ارائه مدل ترکیبی الگوریتم K نزدیک‌ترین 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 نه‌تنها باعث آسیب‌رساندن به منافع کاربران، مصرف زمان و پهنای باند می‌شوند، بلکه به‌عنوان 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 اجتماع 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[1]، دقت[2]، فراخوانی[3] و F-measure 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

۱- مقدمه

امروزه، رایانامه به یکی از رایج‌ترین ابزارهای ارتباطی در زندگی روزمره بشر تبدیل شده است. این ارتباط می‌تواند یک مکالمه ساده دوستانه و یا یک موضوع مهم تجاری باشد. رایانامه روشی سریع و ارزان‌قیمت برای برقراری این ارتباط است و ارسال رایانامه به‌عنوان روشی امن، سریع و کم‌هزینه برای ارسال پیام مورد توجه کاربران اینترنت قرار گرفته است که مزایایی همچون قابلیت اطمینان بالا، هزینه انتقال به‌نسبه کم، تحویل سریع و قابیلت 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. مسأله‌ای که روزی کم‌اهمیت جلوه می‌شد، امروزه به معضلی جدی برای میلیون‌ها کاربر رایانامه بدل شده و استفاده از ابزار و روش‌هایی برای شناسایی و فیلتر رایانامه هرزنامه‌ها ضرورتی 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 اخلال ایجاد می‌کنند. از آسیب‌های هرزنامه می‌توان به فیشینگ اطلاعات، کاهش پهنای باند و انتشار بدافزارها اشاره کرد؛ به‌علاوه افزایش

[1] Accuracy [2] Precision [3] Recall

*نویسندهٔ مسؤول

محاسبات و هزینه به‌روزرسانی شبکه‌ها و تأثیر منفی بر سرویس‌های ارائه‌شده برای رایانامه‌های درست را نیز می‌توان شمرد. رایانامه‌های هرزنامه به پیام‌های ناخواسته که با استفاده از رایانامه، پیام‌رسان‌ها، وبلاگ‌ها، گروه‌های خبری، شبکه‌های اجتماعی، جستجو در وب، تلفن همراه و غیره به‌منظور تبلیغات، فیشینگ ارسال و یا دریافت الکترونیکی، انتشار بدافزار و غیره است [۱]. هدف و انگیزه هرزنامه این است که اطلاعاتی را به گیرنده منتقل کند که این اطلاعات (مانند تبلیغات برای یک محصول) که به‌احتمال 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 به‌تازگی بررسی شده اما هرزنامه‌نویس‌ها 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 فرستندگان هرزنامه تا حد زیادی از رایانامه سوء استفاده می‌کنند؛ بنابراین، هرزنامه را می‌توان به‌عنوان 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، فیشینگ اطلاعات، کاهش پهنای باند، انتشار بدافزارها، اتلاف فضای ذخیره‌سازی در خدمات‌دهندگان رایانامه، اتلاف منابع پردازش و اتلاف وقت کاربران را نام برد [۲].

خدمات‌دهندگان بزرگ به‌طور عمومی از ترکیب چند راه برای تشخیص هرزنامه استفاده می‌کنند و هم هرزنامه‌ها و هم مقابله‌کنندگان با هرزنامه در تلاش برای ایجاد راه‌های جدید برای غلبه بر روش‌های توسعه‌یافته توسط یکدیگر هستند. در [۳و۴]، 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 غیرهرزنامه است، استفاده می‌شود و فیلتر براساس مبداء هم نامیده می‌شود. در رویکرد یادگیری ماشین، مجموعه‌ای از قوانین به‌منظور تعیین دسته‌بندی رایانامه به‌عنوان هرزنامه یا غیرهرزنامه، مورد استفاده قرار می‌گیرد.

شبکه جهانی اینترنت، زمان و فاصله ارتباطات را به‌وسیله رایانامه کاهش داده است [۵]. به‌همین دلیل، کاربران اغلب برای برقراری ارتباط با دیگران و ارسال یا دریافت اطلاعات از رایانامه استفاده می‌کنند. با توجه به پژوهش‌ها، بیشتر ترافیک و پهنای باند، مربوط به رایانامه‌های هرزنامه است. در این راه، مقدار بزرگی از پهنای باند به هدر می رود و در سامانه ارسال رایانامه سرریز رخ می‌دهد. درواقع، فیلتر هرزنامه‌ها، یک برنامه رایانه‌ای است که توانایی دسته‌بندی رایانامه‌ها 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 به‌تنهایی فیلتری مؤثر باشند، ولی در کاربردهای تجاری از ترکیب این روش‌ها استفاده می‌شود. برخی از این روش‌ها، به‌صورت دستی توسط کاربر تعیین می‌شوند. از این نوع فیلترها، می‌توان به فیلترهای استفاده‌شده در رایانامه یاهو اشاره کرد. این روش‌ها یک ایراد 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 به‌طورمعمول پیام‌های تبلیغاتی هستند و ویژگی‌های مشابهی دارند؛ برای مثال آنهایی که محصولی را تبلیغ می‌کنند از قیمت آن حرف می‌زنند و یا می‌گویند که فرصتتان چقدر استثنایی است. حتی رنگارنگ‌بودن بخش‌های نوشته می‌تواند نشان از بی‌ارزش‌بودن آن باشد [۶ و ۷].

تاکنون فیلترهای ضدهرزنامه مختلفی عرضه شده‌اند که بیشتر آنها براساس تطبیق قوانین ثابت عمل می‌کنند [۸]. قوانین این سامانه‌ها به‌صورت دستی توسط کاربر تعیین می‌شوند و شامل ویژگی‌ها و مشخصات ثابت رایانامه نامعتبر یا هرزنامه هستند و عمل حذف هرزنامه براساس آنها انجام می‌شود. برای مثال فیلترهای استفاده‌شده در یاهو که کاربر نشانی‌هایی را به‌عنوان فرستنده هرزنامه معرفی کرده و سرویس‌دهنده رایانامه براساس نشانی‌های معرفی‌شده، رایانامه 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 رایانامه، هرزنامه‌ها را شناسایی و فیلتر می‌کنند. این روش‌ها بدون نیاز به نرم‌افزار فیلترکردن، توسط خدمات‌دهندگان رایانامه استفاده می‌شوند. اشکال چنین روش‌هایی این است که قوانین آنها ثابت است و باید توسط کاربر تعیین شوند و فرستندگان هرزنامه نیز می‌توانند با ترفندهای مختلفی این فیلترها را فریب داده و هرزنامه‌های خویش را از آنها عبور دهند [۱۰]. برخی از پژوهش‌گران برای

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

فیلترکردن هرزنامه‌ها به‌صورت 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 به‌صورت 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) مختصری 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 ارایهٔ 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 دسته‌بندی‌کردن 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 Ripper، 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 بیزین ساده، شبکه‌های عصبی مصنوعی، روش‌های دسته‌بندی سامانه ایمنی مصنوعی و منطق فازی را مورد بررسی قرار دادند. منطق فازی محتوای پیام را برای پیش بینی دسته آن، بر یک مجموعه واژگان کلیدی ثابت ازپیش تعیین‌شده تکیه نمی‌کند؛ بنابراین، می‌تواند با روش‌های هرزنامه سازگار باشد و دانش خود را به‌صورت پویا ایجاد کند. براین اساس، ایشان تلاش کرده‌اند تا عملکرد مدل پیشنهادی را بهبود بخشند و با شناسایی ویژگی رایانامه‌های هرزنامه در حساب رایانامه و حذف توسط خود فرستده عملکرد بهتری را در تشخیص رایانامه‌ها از هرزنامه‌ها ایفا کنند [۱۲].

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 پوشه‌ها 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 حمل‌بودن 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 موفقیت مورد استفاده قرار گیرد [۱۳].

بو و ژانگ، 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 انتخاب‌شدهٔ مختلف، انجام شده است. نتایج نشان می‌دهد که دسته‌بندی شبکه‌های عصبی مصنوعی به‌تنهایی 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 کرده‌اند. 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 به‌دست‌آمده از الگوریتم‌ها را مقایسه کرده‌اند و در پایان از معیار TF-IDF و ماشین بردار پشتیبان برای استخراج ویژگی‌ها از رایانامه و دسته‌بندی هرزنامه‌ها استفاده کرده‌اند [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

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

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

بلاک‌نیک 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 پوششی[1]) برای حذف افزونگی مناسب‌تر تشخیص داده شده است. نتایج به‌دست‌آمده 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 داده‌شده به‌طور قابل توجهی می‌تواند ابعاد مسأله براساس انتخاب ویژگی را برای حل مسایل با ابعاد بزرگ کاهش دهد [۱۷].

آواد و همکاران، از روش‌های یادگیری ماشین (دسته‌بندی بیزین‌ساده، الگوریتم K نزدیک‌ترین همسایه، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان، سامانه ایمنی مصنوعی و مجموعه‌های راف[2]) برای دسته‌بندی رایانامه‌های هرزنامه استفاده کردند و برخی از محبوب‌ترین روش‌های یادگیری ماشین و کاربرد آنها را در رابطه با دسته‌بندی رایانامه 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؛ درحالی‌که در شرایط دقت ما می‌توان دریافت که روش‌های بیزین ساده 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 $O(n^2)$ دارند با این حال، اینها یک الگوریتم جدید ارائه دادند که در آن مورچه‌های مصنوعی به یک گراف فقط با قوس $O(2n)$ هدایت می‌شوند. الگوریتم شامل عملکرد دسته‌بندی و اندازه مجموعه‌ای از ویژگی‌ها را به‌صورت فرااکتشافی انجام می‌دهد و مجموعه‌ای از ویژگی‌های 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 رایانامه، پژوهشی را انجام داده‌اند و نتایج آن‌ها را به‌وسیلهٔ 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 انتخاب‌شده برای دسته‌بندی مجموعه‌داده استفاده شده است. نتایج به‌دست‌آمده 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 توانسته است میزان دقت را افزایش دهد [ ].

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 برروی دقت و عملکرد الگوریتم برای دسته‌بندی رایانامه‌های هرزنامه و غیرهرزنامه تمرکز کردند. نتایج تجربی بر روی داده‌های Spambase نشان می‌دهد که عملکرد روش پیشنهادی بهتر است و دقتی برابر با ۹۴/۲۸ را به‌دست آوردند که از تمام روش‌های دیگر بالاتر است [۲۲].

[1] Wrapper
[2] Rough

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

ادریس و همکاران، یک مدل بهبودیافته جدید را که ترکیبی از الگوریتم انتخاب منفی با الگوریتم بهینه‌سازی ازدحام ذرات است، پیشنهاد داده‌اند. ویژگی منحصربه‌فرد این مدل این است که الگوریتم بهینه‌سازی ازدحام ذرات در مرحله تصادفی، الگوریتم انتخاب منفی تصادفی را اجرا می‌کند. مدل پیشنهادی به‌عنوان یک جایگزین بهتر برای مدل الگوریتم انتخاب منفی عمل می‌کند. مجموعه‌داده Spambase برای بررسی عملکرد الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی ازدحام ذرات در برابر بهبود مدل ترکیبی الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی ازدحام 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 ازدحام 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 ازدحام ذرات بر مدل ترکیبی الگوریتم انتخاب منفی و الگوریتم بهینه‌سازی ازدحام 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 کردند. نکتهٔ منحصربه‌فردبودن 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 به‌عنوان تابع تناسب برای به بیشینه‌رساندن فاصله شناسه‌های هرزنامهٔ تولیدشده از فضای غیرهرزنامه 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 به‌جای 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. مجموعه‌داده Spamassasian برای تأیید عملکرد سامانه پیشنهادی استفاده شده است. مجموعه راف، پیشرفت قابل توجهی در شبکه عصبی مصنوعی، مجموعه راف و الگوریتم ماشین بردار پشتیبان از لحاظ 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].

مارسونا و همکاران نشان دادند که دسته‌بندی رایانامه یا دسته‌بند 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 دسته‌بند بیزین‌ساده را پیشنهاد دادند، که این دسته‌بند پیام‌های رایانامه را براساس محتوای پیام‌ها به‌عنوان هرزنامه یا رایانامه شناسایی می‌کند. هر رایانامه به‌عنوان یک مجموعه‌ای از واژگان (ویژگی‌ها) آن نشان داده شده است. مجموعه‌دادهٔ CSDMC2010 برای آموزش و آزمون دسته‌بند بیزین‌ساده مورد استفاده قرار گرفته است [۲۷].

تانگ و همکاران سامانه‌ای را پیشنهاد کرد که از دسته‌بند ماشین بردار پشتیبان برای مقاصد دسته‌بندی استفاده کردند؛ 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 کردند. آنها دو دسته‌بندی به‌نام الگوریتم ماشین بردار پشتیبان و الگوریتم جنگل تصادفی برای شناسایی سرورهای رایانامه مخرب را معرفی کردند. ویژگی‌های رفتار ارسالی جهانی استخراج‌شده از سوابق پرس‌وجوی ساده به‌دست 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 منابع مؤثر برای ایجاد دسته‌بندی دقیق هستند. بین آنها، جنگل تصادفی به‌مراتب دقیق‌تر، اما از لحاظ زمان و فضا 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

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

پشتیبان دسته‌بندی مؤثر، دقیق و بسیار سریع‌تر از دسته‌بندی تصادفی جنگل‌ها انجام می‌دهد [۲۸].

یانگ و همکاران نخستین روش اولویت‌بندی رایانامه را توسعه دادند که به‌طور خاص بر تجزیه و تحلیل شبکه‌های اجتماعی شخصی برای جذب گروه‌های کاربر و به‌دست‌آوردن ویژگی‌های غنی که نقش اجتماعی را از نظر کاربر خاص نشان می‌دهند، تمرکز می‌کنند و همچنین آنها یک چارچوب دسته‌بندی نظارت‌شده برای مدل‌سازی اولویت‌های شخصی بر روی پیام‌های رایانامه و برای پیش‌بینی سطوح اهمیت برای پیام‌های جدید ایجاد کردند [۲۹].

گوزلا و همکاران، یک مدل الهام‌گرفته از سامانه ایمنی به نام سامانه ایمنی 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 به‌دست آورده‌اند و می‌توان نتیجه گرفت که IA-AIS دارای توانایی بیشتری برای شناسایی پیام‌های هرزنامه است.

فریس و همکاران، روشی 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 یک دسته‌بند جنگل تصادفی با استفاده از ویژگی‌های انتخاب‌شده در مرحله نخست دادهٔ آزمون را دسته‌بندی کرده و درنهایت 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].

راته در پژوهشی، روش ترکیبی الگوریتم فازی C-Mean و دسته‌بند ماشین بردار پشتیبان را برای مسأله رایانامه‌های هرزنامه ارائه داد. در این روش ترکیبی، الگوریتم FCM در ایجاد خوشه‌هایی از داده رایانامه برای آموزش دسته‌بند ماشین بردار پشتیبان به کار می‌رود در نهایت عمل دسته‌بندی انجام می‌شود. در این روش از مجموعه‌داده LingSpam برای ارزیابی روش ترکیبی استفاده شده است و درواقع مجموعه‌داده رایانامه‌های متنی با روش ترکیبی فازی C-Means و دسته‌بند ماشین بردار پشتیبان به‌صورت رایانامه‌های هرزنامه و غیرهرزنامه دسته‌بندی شده‌اند و نتایج ارزیابی نشان داده است که این روش در مقایسه با روشی که تنها از دسته‌بند ماشین بردار پشتیبان برای دسته‌بندی استفاده کرده‌اند، عملکرد بهتری داشته است [۳۲].

فوقا و آواد، الگوریتم بهینه‌سازی ازدحام ذرات را برای فیلترینگ رایانامه‌های هرزنامه با الگوریتم شبکه‌های عصبی تابع پایه رادیال (که به‌اختصار 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 الگوریتم دسته‌بندی مختلف استفاده‌شده را برای تشخیص رایانامه‌های هرزنامه بررسی کردند و نتایج به‌دست‌آمده نشان می‌دهد که میزان صحت الگوریتم دسته‌بندی جنگل تصادفی برابر با ۹۴/۳ درصد از بقیه دسته‌بندها عملکرد بهتری را نشان داده است.

۳- روش پیشنهادی

۳-۱- انتخاب ویژگی

پیداکردن زیرمجموعه بهینه از یک مجموعه‌بزرگ ویژگی، مسأله‌ای 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. مسأله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مسأله 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 این‌که بار اطلاعاتی چندانی ندارند [۳۵]. حذف‌نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند، ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد و علاوه‌براین 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 [۳۷,۳۶]. به عبارت دیگر، کاهش اندازه ساختار بدون کاهش قابل ملاحظه در دقت پیش‌گویی دسته‌بندی‌کننده‌ای که با

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

استفاده از ویژگی‌های داده‌شده به‌دست می‌آید. به‌دست‌آوردن یک زیرمجموعه بهینه از مجموعه ویژگی‌ها، به‌صورت مستقیم با انتخاب تابع ارزیابی بستگی دارد؛ چون که اگر تابع ارزیابی به زیرمجموعه ویژگی بهینه یک مقدار نامناسب نسبت دهد، این زیرمجموعه هیچگاه بعنوان زیرمجموعه بهینه انتخاب نمی‌شود.

ما در این مقاله، با استفاده از الگوریتم ازدحام ذرات، انتخاب ویژگی را انجام داده‌ایم و سپس ویژگی‌های انتخاب‌شده را به دسته‌بند K نزدیک‌ترین همسایه داده تا دقت، صحت و حساسیت آن را محاسبه کنیم.

۲-۳- اجرای الگوریتم بهینه‌سازی ازدحام ذرات

در برخی مراجع، در نحوه تفکیک مراحل اجرای الگوریتم، اختلافاتی دیده می‌شود؛ یعنی در برخی مواقع مراحل به‌صورت تفکیک‌شده‌تری 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، اجراشدن مراحل برنامه به ترتیبی که در ادامه خواهد آمد، و نحوه تفکیک این مراحل بدین صورت است.

مرحلهٔ ۱، تولید تصادفی جمعیت اولیه

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

$$V_{min} \leq V \leq V_{max} \qquad (۱)$$

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

مرحلهٔ ۲، انتخاب تعداد ذرات اولیه

افزایش تعداد ذرات اولیه موجب کاهش تعداد تکرارهای لازم برای هم‌گرا شدن الگوریتم می‌شود؛ اما گاهی مشاهده می‌شود که کاربران الگوریتم‌های بهینه‌سازی تصور می‌کنند، که این کاهش در تعداد تکرارها به معنی کاهش زمان اجرای برنامه برای رسیدن به هم‌گرایی است. درحالی‌که چنین تصوری به‌طورکامل غلط است. هرچند که افزایش تعداد ذرات اولیه، 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 نیابد؛ 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 گیرافتادن 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 به‌دست‌آمده، حل بهینه‌ای برای مسأله نخواهد بود؛ زیرا الگوریتم به‌صورت ناقص اجرا شده است.

به‌اختصار، تعداد جمعیت اولیه با توجه به مسأله تعیین می‌شود. در حالت کلی، تعداد ذرات اولیه مصالحه‌ای بین پارامترهای درگیر در مسأله است. به‌طور تجربی انتخاب جمعیت اولیه ذرات به تعداد بیست تا سی ذره انتخاب مناسبی است که به‌طور تقریبی برای تمامی مسائل آزمون به‌خوبی 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 به‌دست بیاید.

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

در این مرحله، باید هر یک از ذرات را که نشان‌دهندهٔ یک حل برای مسأله مورد بررسی است، ارزیابی کنیم. بسته به مسأله 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، به‌راحتی مقدار هزینه

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

این ذره محاسبه خواهد شد. توجه داشته باشید که هر ذره حاوی اطلاعات کاملی از پارامترهای ورودی مسأله 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 این‌گونه 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 و 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.

مرحلهٔ۴، ثبت بهترین موقعیت برای هر ذره ($P_{i,best}$) و بهترین موقعیت در بین کل ذره‌ها ($p_{g.best}$)

در این مرحله، با توجه به شماره تکرار، دو حالت قابل بررسی است:

اگر در تکرار نخست باشیم (t=1). موقعیت فعلی هر ذره را به‌عنوان بهترین محل یافت‌شده برای آن ذره در نظر می‌گیریم:

$$P_{i,\,best} = X_i(t)\ ,\ i=1,2,3,\ldots,d \qquad (۲)$$

$$cost\left(P_{i,best}\right) = cost\left(X_j(t)\right) \qquad (۳)$$

در سایر تکرارها، مقدار هزینهٔ به‌دست‌آمده برای ذرات در مرحلهٔ ۲ را با مقدار بهترین هزینه به‌دست‌آمده برای تک‌تک ذرات مقایسه می‌کنیم. اگر این هزینه، کمتر از بهترین هزینهٔ ثبت‌شده برای این ذره باشد، آنگاه محل و هزینه این ذره جایگزین مقدار قبلی می‌شود. در غیر این صورت تغییری در محل و هزینه ثبت‌شده برای این ذره ایجاد نمی‌شود؛ یعنی:

$$\begin{cases} \text{if } cost(X_i(t)) < cost\left(P_{i,best}\right) \\ \quad \text{else Not change} \end{cases} \Rightarrow \begin{cases} cost\left(P_{i,best}\right) = cost\left(X_j(t)\right) \\ P_{i,best} = X_i(t) \end{cases}_{i = 1,2,3,\ldots,d} \qquad (۴)$$

مرحلهٔ ۵، به‌روزرسانی بردار سرعت تمامی ذره‌ها

(۵)

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

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

ضرایب c_1، c_2 و w با توجه به مسأله مورد نظر به روش تجربی تعیین می‌شوند؛ اما به‌عنوان یک قانون کلی، w باید کمتر از یک باشد؛ زیرا، اگر بزرگ‌تر از یک انتخاب شود، V(t) به‌طور 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 منفی‌بودن 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 ازدحام ذرات مقدار این ضریب را مثبت و در بازه ۰/۷ تا ۰/۸ در نظر می‌گیرند. پارامترهای c_1 و c_2 نیز نباید زیاد بزرگ انتخاب شوند؛ زیرا انتخاب مقادیر بزرگ برای این دو ضریب باعث انحراف شدید ذره از مسیر خودی می‌شود. اغلب در الگوریتم ازدحام ذرات مقدار این ضرایب را مثبت و در بازه ۱/۵ الی ۱/۷ در نظر می‌گیرند.

(شکل-۱): روندنمای الگوریتم بهینه‌سازی ازدحام ذرات

مرحلهٔ۶، آزمون هم‌گرایی

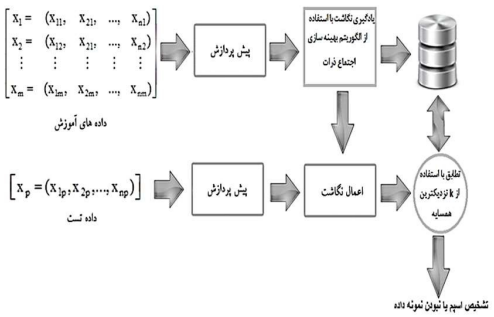
آزمون هم‌گرایی در این الگوریتم، مانند سایر الگوریتم‌های بهینه‌سازی است. برای بررسی روش‌های گوناگونی وجود دارد.

برای مثال: می‌توان تعداد مشخصی تکرار را از همان ابتدا معلوم کرد و در هر مرحله بررسی کرد که آیا تعداد تکرارها به مقدار تعیین‌شده رسیده است؟ اگر تعداد تکرارها کوچک‌تر از مقدار تعیین‌شده اولیه باشد، آنگاه باید به مرحله ۲ بازگردید، درغیر این‌صورت، الگوریتم پایان می‌پذیرد. روش دیگری که اغلب در آزمون هم‌گرایی الگوریتم استفاده می‌شود، این است که اگر در چند تکرار متوالی برای مثال پانزده یا بیست تکرار تغییری در مقدار هزینه بهترین ذره ایجاد نشود، آنگاه الگوریتم پایان می‌یابد، در غیر این صورت، باید به مرحله ۲ بازگردید. روندنمای 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 دسته‌بندی‌کردن دو نمونه رایانامه‌های هرزنامه و غیرهرزنامه است. بدین صورت که ابتدا داده‌ها را به دو دسته آموزش و آزمون تقسیم کرده، که ۸۰٪ داده‌ها به آموزش و ۲۰٪ به داده‌های آزمون تخصیص داده شده است. الگوریتم بهینه‌سازی، اجتماع ذرات را روی داده‌های آموزش اجرا کرد. به این صورت که ابتدا جمعیت اولیه به‌صورت تصادفی انتخاب کرده و موقعیت هر ذره را به‌صورت اعداد اعشاری بین (۰و۱) تعریف کرده و در زمان محاسبه برازندگی این اعداد گرد می‌شوند و در موقعیت یک ذره برای نگاشت داده‌ها استفاده می‌کنیم. که اگر داده‌ها به‌صورت 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 برازندگی موقعیت هر ذره، ابتدا فرض می‌کنیم، خود ذره بهتر است. اگر در تکرار نخست باشیم، موقعیت فعلی هر ذره را به‌عنوان بهترین محل یافت‌شده برای آن ذره در نظر می‌گیریم. در سایر تکرارها، مقدار هزینه به‌دست‌آمده برای ذرات در مرحله دو را با مقدار بهترین هزینه به‌دست‌آمده برای تک‌تک ذرات مقایسه شده است. اگر این هزینه کمتر از بهترین هزینه ثبت‌شده برای این ذره باشد، آنگاه محل و هزینه این ذره جایگزین مقدار قبلی می‌شود. در غیر این صورت، تغییری در محل و هزینه ثبت‌شده برای این ذره ایجاد نمی‌شود.

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 پیش‌بینی‌شده و از خطای پیش‌بینی آن به‌عنوان معیار برازندگی یک ذره استفاده می‌کنیم؛ بنابراین ذره‌ای بهتر است که خطای کمتری داشته باشد؛ سپس با gbest و با گردکردن مقادیر آن نگاشت (۰و۱) نهایی به‌دست آمده و داده‌های آموزش و آزمون را متناظر 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 فراهم‌کردن مجموعه‌داده، نوبت به مرحله پیش‌پردازش می‌رسد و در این مرحله اسناد متنی 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 به‌صورت 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.

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

۲-۳-۳- یادگیری نگاشت با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات

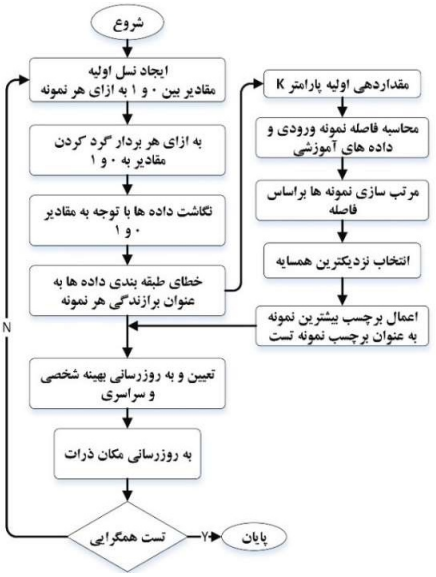
در این مرحله، با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات انتخاب ویژگی انجام می‌گیرد. این الگوریتم را روی داده‌های آموزش اجرا کرده تا ویژگی‌های مهم انتخاب شده و بتوانیم ویژگی‌های اضافی را حذف کرده با این کار ابعاد ویژگی‌ها کم می‌شود.

۳-۳-۳- اعمال نگاشت

بعد از انتخاب ویژگی که نگاشت نهایی به‌صورت (۰و۱) به‌دست آمد، داده‌های آزمون و آموزش نگاشت می‌شود؛ سپس داده‌های نگاشت‌داده‌شده را به الگوریتم K نزدیک‌ترین همسایه داده می‌شوند.

۴-۳-۳- تشخیص هرزنامه یا نبودن رایانامه

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



(شکل-۳): روندنمای نگاشت داده‌ها و انتخاب ویژگی‌ها با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات

در مدل پیشنهادی، ویژگی‌هایی انتخاب می‌شوند که بهترین برازندگی را دارند. تابع برازندگی 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 به‌ترتیب برابر با ۹۹ و ۱ است.

$$Fitness = \delta . Accuracy + \rho . \frac{|n| - |S|}{|n|} \quad (۶)$$

۴- ارزیابی و نتایج

۱-۴- 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 به‌عنوان ویژگی ورودی و ویژگی ۵۸ام ویژگی خروجی است که نشان‌دهنده هرزنامه یا غیرهرزنامه‌بودن رایانامه است. در صورت صفربودن رایانامه، هرزنامه و اگر یک باشد، رایانامه مربوطه عادی است.

۲-۴- معیارهای ارزیابی

معیارهایی که به‌کمک 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 مسأله دسته‌بندی نشان می‌دهد. TN: تعداد رایانامه‌های غیرهرزنامه که درست دسته‌بندی شده‌اند.TP: تعداد رایانامه‌های هرزنامه که درست دسته‌بندی شده‌اند. FN: تعداد رایانامه‌های هرزنامه که به‌اشتباه به‌عنوان رایانامه غیرهرزنامه دسته‌بندی شده‌اند. FP: تعداد رایانامه‌های غیرهرزنامه که به‌اشتباه به‌عنوان رایانامه هرزنامه دسته‌بندی شده‌اند.

اشتباه برچسب‌زدن هرزنامه، دارای عوارض جانبی است. یک برچسب هرزنامه به‌عنوان 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 به‌عنوان هرزنامه مشخص شده‌اند، به‌طورمعمول با دنباله‌ای از حذف به‌صورت خودکار دنبال و باعث می‌شود کاربر رایانامه‌های مهم را از دست بدهد و یا با انتقال خودکار به جعبه هرزنامه کاربر به‌احتمال زیاد آن را نادیده بگیرد؛ بنابراین، نیاز به استفاده از ماتریس درهم‌ریخته برای توصیف انواع مختلف خطاها و اندازه‌گیری این‌که چقدر جدی است، وجود دارد. در ماتریس درهم‌ریخته هر ستون نشان‌دهندهٔ موارد در یک رده پیش‌بینی شده و هر سطر نشان‌دهندهٔ موارد در یک ردهٔ واقعی. دسته مثبت به نمایندگی از هرزنامه و دسته منفی به نمایندگی از غیرهرزنامه است. در اینجا، TP نشان صحیح مثبت، FP نشان غلط مثبت، FN نشان غلط منفی و TN نشان صحیح منفی است.

(جدول-۱): ماتریس در هم‌ریختگی

رده	غیر هرزنامه	هرزنامه
پیش‌بینی هرزنامه	FP	TP
پیش‌بینی غیرهرزنامه	TN	FN

مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی، نرخ صحت است. این معیار دقت کل یک دسته‌بند را محاسبه می‌کند. این معیار نشان‌دهندهٔ این موضوع است که چند درصد از کل مجموعه‌داده‌ها به‌درستی دسته‌بندی شده است. معیار فراخوانی، کارایی دسته‌بند را با توجه به تعداد رخ داد دسته i نشان می‌دهد؛ درحالی‌که معیار دقت به‌طوراساسی مبتنی بر دقت پیش‌بینی دسته و بیان‌گر آن است که به چه میزان می‌توانیم به خروجی دسته‌بند اعتماد کرد. معیار F-measure از ترکیب معیارهای دقت و فراخوانی به‌دست می‌آید، و در مواردی استفاده می‌شود که نتوان اهمیت ویژه‌ای برای هریک از دو معیار دقت و فراخوانی نسبت به یکدیگر قائل شد.

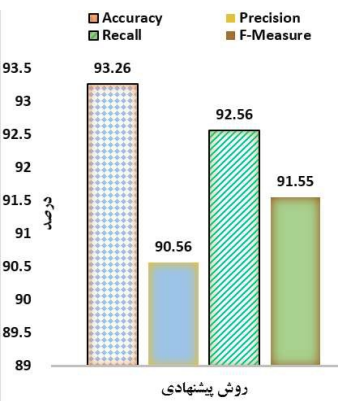
$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (۷)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (۸)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (۹)$$

$$\text{F-Measure}=2*\frac{Precision*Recall}{Precision+Recall} \quad (۱۰)$$

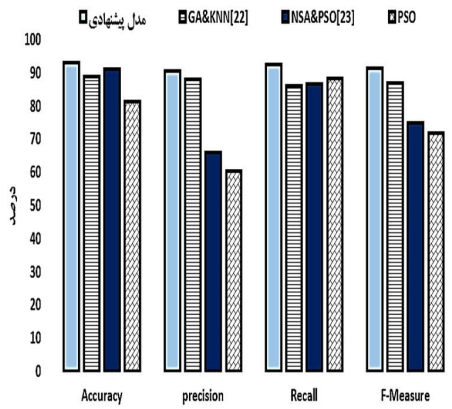
بر مبنای الگوریتم k نزدیک‌ترین همسایه، در ابتدا نمونه‌های آموزشی ارزیابی و یک مدل به‌منظور تشابه بین نمونه‌ها ایجاد می‌شود. در مدل پیشنهادی، مقدار k برابر با ۳ فرض شده است. نمونه‌ها برمبنای نزدیکی فاصله و تشابه به دسته‌ها (دسته هرزنامه و غیرهرزنامه) اختصاص داده می‌شوند. در ابتدا باید نمونه‌ای برای دسته‌بندی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



(شکل-۴): نتایج حاصل از مدل پیشنهادی

درشکل (۴)، نمودار مقادیر معیارهای دقت، فراخوانی، F-measue و درصد صحت روش پیشنهادی ما نشان داده شده است که مقادیر آنها به‌ترتیب از چپ به راست ۹۱/۵۵، ۹۲/۵۶، ۹۰/۵۶ و ۹۳/۲۶ درصد است که این مقادیر با تکرار بالای پنجاه به‌دست آمده‌اند.

در شکل (۵)، مقادیر به‌دست‌آمده مدل پیشنهادی ما با دیگر مدل‌های مشابه نظیر ترکیب الگوریتم ژنتیک با دسته‌بندی K نزدیک‌ترین همسایه [۲۱] و مدل الگوریتم انتخاب منفی با الگوریتم بهینه‌سازی ذرات و مدلی که تنها از الگوریتم ازدحام ذرات استفاده کرده [۲۳]، مقایسه شده است و همان‌طوری که مشاهده می‌کنید هر گروه نمودار نشان‌دهنده یک معیار است، که نتایج هر چهار معیار یادشده با هم مقایسه شده است که نتایج مدل پیشنهادی ما مقادیر بیشتری نسبت به مدل‌های دیگر به‌دست آورده است. مدلی که از الگوریم ژنتیک برای انتخاب ویژگی و بهینه‌سازی دسته‌بندی K نزدیک‌ترین همسایه استفاده کرده، دقتی معادل ۸۹درصد و مدلی که از ترکیب الگوریتم انتخاب منفی و بهینه‌سازی ذرات برای تشخیص رایانامه‌های هرزنامه استفاده کرده، دقتی معادل ۹۱درصد و همچنین خود الگوریتم بهینه‌سازی ذرات قادر به تشخیص رایانامه‌های هرزنامه با دقتی معادل ۸۱درصد است. با این مقایسه‌ها الگوریتم پیشنهادی، ما با دقتی برابر با ۹۳درصد عملکرد بهتری نشان 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 به‌دست‌آمده با تکرار بالای پنجاه به‌دست آمده و در هر بار تعداد ویژگی‌های انتخاب‌شدهٔ ثبت‌شده را گرد هم آورده‌ایم. برای مثال، هنگامی که تعداد

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

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

در جدول (۳)، مدل‌های استفاده‌شده برای تشخیص رایانامه‌های هرزنامه با ذکر منابع آنها و مقادیر به‌دست‌آمدهٔ آنها برای معیارهای صحت، دقت، فراخوانی و F-Measure با مقادیر به‌دست‌آمده توسط مدل پیشنهادی ما با هم مقایسه شده‌اند.

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

F-Measure	فراخوانی	دقت	صحت	تعداد ویژگی‌ها
۸۴/۴۳	۸۸/۱۵۴	۸۱/۰۱	۸۷/۱۸	۲۷
۹۱/۵۵	**۹۲/۵۶**	**۹۰/۵۶**	**۹۳/۲۶**	۲۸
۸۲/۵۸	۸۸/۱۵	۷۷/۶۶	۸۵/۳۴	۲۹
۸۷/۱۵	۸۸/۱۵	۸۶/۷۲	۹۰/۰۱	۳۰
۸۸/۵۲	۹۰/۳۵	۸۶/۷۷	۹۰/۷۷	۳۱
۸۳/۴۸	۸۸/۴۲	۷۹/۰۶	۸۶/۲۱	۳۲
۹۰/۱۷۶	۸۹/۸۰	۹۰/۵۵	۹۲/۲۹	۳۳
۸۵/۴۸	۸۷/۴۲	۸۶/۰۱	۸۹/۲۱	۳٤
۸۸/۹۱	۹۰/۶۳	۸۷/۲۶	۹۱/۰۹	۳۵
۷۹/۱۷	۷۹/۶۱	۷۸/۷۴	۸۳/۴۹	**کل ویژگی‌ها**

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

مجموعه داده	F-measure	فراخوانی	دقت	صحت	Models	Ref
Spambase	۸۷	۷۸	۸۹	۸۴	Aco-NB	[20]
	۷۵	۷۱	۸۵	۷۷	GA-NB	
	۹۰	۸۶	۹۳	۹۱	GA-BN	[21]
	۸۷	۸۶	۸۸	۸۹	GA-Knn	
	۹۰/۸	۸۸/۶	۹۳/۱۱۳	۸۸/۵۶	BN	
	۹۱/۶۶	۹۰/۴	۹۲/۸۶۲	۸۹/۷	LB	
	۹۳	۹۳/۲	۹۲/۷۵۴	۹۱/۵۴	RT	
	۹۳/۷۱	۹۲/۹	۹۴/۴۷۶	۹۲/۳۲	JR	[22]
	۹۳/۷	۹۳/۵	۹۳/۹۰۲	۹۲/۳۴	J48	
	۹۴/۴۵	۹۴/۵	۹۴/۳۳۲	۹۳/۲۸	MP	
	۹۴/۷۳	۹۳/۹	۹۵/۵۸۸	۹۳/۵۶	KS	
	۹۵	۹۴/۱	۹۵/۸۷۵	۹۳/۸۹	RF	
	۹۵/۳۲	۹۴/۵	۹۶/۱۲۶	۹۴/۲۸	RC	
	۷۱/۸۴	۸۸/۴۴	۶۰/۴۸	۸۱/۳۲	PSO	
Spamassasian	۳۶/۰۱	۹۴/۵۳	۲۲/۲۴	۶۸/۸۶	NSA	[23]
	۷۴/۹۵	۸۶/۷۲	۹۹/۶۵	۹۱/۲۲	NSA-PSO	
	۳۶/۰۱	۹۴/۵۳	۲۲/۲۴	۶۸/۸۶	NSA	[24]
	۶۹/۷۶	۹۰/۸۶	۵۶/۶۲	۸۰/۶۶	NSA-DE	
	-	۹۸/۴۶	۹۹/۶۶	۹۹/۴۶	NB	[25]
	-	۹۷/۱۴	۸۷/۰۰	۹۶/۲۰	KNN	
	-	۹۶/۹۲	۹۶/۰۲	۹۶/۸۳	NN	
	-	۹۵/۰۰	۹۳/۱۲	۹۶/۹۰	SVM	
	-	۹۳/۶۸	۹۷/۷۵	۹۶/۲۳	AIS	
Spambase	۹۱/۵۵	۹۲/۵۶	۹۰/۵۶	۹۳/۲۶	KNN-PSO	روش پیشنهادی
Enron	۹۲/۹۸	۹۳/۱۵	۹۲/۸۴	۹۴/۵۶	KNN-PSO	روش پیشنهادی

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

۵- نتیجه‌گیری و کارهای آینده

رایانامه یکی از راحت‌ترین و سریع‌ترین راه‌های انتقال پیام و همچنین بستری مناسب برای ارسال رایانامه‌های هرزنامه است. هرزنامه که بیشتر با هدف تجاری فرستاده می‌شود، به‌عنوان تهدید امنیتی و عاملی منفی در فضای مجازی مطرح است، که تشخیص آن از بروز مشکلات احتمالی جلوگیری می‌کند. برای مقابله با رایانامه‌های ناخواسته تاکنون روش‌های متعددی ایجاد و این روند با توجه به ابعاد گسترده آن، همچنان ادامه دارد. هدف این مقاله، استفاده از الگوریتم بهینه‌سازی ازدحام ذرات برای انتخاب ویژگی و دسته‌بند K نزدیک‌ترین همسایه برای تشخیص و دسته‌بندی هرزنامه‌ها 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 به‌کار 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 Eruopean Conference in 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 forSpam Categorization", In IEEE Transactions on NeuralNetworks, 1999.

[12] M.S. yeganeh, "A Model for Fuzzy Logic Based Machine Learning Approach forSpam 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 Approachto Automated E-Mail Classification", *IEEE/WICInternational 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. ELseuofi, "Machine learning methods for Spam E-mail Classification", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol 3, No 1, February 2011.

[19] Bolun Chen, LingChen and YixinChen, "Efficient ant colony optimization for image

[DOR: 20.1001.1.24763047.1398.8.1.5.1]

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] Seyed Naser Razavi, Sorraya 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] Sumant 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 whit Negatative 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. ELseuofi, 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 & DigitalTechniques*, 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 VectorMachines 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 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 ofNetwork 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. Gharehchopogh, "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. Gharehchopogh, "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. Gharehchopogh, "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.

عاطفه مرتضوی کارشناسی خود را در رشته مهندسی کامپیوتر (نرم‌افزار) 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 رایانامه‌های هرزنامه شروع کرده و درنهایت موضوعی تحت عنوان بهبود الگوریتم K نزدیک‌ترین همسایه با الگوریتم ازدحام ذرات برای تشخیص رایانامه‌های هرزنامه ارائه کرده است. زمینه‌های مورد علاقه ایشان عبارتند از: تشخیص رایانامه‌های هرزنامه و الگوریتم‌های بهینه‌سازی.

فرهاد سلیمانیان قره چپق کارشناسی ارشد و دکترای تخصصی خود را در رشته مهندسی کامپیوتر به‌ترتیب از دانشگاه چوکوروا و دانشگاه حاجت‌تپه در کشور ترکیه گرفته است و تخصص وی 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 و مدیرمسئولی نشریه International Journal of Academic Research in Computer Engineering را برعهده دارد.

[DOR: 20.1001.1.24763047.1398.8.1.5.1]