طبقه‌بندی و شناسایی وب سایت‌های فیشینگ به کمک مجموعه قوانین فازی و الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار

مجید عبدالرزاق نژاد

چکیده: یکی از تهدیدات پیش روی توسعه فناوری اطلاعات در فضای مجازی، سرقت اطلاعات شخصی و مالی کاربران می‌باشد که این تهدید امنیتی، فیشینگ نامیده می‌شود. بررسی و تحلیل روش‌های موجود نشان می‌دهد که ایجاد انعطاف‌پذیری در انتخاب ویژگی‌های اثرگذار در فرایند شناسایی وب سایت‌های فیشینگ، پویاسازی رفتار الگوریتم طبقه‌بندی کننده وب سایت‌های هدف و نیز امکان تحلیل و کنترل حجم گسترده‌ای از وب سایت‌ها مورد توجه قرار نگرفته است. لذا در این مقاله به منظور تحقق هم‌زمان سه هدف یادشده، ابتدا مکانیزمی بر اساس طراحی یک آستانه تغییر برای کاهش انعطاف‌پذیر ویژگی‌های مورد ارزیابی در شناسایی وب سایت‌های فیشینگ تعریف شده است. سپس با حافظه‌مند نمودن الگوریتم بهینه‌سازی صفحات شیب‌دار، کاهش نرم اثر حافظه بر عملکرد الگوریتم در تکرارهای بالا و نیز تعریف ۱۲ قانون فازی در یک سیستم استنتاج فازی اقدام به پویاسازی هوشمند این الگوریتم به منظور طبقه‌بندی وب سایت‌های جامعه ارزیابی به سه طبقه قانونی، مشکوک و فیشینگ می‌نماید. نتیجه پیاده‌سازی رویکرد هوشمند جدید پیشنهادی بر روی داده محک استاندارد در این حوزه و نیز مقایسه عملکرد این الگوریتم با عملکرد بهترین الگوریتم‌های موجود، نشان از تحقق اهداف سه‌گانه فوق‌الذکر برای این تحقیق را دارد.

کلیدواژه: **انتخاب هوشمند ویژگی، استنتاج فازی، الگوریتم بهینه‌سازی صفحات شیبدار، شناسایی وب سایت‌های فیشینگ، طبقه‌بندی.**

۱- مقدمه

توسعه کاربردهای فناوری اطلاعات در قالب رشد نفوذ شبکه جهانی اینترنت در سراسر جهان امری انکارناپذیر می‌باشد که تأثیر بسیار مثبتی روی تجارت الکترونیک داشته است. یکی از مهم‌ترین تهدیدات و چالش‌های پیش روی این حوزه بسیار پرکاربرد فناوری اطلاعات، امنیت خدمات خرید و بانکی می‌باشد. فیشینگ[1] آشکارترین نوع حملات به این نوع خدمات است. سرقت اطلاعات هویتی کاربران فضای مجازی و مشتریان فروشگاه‌های الکترونیکی و مؤسسات مالی را فیشینگ می‌نامند. اخیراً حملات فیشینگ با هدف قرار دادن خدمات مالی به یک کسب و کار جدی تبهکارانه تبدیل شده‌اند. با توجه به حجم گسترده این تهدید، راهکارهای متنوعی به منظور شناسایی وب سایت‌های فیشینگ تا کنون ارائه شده که می‌توان آنها را در ۵ گروه- رویکرد دسته‌بندی و تحلیل نمود.

رویکرد اول: ابزارهای ضد فیشینگ مبتنی بر لیست سیاه [۱] می‌باشند که با نصب در مرورگرها، اقدام به تشخیص سایت‌های فیشینگ می‌نماید. این ابزارها بر اساس ویژگی‌هایی از قبیل طول آدرس اینترنتی [۲]، محبوبیت سایت [۳]، طول مدتی که سایت ثبت شده و همچنین جستجوی سایت در لیست سیاه[2] [۱]، سایت فیشینگ را شناسایی و در صورت برخورد با سایت فیشینگ فعالیت‌های کاربر را مسدود می‌کنند و به او هشدار می‌دهند. از ابزارهای ضدفیشینگ می‌توان به ابزار Calling ID و Net Craft [۱] و EarthLink و Cloud mark [۴] اشاره کرد. مزیت اصلی استفاده از روش‌های مبتنی بر لیست سیاه پیاده‌سازی راحت آن می‌باشد [۱]. این ابزارها اغلب وابسته به مرورگر خاصی هستند و نمی‌توانند بر روی چند مرورگر نصب شوند [۱]. روش‌های مبتنی بر لیست سیاه معمولاً همه وب سایت‌های فیشینگ را شناسایی نمی‌کنند [۴]. در واقع لیست سیاه باید به روز رسانی شود و در غیر این صورت نمی‌تواند جدیدترین تهدیدات فیشینگ را در اختیار بگیرد.

رویکرد دوم: شناسایی فیشینگ در روش حل مبتنی بر تکنیک‌های داده‌کاوی [۵] که بر خلاف ابزارهای ضد فیشینگ مبتنی بر لیست سیاه، این رویکرد بر روی حجم زیادی از سایت‌ها انجام می‌گیرد. از روش‌های مبتنی بر داده‌کاوی می‌توان به روش داده‌کاوی فازی[3] و طبقه‌بندی انجمنی[4] اشاره کرد. در روش داده‌کاوی فازی [۲] از ترکیب تکنیک‌های داده‌کاوی و سیستم‌های فازی و نیز استخراج ۲۷ ویژگی به منظور بررسی بانک‌های اینترنتی که در معرض خطر وب سایت‌های فیشینگ هستند استفاده شده است. در این روش از تعدادی تکنیک طبقه‌بندی دقیق در داده‌کاوی مانند PART[5]، Prism و C۴.۵ استفاده شده است. در روش طبقه‌بندی انجمنی [۳] از تکنیک‌های CAR[6] و MCAR[7] برای شناسایی وب سایت فیشینگ در بانکداری اینترنتی استفاده شده است. چالش جدی این روش‌ها، ضعف آنها در مواجعه با حجم انبوه سایت‌های جامعه هدف می‌باشد که باعث بزرگ‌شدن ابعاد مسئله طبقه‌بندی و افزایش خطای طبقه‌بندی می‌شود.

رویکرد سوم: آنالیز میزان تشابه ویژگی‌های سایت‌های فیشینگ و قانونی به روش حل مبتنی بر تکنیک‌های ابتکاری[8] [۶] و [۷] منتهی می‌شود. اگر میزان تشابه بیش از یک آستانه از پیش تعیین شده باشد وب سایت، فیشینگ تشخیص داده می‌شود. از روش‌های ابتکاری می‌توان به روش مبتنی بر تشابه بصری[9] [۸]، الگوریتم Link Guard [۹]، آنالیز تصویر و مشخصات سایت[10] [۱۰]، روش مبتنی بر تشابه طرح[11] [۱۱] و

این مقاله در تاریخ ۱۴ مهر ماه ۱۳۹۴ دریافت و در تاریخ ۱۵ خرداد ماه ۱۳۹۵ بازنگری شد.

مجید عبدالرزاق نژاد، گروه کامپیوتر، دانشکده فنی و مهندسی، دانشگاه بزرگمهر قائنات، (email: abdolrazzagh@buqaen.ac.ir).

1. Phishing
2. Black List
3. Fuzzy Data Mining
4. Associative Classification
5. Projective Adaptive Resonance Theory
6. Classification Based on Association Rules
7. Multi-Class Classification Based on Association Rules
8. Heuristic
9. Visual Similarity
10. Site Characteristics and Image Analysis
11. Layout-Similarity

روش مبتنی بر EMD[1] [۱۲] اشاره کرد. مزیت روش‌های ابتکاری بر خلاف روش مبتنی بر لیست سیاه سایت‌های فیشینگ جدید را تشخیص می‌دهند و مشکل روش‌های مبتنی بر لیست سیاه را نخواهند داشت [۸]. در روش‌های ابتکاری در صورتی که فیشر وب سایت فیشینگ را متفاوت از وب سایت قانونی ایجاد کند با شکست مواجه می‌شود و نرخ دقت بسیار کاهش می‌یابد [۱۱].

رویکرد چهارم: روش حل مبتنی بر تکنیک‌های یادگیری ماشین طبق تجربیاتی که به دست می‌آورند سایت فیشینگ را شناسایی می‌کنند. از روش‌های مبتنی بر یادگیری ماشین می‌توان به روش مبتنی بر عصبی- فازی[2] [۱۳]، روش طبقه‌بندی رگرسیون لجیستیک [۱۴] و [۱۵]، روش پیشنهادی pagesafe [۱۶]، رگرسیون منطقی [۱۷]، ماشین بردار پشتیبان [۱۸] و [۱۹]، درخت رگرسیون افزایشی بیزین [۱۸] و جنگل تصادفی [۱۷] و [۲۰] اشاره کرد. چالش‌های اصلی این روش‌ها، پیچیدگی در یادگیری و حجم محاسباتی بالا و به تبع عدم موفقیت در مینیمم‌سازی زمان عملیاتی می‌باشند [۱۳].

رویکرد پنجم: روش‌های مبتنی بر الگوریتم‌های فوق ابتکاری [۱۶] که حجم گسترده‌ای از سایت‌ها را بر اساس ویژگی‌های آنها به سه دسته قانونی، مشکوک و فیشینگ طبقه‌بندی کرده و طبقه‌بند بهین را محاسبه می‌کنند. از روش‌های فوق ابتکاری که در تشخیص فیشینگ مورد استفاده قرار گرفته‌اند می‌توان به الگوریتم کلونی مورچه‌ها (ACO)[3] [۲۱] [۲۱] و [۲۲]، بهینه‌سازی گروه ذرات (PSO)[4] [۲۱] و [۲۲]، الگوریتم غذایابی باکتری (BFOA)[5] [۲۱] و الگوریتم خفاش بهبودیافته (MBAT)[6] [۲۱] و [۲۳] اشاره کرد. روش‌های این رویکرد دارای پیچیدگی کم می‌باشند و اغلب به سرعت به جواب می‌رسند [۱۶]. ساختار ساده و کاربرد آنها در طیف گسترده‌ای از مسایل، دلیل استقبال محققان از الگوریتم‌های فوق ابتکاری می‌باشد. از این جهت تقریباً این روش‌ها مشکل روش‌های یادگیری ماشین را نخواهند داشت و به سرعت به جواب می‌رسند.

در این تحقیق الگوریتم بهینه‌سازی صفحات شیب‌دار (IPO)[7] [۲۴] به به عنوان یکی از روش‌های فوق ابتکاری کارآمد با توجه به مزایای ویژه‌ای همچون مفهوم ساده، تنظیم پارامترهای کمتر، شکل‌گیری سریع و قابلیت بالا در جستجوی سراسری در مقایسه با دیگر الگوریتم‌های فوق ابتکاری در نظر گرفته شده است. این الگوریتم توسط ایده ذخیره سوابق موقعیت‌های گوی‌ها، کاهش نرم میزان اثرگذاری سوابق گوی‌ها در تعیین موقعیت جدید آنها و نیز فازی‌سازی ضرایب حرکت گوی‌ها، اصلاح و بهبود یافته است. الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار (MIPO) به منظور شناسایی وب سایت‌های فیشینگ و با طبقه‌بندی وب سایت ها در یکی از طبقات فیشینگ، قانونی و مشکوک مورد استفاده قرار گرفته شده است.

در بحث انتخاب ویژگی‌های به کار گرفته شده در شناسایی وب سایت‌های فیشینگ، در اکثر روش‌های ارائه‌شده [۳]، [۵]، [۲۱] و [۲۳] با انتخاب قطعی ویژگی‌ها، انتخاب پویای ویژگی‌های تاثیرگذار تا کنون مورد توجه قرار نگرفته است. با آنالیز ویژگی‌های موجود، مشاهده شد که برخی از آنها اثر چندانی در تشخیص وب سایت فیشینگ ندارند و فیشرها به آنها توجه چندانی نداشته‌اند. لذا روال کلاسیک در انتخاب ثابت و غیر منعطف ویژگی‌ها، حجم محاسباتی را بالا می‌برد و کارآمدی فرایند طبقه‌بندی را کاهش می‌دهد. در این تحقیق جهت رفع این مشکل، مجموعه آستانه‌هایی در دل مجموعه قواعد فازی تعریف گردیده است که به گونه‌ای سطح اثرگذاری ویژگی‌ها را تعیین می‌نماید و باعث می‌گردد الگوریتم در مواجهه با گروه سایت‌های مختلف، مجموعه ویژگی‌های متفاوتی را به عنوان ویژگی‌های انتخابی برای طبقه‌بندی انتخاب نماید.

در ادامه، سازماندهی مقاله بدین شرح تدوین گردیده است. تشریح مسئله شناسایی وب سایت‌های فیشینگ در بخش ۲ مورد بحث قرار گرفته است. بیان نحوه انتخاب پویای ویژگی‌ها و فرایند پیاده‌سازی الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار به منظور طبقه‌بندی سایت‌های مورد مطالعه در بخش ۳ مورد توجه قرار گرفته و در بخش ۴ به بررسی نتایج به دست آمده برای رویکرد پیشنهادی پرداخته شده است. در نهایت در بخش ۵ خلاصه و نتیجه‌گیری این تحقیق آمده است.

۲- تشریح مسئله شناسایی وب سایت فیشینگ

مسئله شناسایی وب سایت‌های فیشینگ در این تحقیق، نوعی مسئله طبقه‌بندی می‌باشد. این مسئله نیز خود وابسته به ویژگی‌های انتخابی وب سایت‌ها تعریف می‌شود. لذا در این بخش، نخست به تشریح ویژگی‌های یک وب سایت به عنوان بستر تعریف مسئله تحقیق تمرکز کرده و سپس مسئله طبقه‌بندی وب سایت‌ها و نیز انواع معیارهای ارزیابی تدوین می‌گردد. معرفی داده‌های استاندارد به منظور پیاده‌سازی و آزمون الگوریتم مطرح شده در شناسایی وب سایت‌های فیشینگ آخرین زیربخش این بخش می‌باشد.

۲-۱ ویژگی‌های یک وب سایت

هر وب سایت دارای یک سری ویژگی‌های منحصر به فرد است که باعث تمایز آن از سایر سایت‌ها و حتی تمایز با سایت‌های بسیار متشابه خود می‌شود. لذا همواره فیشر در ساخت یک سایت جعلی، به صورت آگاهانه یا ناآگاهانه، اقدام به تغییر برخی از ویژگی‌ها از حالت اولیه سایت قانونی می‌نماید. این ویژگی‌ها را می‌توان در پنج پارامتر URL و هویت دامنه، امنیت و رمزگذاری، کد منبع و جاوا اسکریپت، سبک و محتوای صفحه و آدرس اینترنتی وب سایت جستجو کرد. از این پارامترهای پنج‌گانه نیز، ۲۴ ویژگی با توجه به تحقیقات انجام‌گرفته [۲] و [۲۵] تا [۲۷] استخراج می‌گردد که در جدول ۱ ارائه شده‌اند.

۲-۲ مسئله طبقه‌بندی وب سایت‌ها

مسئله طبقه‌بندی از نوع مسایل بانظارت[8] می‌باشد که هدفش یافتن مدلی که کلاس‌های داده‌ای را متمایز نماید، است. در فرایند این مسئله، داده‌ها به دو بخش آموزش و آزمون تقسیم می‌شوند. از داده‌های آموزش جهت شناسایی و تنظیم حدود طبقات (ساختار مدل) استفاده می‌شود و از داده‌های آزمون جهت ارزیابی و محاسبه کیفیت طبقه‌بندی که در مرحله آموزش ترسیم شده است بهره گرفته می‌شود. نهایتاً مدل آموزش‌یافته و آزمون‌شده با یک میزان دقت مشخص می‌تواند جهت پیش‌بینی کلاس یا اشیایی که برچسب کلاس آنها شناخته نمی‌باشد مورد استفاده قرار

1. Earth Mover's Distance
2. Neuro-Fuzzy
3. Ant Colony Optimization
4. Particle Swarm Optimization
5. Bacteria Foraging Algorithm
6. Modified Bat Algorithm
7. Inclined Planes Optimization
8. Supervise

جدول ۱: انواع ویژگی‌های یک وب سایت.

پارامترها	ویژگی‌های مورد استفاده در فیشینگ
URL & هویت دامنه	استفاده از آدرس IP، درخواست URL، URL غیر طبیعی، نام دامنه URL
امنیت و رمزگذاری	استفاده از گواهی‌نامه، مجوز صدور گواهی‌نامه، کوکی غیر طبیعی، گواهی‌نامه‌های مهم (DN)
کد منبع & جاوا اسکریپت	تغییر مسیر صفحات، استفاده از onMouseOver برای مخفی‌کردن لینک، کنترل فرم سرور (SFH)
سبک و محتوای صفحه	اشتباهات املایی، استفاده از فرم‌های ثبت نام، مرتبه صفحه، بلوک‌های عکس (ویژگی‌های آن شامل رنگ و اندازه تصویر می‌باشد)، فونت، رنگ پس‌زمینه، چیدمان متن و فاصله بین خطوط
آدرس اینترنتی وب سایت	آدرس طولانی URL، جایگزینی شخصیت‌های مشابه برای URL، اضافه‌کردن یک پیشوند یا پسوند، استفاده از نماد @، استفاده از کد کاراکتر هگزادسیمال

جدول ۲: کدگذاری ۳^k ناحیه توسط کد باینری k بیتی.

شماره دهدهی ناحیه	حدود کرانی ناحیه	کد باینری k بیتی ناحیه
۰	$D_۱(X)<۰$ $D_۲(X)<۰$ $D_۳(X)<۰$ ⋮ $D_k(X)<۰$	۰۰۰...۰۰۰
۱	$D_۱(X)\geq ۰$ $D_۲(X)<۰$ $D_۳(X)<۰$ ⋮ $D_k(X)<۰$	۰۰۰...۰۰۱
۲	$D_۱(X)<۰$ $D_۲(X)\geq ۰$ $D_۳(X)<۰$ ⋮ $D_k(X)<۰$	۰۰۰...۰۱۰
⋮	⋮	⋮
$۳^k-۲$	$D_۱(X)<۰$ $D_۲(X)\geq ۰$ $D_۳(X)\geq ۰$ ⋮ $D_k(X)\geq ۰$	۱۱۱...۱۱۰
$۳^k-۱$	$D_۱(X)\geq ۰$ $D_۲(X)\geq ۰$ $D_۳(X)\geq ۰$ ⋮ $D_k(X)\geq ۰$	۱۱۱...۱۱۱

گیرد. در این مقاله هدف طبقه‌بندی وب سایت‌ها در سه کلاس قانونی، مشکوک و فیشینگ می‌باشد به گونه‌ای که کمترین وب سایت به اشتباه طبقه‌بندی شود.

فرض کنید، تعداد k طبقه‌بند $D_۱(X)$، $D_۲(X)$، ... و $D_k(X)$، وظیفه طبقه‌بندی سه کلاس قانونی $(C_{legitimate})$، مشکوک $(C_{suspicious})$ و فیشینگ $(C_{phishing})$ را بر عهده داشته باشند. تعداد نواحی تولیدشده توسط k طبقه‌بند برابر با ۳^k می‌باشند که هر کدام به یک کلاس مشخص تعلق دارد. با مفروض قرار دادن n ویژگی در فضای جواب، شاهد یک فضای جستجوی $n+۱$ بعدی خواهیم بود که معادله ابرصفحه یک طبقه‌بند خطی i ام در آن به صورت زیر فرموله می‌شود [۲۸]

$$D_i(X)=W_{i۱}X_۱+W_{i۲}X_۲+\ldots+W_{in}X_n+W_{i(n+۱)} \quad (۱)$$
$$, \quad i=۱,۲,\ldots,k$$

که در آن $(X_۱,X_۲,\ldots,X_n)$ بردار ویژگی می‌باشد که برای هر وب سایت شامل مقادیر خاص و معین خود بوده و $(W_{i۱},W_{i۲},\ldots,W_{i(n+۱)})$ بردار ضرایب ابرصفحه طبقه‌بند i ام است. هدف مسئله طبقه‌بندی وب سایت‌ها، کمینه‌نمودن تعداد سایت‌هایی می‌باشد که به اشتباه طبقه‌بندی شده‌اند که توسط تابع هدف تعریف‌شده در (۲) محاسبه می‌شود

$$F(W)=\sum_{j=۰}^{۳^k-۱} Miss_j \quad (۲)$$

در معادله فوق، W بردار ضرایب تمام طبقه‌بندها و $Miss_j$ تعداد وب سایت‌هایی که در ناحیه j ام به اشتباه طبقه‌بندی شده‌اند می‌باشد. به منظور محاسبه $Miss_j$، نخست می‌بایست نواحی ایجادشده را کدگذاری و سپس درباره کلاس هر ناحیه تصمیم‌گیری نمود و در نهایت تعداد سایت‌های ناحیه j ام را که کلاس آنها، همسان با کلاس نسبت داده شده به ناحیه j ام نمی‌باشد، به عنوان $Miss_j$ محاسبه نمود.

به طور بدیهی، کرانه‌ها یا دیواره‌های هر ناحیه توسط یکی از نامساوی‌های $D_i(X)\geq ۰$ یا $D_i(X)<۰$ که در آن $i=۱,۲,\ldots,k$، تعیین می‌شوند. اگر برای هر ناحیه، به ازای هر کرانه $D_i(X)\geq ۰$ کد ۱ و به ازای هر کرانه $D_i(X)<۰$ کد صفر در نظر گرفته شود، آن گاه برای هر ناحیه یک کد باینری به طول k بیت مانند جدول ۲ خواهیم داشت. با قراردادن ویژگی‌های هر وب سایت در (۱) طبقه‌بندها، می‌توان به آسانی عضویت آن وب سایت را به یکی از کدهای باینری تعریف‌شده، شناسایی کرد. فرض کنید تعداد کل سایت‌های به کار گرفته شده در بخش آموزش طبقه‌بندی برابر با N باشند. آن گاه مشخص شود که مثلاً N_j وب سایت در ناحیه j ام $(j=۰,۱,۲,\ldots,۳^k-۱)$ قرار دارند. اگر از تعداد N_j وب سایت، φ_t سایت متعلق به کلاس C_t باشد $(t=۱,۲,۳)$، آن گاه ناحیه ۰۱۱...۰ که معادل کد باینری اندیس j می‌باشد، متعلق به کلاسی است که بیشترین فراوانی را داشته باشد. یعنی اگر $\varphi_h=\max(\varphi_۱,\varphi_۲,\varphi_۳)$، ناحیه ۰۱۱...۰ متعلق به کلاس h خواهد بود.

۳-۲ انواع معیارهای ارزیابی

سه معیار اصلی به منظور ارزیابی عملکرد طبقه‌بندی وب سایت‌ها عبارتند از نرخ دقت تشخیص، نرخ خطا و نرخ زمان تشخیص که به ترتیب در (۳) تا (۵) نمایش داده شده است

$$Acc\%=۱۰۰-\frac{F(W)}{N}\times ۱۰۰ \quad (۳)$$

$$Err\%=\frac{F(W)}{N}\times ۱۰۰ \quad (۴)$$

$$T_{detect}=\frac{T}{N} \quad (۵)$$

جدول ۳: انواع ویژگی‌های استخراج‌شده جهت ارزیابی وب سایت فیشینگ.

معیارهای در نظر گرفته شده	ویژگی‌های مرتبط با آن
ویژگی‌های مبتنی بر جاوا اسکریپت و HTML	تغییر مسیر صفحات، استفاده از OnMouseOver جهت پنهان‌کردن لینک، غیر فعال کردن کلیک راست، استفاده از pop-up
ویژگی‌های مبتنی بر آبنرمال	درخواست URL، URL of Anchor، کنترلگر فرم سرور (SFH)، URL غیر طبیعی
ویژگی‌های مبتنی بر نوار آدرس	آدرس IP، طول URL، استفاده از @، پسوند یا پیشوند مجزاشده توسط "-"، دامنه و زیردامنه، HTTPS (پروتکل انتقال ابرمتن)، SSL (لایه سوکت امن)

که در معادلات فوق، N تعداد کل وب سایت‌ها و T زمان محاسبه ضرایب ابرصفحات طبقه‌بندها می‌باشد.

نرخ مثبت کاذب، نرخی از وب سایت قانونی که به نادرستی به عنوان وب سایت فیشینگ طبقه‌بندی می‌شود و نرخ منفی کاذب، نرخی از وب سایت‌های فیشینگ که اشتباهاً به عنوان وب سایت قانونی طبقه‌بندی می‌شوند، می‌باشند. نرخ مثبت کاذب توسط (۶)، نرخ منفی کاذب توسط (۷)، نرخ مثبت درست توسط (۸) و نرخ منفی درست توسط (۹) به دست می‌آید

$$FP = \frac{N_{LP}}{N_{LL} + N_{LP}} \quad (۶)$$

$$FN = \frac{N_{PL}}{N_{PP} + N_{PL}} \quad (۷)$$

$$TP = \frac{N_{PP}}{N_{PP} + N_{PL}} \quad (۸)$$

$$TN = \frac{N_{LL}}{N_{LL} + N_{LP}} \quad (۹)$$

در معادلات بالا N_{LP} تعداد وب سایت‌های قانونی که به عنوان فیشینگ طبقه‌بندی گردیده و N_{LL} تعداد وب سایت‌های قانونی که به درستی طبقه‌بندی شده است می‌باشند. همچنین N_{PL} تعداد وب سایت‌های فیشینگ که به عنوان قانونی طبقه‌بندی شده و N_{PP} تعداد وب سایت‌های فیشینگ که به درستی طبقه‌بندی شده است می‌باشند.

۲-٤ داده‌های استاندارد

از مجموعه داده http://www.phishtank.com که شامل لیست url وب سایت‌های فیشینگ می‌باشند و به طور منظم به روز رسانی می‌شوند جهت شناسایی وب سایت‌های فیشینگ استفاده شده است که این مجموعه داده توسط مقالات زیادی [۲]، [۳]، [۱۱]، [۱۳]، [۱۵]، [۲۱] تا [۲۳]، [۲۹] و [۳۰] استفاده شده‌اند. نکته قابل توجه به روز رسانی لیست وب سایت‌های فیشینگ در بانک اطلاعاتی معرفی‌شده است که باعث می‌شود مجموعه داده‌های مستخرج از این بانک اطلاعاتی در بازه‌های زمانی مختلف، تفاوت‌هایی با یکدیگر داشته باشند. در این پژوهش حدود ۱۰۰۰ وب سایت از این سایت استخراج شده است. تعداد وب سایت‌های فیشینگ، وب سایت‌های قانونی و وب سایت‌های مشکوک در پایگاه داده به ترتیب ۵۰۰، ۳۰۰ و ۲۰۰ می‌باشد. پایگاه داده موجود در تاریخ ۱۱ جولای ۲۰۱۴ تا ۱۸ جولای ۲۰۱۴ و ۲۱ و ۲۲ آگوست ۲۰۱۴ جمع‌آوری شده است.

۳- رویکرد پیشنهادی

در این مقاله شناسایی وب سایت‌های فیشینگ طی سه مرحله استخراج ویژگی‌های وب سایت‌ها، انتخاب هوشمند ویژگی‌ها و طبقه‌بندی توسط الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار، طراحی شده است. لذا در چهار زیربخش، ابتدا به تشریح نحوه استخراج ویژگی‌ها پرداخته و سپس فرایند ابتکاری انتخاب هوشمند ویژگی بیان می‌شود. ترسیم فرایند الگوریتم اصلاح‌شده و نهایتاً نحوه طبقه‌بندی با بهره‌گیری از این الگوریتم قسمت‌های پایانی این بخش را در بر خواهند داشت.

۳-۱ استخراج ویژگی

جهت استخراج ویژگی‌های وب سایت از ابزار Googbar [۳۱] که با زبان جاوا اسکریپت و PHP[1] ایجاد شده، استفاده شده است. این ابزار بر روی مرورگر موزیلا نصب می‌شود و با فراخوانی هر وب سایت ویژگی‌های آن را استخراج می‌کند. ویژگی‌هایی که از طریق این ابزار استخراج شدند به سه معیار از ویژگی‌های مبتنی بر نوار آدرس، ویژگی‌های مبتنی بر آبنرمال و ویژگی‌های مبتنی بر جاوا اسکریپت و HTML بخش‌بندی می‌شوند که در جدول ۳ قابل مشاهده می‌باشند. ویژگی‌های استخراج‌شده از اهمیت خاصی برخوردار هستند و اغلب مقالات [۲]، [۶]، [۲۵] و [۳۱] به پراهمیت بودن آنها در تشخیص وب سایت فیشینگ اشاره کرده‌اند. از این رو به ارزیابی وب سایت مبتنی بر این تعداد ویژگی اکتفا شده که نتایج خوبی هم از این تعداد ویژگی حاصل شده است.

۳-۲ انتخاب هوشمند ویژگی‌ها

جهت شناسایی وب سایت‌های فیشینگ ۱۵ ویژگی با میزان اثرگذاری بیشتر، استخراج شدند که در نظر گرفتن همه آنها برای طبقه‌بندی مجموعه سایت‌ها، زمان محاسبات را افزایش می‌دهد. لذا به منظور کاهش زمان محاسبات از این تعداد ویژگی با استفاده از مجموعه قوانین تعریف‌شده اقدام به انتخاب یک زیرمجموعه منعطف و اثرگذارتر در بحث طبقه‌بندی شده است.

در اکثر روش‌های ارائه‌شده تعداد معینی ویژگی برای شناسایی وب سایت‌های فیشینگ در نظر گرفته شده است. برخی از این ویژگی‌ها اثر چندانی در تشخیص وب سایت‌های فیشینگ ندارند. این انتخاب ثابت و غیر منعطف، حجم محاسباتی را بالا و کارآمدی فرایند طبقه‌بندی را کاهش خواهد داد. در این پژوهش، جهت رفع این مشکل، پیشنهاد تعریف مجموعه آستانه‌هایی در غالب قواعدی برای خصوصیات سایت‌ها مطرح می‌گردد (جدول ۴). کار این قواعد این خواهد بود که خصوصیاتی از برآیند سایت‌ها را انتخاب کند که به گونه‌ای با رد آستانه‌های تعریف‌شده، یک گونه حساسیت و توجه ویژه را به خود جلب می‌کنند. لذا خصوصیاتی که حساسیت قواعد تعریف‌شده را از حالت تعادل خارج نمایند به عنوان خصوصیات هدف برای طبقه‌بندی انتخاب خواهند شد.

در جدول ۴، انتخاب ویژگی‌های مناسب بر اساس دو آستانه ۰٫۶ و ۰٫۳ که با سعی و خطا به دست آمده‌اند صورت گرفته است. آستانه ۰٫۳ ویژگی‌هایی که حداقل ۳۰ درصد وب سایت‌های فیشینگ به آنها توجه ویژه‌ای داشته‌اند و جهت جعل یک وب سایت بر روی این ویژگی‌ها متمرکز شده‌اند را انتخاب می‌کند. در نتایج به دست آمده توسط آستانه

1. Perl Hypertext Preprocessor

جدول ۴: قواعد تعریف‌شده روی ویژگی‌ها بر اساس سطح تهدید فیشینگ.

ویژگی وب سایت	قاعده تعریف‌شده برای انتخاب ویژگی
تغییر مسیر صفحات	If [Phishing dataset,redirect_page] >= 4 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
استفاده از onMouseOver جهت پنهان‌کردن لینک	If [phishing dataset,onMouseOver] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
غیر فعال کردن کلیک راست	If [phishing dataset,right click disabled] = 1 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
استفاده از پنجره pop-up	If [phishing dataset,using popup] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
آدرس IP	If [phishing dataset, IP address] = 1 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
طول URL	If [phishing dataset, URL_length] > 75 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
استفاده از سیمبل @	If [phishing dataset, @ symbol] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
پسوند یا پیشوند مجزاشده توسط "-"	If [phishing dataset, "-" symbol] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
زیردامنه	If [phishing dataset, dots in the domain] > 3 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
HTTPS	If [phishing dataset, Not Using Https] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
SSL	If [phishing dataset, Not Using SSL] = 1 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
درخواست URL	If [phishing dataset, Request URL] > 61 Counter = counter + 1 If counter > size [phishing dataset]*30% → Feature selection
URL of Anchor	If [phishing dataset, URL of anchor] > 67 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
SFH	If [phishing dataset, SFH refers to a different domain] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection
URL غیر طبیعی	If [phishing dataset, the host name is not included in URL] = 1 Counter = counter + 1 If counter > size [phishing dataset]*60% → Feature selection

۰/۳ ویژگی غیر فعال کردن کلیک راست که در همه موارد وب سایت فیشینگ وجود داشته است انتخاب شده است. همچنین ویژگی‌های HTTPs و درخواست URL که در اکثر وب سایت‌های فیشینگ به آنها توجه شده است و ویژگی‌های طول URL، داشتن زیردامنه، تغییر مسیر صفحات، SSL و آدرس IP از دیگر ویژگی‌های مهم هستند که انتخاب شده‌اند و اکثر وب سایت‌های فیشینگ به آنها توجه ویژه‌ای داشته‌اند. آستانه ۰/۶ ویژگی‌هایی که حداقل ۶۰ درصد وب سایت‌های فیشینگ به آنها توجه ویژه‌ای داشته‌اند را انتخاب می‌کند. این ویژگی‌ها شامل ویژگی‌های درخواست URL، غیر فعال کردن کلیک راست و HTTPs می‌باشند.

۳-۳ الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار

فرایند الگوریتم بهینه‌سازی سیستم صفحات شیبدار (IPO)[1] که توسط مظفری و همکارانشان [۲۴] ارائه شده، بر اساس دینامیک حرکت اجسام گرد بر روی سطوح شیب‌دار بدون اصطکاک بیان شده است. در این الگوریتم، ذرات (جواب‌های کاندید) به صورت گوی‌هایی کوچک در نظر گرفته شده‌اند که روی سطح شیبداری بدون اصطکاک (فضای جستجو) با تپه‌ها، دره‌ها و شانه‌های متعدد به سمت نقاط بهینه حرکت می‌کنند. موقعیت هر گوی معادل جزئیات پارامتریک یک جواب مسئله و ارتفاع متناظر آن، معادل کیفیت آن جواب می‌باشد. گوی‌ها (جواب‌ها) در IPO با شتاب‌های ثابت، بر اساس موقعیتشان نسبت به گوی‌های دیگر در ارتفاع پایین‌تر، به دنبال کاهش ارتفاع (بهینه‌کردن کیفیت جواب) خود هستند (شکل ۱).

مطابق آنچه که برای فرایند الگوریتم IPO در [۲۴] مطرح شده است، این الگوریتم بدون حافظه بوده و کیفیت مکانیزم‌های اکتشاف و استخراج آن بسیار وابسته به ضرایب شتاب (θ_1) و سرعت (θ_2) دارد. لذا در این مقاله با تعریف یک حافظه به منظور ذخیره سوابق موقعیت گوی‌ها به منظور اعمال در محاسبه شتاب و سرعت گوی‌ها و نیز فازی‌سازی ضرایب شتاب و سرعت در محاسبه موقعیت‌های جدید، ساختار الگوریتم اصلاح و عملکرد آن بهبود یافته است.

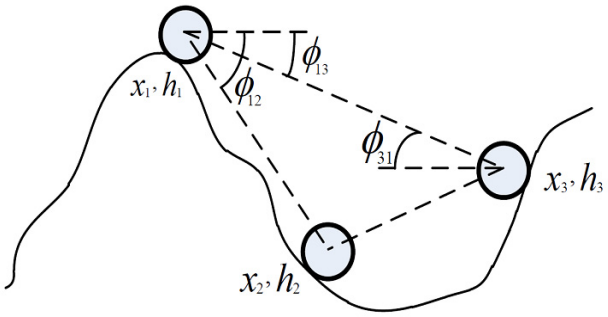
برای توصیف دقیق‌تر الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیبدار (MIPO)[2]، تعداد K گوی (جواب) را به صورت تصادفی بر روی یک سطح شیبدار (فضای جستجو) nبعدی در نظر بگیرید. موقعیت جاری گوی iام در زمان (تکرار) tام به صورت زیر نمایش داده می‌شود

$$X_i^t = (x_{i1}^t, x_{i2}^t, \ldots, x_{in}^t) \quad , \quad 1 \le i \le K \qquad (10)$$

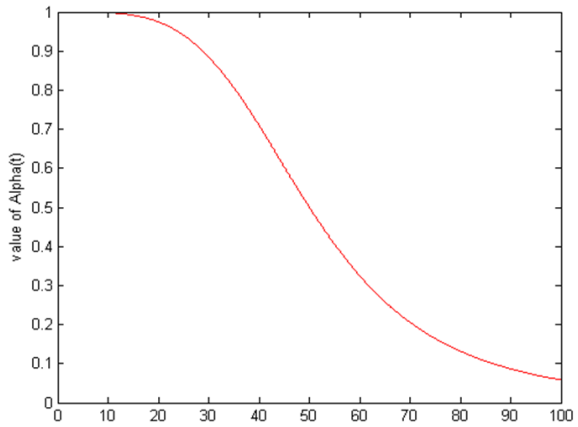
برای الگوریتم MIPO حافظه‌ای تعریف می‌شود تا اطلاعات موقعیت و ارتفاع گوی‌ها در تکرار قبل در آن ذخیره شود. هدف الگوریتم کمینه‌نمودن ارتفاع گوی‌ها ($f(X_i^t)$) می‌باشد. برای رسیدن به این هدف در گام نخست و مطابق با فرایند IPO ارائه‌شده در [۲۴]، زاویه بین دو گوی i و j که زاویه خط راست واصل بین این دو گوی با خط افق می‌باشد به کمک فرمول زیر محاسبه می‌شود

1. Inclined Planes System Optimization

2. Modified Inclined Planes System Optimization



شکل ۱: مثالی از فضای جستجوی الگوریتم IPO به همراه ۳ گوی [۲۴].



شکل ۲: نمودار تغییرات $\alpha(t)$ اگر حداکثر تعداد تکرارها برای خاتمه الگوریتم ۱۰۰ باشد.

1. If (UN is high) and (IN is high) then (Teta1 is high)(Teta2 is high) (1)
2. If (UN is high) and (IN is med) then (Teta1 is med)(Teta2 is low) (1)
3. If (UN is med) and (IN is low) then (Teta1 is low)(Teta2 is high) (1)
4. If (UN is med) and (IN is med) then (Teta1 is med)(Teta2 is high) (1)
5. If (UN is low) and (f(Xdiffer) is low) and (IN is low) then (Teta1 is high)(Teta2 is high) (1)
6. If (UN is low) and (f(Xdiffer) is med) and (IN is low) then (Teta1 is med)(Teta2 is med) (1)
7. If (UN is low) and (f(Xdiffer) is high) and (IN is low) then (Teta1 is med)(Teta2 is low) (1)
8. If (UN is low) and (f(Xdiffer) is med) and (IN is med) then (Teta1 is high)(Teta2 is med) (1)
9. If (UN is low) and (f(Xdiffer) is med) and (IN is high) then (Teta1 is low)(Teta2 is med) (1)
10. If (UN is low) and (f(Xdiffer) is high) and (IN is high) then (Teta1 is high)(Teta2 is high) (1)
11. If (UN is low) and (f(Xdiffer) is low) and (IN is med) then (Teta1 is high)(Teta2 is low) (1)
12. If (UN is low) and (f(Xdiffer) is high) and (IN is med) then (Teta1 is high)(Teta2 is high) (1)

شکل ۳: مجموعه قوانین فازی دوازده‌گانه جهت تنظیم مقادیر ضرایب θ_1 و θ_2.

$$\phi_d^t(i,j) = \tan^{-1}\frac{f(X_i^t) - f(X_j^t)}{x_{id}^t - x_{jd}^t} \qquad (۱۱)$$

که $i,j = 1,2,\ldots,K$ به شرط $i \neq j$ و $d = 1,2,\ldots,n$ نمایشگر ابعاد هر موقعیت می‌باشد. زاویه بین دو گوی i و j در تکرار اول الگوریتم MIPO توسط (۱۱) و از تکرار دوم توسط فرمول توسعه‌یافته زیر محاسبه خواهد شد

$$H_{id} = \frac{f(X_i^{t-1}) - f(X_i^t)}{x_{id}^{t-1} - x_{id}^t} \qquad (۱۲)$$

$$\alpha(t) = \frac{1}{1 + (\frac{t}{0.5 \max Iter})^4} \qquad (۱۳)$$

$$\phi_d^t(i,j) = \tan^{-1}(\alpha(t) H_{id} \frac{f(X_i^t) - f(X_j^t)}{x_{id}^t - x_{jd}^t}) \qquad (۱۴)$$

که H_{id} شیب خط واصل بین موقعیت تکرار قبلی $(t-1)$ گوی iام و تکرار جاری (t) آن در بعد dام می‌باشد. هرچه این شیب بیشتر باشد، زاویه بین گوی i و j را در بعد dام بازتر خواهد کرد. ضریب $\alpha(t)$ نیز به منظور کنترل این میزان اثرگذاری سوابق گذشته هر گوی بر اساس (۱۳) و وابسته به زمان بین بازه $[0,1]$ طبق شکل ۲ انتخاب می‌شود.

هر گوی در بعد dام خود $(d = 1,2,\ldots,n)$، متناسب با زاویه‌اش با بعد dام سایر گوی‌ها و نیز اختلاف ارتفاع آن گوی با سایر گوی‌ها، یک شتاب برای لغزش روی سطح شیبدار (فضای جستجو) می‌گیرد که به صورت زیر محاسبه می‌شود

$$a_{id}^t = \sum_{j=1, j\neq i}^{K} U(f(X_i^t) - f(X_j^t)) \sin(\phi_d^t(i,j)) \qquad (۱۵)$$

که $i = 1,2,\ldots,K$ و $d = 1,2,\ldots,n$ و تابع $U(f(X_i^t) - f(X_j^t))$، تابعی پله‌ای به صورت (۱۶) بوده و اجازه می‌دهد گوی iام در بعد dام خود شتابی متناسب با زاویه این گوی با سایر گوی‌ها بگیرد

$$U(w) = \begin{cases} 1 & , \quad w > 0 \\ 0 & , \quad w \leq 0 \end{cases} \qquad (۱۶)$$

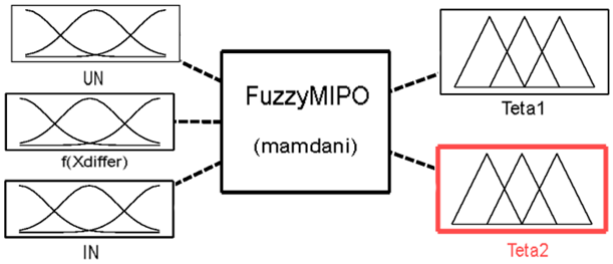
اگر گویی که دارای کمترین ارتفاع در زمان tام باشد با $X_{best}^t = (x_{best,1}^t, x_{best,2}^t, \ldots, x_{best,n}^t)$ نمایش داده شود، آن گاه سرعت ثابت گوی iام در بعد dام در زمان tام به کمک (۱۷) تعیین می‌گردد

$$v_{id}^t = \frac{x_{best,d}^t - x_{id}^t}{\Delta t} \qquad (۱۷)$$

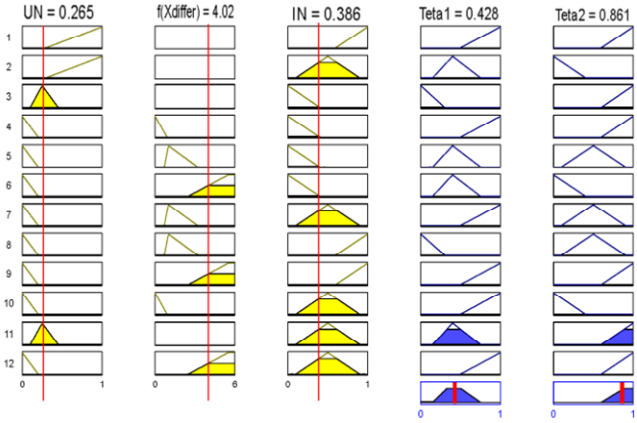
حال با هدف کمینه‌نمودن ارتفاع گوی‌ها، موقعیت جدید گوی i در بعد dام که با x_{id}^{t+1} نمایش داده می‌شود بر اساس حرکت با شتاب ثابت a_{id}^t و سرعت v_{id}^t به صورت زیر محاسبه می‌گردد

$$x_{id}^{t+1} = \theta_1 a_{id}^t \Delta t^2 + \theta_2 v_{id}^t \Delta t + x_{id}^t \qquad (۱۸)$$

ضرایب θ_1 و θ_2 در (۱۸) به ترتیب ضرایب تنظیم‌کننده شتاب (جستجوی اکتشافی) و سرعت (جستجوی استخراجی) در MIPO می‌باشند. در الگوریتم IPO، معادله محاسبه موقعیت جدید (معادله (۸) از [۲۴]) علاوه بر پارامترهای فوق، حاوی دو ثابت تصادفی R_1 و R_2 بوده که به صورت



شکل ۴: سیستم استنتاج فازی مقادیر θ_1 و θ_2.



شکل ۶: پیاده‌سازی مجموعه قوانین فازی در سیستم استنتاج فازی جهت تنظیم ضرایب θ_1 و θ_2.

تصادفی در بازه [۰،۱] انتخاب می‌شوند. لذا به منظور کاهش تعداد پارامترهای الگوریتم، در این نسخه ثابت‌های R_1 و R_2 حذف شده و ضرایب θ_1 و θ_2 بر اساس مجموعه قوانین فازی ۱۲گانه نمایش داده شده در شکل ۳ و یک سیستم استنتاج فازی ممدانی (شکل ۴) تعیین می‌گردند.

در این سیستم فازی برای تعیین مقادیر دو خروجی θ_1 و θ_2، ۳ ورودی UN، $f(Xdiffer)$ و IN به شرح زیر تعریف شده است

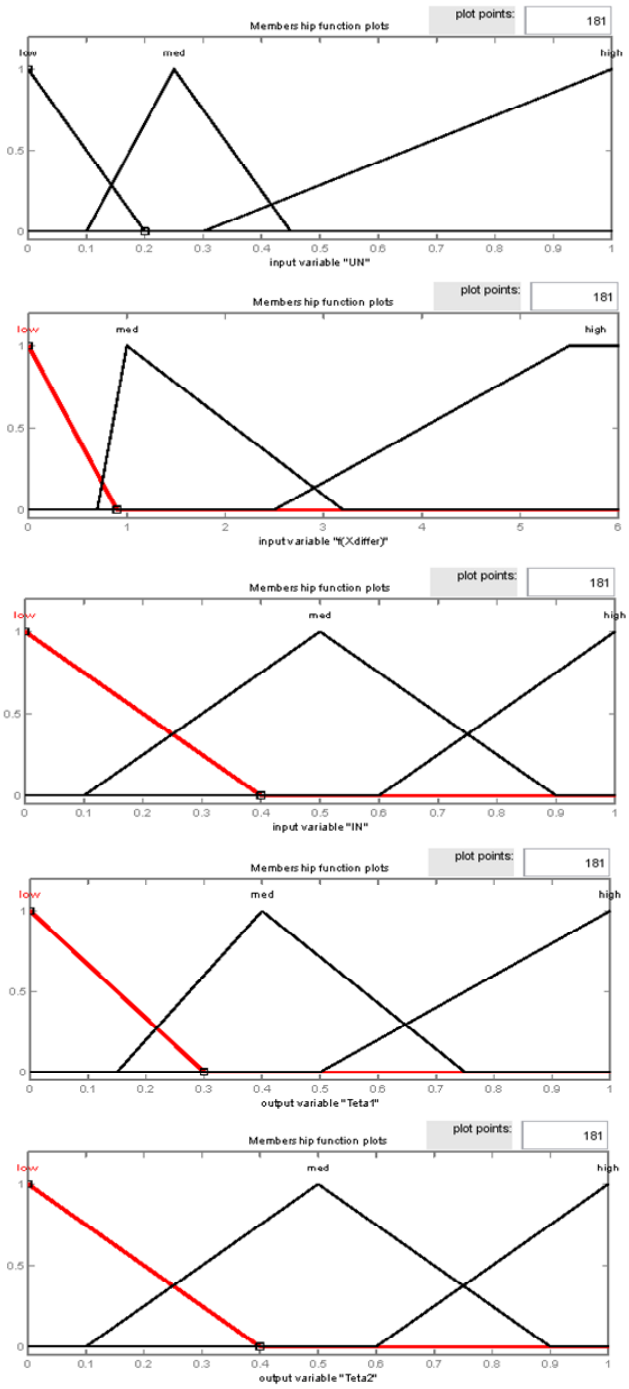
$$UN = \frac{Fail\_Iter}{\max\_Iter} \qquad (۱۹)$$

$$f(Xdiffer) = \frac{f(X_{Best}^{t-1}) - f(X_{Best}^{t})}{f(X_{Best}^{t-2}) - f(X_{Best}^{t-1}) + \varepsilon} \qquad (۲۰)$$

$$IN = \frac{t}{\max\_Iter} \qquad (۲۱)$$

که در آنها، $Fail\_Iter$ تعداد دفعاتی که بهترین جواب یافت‌شده توسط الگوریتم تغییری نکرده، t شماره تکرار جاری الگوریتم، $\max\_Iter$ حداکثر تکرارهای الگوریتم، ε یک عدد ثابت بسیار کوچک بزرگ‌تر از صفر برای اجتناب از صفرشدن کسر و به ترتیب مقادیر بهترین جواب‌های یافته‌شده در تکرارهای tام، $t-1$ام و $t-2$ام می‌باشند. تابع عضویت پارامترهای ورودی و خروجی این سیستم فازی در شکل ۵ نمایش داده شده است. همچنین پیاده‌سازی مجموعه قوانین ۱۲گانه فازی در این سیستم نیز به منظور محاسبه ضرایب در شکل ۶ قابل مشاهده می‌باشد.

در ادامه فرایند الگوریتم با تغییر موقعیت هر گوی، مجدداً میزان ارتفاع آن تعیین و فرایند فوق‌الذکر تا رسیدن به تکرار $\max\_Iter$ ادامه یافته و در نهایت بهترین گوی در تکرار $\max_Iter$ام به عنوان پاسخ الگوریتم استخراج می‌شود.

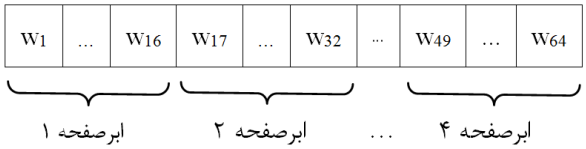


شکل ۵: توابع عضویت پارامترهای ورودی و خروجی.

۳-٤ طبقه‌بندی وب سایت‌ها با الگوریتم MIPO

مسئله طبقه‌بندی وب سایت‌ها در ۳ کلاس وب سایت‌های قانونی، مشکوک و فیشینگ در بخش ۳-۱ ارائه شد. با بررسی و در نظر گرفتن تعداد ۲، ۳، ۴ و ۵ طبقه‌بند بر روی یک نمونه کوچک تصادفی، مناسب‌ترین تعداد طبقه‌بند برای طبقه‌بندی وب سایت‌ها را ۴ انتخاب کردیم. همچنین در ادامه با در نظر گرفتن ۱۵، ۸ و ۳ ویژگی اثرگذار برتر بر اساس قواعد تعریف‌شده در بخش ۳-۲ اقدام به اجرای سه سطح طبقه‌بندی نمودیم تا اثر انتخاب پویای ویژگی‌ها را بر روی طبقه‌بندی ارزیابی نماییم.

وظیفه الگوریتم MIPO در طبقه‌بندی وب سایت‌ها تخمین مناسب بردار ضرایب (۱) به عنوان ضرایب ابرصفحات طبقه‌بندی می‌باشد به طوری که میزان خطای وب سایت‌های به اشتباه طبقه‌بندی شده (معادله



شکل ۷: تعداد ابعاد یک گوی در فضای جستجو بر اساس ۱۵ ویژگی.

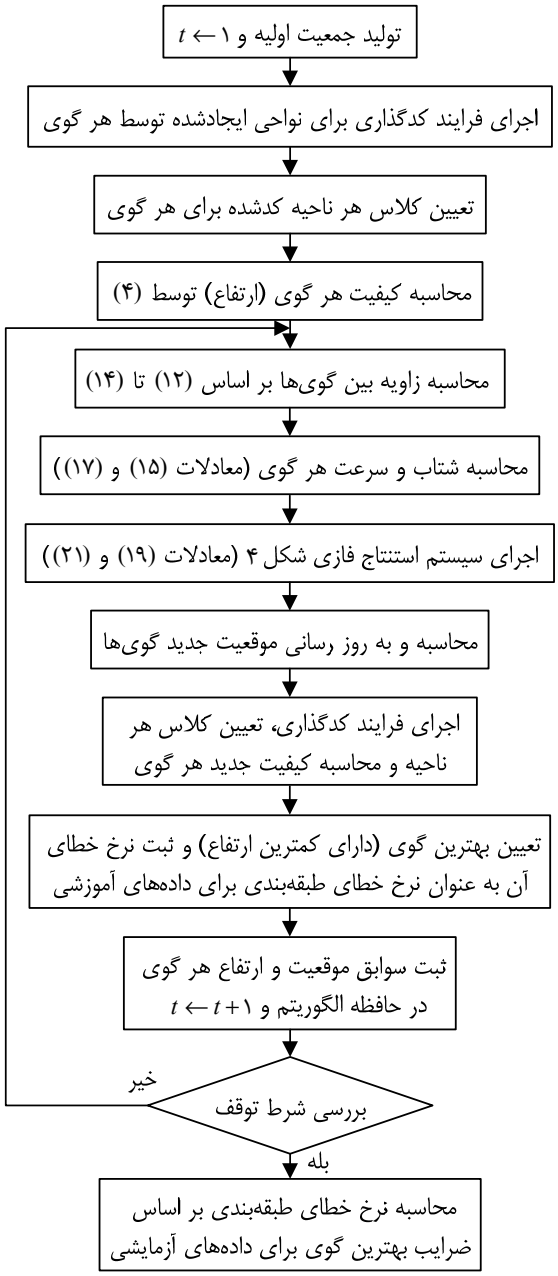
(۴)) حداقل شود. لذا هر گوی حاوی ضرایب ۴ ابرصفحه خواهد بود، بدین معنا که در طبقه‌بندی بر اساس ۱۵ ویژگی، هر گوی حاوی ۶۴ مؤلفه بوده و جستجو در فضایی ۶۴بعدی مانند شکل ۷ انجام خواهد گرفت. همچنین در طبقه‌بندی بر اساس ۸ و ۳ ویژگی به ترتیب هر گوی حاوی ۳۶ و ۱۶ مؤلفه می‌باشد و جستجو در فضای ۳۶ و ۱۶بعدی صورت خواهد گرفت. روندنمای فرایند طبقه‌بندی وب سایت‌ها به کمک الگوریتم MIPO در شکل ۸ نمایش داده شده است.

در گام نخست این فرایند، یک جمعیت اولیه از گوی‌ها مقدار اولیه می‌گیرند. مقدار هر مؤلفه گوی‌ها به طور تصادفی بین ۱- و ۱ انتخاب می‌شود. تعداد مؤلفه‌های گوی‌ها نیز متناسب با تعداد ویژگی‌های انتخاب‌شده برای طبقه‌بندی می‌باشد که قبل‌تر بدان اشاره شد. دومین گام، محاسبه کیفیت جواب‌های تولیدشده یا ارتفاع هر گوی می‌باشد. برای این منظور می‌بایست مقادیر مؤلفه‌های یک گوی که خود ضرایب ابرصفحات ۴ طبقه‌بند می‌باشند را در (۱) قرار داده و با قراردادن مقایر ویژگی‌های هر وب سایت از داده‌های آموزشی، اقدام به شناسایی 2^4 ناحیه ایجادشده در اثر تلاقی ۴ ابرصفحه طبقه‌بند یک گوی نمود. به وب سایت‌هایی که متعلق به یک ناحیه یکسان می‌باشند، یک کد باینری مطابق با آنچه در بخش ۲-۲ و جدول ۲ اشاره شد اختصاص داده می‌شود. بدین ترتیب هر ناحیه حاوی تعدادی وب سایت‌ها خواهد بود که برچسب کلاس آن ناحیه به کلاسی تعلق خواهد گرفت که بیشتر وب سایت از آن کلاس در ناحیه مذکور حضور داشته باشند. با شناسایی برچسب کلاس هر ناحیه می‌توان نرخ خطای طبقه‌بندی ایجادشده در اثر مؤلفه‌های آن گوی را محاسبه و به عنوان ارتفاع آن گوی در نظر گرفت.

گام بعدی محاسبه پارامترهای حافظه و زاویه بین هر گوی بوده که به وسیله آنها بتوان مقادیر شتاب و سرعت هر گوی را محاسبه کرد. برای این منظور از (۱۲) تا (۱۷) می بایست بهره برد. اجرای سیستم استنتاج فازی با سه پارامتر ورودی تعریف‌شده در (۱۹) تا (۲۱) به منظور تخمین مقادیر ضرایب θ_1 و θ_2 گام بعدی است که پیشگام محاسبه موقعیت‌های جدید گوی‌ها و به روز رسانی آنها می‌باشد. ثبت نرخ خطای بهترین گوی (دارای کمترین ارتفاع) به عناون نرخ خطای طبقه‌بندی داده‌های آموزشی و ثبت سوابق موقعیت و ارتفاع هر گوی در حافظه آخرین مراحل حلقه تکراری این فرایند می‌باشند. در پایان حلقه و در صورتی که شرط توقف محقق نشود، حلقه مجدداً با اجرای گام محاسبه زاویه بین گوی‌ها تکرار خواهد شود و در غیر این صورت حلقه خاتمه یافته و با استفاده از داده‌های آزمایشی اقدام به تست و محاسبه نرخ خطای طبقه‌بندی حاصله از ضرایب بهترین گوی می‌نماییم.

٤- نتایج شبیه‌سازی رویکرد پیشنهادی

پیاده‌سازی رویکرد پیشنهادی شامل انتخاب پویای ویژگی‌ها و نیز طبقه‌بندی وب سایت‌ها توسط MIPO به کمک نرم‌افزار متلب در محیط سیستم عامل ویندوز ۷ و پردازشگر Intel Core i۳ و RAM ۴ GB انجام گرفته است. شناسایی وب سایت‌های فیشینگ از طریق انتخاب پویای سه سطح از ویژگی‌ها با تعداد ۱۵، ۸ و ۳ ویژگی مورد بررسی قرار



شکل ۸: روندنمای طبقه‌بندی وب سایت‌ها با الگوریتم MIPO.

گرفته‌اند. نتایج حاصله جهت ارزیابی وب سایت‌های فیشینگ در جدول ۵ قابل مشاهده است. این نتایج با نتایج روش‌های فوق ابتکاری دیگر مثل الگوریتم کلونی مورچه‌ها (ACO)، بهینه‌سازی گروه ذرات (PSO)، الگوریتم غذایابی باکتری (BFOA) و الگوریتم خفاش بهبودیافته (MBAT) که توسط [۲۱] تا [۲۳] به دست آمده، در جدول ۶ مقایسه شده است. همچنین داده‌های محک این مقاله از مجموعه داده‌های http://www.phishtank.com که شامل لیست url وب سایت‌های فیشینگ بوده و به طور منظم به روز رسانی می‌شوند مستخرج گردیده است. تعداد وب سایت‌های فیشینگ، وب سایت‌های قانونی و وب سایت‌های مشکوک در داده‌های محک به ترتیب ۵۰۰، ۳۰۰ و ۲۰۰ می‌باشد. این داده‌ها در تاریخ ۱۱ جولای ۲۰۱۴ تا ۱۸ جولای ۲۰۱۴ و ۲۱ و ۲۲ آگوست ۲۰۱۴ جمع‌آوری شده است.

به منظور پیاده‌سازی فرایند طبقه‌بندی داده‌های محک، ۱۰۰۰ وب سایت کاندید برای طبقه‌بندی به دو دسته داده‌های آموزشی، مشتمل بر ۶۰۰ وب سایت و داده‌های آزمون، مشتمل بر ۴۰۰ وب سایت تقسیم‌بندی

جدول ۵: نتایج به دست آمده از شناسایی وب سایت فیشینگ توسط الگوریتم MIPO و IPO.

معیار ارزیابی	ارزیابی وب سایت مبتنی بر ۱۵ ویژگی		ارزیابی وب سایت مبتنی بر ۸ ویژگی		ارزیابی وب سایت مبتنی بر ۳ ویژگی	
	MIPO	IPO	MIPO	IPO	MIPO	IPO
نرخ خطا در طبقه‌بندی وب سایت‌های فیشینگ	٪۲	٪۲٫۸	٪۵	٪۶٫۸	٪۷٫۴	٪۹٫۸
نرخ خطا در طبقه‌بندی وب سایت‌های قانونی	٪۳	٪۲	٪۵٫۳۳	٪۸	٪۱۰	٪۱۳
نرخ خطا در طبقه‌بندی وب سایت‌های مشکوک	٪۴	٪۶٫۵	٪۵٫۵	٪۱۰	٪۱۳٫۵	٪۱۷٫۵
نرخ خطای طبقه‌بندی کل	٪۲٫۷	٪۳٫۳	٪۵٫۲	٪۷٫۸	٪۹٫۴	٪۱۲٫۳
تعداد کل داده‌ها	۱۰۰۰	۱۰۰۰	۱۰۰۰	۱۰۰۰	۱۰۰۰	۱۰۰۰
تعداد داده‌هایی که اشتباه طبقه‌بندی شده‌اند	۲۷	۳۳	۵۲	۷۸	۹۴	۱۲۳
تعداد داده‌هایی که درست طبقه‌بندی شده‌اند	۹۷۲	۹۶۷	۹۶۰	۹۲۲	۹۰۶	۸۷۷
زمان تشخیص الگوریتم (ms)	۶۴٫۵۲	۷۹٫۰۷	۳۸٫۱۱	۴۲٫۳۵	۱۸٫۷۸	۲۱٫۳

جدول ۶: خلاصه نتایج به دست آمده از MIPO و IPO و مقایسه آنها با الگوریتم‌های فوق ابتکاری دیگر.

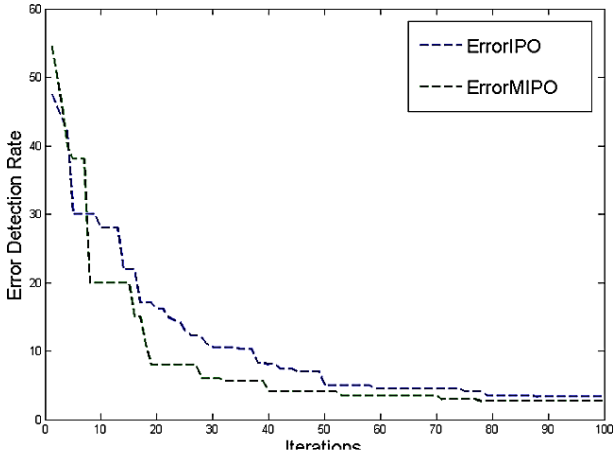
معیار ارزیابی	MIPO	IPO	MIPO	IPO	MIPO	IPO	PSO	ACO	BFOA	MBAT
	مبتنی بر ۱۵ ویژگی		مبتنی بر ۸ ویژگی		مبتنی بر ۳ ویژگی			مبتنی بر ۲۷ ویژگی		
نرخ دقت تشخیص	٪۹۷٫۳	٪۹۶٫۷	٪۹۴٫۸	٪۹۲٫۲	٪۹۰	٪۸۷٫۷	٪۹۲٫۱۳	٪۸۸٫۹۲	٪۹۱٫۳۸	٪۹۷٫۹۸
نرخ خطا	٪۲٫۷	٪۳٫۳	٪۵٫۲	٪۷٫۸	٪۹٫۴	٪۱۲٫۳	٪۷٫۸۷	٪۱۱٫۰۸	٪۸٫۶۲	٪۲٫۰۲
زمان تشخیص (ms)	۶۰٫۵۲	۷۹٫۰۷	۳۴٫۱۱	۴۲٫۳۵	۱۸٫۷۸	۲۱٫۳	۱۱۶٫۹	۱۲۴٫۵۴	۸۲٫۱۱	۸۲٫۲۸
شرط خاتمه	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰

شده‌اند. نتایج به دست آمده از داده‌های آموزشی و آزمایشی به طور جداگانه محاسبه شده و متوسط این نتایج به عنوان خروجی الگوریتم در جدول ۵ قابل مشاهده است.

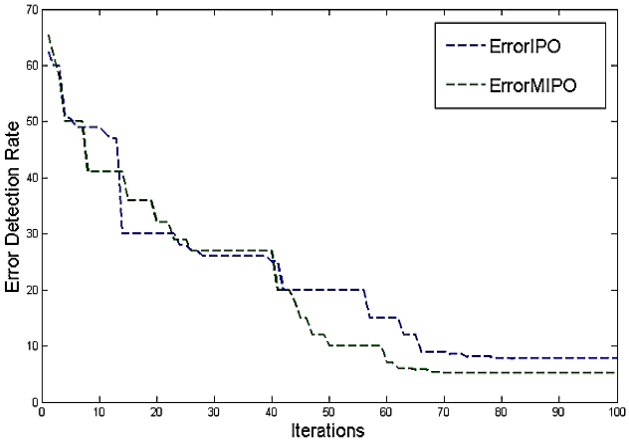
مقایسه نتایج پیاده‌سازی IPO و MIPO که در جدول ۵ ارائه شده، گویای عملکرد بهتر MIPO چه در نرخ خطای طبقه‌بندی و چه در سرعت همگرایی الگوریتم نسبت به نسخه اصلی خود یعنی IPO می‌باشد. این بدان معناست که طراحی ایده‌های حافظه‌مندی IPO و نیز امکان تغییر پویای پارامترهای الگوریتم که وظیفه تنظیم میزان وسعت جستجوهای سراسری و موضعی را به کمک یک مکانیزم استنتاج فازی بر عهده دارند تأثیر مثبت قابل توجهی در عملکرد الگوریتم بهینه‌سازی صفحات شیبدار داشته است. امری که در نمودارهای همگرای IPO و MIPO در شکل‌های ۹ تا ۱۱ قابل مشاهده می‌باشد. از سوی دیگر عملکرد مناسب هر دو نسخه اصلی و اصلاح‌شده الگوریتم بهینه‌سازی صفحات شیبدار روی ویژگی‌های کم می‌باشد به طوری که نرخ خطای طبقه‌بندی کل مبتنی بر ۱۵ برای IPO عدد ٪۳٫۳ و برای نسخه اصلاح‌شده آن ٪۲٫۷ را می‌دهد و هنگامی که با کاهش ۴۷ درصدی و ۸۰ درصدی در تعداد ویژگی‌ها روبه‌رو می شود، کیفیت طبقه‌بندی به ترتیب برای MIPO و IPO در حالت اول ٪۲٫۵ و ٪۴٫۵ و در حالت دوم تنها ٪۶٫۷ و ٪۹ کاهش یافته است. البته این افت کیفیت طبقه‌بندی را نیز تا حدود قابل توجهی زمان تشخیص الگوریتم‌های MIPO و IPO جبران نموده و با بهبود زمان تشخیص به ترتیب در حالت اول ٪۴۱ و ٪۴۶ و در حالت دوم ٪۶۷ و ٪۷۶ روبه‌رو می‌شود.

در جدول ۶ خلاصه نتایج پیاده‌سازی MIPO و IPO به مقایسه با نتایج روش‌های فوق ابتکاری معتبری که تا کنون برای این مسئله پیاده‌سازی شده‌اند، بر آمده است. خلاصه نتایج روش‌هایی همچون الگوریتم بهینه‌سازی ازدحام ذرات (PSO)، الگوریتم بهینه‌سازی کلونی مورچگان (ACO)، الگوریتم بهینه‌سازی غذایابی باکتری (BFOA) و الگوریتم اصلاح‌شده خفاش (MBAT) که در شناسایی وب سایت‌های فیشینگ مورد توجه قرار گرفته‌اند در جدول ۶ قابل مشاهده می‌باشند. شایان ذکر است که نتایج مقایسه‌شده MIPO و IPO از میانگین ارزیابی داده‌های آموزشی و داده‌های آزمایشی به دست آمده است در حالی که در [۲۱] تا [۲۳] اشاره‌ای به چگونگی دقت اعلام‌شده در این مقالات نشده است. همچنین تعداد ویژگی‌های مورد استفاده قرار گرفته شده توسط الگوریتم‌های PSO، ACO، BFOA و MBAT در طبقه‌بندی وب سایت‌ها، عدد ۲۷ را نشان می‌دهد در حالی که تعداد ویژگی‌های مورد استفاده در این تحقیق ۱۵، ۸ و ۳ ویژگی می‌باشند که به صورت انعطاف‌پذیر و متناسب با میزان تأثیرگذاری آنها انتخاب شده‌اند. به طور طبیعی با افزایش تعداد ویژگی‌ها در امر طبقه‌بندی، سطح کیفی طبقه‌بندی نیز با بهبود روبه‌رو می‌شود که در موارد بسیاری ایجاد هزینه زمانی و پیچیدگی محاسباتی آن توجیه استفاده از تعداد بیشتر ویژگی‌ها را رد می‌کند. لذا با توجه به این مهم، بهترین کیفیت طبقه‌بندی به الگوریتم MBAT با خطا ٪۲٫۰۲ تعلق دارد که شناسایی و طبقه‌بندی وب سایت‌ها را بر اساس ۲۷ ویژگی و در زمان ۸۲٫۲۸ میلی‌ثانیه به دست آورده است. این در حالی است که MIPO با بهره‌گیری از ۱۵ ویژگی توانسته نرخ خطا ٪۲٫۷ را در زمان ۶۰٫۵۲ میلی‌ثانیه به خود اختصاص دهد. انعطاف‌پذیری در انتخاب ویژگی‌ها و حجم محاسبات کمتر را می‌توان برتری MIPO در برابر MBAT نام برد. البته می‌بایست به دو نکته مجدداً اشاره شود که توجه به آنها، امر مقایسه عملکرد MIPO با سایر الگوریتم‌های ACO، BFOA و MBAT را با چالش روبه‌رو می‌سازد: اولاً وجود تفاوت‌هایی در داده‌های محک مورد استفاده از یک بانک اطلاعاتی مشابه به علت به روز رسانی بانک اطلاعاتی مذکور در تاریخ‌های مختلف و ثانیاً عدم درج این مهم که دقت شناسایی اعلام‌شده در [۲۱] تا [۲۳] نتیجه مستخرج از داده‌های تست است یا آزمون و یا میانگین هر دو.

لذا فارغ از توجه به نکات مهم و قابل تأمل طرح‌شده و با یک نگاه کلی مشاهده می‌شود که اگرچه بهترین کیفیت طبقه‌بندی به MBAT با ۲۷ ویژگی اختصاص یافته ولی الگوریتم‌های MIPO و IPO با ۱۵ و ۸ ویژگی موفق به طبقه‌بندی و شناسایی وب سایت‌های فیشینگ با کیفیت بهتری از الگوریتم‌های PSO، ACO و BFOA شده و زمان‌های بسیار بهتری را داشته‌اند. حتی عملکرد MIPO با ۳ ویژگی نیز قابلیت‌های این الگوریتم و ایده‌های طراحی‌شده در فرایند طبقه‌بندی و شناسایی وب



شکل ۹: نمودار همگرایی نرخ خطا در طبقه‌بندی مبتنی بر ۱۵ ویژگی برای الگوریتم‌های MIPO و IPO.

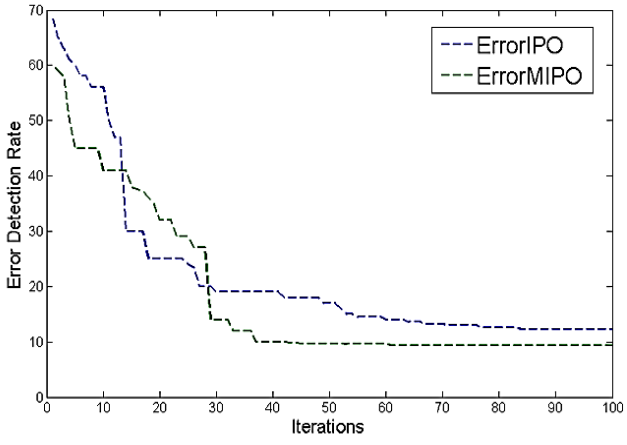


شکل ۱۰: نمودار همگرایی نرخ خطا در طبقه‌بندی مبتنی بر ۸ ویژگی برای الگوریتم‌های MIPO و IPO.

سایت‌های فیشینگ را در مقایسه با ACO با ۲۷ ویژگی به طور آشکاری نشان می‌دهد.

۵- نتیجه‌گیری

تهدید فیشینگ یا سرقت اطلاعات شخصی و مالی کاربران در فضای مجازی یکی از جدی‌ترین چالش‌های امنیتی پیش روی توسعه فناوری اطلاعات می‌باشد. محققان فراوانی تا کنون روی موضوع شناسایی وب سایت‌های فیشینگ تحقیق کرده و روش‌های متعددی را ارائه کرده‌اند که در این مقاله آنها را به ۵ گروه- رویکرد دسته‌بندی و تحلیل نمودیم. با مطالعه روش‌های موجود، مشاهده گردید که در این روش‌ها به طور هم‌زمان به ایجاد انعطاف‌پذیری در انتخاب ویژگی‌های اثرگذار در فرایند شناسایی وب سایت‌های فیشینگ، پویاسازی رفتار الگوریتم طبقه‌بندی کننده وب سایت‌های هدف و نیز امکان تحلیل و کنترل حجم گسترده‌ای از وب سایت‌ها مورد توجه قرار نگرفته است. لذا در این مقاله به منظور تحقق هم‌زمان سه هدف یادشده، یک رویکرد هوشمند جدید ارائه گردیده است. گام نخست با طراحی آستانه تغییر بر اساس ضرایب ۳۰٪ و ۶۰٪ و تعریف مجموعه قواعد شرطی روی ویژگی‌های بر اساس سطح تهدید فیشینگ (جدول ۴)، اقدام به انتخاب هوشمند و منعطف ویژگی‌های استخراج‌شده می‌نماید. در گام بعدی به منظور حافظه‌مندی IPO و انتقال اطلاعات موقعیت گوی‌ها از تکرارهای گذشته به تکرار جاری، (۱۲) و نیز به منظور کنترل اثرگذاری حافظه نیز یک تابع نزولی نرم (معادله (۱۳))



شکل ۱۱: نمودار همگرایی نرخ خطا در طبقه‌بندی مبتنی بر ۳ ویژگی برای الگوریتم‌های MIPO و IPO.

تعریف و در محاسبه موقعیت آتی هر گوی در نظر گرفته شده است. سومین گام نیز به منظور هوشمندسازی تنظیم ضرایب شتاب (اکتشاف) و سرعت (استخراج) بر اساس طراحی یک سیستم استنتاج ممدانی فازی (شکل‌های ۴ و ۵) و با استفاده از ۱۲ قانون فازی (شکل ۳) برداشته شده است.

نتایج پیاده‌سازی ایده‌های طرح‌شده در قالب نسخه کلاسیک IPO و نسخه بهبودیافته آن MIPO در جدول ۵ و مقایسه آن با الگوریتم‌های PSO، ACO، BFOA و MBAT [۲۱] تا [۲۳] در جدول ۶ نمایش داده شده است. اگرچه نکاتی همچون وجود تفاوت‌هایی در داده‌های محک مستخرج از یک بانک اطلاعات یکسان در نتیجه به روز رسانی آن و نیز عدم توضیح نحوه محاسبه نرخ تشخیص در منابع یادشده امر مقایسه نتایج حاصله را با چالش روبه‌رو می‌سازد ولی به طور کلی می‌توان نتیجه گرفت که مقایسه نتایج حاصل از پیاده‌سازی رویکرد هوشمند جدید پیشنهادی بر روی داده محک استاندارد در این حوزه و نیز مقایسه عملکرد این الگوریتم با عملکرد بهترین الگوریتم‌های موجود نشان می‌دهد که ایجاد انعطاف‌پذیری در انتخاب ویژگی‌ها ضمن بسترسازی برای انجام شناسایی وب سایت‌های فیشینگ با دقت بالا، باعث افزایش سرعت محاسبه با کاهش ابعاد مسئله می‌شود و نیز امکان انطباق‌پذیری با جامعه هدف‌های گوناگون را فراهم سازد. این مقاله راه را برای ورود به شناسایی خودکار وب سایت‌های فیشینگ بر روی حجم بزرگی از سایت‌ها باز می‌نماید.

مراجع

[1] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, *Phinding Phish: Evaluating Anti-Phishing Tools*, 2006.

[2] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7913-7921, Dec. 2010.

[3] M. D. I. A. Ajlouni, W. E. Hadi, and J. Alwedyan, "Detecting phishing websites using associative classification," *European J. of Business and Management*, vol. 5, no. 23, pp. 36-40, Aug. 2013.

[4] L. F. Cranor, S. Egelman, J. I. Hong, and Y. Zhang, "Phinding Phish: An Evaluation of Anti-Phishing Toolbars," in *Proc. of the 14th Annual Network & Distributed System Security Symp., NDSS'07,* San Diego, CA, 28 Feb.-2 Mar. 2007.

[5] M. Sirajuddin, "Data mining approach for deceptive phishing detection system," *International Journal of Scientific Research Engineering &Technology*, vol. 2, no. ???, pp. 337-334, ???. 2013.

[6] E. Medvet, E. Kirda, and C. Kruegel, "Visual-similarity-based phishing detection," in *Proc. of the 4th Int. Conf. on Security and Privacy in Communication Netowrks*, p. 22-28, Sept. 2008.

[۲۴] م. ح. مظفری، ح. عبدی و س. ح. ظهیری، "الگوریتم جدید بهینه سازی سیستم صفحات شیبدار،" *مجله رایانش نرم و فناوری اطلاعات*، سال ۱، شماره ۱، صص. ۲۰-۳، ۱۳۹۱.

[25] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah, "Associative classification techniques for predicting e-banking phishing websites," in *Proc. Int. Conf. on Multimedia Computing and Information Technology, MCIT'10,* pp. 9-12, Apr. 2010.

[26] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabatah, "Modelling intelligent phishing detection system for e-banking using fuzzy data mining," in *Proc. Int. Conf. on CyberWorlds, CW'09,* vol. ???, pp. 265-272, ???. 2009.

[27] P. Barraclough, M. Hossain, M. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4697-4706, Sept. 2013.

[28] S. H. Zahiri and S. A. Seyedin, "Intelligent particle swarm classifiers," *Iranian J. of Electrical and Computer Engineering*, vol. 4, no. 1, pp. 63-70, Winter-Spring 2005.

[29] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Experimental case studies for investigating E-banking phishing techniques and attack strategies," *Cognitive Computation*, vol. 2, no. 3, pp. 242-253, Sep. 2010.

[30] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," *IEEE Trans. on Dependable and Secure Computing*, vol. 3, no. 4, pp. 301-311, Oct.-Dec. 2006.

[31] R. Mohammad, T. McCluskey, and F. A. Thabtah, "Intelligent rule based phishing websites classification," *IET Information Security*, vol. 8, no. 3, pp. 153-160, May 2013.

مجید عبدالرزاق‌نژاد در سال ۱۳۸۲ مدرک کارشناسی ریاضیات کاربردی خود را از دانشگاه بیرجند و در سال ۱۳۸۴ مدرک کارشناسی ارشد ریاضیات کاربردی خود را از دانشگاه سیستان و بلوچستان با کار برروی حوزه محاسبات نرم و شبکه‌های عصبی دریافت نمود. از سال ۱۳۸۴ الی ۱۳۸۶ نام‌برده به عنوان مربی در دانشگاه صنایع و معادن شهرستان بیرجند و پس از آن تا سال ۱۳۸۸ در جهاد دانشگاهی مشهد به‌عنوان مدرس مدعو خدمت کرده و پس از آن به دوره دکترای علوم کامپیوتر گرایش محاسبات هوشمند در دانشگاه ملی مالزی وارد گردید. در سال ۱۳۹۲ موفق به اخذ درجه دکتری در علوم کامپیوتر از دانشگاه مذکور گردید. دکتر عبدالرزاق نژاد از سال ۱۳۹۲ در دانشکده فنی مهندسی قائن دانشگاه بیرجند به عضویت هیأت علمی درآمد و پس از استقلال این دانشکده از دانشگاه بیرجند تحت عنوان دانشگاه بزرگمهر قائنات تا کنون به‌عنوان عضو هیأت علمی دانشکده فنی و مهندسی و رئیس دانشکده علوم پایه این دانشگاه مشغول به فعالیت می‌باشد. زمینه‌های علمی مورد علاقه نام‌برده، انواع مسائل بهینه، مسائل زمانبندی، سیستم‌های نادقیق و منطق فازی، شبکه‌های عصبی، داده کاوی و کاربردهای آن، الگوریتم‌های فوق ابتکاری و ابرابتکاری می‌باشد.

[7] W. Zhang, H. Lu, B. Xu, and H. Yang, "Web phishing detection based on page spatial layout similarity," *Informatica*, vol. 37, no. 3, pp. 231-244, Sept 2013.

[8] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," in *Proc. Special Interest Tracks and Posters of the 14th Int. Conf. on World Wide Web*, pp. 1060-1061, Chiba, Japan, 10-14 May 2005.

[9] S. T. Kumar, V. Kumar, and A. Kumar, "Detection and Prevention of Phishing Attacks Using Linkguard Algorithm," 2008.

[10] J. S. White, J. N. Matthews, and J. L. Stacy, "A method for the automated detection phishing websites through both site characteristics and image analysis," in *Proc. SPIE Defense, Security, and Sensing*, 11 pp., May 2012.

[11] A. P. Rosiello, E. Kirda, C. Kruegel, and F. Ferrandi, "A layout-similarity-based approach for detecting phishing pages," in *Proc. 3rd Int.l Conf. on Security and Privacy in Communications Networks and the Workshops, SecureComm'07.*, pp. 454-463, Sept. 2007.

[12] N. R. T. Guhan, "Analyzing and detecting phishing webpages with visual similarity assessment based on earth mover's distance with linear programming model," *International J. of Advanced Engineering Technology*, vol. 3, no. 4, pp. 327-330, Nov. 2012.

[13] P. Barraclough, M. Hossain, M. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4697-4706, Sept. 2013.

[14] A. DeMaris and S. H. Selman, *Logistic Regression*, in Converting Data into Evidence, Ed: Springer, pp. 115-136, 2013.

[15] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proc. of the ACM Workshop on Recurring Malcode*, 8 pp., Nov. 2007.

[16] P. Sengar and V. Kumar, "Client-side defense against phishing with pagesafe," *International J. of Computer Applications*, vol. 4, no. 4, pp. 6-10, Jul. 2010.

[17] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, pp. 60-69, Oct. 2007.

[18] J. P. Marques de Sa, Pattern Recognition: Concepts, Methods, and Applications, Springer, 2001.

[19] H. M. Deylami and Y. P. Singh, "Cybercrime detection techniques based on support vector machines," *Artificial Intelligence Research*, vol. 2, no. 1, 12 pp., 2013.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[21] D. M. L. V. Radha Damodaram, "Experimental study on meta heuristic optimization algorithms for fake website detection," *International Association of Scientific Innovation and Research*, pp. 43-53, 2012.

[22] M. Radha Damodaram and M. Valarmathi, "Phishing website detection and optimization using particle swarm optimization technique," *International J. of Computer Science and Security*, vol. 5, no. 5, p. 477-490, Dec. 2011.

[23] M. Radha Damodaram and M. Valarmathi, "Bacterial foraging optimization for fake website detection," *International J. of Computer Science & Applications*, vol. 1, no. 11, pp. 116-127, Jan. 2013.