یک مدل جدید برای تشخیص ایمیل هرزنامه با استفاده از ترکیب الگوریتم بهینه‌سازی مغناطیسی با الگوریتم جستجوی هارمونی

فرهاد سلیمانیان قره چپق*[1] و محمد سخی دل[2]

[1]استادیار، گروه مهندسی رایانه، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
bonab.farhad@gmail.com

[2]کارشناس ارشد مهندسی رایانه، گروه مهندسی رایانه، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
mohammad.sakhidel@gmail.com

چکیده

متأسفانه در میان خدمات اینترنت، کاربران با یک‌سری پیام‌های ناخواسته که حتی به علایق و حیطه کاری آنان مرتبط نیست و حاوی مطالب تبلیغاتی یا حتی مخرب هستند، مواجه می‌شوند. هرزنامه شامل مجموعه گسترده‌ای از رایانامه‌های تبلیغاتی آلوده و مخرب است که با اهداف خراب‌کارانه موجب زیان، از بین رفتن داده‌ها و سرقت اطلاعات شخصی می‌شود. در اغلب موارد، ایمیل‌های هرزنامه حاوی بدافزارهایی هستند که به‌طورمعمول در قالب اسکریپت یا فایل‌های ضمیمه برای کاربران ارسال می‌شوند و کاربر با بارگیری و اجرای فایل ضمیمه‌شده، رایانه خود را به بدافزار آلوده می‌کند. در این مقاله یک روش جدید برای تشخیص ایمیل هرزنامه برمبنای ترکیب الگوریتم جستجوی هارمونی با الگوریتم بهینه‌سازی مغناطیسی پیشنهاد می‌شود. روش پیشنهادی به‌منظور انتخاب ویژگی‌های تأثیرگذار استفاده و سپس طبقه‌بندی با استفاده از الگوریتم K نزدیکترین همسایه انجام می‌شود. در روش پیشنهادی با استفاده از الگوریتم بهینه‌سازی مغناطیسی، بهترین ویژگی‌ها را برای الگوریتم جستجوی هارمونی پیدا می‌کنیم و ماتریس هارمونی برمبنای آنها تشکیل می‌شود؛ سپس الگوریتم جستجوی هارمونی برمبنای به‌روزرسانی و نرخ تغییرات گام در هر مرحله بردارهای هارمونی را تغییر می‌دهد تا در میان آنها بهترین بردار به‌عنوان بردار ویژگی‌ها انتخاب شود. نتایج ارزیابی‌ها برروی مجموعه‌داده Spambase نشان می‌دهد که روش پیشنهادی با تعداد تکرارهای بیشتر، درصد صحت بیشتری دارد و با دویست بار تکرار، دقت تشخیص آن برابر با ۹۴/۱۷ درصد است.

واژگان کلیدی: تشخیص ایمیل هرزنامه، انتخاب ویژگی، الگوریتم جستجوی هارمونی، الگوریتم بهینه‌سازی مغناطیسی، درصد صحت

۱- مقدمه

با توجه به خدمات متعدد روی شبکه اینترنت، تاکنون مهم‌ترین خدمت از میان خدمت‌های گوناگون اینترنت، خدمت رایانامه بوده است [۱]. بسیاری از کاربران اینترنت از این بخش بیشتر از سایر امکانات فراهم‌آمده به‌وسیلهٔ این شبکه جهانی استفاده می‌کنند؛ اما در مقابل یک‌سری از نفوذگران اینترنتی از سامانه رایانامه برای مقاصد دیگری مانند سرقت اطلاعات و نفوذ به سیستم‌های کاربران استفاده می‌کند. ایمیل هرزنامه درواقع عمل ارسال پیام‌های بدون درخواست و بطور مکرر با محتوای تجاری در مقیاس بزرگ به کاربران است [۲]. ایمیل هرزنامه در اواسط ۱۹۹۰ وقتی که وارد اینترنت شد، تبدیل به یک معضل شد و روزبه‌روز گسترش یافت؛ به‌صورتی که امروزه با یک تخمین محافظه کارانه ۸۰ تا ۸۵ درصد ایمیل‌ها را در بر می‌گیرد و در بعضی موارد از ۹۵ درصد بالاتر می‌رود [۳].

تشخیص بر اساس محتوا و واژگان، ساده‌ترین و رایج‌ترین راه برای شناسایی ایمیل هرزنامه است. اگر محتوای ایمیل و یا محتوای اجزای تشکیل‌دهنده رایانامه مانند عنوان، تگ‌ها، پیوندهای موجود در صفحه و نشانی اینترنتی شامل واژگان خاصی باشد، به‌عنوان ایمیل هرزنامه تشخیص داده می‌شود [۴]. هرزنامه‌نویسان اغلب از عبارات خاص و جذاب مانند رضایت، ارزان، خرید، برنده و... برای جلب توجه کاربران در رایانامه یا تارنما استفاده می‌کنند. به‌طورمعمول هرزنامه‌نویسان واژگان مورد استفاده خود را دائماً به شیوه‌های مختلف تغییر می‌دهند.

این تغییر مکرر باعث کاهش دقت کاربران می‌شود. همچنین احتمال از دست‌رفتن رایانامه‌ها و یا تارنماهای واقعی و قانونی به‌علت استفاده انبوه از این واژگان بالا است.

انتخاب ویژگی مرحله حساسی در بسیاری از مسائل طبقه‌بندی برای انتخاب زیرمجموعه بهینه از ویژگی‌ها است که سبب افزایش دقت یادگیری و کاهش زمان محاسباتی می‌شود [۶,۵]. در مجموعه‌داده‌ها همه ویژگی‌ها برای طبقه‌بندی مفید نیستند، ویژگی‌های وابسته و زائد هیچ اطلاعات اضافه درباره رده‌ها تهیه نمی‌کنند و اغلب تعدادی از ویژگی‌ها نوفه‌ای هستند؛ بنابراین روش‌های انتخاب ویژگی برای تعیین یک زیرمجموعه ویژگی بهینه استفاده می‌شوند؛ بنابراین شناسایی ویژگی‌هایی که وابستگی زیادی با خروجی دارند، مهم است. انتخاب ویژگی، ویژگی‌هایی از مجموعه‌داده‌ها که برای برای پیش‌گویی خروجی مؤثرتر هستند، انتخاب می‌کند [۷].

در این مقاله یک روش ترکیبی برمبنای الگوریتم جستجوی هارمونی [۸] و الگوریتم بهینه‌سازی مغناطیسی [۹] برای انتخاب ویژگی پیشنهاد می‌شود. الگوریتم جستجوی هارمونی [۸] به‌دلیل کاربردی‌بودن برای مسائل بهینه‌سازی گسسته و پیوسته، محاسبات ریاضیاتی کم، مفهوم ساده، پارامترهای کم و اجرای آسان به یکی از پرکاربردترین الگوریتم‌های بهینه‌سازی در سال‌های اخیر در مسائل مختلف تبدیل شده است. این الگوریتم در مقایسه با سایر روش‌های فراابتکاری، فرمول‌های ریاضی کمتری دارد و می‌توان آن را در مسائل مختلف مهندسی با تغییر در پارامترها و عملگرها تطبیق کرد. الگوریتم بهینه‌سازی مغناطیسی [۹]، یکی از الگوریتم‌های فراابتکاری است که برای حل مسائل بهینه‌سازی کاربرد دارد. این الگوریتم برای حل مسائل از خاصیت نیروی ذرات استفاده می‌کند. در این الگوریتم هر پاسخ به‌عنوان یک ذره در نظر گرفته می‌شود. حال ذره‌هایی که بهینه‌تر باشند، نیروی بیشتری دارند و می‌توانند ذرات دیگر را به سمت خود جذب کنند. ایده اصلی در این الگوریتم بر این پایه استوار است که در اطراف نقاط خوب ممکن است نقاط بهتری یافت شود. به همین دلیل نقاط ضعیف به سمت نقاط بهینه حرکت داده می‌شوند.

در روش ترکیبی برای ایجاد حافظه هارمونی هوشمند از الگوریتم بهینه‌سازی مغناطیسی استفاده می‌شود. بدین معنی، ویژگی‌هایی برای حافظه هارمونی انتخاب می‌شوند که فاصله بین آنها کمتر باشد. همچنین برای طبقه‌بندی نمونه‌ها از الگوریتم k نزدیک‌ترین همسایه استفاده می‌شود [۱۰]. طبقه‌بندی در دو مرحله آزمایش و آموزش انجام می‌گیرد. در مرحله آموزش، روش برمبنای یک‌سری قوانین از قبیل محاسبه فاصله اقلیدسی و تشخیص نمونه‌های مشابه به هم ساخته می‌شود و سپس نمونه‌های آموزش برمبنای رتبه‌بندی مرتب می‌شوند و نمونه‌های آزمایش برمبنای بیشینه تشابه به یکی از رده‌های غیرهرزنامه و هرزنامه نسبت داده می‌شوند. در این مقاله مجموعه‌داده Spambase [۱۱] با ۴۶۰۱ نمونه و ۵۷ ویژگی را مورد ارزیابی قرار می‌دهیم. مجموعه‌داده Spambase شامل دو رده هرزنامه با ۱۸۱۳ نمونه (۳۹/۴٪) و غیرهرزنامه با ۲۷۸۸ نمونه (۶۰/۶٪) است.

۲- کارهای قبلی

روش ترکیبی بهینه‌سازی اجتماع ذرات با شبکه عصبی مصنوعی چندلایه به‌منظور تشخیص ایمیل هرزنامه پیشنهاد شده است [۱۲]. از الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی‌ها و از شبکه عصبی مصنوعی چندلایه برای آموزش و آزمایش داده‌ها استفاده شده است. در روش ترکیبی بهینه‌سازی اجتماع ذرات با شبکه عصبی مصنوعی چندلایه از شبکه عصبی پرسپترون با تابع فعال‌سازی سیگموئید برای لایه پنهان و هشتاد درصد داده‌ها برای آموزش و بیست درصد برای آزمایش استفاده شده است. تعداد لایه‌های مخفی در روش شبکه عصبی مصنوعی چندلایه بین ۳-۱۵ لحاظ شده و تکرار الگوریتم بهینه‌سازی اجتماع ذرات برای انتخاب ویژگی برابر دویست است. ارزیابی برروی مجموعه‌داده Ling-Spam با ۲۵۸۹ رایانامه (۴۸۱ ایمیل هرزنامه و ۲۱۷۱ ایمیل غیرهرزنامه) و مجموعه‌داده Spam-Assassin با شش‌هزار نمونه رایانامه انجام شده است. ارزیابی برروی مجموعه‌داده Spam-Assassin و Ling-Spam نشان داده که مقدار صحت در روش ترکیبی بهینه‌سازی اجتماع ذرات با شبکه عصبی مصنوعی چندلایه به‌ترتیب برابر ۹۹/۹۸ و ۹۹/۷۹ است. مقایسه‌ها نشان داده که روش ترکیبی بهینه‌سازی اجتماع ذرات با شبکه عصبی مصنوعی چندلایه در مقایسه با روش‌های ماشین بردار پشتیبان با تابع کرنل، ماشین بردار پشتیبان با تابع شعاعی پایه و شبکه عصبی مصنوعی با تابع شعاعی پایه دقت تشخیص بهتری دارد. مهم‌ترین

مزیت روش ترکیبی این است که چندین بردار به‌عنوان بردارهای ویژگی انتخاب شده‌اند و برای هر اجرا یک بردار متفاوت به شبکه عصبی مصنوعی چندلایه داده می‌شود و برداری که بیشترین مقدار صحت را دارد انتخاب می‌شود.

روش ترکیبی بهینه‌سازی اجتماع ذرات با الگوریتم انتخاب منفی به‌منظور طبقه‌بندی ایمیل هرزنامه پیشنهاد شده است [۱۳]. ارزیابی برروی مجموعه‌داده Spambase انجام شده است. در روش ترکیبی از الگوریتم بهینه‌سازی اجتماع ذرات به‌منظور بهبوددادن الگوریتم انتخاب منفی استفاده شده است. ایجاد ماتریس اولیه برای تشخیص نمونه‌ها برمبنای سرعت ذرات تشکیل شده و برای تشابه بین نمونه‌ها از معیار فاصله تشابه استفاده شده است. نتایج نشان داده که روش بهینه‌سازی اجتماع ذرات با الگوریتم انتخاب منفی در مقایسه با روش‌های نیوی بیز، تمایز انتخاب ویژگی با ماشین بردار پشتیبان و الگوریتم انتخاب منفی دقت تشخیص بالاتر و در مقایسه با ماشین بردار پشتیبان دقت تشخیص کمتری دارد. مقدار صحت در روش بهینه‌سازی اجتماع ذرات با الگوریتم انتخاب منفی برابر ۸۳/۲۰ و در روش‌های نیوی بیز، تمایز انتخاب ویژگی با ماشین بردار پشتیبان و الگوریتم انتخاب منفی به‌ترتیب برابر ۷۸/۸، ۷۱ و ۶۸/۸۶ است. مهم‌ترین مزیت این روش این است که بهترین نمونه‌ها در هر تکرار ذخیره می‌شوند و با پایان تکرار برنامه می‌توان نمونه‌های ذخیره‌شده در بردار اولیه را برای طبقه‌بندی استفاده کرد.

روش ترکیبی الگوریتم تکامل تفاضلی و الگوریتم انتخاب منفی برای تشخیص ایمیل هرزنامه پیشنهاد شده است [۱۴]. مشکل اصلی در تشخیص ایمیل هرزنامه، ضعف در تشخیص فاصله بین ویژگی‌ها است. در روش ترکیبی از الگوریتم تکامل تفاضلی برای تشخیص فاصله بین ویژگی‌ها استفاده شده و از الگوریتم انتخاب منفی برای بهبوددادن تکامل تفاضلی استفاده شده است. نتایج برروی مجموعه‌داده Spambase با ۴۶۰۱ نمونه نشان داده که دقت تشخیص روش ترکیبی الگوریتم انتخاب منفی به‌ترتیب برابر با ۸۳/۰۶ و ۶۸/۸۶ درصد است.

روش ترکیبی برمبنای شبکه عصبی مصنوعی و درخت تصمیم است و برای تشخیص ایمیل هرزنامه پیشنهاد شده است [۱۵]. در این روش داده‌ها با استفاده از شبکه عصبی مصنوعی آموزش و آزمایش داده می‌شوند و با استفاده از درخت C4.5 طبقه‌بندی می‌شوند. از الگوریتم درخت C4.5 جهت تحلیل ویژگی‌های اصلی مؤثر بر ایمیل هرزنامه و تشخیص استفاده شده است. در درخت C4.5 هر مسیر از ریشه به سمت یک گره، نمایان‌گر یک قانون طبقه‌بندی است. ارزیابی برروی دو مجموعه‌داده Spam-Assassin و Corpus 2006 انجام شده است. نتایج برروی مجموعه داده Spam-Assassin نشان داده که مقدار صحت در روش ترکیبی برابر ۸۹/۱۵ و در روش‌های نیوی بیز، بهینه‌سازی کمینه ترتیبی و C4.5 به‌ترتیب برابر ۸۱/۰۸، ۸۸/۶۲ و ۷۳/۰۸ است. و همچنین برروی مجموعه داده Corpus مقدار صحت در روش ترکیبی برابر ۹۰/۸۷ و در روش‌های نیوی بیز، بهینه‌سازی کمینه ترتیبی و C4.5 به‌ترتیب برابر ۸۸/۱۵، ۸۹/۷۹ و ۸۸/۱۵ است.

روش‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی و ترکیب آنها بر روی دو مجموعه داده با چهارده ویژگی آزمایش و اجر شده‌اند [۱۶]. مجموعه‌داده اولی شامل ۵۰۴ رایانامه (۳۳۶ غیرهرزنامه و ۱۶۸ هرزنامه) و مجموعه‌داده دومی شامل ۶۵۷ رایانامه (۳۸۷ غیرهرزنامه و ۲۷۰ هرزنامه) است. در درخت تصمیم‌گیری از آنتروپی، ماشین بردار پشتیبان از تابع کرنل و شبکه عصبی مصنوعی از خطای میانگین استفاده شده است. نتایج بر روی مجموعه داده اولی نشان داده که مقدار صحت در روش ترکیبی برابر ۹۱/۱۷ و در روش‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی به‌ترتیب برابر ۸۹/۸۸، ۸۸/۶۹ و ۸۹/۸۸ است. و برروی مجموعه‌داده دومی درصد صحت در روش ترکیبی برابر ۹۱/۷۸ و در روش‌های درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه عصبی مصنوعی به‌ترتیب برابر ۹۰/۸۷، ۹۰/۸۷ و ۸۹/۱۴ است.

الگوریتم سامانهٔ ایمنی مصنوعی برای تشخیص ایمیل هرزنامه برمبنای انتخاب ویژگی‌های مهم پیشنهاد شده است [۱۷]. روال روش این‌گونه است که در ابتدا عملیات پیش‌پردازش انجام می‌شود و واژگان مهم در متن ایمیل شناسایی و تکرار آنها شمارش و همچنین نمادها و واژگان مشکوک شناسایی و برای طبقه‌بندی از آنها استفاده می‌شود. ارزیابی برروی مجموعه‌داده TREC07 با ۷۵۴۱۹ رایانامه (۲۵۲۲۰ غیرهرزنامه و ۵۰۱۹۹ هرزنامه) نشان داده که دقت تشخیص برابر ٪۸۶/۲۰ است.

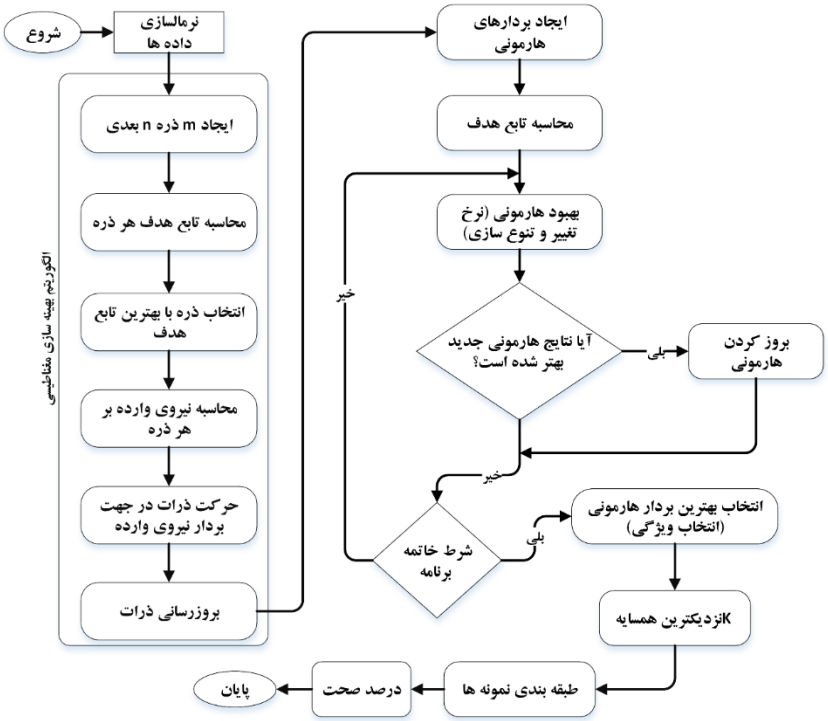
طبقه‌بندی بیزین [۱۸] برای تشخیص ایمیل هرزنامه برروی سه مجموعه‌داده که شامل ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ رایانامه هستند، اجرا و آزمایش شده است. در طبقه‌بندی بیزین از روابط احتمالاتی استفاده می‌شود. نتایج نشان داده دقت تشخیص برای سه روش بیشتر از

۹۳درصد است. برای ارزیابی از معیارهای صحت، دقت و بازخوانی استفاده شده است. مقدار صحت به‌ترتیب برای سه مجموعه‌داده برابر ۹۳/۹۸، ۹۴/۸۹ و ۹۶/۴۶ است.

روش ترکیبی شبکه عصبی مصنوعی با الگوریتم انتخاب منفی [۱۹] به‌منظور انتخاب ویژگی و طبقه‌بندی ایمیل هرزنامه پیشنهاد شده است. در روش ترکیبی شبکه عصبی مصنوعی با الگوریتم انتخاب منفی از الگوریتم انتخاب منفی برای انتخاب ویژگی و از شبکه عصبی مصنوعی برای طبقه‌بندی نمونه‌ها استفاده شده است. الگوریتم انتخاب منفی یکی از الگوریتم‌های انتخاب ویژگی است که از بردارهای اولیه و وزندهی برای انتخاب ویژگی‌ها استفاده می‌کند. روش شبکه عصبی مصنوعی که مهم‌ترین نوع آن شبکه عصبی مصنوعی چندلایه است برای طبقه‌بندی و کاهش خطا دقت بالایی دارد. نتایج برروی ۴۶۰۱ نمونه ایمیل هرزنامه نشان داده که دقت روش ماشین بردار پشتیبان با الگوریتم انتخاب منفی برابر ۹۴/۳۰ درصد است که در مقایسه با ماشین بردار پشتیبان دقت بیشتری دارد.

۳-روش پیشنهادی

در این بخش، مراحل روش پیشنهادی که ترکیبی از الگوریتم جستجوی هارمونی با الگوریتم بهینه‌سازی مغناطیسی است توضیح داده شده است. در روش پیشنهادی هدف این است که با استفاده از الگوریتم بهینه‌سازی مغناطیسی بهترین ویژگی‌ها برای ماتریس حافظه هارمونی کشف شوند. در هر تکرار بهترین ویژگی‌ها در فضای مسأله برمبنای نزدیکی پیدا می‌شوند و وارد حافظه هارمونی می‌شوند. تعداد ویژگی‌های وارد شده به الگوریتم جستجوی هارمونی به منزله جمعیت اولیه هستند. دلیل انجام این کار این است که در هر تکرار حافظه هارمونی شامل بهترین ویژگی‌ها باشد. در روش پیشنهادی، نسل اولیه از بردارهای هارمونی برمبنای الگوریتم بهینه‌سازی مغناطیسی پر می‌شوند. در مرحله دوم، ماتریس حافظه هارمونی با استفاده از لحاظ‌کردن قواعد بازبینی حافظه، تطبیق گام و ایجاد نسل تصادفی از راه‌حل‌های موجود در حافظه الگوریتم ایجاد می‌شود. در مرحله سوم، اگر هارمونی‌های جدید، از بدترین هارمونی‌های موجود در حافظه بهتر باشد، جایگزین آنها و بدترین هارمونی‌های موجود در حافظه حذف می‌شوند؛ لذا در روش پیشنهادی برای دستیابی به بهترین هارمونی‌ها از الگوریتم بهینه‌سازی مغناطیسی استفاده می‌شود. همچنین الگوریتم جستجوی هارمونی، هارمونی‌های ارزیابی‌شده در مراحل قبلی را ذخیره می‌کند و اگر بعضی از ویژگی‌ها باعث کاهش دقت تشخیص شوند عمل جابه‌جایی را انجام می‌دهد و از آنها به‌عنوان ویژگی جایگزین استفاده می‌کند. در شکل (۱)، روندنمای روش پیشنهادی نشان داده شده است.



(شکل-۱): روندنمای روش پیشنهادی

در الگوریتم بهینه‌سازی مغناطیسی، سرعت و موقعیت ذرات براساس نقاط بهینه محلی و سراسری به‌روز می‌شوند. در هر بار تکرار، تمامی ذرات در فضای nبعدی مسأله حرکت می‌کنند تا نقطه بهینه سراسری پیدا شود. امتیاز الگوریتم بهینه‌سازی مغناطیسی در پیاده‌سازی ساده و تعداد کم پارامترهای مورد نیاز است. در الگوریتم جستجوی هارمونی توسط حافظه هارمونی، تمام بردارهای راه حل جاری (مجموعه متغیرهای تصمیم‌گیری) ذخیره می‌شوند. در هر دور تکامل الگوریتم جستجوی هارمونی، اگر برداری تولید شود که نسبت به بردارهای موجود در حافظه برازنده‌تر باشد، در حافظه ذخیره می‌شود و ممکن است در نسل‌های بعدی مورد استفاده قرار گیرد. با آغاز الگوریتم جستجوی هارمونی، حافظه هارمونی با ترکیب‌هایی که به‌صورت تصادفی تولید می‌شوند، مقداردهی اولیه می‌شود. مقداردهی اولیه حافظه هارمونی با ترکیب‌های تصادفی احتمال گیرافتادن در بهینه محلی را افزایش می‌دهد. در روش پیشنهادی هدف الگوریتم بهینه‌سازی مغناطیسی این است که از تولیدهای مبنی بر تولید تصادفی در الگوریتم جستجوی هارمونی جلوگیری شود و بردارهایی برازنده در مرحله نخست به‌وسیلهٔ الگوریتم بهینه‌سازی مغناطیسی تولید شوند. در طول مراحل الگوریتم جستجوی هارمونی، حافظه هارمونی با ترکیب‌هایی که برازندگی بهتری دارند، پر و در انتهای مراحل تکامل، بهترین ترکیب به‌عنوان ترکیب انتخاب‌شده نهایی معرفی می‌شود. اگر چندین ترکیب با برازندگی بهینه یافت شوند، راه حل نهایی به‌طور تصادفی از میان آنها انتخاب می‌شود. این حالت، همانند زمانی است که تابع هدف یک مسأله بهینه‌سازی، دارای چندین نقطه بهینه سراسری باشد.

۳-۱- نرمال‌سازی داده‌ها

مهمترین عملی که باید در ابتدای روش پیشنهادی انجام شود، نرمال‌سازی است. از آنجایی‌که ویژگی‌های مجموعه داده‌ها دارای مقادیر مختلفی هستند، به‌منظور یکسان‌سازی، مقدار ویژگی‌ها به فاصله بین صفر و یک تبدیل می‌شوند. در فرایند نرمال‌سازی، کوچک‌ترین عدد مربوط به هر ویژگی با min و بزرگ‌ترین عدد با max نشان داده می‌شود. و x مقدار هر ویژگی است. عمل نرمال‌سازی طبق معادله (۱)، انجام می‌شود [۲۰]. مقدار x_{max} و x_{min} به‌ترتیب برابر یک و صفر می‌باشند.

$$y = \frac{x - \min}{\max - \min} * [x_{\max} - x_{\min}] + x_{\min} \qquad (۱)$$

در روش پیشنهادی به‌منظور بازنمایی داده‌ها در ابتدا در فضای جستجو عامل‌ها برمبنای مقدار صفر و یک مشخص می‌شوند. عامل‌هایی که فقط مقدار یک دارند از مجموعه داده‌ها خوانده، عملیات طبقه‌بندی برروی آنها انجام و بازنمایی اطلاعات به‌صورت انتخابی انجام می‌شود.

۳-۲- الگوریتم بهینه‌سازی مغناطیسی: ایجاد جمعیت اولیه

در الگوریتم بهینه‌سازی مغناطیسی، تمام ذرات مغناطیسی در ساختاری شبیه به شبکه توری نشان داده می‌شوند و موقعیت اولیه و سرعت آنها برابر با صفر است. در فضای مسأله، ذرات مغناطیسی می‌توانند نیرو را تنها به همسایگان خود اعمال کنند، نیرویی که توسط همسایه X_{uv}^t به ذره $X_{ij,k}^t$ اعمال می‌شود، طبق معادله (۲) تعریف می‌شود. در معادله (۲)، $D(X_{uv}^t, X_{ij,k}^t)$ فاصله بین ذره $X_{ij,k}^t$ و همسایه‌اش X_{uv}^t است. مقدار هر ذره در جمعیت در میدان مغناطیسی B_{ij}^t ذخیره می‌شود. ذرات در فضای جستجو به‌دنبال بهترین نقاط بهینه سراسری هستند. ذره‌ای که بهترین مقدار را داشته باشد، به‌عنوان ذره سراسری در بین جمعیت شناخته می‌شود و بقیه ذرات به‌دنبال بهترین نقطه به‌دست‌آمده توسط ذره سراسری هستند. مقدار min و max بیان‌گر کمترین و بیشترین نقاط پیداشده به‌وسیلهٔ ذره iام هستند.

$$F_{ij,k}^t = F_{ij,k}^t + \frac{(X_{uv}^t - X_{ij,k}^t) * B_{ij}^t}{D(X_{uv}^t, X_{ij,k}^t)} \qquad (۲)$$

$$where \quad B_{ij}^t = \frac{B_{ij}^t - \min}{B_{ij}^t - \max}$$

در معادله(۲)، $D(X_{uv}^t, X_{ij,k}^t)$ فاصله بین ذره $X_{ij,k}^t$ و همسایه‌اش X_{uv}^t است و طبق معادله (۳) محاسبه می‌شود. در معادله (۳)، X_{uv}^t و $X_{ij,k}^t$ موقعیت ذرات هستند و k ابعاد مسأله است. پارامترهای Uk و Lk کران بالا و پایین جمعیت هستند.

$$D(X_{uv}^t, X_{ij}^t) = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{X_{ij,k}^t - X_{uv}^t}{U_k - L_k}\right| \qquad (۳)$$

طبق معادله (۳) ذراتی که فاصله نزدیک‌تری دارند، به‌عنوان ذرات برازنده انتخاب و وارد الگوریتم جستجوی

هارمونی می‌شوند و مرحله بهینه‌سازی و کشف انتخاب ویژگی شروع می‌شود.

۳-۳- الگوریتم جستجوی هارمونی: انتخاب ویژگی

در الگوریتم جستجوی هارمونی، جهت حفظ بهترین راه‌حل‌های پیشین، از "حافظه هارمونی" استفاده می‌شود. ابعاد ماتریس حافظه هارمونی برمبنای N × HMS تعیین می‌شود که N برابر با تعداد نمونه‌ها یا متغیرهای مسأله مورد نظر است و مقدار HMS برمبنای مسأله تعیین می‌شود. ابعاد شامل مجموعه از بردارهای راه حل است. حافظه هارمونی درحقیقت یک ماتریس است که در بردارنده‌ی هارمونی‌ها است. هر هارمونی یک بردار است که دربردارندهٔ متغیرهای (ویژگی‌ها) تصمیم‌گیری مسأله بهینه‌سازی است. در گام آغازین الگوریتم می‌توان به تعداد دلخواه هارمونی‌های اولیه ایجاد کرده و در حافظهٔ هارمونی ذخیره کرد. ماتریس حافظه هارمونی طبق معادله (۴)، با ویژگی‌های تولیدشده به‌وسیلهٔ الگوریتم بهینه‌سازی مغناطیسی پر می‌شود.

$$HM=\begin{bmatrix} x_1(1) & x_1(2) & \dots & x_1(n-1) & x_1(n) \\ x_2(1) & x_2(2) & \dots & x_2(n-1) & x_2(n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{HMS}(1) & x_{HMS}(2) & \dots & x_{HMS}(n-1) & x_{HMS}(n) \end{bmatrix} \quad (۴)$$

فرض کنید $T=\{T1,T2,\dots,Tn\}$ مجموعه ویژگی‌ها باشند که هدف، انتخاب یک زیرمجموعه بهینه از ویژگی‌ها است. هدف اصلی این مقاله، تشکیل ماتریس ویژگی‌های بهینه است که بتواند شرایط حاکم بر هرزنامه و غیرهرزنامه را به‌خوبی تفکیک کند. بنابراین، در این مقاله از الگوریتم جستجوی هارمونی برای انتخاب ویژگی‌های بهینه استفاده می‌شود تا دقت طبقه‌بندی به بیشینه برسد.

مطابق با مرحله نخست الگوریتم جستجوی هارمونی، یک نسل اولیه از بردارهای هارمونی به طور تصادفی ایجاد و در حافظه هارمونی ذخیره می‌شود. در مرحله دوم یک بردار هارمونی جدید (حل جدید) با استفاده از لحاظ‌کردن حافظه، تطبیق گام و ایجاد نسل تصادفی از راه حل‌های موجود در حافظه الگوریتم ایجاد می‌شود. در مرحله سوم، اگر هارمونی جدید، از بدترین هارمونی موجود در حافظه بهتر باشد، جایگزین آن و بدترین هارمونی موجود در حافظه حذف می‌شود. در مرحله چهارم، شرط پایان اجرای الگوریتم بررسی می‌شود که در الگوریتم جستجوی هارمونی بررسی تعداد تکرار است. در صورت عدم برقراری شرط پایان، مراحل ۳ و ۴ مجدداً تکرار می‌شود. البته می‌توان شرط پایان‌یافتن الگوریتم را برای مقدار بهینه مشخصی تنظیم کرد و تا پایان آن مقدار، مراحل الگوریتم تکرار شود.

جمعیت اولیه در الگوریتم جستجوی هارمونی شامل مقادیری است که درقبل به‌وسیلهٔ الگوریتم بهینه‌سازی مغناطیسی تولید شده‌اند و در الگوریتم جستجوی هارمونی در حالت پیوسته قرار دارند و باید به حالت گسسته بروند. با به‌روزرسانی نقاط فضای جستجو، عمل یافتن بهترین ویژگی‌ها انجام می‌شود. برای انتخاب ویژگی‌ها به‌وسیلهٔ الگوریتم جستجوی هارمونی باید فضای مسأله از پیوسته به گسسته تبدیل شود. در نسخه دودویی، موقعیت هر هارمونی باید تغییر یابد و به حالت [۱،۰] نگاشت داده شود که این مقدار بیان‌گر احتمال صفر یا یک‌بودن x است؛ لذا، در نسخه ددودیی، موقعیت هر هارمونی در هر بعد به دو مقدار صفر و یک تغییر پیدا می‌کند. یعنی هر هارمونی، در یک فضا محدود به صفر و یک حرکت می‌کند. برای پیاده‌سازی روش دودویی، ابتدا موقعیت هر هارمونی در هر بُعد، محاسبه، سپس، این مقدار با استفاده از تابع محدود کننده سیگموید به مقداری بین صفر و یک نگاشت می‌شود. تابع سیگموئید، برای تعیین مقادیر موقعیت به مقادیر احتمالی برای به روزسانی موقعیت‌ها ضروری است. در بیش‌تر موارد برای تبدیل فضای پیوسته به دودویی از تابع سیگموئید استفاده می‌شود؛ لذا تغییرات الگوریتم جستجوی هارمونی طبق معادله (۵)، انجام می‌شود. تابع rand اعداد تصادفی در بازه [۱،۰] تولید می‌کند.

$$S(X_{ij})=\frac{1}{1+e^{-x_{ij}}} \quad (۵)$$

$$if\ rand < S(X_{ij})\ then\ x_{ij}=1\ else\ x_{ij}=0 \quad (۶)$$

۴-۳- تابع برازندگی

در روش پیشنهادی، ویژگی‌هایی انتخاب می‌شوند که بهترین برازندگی را دارند. تابع برازندگی برای انتخاب ویژگی‌ها طبق معادله (۷) تعریف می‌شود. در معادله (۷)، |n| تعداد کل ویژگی‌ها و |S| تعداد ویژگی‌های انتخاب شده است. پارامتر accuracy درصد صحت و مقدار

پارامترهای δ و ρ ثابت هستند و مقدار آنها به‌ترتیب برابر با ۹۹ و ۱ است.

$$Fitness = \delta.Accuracy + \rho.\frac{|n| - |S|}{|n|} \quad (۷)$$

۳-۵- طبقه‌بندی با k نزدیکترین همسایه

در الگوریتم k نزدیک‌ترین همسایه، مهم‌ترین معیار برای تشخیص نمونه‌های مشابه استفاده از معیار فاصله است. اگر یک بردار ویژگی به‌صورت $< z_1(x), z_2(x), ..., z_n(x) >$ تعریف شود برای تشخیص همسایه‌های آن از فاصله اقلیدسی طبق معادله (۸) به‌منظور به‌دست‌آوردن فاصله بین دو ویژگی x_i و x_j استفاده می‌شود.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (z_k(x_i) - z_k(x_j))^2} \quad (۸)$$

برمبنای الگوریتم k نزدیک‌ترین همسایه در ابتدا نمونه‌های آموزشی، ارزیابی و یک روش به منظور تشابه بین نمونه‌ها ایجاد می‌شود. نمونه‌ها برمبنای نزدیکی فاصله و تشابه به دسته‌ها (دسته هرزنامه و غیرهرزنامه) اختصاص داده می‌شوند. در ابتدا باید نمونه‌ای برای طبقه‌بندی انتخاب شود که در همسایگی‌اش بیشترین تعداد نمونه منتسب برای آن وجود داشته باشد 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 اکثریت است به نمونه ناشناخته داده می‌شود.

برای تشخیص ایمیل‌های هرزنامه در روش پیشنهادی از معیار درجه تشابه استفاده شده، که خود معیاری برای اندازه‌گیری فاصله بین نمونه‌های رایانامه است.

۳-۶- معیارهای ارزیابی

در این مقاله، جهت ارزیابی کارایی طبقه‌بندی از چهار معیار دقت، بازخوانی، اندازه‌گیری و صحت استفاده شده است. دقت نشان دهنده تعداد نمونه‌های درست به تعداد کل نمونه‌ها است. معیار بازخوانی نشان می‌دهد که اگر نمونه‌ای، نوع A تشخیص داده شد با چه احتمالی نوع A است. دقت و بازخوانی طبق معادله (۹) و (۱۰) محاسبه می‌شوند [۲۱].

شماره	توضیحات	معادلات
(۹)	دقت	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100$
(۱۰)	بازخوانی	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100$
(۱۱)	F-Measure	$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
(۱۲)	صحت	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$

پارامترهای درست مثبت(TP)، درست منفی(TN)، کاذب مثبت(FP)، کاذب منفی(FN) از پارمترهای اصلی برای معیارهای ارزیابی هستند. پارامتر درست مثبت، بیان‌گر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم طبقه‌بندی نیز دسته آن‌ها را به‌درستی مثبت تشخیص داده است. پارامتر درست منفی، بیان‌گر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است. پارامتر کاذب مثبت، بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم طبقه‌بندی دسته آن‌ها را به اشتباه مثبت تشخیص داده است. پارامتر کاذب منفی، بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم طبقه‌بندی دسته آن‌ها را به‌اشتباه منفی تشخیص داده است.

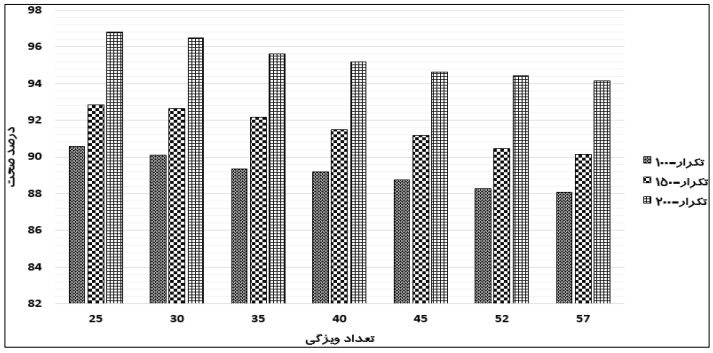
۴- ارزیابی و نتایج

ارزیابی و نتایج روش پیشنهادی برروی مجموعه‌داده Spambase با ۴۶۰۱ نمونه و ۵۷ ویژگی که شامل ۳۹/۴ درصد ایمیل هرزنامه و ۶۰/۶ درصد ایمیل غیرهرزنامه است در محیط برنامه‌نویسی VC#.NET 2017 انجام شده است. بیشینه تعداد تکرار در روش پیشنهادی برابر ۲۰۰ است و مقدار k در الگوریتم k نزدیکترین همسایه برابر سه است. همچنین تعداد حافظه هارمونی برای اجرای اولیه برابر با صد است. در روش پیشنهادی هفتاد درصد از داده‌ها برای قسمت آموزش روش و سی درصد از داده‌ها برای آزمون روش در نظر گرفته شده است. در ابتدا روش پیشنهادی را با انتخاب ویژگی‌های مختلف ارزیابی کردیم و با هر تعداد ویژگی نتایج متفاوتی حاصل شده است. در جدول (۱)، ارزیابی روش پیشنهادی با انتخاب ویژگی‌ها مختلف برمبنای معیارهای مختلف نشان داده شده است.

(جدول-۱): ارزیابی روش پیشنهادی برمبنای انتخاب ویژگی‌ها و تکرارهای مختلف

تعداد تکرار	تعداد ویژگی	دقت	بازخوانی	F-measure	صحت
۱۰۰	۲۵	۸۹/۲۵	۸۹/۶۲	۸۹/۴۳	۹۰/۵۸
	۳۰	۸۹/۰۳	۸۹/۱۶	۸۹/۰۹	۹۰/۱۲
	۳۵	۸۸/۹۲	۸۹/۰۸	۸۹/۰۰	۸۹/۳۵
	۴۰	۸۸/۵۸	۸۸/۹۵	۸۸/۷۶	۸۹/۲۱
	۴۵	۸۸/۴۹	۸۸/۷۴	۸۸/۶۱	۸۸/۷۴
	۵۲	۸۷/۵۶	۸۸/۱۵	۸۷/۸۵	۸۸/۲۹
	۵۷	۸۷/۲۳	۸۸/۰۹	۸۷/۶۶	۸۸/۰۷
۱۵۰	۲۵	۹۱/۲۵	۹۱/۶۸	۹۱/۴۶	۹۲/۸۴
	۳۰	۹۱/۰۳	۹۱/۴۶	۹۱/۲۴	۹۲/۶۵
	۳۵	۹۱/۹۲	۹۱/۹۵	۹۱/۹۳	۹۲/۱۵
	۴۰	۹۰/۵۸	۹۰/۷۸	۹۰/۶۸	۹۱/۴۸
	۴۵	۹۰/۴۹	۹۰/۹۳	۹۰/۷۱	۹۱/۱۶
	۵۲	۸۹/۵۶	۹۰/۱۹	۸۹/۸۷	۹۰/۴۵
	۵۷	۸۹/۲۲	۸۹/۶۱	۸۹/۴۱	۹۰/۱۶
	۲۵	۹۵/۴۸	۹۵/۸۳	۹۵/۶۵	۹۶/۸۲
۲۰۰	۳۰	۹۵/۲۱	۹۵/۴۷	۹۵/۳۴	۹۶/۵۱
	۳۵	۹۵/۴۲	۹۵/۶۹	۹۵/۵۵	۹۵/۶۰
	۴۰	۹۴/۸۵	۹۵/۰۷	۹۴/۹۶	۹۵/۱۷
	۴۵	۹۴/۷۴	۹۵/۱۴	۹۴/۹۴	۹۴/۶۲
	۵۲	۹۳/۴۶	۹۳/۸۶	۹۳/۶۶	۹۴/۴۲
	۵۷	۹۳/۲۵	۹۳/۶۸	۹۳/۴۶	۹۴/۱۷

در شکل (۲)، نتایج روش پیشنهادی برمبنای انتخاب ویژگی و تکرارهای مختلف نشان داده شده است. شکل (۲) نشان می‌دهد که درصد صحت با دویست بار تکرار و ۲۵ ویژگی در مقایسه با حالت‌های دیگر بیشتر است. اگر تعداد ویژگی‌ها کمتر باشد، درصد صحت بیشتر است. به‌دلیل این‌که برای طبقه‌بندی از ویژگی‌های تأثیرگذار استفاده و هر نمونه به دسته‌ای اختصاص داده می‌شود که متعلق به نمونه‌ها آن دسته باشد و برمبنای نزدیکی، تشابه بیشتری داشته باشد؛ لذا ایجاد قوانین برای یافتن نمونه‌های مشابه برمبنای ویژگی‌های اصلی انجام می‌شود.



(شکل-۲): نتایج روش پیشنهادی برمبنای انتخاب ویژگی و تکرارهای مختلف

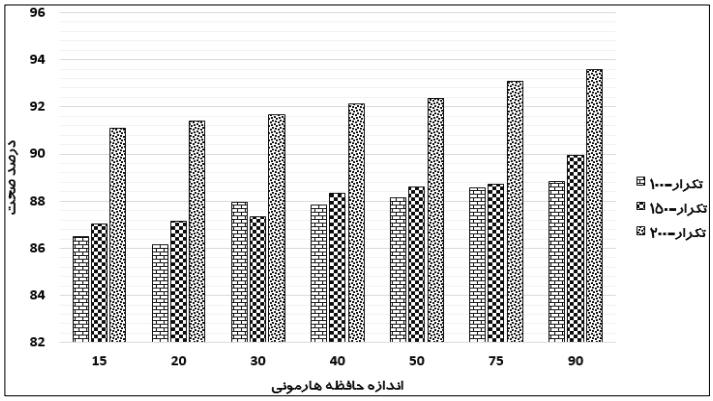
در جدول (۲)، روش پیشنهادی برمبنای تعداد تکرار و اندازه حافظه هارمونی نشان داده شده است. همان‌طورکه مشاهده می‌کنید اگر اندازه حافظه هارمونی افزایش یابد، مقدار صحت هم افزایش می‌یابد. اگر تعداد هارمونی‌ها زیاد باشد، برای انتخاب ویژگی‌هایی با برازندگی بهتر، شانس بیشتری وجود دارد و به عبارتی فضای انتخاب ویژگی گسترده‌تر خواهد بود.

(جدول-۲): ارزیابی روش پیشنهادی برمبنای تکرارهای مختلف و اندازه حافظه هارمونی

تعداد تکرار	اندازه حافظه هارمونی	دقت	بازخوانی	F-measure	صحت
۱۰۰	۱۵	۸۷/۱۲	۸۷/۴۹	۸۷/۳۰	۸۶/۴۹
	۲۰	۸۷/۶۳	۸۹/۰۵	۸۸/۳۳	۸۶/۱۵
	۳۰	۸۸/۲۲	۸۸/۹۴	۸۸/۵۸	۸۷/۹۴
	۴۰	۸۸/۷۶	۸۹/۳۶	۸۹/۰۳	۸۷/۸۲
	۵۰	۸۹/۱۱	۸۹/۸۴	۸۹/۴۷	۸۸/۱۶
	۷۵	۸۹/۶۲	۹۰/۱۶	۸۹/۸۹	۸۸/۵۵
	۹۰	۹۰/۳۶	۹۰/۹۵	۹۰/۶۵	۸۸/۸۵
	۱۵	۸۷/۴۸	۸۷/۸۱	۸۷/۶۴	۸۷/۰۵
	۲۰	۸۷/۹۴	۸۹/۱۳	۸۸/۵۳	۸۷/۱۶
	۳۰	۸۸/۱۶	۸۸/۹۴	۸۸/۵۵	۸۷/۳۵

۱۵۰	۴۰	۸۸/۳۸	۸۸/۹۶	۸۸/۶۷	۸۸/۳۴
	۵۰	۸۹/۱۷	۹۰/۴۷	۸۹/۸۲	۸۸/۶۱
	۷۵	۸۹/۶۴	۹۰/۳۶	۹۰/۰۰	۸۸/۷۲
	۹۰	۹۰/۳۰	۹۱/۶۵	۹۰/۹۷	۸۹/۹۴
۲۰۰	۱۵	۹۱/۱۶	۹۱/۴۸	۹۱/۳۲	۹۱/۰۸
	۲۰	۹۱/۷۸	۹۲/۰۶	۹۱/۹۲	۹۱/۴۱
	۳۰	۹۲/۳۶	۹۲/۸۳	۹۲/۵۹	۹۱/۶۸
	۴۰	۹۲/۴۶	۹۲/۷۱	۹۲/۵۸	۹۲/۱۳
	۵۰	۹۳/۱۸	۹۳/۴۵	۹۳/۳۱	۹۲/۳۶
	۷۵	۹۳/۶۴	۹۳/۹۰	۹۳/۷۷	۹۳/۰۷
	۹۰	۹۴/۰۵	۹۴/۲۱	۹۴/۱۳	۹۳/۵۹

در شکل (۳)، نمودار مقایسه معیار صحت برمبنای اندازه حافظه هارمونی و تعداد تکرار نشان داده شده است. شکل (۳)، نشان می‌دهد که درصد صحت با دویست بار تکرار و نود هارمونی در مقایسه با حالت‌های دیگر بیشتر است. همچنین درصد صحت با ۱۰۰ و ۱۵۰ بار تکرار با ۹۰ هارمونی، بیشتر است. درصد صحت با پانزده هارمونی در مقایسه با حالت‌های دیگر، کمتر است. در تمامی تکرارها، درصد صحت با پانزده هارمونی در حالت بهینه نیست و می‌توان نتیجه گرفت که اگر تعداد هارمونی‌ها بیشتر باشد، دقت تشخیص و طبقه‌بندی افزایش می‌یابد.



(شکل-۳): نمودار مقایسه معیار صحت برمبنای اندازه حافظه هارمونی و تعداد تکرار

۴-۱- مقایسه و ارزیابی

جدول (۳)، مقایسه روش پیشنهادی با روش‌های دیگر را نشان می‌دهد. روش پیشنهادی در مقایسه با ترکیب الگوریتم علف‌های هرز مهاجم با k نزدیک‌ترین همسایه [۲] و ترکیب الگوریتم بهینه‌سازی کلونی مورچه با کرم شب‌تاب [۴] درصد صحت بیشتری دارد. به‌دلیل اینکه کشف راه حل‌ها، نزدیک‌شدن به راه حل بهینه و همچنین انتخاب ویژگی‌ها به نوع الگوریتم‌های فراابتکاری بستگی دارد در نتیجه در مقاله‌های ارائه شده (ترکیب الگوریتم علف‌های هرز مهاجم با k نزدیک‌ترین همسایه [۲] ترکیب الگوریتم بهینه‌سازی کلونی مورچه با کرم شب‌تاب [۴]) درصد صحت با این مقاله متفاوت است.

به‌منظور مقایسه و ارزیابی روش پیشنهادی با روش‌های دیگر در ابتدا باید فاکتورهایی مانند تعداد جمعیت، تعداد تکرار و نرخ عملگرهای بهینه‌سازی در هر الگوریتم تنظیم شود. در این مقاله به‌منظور یک مقایسه عادلانه تمامی فاکتورها در ابتدا تنظیم و سپس مراحل آزمایش انجام شده است.

در الگوریتم‌های فراابتکاری هر الگوریتم یک روش توازن بین کاوش و بهره‌برداری دارد. در الگوریتم کرم شب‌تاب تغییر چگونگی حرکت کرم‌ها برای یافتن راه حل بهینه به مقدار بهینه سراسری و محلی بستگی دارد. کرم شب‌تاب درخشنده‌تر سایر کرم‌های شب‌تاب که در همسایگی‌اش است را به سمت خود جذب می‌کند و اگر هیچ کدام درخشان‌تر از دیگری نبودند، حرکت آنها به‌صورت تصادفی انجام می‌شود.

در [۱۳] درصد صحت در روش ترکیبی بهینه‌سازی اجتماع ذرات با الگوریتم انتخاب منفی، نیوی بیز، ماشین بردار پشتیبان و الگوریتم انتخاب منفی به‌ترتیب برابر با ۸۳/۲۰، ۷۸/۸۰، ۷۱/۰۰، ۶۸/۸۶ است. روش پیشنهادی

در مقایسه با بیش‌تر روش‌ها درصد صحت بیشتری دارد؛ زیرا در انتخاب ویژگی از روش انتخاب بهترین ذره بهره گرفته است و در هر مرحله روش ترکیبی به‌روزرسانی می‌شود و ماتریس هارمونی به‌دنبال ایجاد بهترین بردارهای راه حل بوده است. روش پیشنهادی در مقایسه با الگوریتم جستجوی گرانشی دودویی پیشرفته [۲۴] و بهینه‌سازی اجتماع ذرات دودویی [۲۴] درصد صحت بیشتری دارد. ترکیب شبکه عصبی مصنوعی با الگوریتم انتخاب منفی [۱۹] دارای درصد صحت ۹۴/۳۰ است. همچنین درصد صحت شبکه عصبی مصنوعی چند لایه و جنگل تصادفی به‌ترتیب برابر با ۹۳/۲۸ و ۹۳/۸۹ است [۲۳]. طبق مقایسه‌ها می‌توان نتیجه گرفت که درصد صحت روش پیشنهادی در مقایسه با تکنیک‌های داده‌کاوی و الگوریتم‌های فرا ابتکاری بیشتر است.

همچنین به‌منظور نشان‌دادن کارایی روش پیشنهادی از سه مجموعه‌داده مختلف (LingSpam، SpamAssassin و Enron) برای مقایسه استفاده شده است. مجموعه‌داده LingSpam [۲۵] شامل ۴۸۱ نمونه ایمیل هرزنامه و ۲۱۷۱ نمونه ایمیل غیرهرزنامه است. مجموعه داده SpamAssassin [۲۵] شامل شش‌هزار نمونه ایمیل هرزنامه و غیرهرزنامه است. مجموعه داده Enron [۲۷] شامل ۱۵۰۰ نمونه اسپم و ۳۶۷۲ نمونه غیراسپم است. روش پیشنهادی برروی مجموعه‌داده LingSpam و Enron نتایج بهتری داشته است و در مقایسه با روش‌های دیگر توانسته که درصد صحت بیشتری داشته باشد.

(جدول-۳): مقایسه روش پیشنهادی با روش‌های دیگر برمبنای درصد صحت

مجموعه داده	رفرنس	روش‌ها	درصد صحت
Spambase	[۲]	ترکیب الگوریتم علف‌های هرز مهاجم با k نزدیک‌ترین همسایه	۹۱/۱۱
	[۴]	ترکیب الگوریتم بهینه‌سازی کلونی مورچه با کرم شب‌تاب	۹۲/۰۵
	[۱۳]	ترکیبی بهینه‌سازی اجتماع ذرات با الگوریتم انتخاب منفی	۸۳/۲۰
		نیوی بیز	۷۸/۸۰
		ماشین بردار پشتیبان	۷۱/۰۰
		الگوریتم انتخاب منفی	۶۸/۸۶
	[۱۴]	ترکیب الگوریتم تکامل تفاضلی و الگوریتم انتخاب منفی	۸۳/۰۶
	[۱۹]	ترکیب شبکه عصبی مصنوعی با الگوریتم انتخاب منفی	۹۴/۳۰
	[۲۲]	ترکیب الگوریتم جستجوی هارمونی با درخت تصمیم‌گیری	۹۵/۲۵
	[۲۳]	شبکه عصبی مصنوعی چندلایه	۹۳/۲۸
		جنگل تصادفی	۹۳/۸۹
	[۲۴]	الگوریتم جستجوی گرانشی دودویی پیشرفته	۹۲/۲۰
		بهینه‌سازی اجتماع ذرات دودویی	۹۰/۰۰
		الگوریتم ژنتیک	۹۰/۶۰
	-	**روش پیشنهادی**	۹۴/۱۷
LingSpam	[۱۲]	ماشین بردار پشتیبان تابع شعاعی پایه	۹۹/۷۹
	[۲۸]	K نزدیکترین همسایه	۹۵/۰۴
		K نزدیکترین همسایه وزنی	۹۵/۰۴
		درخت تصمیم‌گیری	۹۵/۰۴
	-	**روش پیشنهادی**	۹۹/۴۳
SpamAssassin	[۱۲]	ماشین بردار پشتیبان تابع شعاعی پایه	۹۹/۹۸
	[۲۹]	الگوریتم بهینه‌سازی شیرمورچه با k نزدیکترین همسایه	۹۴/۸۰
		الگوریتم بهینه‌سازی شیرمورچه با ماشین بردار پشتیبان	۹۵/۳۷
		الگوریتم بهینه‌سازی شیرمورچه با بوستینگ	۹۸/۹۱
	-	**روش پیشنهادی**	۹۷/۲۵
Enron	[۲۸]	K نزدیکترین همسایه	۸۹/۷۵
		K نزدیکترین همسایه وزنی	۸۹/۷۵
		درخت تصمیم‌گیری	۸۹/۷۵
	-	**روش پیشنهادی**	۹۰/۱۳

نتایج به‌دست‌آمده از تکرارهای مختلف نشان داد که درصد صحت روش پیشنهادی با افزایش تعداد تکرار رابطه مستقیم دارد و اگر تعداد تکرار افزایش یابد درصد صحت هم افزایش یافته است. اگر تعداد تکرارها بیشتر باشد، ذرات می‌توانند فضای جستجو را با دقت بیشتری جستجو کنند و همسایگان برازنده را برای رسیدن به نقطه بهینه کشف کنند. همچنین طبق نتایج دریافتیم که اگر اندازه حافظه هارمونی بیشتر بوده درصد صحت بیشتر شده است.

۵- نتیجه‌گیری و کارهای آینده

در بیش‌تر موارد، هدف ایمیل هرزنامه قصد نفوذ به رایانامه کاربران یا کلاه‌برداری است که این‌گونه هرزنامه‌ها اغلب حاوی نرم‌افزارها و پیوندهای مخرب هستند که در درون خود انواع بدافزارها، ویروس‌ها و باج افزارها را دارند. تاکنون روش‌های مختلفی برای مبارزه با هرزنامه ایجاد شده که یکی از راه‌کارها استفاده از نرم‌افزارهای فیلترینگ است که در این برنامه‌ها با استفاده از کلید واژه‌های خاصی، موضوع پیام و رایانامه‌ها بررسی می‌شود و در صورت شناسایی آن‌ها رایانامه دریافت‌شده حذف می‌شود. در چند سال گذشته برمبنای مشخصه‌های اینترنتی و محتویات رایانامه‌ها، یک‌سری مجموعه داده تهیه شده‌اند که یکی از آنها Spambase است. این مجموعه‌داده حاوی مقادیر ویژگی‌های مهم ایمیل‌های هرزنامه و غیرهرزنامه است. در این مقاله به‌منظور طبقه‌بندی و تشخیص ایمیل هرزنامه از روش ترکیبی برمبنای الگوریتم جستجوی هارمونی و الگوریتم بهینه‌سازی مغناطیسی استفاده شد. در روش پیشنهادی، هدف انتخاب ویژگی‌های تأثیرگذار برای طبقه‌بندی بود. طبق نتایج به‌دست‌آمده برروی Spambase، درصد صحت با دویست بار تکرار و ۵۷ ویژگی برابر با ۹۴/۱۷ درصد بود. همچنین نتایج نشان داد که اندازه حافظه هارمونی در دقت تشخیص مؤثر است و افزایش اندازه حافظه هارمونی با درصد صحت رابطه مستقیم دارد. همچنین ارزیابی برروی سه نوع مجموعه‌داده مختلف نشان داد که روش پیشنهادی از کارایی لازم برای تشخیص ایمیل هرزنامه برخوردار بوده است. برای کارهای آینده در نظر داریم از ترکیب درختان تصمیم‌گیری و الگوریتم‌های فراابتکاری برای تشخیص ایمیل هرزنامه استفاده کنیم.

۶- مراجع

[1] A. Mortazavi, F.S. Gharehchopogh, "A New Hybrid Approach of K-Nearest Neighbors Algorithm with Particle Swarm Optimization for E-Mail Spam Detection", Biannual Journal Monadi for Cyberspace Security (AFTA), Vol. 8, No. 1, pp. 57-72, 2019

[2] F.S. Gharehchopogh, M. Vafadar, M. Motamanfar, "Improving Invasive Weed Optimization Using K-Nearest Neighbors in Email Spam Classification", Computing Science Journal (CSJ), No. 10, pp. 54-64, 2018.

[3] H. Faris, A.M. Al-Zoubi, A.A. Heidari, I. Aljarah, H. Fujita, "An Intelligent System for Spam Detection and Identification of the most Relevant Features based on Evolutionary Random Weight Networks", Information Fusion, Vol. 48, pp. 67-83, 2019

[4] F.S. Gharehchopogh, N. Karimpour, "A New Approach to Detect Spam Emails Using the Hybrid Model of Ant Colony and Firefly Algorithms", Computing Science Journal (CSJ), in press, 2019.

[5] F.S. Gharehchopogh, H. Gholizadeh, "A comprehensive survey: Whale Optimization Algorithm and its applications", Swarm and Evolutionary Computation, Vol. 48, 1-24, 2019

[6] H. Majidpour, 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, No. 1, pp. 29-40, 2018.

[7] P. Zhou, X. Hu, P. Li, X. Wu, "OFS-Density: A novel online streaming feature selection method, Pattern Recognition", Vol. 86, pp. 48-61, 2019.

[8] Z.W. Geem, Optimal Design of Water Distribution Networks using Harmony Search, Ph.D, Thesis, Korea University, 2000.

[9] M.H. Tayarani-N, M.R. Akbarzadeh-T, "Magnetic Optimization Algorithms a new synthesis", IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), pp. 2659-2664, 2008.

[10] Martin, "Instance-Based Learning: Nearest Neighbour with Generalisation", Doctoral dissertation, University of Waikato, 1995.

[11] https://archive.ics.uci.edu/ml/datasets/spambase

[12] A.R. Behjat, A. Mustapha, H. Nezamabadi-pour, M.N. Sulaiman, N. Mustapha, "A PSO-Based Feature Subset Selection for Application of Spam /Non-spam Detection", International Multi-Conference on Artificial Intelligence Technology, M-CAIT 2013: Soft

Computing Applications and Intelligent Systems, pp. 183-193, 2013.

[13] I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system", Engineering Applications of Artificial Intelligence, Vol. 39, pp. 33-44, 2015.

[14] I. Idris, A. Selamat, S. Omatu, "Hybrid email spam detection model with negative selection algorithm and differential evolution", Engineering Applications of Artificial Intelligence, Vol. 28, pp. 97-110, 2014.

[15] M.C. Su, H.H. Lo, F.H. Hsu, "A neural tree and its application to spam e-mail detection", Expert Systems with Applications, Vol. 37, Issue 12, pp. 7976-7985, 2010.

[16] K.C. Ying, S.W. Lin, Z.J. Lee, Y.T. Lin, "An ensemble approach applied to classify spam e-mails, Expert Systems with Applications", Vol. 37, Issue 3, pp. 2197-2201, 2010.

[17] W. Ma, D. Tran, and D. Sharma, "A Novel Spam Email Detection System Based on Negative Selection", Fourth International Conference on Computer Sciences and Convergence Information Technology, IEEE, pp. 987-992, 2009.

[18] S.B. Rathod, T.M. Pattewar, "Content Based Spam Detection in Email using Bayesian Classifier", International Conference on Communications and Signal Processing (ICCSP), IEEE, pp. 1257-1261, 2015.

[19] I. Idris, "E-mail Spam Classification with Artificial Neural Network and Negative Selection Algorithm", International Journal of Computer Science & Communication Networks, Vol.1, No. 3, pp. 227-231, 2010.

[20] S. Jain, S. Shukla, R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures", Expert Systems with Applications, Vol. 106, pp. 252-262, 2018.

[21] A. Tharwat, "Classification assessment methods, Applied Computing and Informatics", In press, corrected proof, Available online 21 August 2018.

[22] M.Z. Gashti, "Detection of Spam Email by Combining Harmony Search Algorithm and Decision Tree", Engineering, Technology & Applied Science Research, Vol. 7, No. 3, pp. 1713-1718, 2017

[23] S. Sharma, A. Arora, "Adaptive Approach for Spam Detection", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No. 1, pp. 23-26, July 2013.

[24] F. Barani, M. Mirhosseini, H. Nezamabadi-pour, "Application of binary quantum-inspired gravitational search algorithm in feature subset selection", Applied Intelligence, Vol. 47, Issue 2, pp 304-318, 2017.

[25] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, "Stacking classifiers for anti-spam filtering of E-mail", in: Empirical Methods in Natural Language Processing, pp. 44-50, 2001.

[26] J.R. Mendez, F. Diaz, E.L. Iglesias, J.M. Corchado, "A comparative performance study of feature selection methods for the anti-spam filtering domain", in: Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Springer Berlin Heidelberg, pp. 106–120, 2006.

[27] I. Koprinska, J. Poon, J. Clark, J. Chan, "learning to classify e-mail", Information Sciences: An International Journal, 177(10): 2167-2187, 2007.

[28] G. Hnini, J. Riffi, M.A. Mahraz, A. Yahyaouy, H. Tairi, "Spam Filtering System Based on Nearest Neighbor Algorithms", International Conference on Artificial Intelligence & Industrial Applications, A2IA 2020: Artificial Intelligence and Industrial Applications, 144: 36-46, 2021.

[29] A.A. Naem, N.I. Ghali, A.A. Saleh, "Antlion optimization and boosting classifier for spam email detection", Future Computing and Informatics Journal, 3(2): 436-442, 2018.

فرهاد سلیمانیان قره چپق

کارشناسی ارشد و دکترای تخصصی خود را در رشته مهندسی رایانه 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 رایانه 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.