ICWR2020

کنفرانس بین المللی وب پژوهی
ایران، تهران، ۲۲ خرداد ۱۳۹۹

A Novel Phishing Detection Method Based on the Combination of Penguin Algorithm and Data Mining

Saba Malekpour Bejandi[1], Mohammad Reza Taghva[2], Payam Hanafizadeh[2]

[1]MSc Graduate in Management Information Systems, Industrial Management Department, Allameh Tabataba University, Tehran, Iran
saba.malekpour1993@gmail.com

[2]Associate Professor, Industrial Management Department, Allameh Tabataba University, Tehran, Iran
taghva@atu.ac.ir, hanafizadeh@gmail.com

Abstract

Recently, facility on the internet access in worldwide has caused many businesses take their activities on the internet affiliate networks. But security threats, such as phishing attacks, have always threaten these businesses. The multiplicity of web pages features has led to use of feature selection methods and their combination with machine learning methods to detect phishing. In this paper, Penguin metaheuristic algorithm and its performance are investigated to find the optimal response to phishing detection as the main contribution. Therefore, we propose a combination of Penguin algorithm in feature selection phase with artificial neural network in the phishing detection phase. Also, in order to train and evaluate our proposed method, a dataset with 11055 samples of phishing and normal websites is used. The results of our proposed method using the implementation in MATLAB software present that with increasing the population size and the number of iterations in penguin optimization algorithm, the average value of the feature selection function decreased by 69.57% and the RMSE index reduced by 24.56%. Finally, our proposed method shows about 29.16% lower error in phishing detection in comparison to multilayer artificial neural network.

Keywords: Feature Selection, Machine Learning Methods, Metaheuristic Algorithms, Phishing.

JCWR2020
کنفرانس بین المللی وب پژوهی
ایران، تهران، ۲۸ خرداد ۱۳۹۹

روش جدید تشخیص فیشینگ مبتنی بر ترکیب الگوریتم پنگوئن و داده‌کاوی

صبا ملک پور بجندی[۱]، محمدرضا تقوا[۲]، پیام حنفی زاده[۲]

[۱] دانش آموخته کارشناسی ارشد مدیریت فناوری اطلاعات،
گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران
saba.malekpour1993@gmail.com

[۲]دانشیار، دانشکده مدیریت و حسابداری، گروه مدیریت صنعتی، دانشگاه علامه طباطبائی، تهران
taghva@atu.ac.ir, hanafizadeh@gmail.com

چکیده

با دسترسی آسان به اینترنت، بسیاری از کسب و کارها فعالیت‌های خود را در شبکه‌های وابسته به اینترنت انجام می‌دهند. اما همواره مخاطرات امنیتی از جمله حملات فیشینگ این کسب و کارها را تهدید می‌کنند. تعدد ویژگی‌های صفحات وب، منجر به استفاده از روش‌های انتخاب ویژگی و ترکیب آنها با روش‌های یادگیری به منظور تشخیص فیشینگ شده است. عملکرد مناسب الگوریتم فراابتکاری پنگوئن در یافتن پاسخ بهینه، ایده اصلی این مقاله جهت بررسی نحوه عملکرد این الگوریتم در مسئله تشخیص فیشینگ بوده است. بنابراین از ترکیب الگوریتم پنگوئن در فاز انتخاب ویژگی با شبکه عصبی مصنوعی در فاز تشخیص فیشینگ استفاده شده است. برای آموزش و ارزیابی روش پیشنهادی از یک مجموعه داده با ۱۱۰۵۵ نمونه وبسایت‌های فیشینگ و عادی استفاده شده است. نتایج پیاده‌سازی در محیط متلب نشان می‌دهد با افزایش اندازه جمعیت و تعداد تکرار در الگوریتم بهینه‌سازی پنگوئن، مقدار متوسط تابع انتخاب ویژگی ۶۹.۵۷٪، و شاخص RMSE حدود ۲۴.۵۶٪ کاهش یافته است. همچنین روش پیشنهادی در مقایسه با شبکه عصبی مصنوعی چند لایه حدود ۲۹.۱۶٪ خطای کمتری در تشخیص فیشینگ را نشان می‌دهد.

کلمات کلیدی

فیشینگ، انتخاب ویژگی، الگوریتم‌های فراابتکاری، روش‌های یادگیری ماشین.

۱- مقدمه

فیشینگ یا سرقت اطلاعات[۱] به یک شیوه سرقت در فضای مجازی و اینترنت گفته می‌شود که در آن کاربران به سمت صفحات جعلی[۲] که ظاهر بسیار شبیه به صفحات قانونی[۳] دارند سوق داده می‌شوند. در این نوع حملات بیشتر سایت‌های تجارت الکترونیک مانند سایت آمازون[۴] مورد حمله قرار می‌گیرد تا اطلاعات مالی کاربران توسط هکر یا سارق آنلاین[۵] سرقت شود و سپس در ادامه این اطلاعات برای سوء استفاده سارق استفاده می‌شود[۱]. در فیشینگ نوعی جعل هویت صفحات وب انجام می‌شود به گونه‌ای که یک سایت که به کاربران اینترنت ارائه خدمات می‌دهد جعل می‌شود و صفحات مشابه آن در اینترنت بارگذاری می‌شود. در حملات فیشینگ هرزنامه[۶] و ایمیلهای جعلی نقش مهمی را بر عهده دارند زیرا به کمک این ابزار ارتباطی لینک‌های مرتبط با صفحات جعلی برای کاربران ارسال شده و با روش‌های فریب مانند مهندسی اجتماعی[۷] کاربران را ترغیب می‌کنند که به سمت صفحات جعلی هدایت شوند[2]. در برخی موارد روش‌های سرقت اطلاعات بر اساس انواع بدافزار[۸] مانند ویروس و تروجان[۹] انجام می‌شود؛ در این حالت بدافزاری که بر روی سیستم کاربران وجود دارد آنها را به صورت خودکار به سمت صفحات فیشینگ هدایت می‌نماید. روش‌های فیشینگ بیشتر مبتنی بر ارسال ایمیل به کاربران و فریب آنها می‌باشد.

روش‌های مختلفی برای مقابله با فیشینگ و حملات مبتنی بر آن وجود دارد تا تاثیر این حملات را کاهش دهند. یکی از روش‌های مقابله با فیشینگ استفاده از روش‌های مبتنی بر لیست سیاه[۱۰] است که در آن الگوهای صفحات جعلی و فیشینگ در یک پایگاه داده ذخیره می‌شوند. اگر یک وب‌سایت ملاقات شود، قبل از باز نمودن صفحات در مرورگر الگوی آن در لیست سیاه کنترل می‌شود و اگر تطبیق وجود داشته باشد آنگاه صفحه مورد نظر فیشینگ تشخیص داده می‌شود[3]. یکی دیگر از روش‌های تشخیص

فیشینگ استفاده از اطلاعات بصری و ظاهر سایت‌های بارگذاری شده است که می‌تواند به شناسایی فیشینگ کمک نماید اما این روش‌ها نیز تقریبی بوده و دقت بالایی ندارند[4]. روش‌های اکتشافی[11] یک نمونه دیگر از روش‌های تشخیص فیشینگ است اما دقت آنها به تعریف دقیق تابع اکتشافی و قوانین آن بستگی دارد[5]. روش‌های مبتنی بر یادگیری ماشین[12] و داده‌کاوی[13] نظیر شبکه عصبی مصنوعی[14][6] ، درخت تصمیم‌گیری[15][7] و ماشین بردار پشتیبان[16] [8] از جمله روش‌های تشخیص فیشینگ به شمار می‌روند که نسبت به روش‌های دیگر کارایی بیشتری دارند. ازگور و همکاران[9] برای تشخیص لینکهای جعلی مبتنی بر فیشینگ از یک روش مبتنی بر یادگیری ماشین استفاده نمودند. در این مقاله، یک سیستم ضد فیشینگ در زمان واقعی که از هفت الگوریتم طبقه بندی متفاوت و ویژگی‌های مبتنی بر پردازش زبان طبیعی (NLP) استفاده می‌کند، پیشنهاد شده است. نتایج تجربی و مقایسه‌ای این پژوهش نشان داد الگوریتم جنگل تصادفی با ویژگی‌های مبتنی بر NLP دارای دقتی مناسبی برای شناسایی URL های فیشینگ است.

به نظر می‌رسد مشکل بسیاری از روش‌های یادگیری ماشین برای تشخیص فیشینگ آن است که در این روش‌ها از بیشتر ویژگی‌های مرتبط با دامنه، کد منبع، لینک و موتورهای جستجو و غیره، که اهمیت و وزن یکسانی ندارند استفاده می‌شود. بنابراین میزان دقت روش‌های یادگیری ماشین به انتخاب بهینه ورودی‌ها بستگی دارد به گونه‌ای که اگر عمل یادگیری بر روی ویژگی‌های مهم صفحات جعلی انجام نشود، دقت این روش‌ها کاهش خواهد یافت. انتخاب ویژگی[17] مرتبط با صفحات فیشینگ و حملات مبتنی بر آن، دقت تکنیک‌های یادگیری ماشین مانند شبکه عصبی مصنوعی را افزایش می‌دهد.

در این راستا زبیح و دوران[10]، برای انتخاب ویژگی در حملات فیشینگ یک روش مبتنی بر یادگیری و تئوری ریاضیات راف[18] ارایه نمودند. در این پژوهش تلاش شده ویژگی‌های قطعی کاربردی در تشخیص فیشینگ با تئوری فازی راف انتخاب و شناسایی شوند. با وجود مزایای این روش در قطعیت مناسب برای انتخاب ویژگی، استفاده از تئوری ریاضی می‌تواند هوشمندی الگوریتم‌های فراابتکاری برای انتخاب ویژگی را نداشته باشد و از طرفی به علت استفاده از تئوری راف، دارای پیچیدگی بالایی نیز باشد.

در پژوهش حاضر برای انتخاب ویژگی مرتبط با حملات فیشینگ و استفاده موثر از آنها از الگوریتم‌های فراابتکاری[19]، و برای تشخیص صفحات جعلی یا فیشینگ از روش‌های یادگیری ماشین استفاده می‌شود. ادامه‌ی مقاله به صورت زیر سازماندهی شده است: در بخش دوم به مرور کارهای مرتبط پیشین میپردازیم. در بخش سوم روش پیشنهادی تشریح می‌شود. نتایج تجربی در بخش چهارم آورده شده است و در بخش پنجم به نتیجه‌گیری و کارهای آینده پرداخته شده است.

۲- مرورکارهای مرتبط پشین

انتخاب ویژگی یکی از روش‌های جدید و هوشمندانه برای افزایش دقت روش‌های یادگیری ماشین جهت تشخیص صفحات جعلی و فیشینگ است. تاکنون چندین مطالعه در این زمینه انجام شده است که از جمله آنها پژوهش چیو و همکاران[11] است که در آن برای تشخیص صفحات فیشینگ از یک چارچوب انتخاب ویژگی مبتنی بر الگوریتم توزیع توابع با مکانیزم روش-های گرادیان استفاده نمودند. در این روش در مرحله اول تعدادی بردار ویژگی تصادفی ایجاد شده و در فاز دوم با اعمال الگوریتم مبتنی بر گرادیان تلاش شده است تا بردارهای ویژگی به روزرسانی شوند تا بهینه‌ترین بردار ویژگی جهت تشخیص فیشینگ استخراج شود. در روش پیشنهادی آنها در فاز یادگیری از جنگل تصادفی برای طبقه‌بندی و آموزش استفاده شده است.

باباگلی[12] برای تشخیص صفحات و لینک‌های جعلی از یک روش مبتنی بر انتخاب ویژگی استفاده نمود. نتایج نشان می‌دهد الگوریتم wrapper در مقایسه با درخت تصمیم گیری با دقت بهتری انتخاب ویژگی را انجام داده است. الگوریتم جستجوی هارمونی نیز در مقایسه با الگوریتم ماشین بردار پشتیبان در تشخیص صفحات جعلی از دقت بالاتری برخوردار است.

فاریس و همکاران[13]، برای تحلیل و ارزیابی هرزنامه‌ها و تشخیص آنها یک روش مبتنی بر انتخاب ویژگی ارایه دادند و با استفاده از روش‌های تکاملی و شبکه عصبی مصنوعی تلاش کردند تا ویژگی‌های مهم در استخراج هرزنامه را محاسبه نمایند. روش آنها برای تشخیص هرزنامه در واقع تکیه بر بخش اول مراحل کشف دانش یعنی فاز استخراج ویژگی و سپس انتخاب ویژگی است که باعث می‌شود در مرحله اول کشف دانش ویژگی‌هایی برای یادگیری داده‌کاوی در نظر گرفته شود که در یادگیری و بهبود کیفیت الگوها نقش مهمی دارند.

وآنگ و همکاران[14]، برای تشخیص صفحات و لینکهای جعلی از روش‌های یادگیری عمیق مبتنی بر شبکه عصبی مصنوعی استفاده کردند. در این پژوهش الگوریتم LSTM دو طرفه مبتنی بر شبکه عصبی حلقوی و شبکه عصبی مکرر مستقل ارائه شده است. نتایج تجربی نشان می‌دهد که در مقایسه با سایر روشها، الگوریتم ترکیبی پیشنهادی آنها دقت تشخیص صفحات وب مخرب را بهبود بخشیده است.

با توجه به رشد حملات فیشینگ و متدهای انجام آنها در سال‌های اخیر، استفاده از شیوه‌های جدید برای تشخیص و کاهش خسارت‌های ناشی از این حملات بسیار حائز اهمیت است.

۳- روش پیشنهادی

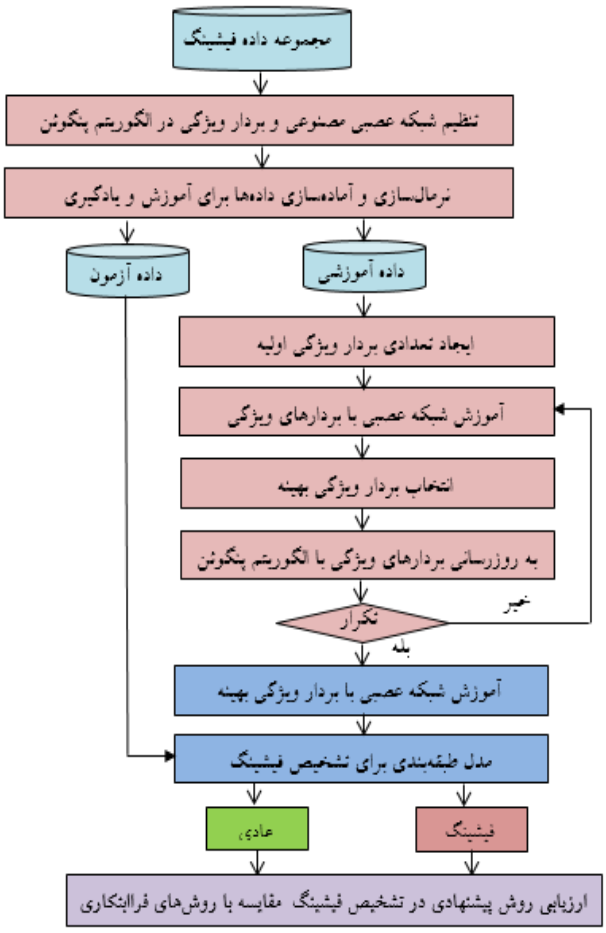
در این پژوهش، مسئله انتخاب ویژگی را در قالب یک مسئله بهینه‌سازی[20] مدلسازی نموده و تلاش شده است تا ویژگی‌های بهینه به گونه‌ای انتخاب شود که همزمان تعداد ویژگی و خطای تشخیص فیشینگ کمینه شود.

این روش توانایی یادگیری در داده‌کاوی و الگوریتم‌های فراابتکاری در تشخیص فیشینگ را در دو فاز ترکیب می‌نماید. در فاز انتخاب ویژگی از

الگوریتم بهینه‌سازی پنگوئن[15]، که یک الگوریتم هوش ازدحامی[۲۱] است استفاده می‌شود. با استفاده از این الگوریتم ویژگی بهینه انتخاب می‌شود و در فاز تشخیص صفحات جعلی یا فیشینگ از روش شبکه عصبی مصنوعی استفاده می‌شود. شکل (۱)، چارچوب کلی پیشنهادی برای تشخیص فیشینگ و لینک‌های جعلی را نمایش می‌دهد.

بر اساس این شکل، یک شبکه عصبی مصنوعی چند لایه برای تشخیص فیشینگ در نظر گرفته می‌شود. در اینجا شبکه عصبی مصنوعی دو لایه که در هر لایه آن ۵ نورون وجود دارد استفاده می‌شود. یک بردار ویژگی به عنوان یک پنگوئن یا عضوی از جمعیت الگوریتم پنگوئن در نظر گرفته می‌شود و این پنگوئن یک بردار باینری با مقادیر صفر و یک است.

تعدادی از این بردارهای ویژگی در مرتبه اول تصادفی ایجاد شده و هر کدام از آنها یک عضو الگوریتم پنگوئن به شمار می‌رود. هر بردار ویژگی بر روی مجموعه داده تاثیر گذاشته و داده‌های متناظر با آن برای آموزش و یادگیری در شبکه عصبی مصنوعی استفاده می‌شوند. هر بردار ویژگی مورد ارزیابی قرار می‌گیرد و اگر دو فاکتور متوسط خطای تشخیص صفحات فیشینگ و تعداد ویژگی آن کمینه باشد به عنوان بردار ویژگی بهینه انتخاب می‌شود.



شکل (۱): چارچوب پیشنهادی برای تشخیص فیشینگ و لینک‌های جعلی

با استفاده از الگوریتم بهینه‌سازی پنگوئن هر بردار ویژگی به‌روزرسانی می‌شود که در نتیجه آن این احتمال وجود دارد که مولفه‌های بردارهای ویژگی از حالت صفر و یک خارج شوند و مقادیر آنها غیرباینری شوند؛ لذا این بردارها با استفاده از تابع V مجدد باینری می‌شوند تا بتوان از آنها برای انتخاب ویژگی استفاده نمود. با تکرار متوالی الگوریتم بهینه‌سازی پنگوئن بردارهای ویژگی به‌روزرسانی شده و در تکرار آخر بهینه‌ترین بردار ویژگی با مقادیر بهینه انتخاب می‌شود. در نهایت بردار ویژگی بهینه در تکرار آخر برای آموزش شبکه عصبی مصنوعی برای تشخیص صفحات جعلی استفاده می‌شود.

۳-۱- شاخص‌های ارزیابی

برای پیاده‌سازی الگوریتم مورد استفاده به عنوان یک الگوریتم یادگیری و طبقه‌بندی کننده از شاخص مقایسه خطا استفاده می‌شود. برای این منظور از متوسط مجذور خطا، طبق رابطه (۱) استفاده می‌شود:

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(\tilde{o}_i - o_i)^2\right)} \qquad (۱)$$

در این رابطه، o_i و $\tilde{o}_i$ به ترتیب شماره کلاس واقعی و تقریبی یک وبسایت است که می‌تواند فیشینگ یا عادی باشد و در اینجا n تعداد نمونه‌ها در نظر گرفته می‌شود. برای ارزیابی روش پیشنهادی و سایر الگوریتم‌های تشخیص فیشینگ از شاخص‌های ارزیابی طبقه‌بندی صفحات جعلی از قانونی نظیر recall، precision و f-score به مانند رابطه (۲)، (۳) و (۴) استفاده می‌شود:

$$recall = Sensitivity = \frac{TP}{TP+FN} \qquad (۲)$$

$$Precision = \frac{TP}{TP+FP} \qquad (۳)$$

$$F1 = \frac{2*precision*recall}{precision+recall} \qquad (۴)$$

حساسیت نشان می‌دهد که دقت شناسایی صفحات فیشینگ از صفحات عادی چقدر است و صحت نیز نشان می‌دهد که روش پیشنهادی، صفحات وبی را که جعلی تشخیص داده است آیا واقعاً جعلی بوده‌اند یا به اشتباه جعلی تشخیص داده است.

۳-۲- مدلسازی روش ارائه شده

در این بخش به توصیف روش پیشنهادی پرداخته می‌شود. برای انتخاب ویژگی در این روش، باید معادلات و روابط موجود در الگوریتم بهینه‌سازی پنگوئن[15] استفاده شود، اما این الگوریتم دارای حالت پیوسته است و از این جهت نسخه باینری آن ارائه می‌شود تا بتوان به کمک آن انتخاب ویژگی موثری را انجام داد. یک بردار ویژگی از صفحات فیشینگ به مانند رابطه(۵)، تعریف شده و این بردار ویژگی یک پنگوئن است که دارای مقادیر باینری است و در مرحله اول تعدادی از این بردارهای ویژگی به صورت تصادفی و به عنوان جمعیت اولیه الگوریتم بهینه‌سازی پنگوئن مانند رابطه(۶)، در نظر گرفته می‌شوند:

$$EPO^i = [EPO_1^i, EPO_2^i, \dots, EPO_D^i] \quad (۵)$$

$$Pop = \{EPO^1, EPO^2, \dots, EPO^n\} \quad (۶)$$

در این روابط، $\boldsymbol{EPO^i}$ بردار ویژگی i – ام بوده و D تعداد ویژگی‌هایی است که در بردار ویژگی استفاده شده است و $\boldsymbol{Pop}$ جمعیت اولیه و n تعداد بردار ویژگی یا جمعیت پنگوئن‌ها است. در روش ارائه شده بردارهای ویژگی تحت تاثیر الگوریتم بهینه‌سازی پنگوئن قرار می‌گیرند تا به‌روزرسانی شوند. در روش ارائه شده نیاز است که هر بردار ویژگی برای آموزش شبکه عصبی مصنوعی برای تشخیص فیشینگ استفاده شود و هر بردار ویژگی دارای یک خطا است که می‌توان آن را طبق رابطه(۷)، تعریف نمود:

$$E = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k| \quad (۷)$$

در این رابطه، y_k شماره کلاس واقعی یک نمونه وب‌سایت است که می‌تواند دارای دو مقدار باشد که مقدار صفر و یک آن به ترتیب سایت عادی و غیرعادی فرض می‌شود و همچنین $\hat{y}_k$ شماره تخمینی یک نمونه وب‌سایت است که توسط روش ارائه شده و n نمونه محاسبه شده و E نیز متوسط خطای تشخیص فیشینگ از سایتهای عادی است. این رابطه فقط برای ارزیابی بردار ویژگی بکار گرفته نمی‌شود بلکه می‌توان تعداد ویژگی یک بردار ویژگی را هم ملاک قرار داد و اگر تعداد ویژگی انتخاب شده در یک بردار کمینه باشد نشان‌دهنده بهینه بودن بردار ویژگی است. برای کمینه نمودن این دو شاخص می‌توان تابع هدف را به صورت رابطه(۸)، برای ارزیابی بردارهای ویژگی معرفی نمود:

$$f = \alpha\frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k| + (1 - \alpha)\frac{FeatureSelect}{D} \quad (۸)$$

در این رابطه، α و β دو پارامتر تصادفی بین صفر و یک هستند که مجموع آنها برابر یک است. هر بردار ویژگی که بتواند این تابع را کمینه‌تر نماید به‌عنوان بردار ویژگی بهینه در هر تکرار برای تشخیص حملات فیشینگ استفاده می‌شود. هر کدام از بردارهای ویژگی مورد ارزیابی قرار می‌گیرند و بهینه‌ترین بردار ویژگی در تکرار t، اگر برابر $EPO^*(t)$ باشد و هر بردار ویژگی نیز با $EPO^i(t)$ نمایش داده شود، می‌توان از الگوریتم پنگوئن برای تغییر آنها استفاده نمود. برای این منظور در هر مرحله دمایی که دسته و گروه پنگوئن‌های امپراطور دارند باید طبق رابطه(۹)، به روزرسانی شوند و برای تغییر دما در هر تکرار از رابطه(۱۰)، نیز استفاده می‌شود[11]:

$$T' = T - \frac{Max_{iteration}}{t - Max_{iteration}} \quad (۹)$$

$$T = \begin{cases} 0 & rand > 0.5 \\ 1 & rand \le 0.5 \end{cases} \quad (۱۰)$$

در این رابطه‌ها، دمای کلونی برابر T' فرض شده و بیشترین تکرار الگوریتم نیز برابر $Max_{iteration}$ است و شماره تکرار الگوریتم نیز برابر t است. در رابطه دوم T دمایی است که می‌تواند افزایش یا تغییر ننماید. در الگوریتم بهینه‌سازی پنگوئن موقعیت فعلی یک پنگوئن یا بردار ویژگی در تکرار t با $EPO^i(t)$ در نظر گرفته می‌شود و فاصله آن در ابتدا از پنگوئن بهینه یا بردار ویژگی بهینه $EPO^*(t)$ نشان داده می‌شود مطابق رابطه(۱۱)، محاسبه می‌شود :

$$\vec{D}_{ep} = Abs\left(S(\vec{A}).EPO^*(t) - \vec{C}.EPO^i(t)\right) \quad (۱۱)$$

در این رابطه، $\vec{A}$ و $\vec{C}$ دو پارامتر برای هدایت جمعیت به سمت نقطه بهینه بوده به گونه‌ای که پارامتر $\vec{C}$ یک عدد تصادفی بین صفر و یک بوده و S نیز یک تابع به نام فاکتور تمایل به پیوستن در دسته اجتماعی است. برای توزیع پنگوئن‌ها در هر تکرار پیرامون پنگوئن بهینه نیاز است که یک فاصله برداری مانند رابطه(۱۲)، تعریف شود:

$$Pgrid = \left\|EPO^*(t) - EPO^i(t)\right\| \quad (۱۲)$$

با استفاده از پارامتر $Pgrid$ می‌توان پارامتر A در الگوریتم بهینه‌سازی پنگوئن طبق رابطه(۱۳)، محاسبه نمود که نقش آن فاصله گرفتن پنگوئن‌ها در پیرامون پنگوئن بهینه است:

$$A = M * (T' + Pgrid) * rand - T' \quad (۱۳)$$

در این رابطه، M در الگوریتم بهینه‌سازی پنگوئن برابر ۲ فرض می‌شود. برای سنجش فاصله یک پنگوئن از پنگوئن بهینه یا پارامتر $\vec{D}_{ep}$ نیاز است که تابع $S(\vec{A})$ تعریف شود که می‌توان مطابق رابطه(۱۴)، آن را ارائه داد:

$$S = \left(\sqrt{f * e^{-\frac{t}{l}} - e^{-t}}\right)^2 \quad (۱۴)$$

در تابع مورد نظر، f و l پارامترهای تصادفی در بازه $[۳،۲]$ و $[۱،۲،۵]$ بوده و e عدد نپر است. با تعاریف اولیه و مشخص شدن پارامترهای مهم می‌توان موقعیت جدید پنگوئن با رابطه(۱۵)، محاسبه نمود:

$$EPO^i(t+1) = EPO^*(t) - \vec{A}.\overrightarrow{D_{ep}} \quad (۱۳)$$

در این رابطه، $EPO^i(t+1)$ موقعیت جدید یک راه‌حل یا یک پنگوئن یا یک بردار ویژگی است که قبلاً در موقعیت $EPO^i(t)$ قرار داشته است.

٤- نتایج

برای پیاده‌سازی روش پیشنهادی از محیط برنامه‌نویسی متلب نسخه ۲۰۱۸ یا ۲۰۱۹، و برای ارزیابی و تحلیل آن از مجموعه داده‌های استاندارد رامی و همکاران[17,16] که از صفحات وب جعلی و قانونی گردآوری شده استفاده می‌شود. این مجموعه داده شامل ۱۱۰۵۵ نمونه وب‌سایت در اینترنت و ۳۰ ویژگی برای صفحات وب است که به عنوان ویژگی‌های ورودی در نظر گرفته شده و این مجموعه داده یک ویژگی خروجی نیز دارد که ویژگی ۳۱ مجموعه داده است. در این مجموعه داده خروجی می‌تواند منفی یک یا مثبت یک باشد که به ترتیب نشان دهنده سایت فیشینگ یا عادی است.

٤-١- تحلیل مجموعه داده در وکا

برای تحلیل و ارزیابی مجموعه داده فیشینگ از نرم‌افزار وکا استفاده شده است که می‌تواند تکنیک‌های یادگیری ماشین را بر روی مجموعه داده پیاده‌سازی نماید. برای این منظور از روابط (۲) و (۳) و (۴)، برای مقایسه الگوریتم پیشنهادی با سایر الگوریتم‌ها استفاده می‌شود. برای محاسبه هر یک از شاخص‌های این معادلات نیاز است که تعداد نمونه‌های صحیح مثبت[۲۲](TP) ، تعداد نمونه‌های غلط مثبت[۲۳](FP) ، تعداد نمونه‌های صحیح منفی[۲۴](TN) و تعداد نمونه‌های غلط منفی[۲۵](FN) محاسبه گردد. خروجی روش شبکه عصبی مصنوعی چند لایه، درخت تصمیم‌گیری، ماشین بردار پشتیبان و شبکه بیزین برای تشخیص فیشینگ در وکا در جدول (۱) نمایش داده شده است:

تجزیه و تحلیل مجموعه داده فیشینگ در وکا نشان می‌دهد حساسیت شبکه عصبی مصنوعی برای تشخیص فیشینگ بیش از همه روش‌ها بوده و در حدود۹۷.۴۰٪ است و کمترین میزان حساسیت مربوط به شبکه بیزین با مقدار ٪۹۲.۴۰ است. در مورد شاخص صحت نیز مشاهده می‌شود که شبکه عصبی مصنوعی چند لایه دارای بیشترین صحت بوده و این مقدار برابر ۹۷.۴۰٪ است و بدترین عملکرد هم مرتبط با شبکه بیزین می‌باشد. در شاخص RMSE کمترین خطای خروجی مربوط به عصبی مصنوعی چند لایه با مقدار ۱۴۴.۰ و بیشترین خطا مربوط به شبکه بیزین با مقدار ۰.۲۴۲ می‌باشد. بنابراین شبکه عصبی مصنوعی در سه شاخص حساسیت، صحت و RMSE بهترین عملکرد را در بین روش‌های مورد بررسی نشان می‌دهد.

٤-٢- خروجی الگوریتم مورد استفاده

خروجی روش پیشنهادی در واقع مقدار تابع هدف انتخاب ویژگی است. در پیاده‌سازی‌ها از شبکه عصبی مصنوعی دو لایه که در هر لایه آن ۵ نورون وجود دارد استفاده شده است و ۷۰٪ از داده‌ها آموزشی و ۳۰٪ از نمونه‌ها نیز آزمون است و سایر پارامترها نیز در جدول(۲)، برای پیاده‌سازی نمایش داده شده است.

دو نمونه خروجی یکی با جمعیت برابر ۵ و تکرار ۱۰ و دیگری با جمعیت ۱۰ و تکرار ۳۰ در محیط متلب برای پیاده سازی روش پیشنهادی در نظر گرفته شده و در این خروجی‌ها مقدار تابع هدف انتخاب ویژگی بر حسب تکرار در خروجی نمایش داده شده است. علت بررسی تابع هدف انتخاب ویژگی در خروجی آن است که اگر الگوریتم مورد استفاده بتواند این مقدار را بر حسب تکرار کاهش دهد نشان‌دهنده آن است که به طور غیرمستقیم عوامل آن را که متوسط خطای تشخیص فیشینگ از غیرفیشینگ و تعداد ویژگی انتخاب شده می‌باشد را کاهش داده است. در شکل‌های (۲) و (۳) مشاهده می‌شود که در هر دو خروجی روند کاهش تابع هدف انتخاب ویژگی نزولی و کاهشی است و این کاهش نشان می‌دهد که متوسط خطای تشخیص فیشینگ و تعداد ویژگی انتخاب شده در حال کاهش است.

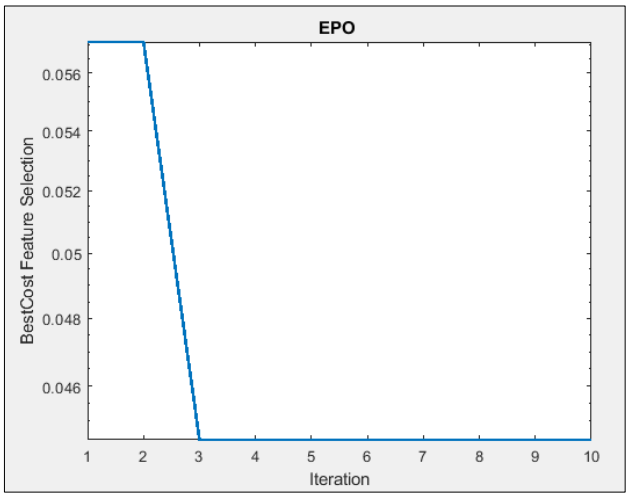
همانطور از خروجی نتایج آزمایشات می‌توان مشاهده کرد، در آزمایش اول خطا تا حدود ۰.۱۳۷، و در آزمایش دوم تا حدود ۰.۱۰۴ کاهش یافته

جدول(۱): مقایسه روش‌های پایه برای تشخیص فیشینگ

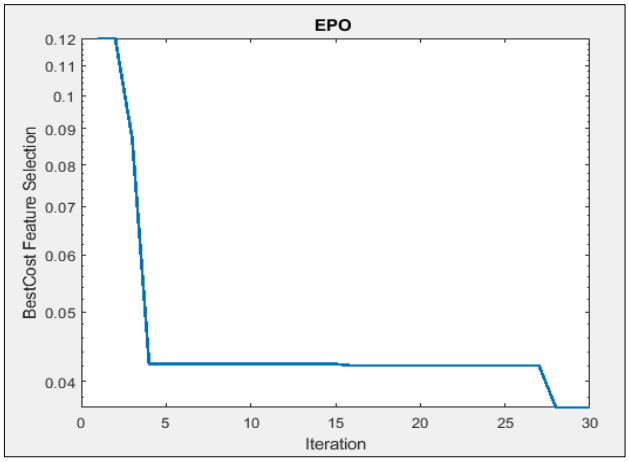
روش	شاخص صحت	شاخص حساسیت	RMSE
MLP	۹۷.۴۰	۹۷.۴۰	۰.۱۴۴
J48	۹۶.۱۰	۹۶.۱۰	۰.۱۸۱
SVM	۹۴.۰۰	۹۴.۰۰	۰.۲۴۴
BN	۹۲.۴۰	۹۲.۴۰	۰.۲۴۲

جدول(۲): پارامترهای پیاده‌سازی الگوریتم پیشنهادی

پارامترهای پیاده‌سازی	مقدار پیشنهادی هر پارامتر
N pop	اندازه جمعیت اولیه که در اینجا ۱۰ در نظر گرفته شده
Max iter	حداکثر تعداد تکرار الگوریتم مورد استفاده که برابر ۳۰ است
f	۱.۵
I	۲
C	یک عدد تصادفی بین صفر و یک
L	تعداد لایه شبکه عصبی که برابر ۲ است
n	تعداد نورون‌های لایه پنهان که برابر ۵ است



شکل(۲): کاهش یافتن مقدار تابع هدف با جمعیت ۵ و تکرار ۱۰



شکل(۳): کاهش یافتن مقدار تابع هدف با جمعیت ۱۰ و تکرار ۳۰

است. همچنین مقدار تابع هدف انتخاب ویژگی بر حسب تکرار کاهش یافته که این نشان می‌دهد مولفه خطای تشخیص فیشینگ و تعداد ویژگی نیز کاهش یافته است.

٤-٣- تحلیل تابع انتخاب ویژگی و شاخص متوسط خطا (RMSE) در روش پیشنهادی

برای ارزیابی روش پیشنهادی جمعیت الگوریتم مورد استفاده تغییر یافت و مقدار تابع هدف برای انتخاب ویژگی و شاخص متوسط خطای تشخیص فیشینگ مورد بررسی قرار گرفت. در جدول(۳)، متوسط تابع هدف انتخاب ویژگی و خطای RMSE به ازای جمعیت‌های مختلف و تعداد تکرار ۳۰ نمایش داده شده است و هر آزمایش ۲۵ مرتبه تکرار شده و متوسط آزمایشات بیان شده است.

مشاهده می‌شود که با افزایش اندازه جمعیت الگوریتم مورد استفاده مقدار تابع انتخاب ویژگی در حدود۶۹.۵۷٪ کاهش یافت و در واقع افزایش جمعیت با وجود افزایش زمان اجراء می‌تواند تابع انتخاب ویژگی را کاهش داده و از این طریق خطا نیز بیشتر کاهش داده می‌شود. همچنین متوسط خطای تشخیص فیشینگ با شاخص RMSE به ازای جمعیت‌های مختلف در حدود ۲۴.۵۶٪ کاهش را نشان داد.

در واقع با افزایش اندازه جمعیت می‌توان احتمال یافتن بردارهای ویژگی بهینه برای تشخیص فیشینگ را افزایش داد و از این طریق خطای تشخیص فیشینگ کاهش خواهد یافت. با افزایش اندازه جمعیت خطای تشخیص فیشینگ به علت جستجوی بهینه فضای ویژگی کاهش مطلوبی نشان داده است.

٤-٤- مقایسه تابع هدف و شاخص RMSE روش پیشنهادی با سایر روش‌ها

در این بخش روش پیشنهادی با الگوریتم‌های مشابه مانند الگوریتم بهینه‌سازی پروانه، بهینه‌سازی کفتار و بهینه‌سازی وال در مقدار تابع هدف و شاخص RMSE، مقایسه می‌شود. مطابق شکل (۴)، روش پیشنهادی در مقدار تابع هدف با سایر الگوریتم‌ها به ازای مقدار جمعیت ۱۰ و تعداد تکرار ۳۰ مورد مقایسه قرار گرفته و هر آزمایش ۲۵ مرتبه تکرار شده است.

بر اساس شکل (۴)، نتایج نشان می‌دهد که مقدار تابع انتخاب ویژگی در تشخیص فیشینگ در روش پیشنهادی با مقدار ۰.۰۱۸۵۴ نسبت به سایر الگوریتم‌های فرابتکاری دارای مقدار کمینه‌تری است و ضعیف‌ترین عملکرد نیز مربوط به الگوریتم بهینه‌سازی پروانه است. بنابراین کمینه بودن مقدار تابع هدف توسط روش پیشنهادی نشان می‌دهد در مجموع خطای کمتری برای تشخیص فیشینگ دارد و از طرفی ویژگی‌های کمتری انتخاب نموده است.

در شکل (۵)، روش پیشنهادی در شاخص متوسط خطا(RMSE) با سایر الگوریتم‌ها مقایسه می‌شود. مقدار جمعیت برابر ۱۰ و تعداد تکرار برابر ۳۰ است و هر آزمایش ۲۵ مرتبه تکرار شده است. خطای تشخیص فیشینگ با شاخص RMSE در روش پیشنهادی با مقدار ۰.۱۰۲ نسبت به سایر الگوریتم‌های مورد بررسی دارای مقدار کمینه‌تری است و الگوریتم بهینه‌سازی پروانه در مقایسه با سایر روش‌ها ضعیف‌ترین عملکرد را نشان می‌دهد.

همچنین بر اساس شکل(۶)، نتایج مقایسه شاخص متوسط خطای RMSE در روش پیشنهادی با روش‌های پایه داده‌کاوی از جمله شبکه عصبی مصنوعی چند لایه با نرخ یادگیری ۰.۳ و حداکثر زمان یادگیری برابر ۵۰۰ ثانیه، درخت تصمیم‌گیری با ضریب اطمینان برابر ۰.۲۵ و seed برابر ۱، ماشین بردار پشتیبان با ضریب جریمه ۱.۰، آستانه خطا برابر 10^{-12} و کرنل از نوع چند جمله‌ای، و شبکه بیزین نشان می‌دهد خطای روش پیشنهادی نسبت به خطای روش پایه یا شبکه عصبی چند لایه در حدود ۲۹.۱۶٪ کاهش یافته است.

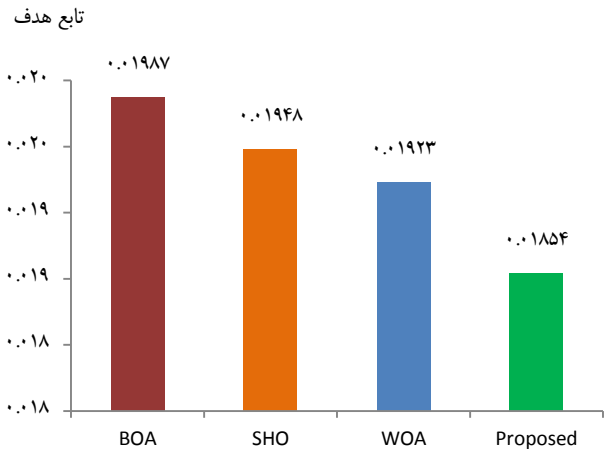
برای ارزیابی روش پیشنهادی می‌توان در شاخص‌های recall، precision و f-score روش پیشنهادی را با سایر روش‌های فرابتکاری مورد مقایسه و تحلیل قرار داد که در اینجا اندازه جمعیت روش پیشنهادی و هر یک از الگوریتم‌ها و تعداد تکرار آنها به ترتیب ۱۵ و ۳۰ در با توجه به نمودار مقایسه‌ای شکل(۷)، شاخص recall، precision و f-score در روش پیشنهادی به ترتیب برابر ۹۸.۳۳٪، ۹۸.۳۸٪ و ۹۸.۲۶٪ است و در این شاخص‌ها از الگوریتم بهینه‌سازی پروانه، بهینه‌سازی کفتار و بهینه‌سازی وال عملکرد بهتری را نشان می دهد.

۵- جمع‌بندی و کارهای آتی

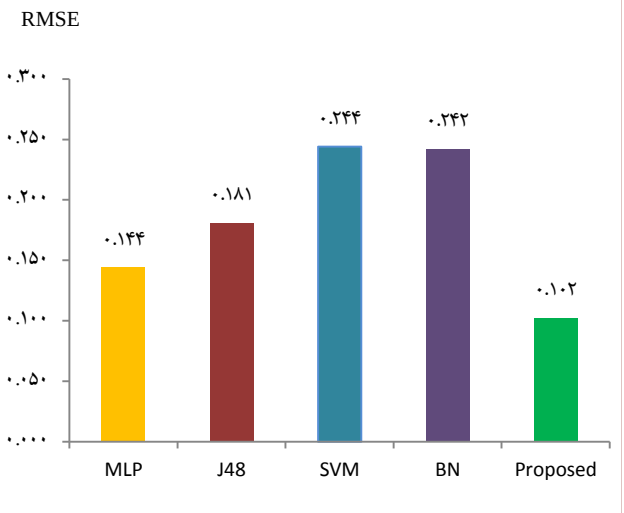
استفاده از شبکه عصبی مصنوعی چند لایه به دلیل سادگی و توانایی در تشخیص الگو و طبقه‌بندی باعث شد که در پژوهش حاضر از این روش استفاده شود. برای کاهش دادن خطا نیز از فرآیند انتخاب ویژگی به کمک الگوریتم بهینه‌سازی پنگوئن استفاده شد که یک الگوریتم فرابتکاری جدید است و می‌توان از آن در مسائل بهینه‌سازی استفاده کرد. در روش پیشنهادی ویژگی‌های مختلفی توسط الگوریتم بهینه‌سازی پنگوئن ارزیابی شد تا دو هدف خطا و تعداد ویژگی همزمان کاهش داده شود. مشاهده شد که تابع هدف که بر اساس این دو مولفه است مرتباً در حال کاهش است. بنابراین هم تعداد ویژگی و هم متوسط خطا کاهش یافته است. نتایج نشان داد در مسئله تشخیص فیشینگ، ترکیب الگوریتم هوش ازدحامی پنگوئن برای انتخاب بردار ویژگی بهینه و آموزش شبکه عصبی بر روی مهمترین ویژگی‌ها، از دقت مناسبی برخوردار است.

جدول(۳): خروجی روش پیشنهادی به ازای جمعیت‌های مختلف

	۲۰	۱۵	۱۰	۵
تابع هدف	۰.۰۰۸۶۴	۰.۰۱۲۳۷	۰.۰۱۸۵۴	۰.۰۲۸۴۰
RMSE	۰.۰۸۶	۰.۰۹۲	۰.۱۰۲	۰.۱۱۴



شکل(٤): مقایسه متوسط تابع انتخاب ویژگی در الگوریتم مورد استفاده و سایر روشهای فراابتکاری



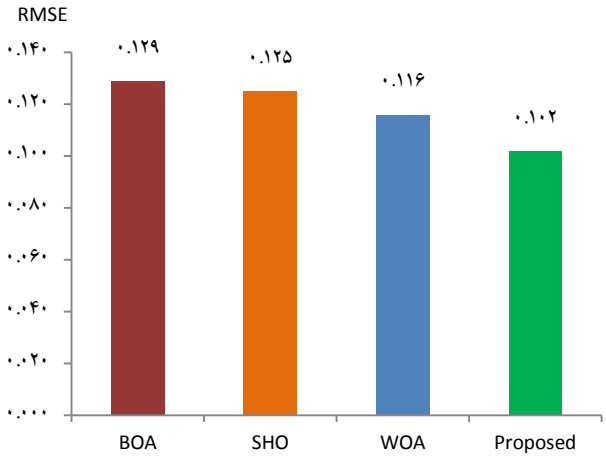
شکل(٦): مقایسه متوسط خطای الگوریتم مورد استفاده با روشهای داده کاوی

در پژوهش های آتی می‌توان از فاز استخراج ویژگی قبل از انتخاب ویژگی استفاده کرد تا از ویژگی های جدیدتر در فرایند یادگیری استفاده شود و دقت روش‌های یادگیری افزایش یابد. همچنین به منظور بهبود سرعت پردازش و انتخاب ویژگی می‌توان از روش‌های موازی‌سازی استفاده کرد.

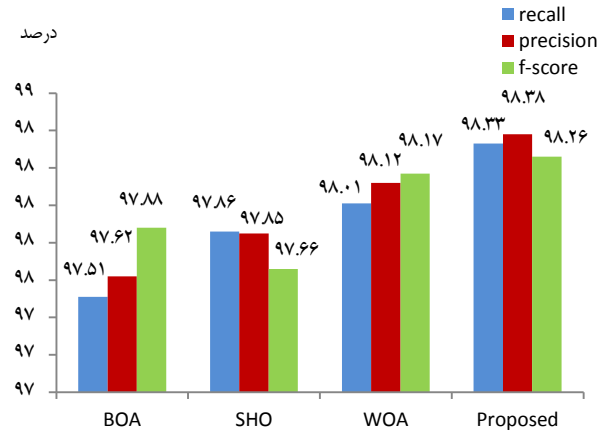
مراجع

[1] Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018, August). *Detection and Prevention of Phishing Websites Using Machine Learning Approach.* In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[2] Aldawood, H., & Skinner, G. (2019, January). *An academic review of current industrial and commercial cyber security social engineering solutions.* In Proceedings of the 3rd International Conference on Cryptography, Security and Privacy (pp. 110-115). ACM.



شکل(٥): مقایسه متوسط خطای الگوریتم مورد استفاده و سایر روشهای فراابتکاری در تشخیص فیشینگ



شکل(۷): مقایسه شاخص‌های طبقه بندی الگوریتم پیشنهادی و سایر الگوریتم های فراابتکاری

[3] Jain, A. K., & Gupta, B. B. (2016). *A novel approach to protect against phishing attacks at client side using auto-updated white-list.* EURASIP Journal on Information Security, 2016(1), 9.

[4] Jain, A. K., & Gupta, B. B. (2017). Phishing detection: *analysis of visual similarity based approaches. Security and Communication Networks, 2017.*

[5] Mao, J., Tian, W., Li, P., Wei, T., & Liang, Z. (2017). *Phishing-alarm: robust and efficient phishing detection via page component similarity.* IEEE Access, 5, 17020-17030.

[6] Gajera, K., Jangid, M., Mehta, P., & Mittal, J. (2019, June). *A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection.* In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 196-200). IEEE.

[7] Kulkarni, A. (2019). *Phishing Websites Detection using Machine Learning.*

[8] Wu, S., Tong, X., Wang, W., Xin, G., Wang, B., & Zhou, Q. (2018, May). *Website Defacements Detection Based on Support Vector Machine Classification*

Method. In Proceedings of the 2018 International Conference on Computing and Data Engineering (pp. 62-66). ACM.

[9] Sahingoz, Ozgur Koray, Ebubekir Buber, Onder Demir, and Banu Diri. *"Machine learning based phishing detection from URLs."* Expert Systems with Applications 117 (2019): 345-357.

[10] Zabihimayvan, Mahdieh, and Derek Doran. *"Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection."* arXiv preprint arXiv:1903.05675 (2019).

[11] Chiew, Kang Leng, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong. *"A new hybrid ensemble feature selection framework for machine learning-based phishing detection system."* Information Sciences 484 (2019): 153-166.

[12] Babagoli, M., Aghababa, M. P., & Solouk, V. (2019). *Heuristic nonlinear regression strategy for detecting phishing websites. Soft Computing*, *23*(12), 4315-4327.

[13] Faris, H., Ala'M, A. Z., Heidari, A. A., Aljarah, I., Mafarja, M., Hassonah, M. A., & Fujita, H. (2019). *An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks*. Information Fusion, 48, 67-83.

[14] Wang, H. H., Yu, L., Tian, S. W., Peng, Y. F., & Pei, X. J. (2019). *Bidirectional LSTM Malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network.* Applied Intelligence, 1-11.

[15] Dhiman, G., & Kumar, V. (2018). *Emperor penguin optimizer: A bio-inspired algorithm for engineering problems.* Knowledge-Based Systems, 159, 20-50.

[16] Mohammad, R., Thabtah, F. A., & McCluskey, T. L. (2015). *Phishing websites dataset.*

[17] https://archive.ics.uci.edu/ml/datasets/phishing+websites

٦- زیرنویس‌ها

[1] Phishing
[2] Fake Pages
[3] Legal pages
[4] Amazon
[5] Phisher
[6] Spam
[7] Social Engineering
[8] Malware
[9] Trojan
[10] Blacklist
[11] Heuristic methods
[12] Machine learning
[13] Data mining
[14] Artificial neural network
[15] Decision tree
[16] Support Vector Machine
[17] Feature selection
[18] Fuzzy Rough Set (FRS)
[19] Metaheuristic algorithms
[20] Optimization problem
[21] Swarm intelligence
[22] True Positive(TP)
[23] False Positive(FP)
[24] True Negative (TN)
[25] False Negative(FN)