تاریخ دریافت: ۱۳۹۷/۰۸/۱۳
تاریخ پذیرش: ۱۳۹۸/۰۱/۰۸

رویکرد ترکیبی نوین برای تشخیص هرزنامه با استفاده از الگوریتم‌های کلونی مورچه و کرم شب‌تاب

ناصر کریم‌پور
کارشناس ارشد، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: nasser.karimpour@gmail.com

فرهاد سلیمانیان قره چپق*
استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: bonab.farhad@gmail.com

چکیده:

با گسترش اینترنت و شبکه‌های کامپیوتری، دسترسی به اطلاعات سریع‌تر انجام می‌شود. اما از طرفی دیگر، دنیای اینترنت با وجود نفوذگرها و هرزنامه نویسان محیط امنی برای کاربران نیست. هر لحظه ممکن است خرابکاری‌هایی از طرف مهاجمان اینترنتی رخ بدهد. میزان آسیب‌ها و نفوذهای گزارش شده به سیستم‌های کامپیوتری در سرتاسر جهان روز به روز در حال افزایش است. با گذشت زمان ابزارها و روش‌های نفوذ به شبکه‌های کامپیوتری ساده و ساده‌تر می‌شوند و نفوذگرها با حداقل ابزارها، مجال نفوذ را می‌یابند. چند سالی است که هرزنامه در صندوق ایمیل‌های کاربران خیلی زیاد شده است و حتی در برخی مواقع کاربران را مجبور به ترک استفاده از ایمیل می‌کند. هرزنامه یکی از جنبه‌های منفی و آزار دهنده استفاده از خدمات ایمیل و در بسیاری از مواقع یکی از راه‌های آلوده شدن سیستم‌های کامپیوتری به انواع ویروس است. در این مقاله، مدل ترکیبی برمبنای بهینه‌سازی کلونی مورچه و الگوریتم کرم شب‌تاب برای طبقه‌بندی ایمیل هرزنامه پیشنهاد می‌شود. در مدل پیشنهادی از بهینه‌سازی کلونی مورچه به منظور انتخاب ویژگی و از الگوریتم کرم شب‌تاب برای آموزش و آزمایش نمونه‌ها استفاده شده است. نتایج نشان می‌دهد که درصد صحت مدل پیشنهادی برروی مجموعه داده Spambase با ۲۰۰بار تکرار و انتخاب همه ویژگی‌ها برابر ۹۲,۰۵ است و همچنین درصد صحت مدل پیشنهادی در مقایسه با بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی، الگوریتم انتخاب منفی، ترکیب تکاملی تفاضلی-الگوریتم انتخاب منفی، درخت تصادفی، شبکه بیزین، ماشین بردار پشتیبان و الگوریتم بیزین ساده بیشتر است.

واژه‌های کلیدی: تشخیص ایمیل هرزنامه، طبقه‌بندی، بهینه‌سازی کلونی مورچه، الگوریتم کرم شب‌تاب

۱. مقدمه

اگر دنیای دیروز را دنیای فناوری بدانیم، دنیای امروز را می‌توان دنیای ابزارها فرض کرد. گسترش فناوری و کاهش محدودیت دسترسی به اطلاعات، عملاً باعث گردیده

* نویسندهٔ مسئول

است که سطح فناوری در فرهنگ‌ها و جوامع مختلف، مشابه و یا حداقل قابل مقایسه باشد. آنچه که امروزه می‌تواند رشد و توسعه یا توقف و پیشرفت یک فرهنگ و جامعه را مشخص نماید، ابزارسازی و ابزارشناسی در آن جامعه است. ایمیل در میان ابزارهای مورد استفاده در جامعه امروزی جایگاه ویژه‌ای دارد که ابزاری بسیار ارزان قیمت با کمترین هزینه اطلاع رسانی، تبلیغات و روابط عمومی می‌باشد.

در دهه‌های اخیر به دلایل استفاده ارزان و راحت و عدم اتلاف زمان ارسال پیام از طریق پست الکترونیکی، بنابر نیازهای جوامع مختلف از جمله اهداف علمی، تبلیغاتی به شدت افزایش پیدا کرده است که این امر باعث پر شدن صندوق‌های پست الکترونیکی کاربران اینترنت از ایمیل‌های هرز گردیده است [۱]. هرزنامه به معنای پیام یا نامه الکترونیکی است که بدون درخواست گیرنده و به‌صورت ناخواسته در حجم انبوه برای کاربران بی‌شماری فرستاده می‌شود. امروزه هرزنامه تبدیل به یک معضل برای کاربران و شرکت‌های حوزه فناوری اطلاعات و سازمان‌های اجتماعی، اقتصادی و دولتی شده است، زیرا در این زمینه مقدار بزرگی از پهنای باند به هدر می‌رود و در سیستم فرستادن ایمیل سرریز رخ می‌دهد [۲]. همچنین از نظر فنی ارسال هرزنامه تقریبا بدون هزینه است و این مسئله باعث شده است که شرکت‌های بازاریابی به سمت آن حرکت کنند.

در برخی از مواقع هرزنامه‌ها حاوی اطلاعاتی مخرب برای کاربران می‌باشند که با اهداف خاصی طراحی می‌شوند و می‌توانند شامل ویروس یا نرم‌افزارهای مخرب باشد که منجر به خرابی در کامپیوترها و شبکه‌ها و سرقت اطلاعات شخصی، سرقت داده، دزدی هویت کاربران و دستاوردهای فکری دیگران می‌گردد [۳]. در بیشتر مواقع، فضایی از صندوق ایمیل کاربران را اشغال می‌کنند و از کاربر می‌خواهند که ایمیل را به بقیه افراد ارسال نماید و یا وبگاه مشخصی را ملاقات کنند که در جامعه اینترنتی چنین ایمیل‌هایی، نمونه بارز هرزنامه می‌باشند. هرزنامه وقت و انرژی زیادی از کاربران را هدر داده و می‌تواند سرچشمه کلاهبرداری و سرقت باشد و همچنین تاثیرات نامطلوب فرهنگی نیز دارند. درصورتی که هر کاربر اینترنت دارای یک نشانی پست الکترونیکی عمومی باشد، ممکن است صدها پیام ناخواسته را با ظاهری کاملاً متفاوت و نامعقول و غیرمنطقی دریافت نماید [٤].

بیشتر تولید کننده‌های هرزنامه سال‌های زیادی برای پنهان کردن منشأ تولید هرزنامه فعالیت می‌کنند و معمولاً هرزنامه ایمیل‌ها را از طریق کارسازهای پراکسی ناامن ارسال می‌کنند. هرزنامه‌نویسان مکرراً از نام‌ها، نشانی‌ها، شماره تلفن‌ها و دیگر اطلاعات جعلی برای پرداخت حساب‌ها استفاده می‌کنند که این کار به آن‌ها اجازه می‌دهد سریعاً از حسابی به حساب دیگر در صورت برملا شدن حرکت کنند [٥]. هدف از ارسال هرزنامه می‌تواند جنبه‌های مختلفی داشته باشد که عبارتند از: اطلاع رسانی، تبلیغاتی، ایجاد ترافیک در شبکه و دستیابی به رمزهای کاربران. شاید این سؤال به وجود آید که آیا همه پیام‌هایی که در ایمیل هرزنامه قرار می‌گیرند مخرب یا ویروسی هستند؟ می‌توان گفت خیر، همه پیام‌ها در زمره پیام‌های مخرب و یا ویروسی نیستند، بلکه بعضی از آن‌ها می‌توانند مخرب و یا ویروسی باشد. بنابراین، راه‌هایی برای مقابله و جلوگیری از قرار نگرفتن پیام‌ها در هرزنامه وجود دارد. اما این راهکارها هم نمی‌تواند به طور کامل کارگشا باشند و جلوی همه آن‌ها را بگیرد، بلکه می‌توان تعداد پیام‌های هرزنامه را کاهش داد. ما در این مقاله می‌خواهیم ایمیل هرزنامه را با استفاده از مدل ترکیبی الگوریتم بهینه‌سازی کلونی مورچه [٦] و الگوریتم کرم شب‌تاب [٧] که هردو از الگوریتم‌های فراابتکاری می‌باشند، تشخیص دهیم. در این مقاله مجموعه داده‌های Spambase [٨] با ٤٦٠١ نمونه و ٥٧ ویژگی را مورد ارزیابی قرار می‌دهیم. مجموعه Spambase شامل دو رده هرزنامه با ۱۸۱۳ نمونه(٣٩,٤٪) و غیرهرزنامه با ۲۷۸۸ نمونه(٦٠,٦٪) است.

الگوریتم بهینه‌سازی کلونی مورچه [٦]، یکی از جدیدترین و کارآمدترین الگوریتم‌های تکاملی می‌باشد که از زندگی مورچه‌ها الهام گرفته است. الگوریتم بهینه‌سازی کلونی مورچه برای اولین بار توسط دوریگو و همکاران به عنوان یک راه حل چند عامله برای حل مسائل مشکل بهینه‌سازی مثل فروشنده دوره‌گرد ارائه شد. مورچه‌ها هنگام راه رفتن از خود ردی از ماده شیمیایی فرمون به‌جای می‌گذارند، این ماده قدرت تبخیر بالایی دارد ولی در کوتاه مدت به‌عنوان رد مورچه بر سطح زمین باقی می‌ماند. یک رفتار پایه‌ای ساده در مورچه‌های وجود دارد. آن‌ها هنگام انتخاب بین دو مسیر به‌صورت احتمالاتی مسیری را انتخاب می‌کنند که فرمون بیشتری داشته باشد یا به‌عبارت دیگر مورچه‌های بیشتری قبلا از آن عبور کرده باشند. الگوریتم کرم شب‌تاب [٧]، یکی از الگوریتم‌های الهام گرفته شده از طبیعت است. کرم‌های شب‌تاب، درک پیرامون خود را از طریق حسگرها انجام می‌دهند و مبتنی بر هوش جمعی می‌باشند.

الگوریتم بهینه‌سازی کلونی مورچه و کرم شب‌تاب به دلیل قابلیت‌هایی مانند تعداد پارامترهای کم، معادلات آسان، فرار از بهینه‌های محلی و قابلیت جستجوی ناحیهٔ اطراف جواب بهینهٔ یافته شده از کیفیت بالایی در انتخاب ویژگی و طبقه‌بندی بهره‌مند هستند. این الگوریتم‌ها می‌توانند احتمال به دام افتادن در کمینه محلی را کاهش دهند و همچنین دارای نرخ همگرایی (کشف راه حل بهینه) بالایی هستند. در تمام الگوریتم‌های فرا ابتکاری از جمله الگوریتم بهینه‌سازی کلونی مورچه و کرم شب‌تاب دو ویژگی ذاتی باید در نظر گرفته شود: توانایی الگوریتم برای جستجوی تمام بخش‌های فضای جستجو (اکتشاف) و توانایی در بهره‌برداری از بهترین راه حل. در الگوریتم بهینه‌سازی کلونی مورچه، توانایی اکتشاف با استفاده از به‌روزرسانی فرمون و توانایی بهره‌برداری با استفاده از احتمال حرکت انجام می‌گیرد. در الگوریتم کرم شب‌تاب، با انتخاب مقدار مناسب برای پارامترهای اولیه، اکتشاف می‌تواند تضمین شود و حرکت کرم‌ها برمبنای فاصله می‌تواند توانایی بهره‌برداری را تضمین کند.

ساختار کلی مقاله به‌شرح زیر سازماندهی شده است: در بخش دوم کارهای قبلی را توضیح می‌دهیم. در بخش سوم، مدل پیشنهادی و مراحل آن را به تفکیک توضیح می‌دهیم. در بخش چهارم، ارزیابی و مقایسه مدل پیشنهادی را توضیح می‌دهیم و نهایتاً در بخش پنجم نتیجه‌گیری و کارهای آینده را توضیح خواهیم داد.

۲. کارهای قبلی

تاکنون تکنیک‌های متعددی برای تشخیص هرزنامه ارائه شده است که شامل تکنیک‌های رگرسیون، داده کاوی، شبکه‌های احتمالاتی، شبکه‌های عصبی مصنوعی، منطق فازی و الگوریتم‌های هوش جمعی می‌باشد که در این بخش برخی از آن‌ها را بررسی و تحلیل می‌کنیم.

مدل RAN-LSH[1] [۹] که ترکیبی از شبکه عصبی مصنوعی تابع شعاعی پایه و ماشین بردار پشتیبان است برای تشخیص ایمیل هرزنامه پیشنهاد شده است. ارزیابی برروی دو مجموعه داده که در سال ۲۰۱۳ گردآوری شده‌اند و هر کدام شامل ۸۹٤۸ (۱۳۱۱ ایمیل هرزنامه و ۷٦۳۷ غیرهرزنامه) و ۲۰۹۰۵ (۸۸٦۳ و ۱۲۰٤ ایمیل غیرهرزنامه) نمونه هرزنامه هستند انجام شده است. مدل RAN-LSH از تابع درهم‌ساز به منظور جستجو و ذخیرهٔ نمونه‌ها استفاده می‌کند. در مدل RAN-LSH [۹]، با استفاده از جدول درهم‌ساز یک گروه از نمونه‌های مشابه به یک سطر خاص در جدول درهم‌ساز اختصاص داده می‌شوند که می‌توانند با سرعت با استفاده از مقدار نمایه بازیابی شوند. همچنین افزودن داده‌های جدید در جدول درهم‌ساز به زمان کمی نیاز دارد. لذا هدف از تابع درهم‌ساز ایجاد یک جدول برای نمونه‌های هرزنامه و غیرهرزنامه است. جدول درهم‌ساز شامل وزن نمونه‌ها و تعداد نمونه‌ها است و ارزیابی آن برمبنای معیارهای دقت و بازخوانی انجام

1- Resource Allocating Network with Locality Sensitive Hashing (RAN-LSH)

شده است. نتایج نشان داده که مدل RAN-LSH در پیش‌بینی دقت مناسبی دارد. به دلیل این‌که مجموعه داده‌ها شامل ویژگی‌های زیادی هستند از انتخاب ویژگی استفاده شده است. در مدل RAN-LSH برای انتخاب ویژگی از ماشین بردار پشتیبان و برای آموزش و آزمایش از شبکه عصبی مصنوعی تابع شعاعی پایه استفاده شده است. همچنین برای انتخاب ویژگی‌های برازنده از لایه وسطی مدل RAN-LSH استفاده می‌شود.

مدل شبکه عصبی مصنوعی نوع روش گروهی مدیریت داده‌ها [۱۰] برای طبقه‌بندی هرزنامه پیشنهاد شده است. شبکه عصبی نوع روش گروهی داده‌گردانی [۲] یکی از پرکاربردترین شبکه‌های عصبی مصنوعی است که از توانایی بالایی در مدل‌سازی داده‌های پیچیده برخوردار است. ارزیابی مدل روش گروهی داده‌گردانی برروی مجموعه داده جهانی و معتبر Spambase انجام شده است. مجموعه داده Spambase شامل ۵۷ ویژگی و ٤٦۰۱ نمونه است. نمونه‌ها در دو ردهٔ هرزنامه و غیرهرزنامه طبقه‌بندی شده‌اند. نتایج با انتخاب ویژگی‌های مختلف نشان داده که درصد صحت در مدل روش گروهی داده‌گردانی در مقایسه با مدل‌های پرسپترون چندلایه و بیزین ساده بیشتر است. بیشترین درصد صحت در مدل روش گروهی داده‌گردانی برابر ۹۲,٤ و در مدل‌های پرسپترون چندلایه و بیزین ساده به‌ترتیب برابر ۹۱,۷ و ۷۵,٤ است.

مدل ترکیبی بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی به‌منظور طبقه‌بندی هرزنامه پیشنهاد شده است [۱۱]. ارزیابی برروی مجموعه داده Spambase انجام شده است. در مدل ترکیبی بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی از بهینه‌سازی اجتماع ذرات برای جستجوی مقدار ویژگی‌ها در فضای مسئله و از مدل الگوریتم انتخاب منفی برای انتخاب ویژگی استفاده شده است. روش پیشنهادی دارای دو مرحلهٔ آموزش و آزمایش است. نتایج نشان داده که مدل ترکیبی بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی در مقایسه با مدل‌های بیزین ساده، ماشین بردار پشتیبان-انتخاب ویژگی متمایز و الگوریتم انتخاب منفی دقت تشخیص بالاتر و در مقایسه با مدل ماشین بردار پشتیبان دقت تشخیص کمتری دارد. درصد صحت در مدل بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی برابر ۸۳,۲۰ و در مدل‌های بیزین ساده، ماشین بردار پشتیبان-انتخاب ویژگی متمایز و الگوریتم انتخاب منفی به‌ترتیب برابر ۷۸,۸، ۷۱ و ٦۸,۸٦ است.

مدل MOEA[۳] [۱۲] با الهام از الگوریتم‌های تکاملی برای تشخیص هرزنامه بر روی دو مجموعه داده که شامل ۲۷۲ و ٤۲٦ ایمیل هستند اجرا و آزمایش شده است. در این مدل از قوانین برای تشخیص هرزنامه استفاده شده است. نتایج نشان داده که دقت تشخیص MOEA در مقایسه با NSGA و SOOA[٤] بیشتر است و همچنین خطای پیش‌بینی در مدل MOEA کمتر است. مدل QUANT[٥] [۱۳] که ترکیبی از شبکه عصبی مصنوعی و درخت تصمیم‌گیری است برای تشخیص هرزنامه پیشنهاد شده است. در این مدل داده‌ها با استفاده از شبکه عصبی مصنوعی آموزش داده و آزمایش می‌شوند و با استفاده از درخت C4.5 طبقه‌بندی می‌شوند. از الگوریتم C4.5 جهت تحلیل ویژگی‌های اصلی موثر بر هرزنامه و تشخیص استفاده شده است. در درخت C4.5 هر مسیر از ریشه به سمت یک گره، نمایانگر یک قانون طبقه‌بندی می‌باشد. ارزیابی برروی دو مجموعه داده Spam-Assassin و Corpus 2006 انجام شده است. نتایج برروی مجموعه داده Spam-Assassin نشان داده که درصد صحت در مدل QUANT برابر ۸۹,۱۵ و در مدل‌های بیزین ساده، SMO و درخت C4.5 به‌ترتیب برابر ۸۱,۰۸، ۸۸,٦۲ و ۷۳,۰۸ است. و همچنین برروی مجموعه داده Corpus درصد صحت مدل QUANT برابر ۹۰,۸۷ و در مدل‌های بیزین ساده، SMO و C4.5 به‌ترتیب برابر ۸۸,۱۵، ۸۹,۷۹ و ۸۸,۱۵ است.

مدل‌های شبکه عصبی مصنوعی و درخت تصمیم‌گیری برروی ۱۲۰۰ نوع ایمیل متفاوت آزمایش و اجرا شده‌اند

2- Group Method of Data Handling (GMDH)

3- Multi-Objective Evolutionary Algorithms (MOEA)
4- Single Objective Optimization Algorithm (SOOA)
5- Quadratic-Neuron-based Neural Tree (QUANT)

[۱٤]. در مدل درخت تصمیم‌گیری از قوانین کشف و در مدل شبکه عصبی مصنوعی از آموزش و آزمایش داده‌ها برای تشخیص استفاده شده است. در لایه‌های مختف شبکه عصبی مصنوعی از معیار وزن‌دهی برای انتخاب ویژگی استفاده شده است. در مدل درخت تصمیم‌گیری انتخاب یک ویژگی برازنده به عنوان ریشه درخت به دقت تشخیص کمک می‌کند. نتایج نشان داده که مدل شبکه عصبی مصنوعی در مقایسه با درخت تصمیم‌گیری دقت بهتری دارد. الگوریتم سیستم ایمنی مصنوعی [۱۵] که یکی از الگوریتم‌های تکاملی است برای تشخیص ایمیل هرزنامه برروی ٤٥٤۱۹ ایمیل از مجموعه TREC07 اجرا و آزمایش شده است. از مدل الگوریتم سیستم ایمنی مصنوعی برای انتخاب ویژگی و حذف ویژگی‌های نامرتبط استفاده شده است. ارزیابی با ٤ اجرا نشان داده که دقت تشخیص در الگوریتم سیستم ایمنی مصنوعی متفاوت است و دقت به حدود ۹۹درصد نزدیک شده است.

مدل ترکیبی «شبکه عصبی مصنوعی چندلایه-بهینه‌سازی اجتماع ذرات» به منظور تشخیص هرزنامه پیشنهاد شده است [۱٦]. از الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی و از مدل شبکه عصبی مصنوعی چندلایه برای آموزش و آزمایش داده‌ها و طبقه‌بندی استفاده شده است. در مدل شبکه عصبی مصنوعی چندلایه-بهینه‌سازی اجتماع ذرات از شبکه عصبی پرسپترون با تابع فعال‌سازی سیگموئید برای لایه پنهان و ۸۰درصد داده‌ها برای آموزش و ۲۰درصد برای آزمایش استفاده شده است. ارزیابی برروی مجموعه داده Ling-Spam با ٤۸۱ هرزنامه و ۲۱۷۱ غیرهرزنامه و مجموعه داده Spam-Assassin با ٦۰۰۰ نمونه ایمیل انجام شده است. ارزیابی برروی مجموعه داده Spam-Assassin و Ling-Spam نشان داده که درصد صحت در مدل شبکه عصبی مصنوعی چندلایه-بهینه‌سازی اجتماع ذرات به‌ترتیب برابر ۹۹,۹۸ و ۹۹,۷۹ است. مقایسه‌ها نشان داده که مدل شبکه عصبی مصنوعی چندلایه-بهینه‌سازی اجتماع ذرات در مقایسه با مدل‌های ماشین بردار پشتیبان با تابع کرنل، ماشین بردار پشتیبان با تابع RBF و شبکه عصبی مصنوعی چندلایه دقت تشخیص بهتری دارد.

مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی چندلایه و ترکیب آن‌ها بر روی دو مجموعه داده با ۱٤ ویژگی آزمایش و اجرا شده‌اند [۱۷]. مجموعه داده اولی شامل ٥۰٤ ایمیل (۳۳٦ ایمیل و ۱٦۸ هرزنامه) و مجموعه داده دومی شامل ٦٥۷ ایمیل(۳۸۷ ایمیل و ۲۷۰ هرزنامه) است. در مدل درخت تصمیم‌گیری از آنتروپی، مدل ماشین بردار پشتیبان از تابع کرنل و شبکه عصبی مصنوعی چندلایه از خطای میانگین استفاده شده است. نتایج بر روی مجموعه داده اولی نشان داده که درصد صحت در مدل ترکیبی برابر ۹۱,۰۷ و در مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی چندلایه به‌ترتیب برابر ۸۹,۸۸، ۸۸,٦۹ و ۸۹,۸۸ است. و برروی مجموعه داده دومی درصد صحت در مدل ترکیبی برابر ۹۱,۷۸ و در مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی چندلایه به‌ترتیب برابر ۹۰,۸۷، ۹۰,۸۷ و ۸۹,۰٤ است.

ادریس و همکاران [۱۸] مدل ترکیبی تکاملی تفاضلی-الگوریتم انتخاب منفی را برای تشخیص ایمیل هرزنامه پیشنهاد کرده‌اند. در این مدل از الگوریتم تکامل تفاضلی برای جستجوی ویژگی‌ها در فضای مسئله استفاده شده است. الگوریتم تکامل تفاضلی یکی از جدیدترین روش‌های جستجو است. الگوریتم تکامل تفاضلی به عنوان روشی قدرتمند و سریع برای مسائل بهینه‌سازی در فضاهای پیوسته معرفی شده است. ارزیابی مدل تکاملی تفاضلی-الگوریتم انتخاب منفی برروی مجموعه داده Corpus با ۱۸۱۳ هرزنامه و ۲۷۸۸ ایمیل غیرهرزنامه انجام شده است. نتایج نشان داده که درصد صحت در الگوریتم انتخاب منفی و تکاملی تفاضلی-الگوریتم انتخاب منفی به‌ترتیب برابر ٦۸,۸٦ و ۸۰,٦٦ است. و مقدار ضریب همبستگی در

الگوریتم انتخاب منفی و تکاملی تفاضلی–الگوریتم انتخاب منفی به‌ترتیب برابر ۳۶،۰۶ و ۶۰،۰۸ است.

مدل بهبود داده شده بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی [۱۹] برای تشخیص طبقه‌بندی هرزنامه پیشنهاد شده است. ارزیابی مدل بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی برروی مجموعه داده Corpus با ۱۸۱۳ ایمیل هرزنامه و ۲۷۸۸ ایمیل غیرهرزنامه انجام شده است. نتایج نشان داده است که مدل بهینه‌سازی اجتماع ذرات–الگوریتم انتخاب منفی در مقایسه با مدل‌های بهینه‌سازی اجتماع ذرات، الگوریتم انتخاب منفی، بیزین ساده، ماشین بردار پشتیبان، ماشین بردار پشتبان–انتخاب ویژگی متمایز، شبکه عصبی مصنوعی و فازی دقت تشخیص بهتری دارد.

طبقه‌بندی بیزین [۲۰] برای تشخیص هرزنامه برروی سه مجموعه داده که شامل ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ ایمیل هستند اجرا و آزمایش شده است. در طبقه‌بندی بیزین از روابط احتمالاتی استفاده می‌شود. نتایج نشان داده دقت تشخیص برای سه مدل بیشتر از ۹۳درصد است. برای ارزیابی از معیارهای صحت، دقت و بازخوانی استفاده شده است. درصد صحت به‌ترتیب برای سه مجموعه داده برابر ۹۳،۹۸، ۹۴،۸۵ و ۹۶،۴۶ است.

سیستم تشخیص هرزنامه با استفاده از ماشین بردار پشتیبان پیاده‌سازی شده است [۲۱]. این سیستم برای تشخیص ایمیل‌هایی که روزانه از طرف شبکه‌های اجتماعی ارسال می‌شوند طراحی شده است. به طور میانگین هر روز ۲۸۰۰۰۰ ایمیل با استفاده از ماشین بردار پشتیبان بررسی می‌شود. نتایج نشان داده که مدل ماشین بردار پشتیبان در زمان محاسباتی کم توانسته است حدود ۶۰درصد ایمیل‌ها را برمبنای هرزنامه و غیرهرزنامه تشخیص دهد. مدل بهینه‌سازی اجتماع ذرات برمبنای واحد پردازش گرافیکی برای تشخیص هرزنامه برروی ۸۴۸۱۰ ایمیل(۴۸۳۶۰ هرزنامه و ۳۶۴۵۰ غیرهرزنامه) آزمایش شده است [۲۲]. به منظور کاهش زمان محاسباتی و بهره‌وری مدل بهینه‌سازی اجتماع ذرات از مدل واحد پردازش گرافیکی استفاده شده است. برای محاسبه تشخیص کلماتی که به عنوان هرزنامه تلقی می‌شوند از تعداد فراوانی کلمات استفاده شده است. نتایج نشان داده که مدل بهینه‌سازی اجتماع ذرات با به‌کارگیری واحد پردازش گرافیکی توانسته است دقت تشخیص را در مقایسه با بهینه‌سازی اجتماع ذرات در حالت معمولی افزایش دهد.

جدول(۱)، مزایا و معایب مدل‌های پیشنهاد شده برای تشخیص هرزنامه را نشان می‌دهد.

در میان مدل‌های پیشنهاد شده در جدول(۱) دلیل برتری نسبی شبکه‌های عصبی مصنوعی در طبقه‌بندی به دلیل فرایند یادگیری است. زیرا براساس فرایند یادگیری، تمام ارتباطات پیچیده غیرخطی و اثرهای متقابل میان ویژگی‌های مستقل فراگرفته می‌شوند. لذا قدرت طبقه‌بندی این مدل به طور قابل ملاحظه‌ای بالا است. با وجود این، در صورت عدم انتخاب معماری مناسبی برای شبکه عصبی مصنوعی و بخصوص انتخاب تعداد زیادی از لایه‌های پنهان و گره‌ها، ممکن است شبکه، خطای تصادفی موجود در داده‌های تحت یادگیری را نیز فرابگیرد (بیش‌آموزش) و در نتیجه، با وجود دقت بالای شبکه در طبقه‌بندی داده‌های تحت یادگیری، در خصوص داده‌های تحت آزمون به خوبی عمل نکند. مهم‌ترین ضعف شبکه‌های عصبی مصنوعی، عدم تفسیرپذیری وزن‌ها، پیچیدگی و زمان‌بر بودن مقدار آستانه وزن‌ها است.

۳. مدل پیشنهادی

مدل پیشنهادی که ترکیبی از بهینه‌سازی کلونی مورچه و الگوریتم کرم شب‌تاب است برروی مجموعه داده Spambase اجرا می‌شود (مدل پیشنهادی برای مجموعه داده Spambase طراحی شده است). مهم‌ترین چالش در مجموعه داده‌های بزرگ، ابعاد زیاد یا به عبارتی تعداد زیاد ویژگی‌ها (صفات) است. علاوه بر این، حذف ویژگی‌های اضافی سبب افزایش کارایی الگوریتم‌ها می‌شود و حتی

جدول۱: مزایا و معایب مدل‌های پیشنهاد شده برای تشخیص هرزنامه

رفرنس	مدل‌ها	مزایا	معایب
[۹]	ترکیب شبکه عصبی مصنوعی تابع شعاعی پایه و ماشین بردار پشتیبان	*انتخاب ویژگی‌های باارزش با استفاده از ماشین بردار پشتیبان برای طبقه‌بندی *آموزش و آزمایش نمونه‌ها برمبنای کشف بهترین مقدار برای وزن نرون‌های شبکه عصبی مصنوعی	*کاهش درصد صحت شبکه عصبی مصنوعی تابع شعاعی پایه با افزایش تعداد گره‌های میانی
[۱۰]	شبکه عصبی مصنوعی از نوع روش گروهی مدیریت داده‌ها	*کشف رابطه بین داده‌های ورودی و خطای خروجی * احتمال رسیدن به پاسخ صحیح، حتی اگر بخشی از گره‌های شبکه در فرایند وزن‌دهی مشکل یا عملکرد غلط داشته باشند.	*افزایش میانگین مربعات خطا به دلیل بیش‌آموزش
[۱۱]	بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی	*بهبود الگوریتم انتخاب منفی با استفاده از بهینه‌سازی اجتماع ذرات *افزایش تنوع جمعیتی و کشف ویژگی‌های مناسب برای طبقه‌بندی	*افزایش زمان محاسباتی
[۱۲]	الگوریتم تکاملی چندگانه	*پیدا کردن جواب‌های بهینه متعدد (سراسری و محلی) مرتبط با تابع برازندگی *فضای جستجو به چندین زیر فضا تقسیم می‌شود، در هر زیر فضا عملیات جستجو به منظور کاهش خطای طبقه‌بندی انجام می‌شود.	*افزایش زمان محاسباتی
[۱۳]	ترکیب شبکه عصبی مصنوعی چندلایه و درخت تصمیم C۴,۵	*از میان متغیرهای مستقل، متغیری که دارای بیشترین آنتروپی است برای طبقه‌بندی انتخاب می‌شود.	*افزایش پیچیدگی
[۱۴]	ترکیب شبکه عصبی مصنوعی چندلایه و درخت تصمیم ID۳	* انتخاب یک ویژگی برازنده در درخت تصمیم‌گیری به عنوان ریشه درخت به دقت طبقه‌بندی کمک می‌کند. *حداقل تعداد تکرار از شبکه عصبی مصنوعی به منظور آموزش و تست نمونه‌ها	*افزایش عمق درخت تصمیم‌گیری
[۱۵]	الگوریتم سیستم ایمنی مصنوعی	*الگوریتم از مدل ریاضی آسانی برخوردار است. *انتخاب ویژگی برمبنای نزدیکی مقدار ویژگی‌ها	*گیر افتادن در بهینه محلی
[۱۶]	شبکه عصبی مصنوعی چندلایه- بهینه‌سازی اجتماع ذرات	*انعطاف‌پذیری در برابر مشکل بهینه محلی *همگرایی بالا	-
[۱۷]	مدل‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی چندلایه	*انتخاب ویژگی‌هایی که در طبقه‌بندی موثر هستند و دسته‌ی بیشتری توسط آن‌ها تشکیل می‌گردد.	*حساس به مقدار پارامترهای اولیه
[۱۸]	ترکیب تکاملی تفاضلی-الگوریتم انتخاب منفی	*افزایش کارایی الگوریتم انتخاب منفی با استفاده از تکاملی تفاضلی	*افزایش زمان اجرای الگوریتم ترکیبی *به‌دست آوردن دقیق نقاط بهینه محلی و سراسری با احتمال کم
[۱۹]	بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی	*افزایش تنوع جمعیتی و کشف ویژگی‌های مناسب برای طبقه‌بندی	افزایش زمان محاسباتی
[۲۰]	طبقه‌بندی بیزین	*استفاده از احتمال به منظور کاهش خطای طبقه‌بندی	*کاهش درصد صحت به دلیل انتخاب نمونه‌های مشابه
[۲۱]	ماشین بردار پشتیبان	*طبقه‌بندی نمونه‌ها به دو نمونه هرزنامه و غیرهرزنامه با تعداد تکرار کم *آموزش نسبتا ساده است. *برای داده‌های با ابعاد بالا تقریبا خوب جواب می‌دهد.	*تابع کرنل به منظور طبقه‌بندی نمونه‌ها باید دقیق باشد.
[۲۲]	بهینه‌سازی اجتماع ذرات	*استفاده از تعداد فراوانی کلمات *استفاده از وزن‌دهی به منظور یافتن ویژگی‌های دقیق	*مشکل بهینه محلی دارد

به دلیل کاهش ویژگی‌های نامرتبط می‌توان به نتایج بهتری دست یافت. در مدل پیشنهادی در ابتدا نمونه‌ها از مجموعه داده Spambase خوانده می‌شوند و سپس عملیات نرمال‌سازی انجام می‌گیرد. در مدل پیشنهادی از بهینه‌سازی کلونی مورچه به منظور انتخاب ویژگی و از الگوریتم کرم شب‌تاب برای طبقه‌بندی داده‌ها برمبنای آموزش و آزمایش نمونه‌ها استفاده می‌شود. در شکل(۱)، شمای کلی مدل پیشنهادی نشان داده شده است.

۳-۱ نرمال‌سازی داده‌ها

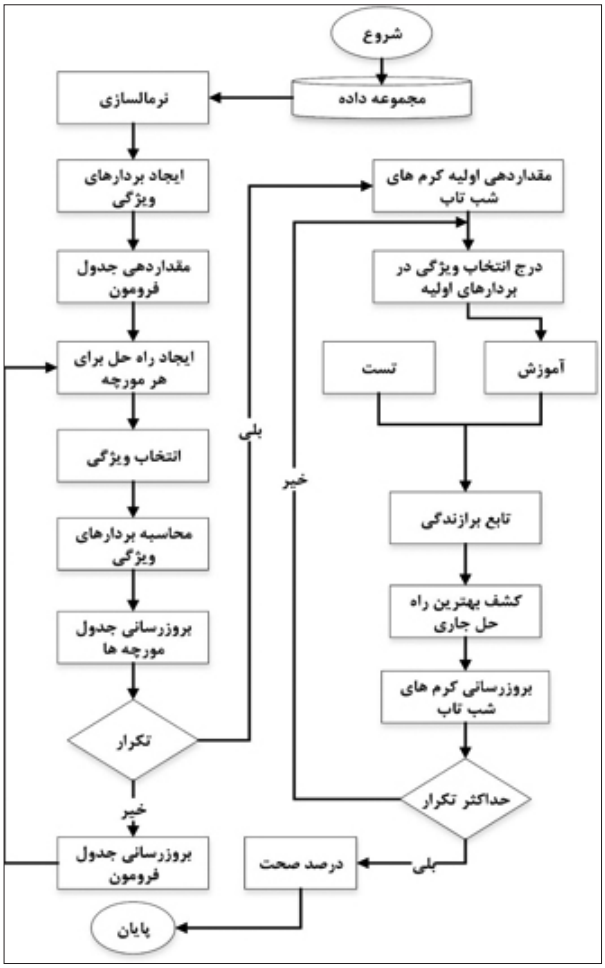
نرمال‌ســـازی داده‌ها گام مهمی در فرآیند شناســـایی ویژگی‌هـــای مهم می‌باشـــد. مجموعه داده‌ها ممکن اســـت چندان کنترل شـــده نباشـــند و مقادیر پـــرت، تکراری و یا اشـــتباه می‌تواند منجر به خروجی نامعتبر شـــوند. وجود داده‌هـــای پرت در اکثر مـــوارد منجر به اختـــلال در دقت تشخیص خواهد شد. داده پرت معمولاً در فاصله دورتری از مقدار ســـایر داده‌ها قرار می‌گیرد. وجود داده‌های پرت اساســـاً به یکی از دلایل زیر اســـت: ۱) غیر صحیح بودن اندازه‌گیری موارد ثبت شـــده یا وارد شـــده در رایانه، ۲) گردآوری اندازه‌گیری‌ها از جوامع مختلف. بنابراین انجام مجموعه فعالیت‌هایی قبل از استفاده از داده‌ها جهت تضمین کیفیت لازم است. اگر اطلاعات دارای افزونگی یا غیرمرتبط در فرآیند آموزش باشند، این مراحل بسیار مشکل خواهند شـــد. نرمال‌ســـازی داده‌ها طبق معادله(۱) انجام می‌شود [۲۳].

$$x' = (x_{\max} - x_{\min}) \times \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} + x_{\min} \qquad (۱)$$

در معادلـــه(۱)، پارامترهـــای $x_{\max}$ و $x_{\min}$ مربـــوط بـــه بیشترین و کمترین مقدار ویژگی‌ها هستند.

۳-۲ انتخاب ویژگی

در بیشتر موارد برای یافتن دانشی که در میان داده‌ها نهفته اســـت، انتخاب همه ویژگی‌ها مهم و حیاتی نیست. به همین دلیل در بسیاری از زمینه‌ها کاهش ابعاد داده یکی از مباحث قابل توجه اســـت. لذا، انتخاب ویژگی برای کاهش فضای ویژگی و در افزایش کارایی طبقه‌بندی موثر اســـت. هدف انتخاب ویژگی، انتخاب زیرمجموعه‌ای از ویژگی‌های مناسب از بین مجموعه ویژگی‌های اولیه، برای بهبود دقت طبقه‌بندی است. در مدل پیشنهادی با استفاده از مورچه‌ها، ویژگی‌ها انتخاب می‌شوند. ویژگی‌های انتخاب شده از هر تکرار با تکرار بعدی مقایســـه می‌شـــوند تا بهترین مسیر فرمون ایجاد شـــود. هر مورچه دارای یک حافظه است که



شکل ۱: روندنمای مدل پیشنهادی

بهترین راه حل به‌دســـت آمده را در آن ذخیره می‌کند. هر مورچه از این حافظه به منظور ســـاخت فرمون اســـتفاده می‌کنـــد. فرایند بهینه‌ســـازی با تعـــدادی راه حل تصادفی شـــروع می‌گردد. در هر تکـــرار از الگوریتم، هر مورچه با نمونه‌برداری از یک توزیع نرمال که ویژگی‌های آن فرمون اصلاح شده توســـط مورچه‌ها می‌باشد، یک راه حل جدید ایجاد می‌کند. حافظه مورچه‌ها برمبنای راه حل‌های ساخته شده در هر تکرار بروزرسانی می‌شود. در صورتی‌که راه حل جدید بهتر از راه حل موجود در حافظـه مورچه باشد جایگزین آن می‌شـــود. این فاز همان بروزرسانی فرمون اســـت که با هدف حرکت بـــه ســـمت بهترین راه حل انجام می‌گیرد.

در بهینه‌ســـازی کلونی مورچه باید فضای جســـتجو بـــرای انتخاب ویژگی به‌صورت یک گـــراف وزن‌دار بدون

جهت نمایش داده شـــود. این گراف به‌صورت G=(F,E) که F={F1,F2,...,Fn} یـــک مجموعـــه از ویژگی‌هایی اصلی و نمایش گره‌های گراف و $E=\{(F_i,F_j):F_i,F_j\in F\}$ یال‌های گـــراف هســـتند. وزن هر یال $(F_i,F_j)\in E$ طبـــق معادله(۲) تعریف می‌شـــود. همبســـتگی بین دو بـــردار F_i و F_j طبق معادله(۲) تعریف می‌شود.

$$w_{ij}=\left|\frac{\sum_p(x_i-\bar{x}_i)(x_j-\bar{x}_j)}{\sqrt{\sum_p(x_i-\bar{x}_i)^2}\sqrt{\sum_p(x_j-\bar{x}_j)^2}}\right| \qquad (۲)$$

در معادله(۲)، x_i و x_j بردارهای ویژگی و متغیرهای $\bar{x}_i$ و $\bar{x}_j$ میانگین مقادیـــر بردارهای x_i و x_j و p تعداد نمونه‌ها اســـت. احتمال این‌که مورچه kام ویژگی بعدی را انتخاب کند طبق معادله(۳) تعریف می‌شود.

$$P_k(i,j)=\begin{cases}\dfrac{[\tau_i]\cdot[\eta_1(F_j)]^\alpha[\eta_2(F_i,F_j)]^\beta}{\sum_{u\in J_i^k}[\tau_u][\eta_1(F_u)]^\alpha[\eta_2(F_i,F_u)]^\beta}, & \text{if } j\in J_i^k\\ 0 & otherwise\end{cases} \qquad (۳)$$

در معادله(۳)، پارامتر J_i^k مجموعه ویژگی‌های ملاقات نشـــده، پارامتر τ_u شـــدت فرمون و پارامترهـــای α و β به‌ترتیب تاثیر مقدار فرمون ریخته شـــده برروی یال‌ها و کشف ویژگی‌ها هستند. ارتباط مقدار ویژگی‌ها با استفاده از η_1 و η_2 تعیین می‌شود.

۳-۳ طبقه‌بندی داده‌ها

مرحلـــه طبقه‌بندی داده‌ها با اســـتفاده از الگوریتم کرم شب‌تاب انجام می‌شود. در این مرحله، آموزش و آزمایش داده‌ها انجام می‌گیرد. در الگوریتم کرم شـــب‌تاب در ابتدا بردارهای ویژگی‌ها را برمبنای الگوریتم بهینه‌سازی کلونی مورچه تولید می‌کنیم. ســـپس در مرحله آموزش برمبنای نزدیکی و بازه ویژگی‌ها عمل نزدیکی و تشـــابه بین مقدار ویژگی‌هـــا انجام می‌شـــود. فاصله اقلیدســـی بین دو کرم شب‌تاب به‌صورت معادله(٤) تعریف می‌شود.

$$r_{ij}=\|x_i-x_j\|=\sqrt{\sum_{k=1}^{n}(x_{i,k}-x_{j,k})^2} \qquad (٤)$$

در الگوریتم کرم شب‌تاب برای بروزرسانی بردارها از معادله(٥) استفاده می‌شود. برای این‌که طبقه‌بندی نمونه‌ها با دقت بالایی انجام شـــود از مقدار قبلی ویژگی‌ها با مقدار پیش‌بینی شـــده به منظور بروزرســـانی بردارها استفاده می‌شود.

$$F=\frac{|x_i-\bar{x}_i|}{x_i}\times(rand-\frac{1}{2}) \qquad (٥)$$

از آنجایی که معیار شـــباهت در تابع هدف بر اســـاس فاصله تعریف می‌شـــود می‌توان از تعاریف مختلفی برای اســـتفاده شـــود که در جدول(۲) چند نمونـــه از این توابع آورده شده است.

۳-٤ معیارهای ارزیابی

نتایج مدل پیشـــنهادی، باید در مرحله ارزیابی توسط معیارهـــای مهم ارزیابی مورد تحلیل قـــرار گیرد تا بتوان کارایی آن را تعیین نمود. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود. در این مقاله مهم‌ترین معیارها را برای دقت پیش‌بینی انتخاب کرده‌ایم که درصد صحت مهم‌ترین معیار در دقت پیش‌بینی است [۲٤][۲٥].

$$\text{Precision}=\frac{TP}{TP+FP} \qquad (۶)\ \text{دقت}$$

$$\text{Recall}=\frac{TP}{TP+FN} \qquad (۷)\ \text{بازخوانی}$$

$$F-Measure=\frac{2*\text{Precision}*\text{Recall}}{(\text{Precision}+\text{Recall})} \qquad (۸)\ \text{اندازه‌گیری F}$$

$$Accuracy=\frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (۹)\ \text{صحت}$$

$$ErrorRate=\frac{(FP+FN)}{(TP+TN+FP+FN)}=1-Accuracy \qquad (۱۰)\ \text{نرخ خطا}$$

پارامتر TP (مثبت درست) نشان دهنده تعداد نمونه‌هایی که جزء دســـته مثبت بوده و درســـت پیش‌بینی شـــده‌اند. پارامتر FP (مثبت نادرست) نشان دهنده تعداد نمونه‌هایی که نادرست به عنوان دسته مثبت پیش‌بینی شده‌اند. پارامتر

جدول۲: محاسبه توابع فاصله برای پیدا کردن داده‌های مشابه

تابع فاصله	فرمول
فاصله اقلیدسی (Euclidean distance)	$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
فاصله همینگ(Hamming)	$d(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$
فاصله چبیشف (chebyshev)	$d(x,y)=\max_{i=1,2,\ldots,n}\lvert x_i-y_i\rvert$
فاصله کنبرا (canberra)	$d(x,y)=\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert}{x_i+y_i}, x_i\_and\_y_i are\_positive$

FN (منفی نادرست) نشـــان دهنده تعـــداد نمونه‌هایی که نادرست به عنوان دســـته منفی پیش‌بینی شده‌اند. پارامتر TN (منفی درســـت) نشـــان دهنده تعداد نمونه‌هایی که جز دسته منفی بوده و درست پیش‌بینی شده‌اند.

۴. ارزیابی و نتایج

ارزیابی و نتایـــج برروی مجموعـــه داده Spambase برگرفتـــه از UCI بـــا ٤٦٠١ نمونـــه و ٥٧ ویژگی در محیط VC#.NET 2017 انجام شده است. به دلیل آن‌که الگوریتم‌های فرا ابتکاری از جستجوی تصادفی پیروی می‌کنند، از نظر سرعت همگرایی و رسیدن به جواب نمی‌توان تنها بر اجرای تکرار یکبار برنامه تکیه کرد. برای ارزیابی مدل پیشنهادی، مقـــدار پارامترهایی مانند جمعیت اولیه، تعداد تکرار، تعداد نسل و مقدار فرمون به‌ترتیب برابر ۱۰۰، ۱۰۰، ۱۰۰ و ۰,۱ است و همچنین مرحله آموزش و آزمایش به‌ترتیب برابر با ۸۰ و ۲۰ درصد می‌باشند. همچنین طول بردار در الگوریتم کرم شب‌تاب به‌صورت متغیر تعیین شده است.

٤-١ ارزیابی مدل‌ها بدون انتخاب ویژگی

در جدول(۳)، ارزیابی مدل‌ها بدون انتخاب ویژگی و با ۱۰۰ بار تکرار انجام شده‌اند. طبقه‌بندی در مدل پیشنهادی برمنبای آموزش و آزمایش نمونه‌ها انجام شـــده اســـت. از بهینه‌ســـازی کلونی مورچه برای انتخـــاب ویژگی و از

جدول۳: ارزیابی مدل‌ها بدون انتخاب ویژگی و با ۱۰۰ بار تکرار

مدل پیشنهادی	الگوریتم کرم شب‌تاب	بهینه‌سازی کلونی مورچه	معیارها
۷۳,۲۵	۶۷,۱۱	۶۹,۵۱	دقت
۷۸,۱۵	۶۹,۰۷	۷۱,۳۲	بازخوانی
۷۵,۶۲	۶۸,۰۸	۷۰,۴۰	اندازه‌گیری F
۸۹,۰۰	۷۶,۲۵	۷۷,۸۴	صحت
۰,۱۱۰۰	۰,۲۳۷۵	۰,۲۳۱۶	نرخ خطا

الگوریتم کرم شـــب‌تاب برای آموزش و آزمایش نمونه‌ها استفاده شده است.

٤-٢ ارزیابی مدل پیشـــنهادی برمبنای انتخاب ویژگی و تعداد تکرار

در جدول(٤)، نتایج ارزیابی مدل پیشـــنهادی برمبنای انتخاب ویژگی و با ۱۰۰ بار تکرار و با استفاده از معیارهای مختلف مورد ارزیابی قرار گرفته است.

در جدول(٥)، نتایج ارزیابی مدل پیشـــنهادی برمبنای انتخاب ویژگی و با ۲۰۰ بار تکرار و با استفاده از معیارهای مختلف مورد ارزیابی قرار گرفته است.

در جـــدول(٦)، نتایج ارزیابی معیـــار صحت، برمبنای انتخاب ویژگی‌های مختلف با ۱۰۰ بار تکرار نشان داده شده است. در جدول(٦) واضح است که با کاهش انتخاب ویژگی مقـــدار دقت افزایش یافته اســـت. زیرا در مدل پیشـــنهادی برای طبقه‌بندی داده‌ها از ویژگی‌های موثر اســـتفاده شده

جدول۴: ارزیابی مدل پیشنهادی برمبنای انتخاب ویژگی و با ۱۰۰ بار تکرار

معیارها	انتخاب ویژگی							
	۲۰	۲۷	۳۲	۴۰	۴۹	۵۰	۵۳	۵۷
دقت	۹۰,۳۶	۹۲,۸۵	۹۱,۳۵	۹۴,۷۹	۹۵,۱۴	۹۲,۴۳	۹۰,۱۲	۸۹,۳۷
بازخوانی	۹۱,۰۶	۹۳,۶۴	۹۲,۰۸	۹۵,۴۸	۹۵,۶۶	۹۳,۱۰	۹۱,۵۲	۹۰,۳۲
اندازه‌گیری F	۸۹,۱۱	۹۱,۰۸	۸۸,۵۷	۹۲,۱۷	۹۲,۸۲	۸۹,۵۹	۸۹,۶۷	۸۸,۱۷
صحت	۹۲,۳۶	۹۱,۸۶	۹۱,۵۲	۹۰,۷۶	۸۹,۷۱	۸۸,۰۹	۹۰,۱۲	۸۹,۰۰
نرخ خطا	۷,۶۴	۸,۱۴	۸,۴۸	۹,۲۴	۱۰,۲۹	۱۱,۹۱	۹,۸۸	۱۱,۰۰

جدول۵: ارزیابی مدل پیشنهادی برمبنای انتخاب ویژگی و با ۲۰۰ بار تکرار

معیارها	انتخاب ویژگی							
	۲۰	۲۷	۳۲	۴۰	۴۹	۵۰	۵۳	۵۷
دقت	۹۵,۰۲	۹۴,۳۶	۹۳,۰۷	۹۶,۰۸	۹۵,۱۴	۹۲,۴۳	۹۰,۱۲	۸۹,۳۷
بازخوانی	۹۶,۱۶	۹۵,۹۷	۹۳,۹۱	۹۷,۹۶	۹۶,۹۴	۹۳,۴۵	۹۱,۰۱	۹۱,۹۰
اندازه‌گیری F	۹۰,۴۴	۹۰,۷۵	۹۰,۳۸	۹۳,۵۷	۹۲,۴۳	۹۱,۴۲	۸۹,۶۶	۹۰,۵۹
صحت	۹۵,۴۷	۹۴,۳۵	۹۴,۰۵	۹۳,۶۱	۹۳,۲۷	۹۳,۶۵	۹۲,۸۶	۹۲,۰۵
نرخ خطا	۴,۵۳	۵,۶۵	۵,۹۵	۶,۳۹	۶,۷۳	۶,۳۵	۷,۱۴	۷,۹۵

جدول ٦: ارزیابی معیار صحت با انتخاب ویژگی‌های مختلف با ۱۰۰ بار تکرار

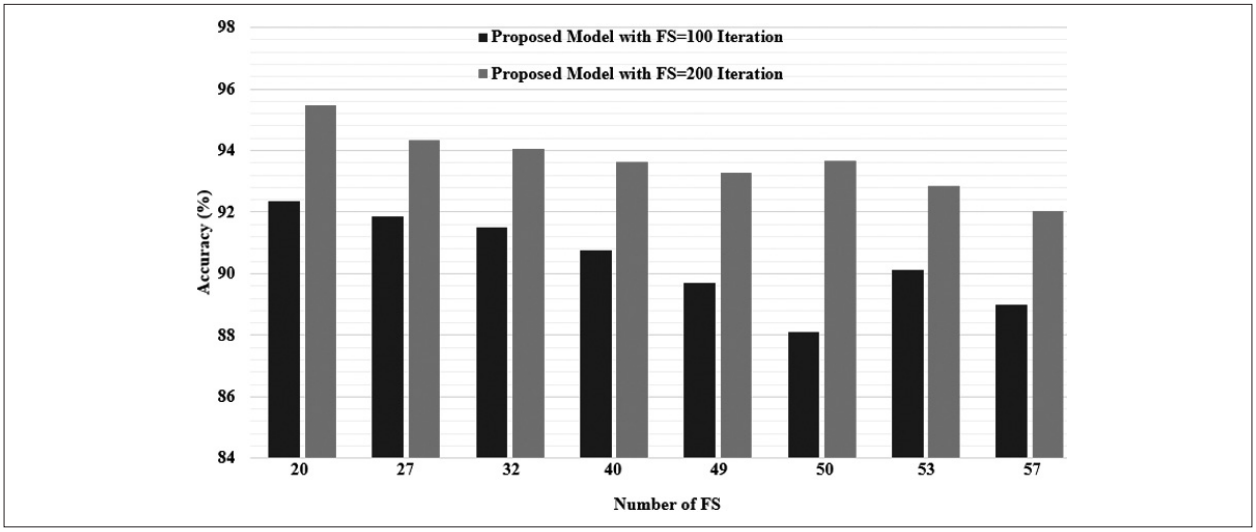
درصد صحت	ویژگی‌های انتخاب شده	تعداد ویژگی
۹۲,۳۶	۲, ۳۵, ۲۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۳۷, ۵, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۱۳, ۱۶, ۱۹, ۵۶	۲۰
۹۱,۸۶	۳, ۱۱, ۲۶, ۱۷, ۱۳, ۹, ۳۵, ۷, ۱۶, ۱۹, ۵۶, ۵, ۲۴, ۲۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۳۷, ۵۷, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰	۲۷
۹۱,۵۲	۱۷, ۳۷, ۵۷, ۸, ۲۰, ۵۵, ۶, ۱۱, ۲۶, ۱۶, ۱۹, ۵۶, ۳۵, ۱۲, ۱۵,۲۸, ۴, ۱۰, ۴۸, ۵۰, ۲۲, ۵, ۲۴, ۲۷, ۳, ۲۳, ۲۵, ۷, ۵۲, ۵۳, ۲۱, ۲	۳۲
۹۰,۷۶	۳۱,۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۵۷, ۸, ۱۲, ۱۵, ۱, ۲۷, ۱, ۲۹, ۴۶, ۱۰, ۱۳, ۲۲, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۲, ۹, ۳۵, ۴۸, ۳۲, ۴۵, ۱۴, ۳۳, ۲۴, ۲۷, ۳, ۲۶	۴۰
۸۹,۷۱	۱۲, ۵۵, ۲۸, ۱, ۲۹, ۱۱, ۱۰, ۱۳, ۳۱, ۳۲, ۴۰, ۴۲, ۲۶, ۱۷, ۳۷, ۵۷, ۸, ۲۰, ۱۵, ۶, ۱۸, ۳۰, ۲, ۹, ۵۴, ۴, ۱۴, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۲۴, ۲۷, ۳, ۴۶, ۳۶, ۳۵, ۵۶, ۲۰, ۲۲, ۳۳, ۳۴, ۳۷, ۳۸, ۷	۴۹
۸۸,۰۹	۵۴, ۴, ۱۴, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۱۲, ۵۵, ۲۸, ۱, ۲۹, ۱۱, ۱۰, ۱۳, ۳۱, ۳۲, ۴۰, ۴۲, ۲۶, ۱۷, ۳۷, ۵۷, ۸, ۲۷, ۳, ۴۶, ۳۶, ۳۵, ۵۶, ۲۰, ۲۲, ۳۳, ۳۴, ۳۷, ۳۸, ۷, ۴۸, ۲۰, ۱۵, ۶, ۱۸, ۳۰, ۲, ۹, ۲۴	۵۰
۹۰,۱۲	۴۹, ۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۴, ۱۴, ۲۲, ۳۶, ۳۹, ۵, ۲۴, ۲۷, ۱۶, ۱۹, ۵۶, ۲۸, ۱, ۲۹, ۴۶, ۱۰, ۱۳, ۳, ۳۲, ۳۱, ۱۱, ۵۷, ۱۷, ۳۷, ۲۶, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۲, ۹, ۳۵, ۱۲, ۱۶, ۳۴, ۴۰, ۴۲, ۳۳, ۴۵, ۵۴, ۴۱, ۴۳, ۴۴, ۴۸	۵۳
۸۹,۰۰	انتخاب همه ویژگی‌ها	۵۷

است. معیار صحت برای مقایسه انتخاب ویژگی‌ها در مدل پیشنهادی استفاده شده است. در بیشتر تحقیقات انجام شده از معیار صحت به عنوان معیار اصلی برای ارزیابی استفاده شده است. در جدول(٦) مقایسه دقت مدل پیشنهادی برمبنای انتخاب ویژگی انجام شده و ویژگی‌های مختلف مقدار دقت متفاوتی دارند. لذا تعداد ویژگی در دقت مدل پیشنهادی تاثیرگذار بوده است.

در جدول(۷)، نتایج ارزیابی معیار صحت برمبنای انتخاب ویژگی‌های مختلف با ۲۰۰ بار تکرار نشان داده شده است. در جدول(۷) واضح است که با کاهش انتخاب ویژگی

جدول۷: ارزیابی معیار صحت با انتخاب ویژگی‌های مختلف با ۲۰۰ بار تکرار

درصد صحت	ویژگی‌های انتخاب شده	تعداد ویژگی
۹۵،۴۷	۲, ۳۵, ۲۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۳۷, ۵, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۱۳, ۱۶, ۱۹, ۵۶	۲۰
۹۴،۳۵	۳, ۱۱, ۲۶, ۱۷, ۱۳, ۹, ۳۵, ۷, ۱۶, ۱۹, ۵۶, ۵, ۲۴, ۲۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۳۷, ۵۷, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰	۲۷
۹۴،۰۵	۱۷, ۳۷, ۵۷, ۸, ۲۰, ۵۵, ۶, ۱۱, ۲۶, ۱۶, ۱۹, ۵۶, ۳۵, ۱۲, ۱۵،۲۸, ۴, ۱۰, ۴۸, ۵۰, ۲۲, ۵, ۲۴, ۲۷, ۳, ۲۳, ۲۵, ۷, ۵۲, ۵۳, ۲۱, ۲	۳۲
۹۳،۶۱	۳۱،۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۵۷, ۸, ۱۲, ۱۵, ۲۷, ۱, ۲۹, ۴۶, ۱۰, ۱۳, ۲۲, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۲, ۹, ۳۵, ۴۸, ۳۲, ۴۵, ۱۴, ۳۳, ۲۴, ۲۷, ۳, ۲۶	۴۰
۹۳،۲۷	۱۲, ۵۵, ۲۸, ۱, ۲۹, ۱۱, ۱۰, ۱۳, ۳۱, ۳۲, ۴۰, ۴۲, ۲۶, ۱۷, ۳۷, ۵۷, ۸, ۲۰, ۱۵, ۶, ۱۸, ۳۰, ۲, ۹, ۵۴, ۴, ۱۴, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۲۴, ۲۷, ۳, ۴۶, ۳۶, ۳۵, ۵۶, ۲۰, ۲۲, ۳۳, ۳۴, ۳۷, ۳۸, ۷	۴۹
۹۳،۶۵	۵۴, ۴, ۱۴, ۵۳, ۲۱, ۲۳, ۲۵, ۱۶, ۱۹, ۵۶, ۵, ۱۲, ۵۵, ۲۸, ۱, ۲۹, ۱۱, ۱۰, ۱۳, ۳۱, ۳۲, ۴۰, ۴۲, ۲۶, ۱۷, ۳۷, ۵۷, ۸, ۲۷, ۳, ۴۶, ۳۶, ۳۵, ۵۶, ۲۰, ۲۲, ۳۳, ۳۴, ۳۷, ۳۸, ۷, ۴۸, ۲۰, ۱۵, ۶, ۱۸, ۳۰, ۲, ۹, ۲۴	۵۰
۹۲،۸۶	۴۹, ۷, ۵۲, ۵۳, ۲۱, ۲۳, ۲۵, ۴, ۱۴, ۲۲, ۳۶, ۳۹, ۵, ۲۴, ۲۷, ۱۶, ۱۹, ۵۶, ۲۸, ۱, ۲۹, ۴۶, ۱۰, ۱۳, ۳, ۳۲, ۳۱, ۱۱, ۵۷, ۱۷, ۳۷, ۲۶, ۸, ۲۰, ۵۵, ۶, ۱۸, ۳۰, ۲, ۹, ۳۵, ۱۲, ۱۶, ۳۴, ۴۰, ۴۲, ۳۳, ۴۵, ۵۴, ۴۱, ۴۳, ۴۴, ۴۸	۵۳
۹۲،۰۵	انتخاب همه ویژگی‌ها	۵۷



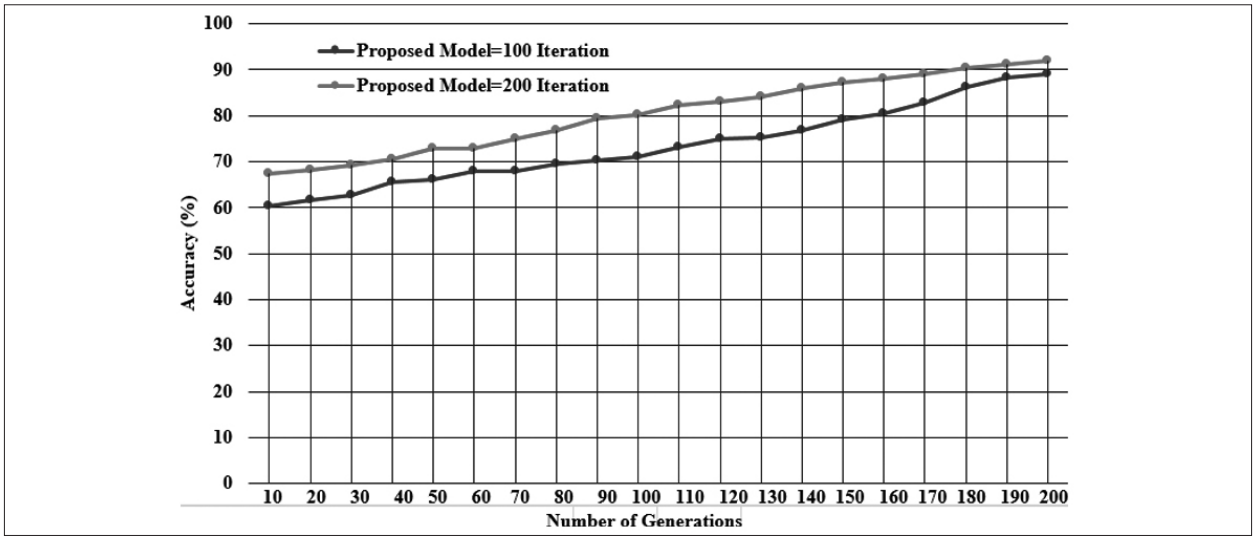
شکل۲: نمودار مقایسه مدل پیشنهادی با انتخاب ویژگی برمبنای تعداد تکرار

مقدار دقت افزایش یافته است. زیرا در مدل پیشنهادی برای طبقه‌بندی داده‌ها از ویژگی‌های موثر استفاده شده است.

در شـــکل(۲)، نمودار مقایسه مدل پیشنهادی با انتخاب ویژگی برمبنای تعداد تکرار نشان داده شده است. همانطور که در شکل(۲)، مشاهده می‌کنید مدل پیشنهادی با انتخاب ویژگی و با ۲۰۰ بار تکرار، درصد صحت بیشتری دارد.

٤-۳ ارزیابی مدل پیشنهادی بر مبنای تعداد نسل

در شـــکل(۳) ارزیابـــی مدل پیشـــنهادی برمبنای ۲۰۰ نسل نشان داده شده است. تعداد نسل بهینه می‌تواند دقت تشخیص را افزایش دهد. زیرا مدل پیشنهادی در هر چرخه از نسل می‌تواند ویژگی‌های متنوع و موثرتری را کشف کند و در نتیجه در طبقه‌بندی نمونه‌ها موثر است. الگوریتم‌های فرا ابتکاری برمبنای تعداد نســـل کاوش در محیط را انجام می‌دهنـــد و تعداد نســـل‌ها بهینه به مدل پیشـــنهادی کمک می‌کنـــد تا محیط جســـتجو را به خوبـــی ارزیابی کند. در شـــکل(۳)، مقایســـه برمبنای ۱۰۰ و ۲۰۰ تکرار انجام شده اســـت. برمبنای اجراهای مختلف به این نتیجه رسیدیم که



شکل۳: ارزیابی مدل پیشنهادی برمبنای تعداد نسل

جدول۸: ارزیابی مدل پیشنهادی برمبنای معیار فاصله همینگ و با ۱۰۰ بار تکرار

معیار فاصله	$d(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$							
انتخاب ویژگی	۲۰	۲۷	۳۲	۴۰	۴۹	۵۰	۵۳	۵۷
دقت	۰.۹۶۲۳	۰.۹۵۱۲	۰.۹۴۶۳	۰.۹۳۴۹	۰.۹۲۵۸	۰.۹۱۶۸	۰.۹۱۳۲	۰.۹۰۱۰
بازخوانی	۰.۹۶۸۵	۰.۹۶۳۵	۰.۹۴۸۷	۰.۹۴۶۸	۰.۹۲۶۸	۰.۹۲۹۸	۰.۹۱۵۶	۰.۹۰۲۱
اندازه‌گیری F	۰.۹۶۵۴	۰.۹۵۷۳	۰.۹۴۷۵	۰.۹۴۰۸	۰.۹۲۶۳	۰.۹۲۳۳	۰.۹۱۴۴	۰.۹۰۱۵
صحت	۰.۹۶۴۲	۰.۹۶۸۴	۰.۹۵۰۸	۰.۹۴۸۲	۰.۹۳۲۱	۰.۹۲۸۷	۰.۹۱۶۵	۰.۹۰۴۶
نرخ خطا	۰.۰۳۵۸	۰.۰۳۱۶	۰.۰۴۹۲	۰.۰۵۱۸	۰.۰۶۷۹	۰.۰۷۱۳	۰.۰۹۳۵	۰.۰۹۵۴

تکرارهـــای ۱۰۰ و ۲۰۰ برای مدل پیشـــنهادی حالت بهینه دارنـــد. در بـــازه ۱۰۰ تا ۲۰۰ مدل پیشـــنهادی می‌تواند به درصد صحت برازنده دســـت یابد و مقادیر بیشـــتر از این بـــازه در درصد صحت تاثیر چندانـــی ندارند و فقط باعث می‌شوند که زمان محاسباتی افزایش یابد.

٤-٤ ارزیابی مدل پیشنهادی برمبنای معیارهای فاصله

در جدول(۸)، نتایج ارزیابی مدل پیشنهادی برمبنای معیـــار فاصله همینگ و با ۱۰۰ بار تکرار و با اســـتفاده از معیارهـــای مختلف مورد ارزیابی قرار گرفته اســـت. در جدول(۸)، بیشـــترین درصد صحت برابر با ۰.۹٦٤۲ است.

در جدول(۹)، نتایج ارزیابی مدل پیشنهادی برمبنای معیار فاصله چبیشـــف و با ۱۰۰ بار تکرار و با استفاده از معیارهـــای مختلف مورد ارزیابی قرار گرفته اســـت. در جدول(۹)، بیشـــترین درصد صحت برابر با ۰.۹۵۲۱ است.

در جدول(۱۰)، نتایج ارزیابی مدل پیشنهادی برمبنای معیـــار فاصله کنبرا و با ۱۰۰ بار تکرار و با اســـتفاده از معیارهای مختلـــف مورد ارزیابی قرار گرفته اســـت. در جدول(۱۰)، بیشترین درصد صحت برابر با ۹٦,۹٤ است.

٤-٥ مقایسه و ارزیابی

مقایســـه و ارزیابـــی به منظور کارایـــی و برتری مدل پیشنهادی در مقایســـه با مدل‌های دیگر انجام شده است. مدل پیشنهادی در مقایسه با اغلب الگوریتم‌های فرا ابتکاری و داده کاوی از دقت بیشتری بهره‌مند است. یکی از دلایلی که باعث شده است که مدل پیشنهادی دقت بیشتری داشته

جدول۹: ارزیابی مدل پیشنهادی برمبنای معیار فاصله چبیشف و با ۱۰۰ بار تکرار

معیار فاصله	$d(x,y)=\max_{i=1,2,\ldots,n}\lvert x_i-y_i\rvert$							
انتخاب ویژگی	۲۰	۲۷	۳۲	۴۰	۴۹	۵۰	۵۳	۵۷
دقت	۹۵,۱۶	۹۴,۹۱	۹۳,۲۱	۹۲,۸۴	۹۱,۳۶	۹۰,۵۴	۸۹,۶۴	۸۷,۹۵
بازخوانی	۹۶,۳۱	۹۵,۸۸	۹۳,۵۴	۹۳,۲۶	۹۱,۵۸	۹۱,۶۴	۸۹,۹۶	۸۹,۱۵
اندازه‌گیری F	۹۵,۷۳	۹۵,۳۹	۹۳,۳۷	۹۳,۰۵	۹۱,۴۷	۹۱,۰۹	۸۹,۸۰	۸۸,۵۵
صحت	۹۵,۲۱	۹۴,۶۸	۹۳,۰۸	۹۲,۶۳	۹۱,۳۵	۹۰,۴۸	۸۹,۱۳	۸۹,۵۱
نرخ خطا	۴,۷۹	۵,۳۲	۶,۹۲	۷,۳۷	۸,۶۵	۹,۵۲	۱۰,۸۷	۱۰,۴۹

جدول۱۰: ارزیابی مدل پیشنهادی برمبنای معیار فاصله کنبرا و با ۱۰۰ بار تکرار

معیار فاصله	$d(x,y)=\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert}{x_i+y_i}$							
انتخاب ویژگی	۲۰	۲۷	۳۲	۴۰	۴۹	۵۰	۵۳	۵۷
دقت	۹۷,۵۲	۹۶,۴۳	۹۴,۶۸	۹۳,۴۸	۹۲,۳۵	۹۱,۸۷	۹۱,۳۴	۹۰,۱۶
بازخوانی	۹۷,۸۲	۹۶,۵۶	۹۵,۱۳	۹۳,۶۷	۹۳,۶۵	۹۱,۹۰	۹۱,۸۴	۹۱,۵۴
اندازه‌گیری F	۹۷,۶۷	۹۶,۴۹	۹۴,۹۰	۹۳,۵۷	۹۳,۰۰	۹۱,۸۸	۹۱,۵۹	۹۰,۸۴
صحت	۹۶,۹۴	۹۵,۸۹	۹۴,۹۷	۹۳,۵۲	۹۲,۸۴	۹۱,۳۴	۹۰,۴۸	۹۰,۶۱
نرخ خطا	۳,۰۶	۴,۱۱	۵,۰۳	۶,۴۸	۷,۱۶	۸,۶۶	۹,۵۲	۹,۳۹

باشد استفاده از ترکیب دو مدل فرا ابتکاری و بهره‌مندی از عملگرهایی مانند بروزرسانی، جستجو برای بهترین بردار ویژگی، انتخاب ویژگی برمبنای تشابه و نزدیکی و تعداد نسل‌ها است. در جدول(۱۱)، مقایسه مدل پیشنهادی با مدل‌های دیگر نشان داده شده است.

در جدول(۱۱)، مدل‌های بهینه‌سازی اجتماع ذرات- الگوریتم انتخاب منفی، الگوریتم انتخاب منفی، بهینه‌سازی اجتماع ذرات، بوستینگ منطقی و بیزین ساده کمترین درصد صحت را در مقایسه با مدل پیشنهادی و مدل‌های دیگر دارند. و همچنین مدل پیشنهادی برمبنای تعداد تکرار و تعداد ویژگی‌ها نتایج متفاوتی دارد. مدل پیشنهادی هم مانند مدل‌های دیگر بنابه دلایلی مانند عدم تنوع در جمعیت اولیه و گیرافتادن در بهینه محلی نمی‌توانند به طور صد در صد پاسخگوی سیستم تشخیص ایمیل هرزنامه باشد. اما درصد صحت مدل پیشنهادی با کاهش تعداد ویژگی‌ها، افزایش می‌یابد.

۵. نتیجه‌گیری و کارهای آینده

بر اساس بررسی انجام شده، هرزنامه بیش از چهل درصد از ترافیک ایمیل در اینترنت را شامل می‌شود. که این ترافیک باعث کاهش افت شدید سرعت اینترنت و همچنین کاهش کارایی کارسازها در پردازش داده‌ها می‌شوند. مهم‌ترین راهکار برای مقابله با هرزنامه، شناسایی و تشخیص نوع ایمیل‌های ارسالی از سوی هرزنامه‌نویسان است. در این مقاله، برای تشخیص هرزنامه از مدل ترکیبی بهینه‌سازی کلونی مورچه و الگوریتم کرم شب‌تاب برروی مجموعه داده Spambase استفاده شد. نتایج نشان داد که مدل پیشنهادی در طبقه‌بندی دقت بالایی دارد. و همچین کارایی مدل پیشنهادی برمبنای انتخاب ویژگی و تکرار نشان داد که با انتخاب ویژگی‌های مهم می‌توان به دقت بیشتری دست یافت. درصد صحت مدل پیشنهادی با ۲۰۰ و ۱۰۰ بار تکرار برای همه ویژگی‌ها به‌ترتیب برابر ۸۹,۰۰ و ۹۲,۰۵ به‌دست آمد. زیرا با ۲۰۰ بار تکرار،

جدول۱۱: مقایسه مدل پیشنهادی با مدل‌های دیگر

رفرنس‌ها	مدل‌ها	صحت	دقت	بازخوانی	اندازه‌گیری F
[۱۰]	روش گروهی مدیریت داده‌ها	۹۱,۷	۹۳,۵	۸۸,۲	۹۰,۵
[۱۱]	بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی	۸۳,۲۰	۸۶,۷۱	۶۵,۹۹	۷۴,۹۵
[۱۸]	الگوریتم انتخاب منفی	۶۸,۸۶	۹۴,۵۳	۲۲,۲۴	۳۶,۰۱
	تکاملی تفاضلی-الگوریتم انتخاب منفی	۸۰,۶۶	۹۰,۸۶	۵۶,۶۲	۶۹,۷۶
[۱۹]	بهینه‌سازی اجتماع ذرات-الگوریتم انتخاب منفی	۹۱,۲۲	۸۱,۳۶	۳۲,۳۱	۴۳,۵۳
	بهینه‌سازی اجتماع ذرات	۸۱,۳۲	۸۱,۷۷	۲۸,۶۴	۳۸,۹۴
	الگوریتم انتخاب منفی	۶۸,۸۶	۸۵,۰۲	۱۳,۶۳	۲۲,۰۹
[۲۶]	شبکه بیزین	۸۸,۵۶	۹۳,۱۱	۸۸,۶	۹۰,۸
	بوستینگ منطقی	۸۹,۷	۹۲,۸۶	۹۰,۴	۹۱,۶۶
	درخت تصادفی	۹۱,۵۴	۹۲,۷۵	۹۳,۲	۹۳,۰۰
	JRIP	۹۲,۳۲	۹۴,۴۷	۹۲,۹	۹۳,۷۱
	درخت J۴۸	۹۲,۳۴	۹۳,۹۰	۹۳,۵	۹۳,۷
	شبکه عصبی مصنوعی چندلایه	۹۳,۲۸	۹۴,۳۳	۹۴,۵	۹۴,۴۵
	الگوریتم Kstar	۹۳,۵۶	۹۵,۵۸	۹۳,۹	۹۴,۷۳
	جنگل تصادفی	۹۳,۸۹	۹۵,۸۷	۹۴,۱	۹۵,۰۰
	کمیته تصادفی	۹۴,۲۸	۹۶,۱۲	۹۴,۵	۹۵,۳۲
[۲۷]	بیزین ساده	۷۹,۲۸	۴۳,۲۶	۹۲,۶۴	۵۹,۹۷
	شبکه بیزین	۸۹,۸۰	۲۱,۴۷	۶۱,۰۸	۳۱,۷۷
	ماشین بردار پشتیبان	۹۰,۴۱	۲۰,۰۷	۶۳,۳۵	۳۰,۴۸
	درخت توابع	۹۳,۳۴	۱۵,۵۲	۵۰,۴۹	۲۳,۷۴
	J۴۸	۹۲,۹۷	۱۸,۶۸	۵۲,۴۳	۲۷,۵۴
	جنگل تصادفی	۹۴,۸۲	۲۰,۱۲	۴۲,۲۳	۲۷,۲۵
	درخت تصادفی	۹۰,۹۳	۱۸,۹۱	۶۱,۴۰	۲۸,۹۱
	درخت نمونه CART	۹۲,۴۳	۲۲,۰۹	۵۳,۳۳	۳۱,۲۴
مدل پیشنهادی	انتخاب همه ویژگی‌ها و تکرار برابر با ۱۰۰	۸۹,۰۰	۸۹,۳۷	۹۰,۳۲	۸۸,۱۷
	انتخاب همه ویژگی‌ها و تکرار برابر با ۲۰۰	۹۲,۰۵	۸۹,۳۷	۹۱,۹۰	۹۰,۵۹
	انتخاب ۲۰ ویژگی و تکرار برابر با ۱۰۰	۹۲,۳۶	۹۰,۳۶	۹۱,۰۶	۸۹,۱۱
	انتخاب ۲۰ ویژگی و تکرار برابر با ۲۰۰	۹۵,۴۷	۹۵,۰۲	۹۶,۱۶	۹۰,۴۴

جستجو در فضای محیط از تنوع و اکتشاف بیشتری بهره‌مند بوده است. برای جلوگیری و کاهش هرزنامه باید از الگوریتم‌های هوشمند استفاده شود. بدین‌گونه که قوانین هر هرزنامه شناسایی شود و از دریافت آن‌ها در دفعات بعدی جلوگیری شود. و همچنین برمبنای آموزش باید مدل‌هایی ارائه دهیم که با استفاده از عنوان و محتوا هرزنامه را تشخیص دهد.

منابع

1. F.S. Gharehchopogh, M. Vafadar, M. Motaman, Improve of Invasive Weed Optimization algorithm with K nearest neighbor for email spam classification, Computing Science Journal (CSJ), in press, March 2018.
2. H. Faris, A.M. Al-Zoubi, A.A. Heidari, I. Aljarah, H. Fujita, An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks, Information Fusion, Vol. 48, pp. 67-83, 2019.
3. A.F. Colladon, P.A. Gloor, Measuring the impact of

spammers on e-mail and Twitter networks, International Journal of Information Management, In press, corrected proof, Available online 24 September 2018
4. A. Heydari, M.A. Tavakoli, N. Salim, Z. Heydari, Detection of review spam: A survey, Expert Systems with Applications, Vol. 42, Issue 7, pp. 3634-3642, 2015
5. G. Dalkılıc, D. Sipahi, Spam filtering with sender authentication network, Computer Communications, Vol. 98, pp. 72-79, 2017.
6. M.Dorigo, L.M. Gambardella, The colony system: A cooperative learning approach to the traveling salesman problem, IEEE Transactions on Evolutionary Computation, Vol. 1, No.1, pp. 53-66, April 1997.
7. X.S. Yang, Nature-Inspired Meta-heuristic Algorithms, Luniver Press, 2008.
8. https://archive.ics.uci.edu/ml/datasets/Spambase
9. Siti-Hajar-Aminah Ali, S. Ozawa, J. Nakazato, T. Ban, J. Shimamura, An autonomous online malicious spam email detection system using extended RBF network, International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2015.
10. El-Sayed M. El-Alfy, R.E. Abdel-Aal, Using GMDH-based networks for improved spam detectionand email feature analysis, Applied Soft Computing, Vol. 11, Issue 1, pp. 477-488, 2011.
11. I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, M. Penhaker, A combined negative selection algorithm–particle swarm optimization for an email spam detection system, Engineering Applications of Artificial Intelligence, Vol. 39, pp. 33-44, 2015
12. L. Nguyen, A.Q. Tran, L. Thu Bui, DMEA-II and its application on spam email detection problems, Seventh IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), IEEE, pp. 1-6, 2014
13. Mu-Chun Su, Hsu-Hsun Lo, Fu-Hau Hsu, A neural tree, and its application to spam e-mail detection, Expert Systems with Applications, Vol. 37, pp. 7976-7985, 2010.
14. S. Ergin, S. Isik, The investigation on the effect of feature vector dimension for spam email detection with a new framework, 9th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-4, 2014
15. W. Ma, D. Tran, D. Sharma, A Novel Spam Email Detection System Based on Negative Selection, 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, pp. 987-992, 2009.
16. A.R. Behjat, A. Mustapha, H. Nezamabadi-pour, Md. Nasir Sulaiman, and N. Mustapha, A PSOBased Feature Subset Selection for Application of Spam /Non-spam Detection, Springer-Verlag Berlin Heidelberg, M-CAIT 2013, CCIS 378, pp. 183-193, 2013.
17. Kuo-Ching Ying, Shih-Wei Lin, Zne-Jung Lee, Yen-Tim Lin, An ensemble approach applied to classify spam e-mails, Expert Systems with Applications, Vol. 37, Issue 3, pp. 2197-2201, 2010.
18. I. Idris, A. Selamat, S. Omatu, Hybrid email spam detection model with negative selection algorithm and differential evolution, Engineering Applications of Artificial Intelligence, Vol. 28, 97-110, 2014
19. I. Idris, A. Selamat, Improved email spam detection model with negative selection algorithm and particle swarm optimization, Applied Soft Computing, Vol. 22, pp. 11-27, 2014.
20. S.B. Rathod, T.M. Pattewar, Content based spam detection in email using Bayesian classifier, International Conference on Communications and Signal Processing (ICCSP), IEEE, pp. 1257-1261, 2015.
21. Chi-Yao Tseng; Ming-Syan Chen, Incremental SVM Model for Spam Detection on Dynamic Email Social Networks, 2009 International Conference on Computational Science and Engineering, Vol. 4, pp. 128-135, 2009.
22. M. Prilepok, T. Jezowicz, J. Platos, V. Snasel, Spam detection using compression and PSO, Fourth International Conference on Computational Aspects of Social Networks (CASoN), pp. 263-270, 2012.
23. B.K. Singh, K. Verma, A.S. Thoke, Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification, International Journal of Computer Applications, Vol. 116, No. 19, pp. 11-15, 2015.
24. R.S. Michalski, I. Bratko, M. Kubat, Machine Learning, and Data Mining: Methods and Applications, New York: Wiley, 1998.
25. D. Francois, Binary classification performances measure cheat sheet, 2009.
26. S. Sharma, A. Arora, Adaptive Approach for Spam Detection, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No. 1, pp. 23-26, July 2013.
27. M. Rathi, V. Pareek, Spam Mail Detection through Data Mining-A Comparative Performance Analysis, I.J. Modern Education and Computer Science, Vol. 12, pp. 31-39, 2013.