

ارائه مدل ترکیبی الگوریتم K نزدیک ترین همسایه با الگوریتم بهینه‌سازی ازدحام ذرات برای تشخیص رایانامه‌های هرزنامه

عاطفه مرتضوی و فرهاد سلیمانیان قره چپق*

گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
Ati.mortazavi67@gmail.com
farhad@iaurmia.ac.ir

چکیده

در دنیای امروزی، رایانامه یکی از روش‌های ارتباطات سریع است و یکی از معضلات رایانامه، هرزنامه‌ها هستند که در سال‌های اخیر رشد فراوانی داشته‌اند. هرزنامه نه تنها باعث آسیب‌رساندن به منافع کاربران، مصرف زمان و پهنه‌ی باند می‌شوند، بلکه به عنوان تهدیدی برای بهره‌وری، قابلیت اطمینان و امنیت شبکه شده‌اند. هرزنامه‌نویس‌ها همیشه سعی می‌کنند راه‌هایی برای فرار از فیلترهای موجود بیابند؛ بنابراین فیلترهای جدید برای تشخیص هرزنامه‌ها نیاز به توسعه دارند. بیشتر این فیلترها از ترکیبی از چندین روش مانند روش‌های سیاه و سفید، استفاده از واژه‌های کلیدی، فیلترهایی بر پایه قانون، روش‌های یادگیری ماشین و غیره برای شناسایی دقیق‌تر هرزنامه‌ها بهره می‌برند. در این مقاله، مدلی جدید با استفاده از دو الگوریتم قدرتمند K نزدیک‌ترین همسایه و الگوریتم بهینه‌سازی ازدحام ذرات برای شناسایی هرزنامه‌ها ارائه شده است که از الگوریتم بهینه‌سازی اجتماعی ازدحام ذرات برای انتخاب ویژگی و از الگوریتم K نزدیک‌ترین همسایه برای دسته‌بندی استفاده شده است. شبیه‌سازی برروی مجموعه‌داده spambase اجرا شده و کارایی مدل پیشنهادی با استفاده از معیارهای صحت، دقت^۱، فراخوانی^۲ و F-measure^۳ ارزیابی شده است. نتایج نشان می‌دهد که مدل پیشنهادی در مقایسه با مدل‌های مشابه و از خود الگوریتم‌های پایه بهتر عمل کرده و کارایی بهتری دارد.

واژگان کلیدی: الگوریتم K نزدیک‌ترین همسایه، الگوریتم بهینه‌سازی ازدحام ذرات، رایانامه هرزنامه، تشخیص هرزنامه، بهینه‌سازی

مسأله‌ای که روزی کم‌اهمیت جلوه می‌شد، امروزه به معضلي جدی برای میلیون‌ها کاربر رایانامه بدل شده و استفاده از ابزار و روش‌هایی برای شناسایی و فیلتر رایانامه هرزنامه‌ها ضروری نیز غیرقابل انکار است. کاربرد فراوان رایانامه با مشکلاتی نیز همراه است. هرزنامه‌های بی‌شماری را که روزانه برای کاربران فرستاده می‌شود، می‌توان به یکی از این مشکلات بهشمار آورد. هرزنامه‌ها پیام‌هایی با اهداف اقتصادی با مخرب هستند، که بدون رضایت گیرندگان پیام ارسال می‌شوند. هرزنامه‌ها علاوه‌بر تأثیرات منفی که برای کاربران دارند، تهدیدی امنیتی بهشمار می‌آیند که در کارایی سیستم‌های شبکه اخلاق ایجاد می‌کنند. از آسیب‌های هرزنامه می‌توان به فیشینگ اطلاعات، کاهش پهنه‌ی باند و انتشار بدافزارها اشاره کرد؛ به علاوه افزایش

۱- مقدمه

امروزه، رایانامه به یکی از رایج‌ترین ابزارهای ارتباطی در زندگی روزمره پسر تبدیل شده است. این ارتباط می‌تواند یک مکالمه ساده دوستانه و یا یک موضوع مهم تجاری باشد. رایانامه روشی سریع و ارزان قیمت برای برقراری این ارتباط است و ارسال رایانامه به عنوان روشی امن، سریع و کم‌هزینه برای ارسال پیام مورد توجه کاربران اینترنت قرار گرفته است که مزایایی همچون قابلیت اطمینان بالا، هزینه انتقال بهنسیه کم، تحويل سریع و قابلیت خودکاربودن را در بر دارد که موجب استفاده روزافزون از رایانامه شده است. متأسفانه همین عمومیت و سادگی استفاده از رایانامه باعث شده تا مورد سوءاستفاده هرزنامه‌نویسان و کلاهبرداران اینترنتی قرار بگیرد.

^۱ Accuracy ^۲ Precision ^۳ Recall

اطلاعات از رایانمۀ استفاده می‌کند. با توجه به پژوهش‌ها، بیشتر ترافیک و پهنای باند، مربوط به رایانمۀ های هرزنامه است. در این راه، مقدار بزرگی از پهنای باند به هدر می‌رود و در سامانه ارسال رایانمۀ سریز رخ می‌دهد. درواقع، فیلتر هرزنامه‌ها، یک برنامه رایانه‌ای است که توانایی دسته‌بندی رایانمۀ ها و تشخیص هرزنامه‌ها را دارد. بیشتر این فیلترها، از ترکیب چندین روش مانند کلاه سیاه و سفید، استفاده از واژه‌های کلیدی فیلتر بر پایه قانون و غیره برای شناسایی دقیق‌تر هرزنامه‌ها بهره می‌برند. این روش‌ها می‌توانند به‌تمهایی فیلتری مؤثر باشند، ولی در کاربردهای تجاری از ترکیب این روش‌ها استفاده می‌شود. برخی از این روش‌ها، به صورت دستی توسط کاربر تعیین می‌شوند. از این نوع فیلترها، می‌توان به فیلترهای استفاده شده در رایانمۀ یاهو اشاره کرد. این روش‌ها یک ابراد عمده دارند و آن ثابت‌بودن قوانین مورد استفاده در این فیلترها است، که باید تحت نظر کاربر تعیین شوند. مشکل دیگر این روش‌ها این است که هرزنامه‌نویسان با ترفندهای مختلفی می‌توانند این فیلترها را فریب دهند. هرزنامه‌ها به طور معمول پیام‌های تبلیغاتی هستند و ویژگی‌های مشابهی دارند؛ برای مثال آنها که محصولی را تبلیغ می‌کنند از قیمت آن حرف می‌زنند و یا می‌گویند که فرستن چقدر استثنایی است. حتی رنگارنگ بودن بخش‌های نوشته می‌تواند نشان از بی‌ارزش بودن آن باشد [۶ و ۷].

تاکنون فیلترهای ضدهرزنامه مختلفی عرضه شده‌اند که بیشتر آنها براساس تطبیق قوانین ثابت عمل می‌کنند [۸]. قوانین این سامانه‌ها به صورت دستی توسط کاربر تعیین می‌شوند و شامل ویژگی‌ها و مشخصات ثابت رایانمۀ نامعتبر یا هرزنامه هستند و عمل حذف هرزنامه براساس آنها انجام می‌شود. برای مثال فیلترهای استفاده شده در یاهو که کاربر نشانی‌هایی را به عنوان فرستنده هرزنامه معرفی کرده و سرویس‌دهنده رایانمۀ براساس نشانی‌های معرفی شده، رایانمۀ و یا هرزنامه دریافتی از آنها را جداگانه ذخیره یا حذف می‌کند [۹]. روش‌های دیگر استفاده از فهرست‌های سیاه یا سفید، و استفاده از واژه‌های کلیدی، فیلترهای قانون، پایه و غیره هستند. اغلب این روش‌ها، با استفاده از ویژگی‌های موجود در سرپیام رایانمۀ هرزنامه‌ها را شناسایی و فیلتر می‌کنند. این روش‌ها بدون نیاز به نرم‌افزار فیلترکردن، توسط خدمات دهنده‌گان رایانمۀ استفاده می‌شوند. اشکال چنین روش‌هایی این است که قوانین آنها ثابت است و باید توسط کاربر تعیین شوند و فرستنده هرزنامه نیز می‌توانند با ترفندهای مختلفی این فیلترها را فریب داده و هرزنامه‌های خویش را از آنها عبور دهند [۱۰]. برخی از پژوهش‌گران برای

محاسبات و هزینه به روزرسانی شبکه‌ها و تأثیر منفی بر سرویس‌های ارائه شده برای رایانمۀ های درست را نیز می‌توان شمرد. رایانمۀ های هرزنامه به پیام‌های ناخواسته که با استفاده از رایانمۀ پیام‌رسان‌ها، بلاگ‌ها، گروه‌های خبری، شبکه‌های اجتماعی، جستجو در وب، تلفن همراه و غیره به منظور تبلیغات، فیشنینگ ارسال و یا دریافت الکترونیکی، انتشار بدافزار و غیره است [۱۱]. هدف و انگیزه هرزنامه این است که اطلاعاتی را به گیرنده منتقل کند که این اطلاعات (مانند تبلیغات برای یک محصول) که با احتمال زیاد بی‌ارزش، غیرقانونی، طعمه برای یک طرح، تقلب، ترویج یک آرمان و یا تخریب رایانه‌ای که اغلب برای نفوذ و یا ربودن رایانه دریافت کننده، است. تشخیص هرزنامه یک فعالیت وقت‌گیر است. روش‌های متعددی به تازگی بررسی شده اما هرزنامه‌نویسان‌ها همیشه سعی می‌کنند، راههایی برای فرار از فیلترهای موجود بیابند؛ بنابراین فیلترهای جدید برای تشخیص هرزنامه‌ها نیاز به توسعه دارند. هرزنامه در قالب متن و تصاویر بوده که می‌تواند به سیستم آسیب برساند. برای پخش حقایق ناخواسته فرستنده‌گان هرزنامه تا حد زیادی از رایانمۀ سوء استفاده می‌کنند؛ بنابراین، هرزنامه را می‌توان به عنوان یکی از مشکلات موجود در محدوده زمانی معین که یک کاربر اینترنت با آن مواجه است، نامید. از آسیب‌های هرزنامه می‌توان تهدید، امنیتی برای کاربران، اخلال در کارایی سیستم‌های شبکه، فیشنینگ اطلاعات، کاهش پهنای باند، انتشار بدافزارها، اتلاف فضای ذخیره‌سازی در خدمات دهنده‌گان رایانمۀ، اتلاف منابع پردازش و اتلاف وقت کاربران را نام برد [۲].

خدمات دهنده‌گان بزرگ به طور عمومی از ترکیب چند راه برای تشخیص هرزنامه استفاده می‌کنند و هم هرزنامه‌ها و هم مقابله‌کننده‌گان با هرزنامه در تلاش برای ایجاد راههای جدید برای غلبه بر روش‌های توسعه یافته توسط یکدیگر هستند. در [۱۳ و ۱۴]، دو روش کلی رویکرد مهندسی دانش و رویکرد یادگیری ماشین برای تشخیص هرزنامه استفاده شده است. در روش مهندسی دانش، استفاده از اطلاعات شبکه‌ها و روش‌های نشان‌دهی پروتکل اینترنت برای تعیین اینکه آیا یک پیام هرزنامه و یا غیره هرزنامه است، استفاده می‌شود و فیلتر براساس مبداء هم نامیده می‌شود. در رویکرد یادگیری ماشین، مجموعه‌ای از قوانین به منظور تعیین دسته‌بندی رایانمۀ به عنوان هرزنامه یا غیره هرزنامه، مورد استفاده قرار می‌گیرد.

شبکه جهانی اینترنت، زمان و فاصله ارتباطات را به وسیله رایانمۀ کاهش داده است [۱۵]. بهمین دلیل، کاربران اغلب برای برقراری ارتباط با دیگران و ارسال یا دریافت

کلارک و همکاران از مجموعه واژگان و شبکه عصبی مصنوعی برای سامانه خودکار فیلترینگ رایانامه استفاده کردند که از یک شبکه عصبی مبتنی بر فیلترینگ برای ارسال رایانامه به پوشیده‌ها و ضدهرزنامه استفاده شده است. آزمایش‌ها نشان می‌دهد که این روش دقیق‌تر از چند روش دیگر است. آنها همچنین تأثیر انتخاب ویژگی‌های مختلف و همچنین قابل حمل بودن فیلتر ضدهرزنامه را در میان کاربران مختلف بررسی کردند و نشان دادند که شبکه عصبی مصنوعی می‌تواند برای فیلترکردن ارسال رایانامه به صندوق‌های پستی و رایانامه هرزنامه با موقعيت مورد استفاده قرار گیرد [۱۳].

بووژنگ، الگوریتم‌های یادگیری ماشین به نام ماشین بردار پشتیبان، شبکه‌های عصبی مصنوعی و شبکه‌های بیزین‌ساده را برای تشخیص رایانامه از هرزنامه باهم مقایسه کردند. آزمایش‌ها برروی اندازه‌های مختلف مجموعه‌داده و ویژگی‌های انتخاب‌شده مختلف، انجام شده است. نتایج نشان می‌دهد که دسته‌بندی شبکه‌های عصبی مصنوعی به تنها برای ابزارهای فیلتر هرزنامه مناسب است. در حالت کلی عملکرد دسته‌بندی ماشین بردار پشتیبان به‌طور واضح بر دسته‌بندی بیزین‌ساده برتری دارد [۱۴].

وانگ و همکاران، برای استخراج ویژگی‌ها از چهار ترکیب متفاوت مثل همه متن رایانامه، سرآیند رایانامه، موضوع رایانامه و بدنه آن استفاده کردند. بعد از استخراج ویژگی‌ها، سامانه را با سه الگوریتم یادگیری ماشین بردار پشتیبان و الگوریتم K نزدیک‌ترین همسایه و شبکه‌های بیزین‌ساده آموزش داده و سپس نتایج به‌دست آمده از الگوریتم‌ها را مقایسه کردند و در پایان از معیار TF-IDF و ماشین بردار پشتیبان برای استخراج ویژگی‌ها از رایانامه و دسته‌بندی هرزنامه‌ها استفاده کردند [۱۵].

سوجاتا و همکاران، ایجاد یک مدل جدید با استفاده از دسته‌بند شبکه عصبی مصنوعی روی بخشی از پیام‌های رایانامه از چندین کاربر را پیشنهاد دادند. در این مدل نیز، از انتخاب ویژگی استفاده شده است. مجموعه ویژگی‌هایی مثل ویژگی‌های توصیفی از واژگان و پیام‌هایی شبیه به آنچه که یک خواننده رایانامه برای شناسایی هرزنامه استفاده می‌کند، استفاده شده است. مدل دسته‌بندی را برای دسته‌بندی رایانامه‌ها به دو دسته هرزنامه و غیره‌رزنامه پیشنهاد و از یک سامانه برای دسته‌بندی داده‌های بدون ساختار در فرمت مناسب استفاده می‌کنند. از الگوریتم ماشین بردار پشتیبان برای دسته‌بندی رایانامه‌ها به هرزنامه و غیره‌رزنامه استفاده

فیلترکردن هرزنامه‌ها به صورت کارا، تنها از یک روش استفاده نمی‌کنند، بلکه تعدادی از روش‌ها را با هم ترکیب می‌کنند تا به دقت بیشتری دست پیدا کنند. انگیزه نگارندگان این مقاله، ارائه روشی کارآمد برای تفکیک هرزنامه‌ها از رایانامه‌های معتبر است. مدل پیشنهادی بر پایه ایده بهبود الگوریتم K نزدیک‌ترین همسایه با الگوریتم بهینه‌سازی اجتماع ذرات است که قادر به تولید نتیجه بهتر نسبت به عملکرد هر یک از روش‌ها به صورت انفرادی است. عملکرد این سامانه روی مجموعه‌داده استاندارد Spambase اندازه‌گیری و گزارش شده است.

ما ساختار این مقاله را بدین شرح ساماندهی کردیم: در بخش (۲) مختصراً به کارهای قبلی اشاره شده است و در بخش (۳) روش پیشنهادی و در بخش (۴) ارزیابی روش پیشنهادی و درنهایت در بخش (۵) به نتیجه‌گیری و ارایه کارهای آینده خواهیم پرداخت.

۲- کارهای قبلی

در سال‌های اخیر، روش‌های زیادی مانند الگوریتم‌های مبتنی بر جمعیت برای انتخاب ویژگی‌های مهم و حذف ویژگی‌های زائد استفاده شده و همچنین الگوریتم‌های دسته‌بندی متفاوتی برای دسته‌بندی کردن رایانامه‌های هرزنامه و غیره‌رزنامه استفاده شده است. دروکر و همکاران، الگوریتم ماشین بردار پشتیبان را برای دسته‌بندی رایانامه‌های هرزنامه پیشنهاد داده‌اند و این الگوریتم را با سه الگوریتم Rocchio و درخت تصمیم‌گیری مقایسه کرده‌اند، این چهار الگوریتم با مجموعه‌داده‌های مختلف (شامل هزار ویژگی و هفت‌هزار بعد) آزمایش شدند [۱۱].

سمیعی یگانه و همکاران، برخی از روش‌های یادگیری ماشینی مثل بیزین‌ساده، شبکه‌های عصبی مصنوعی، روش‌های دسته‌بندی سامانه اینمنی مصنوعی و منطق فازی را مورد بررسی قرار دادند. منطق فازی محتوای پیام را برای پیش‌بینی دسته آن، بر یک مجموعه واژگان کلیدی ثابت از پیش‌تعیین شده تکیه نمی‌کند؛ بنابراین، می‌تواند با روش‌های هرزنامه سازگار باشد و دانش خود را به صورت پویا ایجاد کند. براین اساس، ایشان تلاش کرده‌اند تا عملکرد مدل پیشنهادی را بهبود بخشدند و با شناسایی ویژگی رایانامه‌های هرزنامه در حساب رایانامه و حذف توسط خود فرسته عملکرد بهتری را در تشخیص رایانامه‌ها از هرزنامه‌ها ایفا کنند [۱۲].

یک گراف فقط با قوس (O²ⁿ) هدایت می‌شوند. الگوریتم شامل عملکرد دسته‌بندی و اندازه مجموعه‌ای از ویژگی‌ها به صورت فراکتشافی انجام می‌دهد و مجموعه‌ای از ویژگی‌ها را با اندازه کوچک و دقت دسته‌بندی بالا را انتخاب می‌کند. ایشان آزمایش خود را روی دو پایگاه داده تصویری بزرگ و پانزده مجموعه‌داده غیرتصویری انجام دادند و نشان دادند الگوریتم پیشنهادی دارای سرعت پردازش بالا، دقت دسته‌بندی بالا و با ابعاد ویژگی کوچک نسبت به دیگر روش‌ها است. نتایج در مجموعه‌داده‌های تصویری و غیر تصویری نشان می‌دهند که الگوریتم پیشنهادی دارای عملکرد برتر در مقایسه با دیگر الگوریتم‌های موجود است [۱۹]؛ همچنین، کارتیکا و همکاران، با استفاده از الگوریتم کلونی مورچه انتخاب ویژگی و بوسیله دسته‌بند بیزین‌ساده دسته‌بندی را انجام داده‌اند و نتایج خود را با الگوریتم ژنتیک مقایسه کردند. الگوریتم کلونی مورچه نتایج بهتر را با دقت ۸۴٪ در مقایسه با دسته‌بند بیزین ساده و الگوریتم ژنتیک نشان دادند که دقت آن ۷۷٪ بوده است [۲۰].

رضوی و میرزاپور، بر روی هرزنامه‌ها در رایانمه، پژوهشی را انجام داده‌اند و نتایج آن‌ها را به‌وسیله یک روش انتخاب ویژگی مبتنی بر الگوریتم ژنتیک روی چند الگوریتم یادگیری ماشینی ارزیابی کردند. شبکه‌های بیزین‌ساده و دسته‌بندی K نزدیک‌ترین همسایه برای دسته‌بندی استفاده شده است. از الگوریتم ژنتیک برای پیداکردن یک زیرمجموعه بهینه از ویژگی‌ها استفاده شده که ویژگی‌های انتخاب شده برای دسته‌بندی مجموعه‌داده استفاده شده است. نتایج به دست آمده از روش پیشنهادی با شرایطی که هیچ ویژگی انتخاب نشده مقایسه شده و نتایج نشان داده است که روش پیشنهادی دقت قابل مقایسه نسبت به روش‌های دیگر را دارد. به علاوه، دسته‌بند شبکه بیزین‌ساده نتایج بهتری در مقایسه با دسته‌بندی K نزدیک‌ترین همسایه را دارد. بنابراین روش پیشنهادی برای انتخاب ویژگی توائیته است میزان دقت را افزایش دهد [۱].

شارما و همکاران، پژوهش‌های خود را روی الگوریتم‌های زبان ماشین انجام داده و با استفاده از مجموعه‌داده Spambase آزمایش کردند. آنها ۲۴ الگوریتم را بر روی داده‌های هرزنامه با استفاده از ۵۵ ویژگی هرزنامه مقایسه کردند و بر روی دقت و عملکرد الگوریتم برای دسته‌بندی رایانمه‌های هرزنامه و غیره رزنامه تمرکز کردند. نتایج تجربی بر روی داده‌های Spambase نشان می‌دهد که عملکرد روش پیشنهادی بهتر است و دقیقی برابر با ۹۴/۲۸ را به دست آوردند که از تمام روش‌های دیگر بالاتر است [۲۲].

می‌شود. سامانه پیاده‌سازی شده با استفاده از الگوریتم ماشین بردار پشتیبان توانسته است با موفقیت، رایانمه‌های ورودی را در زمان واقعی به دو دسته هرزنامه و غیره رزنامه دسته‌بندی کند. عملکرد و دقت سامانه پیاده‌سازی شده با استفاده از ماتریس درهم‌یختگی بررسی شده و نتیجه آن مطلوب نشان داده شده است [۱۶].

پلاکتیک و همکاران، به‌منظور دسته‌بندی هرزنامه‌ها، مجموعه‌داده را به شش زیرمجموعه تقسیم کردند و سپس هر کدام از زیرمجموعه را به دو بخش آموزش و آزمایش تقسیم کرده و از الگوریتم K نزدیک‌ترین همسایه برای آموزش مدل استفاده کرده‌اند که یک فیلتر برش سریع براساس آمار اصلاح شده kolmogrov-smirnov پیشنهاد شده است. فیلتر مبتنی بر همبستگی پیشنهاد شده در مقایسه با سایر فیلترها (فیلتر ویژگی سریع مبتنی بر همبستگی، فیلتر مبتنی بر همبستگی و رتبه‌بندی ساده مبتنی بر پوشش^۱) برای حذف افزونگی مناسب‌تر تشخیص داده شده است. نتایج به دست آمده در مقایسه با فیلتر مبتنی بر همبستگی به‌طور قابل توجهی متفاوت نیست، اما بسیاری موارد بهتر از نتایج انتخاب ویژگی سریع مبتنی بر همبستگی است. الگوریتم پیشنهاد داده شده به طور قابل توجهی می‌تواند ابعاد مسئله براساس انتخاب ویژگی را برای حل مسایل با ابعاد بزرگ کاهش دهد [۱۷]. آزاد و همکاران، از روش‌های یادگیری ماشین (دسته‌بندی بیزین‌ساده، الگوریتم K نزدیک‌ترین همسایه، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان، سامانه ایمنی مصنوعی و مجموعه‌های راف^۲) برای دسته‌بندی رایانمه‌های هرزنامه استفاده کردند و برخی از محبوب‌ترین روش‌های یادگیری ماشین و کاربرد آنها را در رابطه با دسته‌بندی رایانمه هرزنامه مورد بررسی قرار دادند و کارایی آنها با مجموعه‌داده SpamAssassin ارزیابی شده است. درصد معیار بازخوانی هرزنامه در شش روش، ارزش کمتری در نسبت به دقت و صحت دارد؛ درحالی که در شرایط دقت ما می‌توان دریافت که روش‌های بیزین‌ساده و مجموعه‌های راف دارای عملکرد بسیار رضایت‌بخش در میان روش‌های دیگر است [۱۸].

چن و همکاران، از الگوریتم کلونی مورچه برای بهینه‌سازی استفاده کردند. روش‌های انتخاب ویژگی‌ها نیاز به یک گراف کامل با لبه‌های O²ⁿ^۲ دارند با این حال، اینها یک الگوریتم جدید ارائه دادند که در آن مورچه‌های مصنوعی به

¹ Wrapper
² Rough

مجموعه‌داده Spamassasian برای تأیید عملکرد سامانه پیشنهادی استفاده شده است. مجموعه راف، پیشرفته قابل توجهی در شبکه عصبی مصنوعی، مجموعه راف و الگوریتم ماشین بردار پشتیبان از لحاظ دقیق دقت دسته‌بندی نشان داده است. با مقایسه عملکرد الگوریتم ژنتیک و بدون استفاده از الگوریتم ژنتیک برای انتخاب ویژگی، مشاهده شده است که روش ترکیبی رویکرد مجموعه راف برای فیلتر کردن هوشمند رایانامه می‌تواند به نتایج دقیق مورد نیاز دست یابد. الگوریتم ژنتیک برای بهینه‌سازی انتخاب زیرمجموعه ویژگی و پارامترهای دسته‌بندی برای دسته‌بندی قانون مجموعه اعمال شده است که ویژگی‌های اضافی و بی‌معنی را حذف می‌کند و در نتیجه ابعاد بردار ویژگی را کاهش می‌دهد و به مجموعه راف برای انتخاب بهینه کمک می‌کند [۲۵].

مارسوتا و همکاران نشان دادند که دسته‌بندی رایانامه یا دسته‌بندی بیزین‌ساده بدون نیاز به جمع‌آوری مجدد می‌تواند برای پردازش لایه ۳ سازگار مفید واقع شود [۲۶]. و همچنین، بیاتی و همکاران برای شناسایی هرزنامه‌ها، یک روش تشخیص هرزنامه با استفاده از دسته‌بندی بیزین‌ساده را پیشنهاد دادند، که این دسته‌بندی پیام‌های رایانامه را براساس محتوای پیام‌ها به عنوان هرزنامه یا رایانامه شناسایی می‌کند. هر رایانامه به عنوان یک مجموعه‌ای از واژگان (ویژگی‌ها) آن نشان داده شده است. مجموعه‌داده CSDMC2010 برای آموزش و آزمون دسته‌بندی بیزین‌ساده مورد استفاده قرار گرفته است [۲۷].

تانگ و همکاران سامانه‌ای را پیشنهاد کرد که از دسته‌بندی ماشین بردار پشتیبان برای مقاصد دسته‌بندی استفاده کردد؛ از جمله این سامانه، استخراج داده‌های رفتار فرستنده رایانامه براساس توزیع ارسال سراسری است، با تجزیه و تحلیل هر نشان IP. ارسال رایانامه را ارزش گذاری کردد. آنها دو دسته‌بندی بهنام الگوریتم ماشین بردار پشتیبان و الگوریتم جنگل تصادفی برای شناسایی سورهای رایانامه مخرب را معرفی کرده‌اند. ویژگی‌های رفتار ارسالی جهانی استخراج شده از سوابق پرس‌وحجی ساده به دست آمده است. این ویژگی‌ها با روش‌های دسته‌بندی پیشرفته همراه است که برای تشخیص هرزنامه قابل اعتماد هستند. هر دو الگوریتم ماشین بردار پشتیبان و الگوریتم جنگل تصادفی متأثر برای ایجاد دسته‌بندی دقیق هستند. بین آنها، جنگل تصادفی به‌هراتقاب دقیق‌تر، اما از لحاظ زمان و فضای هزینه‌بر است. اگر پارامترهای مدل‌سازی درست باشند، دو ماشین بردار پشتیبان، دقت مشابهی را در یک روش بسیار کارآمدتر تولید می‌کند. نتایج تجربی نشان می‌دهد که دو ماشین بردار

ادریس و همکاران، یک مدل بهبودیافته جدید را که ترکیبی از الگوریتم انتخاب منفی با الگوریتم بهینه‌سازی از دحام ذرات است، پیشنهاد داده‌اند. ویژگی منحصر به‌فرد این مدل این است که الگوریتم بهینه‌سازی از دحام ذرات در مرحله تصادفی، الگوریتم انتخاب منفی تصادفی اجرا می‌کند. مدل پیشنهادی به عنوان یک جایگزین بهتر برای مدل الگوریتم Spambase برای انتخاب منفی عمل می‌کند. مجموعه‌داده Spambase بررسی عملکرد الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی از دحام ذرات در برابر بهبود مدل ترکیبی الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی از دحام ذرات مورد استفاده قرار گرفته است. آزمایش‌ها نشان داده است که مدل بهبودیافته در عملکرد و دقت قادر به شناسایی هرزنامه‌های رایانامه بهتر از مدل الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی از دحام ذرات است. دقت الگوریتم انتخاب منفی ۶۸/۸۶ درصد و دقت روش ترکیبی برابر با ۹۱/۲۲ درصد است. بهطور کلی، نتایج تجربی برتری مدل بهبودیافته الگوریتم انتخاب منفی الگوریتم بهینه‌سازی از دحام ذرات بر مدل ترکیبی الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی از دحام ذرات را نشان داده است [۲۳].

ادریس و همکاران، مدل ترکیبی جدیدی را که ترکیبی از الگوریتم انتخاب منفی و الگوریتم تکامل تفاضلی است، پیشنهاد کرده‌اند. نکته منحصر به‌فرد این مدل این است که الگوریتم تکامل دیفرانسیل در مرحله تولید تصادفی الگوریتم انتخاب منفی اجرا شده است؛ همچنین فاصله آشکارساز تولیدشده بیشینه شد و همپوشانی از آشکارسازها نیز به کمینه رسیده است. فاکتور خروجی محلی به عنوان تابع تناسب برای به بیشینه‌رساندن فاصله شناسه‌های هرزنامه تولیدشده از فضای غیرهزنامه اجرا شده و همچنین مشکل آشکارسازهای همپوشانی با محاسبه کمینه و بیشینه فاصله دو آشکارساز همپوشانی در هرزنامه حل شده است. مدل ترکیبی پیشنهادشده، جایگزینی بهتر به جای مدل الگوریتم انتخاب منفی است. برای ارزیابی این روش‌ها از مجموعه‌داده هرزنامه استفاده شده است. نتایج نشان داده است که مدل ترکیبی پیشنهادی عملکرد و دقت بیشتری نسبت به مدل الگوریتم انتخاب منفی دارد. و همچنین، دقت تشخیص الگوریتم ترکیبی برابر با ۸۳/۰۶ درصد و دقت الگوریتم انتخاب منفی استاندارد برابر با ۶۸/۸۶ درصد است [۲۴].

صوفی و همکاران، یک روش ترکیبی الگوریتم ژنتیک با مجموعه راف برای بهینه‌سازی پارامترهای دسته‌بندی، دقت پیش‌بینی و زمان محاسبه مجموعه راف طراحی کردند.

روشی که تنها از دسته‌بند ماشین بردار پشتیبان برای دسته‌بندی استفاده کرده‌اند، عملکرد بهتری داشته است [۳۲]. فوچا و آود، الگوریتم بهینه‌سازی ازدحام ذرات را برای فیلترینگ رایانمehا های هرزنامه با الگوریتم شبکه‌های عصبی تابع پایه رادیال (که با اختصار RBFNN نامیده می‌شود) ترکیب کردند. الگوریتم بهینه‌سازی ازدحام ذرات برای بهینه‌سازی الگوریتم RBFNN از چند جهت مانند معماری شبکه، الگوریتم یادگیری و اتصالات شبکه استفاده شده است. الگوریتم بهینه‌سازی ازدحام ذرات برای پیدا کردن مراکز بهینه نرون‌های مخفی در RBFNN استفاده شده و همچنین الگوریتم K نزدیک‌ترین همسایه برای بهینه‌سازی عرض RBFNN و روش SVD برای بهینه‌سازی وزن RBFNN استفاده شده است. نتایج حاصل از آزمایش نشان می‌دهد که این روش یک روش مؤثر در دسته‌بندی می‌باشد [۳۳].

در پژوهش [۳۴]، چهارده روش دسته‌بندی مختلف استفاده شده را برای تشخیص رایانمehا های هرزنامه بررسی کردند و نتایج به دست آمده نشان می‌دهد که میزان صحت الگوریتم دسته‌بندی جنگل تصادفی برابر با $\frac{94}{3}$ درصد از بقیه دسته‌بندها عملکرد بهتری را نشان داده است.

۳- روش پیشنهادی

۳-۱- انتخاب ویژگی

پیدا کردن زیرمجموعه بهینه از یک مجموعه بزرگ ویژگی، مسئله‌ای است که اغلب با آن مواجه هستیم. از آنجایی که بالابودن تعداد ویژگی‌ها، بار محاسباتی سامانه را بالا می‌برد و دقت سامانه را کاهش می‌دهد؛ لذا نیاز به ارائه سامانه‌ای با کمترین میزان ویژگی و کارایی قابل قبول ضروری است. مسئله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مسئله در بسیاری از کاربردها (مانند دسته‌بندی) اهمیت بهسزایی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آنها یا بدون استفاده هستند و یا این که بار اطلاعاتی چندانی ندارند [۳۵]. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند، ولی بار محاسباتی را برای کاربرد موردنظر بالا می‌برد و علاوه بر این باعث می‌شود که اطلاعات غیر مفید زیادی را به همراه داده‌های مفید ذخیره کنیم. هدف انتخاب ویژگی این است که یک زیرمجموعه از ویژگی‌ها برای افزایش دقت پیش‌گویی انتخاب شوند [۳۶، ۳۷]. به عبارت دیگر، کاهش اندازه ساختار بدون کاهش قابل ملاحظه در دقت پیش‌گویی دسته‌بندی کننده‌ای که با

پشتیبان دسته‌بندی مؤثر، دقیق و بسیار سریع تر از دسته‌بندی تصادفی جنگل‌ها انجام می‌دهد [۲۸]. یانگ و همکاران نخستین روش اولویت‌بندی رایانمeh را توسعه دادند که به طور خاص بر تجزیه و تحلیل شبکه‌های اجتماعی شخصی برای جذب گروه‌های کاربر و به دست آوردن ویژگی‌های غنی که نقش اجتماعی را از نظر کاربر خاص نشان می‌دهند، تمرکز می‌کنند و همچنین آنها یک چارچوب دسته‌بندی نظارت شده برای مدل‌سازی اولویت‌های شخصی بر روی پیام‌های رایانمeh و برای پیش‌بینی سطوح اهمیت برای پیام‌های جدید ایجاد کردند [۲۹].

گوزلا و همکاران، یک مدل الهام‌گرفته از سامانه اینمی به نام سامانه اینمی مصنوعی ذاتی و سازگار (IA-AIS) را پیشنهاد [۳۰] و برای مشکل شناسایی پیام‌های هرزنامه ناخواسته اعمال کردند. در مقایسه با نسخه بهینه‌شده از بیزین‌ساده نرخ دسته‌بندی بسیار دقیق را به دست آورده‌اند و می‌توان نتیجه گرفت که IA-AIS دارای توانایی بیشتری برای شناسایی پیام‌های هرزنامه است.

فریس و همکاران، روش ترکیبی برمبنای الگوریتم بهینه‌سازی ازدحام ذرات و دسته‌بند جنگل تصادفی برای فیلترینگ هرزنامه ارائه دادند. در این روش، در مرحله نخست انتخاب ویژگی و در مرحله دوم دسته‌بندی شده است. در مرحله نخست الگوریتم بهینه‌سازی ازدحام ذرات بهترین ویژگی‌ها را از مجموعه ویژگی‌ها انتخاب کرده و ابعاد داده آموزش را کاهش داده است. در مرحله دوم دسته‌بند جنگل تصادفی با استفاده از ویژگی‌های انتخاب شده در مرحله نخست داده آزمون را دسته‌بندی کرده و در نهایت نتیجه با دیگر الگوریتم‌های یادگیری ماشین مقایسه شده است. نتایج ارزیابی نشان می‌دهد که این روش ترکیبی در مقایسه با دیگر روش‌ها عملکرد بهتری داشته است [۳۱].

رانه در پژوهشی، روش ترکیبی الگوریتم فازی C-Mean و دسته‌بند ماشین بردار پشتیبان را برای مسئله رایانمeh های هرزنامه ارائه داد. در این روش ترکیبی، الگوریتم FCM در ایجاد خوش‌هایی از داده رایانمeh برای آموزش دسته‌بند ماشین بردار پشتیبان به کار می‌رود در نهایت عمل دسته‌بندی انجام می‌شود. در این روش از مجموعه داده LingSpam برای ارزیابی روش ترکیبی استفاده شده است و درواقع مجموعه داده رایانمeh های متنی با روش ترکیبی فازی C-Means و دسته‌بند ماشین بردار پشتیبان افتتا به صورت رایانمeh های هرزنامه و غیره روزنامه دسته‌بندی شده‌اند و نتایج ارزیابی نشان داده است که این روش در مقایسه با

علمی
پژوهشی

که کاربران الگوریتم‌های بهینه‌سازی تصور می‌کنند، که این کاهش در تعداد تکرارها به معنی کاهش زمان اجرای برنامه برای رسیدن به هم‌گرایی است. در حالی که چنین تصوری به طور کامل غلط است. هرچند که افزایش تعداد ذرات، کاهش تعداد تکرارها را در پی دارد، اما افزایش در تعداد ذرات، باعث می‌شود که الگوریتم در مرحله ارزیابی ذرات زمان بیشتری را صرف کند که این افزایش در زمان ارزیابی باعث می‌شود زمان اجرای الگوریتم تا رسیدن به هم‌گرایی با وجود کاهش در تعداد تکرارها، کاهش نیاید؛ پس افزایش تعداد ذرات نمی‌تواند برای کاهش زمان اجرای الگوریتم مورد استفاده قرار گیرد. تصور غلط دیگری وجود دارد و آن این است که برای کاهش زمان اجرای الگوریتم می‌توان تعداد ذرات را کاهش داد. این تصور نیز وجود دارد و آن این است که برای کاهش زمان لازم برای ارزیابی ذرات می‌شود؛ اما برای این که الگوریتم به جواب بهینه برسد، تعداد تکرارها افزایش می‌یابد، که درنهایت باعث می‌شود، زمان اجرای برنامه برای رسیدن به پاسخ بهینه روند کاهشی نداشته باشد. همچنین باید یادآور شد که کاهش تعداد ذرات ممکن است موجب گیرافتدان در کمینه‌های محلی شود و الگوریتم از رسیدن به کمینه اصلی باز ماند. اگر ما شرط هم‌گرایی را تعداد تکرارها در نظر گرفته باشیم، هرچند با کاهش تعداد ذرات اولیه زمان اجرای الگوریتم کاهش می‌یابد، اما جواب به دست آمده، حل بهینه‌ای برای مسئله نخواهد بود؛ زیرا الگوریتم به صورت ناقص اجرا شده است.

با اختصار، تعداد جمعیت اولیه با توجه به مسئله تعیین می‌شود. در حالت کلی، تعداد ذرات اولیه مصالحه‌ای بین پارامترهای درگیر در مسئله است. به طور تجربی انتخاب جمعیت اولیه ذرات به تعداد بیست تا سی ذره انتخاب مناسب است که به طور تقریبی برای تمامی مسائل آزمون به خوبی جواب می‌دهد و می‌توان تعداد ذرات را کمی بیشتر از حد لازم نیز در نظر گرفت تا کمی حاشیه ایمنی در مواجه با کمینه‌های محلی به دست بیاید.

محله ۳، ارزیابی تابع هدف (محاسبه هزینه یا برآزندگی) ذرات

در این مرحله، باید هر یک از ذرات را که نشان‌دهنده یک حل برای مسئله مورد بررسی است، ارزیابی کنیم. بسته به مسئله مورد بررسی، روش ارزیابی متفاوت خواهد بود؛ برای مثال اگر امکان تعریف یک تابع ریاضی برای هدف وجود داشته باشد، با جایگذاری پارامترهای ورودی (که از بردار موقعیت ذره استخراج شده‌اند) در این تابع ریاضی، برآحتی مقدار هزینه

استفاده از ویژگی‌های داده شده به دست می‌آورد. یک زیرمجموعه بهینه از مجموعه ویژگی‌ها، به صورت مستقیم با انتخاب تابع ارزیابی بستگی دارد، چون که اگر تابع ارزیابی به زیرمجموعه ویژگی بهینه یک مقدار نامناسب نسبت دهد، این زیرمجموعه هیچگاه عنوان زیرمجموعه بهینه انتخاب نمی‌شود.

ما در این مقاله، با استفاده از الگوریتم از دحام ذرات، انتخاب ویژگی را انجام داده‌ایم و سپس ویژگی‌های انتخاب شده را به دسته‌بند K نزدیکترین همسایه داده تا دقیق، صحیح و حساسیت آن را محاسبه کنیم.

۳-۲- اجرای الگوریتم بهینه‌سازی از دحام ذرات

در برخی مراجع، در نحوه تفکیک مراحل اجرای الگوریتم، اختلافاتی دیده می‌شود؛ یعنی در برخی مواقع مراحل به صورت تفکیک شده‌تری ذکر می‌شوند و در برخی مواقع دو یا تعداد بیشتری از مراحل را با هم ترکیب کرده و تبدیل به یک مرحله می‌کنند؛ اما این موضوع اشکالی در برنامه‌نویسی‌های انجام‌شده ایجاد نمی‌کند؛ زیرا آنچه که مهم است، اجرای مراحل برنامه به ترتیبی که در ادامه خواهد آمد، و نحوه تفکیک این مراحل بین صورت است.

مرحله ۱، تولید تصادفی جمعیت اولیه

تولید تصادفی جمعیت اولیه به طور ساده عبارت از تعیین تصادفی محل اولیه ذرات با توزیع یکنواخت در فضای حل (فضای جستجو) است. مرحله تولید تصادفی جمعیت اولیه به طور تقریبی در تمامی الگوریتم‌های بهینه‌سازی احتمالاتی وجود دارد. اما، در این الگوریتم علاوه بر محل تصادفی اولیه ذرات، مقداری برای سرعت اولیه ذرات نیز اختصاص می‌یابد. بازه پیشنهادی اولیه برای سرعت ذرات را می‌توان از رابطه (۱) استخراج کرد.

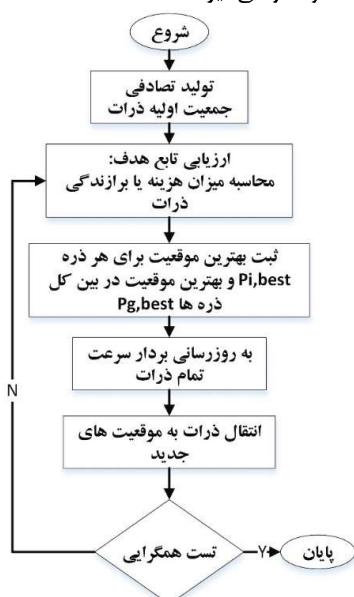
$$V_{\min} \leq V \leq V_{\max} \quad (1)$$

برای جلوگیری از افزایش بیش از حد سرعت حرکت یک ذره در حرکت از یک محل به محل دیگر (واکنشدن بردار سرعت)، تغییرات سرعت را به بازه V_{\min} تا V_{\max} محدود می‌کنیم که حد بالا و پایین سرعت با توجه به نوع مسئله تعیین می‌شود.

مرحله ۲، انتخاب تعداد ذرات اولیه افزایش تعداد ذرات اولیه برای هم‌گرایش دهنده می‌شود؛ اما گاهی مشاهده می‌شود

در مسیر بهترین مقدار ذره مورد بررسی) و پارامتر c_2 ضریب ثابت آموزش (حرکت در مسیر بهترین ذره یافتشده در بین کل جمعیت) و پارامتر $rand_{1,2}$ دو عدد تصادفی با توزیع یکنواخت در بازه صفر تا یک است. پارامتر ($t-1$) V_i بردار سرعت در تکرار ($t-1$) ام و پارامتر ($t-1$) X_i بردار موقعیت در تکرار ($t-1$) ام است.

ضرایب c_1 و w با توجه به مسئله مورد نظر به روش تجربی تعیین می‌شوند، اما به عنوان یک قانون کلی، w باید کمتر از یک باشد؛ زیرا، اگر بزرگ‌تر از یک انتخاب شود، $V(t)$ به طور دائمی افزایش می‌یابد تا جایی که واگرا شود. همچنین هرچند در تئوری، ضریب w می‌تواند منفی نیز باشد، اما در استفاده عملی از این الگوریتم نباید این ضرایب را منفی در نظر گرفت؛ زیرا منفی بودن w موجب ایجاد نوسان در $V(t)$ می‌شود. انتخاب مقدار کوچک برای این ضریب (w) نیز مشکلاتی را در پی خواهد داشت. اغلب در الگوریتم بهینه‌سازی از دحام ذرات مقدار این ضریب را مشیت و در بازه 0.7 تا 0.8 در نظر می‌گیرند. پارامترهای c_1 و c_2 نیز نباید زیاد بازگردانی شوند؛ زیرا انتخاب مقادیر بزرگ برای این دو ضریب باعث انحراف شدید ذره از مسیر خودی می‌شود. اغلب در الگوریتم از دحام ذرات مقدار این ضرایب را مشیت و در بازه 0.1 تا 0.7 در نظر می‌گیرند.



شکل ۱-۱: روند نمای الگوریتم بهینه‌سازی از دحام ذرات

موجله ۶، آزمون هم‌گرایی
آزمون هم‌گرایی در این الگوریتم، مانند سایر الگوریتم‌های بهینه‌سازی است. برای بررسی روش‌های گوناگونی وجود دارد.

این ذره محاسبه خواهد شد. توجه داشته باشید که هر ذره حاوی اطلاعات کاملی از پارامترهای ورودی مسئله است که این اطلاعات استخراج شده و در تابع هدف قرار می‌گیرد. گاهی اوقات، امکان تعریف یک تابع ریاضی برای ارزیابی ذرات وجود ندارد. این حالت زمانی پیش می‌آید که ما الگوریتم را با یک نرم‌افزار دیگر پیوند کرده باشیم و یا الگوریتم را برای داده‌های تجربی (آزمایش) استفاده کنیم. در این گونه موارد، باید اطلاعات مربوط به پارامترهای ورودی نرم‌افزار یا آزمایش را از بردار موقعیت ذرات استخراج کرده و در اختیار نرم‌افزار پیوندشده با الگوریتم جایگذاری یا در آزمایش مربوطه اعمال کرد. با اجرای نرم‌افزار و یا انجام آزمایش و مشاهده و اندازه‌گیری نتایج هزینه‌ای را که هر یک از ذرات در پی دارند، مشخص خواهد شد.

مرحله ۴، ثبت بهترین موقعیت برای هر ذره ($P_{i,best}$) و بهترین موقعیت در بین کل ذراتها ($p_{g,best}$) در این مرحله، با توجه به شماره تکرار، دو حالت قابل بررسی است:

اگر در تکرار نخست باشیم ($i=1$)، موقعیت فعلی هر ذره را به عنوان بهترین محل یافتشده برای آن ذره در نظر می‌گیریم:

$$P_{i,best} = X_i(t), i=1, 2, 3, \dots, d \quad (2)$$

$$\text{cost}(P_{i,best}) = \text{cost}(X_i(t)) \quad (3)$$

در سایر تکرارها، مقدار هزینه به دست آمده برای ذرات در مرحله ۲ را با مقدار بهترین هزینه به دست آمده برای تک تک ذرات مقایسه می‌کنیم. اگر این هزینه، کمتر از بهترین هزینه ثبت شده برای این ذره باشد، آنگاه محل و هزینه این ذره جایگزین مقدار قبلی می‌شود. در غیر این صورت تغییری در محل و هزینه ثبت شده برای این ذره ایجاد نمی‌شود؛ یعنی:

$$\begin{cases} \text{if } \text{cost}(X_i(t)) < \text{cost}(P_{i,best}) \\ \quad \text{else Not change} \\ \Rightarrow \begin{cases} \text{cost}(P_{i,best}) = \text{cost}(X_j(t)), & i \\ P_{i,best} = X_i(t), & = 1, 2, 3, \dots, d \end{cases} \end{cases} \quad (4)$$

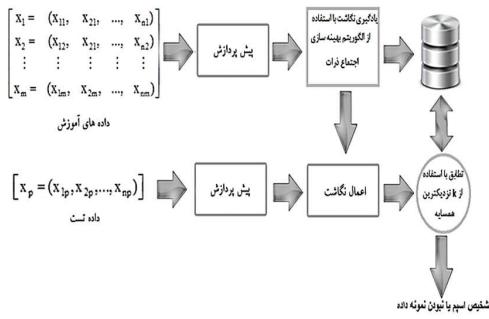
مرحله ۵، به روزرسانی بردار سرعت تمامی ذرات

$$V_i(t) = w * V_i(t-1) + c_1 * rand_1 * (P_{i,best} - X_i(t-1)) + c_2 * rand_2 * (g_{g,best} - X_i(t-1)) \quad (5)$$

پارامتر w ضریب وزنی اینرسی که نشان‌دهنده میزان تأثیر بردار سرعت تکرار قبل ($t-1$) بر روی بردار سرعت در تکرار فعلی ($t+1$) است و پارامتر c_1 ضریب ثابت آموزش (حرکت

علمی
دست‌نامه

در مدل پیشنهادی، بردارهایی به طول ۵۷ خانه ایجاد و سپس مقدار آنها با اعداد مجموعه‌داده پر می‌شوند. ذرات در فضای جستجو به دنبال بهترین و نزدیکترین ویژگی‌ها با فاصله کم هستند و اگر دو ویژگی فاصله کمتری نسبت به هم داشته باشند، انتخاب می‌شوند. هر ویژگی توسط ذرات با ویژگی همسایه مقایسه می‌شود. اگر مقدار ویژگی‌ها نزدیک به هم باشد، درصد دسته‌بندی بهمنظور دسته‌بندی درست داده‌ها افزایش می‌یابد؛ سپس از برچسب داده‌های پیش‌بینی شده و از خطای پیش‌بینی آن به عنوان معیار برآزنده‌گی یک ذره استفاده می‌کنیم؛ بنابراین ذره‌ای بهتر است که خطای کمتری داشته باشد؛ سپس با g_{best} و با گردکردن مقادیر آن نگاشت ($^{(0)}$) نهایی به دست آمده و داده‌های آموزش و آزمون را منتظر با آن نگاشت داده و الگوریتم K نزدیکترین همسایه را بر روی داده‌های نگاشتشده آموزش داده، سپس داده‌های آزمون را برای پیش‌بینی به الگوریتم K نزدیکترین همسایه داده شده است.



(شکل-۲): چارچوب روش پیشنهادی

- ۱-۳-۳- مرحله پیش‌پردازش:**
- برای دسته‌بندی رایانامه‌ها پس از فراهم کردن مجموعه‌داده، نوبت به مرحله پیش‌پردازش می‌رسد و در این مرحله اسناد متنه خام باید به فرمی که قابل استفاده در مرحله انتخاب ویژگی و الگوریتم یادگیری باشند، تبدیل شوند. در این مرحله عملیات زیر انجام می‌شود:
 - تمامی نویسه‌های موجود در متن به شکل یکسان تبدیل می‌شوند.
 - کل متن به صورت واژگان متوالی جدا از هم تقسیم می‌شوند.
 - واژگان زائد در زبان انگلیسی حذف می‌شوند.
 - ریشه واژگان انگلیسی با استفاده از الگوریتم ریشه‌یابی که برای حذف پسوندها و پیشوندهای واژگان در جهت کاهش طول واژگان و تا زمانی که به فرم ریشه‌شان تبدیل شوند استفاده می‌شوند.

برای مثال: می‌توان تعداد مشخصی تکرار را از همان ابتدا معلوم کرد و در هر مرحله بررسی کرد که آیا تعداد تکرارها به مقدار تعیین شده رسیده است؟ اگر تعداد تکرارها کوچکتر از مقدار تعیین شده اولیه باشد، آنگاه باید به مرحله ۲ بازگردید، در غیر این صورت، الگوریتم پایان می‌پذیرد. روش دیگری که اغلب در آزمون هم‌گرایی الگوریتم استفاده می‌شود، این است که اگر در چند تکرار متوالی برای مثال پاتزده یا بیست تکرار تغییری در مقدار هزینه بهترین ذره ایجاد نشود، آنگاه الگوریتم پایان می‌یابد، در غیر این صورت، باید به مرحله ۲ بازگردید. روند نمای الگوریتم ازدحام ذرات در شکل (۱) نشان داده شده است.

۳-۳- مدل پیشنهادی

در این مقاله، برای تشخیص رایانامه‌های هرزنامه از مدلی ترکیبی با بهبود الگوریتم K نزدیکترین همسایه با الگوریتم بهینه‌سازی ازدحام ذرات استفاده شده است. از الگوریتم بهینه‌سازی ازدحام ذرات برای انتخاب ویژگی و از دسته‌بندی کننده K نزدیکترین همسایه برای دسته‌بندی کردن دو نمونه رایانامه‌های هرزنامه و غیره رزنامه است. بدین صورت که ابتدا داده‌ها را به دو دسته آموزش و آزمون تقسیم کرده، که $\%80$ داده‌ها به آموزش و $\%20$ به داده‌های آزمون تقسیم داده شده است. الگوریتم بهینه‌سازی، اجتماع ذرات را روی داده‌های آموزش اجرا کرد. به این صورت که ابتدا جمعیت اولیه به صورت تصادفی انتخاب کرده و موقعیت هر ذره را به صورت اعداد اعشاری بین ($^{(0)}$) تعریف کرده و در زمان محاسبه برآزنده‌گی این اعداد گرد می‌شوند و در موقعیت یک ذره برای نگاشت داده‌ها استفاده می‌کنیم. که اگر داده‌ها به صورت ماتریس در نظر گرفته شوند، در این صورت ستون‌های متناظر با یک‌ها نگه داشته شده و متناظر با صفر حذف می‌شوند. این یعنی تعداد ستون‌های ماتریس کم شده است ولی تعداد سطرها تغییری نمی‌کند. جهت ارزیابی برآزنده‌گی موقعیت هر ذره، ابتدا فرض می‌کنیم، خود ذره بهتر است. اگر در تکرار نخست باشیم، موقعیت فعلی هر ذره را به عنوان بهترین محل یافت شده برای آن ذره در نظر می‌گیریم. در سایر تکرارها، مقدار هزینه به دست آمده برای ذرات در مرحله دو را با مقدار بهترین هزینه به دست آمده برای تک تک ذرات مقایسه شده است. اگر این هزینه کمتر از بهترین هزینه ثبت شده برای این ذره باشد، آنگاه محل و هزینه این ذره جایگزین مقدار قبلی می‌شود. در غیر این صورت، تغییری در محل و هزینه ثبت شده برای این ذره ایجاد نمی‌شود.

۴- ارزیابی و نتایج

۴-۱- مجموعه‌داده

در این مقاله، از مجموعه‌داده Spambase استفاده شده که شامل ۴۶۰۱ پیام است که تعداد ۱۸۱۳ رایانامه این مجموعه هرزنامه و تعداد ۲۷۸۸ رایانامه عادی هستند. ۵۸٪ ویژگی برای نمونه‌ها تعیین شده که ۵۷٪ ویژگی به عنوان ویژگی ورودی و ویژگی ۵۸٪ خروجی است که نشان‌دهنده هرزنامه یا غیرهرزنامه بودن رایانامه است. در صورت صرفبودن رایانامه، هرزنامه و اگر یک باشد، رایانامه مربوطه عادی است.

۴-۲- معیارهای ارزیابی

معیارهایی که به کمک آنها روش پیشنهادی را مورد ارزیابی قرار داده شده است، عبارتند از: صحت، دقت، فراخوانی و F- Measure. در ادامه، عملکرد الگوریتم دسته‌بندی را با توجه به مجموعه‌داده ورودی به تفکیک انواع دسته‌های مسئله دسته‌بندی نشان می‌دهد. TN: تعداد رایانامه‌های غیرهرزنامه که درست دسته‌بندی شده‌اند. TP: تعداد رایانامه‌های هرزنامه که درست دسته‌بندی شده‌اند. FN: تعداد رایانامه‌های هرزنامه که باشتباه به عنوان رایانامه غیرهرزنامه دسته‌بندی شده‌اند. FP: تعداد رایانامه‌های غیرهرزنامه که باشتباه به عنوان رایانامه هرزنامه دسته‌بندی شده‌اند.

اشتباه برچسب زدن هرزنامه، دارای عوارض جانی است یک برچسب هرزنامه به عنوان غیرهرزنامه یکبار از روی کاربرانی که نیاز به خواندن از آغاز تا انتها و حذف آن را دارند، برداشته است. با این حال، یک غیرهرزنامه که به عنوان هرزنامه مشخص شده‌اند، به طور معمول با دنباله‌ای از حذف به صورت خودکار دنبال و باعث می‌شود کاربر رایانامه‌های مهم را از دست بدهد و یا با انتقال خودکار به جعبه هرزنامه کاربر به احتمال زیاد آن را نادیده بگیرد، بنابراین، نیاز به استفاده از ماتریس درهم‌ریخته برای توصیف انواع مختلف خطاهای و اندازه‌گیری این که چقدر جدی است، وجود دارد. در ماتریس درهم‌ریخته هر سنتون نشان‌دهنده موارد در یک رد پیش‌بینی شده و هر سطر نشان‌دهنده موارد در یک رد واقعی. دسته مثبت به نمایندگی از هرزنامه و دسته منفی به نمایندگی از غیرهرزنامه است. در اینجا، TP نشان صحیح مشبت، FN نشان غلط مشبت، FN نشان غلط منفی و TN نشان صحیح منفی است.

(جدول-۱): ماتریس درهم‌ریختگی

هرزنامه	غیر هرزنامه	رد
TP	پیش‌بینی هرزنامه	
FN	TN	پیش‌بینی غیر هرزنامه

۳-۳-۲- یادگیری نگاشت با استفاده از الگوریتم

بهینه‌سازی از دحام ذرات

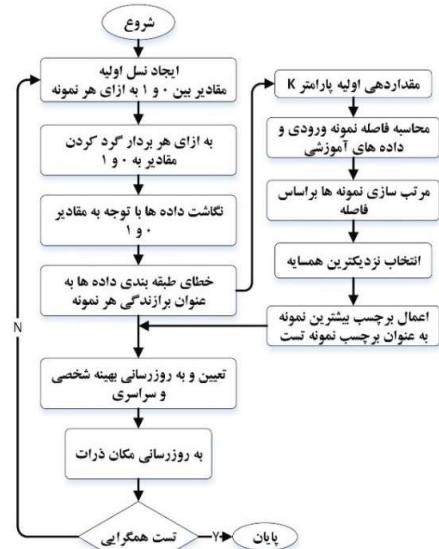
در این مرحله، با استفاده از الگوریتم بهینه‌سازی از دحام ذرات انتخاب ویژگی انجام می‌گیرد. این الگوریتم را روی داده‌های آموزش اجرا کرده تا ویژگی‌های مهم انتخاب شده و بتوانیم ویژگی‌های اضافی را حذف کرده با این کار ابعاد ویژگی‌ها کم می‌شود.

۳-۳-۳- اعمال نگاشت

بعد از انتخاب ویژگی که نگاشت نهایی به صورت (۹۰٪) به دست آمد، داده‌های آزمون و آموزش نگاشت می‌شود؛ سپس داده‌های نگاشت‌داده شده را به الگوریتم K نزدیک‌ترین همسایه داده می‌شوند.

۳-۳-۴- تشخیص هرزنامه یا نبودن رایانامه

در این مرحله، دسته‌بند با توجه به آموزشی که در قبل در مورد هرزنامه یا نبودن رایانامه توسط الگوریتم بهینه‌سازی از دحام ذرات گرفته، دسته‌بندی را انجام می‌دهد.



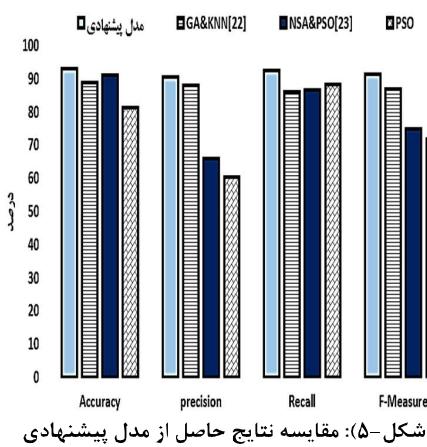
(شکل-۳): روند نمای نگاشت داده‌ها و انتخاب ویژگی‌ها با استفاده از الگوریتم بهینه‌سازی از دحام ذرات

در مدل پیشنهادی، ویژگی‌هایی انتخاب می‌شوند که بهترین برآزندگی را دارند. تابع برآزندگی برای انتخاب ویژگی‌ها طبق معادله (۶) تعریف می‌شود. در معادله (۶)، $|n|$ تعداد کل ویژگی‌ها و $|S|$ تعداد ویژگی‌های انتخاب شده است. مقدار پارامترهای δ و ρ ثابت هستند و مقدار آنها به ترتیب برابر با ۹۹ و ۱ است.

$$Fitness = \delta \cdot Accuracy + \rho \cdot \frac{|n| - |S|}{|n|} \quad (6)$$

در شکل (۴)، نمودار مقادیر معیارهای دقت، فراخوانی، F-measure و درصد صحت روش پیشنهادی ما نشان داده شده است که مقادیر آنها بدتر ترتیب از چپ به راست $91/55$ ، $92/56$ ، $90/56$ و $93/26$ درصد است که این مقادیر با تکرار بالای پنجاه به دست آمدند.

در شکل (۵)، مقادیر به دست آمده مدل پیشنهادی ما با دیگر مدل‌های مشابه نظری ترکیب الگوریتم زنگیک با دسته‌بندی K نزدیکترین همسایه [۲۱] و مدل الگوریتم انتخاب منفی با الگوریتم بهینه‌سازی ذرات و مدلی که تنها از الگوریتم از دحام ذرات استفاده کرده [۲۳] مقایسه شده است که نتایج هر گروه نمودار و همان‌طوری که مشاهده می‌کنید هر گروه معیار نشان‌دهنده یک معيار است، که نتایج هر چهار معيار یادداشده با هم مقایسه شده است که نتایج مدل پیشنهادی ما مقادیر بیشتری نسبت به مدل‌های دیگر به دست آورده است. مدلی که از الگوریتم زنگیک برای انتخاب ویژگی و بهینه‌سازی دسته‌بندی K نزدیکترین همسایه استفاده کرده، دقیق‌ترین معادل درصد و مدلی که از ترکیب الگوریتم انتخاب منفی و بهینه‌سازی ذرات برای تشخیص رایانامه‌های هرزنامه استفاده کرده، دقیق‌ترین معادل درصد و همچنین خود الگوریتم بهینه‌سازی ذرات قادر به تشخیص رایانامه‌های هرزنامه با دقیق‌ترین معادل درصد است. با این مقایسه‌ها الگوریتم پیشنهادی، ما با دقیق‌ترین نیز به این ترتیب درصد کمتری می‌دهد که دیگر مهیارها نیز به این ترتیب درصد کمتری نسبت به روش پیشنهادی ما ارائه داده‌اند.



(شکل-۵): مقایسه نتایج حاصل از مدل پیشنهادی با دیگر مدل‌ها

در جدول (۲)، مقادیر به دست آمده با تکرار بالای پنجاه به دست آمده و در هر بار تعداد ویژگی‌های انتخاب شده ثبت شده را گرد هم آورده‌ایم. برای مثال، هنگامی که تعداد

مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی، نرخ صحت است. این معیار دقت کل یک دسته‌بندی را محاسبه می‌کند. این معیار نشان دهنده این موضوع است که چند درصد از کل مجموعه‌داده‌ها به درستی دسته‌بندی شده است. معیار فراخوانی، کارایی دسته‌بندی را با توجه به تعداد رخداد دسته آن نشان می‌دهد؛ در حالی که معیار دقت به طور اساسی مبتنی بر دقیق‌ترین نتایم است و بیان گر آن است که به چه میزان می‌توانیم به خروجی دسته‌بند اعتماد کرد. معیار F-measure می‌آید، و در مواردی استفاده می‌شود که نتوان اهمیت ویژه‌ای برای هر یک از دو معیار دقت و فراخوانی نسبت به یکدیگر قائل شد.

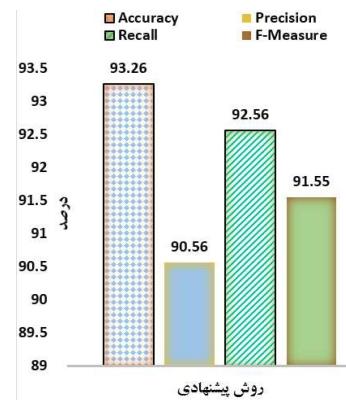
$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{F-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

بر مبنای الگوریتم k نزدیکترین همسایه، در ابتدا نمونه‌های آموزشی ارزیابی و یک مدل به منظور تشابه بین نمونه‌ها ایجاد می‌شود. در مدل پیشنهادی، مقدار k برابر با ۳ فرض شده است. نمونه‌ها بر مبنای نزدیکی فاصله و تشابه به دسته‌ها (دسته هرزنامه و غیره رزنامه) اختصاص داده می‌شوند. در ابتدا باید نمونه‌ای برای دسته‌بندی انتخاب شود که در همسایگی اش بیشترین تعداد نمونه منتبه برای آن وجود داشته باشد. درنتیجه، پس از آنکه فاصله اقلیدسی بین نقاط محاسبه شد، با مرتب‌سازی عناصر بر حسب فاصله اقلیدسی، از میان k همسایه، برچسبی که از میان k همسایه دارای بیشینه است به نمونه ناشناخته داده می‌شود.



(شکل-۶): نتایج حاصل از مدل پیشنهادی

دوفصلنامه علمی ترویجی متدی امنیت فضای تولید و تبادل اطلاعات (افتا)

(جدول-۲): مقایسه نتایج ارزیابی روش پیشنهادی با انتخاب ویژگی‌های مختلف

تعداد ویژگی‌ها	صحت	دقت	فراخوانی	F-Measure
۲۷	۸۷/۱۸	۸۱/۰۱	۸۸/۱۵۴	۸۴/۴۲
۲۸	۹۳/۲۶	۹۰/۵۶	۹۲/۵۶	۹۱/۵۵
۲۹	۸۵/۳۴	۷۷/۶۶	۸۸/۱۵	۸۲/۵۸
۳۰	۹۰/۰۱	۸۶/۷۲	۸۸/۱۵	۸۷/۱۵
۳۱	۹۰/۷۷	۸۶/۷۷	۹۰/۳۵	۸۸/۵۲
۳۲	۸۶/۲۱	۷۹/۰۶	۸۸/۴۲	۸۳/۴۸
۳۳	۹۲/۲۹	۹۰/۵۵	۸۹/۸۰	۹۰/۱۷۶
۳۴	۸۹/۲۱	۸۶/۰۱	۸۷/۴۲	۸۵/۴۸
۳۵	۹۱/۰۹	۸۷/۲۶	۹۰/۶۳	۸۸/۹۱
کل ویژگی‌ها	۸۳/۴۹	۷۸/۷۴	۷۹/۶۱	۷۹/۱۷

۲۷ عدد ویژگی انتخاب شده، میزان دقت به دست آمده ۱۸/۸۷ درصد بوده است. این در حالی است که این برنامه با تعداد تکرار بالا، آزمون شده و هر بار تعداد ویژگی‌های مختلف انتخاب شده است که مانند نتایج آن‌ها در بهترین حالت نوشته‌ایم. با انتخاب تعداد ۲۸ ویژگی نتایج مختلف بهتر شده است و دقت بهینه را به دست آورده‌ایم و اما برای مقایسه با همه ویژگی، ما همه ویژگی‌ها را به صورت دستی انتخاب کرده‌ایم، تا بتوانیم نتایج حاصل از آن را با بقیه مقایسه کنیم که در این حالت دقت به دست آمده ۸۳/۴۹ درصد است.

در جدول (۳)، مدل‌های استفاده شده برای تشخیص رایانمehrای هرزنامه با ذکر منابع آنها و مقادیر به دست آمده آنها برای معیارهای صحت، دقت، فراخوانی و F-Measure با مقادیر به دست آمده توسط مدل پیشنهادی ما با هم مقایسه شده‌اند.

(جدول-۳): مقایسه نتایج ارزیابی روش پیشنهادی با روش‌های دیگر

Ref	Models	صحت	دقت	فراخوانی	F-measure	مجموعه داده
[20]	Aco-NB	۸۴	۸۹	۷۸	۸۷	Spambasc
	GA-NB	۷۷	۸۵	۷۱	۷۵	
[21]	GA-BN	۹۱	۹۳	۸۶	۹۰	
	GA-Knn	۸۹	۸۸	۸۶	۸۷	
	BN	۸۸/۵۶	۹۳/۱۱۳	۸۸/۶	۹۰/۸	
	LB	۸۹/۷	۹۲/۸۶۲	۹۰/۴	۹۱/۶۶	
	RT	۹۱/۵۴	۹۲/۷۵۴	۹۳/۲	۹۳	
[22]	JR	۹۲/۳۲	۹۴/۴۷۶	۹۲/۹	۹۳/۷۱	Spambasc
	J48	۹۲/۳۴	۹۳/۹۰۲	۹۳/۵	۹۳/۷	
	MP	۹۲/۲۸	۹۴/۳۲۲	۹۴/۵	۹۴/۴۵	
	KS	۹۳/۵۶	۹۵/۵۸۸	۹۳/۹	۹۴/۷۳	
	RF	۹۳/۸۹	۹۵/۸۷۵	۹۴/۱	۹۵	
	RC	۹۴/۲۸	۹۶/۱۲۶	۹۴/۵	۹۵/۱۲۲	
	PSO	۸۱/۳۲	۶۰/۴۸	۸۸/۴۴	۷۱/۸۴	
[23]	NSA	۶۸/۸۶	۲۲/۲۴	۹۴/۵۳	۳۶/۰۱	Spamassasian
	NSA-PSO	۹۱/۲۲	۹۹/۶۵	۸۶/۷۲	۷۴/۹۵	
[24]	NSA	۶۸/۸۶	۲۲/۲۴	۹۴/۵۳	۳۶/۰۱	
	NSA-DE	۸۰/۶۶	۵۶/۶۲	۹۰/۸۶	۶۹/۷۶	
[25]	NB	۹۹/۴۶	۹۹/۶۶	۹۸/۴۶	-	Spamassasian
	KNN	۹۶/۲۰	۸۷/۰۰	۹۷/۱۴	-	
	NN	۹۶/۸۳	۹۶/۰۲	۹۶/۹۲	-	
	SVM	۹۶/۹۰	۹۳/۱۲	۹۵/۰۰	-	
	ATIS	۹۶/۲۳	۹۷/۷۵	۹۳/۶۸	-	
روش پیشنهادی	KNN-PSO	۹۳/۲۶	۹۰/۵۶	۹۲/۵۶	۹۱/۵۵	Spambasc
روش پیشنهادی	KNN-PSO	۹۴/۵۶	۹۲/۸۴	۹۳/۱۵	۹۲/۹۸	Enron

- Case-Based Reasoning, ECCBR-04, Springer, Berlin, 2004, pp. 128-141.
- [7] J. Jung, E. Sit, "An empirical study of spam traffic and the use of DNS black lists", Fourth ACM SIGCOMM Conf. on Internet Measurement, Taormina, Sicily, Italy, October, 2004.
- [8] A. Thakkar, R. Lohiya, "Role of swarm and evolutionary algorithms for intrusion detection system: A survey", *Swarm and Evolutionary Computation*, Vol. 53, March 2020.
- [9] Luo, Q., Liu, B., Yan, J., He, Z., (2011), "Design and Implement a Rule-Based Spam Filtering System Using Neural Network," Proc. Int. Computational and Information Sciences (ICCIS) Conf, 398-401.
- [10] Goodman, J. "IP Addresses in Email Clients", Conference on Email and Anti-Spam 2004, July 2004.
- [11] H. Drucker, "Support Vector Machines for Spam Categorization", In IEEE Transactions on Neural Networks, 1999.
- [12] M.S. yeganeh, "A Model for Fuzzy Logic Based Machine Learning Approach for Spam Filtering" *IOSR Journal of Computer Engineering (IOSRJCE)*, Vol. 4, Issue 5, PP. 07-10, Sep-Oct. 2012.
- [13] Clark, et. al. , "A Neural Network Based Approach to Automated E-Mail Classification", *IEEE/WIC International Conference on Web Intelligence*. 2003.
- [14] Bo Yu, Zong-ben Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms", *Knowledge-Based Systems*, vol. 21, pp. 355–362, 2008.
- [15] B. Wang, et al, "Using Online Linear Classifiers to Filter Spam Emails", *Pattern Analysis & Application*, vol. 9, no. 4, pp. 339-351, 2006.
- [16] Miss Sujata S. Damawale, Prof. Vipul V. Bag, "Classification of Unstructured Data Using Machine Learning Algorithm", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vol. 5, Issue 5, September - October 2016.
- [17] M. Blachnik, et al, "Feature Selection for Supervised Classification: A KolmogorovSmirnov Class CorrelationBased Filter", *Sympisium on method of Artificial Intelligence*, Vol. 18-19, pp. 33-40, 2009.
- [18] W.A. Awad and S.M. ELseoufi, "Machine learning methods for Spam E-mail Classification", *International Journal of Computer Science & Information Technology (IJCST)*, Vol 3, No 1, February 2011.
- [19] Bolun Chen, LingChen and YixinChen, "Efficient ant colony optimization for image

۵- نتیجه‌گیری و کارهای آینده

رایانمۀ یکی از راحت‌ترین و سریع‌ترین راه‌های انتقال پیام و همچنین بستری مناسب برای ارسال رایانمدهای هرزنامه است. هرزنامه که بیشتر با هدف تجارتی فرستاده می‌شود، به عنوان تهدید امنیتی و عاملی منفی در فضای مجازی مطرح است، که تشخیص آن از بروز مشکلات احتمالی جلوگیری می‌کند. برای مقابله با رایانمدهای ناخواسته تاکنون روش‌های متعددی ایجاد و این روند با توجه به ابعاد گسترده آن، همچنان ادامه دارد. هدف این مقاله، استفاده از الگوریتم بهینهسازی از دحام ذرات برای انتخاب ویژگی و دسته‌بند K نزدیکترین همسایه برای تشخیص و دسته‌بندی هرزنامه است. از آنجایی که انتخاب ویژگی زمان‌بر است، ابتدا ابعاد ویژگی‌ها را با استفاده از نگاشت داده‌ها کاهش و سپس ویژگی‌های انتخابی را جهت دسته‌بندی به دو گروه هرزنامه و غیرهرزنامه به دسته‌بند K نزدیکترین همسایه دادیم و نتایج حاصل از معیارهای ارزیابی را با دیگر روش‌های به کار رفته مقایسه کردیم، در مقایسه با دیگر روش‌هایی که تنها از یک روش یادگیری ماشین یا یک الگوریتم فرا ابتکاری استفاده کرده‌اند، این روش ترکیبی بهتر عمل می‌کند.

۷- مراجع

- [1] S. Amjad and F.S. Gharehchopogh, "A Novel Hybrid Approach for Email Spam Detection based on Scatter Search Algorithm and K-Nearest Neighbors", *Journal of Advances in Computer Engineering and Technology*, vol.5(3), pp. 181-192, 2019.
- [2] A.A. Naem, N.I. Ghali, A.A. Salch, "Antlion optimization and boosting classifier for spam email detection", *Future Computing and Informatics Journal*, Vol. 3, Issue 2, pp. 436-442, 2018.
- [3] N. Saidani, K. Adi, M.S. Allili, "A Semantic-Based Classification Approach for an Enhanced Spam Detection", *Computers & Security*, In press, journal pre-proof, Available online 9 January 2020, Article 101716, 2020.
- [4] S. Heron, "Technologies for spam detection. Network Security", vol.1, pp. 11-15, 2009.
- [5] S. E. W.A. Awad, "Machine learning methods for spam e-mail classification", *Int. J. Comput. Sci. Inf. Technol.* 3 (1), 2011.
- [6] Delany. S. J, Cunningham. P, "An Analysis of Case-based Editing in a Spam Filtering System," In: P. Funk ad and P. A. Gonzalez-Calero, Editirs, Proceedings of 7th European Conference in

- and Random Forests for E-Mail Spam Filtering", In: Proc. of International Conference on Computational Collective Intelligence, Halkidiki, Greece, pp.498-508, 2016.
- [32] Asha Rathee, "email spam classification using integrated approach fuzzy c-means and support vector machine", International Journal of Computer Engineering and Applications, Vol.XII(I), 2018.
- [33] Foqaha, M. A. M, "Email SpamClassification Using Hybrid Approach of RBF Neural NetworkAnd Particle Swarm Optimization", International Journal of Network Security&ItsApplications, vol.8, No.4, pp.17-28, 2016.
- [34] S.M. Abdulhamid, I. Ismaila, et al, "Comparative Analysis of Classification Algorithms for Email Spam Detection", I. J. Computer Network and Information Security, , vol.1, pp. 60-67, 2018.
- [35] S. Khalandi and F.S. Gharchhopogh, "A New Approach for Text Documents Classification with Invasive Weed Optimization and Naive Bayes Classifier", Journal of Advances in Computer Engineering and Technology, vol.4 (3), pp. 31-40, 2018.
- [36] A. Allahverdipour and F.S. Gharchhopogh, "An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification", Journal of Advances in Computer Research, vol.9 (2), pp. 37-48, 2018.
- [37] H. Majidpour and F.S. Gharchhopogh, "An Improved Flower Pollination Algorithm with AdaBoost Algorithm for Feature Selection in Text Documents Classification", Journal of Advances in Computer Research, vol. 9 (1), pp. 29-40, 2018.
- feature selection", Signal Processing Vol. 93, No. 1566–1576, 2013.
- [20] D. karthikkaRenuka, P. visalakshi, Tsankar," Improving E-Mail Spam Classification using Ant Colony Optimization Algorithm", International Journal of computer Application, (ICICT2015).
- [21] Scyed Naser Razavi, Soraya Mirzapour Kalaibar", Spam filtering by using Genetic based Feature Selection", International Journal of Computer Applications Technology & Research, Vol. 3, Issue12, pp. 839-843, 2014.
- [22] Suman Sharma, Amit Arora, "Adaptive Approach for Spam Detection", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1, July 2013.
- [23] A.S.S. Idris, "Improved Email Spam Detection Model with Negative Selection Algorithm and Particle Swarm Optimization," Applied soft computing, pp.11-27, 2014.
- [24] I. Idris, A. Selamat, SigeruOmatu, "Hybrid email spam detection model with negative selection algorithm and differential evolution," Engineering Applications of Artificial Intelligence, vol.28, pp.97-110. 2014.
- [25] S.M. ElSeoufi, et al, "Enhancing E-mail Filtering Based on GRF", IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 3, No 1, May 2014.
- [26] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, "Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis", IET Computers & Digital Techniques, 2008.
- [27] Maha Adham Bayati, Saadya Fahad Jabbar," Developing a Spam Email Detector", International Journal of Engineering and Innovative Technology , Vol. 5, No. 2, August 2015.
- [28] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch, "Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis", IEEE GLOBECOM, 2008, International Journal of Computer Science & Information Technology (IJCSIT), Vol .3, No 1, Feb 2011,184.
- [29] S. Yoo, Y. Yang, F. Lin, and I. Moon, "Mining social networks for personalized email prioritization". In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France), June 28 - July 01, 2009.
- [30] T. S. Guzella and W.M. Caminhas, "A review of machine learning approaches to Spam filtering." Expert Syst. Appl., 2009.
- [31] Faris H., I. Aljarah, and B. Al-Shboul, "A Hybrid Approach Based on Particle Swarm Optimization



عاطفه مرتضوی کارشناسی خود را در رشته مهندسی کامپیوتر (نرم افزار) در دانشگاه آزاد اسلامی واحد صوفیان در سال ۱۳۹۰ و کارشناسی ارشد خود را در همان رشته در دانشگاه آزاد اسلامی واحد ارومیه در سال ۱۳۹۶ اخذ کرده است. ایشان از مقطع کارشناسی ارشد فعالیت خود را در زمینه تشخیص رایانه‌های هرزنامه شروع کرده و در نهایت موضوعی تحت عنوان بهبود الگوریتم K نزدیک ترین همسایه با الگوریتم ازدحام ذرات برای تشخیص رایانه‌های هرزنامه ارائه کرده است. زمینه‌های مورد علاقه ایشان عبارتند از: تشخیص رایانه‌های هرزنامه و الگوریتم‌های بهینه‌سازی.

افتا
من رویی
علمی
دوفصلنامه

فرهاد سلیمانیان قوه چپق کارشناسی

ارشد و دکترای تخصصی خود را در رشته

مهندسی کامپیوتر بهترتبه از دانشگاه

چوکورووا و دانشگاه حاجت‌آباد در کشور

ترکیه گرفته است و تخصص او پردازش

زبان طبیعی، داده‌کاوی و الگوریتم‌های یادگیری ماشین است.

ایشان عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه

آزاد اسلامی واحد ارومیه است و به تدریس دروس مختلف در

حوزه کاری خویش مشغول است. و در حال حاضر سردبیری

International Journal of Academic

و مدیرمسئولی نشریه Research in Computer Engineering



را بر عهده دارد.

