌آید. این چالش با نام داده‌های نامتوازن[۲] شناخته می‌شود [۷] که در آن تعداد داده‌های یک کلاس در مقایسه با کلاس‌های دیگر بسیار اندک است. الگوریتم‌های یادگیری در مواجهه با اینگونه کلاس‌ها سعی در یادگیری کلاس با نمونه‌های بیشتر را خواهند داشت. در صورتی که معمولا کلاس با داده‌های کمتر از اهمیت بیشتری نسبت به سایر کلاس‌ها برخوردار است. برای مقابله با این مشکل نیاز به تعریف معیارهای جدیدی [۸] است که باعث شود الگوریتم‌های انتخاب ویژگی سعی در یادگیری ویژگی‌هایی نمایند که داده‌ها با نمونه کلاس کمتر نیز توسط الگوریتم یادگیری ماشین در نظر گرفته شوند.

الگوریتم‌های انتخاب ویژگی به صورت کلی به سه نوع: Filter[۹]، Wrapper[۱۰] و Embedded[۱۱] دسته‌بندی می‌شوند. در روش Filter معیارهایی از قبل تعریف می‌شوند که مستقل از فاز یادگیری هستند. این معیارها کیفیت ویژگی‌های انتخاب شده را بررسی می‌کنند و از آنجایی که هیچ ارتباط مستقیمی با الگوریتم یادگیری ندارند، تضمینی بر کیفیت ویژگی‌های انتخاب شده و تاثیر آن برروی بهبود الگوریتم یادگیری وجود ندارد. روش‌های مبتنی بر Filter به دلیل اینکه در فرآیند یادگیری دخالتی ندارند، کم هزینه هستند[۱۲]. روش‌های مبتنی بر Wrapper از فاز یادگیری استفاده می‌کنند و نسبت به روش‌های مبتنی بر Filter عملکرد بهتری دارند. هزینه محاسباتی الگوریتم‌های مبتنی بر Wrapper به دلیل استفاده از فاز یادگیری بیشتر از روش‌های مبتنی بر Filter است. الگوریتم‌های یادگیری بسیاری جهت ارزیابی ویژگی‌های انتخاب شده در روش‌های مبتنی بر Wrapper ارائه شده است [۱۳]. در روش‌های مبتنی بر Embedded از ترکیب دو روش Filter و Wrapper استفاده شده است [۱۴].

الگوریتم‌های تکاملی مبتنی بر روش Wrapper، به دلیل قدرت بهینه‌سازی جهانی خود مورد توجه محققین قرار گرفته اند. از جمله الگوریتم‌های تکاملی که در روش Wrapper مورد استفاده قرار گرفته‌اند می‌توان به الگوریتم‌های ژنتیک [۱۵]، باکتری [۱۶] وPSO[۳] [۱۷] اشاره کرد. الگوریتم‌های تکاملی بدون توجه به فضای مسئله به صورت فرااکتشافی اقدام به یافتن جواب می‌کنند. الگوریتم‌های تکاملی غالبا به صورت تک هدفه یا چند هدفه مورد استفاده قرار می‌گیرند. مسئله انتخاب ویژگی دارای دو هدف است؛ که یکی از اهداف کاهش خطای الگوریتم یادگیری و دیگری کاهش تعداد ویژگی‌های انتخاب شده توسط الگوریتم تکاملی است [۱۸].

یکی از مهم‌ترین چالش‌های الگوریتم‌های تکاملی، گرفتار شدن در نقاط بهینه محلی است. بنابراین الگوریتم‌های تکاملی برای یافتن نقاط نزدیک به بهینه جهانی باید از مکانیزمی برای خروج از نقاط بهینه محلی استفاده کنند. یکی از این مکانیسم‌ها استفاده از روش جهش است. در مکانیسم جهش، مقدار ویژگی‌ها برای تولید جواب‌های جدید تغییر می‌کند. از آنجایی که این مکانیسم به طور تصادفی مقادیر ویژگی‌ها را تغییر می‌دهد، هیچ تضمینی برای یافتن پاسخ بهینه وجود ندارد.

به همین دلیل، این مقاله یک روش جهش هوشمند را ارائه می کند که از فضای مسئله نیز استفاده می‌کند. در این روش کیفیت هر ویژگی مستقیماً بر احتمال جهش آن تأثیر می گذارد.

یکی دیگر از چالش‌های الگوریتم‌های تکاملی وجود پارامترهای زیادی برای بهینه‌سازی الگوریتم برای مسائل خاص است. وجود پارامترهای زیاد باعث می‌شود که الگوریتم برای همه مسائل بهینه‌سازی به خوبی کار نکند. بنابراین، این مقاله یک روش جدید بر اساس الگوریتم باکتری ارائه می‌دهد. دلیل انتخاب الگوریتم باکتری، تعداد پارامترهای کمتر آن به دلیل الهام گرفتن از رفتار آن از طبیعت است.

در این مقاله، روش جدیدی از نوع Wrapper مبتنی بر الگوریتم باکتری ارائه شده است. در این روش، با الهام گرفتن از عملکرد باکتری‌ها دو عملگر جهش[۴] درون سلولی و لوله جنسی[۵] معرفی شده است. باکتری‌ها دارای عملگر برش[۶] جهت ایجاد فرزند جدید که در آن دو والد شرکت دارند، نیستند. باکتری‌ها معمولا چندین کپی از خود ایجاد می‌کنند که در آن‌ها عمل جهش اتفاق می‌افتد. در عمل جهش برخی از DNAها تغییر می‌کند که ممکن است باعث مقاومت باکتری دربرابر آنتی‌بیوتیک‌ها شود. همچنین باکتری‌های برتر از طریق لوله جنسی، برخی از DNAهای خود را به باکتری دیگر انتقال می‌دهند. در الگوریتم پیشنهادی نیز باکتری‌هایی که دارای تعداد ویژگی و خطای کمتری هستند،

برخی از ویژگی‌های برتر خود را به باکتری دیگر انتقال می‌دهند. از مکانیزم آرشیو جهت نگهداری باکتری‌های برتر و امتیازدهی به ویژگی‌های مسئله استفاده شده است. نوآوری‌های الگوریتم پیشنهادی به شرح ذیل است:

- الگوریتم جدید مبتنی بر باکتری
- جهش هوشمند جدید به همراه نخبه‌گرایی

ادامه مقاله به شرح ذیل تنظیم شده است. در بخش دوم، به بررسی کارهای پیشین و پیش‌زمینه‌های لازم می‌پردازیم. در بخش سوم الگوریتم ارائه شده به تفصیل توضیح داده خواهد شد. در بخش چهارم، نتایج شبیه‌سازی به همراه دیتاست‌های مورد استفاده معرفی خواهد شد. و در انتها، در بخش پنجم به نتیجه‌گیری خواهیم پرداخت.

۲- کارهای پیشین و پس‌زمینه

در این قسمت ابتدا پیش‌زمینه الگوریتم پایه باکتری معرفی می‌شود. سپس انواع روش‌های Wrapper بیان شده و در انتها کارهای پیشین بررسی خواهد شد.

۲-۱- الگوریتم پایه باکتری

بسیاری از الگوریتم‌های تکاملی از رفتارهای موجود در طبیعت الهام گرفته‌اند. باکتری‌ها نیز از جمله ساده‌ترین موجودات تک سلولی هستند که الگوهای رفتاری ساده‌ای دارند و به راحتی قابل توصیف هستند. باکتری‌ها دارای قدرت بالایی در تطبیق‌پذیری با محیط پیچیده اطراف خود هستند و غریزه بهینه‌سازی آن‌ها در جهت زنده ماندن در این محیط را نشان می‌دهند. این دو ویژگی باعث شده محققان الگوریتم‌های تکاملی بسیاری مبتنی بر رفتار باکتری‌ها ارائه کنند. باکتری‌ها دارای سه رفتار عمده با نام‌های: کموتاکسی[۷]، تولید مثل[۸] و حذف و پراکندگی [۹]هستند [۱۹].

الگوریتم پایه باکتری [۲۰] به صورت زیر خلاصه شده است:

۱) مقداردهی اولیه: در الگوریتم پایه باکتری، جمعیت اولیه به صورت تصادفی تولید می‌شود. سپس مقدار برازش هر باکتری با توجه به مسئله مورد نظر محاسبه شده و بهترین باکتری انتخاب می شود.

۲) عملگر کموتاکسی: تمامی باکتری‌ها سعی دارند از محیط ناامن دوری کرده و به سمت محیط مورد علاقه خود حرکت کنند. در طی فرآیند کموتاکسی، حرکت باکتری‌ها شامل دو حالت غلت زدن [۱۰]و شنا کردن [۱۱]است. غلت زدن به معنی حرکت باکتری در هر مسیر ممکن است و در صورتی که مقدار برازش آن بیشتر از وضعیت قبلی بود، در جهت مسیر جدید حرکت خواهد کرد.

۳) عملگر تولید مثل: قانون بقا بیانگر این موضوع است که هر گونه‌ای که بهتر باشد زنده خواهد ماند. پس از فرآیند کموتاکسی، باکتری‌هایی که دارای برازش کمتری هستند، از بین خواهند رفت و باکتری‌هایی که قوی‌تر هستند باقی خواهند ماند و جمعیت بعدی را با تقسیم خود به دو گونه یکسان تشکیل خواهند داد. این امر باعث می‌شود کیفیت جمعیت در تکرارهای بعدی حفظ شود.

۴) حذف و پراکندگی: در طی فرآیندهای کموتاکسی و تولید مثل، محیط بقای جمعیت دچار آسیب می شود و برخی از باکتری‌ها از بین رفته و برخی دیگر به محیط‌های دیگر مهاجرت می‌کنند. عملگر حذف و پراکندگی قابلیت کنترل قدرت اکتشاف و خروج از بهینه محلی را کنترل می‌کند.

۲-۲- روش Wrapper

روش‌های مبتنی بر Wrapper از جمله روش‌های مبتنی بر ویژگی هستند که از الگوریتم‌های یادگیری در ارزیابی ویژگی‌های انتخاب شده استفاده می‌کنند. اجزای اصلی و تشکیل دهنده این روش، الگوریتم یادگیری و استراتژی جستجو است. انتخاب ترتیبی و الگوریتم‌های تکاملی از انواع روش‌های استراتژی جستجو هستند که در الگوریتم تکاملی از الگوریتم‌های یادگیری جهت ارزیابی ویژگی‌های انتخاب شده، استفاده شده است [۲۱]. در ادامه به بررسی استراتژی‌های انتخاب ترتیبی و الگوریتم‌های تکاملی در روش Wrapper می‌پردازیم.

۲-۲-۱- روش انتخابی

روش‌های انتخاب ویژگی به سه دسته انتخاب رو به جلو، انتخاب رو عقب و انتخاب دوطرفه دسته‌بندی می‌شوند. در روش رو به جلو، ویژگی‌ها انتخاب و به ویژگی‌های انتخاب شده قبلی اضافه خواهند شد و بر اساس آن تابع هدف بهینه می‌شود. در روش رو به عقب، ویژگی‌ها برخلاف روش رو به جلو حذف خواهند شد. در روش دوطرفه نیز هر دو عمل حذف و اضافه شدن انجام خواهد شد. روش‌های ذکر شده، از مشکل "اثر تودرتو[۱۲]" رنج می‌برند. در مشکل اثر تودرتو، در روش‌های رو به جلو، ویژگی‌هایی که پیش‌تر انتخاب نشده بوند، دیگر شانس انتخاب مجدد نخواهند داشت، در روش رو به عقب نیز، ویژگی‌هایی که انتخاب می‌شوند دیگر امکان حذف شدن نخواهند داشت [۲۲]. برای حل مشکل اثر تودرتو، دو روش SBFS (Sequential Backward Floating

(Selection) و SFFS (Sequential Forward Floating Selection) ارائه شده‌اند [۲۳].

۲-۲-۲- روش الگوریتم‌های تکاملی

الگوریتم‌های تکاملی از رفتار طبیعت الهام گرفته‌اند و در هر نسل سعی در بهبود جمعیت دارند. این الگوریتم‌ها دارای تابع هدف هستند که باعث می‌شود الگوریتم در راستای اهداف مورد نظر حرکت کند. از آنجایی که این الگوریتم‌ها به مسائل به صورت جعبه سیاه نگاه می‌کنند، به راحتی در تمامی مسائلی که قابلیت تعریف تابع هدف دارند، کاربردی هستند. توابع هدف در این الگوریتم‌ها به صورت تک‌هدفه یا چندهدفه درنظر گرفته می‌شود. الگوریتم‌های تکاملی در حل مسئله انتخاب ویژگی نیز کاربرد دارند، اما به دلیل استفاده از فرآیند یادگیری جهت ارزیابی ویژگی‌های انتخاب شده، دارای هزینه محاسباتی و زمانی بالایی هستند.

۲-۳- کارهای مرتبط

الگوریتم‌های بسیاری برای حل مسئله انتخاب ویژگی ارائه شده‌اند. در این قسمت به بررسی برخی از جدیدترین روش‌ها می پردازیم. الگوریتم CMDPFSOFS[۱۳] [۲۴] یک روش مبتنی بر Wrapper است که از دو عملگر جهش یکنواخت و غیر یکنواخت استفاده کرده است. این دو عملگر باعث حفظ تنوع و بهبود جستجوی الگوریتم شده است. در این الگوریتم ذرات به سه دسته مساوی تقسیم شده‌اند که بر روی دو دسته از سه دسته، این دو عملگر انجام می‌شود. در الگوریتم HMDPSOFS [۱۴][۲۵] از یک جهش ترکیبی که در آن در هر تکرار از الگوریتم ده درصد ذرات اجازه دارند مقدار سرعت خود را به مقدار اولیه تنظیم کنند، استفاده شده است. برای افزایش قدرت بهینه‌سازی الگوریتم، جهش پرشی[۱۷] ارائه شده است، که در آن ذرات با احتمال یکنواخت امکان جهش به مکان دیگر دارند. الگوریتم ISRPSO [۱۵][۲۶]، از روش جستجوی محلی جهت بهبود آرشیو (بهترین جواب‌ها در آرشیو نگهداری می‌شود) استفاده می‌کند. این الگوریتم از سه روش درج کردن، مبادله کردن [۱۶]و حذف کردن جهت بهبود آرشیو استفاده می‌کند. در روش درج کردن، ویژگی های موجود در آرشیو بر اساس دقت الگوریتم مرتب شده و آن‌هایی که دارای اهمیت بالایی هستند برای ایجاد جواب جدید استفاده می‌شوند. در روش حذف، الگوریتم حذف ویژگی‌ها را تا جایی که دقت الگوریتم افزایش یابد ادامه می‌دهد. در روش مبادله نیز الگوریتم ویژگی‌ها را با ویژگی‌های دارای رتبه بالا در آرشیو مبادله می‌کند. الگوریتم RSPSOFS [۱۷][۲۷]، یک روش جدید است که در آن از مکانیزم آرشیو جهت ذخیره کردن بهترین جواب‌ها استفاده شده است. ویژگی‌ها بر اساس تعداد تکرارشان در آرشیو رتبه‌بندی می‌شوند. در این روش ذرات به سه دسته مساوی تقسیم شده و بر روی دو دسته از سه دسته، دو عملگر جهش یکنواخت و غیر یکنواخت انجام می‌شود. در IIEFS[۱۸] [۲۸]، نویسندگان از نظریه اطلاعات برای انتخاب ویژگی استفاده کرده‌اند. این الگوریتم از دو گام تشکیل شده است. در گام اول، ابتدا مقدار برازش هر ویژگی محاسبه شده و ویژگی‌هایی که دارای برازش بیشتری هستند به عنوان کاندیدا انتخاب شده و یک مجموعه را به‌وجود می‌آورند. سپس ذرات به صورت جفت انتخاب شده و مقدار تعادل آن‌ها محاسبه می‌شود. آن جفت‌هایی که دارای تعادل بیشتری باشند به مجموعه جدید منتقل می‌شوند. در گام دوم نیز بر اساس مجموعه قبلی فرآیندها مجدد تکرار می شوند تا مجموعه جفت‌های جدید با مقدار برازش بالا ایجاد شوند. این مقدار تا اندازه محدوده مشخص شده ادامه می‌یابد و مجموعه ویژگی‌های انتخاب شده ایجاد می‌شود.

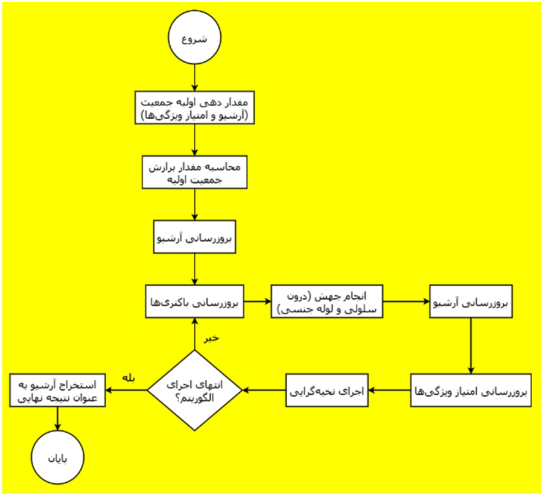
۳- روش پیشنهادی

مسئله انتخاب ویژگی یک مسئله چند هدفه است که در آن اهداف، کاهش تعداد ویژگی و افزایش دقت الگوریتم یادگیری است. در روش پیشنهادی از جبهه پرتو برای حل این مسئله استفاده شده است. روش پیشنهادی یک روش مبتنی بر باکتری است که در آن سعی شده وابستگی به پارامترهای تنظیم کننده کاهش یابد و الگوریتم برای مسائل مختلف با کمترین تغییرات قابل استفاده باشد. الگوریتم پیشنهادی مبتنی بر Wrapper بوده و از الگوریتم یادگیری ماشین جهت ارزیابی مجموعه ویژگی‌های انتخاب شده استفاده خواهد نمود. مکانیزم آرشیو جهت نگهداری باکتری‌های برتر درنظر گرفته شده است. ویژگی‌های موجود در آرشیو بر اساس تعداد تکرار رتبه دهی می‌شوند. گام‌های روش پیشنهادی در جدول (۱) و شکل (۱) به ترتیب آورده شده است.

جدول(۱): گام‌های الگوریتم پیشنهادی

تقسیم داده‌های دیتاست به آموزشی و تست و سپس اعمال گام‌های زیر بر روی داده‌های آموزشی:
گام ۱: مقداردهی اولیه جمعیت بر اساس قسمت ۳-۱.
گام ۲: ارزیابی مقدار برازش جمعیت
گام ۳: بروزرسانی آرشیو براساس مقدار برازش جمعیت
گام ۴: تولید مثل باکتری‌ها
گام ۴: بروزرسانی موقعیت باکتری‌ها براساس (۶)
گام ۵: اعمال جهش هوشمند درون سلولی و لوله جنسی و بروزرسانی باکتری‌ها
گام ۶: بروزرسانی آرشیو
گام ۷: محاسبه و بروزرسانی امتیاز هر ذره بر اساس آرشیو
گام ۸: اعمال عملگر نخبه‌گرایی بر روی جمعیت آرشیو
گام ۹: اگر الگوریتم به بیشینه مقدار تکرار رسیده بود به گام ۱۰ برو در غیر اینصورت به گام ۴ برگرد.

گام ۱۰: استخراج اعضای موجود در آرشیو به عنوان بهترین جواب‌های روش پیشنهادی



شکل ۱: فلوچارت روش پیشنهادی

۳-۱- کدگذاری ذرات

موقعیت باکتری‌ها در روش پیشنهادی به صورت بازه پیوسته بین صفر و یک در نظر گرفته شده است که در رابطه(۱) نشان داده شده است.

$$X_i(t) = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iD})$$
$$x_{ij} \in [0,1], j = 1,2,3,\ldots,D \quad (۱)$$

که در آن $X_i(t)$ نشان دهنده موقعیت باکتری در تکرار t، x_{ij} نشان دهنده ویژگی j ام ذره i ام و D تعداد ویژگی‌های مسئله است. مقدار x_{ij}ها نیز بین اعداد صفر و یک است. از آنجایی که انتخاب ویژگی به صورت دودویی است، نیاز است اعداد پیوسته به دو عدد صفر و یک نگاشت شوند. برای این کار از یک مقدار آستانه با نام θ استفاده می‌کنیم که مقدار آن برابر ۰،۶ است. در نتیجه، x_{ij}هایی که مقدار آن‌ها بیشتر از مقدار آستانه باشد، برابر یک و در غیر اینصورت برابر صفر در نظر گرفته می‌شوند. نسخه دودویی $X_i(t)$ در رابطه (۲) بیان شده است.

$$Z_i(t) = (z_{i1}, z_{i2}, z_{i3}, \ldots, z_{iD})$$
$$z_{i,j} \in \{0,1\}, j = 1,2,3,\ldots,D \quad (۲)$$

که در آن $Z_i(t)$ نسخه دودویی $X_i(t)$ ، z_{ij} نشان‌دهنده انتخاب یا عدم انتخاب ویژگی j ام ذره i ام و D تعداد ویژگی‌های مسئله است.

۳-۲- ارزیابی برازش باکتری

در مسئله انتخاب ویژگی که یک مسئله چندهدفه است، یکی از اهداف افزایش دقت الگوریتم یادگیری است. در داده‌های متوازن معیار دقت خطای الگوریتم یک معیار مناسب بشمار می‌رود. اما در داده‌های نامتوازن که تعداد داده های یک کلاس بیشتر از سایر کلاس‌ها است، این معیار باعث می‌شود الگوریتم یادگیری به سمت آموختن کلاس با داده‌های بیشتر همگرا شود. به‌همین دلیل معیارهای دیگری برای حل مسئله با داده‌های متوازن ارائه شده است. از جمله این روش‌ها، می‌توان به F-measure اشاره نمود که در رابطه (۳) آورده شده است.

$$F-Measure = \frac{(2\times Precision \times Recall)}{(Precision + Recall)}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FP} \quad (۳)$$

که در آن TP مثبت واقعی، FP منفی واقعی و FN منفی کاذب است. در رابطه (۴) نحوه محسابه تعداد ویژگی انتخاب شده، آورده شده است.

$$Feature_{rate}(i) = \frac{f_i}{D}, f_i = \sum_{j=1}^{D} z_{ij} \quad (۴)$$

که در آن f_i تعداد ویژگی‌های انتخاب شده و D تعداد کل ویژگی‌های مسئله است.

برای ارزیابی نهایی باکتری از روش جبهه پرتو استفاده شده است که به صورت رابطه (۵) تعریف می‌شود.

$$x \in pareto\ front$$
$$if \begin{Bmatrix} f_i \le f_i(x') & \forall i \\ f_i < f_i(x') & for\ last\ one \end{Bmatrix} \quad (۵)$$

۳-۳- آرشیو

در دنیای واقعی باکتری‌های برتر همواره در محیط خود زنده می‌مانند و بر روی سایر باکتری‌ها تاثیر می‌گذارند. برای شبیه‌سازی این مکانیزم، از آرشیو استفاده شده است. باکتری‌هایی که دارای برازش بیشتری نسبت به سایر باکتری‌ها هستند در آرشیو نگهداری می‌شوند. اندازه آرشیو محدود است؛ بنابراین در صورتی که اعضای آرشیو به مقدار بیشینه اندازه آرشیو برسد، اعضایی که دارای برازش کمتری هستند با اعضایی که دارای برازش بیشتری هستند، جایگزین می‌شوند. این فرآیند تضمین می‌کند بهترین جواب‌ها در طول اجرای الگوریتم نگهداری شوند.

۴-۳- امتیازدهی به ویژگی‌ها

در مسئله انتخاب ویژگی، هدف یافتن بهترین ویژگی‌هایی است که منجبر به افزایش دقت الگوریتم یادگیری شود. جهت رسیدن به این هدف ویژگی‌ها بر اساس تعداد مشاهده در باکتری‌های موجود در آرشیو، رتبه‌دهی می‌شوند. در رابطه (۶) نحوه محاسبه رتبه هر ویژگی آورده شده است:

$$FR=\sum_{k=1}^{|A|} Z_k(t), Z_k(t)\in A \qquad (۶)$$

که در آن FR یک بردار با D بعد که در آن $FR=[r_1,r_2,r_3,\ldots,r_D]$ مقدار امتیاز هر ویژگی، A نیز نمایانگر باکتری‌های موجود در آرشیو و $Z_k(t)$ مقدار دودویی ویژگی در تکرار t است. در صورتی که ویژگی در آرشیو انتخاب نشده باشد مقدار آن برابر صفر خواهد بود. بیشینه مقدار امتیاز هر ویژگی برابر تعداد اعضای آرشیو است.

۵-۳- تولید مثل باکتری‌ها

باکتری‌ها بر خلاف سایر موجودات دارای جنسیت نیستند و برای تولید مثل یک نمونه از خود را تکثیر می‌کنند. در دنیای واقعی در صورتی که باکتری‌های مضر به موقع شناسایی نشوند، با سرعت نمایی رشد خواهند نمود. این مکانیزم نیز در الگوریتم باکتری درنظر گرفته شده است. پس از هر تکرار الگوریتم، تعداد باکتری‌ها دو برابر خواهد شد.

۶-۳- بروزرسانی موقعیت باکتری‌ها

باکتری‌ها در بدن میزبان به سمت یافتن مواد مغذی حرکت می‌کنند. برای پیاده‌سازی این مکانیزم در الگوریتم باکتری، از تابع حرکت استفاده شده است. حرکت باکتری با سه عامل مرتبط است. عامل اول موقعیت قبلی باکتری، عامل دوم موقعیت برترین باکتری موجود در محیط و عامل سوم موقعیت تصادفی است. این سه عامل باهم ترکیب شده و موقعیت جدید باکتری را مشخص می‌کند. نحوه محاسبه موقعیت جدید باکتری در رابطه (۷) آورده شده است.

$$x^{t+1}=w1\times R+w2\times x_{best}^{t}+w3\times x^{t} \qquad (۷)$$

که در آن $w1$، $w2$ و $w3$ به ترتیب وزن‌های عامل تصادفی، عامل موقعیت برترین باکتری و عامل موقعیت فعلی باکتری است. R بردار اعداد تصادفی تولید شده بین ۰،۶ و -۰،۶، x_{best}^{t} موقعیت برترین باکتری و x^{t} موقعیت فعلی باکتری است.

۷-۳- جهش هوشمند باکتری‌ها

پس از اینکه باکتری‌ها حرکت خود را انجام دادند، باکتری‌ها به کمک الگوریتم انتخاب تورنومنت به دو دسته مساوی تقسیم می‌شوند. دسته اول بدون اینکه تغییری بر روی آن‌ها انجام شود به مرحله بعدی می‌روند. اما در دسته دوم عملگر جهش هوشمند انجام می‌شود. عملگر جهش خود به دو صورت درون سلولی و لوله جنسی امکان پذیر است. برای اینکار الگوریتم یک عدد تصادفی تولید کرده و در صورتی که عدد تولید شده از یک مقدار آستانه با نام زیگما بیشتر بود، جهش درون سلولی اتفاق می‌افتد و در غیر اینصورت از طریق لوله جنسی تعدادی DNA توسط بهترین باکتری به باکتری انتخاب شده منتقل می‌شود. الگوریتم جهش درون سلولی و لوله جنسی به ترتیب در الگوریتم (۱) و (۲) آورده شده است.

الگوریتم (۱): روش جهش درون سلولی

Input: *bacteria, features rank*
Output: *bacteria*
Begin
for i ← ۱ **to** D **do**
If $random,uniform(0,1)\le P_{innercell}$ **and** $feature_{rank(i)}\le 0.5$
If $bacteria.position_{(i)}\ge C$
$bacteria.position(i)\leftarrow bacteria.position(i)-\theta$
else
$bacteria.position(i)\leftarrow bacteria.position(i)+\theta$
end
end
end
end

الگوریتم (۲): روش جهش لوله جنسی

Input: *bacteria, best bacteria, features rank*
Output: *bacteria*
Begin

$Ones_{bestbacteria} \leftarrow$ array of features index which values are one in best bacteria

$Ones_{length} \leftarrow$ length of $Ones_{bestbacteria}$

If $random, uniform(0,1) \le P_{sexpilus}$

$Ones_{bacteria} \leftarrow$ array of features index which values are one in bacteria

Shuffle $Ones_{bacteria}$

for i **to** $Ones_{bacteria}$

if

$bacteria.position(Ones_{bacteria(i)}) \ge \theta$

and $i \le Ones_{bacteria}$

$bacteria.position(Ones_{bacteria}(i)) \leftarrow 1-\theta$

end

If $features_{rank}(Ones_{bestbacteria}(i)) \le 0.5$

$bacteria.position(One_{bestbacteria}(i)) \leftarrow \theta$

end

end

end

end

۳-۸- تزریق آنتی‌بیوتیک

در دنیای واقعی با تزریق آنتی‌بیوتیک باکتری‌هایی که ضعیف هستند از بین می‌روند. این مکانیزم در الگوریتم باکتری پیشنهادی در نظر گرفته شده است. در این مکانیزم از آنجایی که در فاز تولید مثل باکتری‌ها دو برابر شده بودند، نصف باکتری‌هایی که دارای برازش کمتری هستند، از بین می‌رود. این فرآیند باعث می‌شود در انتهای گام‌های الگوریتم میزان جمعیت باکتری‌ها به مقدار اولیه بازگردد.

۳-۹- نخبه‌گرایی

هدف از نخبه‌گرایی ایجاد باکتری‌هایی است که مقاومت آن‌ها در برابر آنتی‌بیوتک افزایش یابد. در این فاز از یک مکانیزم هوشمند استفاده شده است که در آن ابتدا ویژگی‌ها بر اساس امتیازشان رتبه‌بندی می‌شوند و بر اساس رابطه (۸) مجموعه ویژگی‌های برتر انتخاب می‌شوند.

$$BFR = FR(i), FR(i) \ge \max(\frac{FR}{2}) \quad (۸)$$

که در آن BFR لیست بهترین ویژگی‌ها و $FR(i)$ مقدار امتیاز ویژگی i ام است. سپس، این ویژگی‌ها به صورت تصادفی با یک چهارم ویژگی‌هایی که مقدار آنها برابر یک است، جایگزین می شوند. در صورتی که مقدار برازش ذره جدید ایجاد شده از ذره قبلی بهتر بود، با آن ذره تعویض می‌شود. این فرآیند بر روی تمامی اعضای آرشیو اجرا می‌شود. شبه‌کد نخبه‌گرایی در الگوریتم ۳ آورده شده است.

الگوریتم (۳): روش نخبه‌گرایی هوشمند

Input: *archive, features rank*
Output: *updated archive*
Begin

$bestrank_{index} \leftarrow$ select fr_i where fr_i greater than $\max(\frac{FR}{2})$ ، fr_i is feature rank of *jth* feature

$x'_{ij} \leftarrow$ copy of x_{ij} ، x_{ij} is *jth* feature of particle *i* in archive

for i ← ۱ **to** size of archive **do**

$ones_{index} \leftarrow x'_{ij}$ where x'_{ij} greater than θ

$ones_{length} \leftarrow$ sum of x'_{ij} where x'_{ij} greater than θ

$randoms_{ones} \leftarrow$ select ¼ of $ones_{index}$ randomly

for k ← ۱ to size of $randoms_{ones}$ **do**

$x'_{i,randoms_{ones}(k)} \leftarrow 0$

end

for m ← ۱ to size of $\frac{1}{4} ones_{index}$ **do**

$x'_{i,ones_{index}(m)} \leftarrow 1$

end

end

calculate fitness of x'_i

if fitness of x'_i is better than x_i **then**

$x_i \leftarrow x'_i$

end

end

end

end

۴- نتایج شبیه‌سازی

در این بخش به بررسی نتایج شبیه‌سازی روش پیشنهادی، معرفی دیتاست‌های استفاده شده و پارامترهای شبیه‌سازی می پردازیم.

۴-۱- دیتاست

دو دسته دیتاست برای مقایسه روش پیشنهادی با روش‌های موجود ارائه شده است. دیتاست‌های دسته اول، شامل داده‌های متوازن و دیتاست‌های دسته دوم شامل داده‌های نامتوازن هستند. در ادامه دیتاست‌های مورد استفاده معرفی خواهند شد.

۴-۱-۱- دیتاست‌های متوازن

جدول (۲): توضیحات دیتاست متوازن

نام دیتاست	تعداد ویژگی‌ها	تعداد داده‌ها	تعداد کلاس
Isolet	۶۱۷	۷۷۹۷	۲۶
Musk1	۱۶۷	۶۵۹۸	۲
Madelon	۵۰۰	۴۴۰۰	۲
Arrhythmia	۲۷۹	۴۲۵	۱۶
Cnae	۸۵۷	۱۰۸۰	۹
Hillvalley	۱۰۰	۶۰۶	۲
LSVT	۳۰۹	۱۲۶	۲

۴-۱-۲- پارامترها

در روش پیشنهادی که مبتنی بر رفتار باکتری‌ها است، سعی شده از کمترین تعداد پارامترها استفاده شود تا الگوریتم قدرت حل تمامی مسائل بهینه‌سازی بدون نیاز به بهینه کردن پارامترها را داشته باشد. یکی از پارامترهایی که با آزمایش مقادیر مختلف مشخص شد، مقدار θ است. در جدول ۳ نتایج آزمایش های مذکور آورده شده است. بر طبق این آزمایش مقدار θ برابر ۰،۸۵ درنظر گرفته شده است.

جدول (۳): آزمایش مقدار θ بر روی دیتاست CNAE

مقدار θ	بهترین جواب اول		بهترین جواب دوم		بهترین جواب سوم	
	تعداد ویژگی	میزان خطا	تعداد ویژگی	میزان خطا	تعداد ویژگی	میزان خطا
۰،۶	۳۱۶	۰،۲	۳۰۱	۰،۲۱	۱۴۰	۰،۳۴
۰،۶۵	۱۲۴	۰،۳۵	۱۲۷	۰،۳	۱۱۷	۰،۲۹
۰،۷۰	۸۹	۰،۳	۱۱۶	۰،۳۱	۸۶	۰،۳۴
۰،۷۵	۱۱۲	۰،۲۹	۱۳۶	۰،۳۴	۸۹	۰،۳۹
۰،۸۰	۵۹	۰،۳۳	۳۷	۰،۳۳	۵۳	۰،۳۳
۰،۸۵	۴۶	۰،۳۳	۴۰	۰،۳۵	۲۶	۰،۳۹

پارامتر دیگری که از طریق آزمایش مشخص شد، پارامتر $P_{innercell}$ است. ما مقادیر مختلفی را بر روی دیتاست Caltech101 آزمایش کردیم جدول ۴ نمایش دهنده مقادیر آزمایش شده است. بر این اساس، مقدار $P_{innercell}$ برابر ۰،۲ و مقدار $P_{s\exp ilus}$ که همان برابر ۰،۸ در نظر گرفته شده است.

جدول (۴): آزمایش مقدار $P_{innercell}$ بر روی مجموعه داده Caltech101

مقدار $P_{innercell}$	بهترین جواب اول		بهترین جواب دوم		بهترین جواب سوم	
	تعداد ویژگی	میزان خطا	تعداد ویژگی	میزان خطا	تعداد ویژگی	میزان خطا
۰،۱	۴۹	۰،۵۴	۵۵	۰،۵۳	۲۵	۰،۶۱
۰،۲	۵۰	۰،۵۴	۴۴	۰،۵۵	۱۸	۰،۶۶
۰،۳	۴۸	۰،۵۳	۵۸	۰،۵۲	۲۴	۰،۶۴
۰،۴	۵۳	۰،۵۳	۵۵	۰،۵۳	۲۴	۰،۶۱
۰،۵	۳۹	۰،۵۴	۵۳	۰،۵۳	۲۶	۰،۶۱
۰،۶	۴۱	۰،۵۳	۴۶	۰،۵۳	۳۷	۰،۵۶
۰،۷	۴۶	۰،۵۳	۵۹	۰،۵۲	۳۱	۰،۵۷
۰،۸	۴۱	۰،۵۳	۵۰	۰،۵۳	۳۳	۰،۵۹
۰،۹	۵۹	۰،۵۱	۴۸	۰،۵۳	۳۲	۰،۵۹

سایر پارامترهای شبیه‌سازی در جدول (۵) آورده شده است.

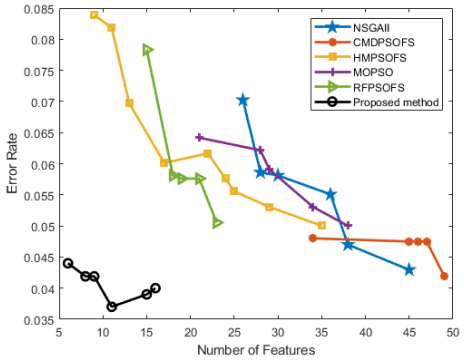
جدول (۵): مقداردهی پارامترهای الگوریتم پیشنهادی

نام پارامتر	مقدار پارامتر
W1	۰،۱۵
W2	۰،۳۵
W3	۰،۵
$P_{s\exp ilus}$	۰،۸
$P_{innercell}$	۰،۲
اندازه جمعیت باکتری‌ها	۱۵
اندازه آرشیو	۶
بیشینه تکرار الگوریتم	۱۰۰

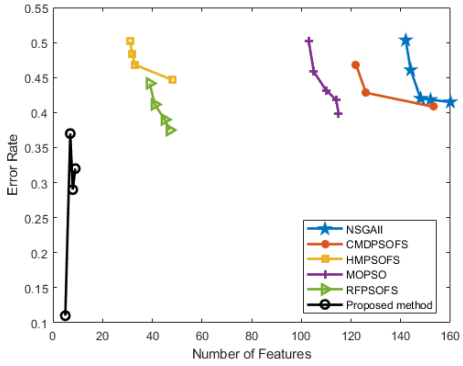
در سایر روش‌های مبتنی بر PSO، سایز جمعیت برابر ۳۰، بیشینه مقدار تکرار برابر ۱۰۰، بیشینه مقدار سرعت یا همان $v\_\max$ برابر ۰،۶، مقادیر $b1$ و $b2$برابر ۰،۱۴۶، مقدار w برابر ۰،۷۲۹ و مقدار θ برابر ۰،۶ در نظر گرفته شده است. پارامترهای مخصوص هر روش بر اساس گزارشات آن‌ها تنظیم شده است. احتمال پرش در روش HMPSOFS برابر ۰،۰۱ است. در روش CMDPSOFS وزن اینرسی به صورت تصادفی بین ۰،۱ و ۰،۵، مقادیر $b1$ و $b2$ اعداد تصادفی بین ۱،۵ و ۲ و همچنین نرخ جهش در این روش برابر $\frac{1}{D}$که در آن D تعداد ویژگی‌های دیتاست است، در نظر گرفته شده است. برای NSGAII نیز پارامترها مشابه روش CMDPSOFS و نرخ برش برابر ۰،۹ است. لازم به ذکر است که در روش NSGAII کروموزوم‌ها به صورت یک آرایه d بعدی است که اگر مقدار آن برابر یک باشد به معنی انتخاب ویژگی و اگر مقدار آن برابر صفر باشد، به معنی عدم انتخاب ویژگی است [۲۹].

۴-۱-۳- نتایج شبیه‌سازی

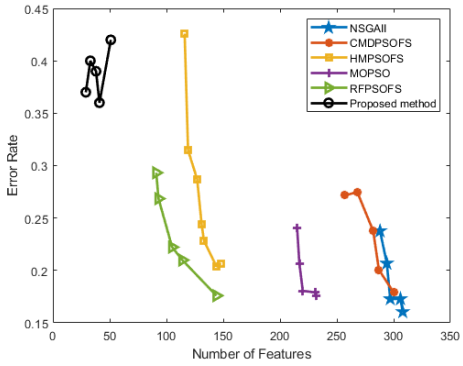
روش پیشنهادی در ۷ دیتاست با نام‌های "isolet"، "Cnae"، "Musk1"، "Madelon"، "Arrhythmia"، "Hillvalley" و "LSVT" با روش‌های CMDPSOFS، HMPSOFS و RFPSOFS که سه روش موثر و جدید هستند و همچنین، با روش‌های NSGAII [۲۹] و MOPSO [۳۰] که جزو روش‌های پایه هستند، مقایسه شده است. میانگین نتایج شبیه‌سازی با ۳۰ اجرا در شکل‌های (۲) تا (۸) آورده شده است.



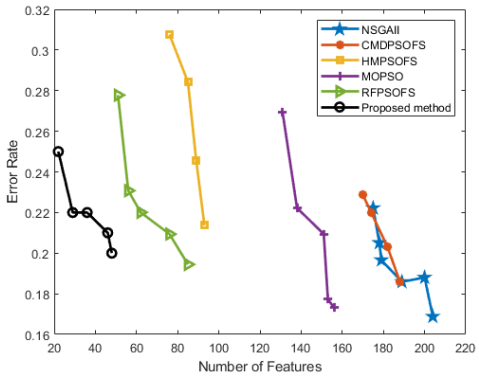
شکل ۲ :نتایج شبیه‌سازی در دیتاست musk



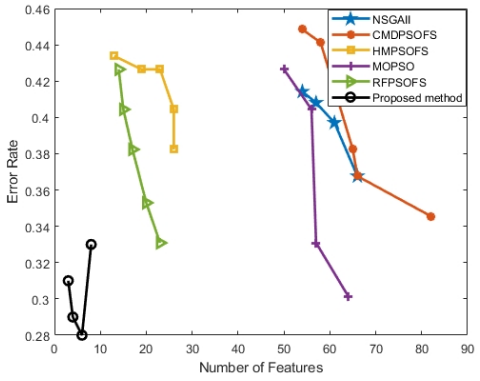
شکل ۳: نتایج شبیه‌سازی در دیتاست Madelon



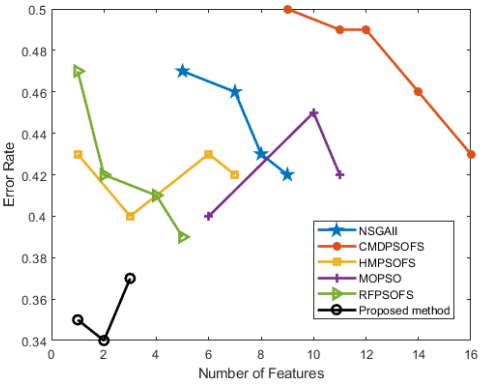
شکل ۴: نتایج شبیه‌سازی در دیتاست Cnae



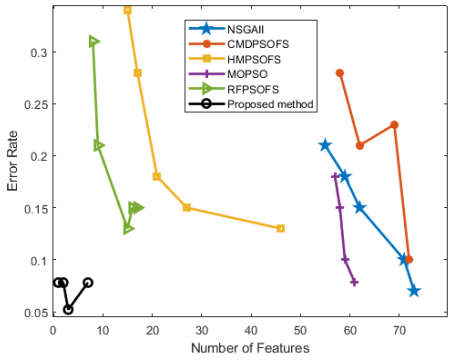
شکل ۵: نتایج شبیه‌سازی در دیتاست Isolte



شکل ۶: نتایج شبیه‌سازی در دیتاست Arrhythmia



شکل ۷: نتایج شبیه‌سازی در دیتاست Hill



شکل ۸: نتایج شبیه‌سازی در دیتاست LSVT

باتوجه به شکل‌ها روش پیشنهادی توانسته تعداد ویژگی‌های کمتری را به نسبت سایر روش‌ها انتخاب کند. همچنین این کاهش تعداد ویژگی تاثیر منفی در خطای الگوریتم یادگیری نداشته است. دلیل این برتری، استفاده از الگوریتم باکتری در کنار جهش و نخبه‌گرایی هوشمند است. استفاده از جهش هوشمند باعث شده الگوریتم در تعداد اجراهای بالا با استفاده از دانش قبلی خود که از فضای مسئله کسب کرده و در آرشیو وجود دارد، بتواند عمل جهش خود را به صورت هوشمند انجام داده و ویژگی با امتیاز بالا را با ویژگی با امتیاز پایین جایگزین کند. این تغییرات باعث می‌شود الگوریتم پیشنهادی ذرات با مقدار برازش بهتری را تولید کند و کیفیت جمعیت را در اجراهای بالا بهتر کند. الگوریتم با گذر زمان امتیاز ذرات را با توجه به برازش ذرات موجود در آرشیو بهبود می‌بخشد. بنابراین، ذرات تولید شده کیفیت بهتری خواهند داشت. در سایر روش ها معمولا از دانش فضای مسئله کمتر استفاده می‌شود و باعث می‌شود عملگرهای جهش در آن‌ها با کیفیت کمتری انجام شود. دلیل برتری الگوریتم پیشنهادی نسبت به سایر روش‌ها استفاده بهینه و هوشمند از فضای مسئله است که منجر به انتخاب تعداد ویژگی‌های کمتر ولی با کیفیت‌تر شده است. در برخی دیتاست‌ها مانند Musk و Madelon هر دو معیار تعداد ویژگی و خطای الگوریتم نسبت به سایر روش‌ها برتری قابل توجهی دارد. در جدول (۶) نتایج شبیه‌سازی آورده شده است. با توجه به نتایج شبیه‌سازی در برخی دیتاست‌ها مانند Isolet روش‌هایی مانند NSGAII دارای خطای الگوریتم کمتری به نسبت الگوریتم پیشنهادی هستند، اما تعداد ویژگی‌هایی که توسط این الگوریتم انتخاب شده است، بسیار بیشتر از روش پیشنهادی است.

جدول (۶): میانگین تعداد ویژگی‌های انتخاب شده و دقت الگوریتم در کلاس‌بند KNN

		Musk	Madelon	Cnae	Isolet	Arrhythmia	Hillvalley	LSVT
NSGAII	ویژگی	۲۶	۱۴۳	۲۸۸	۱۷۵	۵۴	۵	۵۵
	دقت	۹۳	۵۰	۷۶.۴	۷۹	۵۹	۵۳	۷۹
MOPSO	ویژگی	۲۱	۱۲۲	۲۱۵	۱۳۱	۵۰	۶	۵۷
	دقت	۹۴	۵۰	۷۶	۷۴	۵۸	۶۰	۸۲
CMDPSOFS	ویژگی	۳۴	۱۲۲	۲۵۷	۱۷۰	۵۴	۹	۵۸
	دقت	۹۵.۲	۵۴	۷۳	۷۸	۵۶	۵۰	۷۲
HMPSOFS	ویژگی	۹	۳۱	۱۱۶	۷۶	۱۳	۱	۱۵
	دقت	۹۲	۵۰	۵۸	۷۰	۵۷	۵۷	۶۶
RFPSOFS	ویژگی	۱۵	۳۹	۹۱	۵۱	۱۴	۱	۸
	دقت	۹۳	۵۶	۷۱	۷۳	۵۸	۵۳	۶۹
Proposed Method	ویژگی	۶	۵	۲۹	۲۹	۶	۱	۱
	دقت	۹۶	۸۹	۶۳	۷۸	۷۲	۶۵	۹۲.۲

در جداول ۷ الی ۱۱ نیز انحراف معیار و واریانس خطا و تعداد ویژگی روش پیشنهادی در ۳۰ بار تکرار آورده شده است.

جدول (۷): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

	Proposed method	Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۱۹۶	۰.۰۶۱	۰.۱۹۳۷	۰.۳۶۷۷	۰.۲۷۷
	تعداد ویژگی	۳۷.۳۳	۴.۲۸	۷.۱	۹.۲۵	۹۵.۲
انحراف معیار	خطا	۰.۰۲۵۰	۰.۰۰۷	۰.۰۳۴۲	۰.۰۲۷	۰.۰۸۵
	تعداد ویژگی	۱۲.۲۹	۱.۷۹	۱.۸۲	۲.۵	۱۸.۸۸
واریانس	خطا	۰.۰۰۰۶	۰.۰۰۰۰۵	۰.۰۰۱۱	۰.۰۰۰۷۷	۰.۰۰۷۲

	تعداد ویژگی	۱۵۱.۰۶	۳.۲۳	۳.۳۳	۶.۲۵	۳۵۶.۷

جدول (۸): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

RFPSOFS		Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۲۲۶۴	۰.۰۶۰	۰.۴۰۴	۰.۳۷۹۴	۰.۲۳۳
	تعداد ویژگی	۶۶.۱	۱۹.۲	۴۳.۲	۱۷.۸	۱۰۹.۴
انحراف معیار	خطا	۰.۰۳۱۶	۰.۰۱۰۴	۰.۰۲۸	۰.۰۳۸۴	۰.۰۴۶۸
	تعداد ویژگی	۱۴.۱۵	۳.۰۳۱	۳.۶۵	۳.۷۰	۲۱.۴۷
واریانس	خطا	۰.۰۰۱۰	۰.۰۰۰۱۱	۰.۰۰۰۸۳	۰.۰۰۱۴	۰.۰۰۰۲۱
	تعداد ویژگی	۲۰۰.۵	۹.۱۹	۱۳.۳۳	۱۳.۷	۴۶۱.۳

جدول (۹): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

CMDPSOFS		Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۲۰۹۴	۰.۰۴۶۴	۰.۴۳۴۹	۰.۳۷۹۴	۰.۲۳۲۷
	تعداد ویژگی	۱۷۸.۵	۴۴.۲	۱۳۰.۳۳	۶۵.۱	۲۷۸.۸
انحراف معیار	خطا	۰.۰۱۸۹۵	۰.۰۰۲۵	۰.۰۳۰۵	۰.۰۴۹۷۱	۰.۰۴۲۴۶
	تعداد ویژگی	۸.۰۶۲۲	۵.۸۹	۱۶.۳۱	۱۰.۷۲	۱۲.۷۱۳
واریانس	خطا	۰.۰۰۰۳۵	۰.۰۰۰۰۰۰۶	۰.۰۰۰۹	۰.۰۰۲۴۷	۰.۰۰۰۱۸
	تعداد ویژگی	۶۵.۱	۳۴.۷	۲۶۶.۳۳	۱۱۵.۲	۱۶۱.۶۲

جدول (۱۰): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

HMPSOFS		Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۲۶۲۸	۰.۰۶۳۷۲	۰.۴۷۵	۰.۴۱۴۷۰	۰.۲۷۲۹
	تعداد ویژگی	۸۵.۷۵	۲۰.۵۵	۳۶.۲	۲۱.۴	۱۳۱.۱۴
انحراف معیار	خطا	۰.۰۴۱۵۴	۰.۰۱۸۶۲	۰.۰۸۶۳	۰.۰۲۱۱۸	۰.۰۷۸۹۰
	تعداد ویژگی	۷.۲۷۴۳	۸.۷۱۹۳	۸.۰۴۱۵	۵.۵۰۴۵	۱۱.۸۸
واریانس	خطا	۰.۰۰۱۷۲۵	۰.۰۰۰۳۴	۰.۰۰۷۴	۰.۰۰۰۴۴	۰.۰۰۶۲۲
	تعداد ویژگی	۵۲.۹۱۶۶	۷۶.۰۲۷	۶۴.۶۶	۳۰.۳	۱۴۱.۱۴۲

جدول (۱۱): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

MOPSO		Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۲۱۰۲۵۶	۰.۰۵۷۷۰	۰.۴۴۱۶	۰.۳۶۵۸	۰.۱۹۶۶
	تعداد ویژگی	۱۴۵.۸	۳۰.۱	۱۰۹.۴	۵۶.۷۵	۲۲۳.۲
انحراف معیار	خطا	۰.۰۳۹۰۰	۰.۰۰۶۰۰۴	۰.۰۴۰۰	۰.۰۵۵۷۰	۰.۰۲۷۵۹
	تعداد ویژگی	۱۰.۷۵۶۳	۶.۴۴۲۰	۵.۳۱۹۷	۵.۷۳۷۳	۷.۹۶۸۶
واریانس	خطا	۰.۰۰۱۵۲۱	۰.۰۰۰۰۳۶	۰.۰۰۱۶	۰.۰۰۳۱۰۳	۰.۰۰۰۷۶
	تعداد ویژگی	۱۱۵.۷	۴۱.۵	۲۸.۳	۳۲.۹۱۶	۶۳.۵

جدول (۱۲): میانگین، انحراف معیار و واریانس خطا و تعداد ویژگی

NSGAII		Isolet	Musk	Madelon	Arrhythmia	Cnae
میانگین	خطا	۰.۱۹۴۴۴	۰.۰۵۵۳۳	۰.۴۴۳۴	۰.۳۹۶۷	۰.۱۹۰۱
	تعداد ویژگی	۱۸۷.۵	۳۳.۸۳	۱۴۹.۲	۵۹.۵	۲۹۸.۶
انحراف معیار	خطا	۰.۰۱۸۲۳	۰.۰۰۹۶۲۸	۰.۰۳۸۳	۰.۰۲۰۶۶	۰.۰۳۱۶۷
	تعداد ویژگی	۱۲.۲۴۳۳	۷.۱۶۷۰	۷.۱۵۵۴	۵.۱۹۶۱۵	۸.۳۵۴
واریانس	خطا	۰.۰۰۰۳۳	۰.۰۰۰۰۹۲	۰.۰۰۱۴	۰.۰۰۰۴۲	۰.۰۰۱۰
	تعداد ویژگی	۱۴۹.۹	۵۱.۳۶	۵۱.۲	۲۷.۱۲	۶۹.۸

۴-۲- دیتاست‌های نامتوازن

برای بررسی کارایی روش پیشنهادی از دیتاست‌های نامتوازن جدول (۱۳) استفاده شده است.

جدول (۱۳): توضیحات دیتاست نامتوازن

نام دیتاست	تعداد داده‌ها	تعداد ویژگی‌ها	تعداد کلاس	میزان نامتوازن
Leukemia	۷۲	۷۱۲۹	۲	۱.۸۸
Caltech101	۸۶۷۱	۷۸۴	۱۰۱	۲۵.۷۴
Mnist	۷۰۰۰۰	۷۸۴	۱۰	۶
FuelCons	۱۷۶۴	۳۷	۴	۱۳.۰۸
Lymphoma cancer	۶۶	۴۰۲۶	۳	۵.۱۱
DLBCL	۷۷	۵۴۶۹	۲	۳.۰۵
Dermatology	۳۶۶	۳۴	۶	۵.۶

۴-۲-۱- پارامترها

پارامترهای روش پیشنهادی برای دیتاست نامتوازن نیز مطابق جدول (۴) است. در سایر روش‌ها، در روش IIEFS پارامتر Q برابر ۱۰، پارامتر K برابر ۳۰ و درصد ویژگی‌های تکی و درصد ویژگی‌های جفت به ترتیب در دیتاست Lymphoma برابر ۰,۱۵ و ۰,۰۰۷، در دیتاست DLBCL و Leukemia برابر ۰,۱۲۵ و ۰,۰۰۴۳، در دیتاست Caltech101 برابر ۰,۵ و ۰,۱، در دیتاست Mnist برابر ۰,۷ و ۰,۱ و برای سایر دیتاست‌ها برابر ۰,۶ و ۰,۲ است.

۴-۲-۲- نتایج شبیه‌سازی

روش پیشنهادی در ۷ دیتاست با نام‌های "Leukemia"، "Caltech101"، "Mnist"، "FuelCons"، "Lymphoma cancer"، "DLBCL" و "Dermatology" با روش‌های IIEFS،

IWFS[۳۴] [۳۱]، WJMI [۳۲] و mRMR [۳۳] مقایسه شده است. میانگین نتایج شبیه‌سازی با ۳۰ اجرا در جداول (۱۴) الی (۱۷) آورده شده است. در این شبیه‌سازی‌ها از معیار F-measure برای ارزیابی کیفیت ویژگی‌های انتخاب شده و از دو الگوریتم کلاس‌بند KNN و Naïve Bayes در فرآیند یادگیری استفاده شده است.

جدول (۱۴): میانگین تعداد ویژگی‌های انتخاب شده در الگوریتم کلاس‌بند KNN

نام دیتاست	Proposed Method	IIEFS	IWFS	WJMI	mRMR
Leukemia	۱	۳	۵	۲	۷.۱
Caltech101	۲۸	۳۰	۳۰	۲۹	۳۰
Mnist	۳۰	۳۰	۲۹.۸	۳۰	۳۰
FuelCons	۱.۵	۲۰	۲۰.۲	۲۲.۲	۲۹.۳۳
Lymphoma cancer	۱	۴.۴۷	۲	۳	۱۴.۲
DLBCL	۱	۲	۷	۵	۷.۶۷
Dermatology	۳.۵	۱۳.۷	۲۹.۳	۲۴.۱	۲۷.۵۶

جدول (۱۵): میانگین تعداد ویژگی‌های انتخاب شده در الگوریتم کلاس‌بند Naïve Bayes

نام دیتاست	Proposed Method	IIEFS	IWFS	WJMI	mRMR
Leukemia	۱	۴.۶۷	۱۴.۵	۵	۱۷
Caltech101	۱۶	۳۰	۳۰	۲۸.۱۱	۲۷
Mnist	۱۶	۳۰	۱۸	۳۰	۳۰
FuelCons	۲	۱۳.۳۳	۲۷.۶۷	۲۷.۲	۲۹.۶۷
Lymphoma cancer	۱	۷.۴۰	۲	۵	۱۴
DLBCL	۲	۹	۲۴.۶۷	۱۳.۶۷	۳
Dermatology	۴	۱۳.۲۵	۲۲.۶۷	۲۰.۶۷	۲۵.۳۳

جدول (۱۶): میانگین مقدار F-Measure در الگوریتم کلاس‌بند KNN

نام دیتاست	Proposed Method	IIEFS	IWFS	WJMI	mRMR
Leukemia	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
Caltech101	۴۱	۳۵	۱۷.۷۵	۲۲.۲۷	۶.۱۳
Mnist	۸۶	۸۹.۲۶	۷۵.۸۵	۷۰.۲۸	۶۶.۰۴
FuelCons	۸۰	۷۸.۰۱	۷۶.۹۲	۷۷.۴۹	۷۶.۹۴
Lymphoma cancer	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
DLBCL	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۳.۵۹
Dermatology	۹۵	۹۸.۳	۸۶.۲	۸۷.۵۱	۸۷.۶۵

جدول (۱۷): میانگین مقدار F-Measure در الگوریتم کلاس‌بند Naïve Bayes

نام دیتاست	Proposed Method	IIEFS	IWFS	WJMI	mRMR
Leukemia	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
Caltech101	۱۷	۲۹.۳	۱۱.۹۸	۱۷.۸۴	۳.۰۷
Mnist	۶۴	۷۶.۳۱	۵۵	۵۰.۵۲	۵۰.۰۹
FuelCons	۷۷	۸۴.۷۵	۶۲.۶۶	۶۳.۵۸	۶۳.۱۳
Lymphoma cancer	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
DLBCL	۱۰۰	۱۰۰	۹۷.۱۶	۱۰۰	۹۲.۹۴
Dermatology	۹۴	۹۵.۸۴	۹۴.۲۵	۹۴.۶۷	۹۴.۱۹

همانطور که در جداول مشخص است، الگوریتم پیشنهادی در مقایسه با سایر روش‌ها تعداد ویژگی بسیار کمتری را انتخاب کرده است. تعداد کم ویژگی در برخی دیتاست‌ها مانند Mnist، FuelCons و Dermatology منجر به خطای بیشتری شده است. انتخاب ویژگی و میزان خطا دو هدف متضاد هستند. هر چه تعداد ویژگی کمتر شود احتمال افزایش خطا وجود خواهد داشت. الگوریتم پیشنهادی به گونه‌ای طراحی شده است که تعداد ویژگی را با در نظر گرفتن خطای الگوریتم تا حد امکان کاهش دهد. به همین دلیل در برخی موارد الگوریتم پیشنهادی دقت کمتری به نسبت سایر روش‌ها دارد. اما با در نظر گرفتن تعداد ویژگی کمتر این خطا قابل چشم پوشی است.

۵- نتیجه‌گیری

امروزه به دلیل حجم بالای داده‌های تولید شده در کاربردهای مختلف، نیاز به الگوریتم‌های انتخاب ویژگی و کاهش ابعاد داده، بیش از پیش احساس می شود. با بهره مندی از روش های مختلف انتخاب ویژگی، الگوریتم‌های یادگیری می توانند مدل‌های بهتری را ایجاد کنند. از طرف دیگر، برخی داده‌هایی که جمع‌آوری می‌شوند به صورت متوازن نیستند. در اینگونه مواردِ، برخی کلاس‌ها دارای داده‌های کمتر و تواما اهمیت بالاتری نسبت به سایر کلاس‌ها هستند. از این رو الگوریتم‌های انتخاب ویژگی بایستی بتوانند عدم توازن داده را در نظر گرفته و ویژگی‌هایی را انتخاب کنند که کلاسبند بتواند کلاس‌ها با داده‌های کمتر را نیز تشخیص دهد. در این مقاله یک روش تکاملی مبتنی بر باکتری‌ها ارائه شده است . در روش پیشنهادی، از مکانیزم آرشیو در کنار

جهش‌های هوشمند درون سلولی و لوله جنسی استفاده شد. این دو جهش باعث شد الگوریتم پیشنهادی از نقاط بهینه محلی خارج شده و به سمت یافتن بهینه جهانی حرکت کند.

روش پیشنهادی در دو نوع دیتاست با روش‌های موجود مقایسه شد. در دیتاست‌های متوازن که تعداد آن‌ها هفت بود، روش پیشنهادی با روش‌های پایه NSGAII و MOPSO و روش‌های موثر با نام‌های CMDPSOFS، HMPSOFS و RFPSOFS مقایسه شد و عملکرد بهتری به نسبت سایر روش‌ها داشت. در دیتاست Musk روش پیشنهادی در مقایسه با RFPSOFS که از جدیدترین روش‌ها بود، با دستیابی به تعداد ویژگی ۶ و دقت الگوریتم ۹۶ عملکرد بهتری از خود نشان داد. در دیتاست‌های Madelon، Cnae، Isolet، Arrhythmia، HillValley و LSVT نیز به ترتیب تعداد ویژگی‌های انتخاب شده و دقت الگوریتم برابر ۵ و ۸۹، ۲۹ و ۶۳، ۲۹ و ۷۸، ۶ و ۷۲، ۱ و ۶۵ و ۱ و ۹۲،۲ بود. در دیتاست‌های نامتوازن که تعداد آن‌ها نیز هفت بود، روش پیشنهادی با روش‌های IIEFS، IWFS، WJMI و mRMR مقایسه شد. در این مقایسه از دو الگوریتم یادگیری KNN و Naïve Bayes استفاده شد که در هر دو الگوریتم، روش پیشنهادی توانست به تعداد ویژگی کمتری نسبت به سایر روش‌ها دست یابد. در دیتاست FuelCons روش پیشنهادی توانست به تعداد ویژگی انتخاب شده ۲ دست یابد در حالی که دقت الگوریتم KNN در این تعداد برابر ۸۰ بود. سایر روش‌ها در این دیتاست بیشتر از ۲۰ ویژگی انتخاب کرده بودند. در دیتاست‌های Leukemia، Caltech101، Mnist، Lymphoma cancer، DLBCL و Dermatology با در نظر گرفتن الگوریتم KNN به عنوان الگوریتم یادگیری، تعداد ویژگی‌های انتخاب شده و دقت الگوریتم به ترتیب برابر ۱و ۱۰۰، ۲۸ و ۴۱، ۳۰ و ۸۶، ۱ و ۱۰۰، ۱ و ۱۰۰ و ۳،۵ و ۹۵ بود.

۶- منابع و مراجع

[1] R. Ran, J. Feng, S. Zhang, and B. Fang, "A general matrix function dimensionality reduction framework and extension for manifold learning," *IEEE Transactions on Cybernetics,* 2020.

[2] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, 2015: Ieee, pp. 1200-1205.

[3] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR,* vol. 4, no. 3, pp. 211-229, 2014.

[4] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to multi-objective feature selection: A systematic literature review," *IEEE Access,* vol. 8, pp. 125076-125096, 2020.

[5] B. Cao *et al.*, "Multiobjective feature selection for microarray data via distributed parallel algorithms," *Future Generation Computer Systems,* vol. 100, pp. 952-981, 2019.

[6] N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access,* vol. 7, pp. 78533-78548, 2019.

[7] P. Zhou, X. Hu, P. Li, and X. Wu, "Online feature selection for high-dimensional class-imbalanced data," *Knowledge-Based Systems,* vol. 136, pp. 187-199, 2017.

[8] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data--recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*, 2013: IEEE, pp. 245-251.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, pp. 1157-1182, 2003.

[10] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, no. 1-2, pp. 273-324, 1997

[11] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014.

[12] S. S. Darshan and C. Jaidhar, "Performance evaluation of filter-based feature selection techniques in classifying portable executable files," Procedia Computer Science, vol. 125, pp. 346-356, 2018.

[13] S. Karasu, A. Altan, S. Bekiros, and W. Ahmad, "A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series," Energy, vol. 212, p. 118750, 2020.

[14] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," IEEE/CAA Journal of Automatica Sinica, vol. 6, no. 3, pp. 703-715, 2019.

[15] R. Vijayanand, D. Devaraj, and B. Kannapiran, "Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection," Computers & Security, vol. 77, pp. 304-314, 2018.

[16] Y.-P. Chen et al., "A novel bacterial foraging optimization algorithm for feature selection," Expert Systems with Applications, vol. 83, pp. 1-17, 2017.

[17] M. Amoozegar and B. Minaei-Bidgoli, "Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism," Expert Systems with Applications, vol. 113, pp. 499-514, 2018.

[18] A. Lin, W. Sun, H. Yu, G. Wu, and H. Tang, "Global genetic learning particle swarm optimization with diversity enhancement by ring topology," Swarm and evolutionary computation, vol. 44, pp. 571-583, 2019.

[19] B. Niu, W. Yi, L. Tan, S. Geng, and H. Wang, "A multi-objective feature selection method based on bacterial foraging optimization," *Natural Computing,* vol. 20, no. 1, pp. 63-76, 2021.

[20] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," *IEEE control systems magazine,* vol. 22, no. 3, pp. 52-67, 2002.

[21] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," Information Fusion, vol. 50, pp. 158-167, 2019

[22] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," Cybernetics and Information Technologies, vol. 19, no. 1, pp. 3-26, 2019.
[23] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," Pattern recognition letters, vol. 15, no. 11, pp. 1119-1125, 1994.
[24] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," IEEE transactions on cybernetics, vol. 43, no. 6, pp. 1656-1671, 2012.
[25] Y. Zhang, D.-w. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," IEEE/ACM transactions on computational biology and bioinformatics, vol. 14, no. 1, pp. 64-75, 2015.
[26] H. B. Nguyen, B. Xue, I. Liu, P. Andreae, and M. Zhang, "New mechanism for archive maintenance in PSO-based multi-objective feature selection," Soft Computing, vol. 20, no. 10, pp. 3927-3946, 2016.
[27] M. Amoozegar and B. Minaei-Bidgoli, "Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism," Expert Systems with Applications, vol. 113, pp. 499-514, 2018.
[28] E. S. Hosseini and M. H. Moattar, "Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification," Applied Soft Computing, vol. 82, p. 105581, 2019.
[29] T. M. Hamdani, J.-M. Won, A. M. Alimi, and F. Karray, "Multi-objective feature selection with NSGA II," in international conference on adaptive and natural computing algorithms, 2007: Springer, pp. 240-247.
[30] L. Cagnina, S. C. Esquivel, and C. C. Coello, "A particle swarm optimizer for multi-objective optimization," Journal of Computer Science and Technology, vol. 5, no. 04, pp. 204-210, 2005.
[31] Z. Zeng, H. Zhang, R. Zhang, and C. Yin, "A novel feature selection method considering feature interaction," *Pattern Recognition,* vol. 48, no. 8, pp. 2656-2666, 2015.
[32] X. Qi, C. Yin, K. Cheng, X. Liao, and X. Kang, "WJMI: A New Feature Selection Algorithm Based on Weighted Joint Mutual Information," 2015.
[33] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

زیرنویس‌ها

[۱] Micro array
[۲] Imbalanced data
[۳] Particle Swarm Optimization
[۴] Mutation
[۵] Sex pilus
[۶] Crossover
[۷] Chemotaxis
[۸] Reproduction
[۹] Elimination/Dispersal
[۱۰] Tumbling
[۱۱] Swimming
[۱۲] Nesting effect
[۱۳] Crowding Mutation Dominance PSO Feature Selection
[۱۴] Hybrid Mutation Dominance PSO feature selection
[۱۵]Improved Self-Regulating Particle Swarm Optimization
[۱۶] Swapping
[۱۷] Ranked Feature PSO Feature Selection
[۱۸] Interaction Information based Evolutionary Feature subsets Selection