ارائه روشی جدید برای شناسایی رایانامه‌های مخرب و فیشینگ بانکی با استفاده از الگوریتم‌های ژنتیک و جستجوی ممنوع

مجید فانی[1]، محمدامین ترابی[2] و متینه مقدم[3]

[1]استادیار،گروه مدیریت بازرگانی، دانشگاه آزاد اسلامی، بابل، ایران
Fani@baboliau.ac.ir
[2]دانشجوی دکتری مدیریت بازرگانی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران
Torabi628@Gmail.com
[3]دکتری مدیریت بازرگانی، دانشکده مدیریت، اقتصاد و حسابداری، دانشگاه پیام نور، تهران، ایران
Matineh_moghaddam@yahoo.com

چکیده

همه حملات فیشینگ همواره به‌صورت جعل وبگاه و فیشینگ تلفنی انجام نمی‌شود. رایانامه‌ها و پیام‌هایی که به‌ظاهر از طرف بانک فرستاده می‌شود و از کاربر اطلاعات دریافت می‌کنند، نیز می‌تواند حمله فیشینگ باشد. انتخاب ویژگی و انتخاب نمونه دو مسأله بسیار مهم در مرحله پیش‌پردازش داده‌ها در کشف رایانامه‌های مخرب هستند. به‌خصوص، در شناسایی هرزنامه‌ها که بدون کاهش داده تقریباً دقت خوبی در نتایج به‌دست نخواهد آمد. بیشتر مقالات و تحقیقات بر روی یکی از این مسأله تمرکز کرده‌اند و کمتر مقالاتی وجود دارند که به‌صورت ترکیبی در جهت کشف رایانامه‌های مخرب کار کرده باشند. ازاین‌رو هدف از پژوهش حاضر، ارائه روشی است که جهت کاهش داده در شناسایی رایانامه‌ها انتخاب ویژگی و نمونه را به‌صورت هم‌زمان انجام دهد. در روش پیشنهادی در این مقاله از الگوریتم جست‌وجوی ممنوع و الگوریتم ژنتیک به‌صورت ترکیبی و هم‌زمان استفاده شده است. جهت برازندگی این روش نیز از تابع ارزیابی ماشین بردار پشتیبان بهره گرفته شد. نتایج نشان دادکه میزان صحت تشخیص شناسایی هرزنامه‌ها و رایانامه‌ها در مجموعه دادگان لاین‌اسپم و یوسی‌آی، ۹۷/۲۸ است که نسبت به سایر الگوریتم‌های پیشنهاد شده در پژوهش‌های قبلی، دارای بیشترین مقدار ممکن بوده است.

واژگان کلیدی: فیشینگ بانکی، شناسایی هرزنامه، الگوریتم‌های ژنتیک، جستجوی ممنوع

۱- مقدمه

فیشینگ[1] یک نوع حمله رایانامه‌ای است که مهاجم از طریق کانال‌های ارتباط الکترونیکی[2]، با ایجاد ارتباط با انسان‌ها و با استفاده از پیام‌های مهندسی اجتماعی، به ترغیب آن‌ها، برای انجام کارهایی که به نفع مهاجم است، اقدام می‌کند [۱, ۲]. به عبارت دیگر، یک کلاهبرداری است که معمولا از طریق رایانامه انجام می‌شود تا اطلاعات شخصی افراد سرقفت شود. درخصوص نام‌گذاری این حمله، بیشتر متخصصین معتقدند کلمه فیشینگ نمایانگر به دام انداختن قربانیان حمله، به مثابه ماهیگیری[3]، است[۳].

کلمه فیشینگ ابتدا در سال ۱۹۹۶ استفاده گردید که فیشرها، با استفاده از مهندسی اجتماعی، مجوزهای کاربری را مورد حمله قرار دادند. کلاهبردارن اینترنتی، از رایانامه برای صید اطلاعات مالی و گذرواژه‌های کاربران اینترنتی استفاده کردند [۴].

کلاهبرداری فیشینگ به‌صورت مرسوم از طریق جعل وبگاه بانک [۵]، تماس تلفنی جعلی [۶]، فیشینگ نیزه‌ای[4] [۷] و ارسال رایانامه‌های انبوه [۸]، انجام می‌شود.

مطالعات و تلاش‌های زیادی در جهت ارائه راهکارهای مختلف برای مقابله با حملات سایبری فیشینگ مطرح شده است که در دو دسته آگاهی اجتماعی از فیشینگ جهت جلوگیری و شناسایی خودکار هوشمند صفحات جعلی فیشینگ، قرار می‌گیرند [۹].

[1] Phishing
[2] Electronic Communication Channels
[3] Phishers
[4] Spear Phishing

*نویسندهٔ عهده‌دار مکاتبات: محمدامین ترابی

[Downloaded from monadi.isc.org.ir on 2022-08-19]

در خصوص مورد اول، پژوهش‌های گسترده‌ای انجام شده است که نشان‌دهنده‌ای این است که رسانه‌های جمعی [۱۰]، سیاست‌های آگاهی رسانی دولت [۱۱]، آموزش در مدارس و دانشگاه [۱۲]، شبکه‌های اجتماعی [۱۳] و پلیس هر کشور [۱۴]، در اطلاع‌رسانی و آگاهی مردم جهت شناسایی فیشینگ‌ها، نقش مؤثری دارند. اما میزان حملات سایبری فیشینگ به‌خصوص ارسال انبوه رایانامه‌ها به قدری زیاد و پیچیده است که از نظر زمانی و دقت انسانی قابل شناسایی و کشف به صورت دستی و غیرسیستماتیک نیستند [۱۵]، به همین خاطر استفاده از الگوریتم‌های ماشینی جهت کشف و حذف هوشمند سیستماتیک این رایانامه‌ها، مسأله‌ای است که نهادهای امنیت سایبری به خصوص نیروی انتظامی فضای مجازی، به دنبال یافتن روشی کارا و موثر جهت حل آن هستند.

از طرفی دیگر، هرزنامه‌ها، رایانامه‌هایی هستند که به‌صورت انبوه به جهت تبلیغات، افزایش بازدید سایت جهت سئو، افزایش بازدید پست‌های شبکه‌های اجتماعی و اخبار بی اعتبار برای کاربران هدف زیادی ارسال می‌شوند [۱۶].

فرستادن هرزنامه همچنان به دلیل صرفه اقتصادی پایدار می‌ماند؛ زیرا تبلیغ‌کنندگان هیچ هزینه‌ای صرف مدیریت لیست‌های رایانامه‌هایشان نمی‌کنند و این، کار را برای مسئول دانستن فرستندگان رایانامه سخت می‌کند [۱۷].

متأسفانه همین عمومیت و سادگی استفاده از رایانامه باعث شده تا مورد سوءاستفاده هرزنامه‌نویسان و کلاه‌برداران اینترنتی قرار بگیرد، مسأله‌ای که روزی کم‌اهمیت جلوه می‌نمود، امروزه به معضلی جدی برای میلیون‌ها کاربر رایانامه بدل شده است و استفاده از ابزار و متدهایی برای شناسایی و فیلتر هرزنامه‌ها ضرورتی غیرقابل‌انکار است [۱۸].

هرزنامه‌ها که معمولاً تبلیغاتی هستند، ویژگی‌های مشابهی دارند مثلاً آن‌هایی که محصولی را تبلیغ می‌کنند از قیمت آن حرف می‌زنند و یا می‌گویند که فرصتتان چقدر استثنایی است. حتی رنگارنگ بودن بخش‌های نوشته می‌تواند نشان از بی‌ارزش بودن آن باشد [۱۹]. ازآنجایی‌که این نشانه‌های قطعی نیستند نمی‌توان با چند قانون ساده هرزنامه‌ها را جدا کرد، لذا برای شناسایی این هرزنامه‌ها نیاز به کاهش داده‌ها است [۲۰-۲۳]. یکی از طبقه‌بندها برای شناسایی رایانامه‌ها، ماشین بردار پشتیبان است که با داده‌های حجیم، نتایج خوبی به‌دست نمی‌دهد [۲۴-۲۶].

کاهش داده یکی از مراحل داده‌کاوی است که در مرحلهٔ پیش‌پردازش داده استفاده می‌شود [۲۷]. کاهش داده به این معناست که تعدادی از داده‌ها را انتخاب کنیم که در اصل نماینده تمامی داده‌ها باشند. قسمتی از داده‌ها، داده‌های غیر مرتبط و اضافی هستند که وجود آن‌ها هیچ مزیتی ندارد و حتی در برخی از موارد این داده‌های زائد باعث ایجاد نویز و خرابی در الگوریتم‌های یادگیری می‌شوند و آن‌ها را به‌اشتباه و خطا می‌اندازد. ازاین‌رو کاهش داده علاوه بر اینکه باعث کاهش در مصرف حافظه می‌شوند، حتی می‌تواند دقت، سرعت و کارایی بهتری را نیز در برداشته باشند [۲۸].

داده‌ها دارای دو بعد سطر و ستون هستند، ستون‌های آن به‌عنوان ویژگی شناخته می‌شوند و سطرهای آن به‌عنوان نمونه معرفی می‌شوند. حال می‌توان هم در سطر(انتخاب نمونه) و هم در ستون(انتخاب ویژگی) کاهش داده‌ها را اختیار کرد. انتخاب ویژگی‌های مرتبط جهت بازدهی بهتر و دقت بالاتر طبقه‌بندها و الگوریتم‌های یادگیری تأثیر بسزایی دارند و به همین خاطر موردتوجه بسیاری در داده‌کاوی قرارگرفته‌اند [۲۹].

انتخاب نمونه نیز یکی دیگر از روش‌های کاهش ابعاد داده است. نمونه‌هایی که دارای بیشترین ارتباط با کلاس مربوطه هستند گزینش می‌شوند. در این نمونه‌ها، نمونه‌هایی که غیر مرتبط و یا اضافی هستند حذف می‌شوند که هم در نمونه‌ها کاهش به‌دست آید و هم برای طبقه‌بندها عملیات کمتری وجود داشته باشد [۳۰]؛ بنابراین می‌توان در طبقه‌بندی‌ها در بخش آموزش داده‌ها کاهش داده‌ها را انجام داد. منظور از عملیات طبقه بندی، در بخش آموزش داده است که سعی می‌شود نمونه‌های کمتری به طبقه‌بندی برای آموزش ارائه شود.

در بیشتر مقالات طبقه‌بندی رایانامه‌ها از کاهش داده استفاده‌شده است و تقریباً بدون کاهش داده(انتخاب ویژگی) دقت طبقه‌بندی خوبی به‌دست نخواهد آمد [۲۷, ۳۱]. انتخاب ویژگی و انتخاب نمونه هر دو باعث کاهش داده می‌شوند و ویژگی‌ها و نمونه‌های نامرتبط، زائد و نویزدار را حذف می‌کند. درواقع انتخاب ویژگی و نمونه باعث تسریع الگوریتم یادگیری و بهبود کارایی مانند دقت و صحت نتایج می‌شود [۲۸, ۳۲]. انتخاب ویژگی و نمونه از ابزارهای مهم در کاهش داده‌های نامربوط و اضافی هستند. و مدل را ساده‌تر خواهد کرد و درنتیجه هزینه

[Downloaded from monadi.isc.org.ir on 2022-08-19]

محاسباتی را کاهش می‌دهد. زیرا مدلی که ورودی‌های کمتری دارد، ارزیابی کمتری برای نمونه‌های جدید انجام می‌دهد و درنتیجه هزینه محاسباتی کاهش می‌یابد. با حذف ویژگی‌ها و نمونه‌های غیر مهم از مجموعه داده‌های اصلی، مدل شفاف‌تر خواهد شد [۲۸].

در این پژوهش، ،هدف این است که روش مناسبی برای کاهش داده‌ها هم در بعد ویژگی و هم در بعد نمونه برای شناسایی رایانامه‌های هرزنامه، مخرب و فیشینگ ارائه شود. به دلیل افزایش روزافزون ابعاد داده‌ها و کاهش کارایی الگوریتم‌های یادگیری در شناسایی رایانامه‌های مخرب، استفاده از روشی برای کاهش هر دو بعد داده لازم و ضروری است؛ زیرا در تحقیقات به‌عمل‌آمده یکی از بهترین طبقه‌بندها در تشخیص رایانامه‌های مخرب ماشین بردار پشتیبان[1] بوده است که این طبقه‌بندی در داده‌هایی با ابعاد زیاد به‌خوبی عمل نمی‌کند [۳۳]. در تحقیقات و مقالات اغلب بر روی یکی از این دو بعد و مخصوصاً بر روی انتخاب ویژگی کار شده است. ازاین‌رو ما بر این شدیم که انتخاب ویژگی و نمونه را باهم به‌صورت همزمان انجام دهیم. با توجه به کارایی الگوریتم‌های متاهیورستیک(الگوریتم‌های فراابتکاری)[2]، از آن‌ها برای این عمل استفاده کرده‌ایم. در بین این الگوریتم‌ها به الگوریتم ژنتیک[3] در کاهش دوبعدی داده‌ها و الگوریتم جستجوی ممنوعه[4] توجه ویژه‌ای شده است. در نهایت، الگوریتم استفاده‌شده در این پژوهش، ترکیبی از الگوریتم‌های ژنتیک، جستجوی ممنوع و تابع ارزیابی[5] است.

۲- مبانی نظری و پیشینه تحقیق

در دادگان دو نوع نمایش داده وجود دارد که داده‌های با قالب، مانند جداول اکسل، دارای سطر وستون‌اند و داده‌های بدون غالب مثل داده‌های متنی.

داده‌های با قالب دارای دو بعد سطر وستون هستند که سطرها خود رایانامه‌ها و ستون‌ها بیانگر ویژگی‌ها هستند. جهت شناسایی ویژگی‌ها نیز، از تابع انتخاب ویژگی استفاده می‌شود که زیرمجموعه استخراج ویژگی است.

۳- انتخاب ویژگی

ابعاد داده در داده‌کاوی به‌صورت قابل توجهی افزایش یافته است. داده‌ها با ابعاد بسیار بالا، چالش های جدید روش‌های یادگیری موجود به وجود آورده‌اند. برای رسیدگی به مشکل افزایش ابعاد، روش‌های کاهش ابعاد مورد مطالعه قرارگرفته است، که به یک شاخه مهم و پژوهشی در یادگیری ماشین وداده‌کاوی تبدیل شده است [۳۴].

فرایند انتخاب زیردسته‌ای از میان دسته ویژگی‌ها است که در جهت پیداکردن زیرمجموعه‌ای از ویژگی‌ها با حداقل اندازه، لازم وکافی برای مفهوم هدف است [۳۵]. به عبارت دیگر انتخاب یک زیرمجموعه M عنصری از میان N ویژگی، به طوریکه m<n باشد و همچنین مقدار یک تابع معیار برای زیرمجموعه مورد نظر، نسبت به سایر زیرمجموعه‌ها با اندازه M، بهینه باشد [۳۶].

انتخاب ویژگی یک تکنیک گسترده‌ی کاربردی برای کاهش ابعاد درمیان محققین است که هدف آن انتخاب زیرمجموعه کوچکی از ویژگی‌های مرتبط است که از ویژگی‌های اصلی با توجه به معیار ارزیابی ارتباط[6]،که معمولاً به بهبود عملکرد یادگیری و هزینه محاسباتی پایین تر،و مدل قابل تفسیر بهتر منجر می‌شود [۳۷].

۳-۱- فواید انتخاب ویژگی

۱. **افزایش دقت پیش‌بینی:** هدف از انتخاب ویژگی این است که یک زیرمجموعه از ویژگی‌ها را برای بهبود دقت پیش‌بینی و یا کاهش اندازه ساختار انتخاب کند، البته بدون کاهش دقت پیش‌بینی درطبقه‌بند و تنها با استفاده از ویژگی‌های انتخاب شده باشد [۳۸, ۳۹].

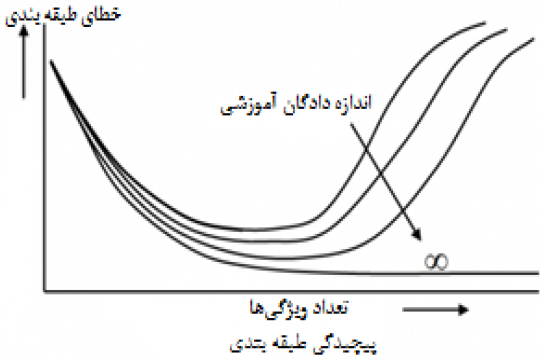
۲. **تقریب توزیع کلاس اصلی:** هدف از انتخاب ویژگی این است که یک زیرمجموعه کوچک از ویژگی‌ها انتخاب شوند، توزیع ویژگی‌هایی که انتخاب می‌شوند، بایستی تا حد امکان به توزیع کلاس اصلی با توجه به تمام مقادیر ویژگی‌های انتخاب شده نزدیک باشد [۳۹].

۳. **پدیده پیکینگ[7] و اهمیت کاهش ابعاد:** بر خلاف رفتار کلاسیفایر بیز[8]، خطای کلاسیفایر طراحی شده توسط داده‌های واقعی رفتار متفاوتی را در حالت افزایش تعداد ویژگی‌ها نمایش می‌دهد. درنتیجه پدیده

[1] Support Vector Machines(SVM)
[2] Approximate Algorithms
[3] Genetic Algorithm
[4] Tabu Search
[5] Evaluation Function
[6] Relevance Evaluation Criterion
[7] Peaking
[8] Base Classifier

[Downloaded from monadi.isc.org.ir on 2022-08-19]

پیک زدن، افزایش تعداد ویژگی‌ها در مواردی می‌تواند به افزایش خطای طبقه‌بندی نیز منجر شود. از این جهت است که در یک مسأله عملی، بهترین تعداد ویژگی‌ها، بیشترین تعداد ویژگی‌ها ناست. شکل زیر پدیده پیک‌زدن را نشان می‌دهد [۴۰, ۴۱].



(شکل-۱):پدیده پیکینگ [۴۰, ۴۱].

با افزایش تعداد ویژگی‌ها از یک نقطه به بعد، خطای طبقه‌بندی افزایش می‌یابد. پدیده پیکینگ به تنهایی برای نمایش دادن اهمیت استفاده از مقدار مناسب (نه کم و نه زیاد) ویژگی‌ها را نشان می‌دهد. البته این همه اهمیت داستان نیست. انتخاب ویژگی می‌تواند به کاهش پیچیدگی محاسباتی (هم از جهت بار اجرایی و هم ذخیره سازی) منجر شود. از سوی دیگر، کاهش ابعاد می‌تواند به ایجاد شناخت بهتر از داده‌ها نیز کمک نماید. به‌عنوان مثال در یک مسأله شناسایی بیماری، دانستن ۳ ویژگی مهم که بیشترین جداسازی داده‌ها را ایجاد می‌کند، مفید خواهد بود (مثلاً سه ژن عامل یک بیماری) [۴۱].

۴-الگوریتم ژنتیک

الگوریتم ژنتیک، الهامی از علم ژنتیک و نظریه تکامل داروین[1] است و بر اساس بقای برترین‌ها یا انتخاب طبیعی استوار است. یک کاربرد متداول الگوریتم ژنتیک، استفاده از آن به‌عنوان تابع بهینه‌کننده است [۴۲, ۴۳]. الگوریتم ژنتیک ابزار سودمندی دربازشناسی الگو [۴۴]، انتخاب ویژگی [۴۵]، درک تصویر [۴۶] و یادگیری ماشینی [۴۷] است. در الگوریتم‌های ژنتیکی[2]، نحوه تکامل ژنتیکی موجودات زنده شبیه‌سازی می‌شود [۴۸]. الگوریتم‌های ژنتیکی را می‌توان یک روش بهینه‌سازی تصادفی جهت‌دار دانست که به تدریج به سمت نقطه بهینه حرکت می‌کند. در مورد ویژگی‌های الگوریتم ژنتیک در مقایسه با دیگر روش‌های بهینه‌سازی می‌توان گفت که الگوریتمی است که بدون داشتن هیچ گونه اطلاعی از مسأله و هیچ گونه محدودیتی بر نوع متغیرهای آن برای هر گونه مسأله ای قابل اعمال است و دارای کارآیی اثبات‌شده‌ای در یافتن بهینه کلی[3] می‌باشد. توانایی این روش در حل مسائل پیچیده بهینه‌سازی, است که روش‌های کلاسیک یا قابل اعمال نیستند و یا دریافتن بهینه کلی قابل اطمینان نیستند [۴۹].

[1] theory of biological evolution
[2] Genetic Algorithm

۵-الگوریتم جستجوی ممنوعه‌ای

یک الگوریتم بهینه‌سازی فراابتکاری است که برای اولین بار در سال ۱۹۸۶ توسط گلوور[4] معرفی شد. واژه تابو از تُنگان زبان مردم جزایر پلینزی در اقیانوس آرام گرفته شده‌است. این واژه به معنای شیء مقدسی است که به دلیل قداست نباید آن را لمس کرد [۵۰]. بر اساس واژه‌نامه‌ی وِبستر، امروزه این واژه در معنای «ممنوعیت ایجاد شده به دلیل فرهنگ اجتماعی برای ایجاد اقدام حفاظتی» یا «ممنوعیت چیزی که دارای ریسک است»، به کار می‌رود. معنای اخیر واژه تابو، با تکنیک جستجوی ممنوعه کاملاً سازگار است. ریسکی که در الگوریتم جستجوی ممنوعه از آن اجتناب می‌شود، خطر مسیرهای نامناسب است [۵۱].

۶- طبقه‌بندی ماشین بردار پشتیبان

ماشین بردار پشتیبان در سال ۱۹۹۵ توسط واپنیک[5] پیشنهاد شد. یکی از روش‌هایی که در حال حاضر به‌صورت گسترده‌ای برای مسأله دسته‌بندی مورداستفاده قرار می‌گیرد [۵۲]. شاید به‌گونه‌ای بتوان محبوبیت کنونی روش ماشین بردار پشتیبان را با شبکه‌های عصبی در دهه گذشته مقایسه کرد. علت این قضیه نیز قابلیت استفاده این روش در حل مسائل گوناگون است، درحالی‌که روش‌هایی مانند درخت تصمیم‌گیری را نمی‌توان به‌راحتی در مسائل مختلف به کار برد [۵۳]. SVMها الگوریتم‌های طبقه‌بندی بسیار قدرتمندی هستند. هنگامی‌که در مواردپیش‌بینی[6] و دیگر ابزار یادگیری ماشین استفاده می‌شوند، ابعاد بسیار گسترده‌ای از مدل‌های کارآمد را ارائه می‌دهند. ازاین‌رو برای مواردی که قدرت پیش‌بینی بسیار بالا موردنیاز است بسیار بااهمیت به شمار می‌آیند [۵۴]. بااینکه تجسم بعضی الگوریتم‌ها به‌واسطهٔ پیچیدگی فرمول آن‌ها کمی سخت است یک ماشین بردار پشتیبان برای

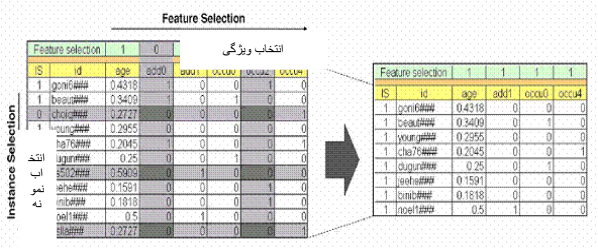
[3] Global Optimum
[4] Glover
[5] Vapnik
[6] Prediction

[Downloaded from monadi.isc.org.ir on 2022-08-19]

تعیین ابر صفحه جداکننده داده‌های آموزش از مکانیسمی استفاده می‌کند که در آن به کمک یک تابع هسته‌ای داده‌ها را در فضا به بعدی انتقال می‌دهد که بتواند این ابر صفحه را بر داده‌های ورودی برازش دهد [۵۵]. بنابراین انتخاب درست یک تابع هسته‌ای می‌تواند تأثیر مستقیم روی کارایی ماشین بردار پشتیبان داشته باشد. در زمان استفاده ماشین بردار پشتیبان دو مسأله مهم وجود دارد که باید مدنظر قرارگرفته شوند. اول اینکه چگونه مجموعه داده ورودی انتخاب شود، یا اینکه آیا انتقالی روی داده‌های ورودی قبل از ورود به ماشین بردار پشتیبان می‌توان ایجاد نمود که نتایج مناسب‌تری ایجاد کند یا خیر، و دوم اینکه چگونه بهترین پارامترهای تابع هسته‌ای تنظیم شود. این دو مسأله باهم کاملاً در ارتباط‌اند. چراکه مجموعه‌ای ورودی روی انتخاب پارامترهای هسته تأثیرگذار است [۵۶].

۷- پیشینه پژوهش

اولین روش کاهش دو بعدی به وسیله کانچوا و جاین[1](۱۹۹۹) مطرح شد. آن‌ها بهینه‌کردن همزمان انتخاب ویژگی و نمونه را با استفاده از الگوریتم ژنتیک پیشنهاد دادند، و مدلشان را با الگوریتم‌های پیشین، که به طور متوالی انتخاب ویژگی و نمونه را باهم ترکیب می‌کردند، مقایسه کردند [۵۷]. بعد از آن رازسپال و کوبات[2] مدل مشابهی ارائه دادند و نشان دادند که مدلشان از مدل کانچوا و جاین بهتر عمل می‌کند [۵۸]. شکل (۲) نمونه‌ای از کاهش دو بعدی را نشان می‌دهد.



(شکل-۲): کاهش دو بعدی دادگان با استفاده از الگوریتم ژنتیک [۵۸].

طبیعی‌ترین راه و روش مستقیم روبه جلو برای ترکیب انتخاب ویژگی و نمونه، اجرای یک فرآیند پس ازدیگری است و در عمل روشی انتخاب شده است که هر دو مشکل را پوشش داده است. به‌عنوان مثال روش انتخاب ویژگی و انتخاب نمونه[3] معرفی شده است، در این روش در ابتدا عمل انتخاب ویژگی انجام شده و به دنبال آن عمل انتخاب نمونه صورت می‌گیرد و برای انتخاب نمونه و انتخاب ویژگی[4] برعکس این روند طی می‌شود. از آنجا که این‌روش به طورکلی و به‌صورت عمومی شناخته شده است، یعنی می‌توان آن را در هر حوزه‌ای به کار برد [۵۹]. در پژوهش [۵۹] پیشنهاد FIS[5] برای استخراج ویژگی از مجموعه دادگان و اسناد متنی ارائه شده است، که در آن دو روش مسأله انتخاب ویژگی و نمونه به‌صورت همزمان، در طبقه‌بندی متن[6] ارائه شده است. برخی از کارهای انجام شده، راه حل های فرااکتشافی[7] را برای حل این دو مسأله(انتخاب ویژگی و انتخاب نمونه) استفاده کرده‌اند. در [۶۰] نویسندگان از روش شبیه‌سازی تبرید[8] [۶۱] برای حل این دو استفاده کردند به این صورت که دو شبیه سازی تبرید را به‌صورت همزمان اجرا می‌کند و هر مسأله را به‌صورت جداگانه حل می‌کند. در اصل در این روش از دو شبیه سازی تبرید تو در تو استفاده می‌شود که کیفیت اجرای هر کدام از فرآیندها بر روی کیفیت یکدیگر تأثیر می‌گذارند. در [۵۷] نویسندگان از روش کدگذاری و مقداردهی بولین به ویژگی‌ها و نمونه‌ها استفاده کرده است. تابع هدف این روش نیز به‌صورت ترکیبی است و شامل دقت1-NN و مقدار جریمه[9] های کاردینالیتی[10]برای هر مجموعه است. در [۶۲] نویسندگان از الگوریتم‌های ژنتیک اقتباس گرفتند که به نام الگوریتم برآورد توزیع[11](EDA) است که برای انتخاب نمونه و ویژگیهای مشکل برآورد احتمال مرگ بیماران مبتلا به سیروز که حدبیشتر ۶ ماه پس از درمانی که TIPS[12] نامیده می‌شود، استفاده می‌شود.

[۶۳] یک مطالعه‌ای است که از طراحی چند هدفه صریح[13] برای مسائل انتخاب ویژگی و نمونه استفاده می‌شود. در این روش سعی شده که عملکرد طبقه بند 1-NN به حدبیشتر برسد و همچنین تعداد ویژگی‌ها و نمونه‌ها به حداقل برسد. در [۶۴] یک رویکرد چند هدفه پیشنهاد شده است که حل آن‌ها توسط الگوریتم ژنتیک و در دو فاز انجام شده است. در [۶۵] نیز ازیک الگوریتم ژنتیک استفاده کرده است که گرایش به کاهش تعداد ویژگی‌های انتخاب شده دارد . این کار با دادن احتمال

[1] Kuncheva & Jain
[2] Rozsypal & Kubat
[3] Feature Selection and Instance Selection(FSIS)
[4] Instance Selection and Feature Selection(ISFS)
[5] Feature and Instance Selection
[6] Text Classification
[7] Meta Heuristic
[8] Simulated Annealing
[9] Penalizes
[10] Cardinality
[11] Estimation of Distribution Algorithm
[12] Transjugular Intrahepatic Portosystemic Shunt
[13] Explicit Multi-Objective

[Downloaded from monadi.isc.org.ir on 2022-08-19]

بزرگتر به ویژگی، برای حذف آن‌ها است که این مقدار در حال تغییر کردن است. در [۶۶] کاربرد الگوریتم ژنتیک را برای انتخاب ویژگی و نمونه باهم و تنها بررسی کرده است. دو روش کلی انتخاب همزمان ویژگی و نمونه به‌صورت زیر است:

- انتخاب ویژگی در ابتدا وسپس انتخاب نمونه(FSIS)
- انتخاب نمونه درابتدا و سپس انتخاب ویژگی(ISFS)

تجزیه و تحلیل پیچیدگی زمانی نشان داد که ترکیب انتخاب ویژگی و نمونه تا حد زیادی هزینه‌های محاسباتی آموزش طبقه بندها را کاهش می‌دهد. حتی ترکیب این دو، از انتخاب ویژگی و نمونه که به تنهایی انجام شده، کاهش بیشتری داشته است (از طبقه بند k-NN و SVM استفاده‌شده است).

۸- روش پژوهش

روش تحقیق پژوهش حاضر از نظر هدف، کاربردی و از نظر ماهیت، آزمایشگاهی-شبیه‌سازی است. ابزار جمع‌آوری اطلاعات نیز به‌صورت مطالعات کتابخانه‌ای و برای بخش شبیه‌سازی نیز، از مجموعه دادگان حاضر در مقاله [۶۷] استفاده شده است. محیط شبیه‌سازی نیز نرم‌افزار متلب[1] نسخه 2020a بوده است.

۸-۱- روش پیشنهادی پژوهش

بیشتر پژوهش‌های انجام‌شده در حوزه تخصصی مقالات مربوط به انتخاب ویژگی و نمونه تلاش می‌کنند برای حل این مشکلات از نسخه‌های تخصصی الگوریتم ژنتیک دریک فرمول گسترده عمومی استفاده و آن‌ها را حل کنند. تلاش برای جدا کردن کروموزوم به دو حوزه مختلف، یکی برای ویژگی‌ها و یکی برای نمونه‌ها و اپراتورهای جداگانه‌ای را به هر حوزه اعمال نمایند. این روش‌ها از آسانی و سهولت طبیعی مدل‌سازی کروموزوم برای حل این مشکلات به وجود می‌آیند و برای مقابله با این دو مشکل به‌صورت جداگانه عمل می‌کنند. نکته دیگر این است که این روش‌ها در یک زمینه خاص از یادگیری نظارت‌شده (به‌عنوان‌مثال متن‌کاوی) کار می‌کنند و یا به یک طبقه‌بندی خاص (مثلاًKNN) بستگی دارند. ما در اینجا از الگوریتم ژنتیک برای همزمانی انتخاب ویژگی و نمونه استفاده می‌کنیم و برای فرار از بهینه محلی از الگوریتم جستجوی ممنوع استفاده می‌کنیم و درنهایت برای طبقه‌بندی رایانامه‌ها از طبقه بند ماشین بردار پشتیبان استفاده خواهیم کرد در مورد انتخاب مجموعه داده‌ی ورودی باید به این نکته اشاره کرد که در بسیاری موارد متعادل نبودن نسبت ویژگی‌های در دست به‌کل اطلاعات موردنیاز در مورد هر کلاس باعث می‌رود کلاس بندهای معمولی در این مورد با شکست مواجه شوند. همچنین در بسیاری از مجموعه داده‌ها برخی از ویژگی‌ها در تصمیم‌گیری نقشی ندارند و به‌نوعی می‌توان آن‌ها را اضافی تلقی نمود.پس انتخاب یک زیرمجموعه‌ی مناسب از ورودی‌ها می‌تواند هم در دقت کلاس‌بندی و هم در سرعت آن مثمر ثمر باشد.

ما نیز در این پژوهش از الگوریتم ژنتیک استفاده می‌کنیم و در پایان با اعمال الگوریتم جستجوی ممنوع را بر روی بهترین کروموزوم به‌دست‌آمده دقت را افزایش می‌دهیم.

مراحل زیر برای رسیدن به هدف در پژوهش انجام می‌گیرد:

- تولید جمعیت اولیه با کدگذاری دودویی و محاسبه‌ی تابع ارزیابی جمعیت با استفاده از آموزش و آزمایش ماشین بردار پشتیبان.
- تکرار مراحل زیر تا رسیدن شرایط خاتمه:
- انجام عمل تقاطع و جهش و محاسبهٔ تابع ارزیابی جمعیت جدید.
- اعمال الگوریتم جستجوی ممنوع بر روی بهترین کروموزوم و بررسی تابع ارزیابی کروموزوم جدید
- انتخاب جمعیت جدید
- اعمال طبقه بند SVM بر روی کروموزوم بهینه

۸-۲- دادگان مورد استفاده

آزمایش‌ها بر دو مجموعه داده مختلف انجام شده است؛ مجموعه دادگان اول شامل: مجموعه داده لینگ اسپم[2] که مبتنی بر رایانامه‌های هرزنامه و مخرب (مانند تبلیغات، توسعه سئوی سایت‌ها، جذب اطلاعات کاربران برای فروش به سایر تبلیغاتچی‌ها و جمع‌آوری بانک اطلاعات) است.

مجموعه دادگان دوم نیز شامل: مجموعه داده‌های یوسی‌آی[3] که مبتنی بر رایانامه‌های شبیه‌سازی فیشینگ با استفاده از منبع یادگیری ماشین است که با آخرین متدهای فیشینگ آموزش داده شده است.

برای مجموعه لینگ اسپم، ۲۴۱۲ رایانامه صحیح و ۴۸۱ هرزنامه در دسترس بوده است. مجموعه داده یوسی‌آی، شامل ۴۶۰۱ نمونه است که ۱۸۱۳ نمونه با

[1] Matlab

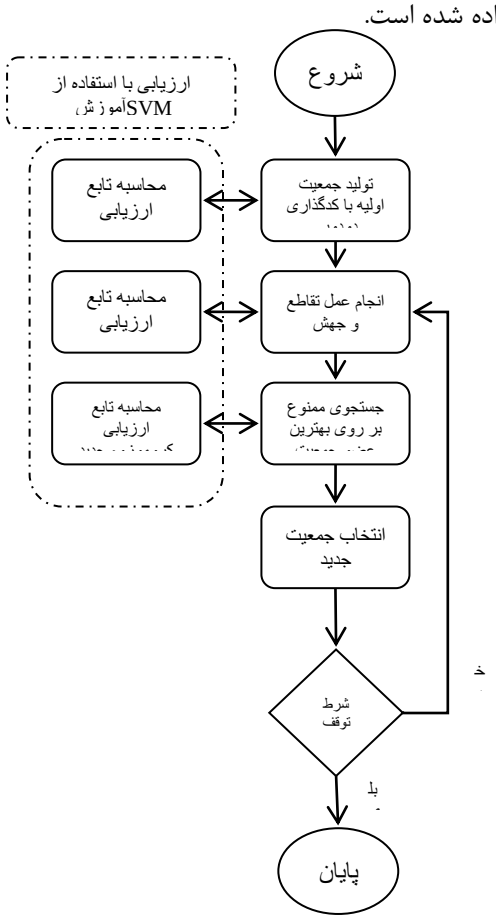
[2] Ling-Spam

[3] UCI

[Downloaded from monadi.isc.org.ir on 2022-08-19]

برچسب اسپم و ۲۷۸۸ نمونه با برچسب صحیح در این مجموعه وجود دارد. هر نمونه با ۵۸ صفت توصیف‌شده است [۶۷].

در شکل (۳)، فلوچارت الگوریتم پیشنهادی، نشان داده شده است.



(شکل-۳): فلوچارت الگوریتم پیشنهادی

۸-۳- معیارهای ارزیابی

برای سنجیدن عملکرد فیلتر کردن اسپم، از معیارهای بازخوانی اسپم[1]، دقت اسپم [2]استفاده‌شده است. این معیارها با صورت زیر تعریف‌شده‌اند:

$$spam\,recall = \frac{n_{s\to s}}{n_{s\to s} + n_{s\to L}}, \quad (۱)$$

$$spam\ precision = \frac{n_{s\to s}}{n_{s\to s} + n_{L\to s}} \quad (۲)$$

$n_{s\to s}$، نشان‌دهنده تعداد رایانامه‌های طبقه‌بندی‌شده به‌عنوان اسپم است و برچسب حقیقی آن‌ها اسپم است، $n_{L\to S}$ نشان‌دهنده تعداد رایانامه‌های با برچسب حقیقی صحیح است که به‌عنوان اسپم طبقه‌بندی‌شده‌اند و $n_{S\to L}$ تعداد رایانامه‌هایی با برچسب حقیقی اسپم را نشان می‌دهد که به‌عنوان رایانامه صحیح طبقه‌بندی‌شده‌اند.

معیار دقت به‌صورت زیر تعریف می‌شوند:

$$accuracy = \frac{n_{L\to L} + n_{S\to S}}{N_L + N_S} \quad (۳)$$

منظور از $n_{L\to L}$، تعداد رایانامه‌های که به‌صورت صحیح به‌عنوان رایانامه صحیح طبقه‌بندی‌شده‌اند، N_L و N_S تعداد رایانامه‌های صحیح و تعداد هرزنامه‌ها است که باید با فیلترینگ اسپم، طبقه‌بندی شوند.

در جدول (۱) پارامترهای مورد استفاده در الگوریتم پیشنهادی، مشاهده می‌شود:

جدول(۱): پارامترهای الگوریتم پیشنهادی

درصد داده‌های آزمایش	۹۰٪
درصد داده‌های آزمون	۱۰٪
تعداد تکرار الگوریتم	۱۰۰
تعداد جمعیت	۲۰
نرخ تقاطع	۷۰٪
نرخ جهش	۳٪
تعداد تکرار جستجوی ممنوع	۱۰
تعداد همسایه‌ها	۵

از صحت به‌عنوان معیار ارزیابی الگوریتم پیشنهادی استفاده می‌شود در جدول (۱) سه معیار صحت، دقت، بازخوانی و زمان سه الگوریتم مختلف در آزمایش بر روی دادگان یوسی‌آی دیده می‌شود.

۹- یافته‌ها

پس از انجام آزمایش، یافته‌های به‌دست آمده در جدول (۲)، با چهار پارامتر صحت، دقت، بازخوانی و زمان در دو مجموعه داده لاین اسپم و یوسی‌آی سنجیده شدند.

(جدول-۲): عملکرد الگوریتم‌ها بر روی دادگان یوسی‌آی

دادگان UCI				
نوع الگوریتم	**صحت**	**دقت**	**بازخوانی**	**زمان (s)**
ماشین بردار پشتیبان بدون انتخاب ویژگی و نمونه	۷۹٪/۵	۷۶٪/۲۶	۶۸٪/۶۷	۸۴
الگوریتم ژنتیک	۹۲٪/۱۶	۹۱٪/۳۰	۸۶٪/۳۰	۵۷۰
الگوریتم ژنتیک و جستجوی ممنوع	۹۷٪/۲۸	۹۷٪/۱۴	۹۳٪/۹۱	۷۸۲
دادگان Linspam				
	صحت	**دقت**	**بازخوانی**	**زمان (s)**
ماشین بردار پشتیبان بدون انتخاب ویژگی و	۸۱٪/۳	۷۹٪/۰۸	۷۰٪/۴۳	۶۸

[1] Spam Recall
[2] Spam Precision

[Downloaded from monadi.isc.org.ir on 2022-08-19]

نمونه				
الگوریتم ژنتیک	۹۴٪/۴۲	۹۲٪/۴۶	۸۹٪/۴۸	۳۸۷
الگوریتم ژنتیک و جستجوی ممنوع	۹۸٪/۲۲	۹۶٪/۸۸	۹۷٪/۱۲	۴۸۸

باتوجه به جدول بالا مشخص شد که میزان دقت، صحت و بازخوانی در الگوریتم‌های ترکیبی ژنتیک و جستجوی ممنوع در بالاترین حد ممکن قرار دارد و نشان دهنده این است که شناسایی هرزنامه‌ها و یا رایانامه‌های فیشینگ در مجموعه دادگان لاین اسپم، با دقت و صحت بالاتری به‌دست آمده است و در نهایت ترکیب الگوریتم ژنتیک و جستجوی ممنوع، می‌تواند بیشترین بازدهی رو داشته باشد. درپایان، نیز الگوریتم کاهش داده پیشنهادی را با سایر روش‌های کاهش داده مقایسه شده است که تمام این مراجع برای دسته‌بندی رایانامه‌ها از طبقه بندSVM و مجموعه داده‌ی اسپم گرفته‌شده از سایت UCI استفاده شده است که نتیجه بدست کارایی و دقت بالای به‌دست‌آمده از روش پیشنهادی را ثابت می‌کند که نتیجه آن، در جدول (۳)، نشان داده شده است:

(جدول-۳): مقایسه روش پیشنهادی با روش‌های مبتنی بر SVM بر روی دادگان UCI در سایر مقالات دیگر.

طبقه‌بند استفاده‌شده	نوع کاهش داده	دقت(Acc)
Svm[68]	انتخاب ویژگی با *CFS*	۹۱/۴۴
Svm[68]	انتخاب ویژگی با *Chi*	۹۳/۰۰
Svm[68]	انتخاب ویژگی با *IG*	۹۳/۰۰
Svm[68]	انتخاب ویژگی با *GR*	۹۳/۳۹
Svm[68]	انتخاب ویژگی با *SU*	۹۳/۳۳
Svm[68]	انتخاب ویژگی با One R	۹۲/۶۵
Svm[68]	انتخاب ویژگی با Relief	۹۳/۱۵
Svm[68]	انتخاب ویژگی با Lda	۹۱/۹۰
Svm[68]	انتخاب ویژگی با Rpart	۹۰/۵۱
Svm[68]	انتخاب ویژگی با SVM	۸۵/۹۵
Svm[68]	انتخاب ویژگی با RF	۹۱/۲۳
Svm[68]	انتخاب ویژگی با NB	۸۰/۰۰
Svm[69]	انتخاب ویژگی با *ACO*	۸۱/۲۵
Svm[70]	انتخاب ویژگی Best-First	۸۶/۵۴
Svm	**روش پیشنهادی**	**۹۷/۲۸**

۱۰- بحث و نتیجه‌گیری

فیشینگ و انواع حملات فیشینگی در سال‌های اخیر موجب شده است بسیاری از افراد و مشتریان بانک‌ها درگیر کلاهبرداری‌های گسترده‌ای شوند و سارقین و افراد سودجو بتوانند به دلیل ناآگاهی، فریب یا شبیه‌سازی رایانامه‌ها به رایانامه‌های بانکی، بسیاری را در دام خود گرفتار کنند. در این شرایط وظیفه پلیس فتا، انجام اقداماتی در جهت پیشگیری و مقابله با فیشینگ است که یکی از مهم‌ترین آن‌ها، رایانامه‌های فیشینگ است، از آنجایی که شناسایی موردی این رایانامه‌ها موجب اتلاف زمان و یا خطای انسانی می‌شود، این پژوهش با هدف شناسایی هوشمند و خودکار رایانامه‌های فیشینگ و هرزنامه‌ها با میزان دقت و صحت بالا، به دنبال ارائه روشی ترکیبی حاصل از الگوریتم ژنتیک و جستجوی ممنوع و سپس بهینه‌سازی آن با استفاده از ماشین بردار پشتیبان بود. در همین راستا با استفاده از دادگان استاندارد یوسی‌آی و لاین‌اسپم، واقع در مقالات پیشین، الگوریتم پیشنهادی، پیاده‌سازی شد. نتایج مبین این قضیه بود که نتایج الگوریتم پیشنهادی از منظر صحت شناسایی رایانامه‌های هرزنامه و فیشینگ، برای دادگان یوسی‌آی برابر با ۹۷/۲۸ درصد و برای لاین‌اسپم برابر با ۹۸/۲۲ است که این مقدار در بین کلیه مقالات مشابه خصوصا مقاله [۶۸-۷۰] که به‌صورت اختصاصی از الگوریتم‌های مشابه استفاده کرده بودند، بیشتر است و می‌توان ادعا کرد که بهترین نتایج موجود در این زمینه بوده است.

۱۱- پیشنهادات

در راستای نتایج به‌دست‌آمده به شرکت‌های توسعه دهنده زیرساخت‌های بومی فناوری اطلاعات پیشنهاد می‌گردد، با استفاده از الگوریتم پیشنهاد شده در این پژوهش، افزونه‌هایی جهت شناسایی رایانامه‌های مخرب با ویژگی‌های بومی فیشینگ طراحی و توسعه دهند تا بتواند با اجرا بر بستر درگاه‌های رایانامه‌های داخلی و خارجی، هرزنامه‌ها و فیشینگ‌های بومی را شناسایی کند و بصورت خودکار آدرس فرستنده را به پلیس فتا گزارش کند.

همچنین پیشنهاد می‌شود پلیس فتا با توسعه این الگوریتم به عنوان یک نرم‌افزار و یا افزونه قابل نصب بر روی کلیه بسترهای رایانامه داخلی و خارجی، و قرار دادن فایل نصب آن در سایت پلیس فتا، بتوانند زمینه ساز امنیت اطلاعات کاربران داخلی شوند.

۱۲- مراجع

[1] X. Fontes, D. Aparício, M. I. Silva, B. Malveiro, J. T. Ascensão, and P. Bizarro, "Finding NeMo: Fishing in banking networks using network motifs," *arXiv preprint arXiv:2108.04494*, 2021.

[Downloaded from monadi.isc.org.ir on 2022-08-19]

[2] O. Ilchenko, V. Chumak, S. Kuzmenko, O. Shelukhin, and A. Dobrovinskyi, "Fishing as a cybercrime in the Internet banking system: economic and legal aspects," 2019.

[3] P. Mowar and M. Jain, "Fishing out the Phishing Websites," in *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 2021: IEEE, pp. 1-6.

[4] Y. Kwak, S. Lee, A. Damiano, and A. Vishwanath, "Why do users not report spear phishing emails?," *Telematics and Informatics*, vol. 48, p. 101343, 2020.

[5] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345-357, 2019.

[6] L. Chen, "Research on Anti-phishing Strategy of Smart Phone," in *Journal of Physics: Conference Series*, 2019, vol. 1314, no. 1: IOP Publishing, p. 012172.

[7] T. Lin *et al.*, "Susceptibility to spear-phishing emails: Effects of internet user demographics and email content," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 5, pp. 1-28, 2019.

[8] H. Gascon, S. Ullrich, B. Stritter, and K. Rieck, "Reading between the lines: content-agnostic detection of spear-phishing emails," in *International Symposium on Research in Attacks, Intrusions, and Defenses*, 2018: Springer, pp. 69-91.

[9] M. J. Miranda, "Enhancing cybersecurity awareness training: A comprehensive phishing exercise approach," *International Management Review*, vol. 14, no. 2, pp. 5-10, 2018.

[10] M. Bossetta, "The weaponization of social media: Spear phishing and cyberattacks on democracy," *Journal of international affairs*, vol. 71, no. 1.5, pp. 97-106, 2018.

[11] I. Qabajeh, F. Thabtah, and F. Chiclana, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," *Computer Science Review*, vol. 29, pp. 44-55, 2018.

[12] R. Röpke, U. Schroeder, V. Drury, and U. Meyer, "Towards Personalized Game-Based Learning in Anti-Phishing Education," in *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, 2020: IEEE, pp. 65-66.

[13] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Computers & Security*, vol. 94, p. 101862, 2020.

[14] D.-Y. Kim, "A Study on Voice Phishing Countermeasures of the Police," *Journal of Digital Contents Society*, vol. 19, no. 1, pp. 193-198, 2018.

[15] S. S. M. M. Rahman, F. B. Rafiq, T. R. Toma, S. S. Hossain, and K. B. B. Biplob, "Performance assessment of multiple machine learning classifiers for detecting the phishing URLs," in *Data Engineering and Communication Technology*: Springer, 2020, pp. 285-296.

[16] A. Akhtar, G. R. Tahir, and K. Shakeel, "A mechanism to detect urdu Spam emails," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2017: IEEE, pp. 168-172.

[17] N. Govil, K. Agarwal, A. Bansal, and A. Varshney, "A machine learning based spam detection mechanism," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020: IEEE, pp. 954-957.

[18] N. Kumar and S. Sonowal, "Email Spam Detection Using Machine Learning Algorithms," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020: IEEE, pp. 108-113.

[19] H. Yang, Q. Liu, S. Zhou, and Y. Luo, "A spam filtering method based on multi-modal fusion," *Applied Sciences*, vol. 9, no. 6, p. 1152, 2019.

[20] A. Shrivastava and R. Dubey, "Review of spam classification using different machine learning algorithms."

[21] M. A. Mohammed *et al.*, "An anti-spam detection model for emails of multi-natural language," *Journal of Southwest Jiaotong University*, vol. 54, no. 3, 2019.

[22] S. Abdulla and A. Altaher, "IHASS: Intelligent and Hybrid Anti-Spam System," *International Journal of Computer Networks and Communications Security*, vol. 5, no. 6, p. 115, 2017.

[23] J. Soyemi and M. Hammed, "Detection and Classification of Legitimate and Spam Emails using K-Neareset," *International Journal of Computer Applications*, vol. 175, no. 18, pp. 28-32, 2020.

[Downloaded from monadi.isc.org.ir on 2022-08-19]

[24] Y. Alamlahi and A. Muthana, "An Email Modelling Approach for Neural Network Spam Filtering to Improve Score-based Anti-spam Systems," *International Journal of Computer Network and Information Security*, vol. 11, no. 12, p. 1, 2018.

[25] A. Subasi, S. Alzahrani, A. Aljuhani, and M. Aljedani, "Comparison of decision tree algorithms for spam e-mail filtering," in *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, 2018: IEEE, pp. 1-5.

[26] S. Sagar, S. Vishwanatha, T. Vishwas, and G. Prabhukumar, "Spam Message Filter Using Naive Bayes Classifier And Intelligent Text Modification Method," CMR Institute of Technology. Bangalore, 2020.

[27] D. Quick and K.-K. R. Choo, "Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive," *Trends and Issues in Crime and Criminal Justice*, no. 480, pp. 1-11, 2014.

[28] I. Czarnowski and P. Jędrzejowicz, "Data reduction algorithm for machine learning and data mining," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2008: Springer, pp. 276-285.

[29] Z. Sun and Y. Huo, "The spectrum of big data analytics," *Journal of Computer Information Systems*, vol. 61, no. 2, pp. 154-162, 2021.

[30] C. Ricciardi *et al.*, "Using gait analysis' parameters to classify Parkinsonism: A data mining approach," *Computer methods and programs in biomedicine*, vol. 180, p. 105033, 2019.

[31] J. W. Grzymala-Busse, "Data reduction: discretization of numerical attributes," in *Handbook of data mining and knowledge discovery*, 2002, pp. 218-225.

[32] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern recognition letters*, vol. 27, no. 5, pp. 414-423, 2006.

[33] J. Wang, P. Wonka, and J. Ye, "Scaling SVM and least absolute deviations via exact data reduction," in *International conference on machine learning*, 2014: PMLR, pp. 523-531.

[34] Z. S. Ageed *et al.*, "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 29-38, 2021.

[35] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Artificial Intelligence Review*, vol. 11, no. 1, pp. 273-314, 1997.

[36] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on computers*, vol. 26, no. 09, pp. 917-922, 1977.

[37] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012.

[38] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[39] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm and Evolutionary Computation*, vol. 54, p. 100663, 2020.

[40] A. Zollanvari, A. P. James, and R. Sameni, "A theoretical analysis of the peaking phenomenon in classification," *Journal of Classification*, vol. 37, no. 2, pp. 421-434, 2020.

[41] C. Sima and E. R. Dougherty, "The peaking phenomenon in the presence of feature-selection," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1667-1674, 2008.

[42] S. Mirjalili, "Genetic algorithm," in *Evolutionary algorithms and neural networks*: Springer, 2019, pp. 43-55.

[43] H. Liang, J. Zou, K. Zuo, and M. J. Khan, "An improved genetic algorithm optimization fuzzy controller applied to the wellhead back pressure control system," *Mechanical Systems and Signal Processing*, vol. 142, p. 106708, 2020.

[44] R. Guha *et al.*, "Deluge based Genetic Algorithm for feature selection," *Evolutionary intelligence*, vol. 14, no. 2, pp. 357-367, 2021.

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

[46] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm," *Journal of digital imaging*, vol. 31, no. 4, pp. 477-489, 2018.

[47] A. Diker, D. Avci, E. Avci, and M. Gedikpinar, "A new technique for ECG signal classification genetic algorithm Wavelet

[Downloaded from monadi.isc.org.ir on 2022-08-19]

Kernel extreme learning machine," *Optik*, vol. 180, pp. 46-55, 2019.

[48] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091-8126, 2021.

[49] J. Genlin, "Survey on genetic algorithm," *Computer Applications and Software*, vol. 2, no. 1, pp. 69-73, 2004.

[50] X. Xue and J. Chen, "Using compact evolutionary tabu search algorithm for matching sensor ontologies," *Swarm and Evolutionary Computation*, vol. 48, pp. 25-30, 2019.

[51] M. Alinaghian, E. B. Tirkolaee, Z. K. Dezaki, S. R. Hejazi, and W. Ding, "An augmented Tabu search algorithm for the green inventory-routing problem with time windows," *Swarm and Evolutionary Computation*, vol. 60, p. 100802, 2021.

[52] M. Ebrahimi, M.-H. Khoshtaghaza, S. Minaei, and B. Jamshidi, "Vision-based pest detection based on SVM classification method," *Computers and Electronics in Agriculture*, vol. 137, pp. 52-58, 2017.

[53] S. Ketu and P. K. Mishra, "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare," *Complex & Intelligent Systems*, pp. 1-19, 2021.

[54] X. Wang, D. Fu, Y. Wang, Y. Guo, and Y. Ding, "The XGBoost and the SVM-based prediction models for bioretention cell decontamination effect," *Arabian Journal of Geosciences*, vol. 14, no. 8, pp. 1-11, 2021.

[55] P. Liu, Y. Han, and C. Dong, "Research on Pipeline Defect Classification Based on SVM," in *Journal of Physics: Conference Series*, 2021, vol. 1944, no. 1: IOP Publishing, p. 012018.

[56] M. Achite, P. Tsangaratos, I. Ilia, and A. K. Toubal, "Applying support vector machines optimized by genetic algorithm for estimating the spatial distribution of mean annual precipitation," *Arabian Journal of Geosciences*, vol. 14, no. 8, pp. 1-16, 2021.

[57] L. I. Kuncheva and L. C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," *Pattern recognition letters*, vol. 20, no. 11-13, pp. 1149-1156, 1999.

[58] A. Rozsypal and M. Kubat, "Selecting representative examples and attributes by a genetic algorithm," *Intelligent Data Analysis*, vol. 7, no. 4, pp. 291-304, 2003.

[59] D. Fragoudis, D. Meretakis, and S. Likothanassis, "Integrating feature and instance selection for text classification," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 501-506.

[60] J. T. De Souza, R. A. F. Do Carmo, and G. A. L. De Campos, "A novel approach for integrating feature and instance selection," in *2008 International Conference on Machine Learning and Cybernetics*, 2008, vol. 1: IEEE, pp. 374-379.

[61] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671-680, 1983.

[62] B. Sierra, E. Lazkano, I. Inza, M. Merino, P. Larranaga, and J. Quiroga, "Prototype selection and feature subset selection by estimation of distribution algorithms. a case study in the survival of cirrhotic patients treated with tips," in *Conference on artificial intelligence in medicine in Europe*, 2001: Springer, pp. 20-29.

[63] J.-H. Chen, H.-M. Chen, and S.-Y. Ho, "Design of nearest neighbor classifiers: multi-objective approach," *International Journal of Approximate Reasoning*, vol. 40, no. 1-2, pp. 3-22, 2005.

[64] F. Ros, S. Guillaume, M. Pintore, and J. R. Chretien, "Hybrid genetic algorithm for dual selection," *Pattern Analysis and Applications*, vol. 11, no. 2, pp. 179-198, 2008.

[65] H. Ishibuchi and T. Nakashima, "Multi-objective pattern and feature selection by a genetic algorithm," in *Proceedings of the 2nd annual conference on genetic and evolutionary computation*, 2000, pp. 1069-1076.

[66] C.-F. Tsai, W. Eberle, and C.-Y. Chu, "Genetic algorithms in feature and instance selection," *Knowledge-Based Systems*, vol. 39, pp. 240-247, 2013.

[67] R. Hassanpour, E. Dogdu, R. Choupani, O. Goker, and N. Nazli, "Phishing e-mail detection by using deep learning algorithms," in *Proceedings of the ACMSE 2018 Conference*, 2018, pp. 1-1.

[68] R. Parimala and R. Nallaswamy, "A study of spam e-mail classification using feature selection package," *Global Journal of Computer Science and Technology*, 2011.

[69] R. Karthika and P. Visalakshi, "A hybrid ACO based feature selection method for email spam classification," *WSEAS Trans. Comput*, vol. 14, no. 2015, pp. 171-177, 2015.

[70] M. Rathi and V. Pareek, "Spam mail detection through data mining-A comparative performance analysis," *International Journal of Modern Education and Computer Science*, vol. 5, no. 12, p. 31, 2013.

[Downloaded from monadi.isc.org.ir on 2022-08-19]

[Downloaded from monadi.isc.org.ir on 2022-08-19]