سال دوازدهم، شماره‌های ۴۳ و ۴۴، بهار و تابستان ۱۳۹۹
صفحات: ۲۷_۳۸

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

بهبود روش شناسایی وب سایت فیشینگ با استفاده از داده‌کاوی روی صفحات وب

مهدیه بهارلو* علیرضا یاری**

*دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران
**استادیار پژوهشگاه ارتباطات و فناوری اطلاعات

تاریخ دریافت: ۱۳۹۹/۰۲/۰۱ تاریخ پذیرش: ۱۳۹۹/۰۷/۳۰
نوع مقاله: پژوهشی

چکیده

فیشینگ یک نوع حمله اینترنتی در سطح وب است که هدف آن سرقت مشخصات فردی کاربران برای سرقت آنلاین است. فیشینگ دارای اثر منفی در از بین بردن اعتماد بین کاربران در کسب‌وکارهای الکترونیکی است؛ بنابراین در این تحقیق سعی بر بررسی روشهای تشخیص وب‌سایت‌های فیشینگ با استفاده از داده کاوی شده‌است. شناسایی ویژگی‌های اصلی از صفحات وب فیشینگ یکی از پیش‌شرط‌های مهم در طراحی یک سیستم تشخیص فیشینگ دقیق است. در پژوهش حاضر، برای افزایش کارایی سامانه تشخیص فیشینگ، یک روش ترکیبی برای کاهش ویژگی‌های وب سایت‌های فیشنگ پیشنهاد شده است. این روش ترکیبی از روش‌های انتخاب و کاهش ویژگی است که در دو مرحله انجام می‌شود. برای پیاده سازی و ارزیابی این روش پیشنهادی، بعد از کاهش ویژگیها دسته بندی داده ها از طریق روش‌های درخت تصمیم‌گیری J48، جنگل تصادفی و بیزین ساده موردبررسی قرار گرفت. نتایج نشان می‌دهند دقت مدل ایجاد شده برای تعیین وب سایت‌های فیشینگ با استفاده از کاهش ویژگی دومرحله‌ای مبتنی بر پوششی و الگوریتم تحلیل مؤلفه اصلی در روش جنگل تصادفی که به میزان ۹۶٫۵۸٪ است، نسبت به سایر روش‌ها نتیجه مناسب‌تری را دارد.

واژگان کلیدی: حملات اینترنتی، فیشینگ، داده‌کاوی، انتخاب ویژگی، استخراج ویژگی

۱- مقدمه

امروزه اینترنت به یک جزء ضروری از زیرساخت‌های اجتماعی و اقتصادی روزمره مردم تبدیل‌شده است. حجم بالای اطلاعات محرمانه و امنیتی اینترنت باعث می‌شود انواع تهدیدات و حملات مختلف در آن به وجود آید که ممکن است باعث خسارت مالی، سرقت هویت، از دست دادن اطلاعات خصوصی، آسیب شهرت نام تجاری و از دست دادن اعتماد مشتریان در تجارت الکترونیک شود [۱]. عواملی که

نویسنده مسئول: علیرضا یاری a_yari@itrc.ac.ir

تهدید و حمله را در یک شبکه اینترنتی به وجود می‌آورند عبارتند از: دسترسی بدون محدودیت به اینترنت، گمنامی افراد، سرعت‌بالای انتشار، عدم ارتباط چهره به چهره، دسترسی آزاد به خدمات و محتویات ارزشمند و همچنین عدم وجود قوانین و توافقات مناسب [۲]؛ بنابراین، مناسب بودن اینترنت به‌عنوان یک کانال برای انجام معاملات تجاری مطرح می‌شود.

در اوایل ۱۹۹۰، با محبوبیت رو به رشد اینترنت، ما شاهد تولد یک نوع جدید از جرائم اینترنتی بودیم؛ که فیشینگ نام دارد [۳].

برخلاف سایر روش‌های هک و ورود به سیستم، در روش فیشینگ معمولاً هیچ نفوذی انجام‌نشده و از رخنه‌ها و آسیب‌پذیری‌ها استفاده نمی‌شود. بلکه خود کاربر است که با استفاده از روش‌های گوناگون فریب‌خورده و اطلاعاتی نظیر نام کاربری، رمز عبور، اطلاعات حساب بانکی را در اختیار حمله‌کننده که به‌اصطلاح فیشر نامیده می‌شود، قرار می‌دهد [۴]. طبق بررسی انجام‌شده توسط "گروه کاری ضد فیشینگ" تعداد حملات فیشینگ سراسر جهان در سه ماهه چهارم سال ۲۰۱۹ کاهش یافته و به میانگین نزدیکتر شده است. البته در همین سال در کشور برزیل تا ۲۳۲ درصد افزایش یافته است. حملات فیشینگ که کاربران وب، ایمیل و سرویسهای نرم افزاری را هدف قرار می‌دهد، همچنان بزرگترین گروه حملات فیشینگ است. تقریباً سه چهارم از همه سایت‌های فیشینگ اکنون از حفاظت SSL استفاده می‌کنند، بالاترین میزان ثبت شده از اوایل سال ۲۰۱۵، و این نشانگر این است که کاربران نمی‌توانند به تنهایی به SSL اعتماد کنند و برای درک درست نیاز به ویژگیهای بیشتری دارند [۵].

برای وب‌سایت فیشینگ تعاریف زیادی ارائه شده است که می‌توان تمام تعاریف را به صورت جامع و کامل در یک جمله بیان نمود: "وب‌سایت فیشینگ عمل ایجاد یک کپی از یک وب‌سایت قانونی و استفاده از مهارت‌های اجتماعی برای فریب قربانی برای ارسال اطلاعات شخصی او است" [۳].

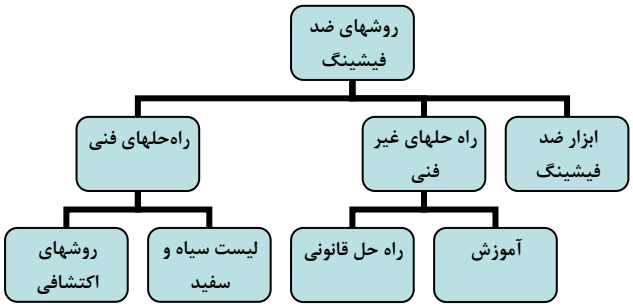
انواع حملات فیشینگ را می‌توان فیشینگ سرنیزه یا فیشینگ هدفمند، کلون فیشینگ، صید نهنگ، فیشینگ تلفن، حمله بایوزی، قاپیدن تب، فارمینگ، فیشینگ موتور جستجو و فیشینگ مبتنی بر بدافزار دسته بندی کرد [۶،۷].

برای تشخیص صفحات فیشینگ از روش‌های متعددی استفاده می‌شود که هریک به دسته‌ای از ویژگی‌های صفحات می‌پردازند. در این مقاله روشی جدیدی پیشنهادشده است که برای دسته‌بندی داده‌های وب‌سایت‌های فیشینگ، از کاهش دومرحله‌ای ویژگی‌ها استفاده می‌نماید. بدین‌صورت که در مرحله اول بر اساس یک روش انتخاب ویژگی، زیرمجموعه‌ای مفید از ویژگی‌ها انتخاب شده و درمرحله دوم با کاهش بیشتر ترکیبی از ویژگی‌های مفید با ابعاد کمتر حاصل می‌شود.

در مقاله جاری، در ابتدا ابزارها و روش‌های مقابله با فیشینگ در قسمت پیشینه تحقیقات معرفی خواهند شد. سپس در بخش ۳ روش انجام کار، ویژگی‌های که توسط فیشرها برای ایجاد وب سایت‌های جعلی استفاده می‌شوند و همچنین روش‌های کاهش ویژگی و روش پیشنهادی کاهش ویژگی در این مقاله معرفی می‌گردد. در ادامه در بخش ۴ داده‌های حاصل از انجام آزمایشات مورد تجزیه و تحلیل قرار می‌گیرند و نتیجه با کارهای مشابه گذشته مورد مقایسه قرار می‌گیرد. در نهایت در بخش نتیجه گیری، نتیجه حاصل از انجام تحقیق بررسی خواهد شد.

۲- پیشینه پژوهش

تکنیک‌های فیشینگ روزبه‌روز در حال افزایش است و درعین‌حال پیچیده‌تر می‌شود. درنتیجه نیاز فوری به یافتن راه‌حل‌های مناسب برای مبارزه با حملات فیشینگ وجود دارد. تاکنون، راه‌حل‌های مختلفی در پاسخ به حملات فیشینگ پیشنهادشده است. این راه‌حل‌ها با توجه به شکل (۱) به سه روش مقابله با فیشینگ شامل ابزارهای ضد فیشینگ، راه‌حلهای فنی و غیر فنی تقسیم می‌شوند[۸]. ازجمله ابزار ضد فیشینگ می‌توان افزونه مرورگر WOT [۸]، سایت ضد فیشینگ فیش تانک [۹]، سازمان جیو تراست، موزیلا تاندر برد [۱۰] را نام برد؛ اما در ادامه به توضیح راه حل‌های فنی و غیر فنی خواهیم پرداخت.



شکل (۱): دسته‌بندی روش‌های شناسایی فیشینگ

۱-۲- راه حل‌های غیرفنی

۱-۱-۲- آموزش

آموزش مشتریان برای افزایش سطح آگاهی آن‌ها از جرائم آنلاین ضروری است تا با درک شاخص‌های امنیتی داخل وب‌سایت‌ها بتوانند به مقابله با فیشینگ بپردازند [۱۱]. اما این کار بسیار مشکل است چراکه کاربران باید زمان زیادی را صرف فراگیری متدهای فیشینگ کنند، علاوه بر آن فیشرها در ایجاد تکنیک‌های جدید هوشمندانه‌تر عمل می‌کنند [۳].

۲-۱-۲- راه حل‌های قانونی

این راه‌حل‌ها در کشورهای زیادی استفاده شده‌اند. ایالت متحده اولین کشوری بود که قانونی را در مورد این وب‌سایت‌ها وضع کرد و همین امر موجب گردید، فعالیت‌های زیادی توقیف گردد و عاملین آن به زندان انداخته شوند. بااین‌حال، راه‌حل‌های قانونی نمی‌توانند مانع عمل وب‌سایت‌های جعلی شوند، چراکه ردیابی آن‌ها به دلیل مخفی شدن سریع آن‌ها در دنیای سایبری مشکل است [۱].

۲-۲- راه حل‌های فنی

علاوه بر روشهای غیر فنی نظیر آموزش و قانون، برای مبارزه با فیشینگ، عموماً، دو متد فنی رایج در مبارزه با حملات فیشینگ، لیست سیاه‌وسفید و روش‌های اکتشافی است [۶].

۱-۲-۲- لیست سیاه‌وسفید

در روش لیست سیاه و سفید تمام URL های درخواستی با URL های موجود در لیست سیاه و سفید فیشینگ مقایسه می‌شوند تا مشخص شود وب‌سایت موردنظر قانونی است یا جعلی؛ لیست سفید در واقع لیست صفحاتی هست که قانونی هستند و اما در مقابل در لیست سیاه سعی می‌شود تمام صفحات جعلی شناسایی و پوشش داده شود. در این روش شناسایی تمام صفحات و به‌روز رسانی اطلاعات کار بسیار دشوار و زمان بر هست، چراکه بطور دائم صفحات زیادی ایجاد شده و یا از بین می‌روند [۱۲].

برخی از ارائه‌دهندگان لیست سیاه مانند کاوش ایمن گوگل [۱]، فیش نت، لیست سیاه مبتنی بر DNS [۶] و نرم‌افزارهای ضد فیشینگ نت‌کرفت، وب‌سن و کلودمارک [۱۳] با استفاده از این روش مانع از حملات فیشینگ می‌شوند.

۲-۲-۲- روش‌های اکتشافی

در این روش از ویژگی‌های وب‌سایت‌های فیشینگ برای تشخیص وب‌سایتهای فیشنیگ جدید استفاده می‌شود. در واقع قدرت اصلی تشخیص‌دهنده‌های فیشینگ بر اساس روش اکتشافی این است که آن‌ها قادر به تشخیص وب‌سایت‌های فیشینگ جدید هستند [۱۰].

۱-۲-۲-۲-روش‌های مبتنی بر الگوریتم‌های فازی

یک روش که توسط ابارس و همکاران در سال ۲۰۱۰ به‌کار گرفته‌شده است، بر مبنای الگوریتم‌های دسته‌بندی مبتنی بر قانون فازی برای تشخیص وب‌سایت‌های فیشینگ بانکداری الکترونیکی است [۱۳]. سعیدی در سال ۱۳۹۴ از روش دسته‌بندی مجموعه‌های فازی و روش ترکیبی تصمیم‌گیری AHP_TOPSIS برای تشخیص سریع‌تر و کارآمدتر وب‌سایت‌های فیشینگ در بانکداری الکترونیکی استفاده کرده است [۱۴]. عبدالحمید و همکاران در سال ۲۰۱۴ متد خاص دسته‌بندی انجمنی به نام دسته‌بند چند برچسبه بر مبنای دسته‌بندی انجمنی را ارائه دادند [۱]. هادی و همکاران در سال ۲۰۱۶ از روش جدید دسته‌بندی انجمنی به نام الگوریتم دسته‌بندی انجمنی سریع استفاده کرده‌اند [۱۵].

۲-۲-۲-۲- روش‌های مبتنی بر یادگیری ماشین

باسنت و همکاران در سال ۲۰۱۲ از الگوریتم‌های یادگیری ماشین ازجمله بیزین ساده، رگرسیون لجستیک و جنگل

تصادفی استفاده کردند. باتوجه به آنکه از ۱۷۷ ویژگی اولیه استفاده شده بود آنها از روش انتخاب ویژگی مبتنی بر همبستگی (CFS)[۱]و پوششی برای کاهش ویژگی‌ها استفاده کردند [۱۶]. خنجی و همکاران بر کاهش ویژگی‌ها و یافتن بهترین زیرمجموعه از آن‌ها برای رسیدن به‌دقت بهتر در دسته‌بندی ایمیل‌های فیشینگ تمرکز داشتند. آن‌ها چهار روش کاهش ویژگی بهره اطلاعات (IG)[۲]، پوششی، ConEval و CFS را با الگوریتم جنگل تصادفی مورد ارزیابی قراردادند [۱۷]. سینگ و همکاران در سال ۲۰۱۵ از چهار الگوریتم پیش‌پردازش CFS، IG، زیرمجموعه مبتنی بر سازگاری[۳] PCA برای کاهش ابعاد و پنج الگوریتم دسته‌بندی؛ J48، بیزین ساده، SVM، جنگل تصادفی و آدابوست برای ارزیابی عملکرد الگوریتم‌های کاهش ویژگی ازنظر دقت و AUC استفاده کردند [۱۸]. رحیمی و همکاران با روش ترکیبی انتخاب ویژگی مبتنی بر محتوا و رفتار و با استفاده از الگوریتم‌های شبکه بیزین، آدابوست، جنگل تصادفی و جدول تصمیم‌گیری، به تشخیص ایمیل‌های فیشینگ پرداختند [۱۹]. در مقاله [۲۰] نیز یک روش کاهش ویژگی ترکیبی هایبرید (HEFS) ارائه شده است که در کنار روش طبقه‌بندی جنگل تصادفی بهترین عملکرد را داشته است. در مقاله [۲۱] با استفاده از الگوریتم‌های یادگیری ماشین علاوه بر استفاده از یک مجموعه داده جدید مربوط به تشخیص فیشینگ که شامل ۵۰۰۰ صفحه وب قانونی و ۵۰۰۰ فیشینگ است، به این مسئله پرداخته که به منظور دستیابی به بهترین نتایج، الگوریتم‌های مختلف یادگیری ماشین مورد آزمایش قرار گرفتند و نهایتا J48، جنگل تصادفی و پرسپترون چند لایه انتخاب شدند. در مقاله [۲۲] نیز روشهای یادگیری ماشین بطور کلی در کنارکاهش ویژگی‌ها برای تشخیص صفحات وب و ایمیل فیشینگ مورد بررسی و ارزیابی قرار گرفته‌اند.

[۱] Correlation feature selection
[۲] Information Gain
[۳] Consistency Subset

۳-۲-۲-۲- روش‌های مبتنی بر متن‌کاوی

پاندی و راوی در سال ۲۰۱۲ از متن‌کاوی و داده‌کاوی برای تشخیص ایمیل‌های فیشینگ استفاده کردند. با استفاده از انتخاب ویژگی مبتنی بر t-statistic، به ۱۲ ویژگی مهم دست یافتند و از دسته‌بندهای SVM، رگرسیون لجستیک، جدول تصمیم‌گیری، MLP، MDH، PNN و برنامه‌نویسی ژنتیک استفاده نمودند، این الگوریتم‌ها با انتخاب ویژگی و بدون انتخاب ویژگی مورد ارزیابی قرار گرفتند [۲۳]. آن‌ها در تحقیق دیگری روش بهبود یافته‌ای ارائه داده و برای تست از اعتبار سنجی تقاطعی با ۱۰ تکرار استفاده کردند [۲۴].

۴-۲-۲-۲- روش‌های مبتنی بر تکنیک‌های ابتکاری

لنگری و عبدالرزاق نژاد در سال ۱۳۹۵، وب سایت‌ها را بر اساس ویژگی‌های استخراج‌شده توسط الگوریتم بهینه‌سازی صفحات شیب‌دار (IPO)[۴]به سه دسته وب‌سایت‌های قانونی، مشکوک و فیشینگ دسته‌بندی کردند [۲۵].

۳- روش کاهش ویژگی

در تمام مجموعه‌های داده که برای مدلسازی و آزمون سامانه‌های ضدفیشینگ استفاده می‌شوند، تعداد زیادی ویژگی در دسترس هستندکه امکان دسته‌بندی را فراهم می‌کنند. با توجه به اینکه تعداد محدودی از آن‌ها اثرات قابل‌توجهی بر متغیر هدف دارند؛ لذا به دنبال کاهش آنها به نحوی هستیم که کمترین تأثیر را در تشخیص فیشینگ بگذارد. در این بخش ویژگیهای وب‌سایتهای فیشینگ و نحوه کاهش آنها مورد بررسی قرار گرفته است.

۳-۱- ویژگی‌ها

هر وب سایت دارای یک سری ویژگی‌های منحصربه‌فرد است که باعث تمایز آن از سایر وب‌سایت‌ها می‌شود. لذا همواره فیشر به دنبال ترفندی برای ساخت یک وب‌سایت جعلی، اقدام به تغییر برخی از ویژگی‌های وب‌سایت‌های قانونی می‌نماید [۲۵]. در این بخش ویژگی‌های درج شده در

[۴] Inclined Planes System Optimization

مجموعه داده سایت داده‌کاوی UCI مورد بررسی قرار گرفته است. این مجموعه داده حاوی ۳۰ ویژگی مبتنی بر ۴ معیار زیر است[۲۶]:

- نوار آدرس
- غیرطبیعی بودن
- جاوا اسکریپت و HTML
- دامنه

در این قسمت با این ویژگی‌ها که در آزمایش‌ها مورد استفاده قرار گرفته‌اندآشنا می‌شویم. ویژگی‌ها به‌صورت صفر و یک و منفی یک کدگذاری شده‌اند.

۱-۱-۳- ویژگی‌های مبتنی بر نوار آدرس

- آدرس مبتنی بر IP
- URL های طولانی
- استفاده از خدمات کوتاه کننده URL
- استفاده از نماد @ در URL
- تغییر مسیر با "//"
- وجود نماد "-" در دامنه
- تعداد نقطه‌ها در دامنه
- HTTPS
- مدت ثبت دامنه
- فاوآیکون
- استفاده از درگاه غیراستاندارد
- وجود توکن HTTPS در دامنه

۲-۱-۳- ویژگی‌های مبتنی بر غیرطبیعی بودن

- درخواست URL
- تگ انکر
- پیوندها در تگ‌های <Meta>, <Script> و<Link>
- فرم هندلر در سرور
- ارسال اطلاعات به ایمیل
- URL غیرطبیعی

۳-۱-۳- ویژگی‌های مبتنی بر جاوا اسکریپت و HTML

- ارسال وب‌سایت
- سفارشی‌سازی نوار وضعیت
- غیرفعال کردن کلیک راست
- استفاده از پنجره پاپ آپ
- تغییر مسیر با تگ آیفریم

۴-۱-۳- ویژگی‌های مبتنی بر دامنه

- طول عمر دامنه
- رکورد DNS
- ترافیک وب‌سایت
- رتبه‌بندی صفحه
- شاخص گذاری گوگل
- تعداد پیوندهای اشاره‌کننده به صفحه
- ویژگی‌های مبتنی بر گزارش‌های آماری

۳-۲- روش‌های کاهش ویژگی

روش‌های کاهش ابعاد داده به دودسته روش‌های مبتنی بر استخراج ویژگی و همچنین روش‌های مبتنی بر انتخاب ویژگی تقسیم می‌شوند [۲۷]. روش‌های انتخاب ویژگی سعی می‌کنند با انتخاب زیرمجموعه‌ای از ویژگی‌های اولیه، ابعاد داده‌ها را کاهش دهند. برخلاف روش‌های مبتنی بر استخراج ویژگی، این روش‌ها معنای اصلی ویژگی‌ها را بعد از کاهش حفظ می‌کنند. روش‌های مبتنی بر انتخاب ویژگی خود به سه روش فیلتر، پوششی و جاسازی‌شده تقسیم می‌شوند [۲۸]. هرکدام از این سه روش انتخاب ویژگی حاوی الگوریتم‌هایی برای اجرا هستند. در این قسمت به توضیح الگوریتم‌هایی که قرار است در این پژوهش مورداستفاده قرار گیرند می‌پردازیم.

۱-۲-۳روش پوششی

در ابن روش انتخاب زیرمجموعه ویژگی با استفاده از الگوریتم یادگیری انجام می‌شود. الگوریتم دسته‌ای از ویژگی‌ها را برای یادگیری انتخاب می‌نماید و نهایتاً آن دسته

از ویژگی‌ها که دقت بالاتری دارند، انتخاب می‌شود. الگوریتمی که کار ارزیابی زیرمجموعه ویژگی‌ها و انتخاب بهترین زیرمجموعه را انجام می‌دهد، خود به‌عنوان بخشی از تابع ارزیابی، کار جستجو برای انتخاب بهترین مدل را انجام می‌دهد [۷].

- روش CFS

CFS مقدار همبستگی بین ویژگی‌ها و کلاس‌هایشان و همچنین همبستگی بین خود ویژگی‌ها را اندازه‌گیری می‌کند. ایده کلی این است که زیرمجموعه ویژگی‌های خوب، همبستگی زیادی باکلاس‌ها دارند، اما نباید با یکدیگر همبستگی داشته باشند [۱۷]. در الگوریتم CFS، هیوریستیکی برای ارزیابی ارزش یا شایستگی یک زیرمجموعه ویژگی وجود دارد [۱۸].

- روش IG

تفاوت بین آنتروپی $H(S)$ از مجموعه داده S و آنتروپی مشروط $H(S|F)$ از مجموعه داده که پس از جداسازی توسط ویژگی F ساخته‌شده، به دست می‌آید. آنتروپی روش اندازه‌گیری ناخالصی در یک مجموعه داده است و اگر یک مجموعه داده تعداد مساوی از نمونه‌ها برای هر کلاس داشته باشد، مقدار آن حداکثر در نظر گرفته می‌شود [۱۷].

- روش PCA

یکی از عمومی‌ترین و شناخته‌شده‌ترین روش‌های آماری برای کاهش ویژگی‌ها است. هدف این روش به دست آوردن یک تبدیل تصویری است که از طریق آن بتوان با ترکیب خطی ویژگی‌های اصلی، تعداد کمتری ویژگی‌های جدید تولید نمود [۲۸].

۳-۳- الگوریتم‌های مدل‌سازی

- بیزین ساده

در روش بیزین ساده، دسته‌بندی بر پایه احتمالات و با فرض استقلال متغیرهای تصادفی ساخته می‌شود. این روش از ساده‌ترین الگوریتم‌های دسته‌بندی است که دقت قابل قبولی داشته و بر پایه احتمال وقوع یا عدم وقوع یک پدیده شکل می‌گیرد[۲۹].

- جنگل تصادفی

جنگل تصادفی مجموعه‌ای از درخت‌های تصمیم است که داده‌های آموزشی جهت ساخت هر درخت از روش انتخاب تصادفی با جایگذاری استفاده می‌کنند. هر درخت یک دسته‌بندی را می‌دهد که گفته می‌شود آن درخت به آن دسته رأی داده است. در انتها، دسته‌بندی که بیشترین رأی را داشته باشد انتخاب می‌شود. درخت‌ها هرس نمی‌شوند و در هر گره تعدادی ویژگی به‌طور تصادفی از مجموعه‌ی کل ویژگی‌ها برای انشعاب بررسی می‌شوند[۳۰].

- J48

این روش از معیار شاخص جینی[۵]جهت انتخاب ویژگی استفاده می‌کند[۲۷]. از میان ویژگی‌ها، هرکدام که مقدار شاخص جینی آن کوچک‌تر است، برای گروه جاری درخت تصمیم در نظر گرفته می‌شود.

۴-۳- روش پیشنهادی

همانطوریکه در شکل ۲ آمده است، روش پیشنهادی شامل سه مولفه هست: مولفه پیش‌پردازش، مولفه مدل‌سازی و مولفه آزمون و ارزیابی.

در این پژوهش در مولفه پیش‌پردازش داده‌ها برای کاهش ویژگی‌ها از روش ترکیبی جدیدی برای کاهش دومرحله‌ای ویژگی‌ها استفاده شده‌است. در این روش سعی شده است از پیچیدگی مسئله تا حد امکان کاسته شود و ویژگی‌هایی انتخاب گردد که در دسته‌بندی آن‌ها دارای بالاترین کارایی باشند و خطای دسته‌بندی نمونه‌ها را حتی‌الامکان کاهش دهد. بدین‌صورت که در مرحله اول بر اساس یک روش انتخاب ویژگی، زیرمجموعه‌ای مفید از ویژگی‌ها انتخاب می‌شود. برای این منظور با استفاده از روش‌های انتخاب ویژگی مبتنی بر پوششی، CFS و IG زیرمجموعه‌ای مهم از

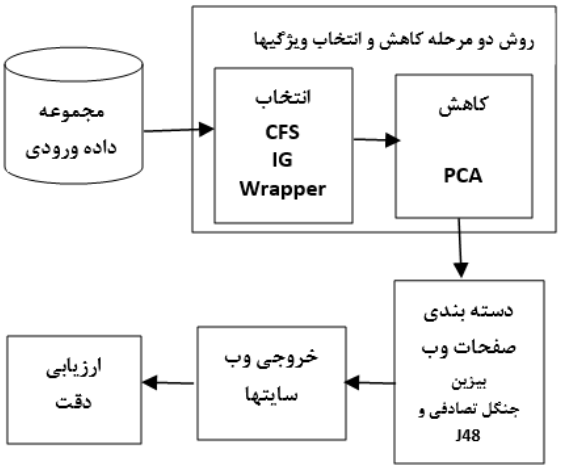
۵Gini coefficient

ویژگی‌ها انتخاب می‌شوند. سپس در مرحله بعد با اعمال روش استخراج ویژگی PCA بر روی ویژگی‌های باقیمانده، ترکیبی از این ویژگی‌ها با ابعاد کمتر به دست می‌آید. بدین‌صورت با کاهش دومرحله‌ای ویژگی‌ها، ترکیبی از ویژگی‌های مفید با ابعاد کمتر حاصل می‌شود. تکنیک‌های کاهش ویژگی دومرحله‌ای پیشنهادی، به‌صورت زیر نام‌گذاری شدند:

۱) کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر CFS و اعمال روش استخراج ویژگی PCA بر روی آن (CFS+PCA)
۲) کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر IG و اعمال روش استخراج ویژگی PCA بر روی آن (IG+PCA)
۳) کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر پوششی و اعمال روش استخراج ویژگی PCA بر روی آن (Wrapper+PCA)

در مولفه دوم برای مدل‌سازی از الگوریتم‌های مختلف داده‌کاوی نظیر J48، جنگل تصادفی و بیزین ساده برای داده‌های آموزشی و آزمون استفاده شده‌است. از این مدل‌ها همچنین برای تشخیص دسته‌ی وب‌سایت استفاده شده‌است. با استفاده از قوانین حاصل از این مدل‌ها می‌توان وب‌سایت‌های ورودی را دسته‌بندی کرده و در دسته فیشینگ یا غیر فیشینگ جای داد.

در بخش آخر هم ارزیابی و اندازه گیری دقت دسته‌بندی با توجه به داده‌های آزمون صورت می‌گیرد.



شکل (۲): الگوریتم شناسایی فیشینگ

۴- آزمایش و ارائه نتیجه

در این قسمت از مقاله ابتدا به معرفی مجموعه داده و ابزار مورداستفاده در پژوهش می‌پردازیم؛ سپس به تحلیل نتایج حاصل از آزمایش‌ها پرداخته و در آخر نتایج حاصل را با چند روش دیگر مقایسه می‌کنیم.

۴-۱- دادگان آزمون

داده‌های مورداستفاده در این پژوهش تعدادی وب‌سایت جعلی و واقعی است که از سایت داده‌کاوی UCI [۲۶] استخراج‌شده است. نوع داده‌های این پژوهش از نوع دسته‌ای است. هر وب‌سایت در مجموعه داده دارای یک ویژگی کلاس یا هدف است که دسته آن را نشان می‌دهد. یک بودن ویژگی نشان‌دهنده وب‌سایت جعلی، صفر بودن آن نشان‌دهنده مشکوک و منفی یک نشان‌دهنده وب‌سایت واقعی است. در این مجموعه داده ۳۰ ویژگی در دسترس است که در بخش سوم تشریح گردید.

این مجموعه داده شامل ۱۱۰۵۵ وب‌سایت فیشینگ و قانونی متعلق به سال ۲۰۱۵ است که شامل ۴۸۹۸ وب‌سایت قانونی و ۶۱۵۷ وب‌سایت فیشینگ است.

۴-۲- ابزارهای آزمون

پیاده‌سازی این پژوهش به کمک نرم‌افزار وکا نسخه ۳,۸,۱؛ یک نرم‌افزار منبع باز از دانشگاه Waikato [۱۲] است که

در محیط سیستم‌عامل ویندوز ۷ و پردازشگر Intel Core 2 Duo و RAM 2GB انجام‌گرفته است.

در این پژوهش از روش اعتبار سنجی تقاطعی با ۱۰ تکرار استفاده شده است که موجب می‌شود نتیجه به‌دست‌آمده دقیق‌تر باشد. اعتبار سنجی تقاطعی با ۱۰ تکرار موجب ارزیابی منطقی از مدل‌ها و کاهش سرریز می‌شود [۱۵]، این روش به این شکل است که داده‌ها به k قسمت تقسیم‌شده و k-1 قسمت آن به‌عنوان آموزش و ۱ قسمت به‌عنوان تست استفاده می‌شود؛ این عمل k مرتبه تکرار می‌شود. ارزیابی دقت نهایی برابر با میانگین k دقت محاسبه می‌شود [۳۱].

۳-۴- نتیجه آزمون

مدل آموزش‌یافته و آزمون‌شده با یک‌میزان دقت مشخص، می‌تواند جهت تشخیص کلاس یا اشیایی که برچسب کلاس آن‌ها ناشناخته است، مورداستفاده قرار گیرد [۳۲]. یکی از اهداف انتخاب ویژگی این است که یک زیرمجموعه از ویژگی‌ها برای افزایش دقت تشخیص، انتخاب شوند. به‌عبارت‌دیگر کاهش اندازه ساختار بدون کاهش قابل‌ملاحظه در دقت تشخیص دسته‌بندی که با استفاده از ویژگی‌های مدل به دست می‌آید، صورت گیرد [۱۵].

پس از اعمال روش‌های کاهش ویژگی بر روی داده‌ها، تعداد ویژگی‌ها به صورت نشان داده شده در جدول (۳) کاهش یافت. همان‌گونه که در جدول (۳) مشخص شده است، تعداد ویژگی‌ها با استفاده از تمام روش‌های کاهش ویژگی دو مرحله‌ای به طور قابل توجهی کاهش یافت. روش CFS+PCA و بیزین ساده در Wrapper+PCA به کمترین مقدار با تعداد ۵ ویژگی دست‌یافته‌اند؛ و مابقی روش‌ها به تعداد ۸ ویژگی رسیده‌اند.

همچنین براساس نتایج از ۴ معیار ذکرشده، معیار مبتنی بر جاوا اسکریپت و HTML با ۵ ویژگی متعلق به آن، کمترین تأثیر را در تشخیص وب‌سایت‌های فیشینگ دارد.

جدول(۳): تعداد ویژگی‌ها پس ازاعمال روش‌های کاهش ویژگی

نام روش	تعداد ویژگی		
داده‌های نرمال بدون کاهش ویژگی	۳۰		
CFS + PCA	۵		
IG + PCA	۸		
Wrapper + PCA	J48	جنگل تصادفی	بیزین ساده
	۸	۸	۵

پس از کاهش ویژگی‌ها، روش‌های دسته‌بندی درخت تصمیم J48، جنگل تصادفی و بیزین ساده به ترتیب با استفاده از تابع‌های J48، RandomForest و NaiveBayes در نرم‌افزار وکا اجرا شدند و معیارهای ارزیابی با استفاده از روش اعتبار سنجی تقاطعی با ۱۰ تکرار محاسبه و مورد ارزیابی قرار گرفتند تا بتوان به بهترین روش کاهش ویژگی و بهترین روش دسته‌بندی دست‌یافت.

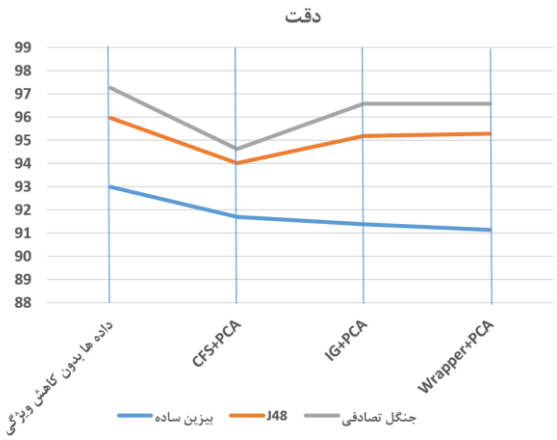
جدول (۴): نتایج بررسی روش‌های دسته‌بندی با استفاده از دقت

روش‌های کاهش ویژگی	روش‌های مدل‌سازی		
	J48	جنگل تصادفی	بیزین ساده
داده‌های نرمال بدون کاهش ویژگی	۹۵،۹۷٤۷	۹۷،۲۵۹۲	۹۲،۹۸۰٦
CFS + PCA	۹٤،۰۱۱۸	۹٤،٦۳۵۹	۹۱،۷۰۵۱
IG + PCA	۹۵،۱۷۸۷	۹٦،۵۷۱۷	۹۱،۳۷۹۵
Wrapper + PCA	۹۵،۲۷۸۲	۹٦،۵۸۰۷	۹۱،۱۳۵۲

در جداول (۴) و (۵) سه الگوریتم داده‌کاوی را با توجه به روش‌های کاهش ویژگی به ترتیب ازنظر دقت و شاخص F1 مقایسه می‌کنیم تا به بهترین الگوریتم دسته‌بندی و بهترین روش کاهش ویژگی برسیم.

جدول (۴) نشان می‌دهد ازنظر دقت که مهم‌ترین معیار ارزیابی است، الگوریتم‌های دسته‌بندی جنگل تصادفی و J48 با تعداد ۸ ویژگی و بیزین ساده با ۵ ویژگی به ترتیب دارای بهترین دقت هستند. ازنظر بهترین روش کاهش

ویژگی‌ها در دو الگوریتم جنگل تصادفی و J48 به ترتیب روش‌های Wrapper+PCA، IG+PCA و CFS+PCA دارای بهترین نتیجه هستند؛ اما در الگوریتم بیزین ساده این ترتیب بالعکس است. در نمودار (۱) دقت روش‌های دسته بندی در مقایسه با یکدیگر آمده است.



نمودار(۱): نتایج دسته‌بندی با استفاده از شاخص دقت

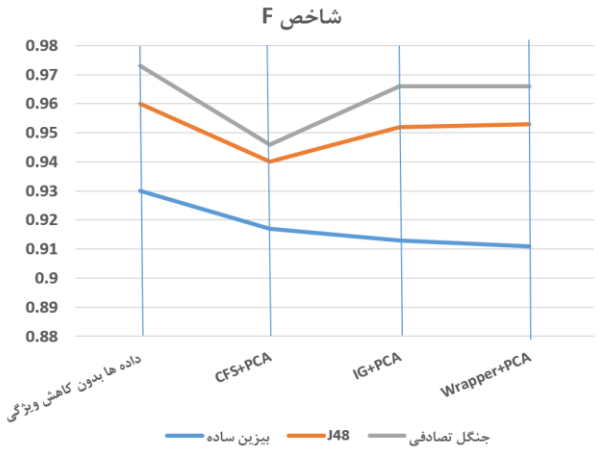
مقدار دقت در الگوریتم جنگل تصادفی، J48 و بیزین ساده در بهترین نتیجه به ترتیب به میزان ۰,۶۷۸۵، ۰,۶۹۶۵ و ۱,۲۷۵۵ نسبت به داده‌ها بدون انتخاب ویژگی کاهش‌یافته‌اند؛ که نشان می‌دهد الگوریتم جنگل تصادفی کمترین کاهش را نسبت به دو الگوریتم دیگر داشته است. همچنین بالاترین دقت با مقدار ۹۶,۵۸۰۷ مربوط به الگوریتم جنگل تصادفی با روش کاهش ابعاد Wrapper+PCA است؛ درحالی‌که تعداد ویژگی‌ها کاهش محسوسی نداشته‌اند.

جدول (۵): نتایج بررسی روش‌های دسته‌بندی با استفاده از F1

روش‌های کاهش ویژگی	روش‌های مدل‌سازی		
	J48	جنگل تصادفی	بیزین ساده
داده‌های نرمال بدون کاهش ویژگی	۰,۹۶	۰,۹۷۳	۰,۹۳
CFS + PCA	۰,۹٤	۰,۹٤٦	۰,۹۱۷
IG + PCA	۰,۹۵۲	۰,۹٦٦	۰,۹۱۳
Wrapper + PCA	۰,۹۵۳	۰,۹٦٦	۰,۹۱۱

همان‌گونه که در جدول (۵) می‌بینیم با استفاده از کاهش ویژگی دومرحله‌ای Wrapper+PCA و IG+PCA الگوریتم جنگل تصادفی بهترین عملکرد نسبت به سایر الگوریتم‌ها داشته و به مقدار یکسان ۰,۹۶۶ دست‌یافته است. پس از الگوریتم جنگل تصادفی الگوریتم J48 و بیزین ساده به ترتیب به بهترین نتیجه دست‌یافته‌اند.

مقدار F1 در بهترین نتیجه در الگوریتم جنگل تصادفی و J48 به میزان ۰,۰۰۷ و در بیزین ساده به میزان ۰,۰۱۳ نسبت به داده‌ها بدون انتخاب ویژگی کاهش‌یافته‌اند؛ که نشان می‌دهد الگوریتم جنگل تصادفی و J48 کمترین کاهش F1 را نسبت به مقدار اولیه داشته‌اند و از این نظر در یک سطح هستند. در نمودار (۲) شاخص F برای روش‌های دسته بندی در مقایسه با یکدیگر آمده است.



نمودار(۲): نتایج دسته‌بندی با استفاده از شاخص F

کاهش نامحسوس در مقدار دقت و F1 پس از کاهش ویژگی‌ها، حاصل افزایش هردو مقدار نرخ FP[۶] و نرخ FN[۷] است. ولی همواره نرخ FN نسبت به نرخ FP پایین‌تر بوده و این امر بدین معنی است که احتمال اینکه یک وب‌سایت فیشینگ به‌عنوان وب‌سایت قانونی در نظر گرفته شود کمتر است. درنتیجه احتمال گیر افتادن در دام فیشینگ کاهش می‌یابد.

[۶] False Positive
[۷] False Negative

۳-۴- مقایسه با روش‌های مشابه

در ادامه برای ارزیابی هرچه بیشتر رویکرد فوق، روش کاهش ویژگی دومرحله‌ای پیشنهادی با چند روش دیگر مقایسه می‌شود. ازآنجاکه در اغلب روش‌های مشابه از معیار دقت برای ارزیابی استفاده کرده‌اند ما برای مقایسه روش خود با روش‌های دیگر از این معیار استفاده کردیم. جدول (۶) دقت به‌دست‌آمده از تحقیقات گذشته را با دقت به‌دست‌آمده از روش پیشنهادی نشان می‌دهد.

جدول (۶): مقایسه روش پیشنهادی با روش‌های مشابه

تاریخ	مراجع	ویژگی‌های اولیه	ویژگی‌های نهایی	تعداد نمونه	دقت
روش پیشنهادی		۳۰	۸	۱۱۰۵۵	۹۶,۵۸
۲۰۱۱	[۱۹]	۷	۷	۳۰۰۰	۹۳
۲۰۱۲	[۲۳]	۲۳	۱۲	۲۵۰۰	۹۷,۶
۲۰۱٤	[۱]	۱٦	۹	۱۳۵۳	۹٤,٤
۲۰۱۵	[۲۵]	۱۵	۱۵	۱۰۰۰	۹٦,۵٦
۲۰۱۵	[۱۸]	۳۰	۱۵	۲٤۵٦	۹۷,٤۷
۲۰۱٦	[۳۳]	۳۰	۱۲	۱۱۰۵۵	۹۲,۰۵
۲۰۱۹	[۲۰]	۳۰	٦	۱۱۰۵۵	۹٤,٦

نتایج مطالعات ذکرشده محققان در مقایسه با روش جدید کاهش ویژگی دومرحله‌ای پیشنهادی، حاکی از برتری روش پیشنهادی است چراکه تعداد ویژگی‌های ما در بهترین عملکرد (۸ ویژگی) از تمامی آن‌ها کمتر بوده درحالی‌که به‌دقت بالاتری نسبت به آن‌ها دست‌یافته‌ایم. مجموعه داده استفاده‌شده در این پژوهش با مجموعه داده استفاده‌شده در [۳۳] یکسان بوده که در آن از یک روش طبقه‌بندی مشارکتی[۸] با الگوریتم FACA استفاده‌شده است.

در خصوص تحقیق‌های جدول ۶ که دقت بالاتری نسبت به روش پیشنهادی دارند، می‌بایست به تعداد ویژگی‌ها و همچنین تعداد نمونه به‌کاربرده شده توجه نمود. تحقیق[۲۰] نیز با کاهش دو مرحله‌ای به تعداد ویژگی ۲۰ درصدی دست یافته‌است.

۵- نتیجه‌گیری

با توجه به مشکلات و پیچیدگی‌های زیاد و سایر چالش‌ها در تشخیص صفحات فیشینگ، بررسی و ارائه روشی هوشمندانه برای تشخیص فیشینگ در صفحات وب ضروری است. در این پژوهش برای ساده سازی، کاهش دومرحله‌ای ویژگی‌ها پیشنهاد شده که در آن ابتدا با استفاده از روشهای پوششی، CFS و IG و ویژگیها انتخاب شده و سپس با اعمال روش استخراج ویژگی PCA از میان آنها بهترین ویژگی‌ها انتخاب می‌شوند که نهایتا لیست خیلی کوتاهی از ویژگیهای صفحات وب حاصل می‌شود.

نتایج حاصل از اجرای روش پیشنهادی بر روی مجموعه داده UCI، در بهترین حالت به‌دقت ۹۶,۵۸٪ با روش دومرحله‌ای Wrapper+PCA و الگوریتم جنگل تصادفی دست‌یافت. در این روش تعداد ویژگی‌ها از ۳۰ ویژگی به ۸ ویژگی کاهش یافت.

مزیت مدل پیشنهادی نسبت به سایر سامانه‌های مشابه، دستیابی به کمترین تعداد ویژگی‌ها پس از کاهش ویژگی است که این امر موجب ساده‌سازی و کاهش پیچیدگی مدل می‌شود، مزیت دیگر استفاده از تعداد نمونه بیشتر است که موجب می‌شود نتایج به واقعیت نزدیک‌تر باشند.

به‌عنوان کارهای آتی به محققان پیشنهاد می‌شود زمان محاسباتی الگوریتم‌های یادگیری را موردبررسی و تحلیل قرار دهند. همچنین پیشنهاد می‌شود از مجموعه داده حاوی وب‌سایت‌های فارسی برای تشخیص وب‌سایت‌های فیشینگ استفاده نمایند.

[۸] Associative classification

مراجع

[۱] Abdelhamid, N., Ayesh, A., Thabtah, F., "Phishing detection based Associative Classification data mining", Expert Systems with Applications 41 5948–5959, 2014.

[۲] معاونی, مسعود، "تشخیص حملات در بانکداری الکترونیکی با استفاده از سیستم ترکیبی فازی-راف (Fuzzy _rough)" گروه کامپیوتر دانشگاه امام رضا (ع)، ۱۳۹۴.

[۳] Mohammad, R. M., Thabtah, F., McCluskey, L., "Tutorial and critical analysis of phishing websites methods", Computer Science Review 17 (2015) 1-24.

[۴] Chaudhry, J. A., Rittenhouse, R. G., "Phishing: Classification and Countermeasures", 7th International Conference on Multimedia, Computer Graphics and Broadcasting, pp. 28-31, IEEE, 2015.

[۵] Anti Phishing Working Group, Phishing activity trends report, http://www.antiphishing.org/resources/apwg-reports/apwg_trends_report_q4_2019.pdf.

[۶] Buber, E., Demir, Ö., Sahingoz, O.K., "Feature Selections for the Machine Learning based Detection of Phishing Websites", International Artificial Intelligence and Data Processing Symposium (IDAP) IEEE, 2017.

[۷] Kohavi, R., John, G. H., "Wrappers for feature subset selection", Artificial Intelligence,Vol. 97, pp. 273-324, 1997.

[۸] Abur-rous, M. R. M., "Phishing Website Detection Using Intelligent Data Minning Techniques", Ph.D, dissertation, Dept. Computing, Bradford Univ, Bradford, 2010.

[۹] PhishTank.http://www.phishtank.com,2017.

[۱۰] Aravindhan, R., Shanmugalakshmi, Dr.R., Ramya, K., Dr.Selvan C, "Certain Investigation on Web Application Security:Phishing Detection and Phishing Target Discovery", 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS -2016), Jan. 22 – 23, 2016, Coimbatore, INDIA, Available: IEEE Xplore, http://www.ieee.org.

[۱۱] محمدی، شهریار، غروی، عرفانه، "کاربرد تکنیک‌های داده‌کاوی جهت تشخیص آدرس‌های فیشینگ"، کنگره ملی مهندسی برق، کامپیوتر و فناوری اطلاعات، مشهد: موسسه آموزش عالی خیام، ۱۳۹۲.

[۱۲] Sanglerdsinlapachai, N., Rungsawang, A., "Using Domain Top-page Similarity Feature in Machine Learning-based Web Phishing Detection", *Third International Conference on Knowledge Discovery and Data Mining*, IEEE, pp. 17-190, 2010.

[۱۳] Aburrous, M., Hossain, M. A., Keshav, D., Thabtah, F., "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies", IEEE Seventh International Conference on Information Technology, pp. 176-181, 2010.

[۱۴] سعیدی، پریسا، "بررسی سیستم‌های هوشمند تشخیص وب‌سایت فیشینگ در بانکداری الکترونیکی به روش منطق فازی"، نخستین کنفرانس بین‌المللی فناوری اطلاعات، تهران: مرکز همایش‌های توسعه ایران، ۱۳۹۴.

[۱۵] حاتمی خواه، نفیسه، "بررسی روش‌های مبتنی بر انتخاب ویژگی"، تهران، دانشگاه صنعتی مالک اشتر، ۱۳۹۲.

[۱۶] Basnet, R. B., Sung, A.H., Liu, Q., "Feature Selection for Improved Phishing Detection", international conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, pp 252-261, 2012, Available: https://link.springer.com.

[۱۷] Khonji, M., Jones, A., Iraqi, Y., "A Study of Feature Subset Evaluators and Feature Subset Searching Methods for Phishing Classification", *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pp.135-144, ACM, 2011.

[۱۸] Singh, P., Jain, N., Maini, A., "Investigating the Effect Of Feature Selection and

Dimensionality Reduction On Phishing Website Classification Problem", *1st International Conference on Next Generation Computing Technologies (NGCT) Dehradun*, India, IEEE, pp. 388-393, 2015.

[۱۹] rahmi A. H., isredza, Abawajy, J., "Phishing Email Feature Selection Approach", 10th International Joint Conference of IEEE TrustCom., pp. 916-921, 2011.

[۲۰] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learningbased phishing detection system," Information Sciences, vol. 484, pp. 153–166, 2019.

[۲۱] M. Almseidin, A. A. Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," International Journal of Interactive Mobile Technologies (iJIM), vol. 13, no. 12, pp. 171–183, 2019.

[۲۲] Meenu , Sunila godara, "Phishing Detection using Machine Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT) , Volume-9 Issue-2, December, 2019.

[۲۳] Pandey, M., Ravi, V., "Detecting phishing e-mails using Text and Data mining", *IEEE International Conference on Computational Intelligence and Computing Research(ICCIC)*, 2012.

[۲۴] Pandey, M., Ravi, V., "Text and Data Mining to Detect Phishing Websites and Spam Emails", *Proceedings of the 4th International Conference on Swarm, Evolutionary, and Memetic Computing*, Vol. 8298, pp.559-573, 2013.

[۲۵] لنگری، نفیسه، عبدالرزاق نژاد، مجید، "شناسایی وبگاه فیشینگ در بانکداری اینترنتی با استفاده از الگوریتم بهینه‌سازی صفحات شیب‌دار"، مجله پدافند الکترونیکی و سایبری. شماره ۱، صفحه ۴۰-۲۹، ۱۳۹۴.

[۲۶] Mohammad, R. M., Thabtah, F., McCluskey, L., Phishing Website Dataset, https://archive.ics.uci.edu/ml/datasets/ Phishing+websites, 2015.

[۲۷] اسماعیلی، مهدی، مفاهیم و تکنیک‌های داده‌کاوی، کاشان: سوره، ۱۳۹۲.

[۲۸] ورسلیز، کارلو، هوش تجاری داده‌کاوی و بهینه‌سازی برای تصمیم‌گیری، ترجمه‌ی احمدی، عباس، محبی، آزاده، ویرایش دوم، تهران، نشر دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، زمستان ۱۳۹۲.

[۲۹] H.John, George, and pat Langley, "Estimating Continuous Distribution in Bayesian Classifiers", In Proceeding of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufman,1995.

[۳۰] Breiman, Leo ."Random Forests", Machine Learning, Kluwer Academic Publishers. Manufactured in The Netherlands, Statistics DepartmentUniversity of CaliforniaBerkeley, 45, 5–32, 2001.

[۳۱] Kohavi, Ron, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI)*, pp. 1137-1143, ACM, 1995.

[۳۲] Lakhita, Yadav, S., Bohra, B., Pooja, "A Review on Recent Phishing Attacks in Internet", *IEEE International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 1312-1315, 2015.

[۳۳] Hadi, W., Aburub, F., Alhawari, S., "A new fast associative classification algorithm for detecting phishing websites", Applied Soft Computing 48 (2016) 729–734.

An efficient method for detecting phishing websites using data mining on web pages

Abstract

Phishing is regarded as a kind of internet attack on the web which aimed to steal the users' personal information for online stealing. Phishing plays a negative role in reducing the trust among the users in the business network based on the E-commerce framework. therefore, in this research, we tried to detect phishing websites using data mining. The detection of the outstanding features of phishing is regarded as one of the important prerequisites in designing an accurate detection system. Therefore, in order to detect phishing features, a list of 30 features suggested by phishing websites was first prepared. A new idea based on two steps: feature selection and feature extraction, has been proposed. To evaluate the proposed method, the performance of decision tree J48, random forest, naïve Bayes methods were evaluated on the reduced features. The results indicated that accuracy of the model created to determine the phishing websites by using the two-stage feature reduction-based Wrapper and Principal Component Analysis (PCA) algorithm in the random forest method of 96.58%, which is a desirable outcome compared to other methods.

***Keywords*:** Internet attack, Phishing, Data Mining, Feature Selection, Feature Extraction.