شناسایی فیشینگ[1]: یک مدل عصبی فازی کارآمد بدون استفاده از مجموعه قوانین

چکیده

اینترنت مزایای فراوانی برای انسان به همراه داشته است، اما می تواند خطرات احتمالی ای نیز به دنبال داشته باشد. جرایم اینترنتی به سرعت در حال رشد هستند، فیشینگ یکی از انواع جدید جرایم آنلاین است. سایت فیشینگ یک سایت جعلی است که هدف آن سرقت اطلاعات شخصی مانند رمز عبور، اطلاعات حساب بانکی و کارت اعتباری و غیره است. اکثر این صفحات فیشینگ از لحاظ لینک و آدرس نشانه وب یا مکان یکنواخت منبع (URL)[2]مشابه صفحات واقعی هستند. تکنیک های بسیاری برای شناسایی سایت های فیشینگ پیشنهاد شده است. با این حال، به دلیل تکنیک های ناکارآمد حفاظت، تعداد قربانیان این موضوع در حال افزایش است. در این مقاله، یک مدل عصبی فازی برای شناسایی موثر فیشینگ توسعه می دهیم. مدل، عوامل ذهنی را حذف می کند تا کارآیی هایی مانند مجموعه قوانین اگر-پس[3]، پارامترهای تابع عضویت و غیره را بهبود بخشد. بعلاوه، ویژگی های کارآمد جهت شناسایی فیشینگ برای مدل عصبی فازی مورد استفاده قرار گرفت. اثربخشی تکنیک پیشنهادی در پایگاه داده های مقیاس بزرگ بررسی شد که از سایت های فیشینگ و سایت های قانونی جمع آوری شد. نتایج نشان داد که تکنیک پیشنهادی م یتواند بیش از ۹۹٪ از سایت های فیشینگ را شناسایی کند.

کلمات کلیدی: فیشینگ، مبتنی بر URL، عصبی فازی.

۱. مقدمه

فیشر[4] از تعدادی تکنیک برای فریب دادن قربانیانش استفاده می کند که هدف آن سرقت اطلاعات شخصی شامل پیام های ایمیل، پیام های فوری، پست های انجمن، تماس های تلفنی و شبکه های اجتماعی است. این فعالیت های

[1] Phishing
[2] Uniform resource locator
[3] If-then rule sets
[4] Phisher

فیشینگ منجر به خسارت شدید اقتصادی در سراسر جهان می شود. بر طبق مجله ی Fortune در سال ۲۰۱۱، ۸۳٪ از آمریکایی ها و ۸۵٪ از اروپایی ها مرتبا به صورت آنلاین خرید می کنند. در همین حال، سایت های فیشینگ نیز به سرعت از لحاظ کیفیت و کمیت در حال رشد هستند. بنابراین، خطر سرقت اطلاعات کاربر بسیار بالا است. به این دلایل، شناسایی مشکلات فیشینگ بسیار فوری، پیچیده و مهم است.

در این مقاله، یک روش کارآمد برای شناسایی سایت های فیشینگ که بر ویژگی های URL (Primary Domain, SubDomain, PathDomain) و پارامترهای گوگل (PageRank, BackLink, GoogleIndex) متمرکز است ارائه شده است. پس، یک مدل عصبی فازی پیشنهادی سیستمی است که خطا را کاهش و عملکرد را افزایش می دهد. مدل عصبی فازی از مدل های محاسباتی برای اجرا بدون استفاده از مجموعه قوانین اگر-سپس (if-then) استفاده می کند. تکنیک پیشنهادی دقت شناسایی بالای ۹۹٪ با سیگنال نادرست پایین را بدست آورد.

باقی مقاله به صورت زیر سازمان دهی شده است: بخش II کارهای مرتبط را ارائه می دهد. طرح سیستم در بخش III نشان داده شده است. بخش IV دقت روش را ارزیابی می کند. در نهایت، بخش V از مقاله نتیجه گیری می کند و کارهای آتی را مورد بررسی قرار می دهد.

II. کارهای مربوطه

روش های شناسایی فیشینگ می تواند به سه گروه تقسیم شود: فهرست سیاه[5]، اکتشافی و یادگیری ماشین. تکنیک مبتی بر فهرست سیاه (بلک لیست) (۱،۲،۳،۴) دارای فهرستی از وبسایت های فیشینگ به نام فهرست سیاه است. تکنیک به دلیل رشد سریع تعداد سایت های فیشینگ ناکارآمد است. بنابراین، رویکردهای اکتشافی و یادگیری ماشین توجه محققان بیشتری را به خود جلب کرده است.

Cantina (۵) الگوریتم TF-IDF را بر مبنای ۲۷ ویژگی صفحات وب ارائه کرده است.این تکنیک می تواند ۹۷٪ از سایت های فیشینگ با ۶٪ مثبت کاذب را شناسایی کند. اگرچه این تکنیک کارآمد است، اما زمان استخراج ۲۷ ویژگی صفحه وب بسیار طولانی است تا تقاضای زمان واقعی را برآورده کند و برخی از ویژگی های برای بهبود دقت

[5] Blacklist

شناسایی فیشینگ ضروری نیستند. به همین نحو، Cantina+ (6) از تکنیک های یادگیری ماشین بر اساس ۱۵ ویژگی صفحه وب استفاده کرد و تنها شش مورد از ۱۵ ویژگی برای شناسایی فیشینگ مانند فرم بد، فیلد فعالیت بعد، URL غیرمنطبق، صفحه در بالای نتیجه ی جستجو، حق کپی جستجو به اضافه ی برند حق کپی جستجو و حوزه به اضافه نام میزبان می باشد. در (۷)، نویسنده از URL برای شناسایی سایت های فیشینگ به صورت خودکار توسط استخراج و تصدیق عبارات متفاوت URL از طریق موتور جستجو استفاده کرد. حتی اگرچه این مقاله تکنیک جالب و جدیدی را پیشنهاد کرد، نرخ شناسایی نسبتا پایین است (۵۴,۳٪). تکنیک (۸) یک رویکرد مبتنی بر محتوا برای شناسایی فیشینگ به نام CATINA توسعه داد که ارزش Google PageRank از صفحه را در نظر می گیرد، مجموعه داده های ارزیابی نسبتا کوچک است. ویژگی کد منبع برای شناسایی سایت های فیشینگ در (۹) مورد استفاده قرار گرفته است.

نویسندگان در (۱۰) تکنیک فازی مبتنی بر ۲۷ ویژگی صفحه وب پیشنهاد کنند، که به ۳ لایه طبقه بندی می شود. هر ویژگی سه ارزش زبانی دارد: پایین، متوسط و بالا. تکنیک مجموعه قوانین، توابع عضویت مثلثی و ذوزنقه ای ایجاد کرده است. نرخ بدست آمده از تکنیک ۸۶,۲٪ است. اما نقص های بسیاری در(۱۰) وجود دارد. نخست، مجموعه قوانین عینی نیستند و بسیار بر سازنده بستگی دارند. دوم، وزن هر معیار اصلی بدون هیچ توضیحی مورد استفاده قرار گرفته است.در نهایت، اکتشافات مورد استفاده بهینه و موثر نیستند.

نویسندگان در (۱۱) تکنیک شبکه عصبی را پیشنهاد کردند. سه لایه در شبکه عصبی مورد استفاده بود که شامل لایه ورودی، لایه پنهان و لایه خروجی بود. بهترین نرخ بدست آمده از تکنیک ۹۵٪ است. با این حال، نقص هایی در (۱۱) وجود دارد. نخست، تعدادی از گره های پنهان و تابع فعال سازی باید از طریق آزمایش تعیین شوند. دوم، نویسندگان توضیح نمی دهند که چرا از یک لایه پنهان استفاده می کنند. سوم، ارزش ویژگی ها مشخص نمی سازد که چگونه محاسبه شده است. در نهایت، مجموعه داده برای بررسی دقت و درستی کافی نیست.

با توجه به تکنیک های قبلی، URL نقش کمی در شناسایی صفحات وب فیشینگ ایفا می کند. در این مقاله، مدل فازی عصبی بر مبنای ویژگی های URL و پارامترهای Google طراحی کردیم تا سایت های فیشینگ را شناسایی

کنیم. مدل بدون تعریف مجموعه قوانین if-then از (۱۲) با جنبه های جدید توسعه یافته است: i) اکتشافات جدید پیشنهاد شد تا صفحات وب را به صورت موثرتر و سریع تر شناسایی شود. ii) پارامترهای توابع عضویت حذف شده اند، بنابراین ارزش های فازی به صورت عینی تر محاسبه شدند. iii) ارزش های ورودی نرمال شده اند تا دقت و همگرایی فاز آموزش را افزایش دهد. بعلاوه، جنبه های موجود در (۱۲) نیز مدل جدید را به صورت زیر پشتیبانی می کند: i) وزن ها توسط شبکه عصبی آموزش می بینند، بنابراین مدل کارآمد تر است. ii) قوانین if-then مورد استفاده نیستند. بنابراین، نتیجه دقیق و عینی تر است.

III. طرح سیستم

A. URL

یک URL (مکان یکنواخت منبع) برای مکانیابی منابع استفاده می شود (۱۳).

ساختار URL به صورت زیر است:

$$< protocol >: // < subdomain > . < primarydomain > . < TLD > / < pathdomain >$$

برای مثال، با URL: http://www.paypal.abc.net/login/web/index.html ممکن است شش مولفه به شرح زیر وجود داشته باشد: پروتکل http است، Subdomain به صورت paypal است، primarydomain به صورت abc است، TLD شبکه است، Domain به صورت abc.net و Pathdomain به صورت login/web/index.html است.

B. ویژگی های URL

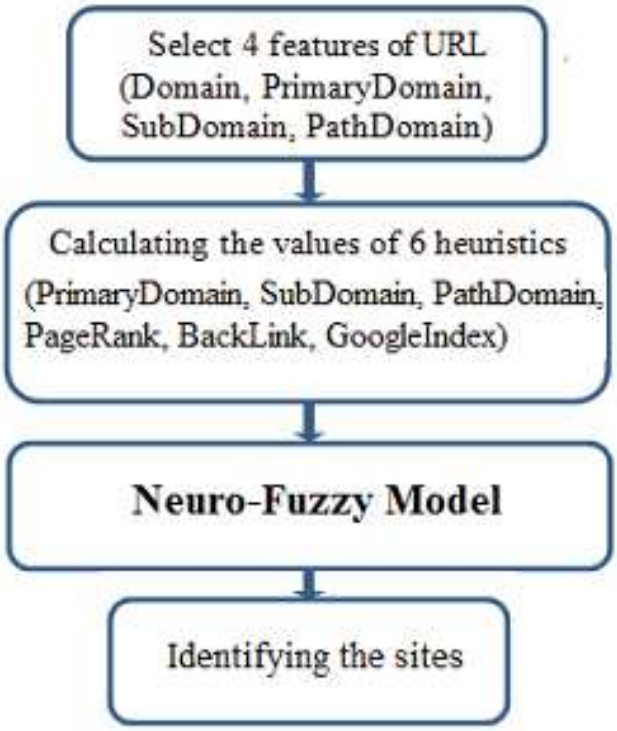
فیشرها معمولا سعی دارند که آدرس اینترنتی (URL سایت های فیشینگ را مشابه سایت های قانونی بسازند تا کاربران آنلاین را فریب دهند. آن ها نمی توانند از URL دقیق سایت قانونی استفاده کنند، آنها خطاهای املایی بیشتری از ویژگی های URL ایجاد می کنند مانند PrimaryDomain, SubDomain, PathDomain. برای مثال، URL www.Paypall.com مانند وبسایت معروف www.Paypal.com یا http://www.Paypal.attack.com می باشد. اگر کاربران دقیق نباشند، فکر می کنند که در سایت Paypal هستند.

C. ویژگی های پارامترهای گوگل

بدیهی است که صفحه ی فیشینگ صفحه ی جدیدی است و برای مدت زمان کوتاهی وجود دارد، بنابراین رتبه بندی صفحه ی فیشینگ بسیار پایین است و تعداد لینک ها از صفحات دیگر محدود می باشد. بنابراین، پارامترهای گوگل مانند PageRangk, BackLink و GoogleIndex می تواند برای شناسایی فیشینگ پشتیبانی شود.

D. طرح مدل سیستم

مدل می تواند در شکل ۱ نشان داده شود.



(ترجمه شکل:انتخاب ۴ ویژگی URL- محاسبه مقادیر ۶ اکتشاف-مدل عصبی فازی-شناسایی سایت ها)

شکل۱. مدل سیستم

۱) **فاز I**– انتخاب چها ویژگی URL: چهار ویژگی مانند دامنه، دامنه اصلی، زیردامنه و دامنه مسیر از URL استخراج شده است.

۲) **فاز II**– محاسبه ی شش مقدار اکتشافی: شش مقدار اکتشافی محاسبه می شوند و ۶ مورد اکتشافی ۶ گره ورودی شبکه عصبی فازی هستند.

۳) **فاز III**– شبکه عصبی فازی: شبکه عصبی فازی برای محاسبه ی مقدار گره خروجی اجرا می شود.

۴) **فاز IV**– شناسایی سایت ها: بر اساس مقدار گره خروجی تصمیم میگیریم که آیا یک سایت، سایت فیشینگ است.

E. مدل شبکه عصبی فازی

۱) مدل: مدل شبکه عصبی فازی به صورت شکل ۲ نشان داده شده است. مدل با پنج لایه به شرح زیر می باشد:

- لایه نخست، به نام لایه ورودی، شامل شش گره است که شش اکتشافی مانند دامنه اصلی، زیردامنه، دامنه مسیر، رتبه صفحه، لینک برگشت، شاخص گوگل(PageRank, BackLink, GoogleIndex) می باشد.
- لایه دوم شامل ۱۲ گره است. ارزش هر گره از تابع عضویت سیگموئید چپ و تابع عضویت سیگموئید راست محاسبه می شود.
- سومین لایه شامل دو گره است که π_L و π_P می باشد. π_L و π_P توسط (۱) و (۲) محاسبه می شوند.

$$\pi_L = \prod_{i=1}^{6} L_i \qquad (1)$$

$$\pi_P = \prod_{i=1}^{6} P_i \qquad (2)$$

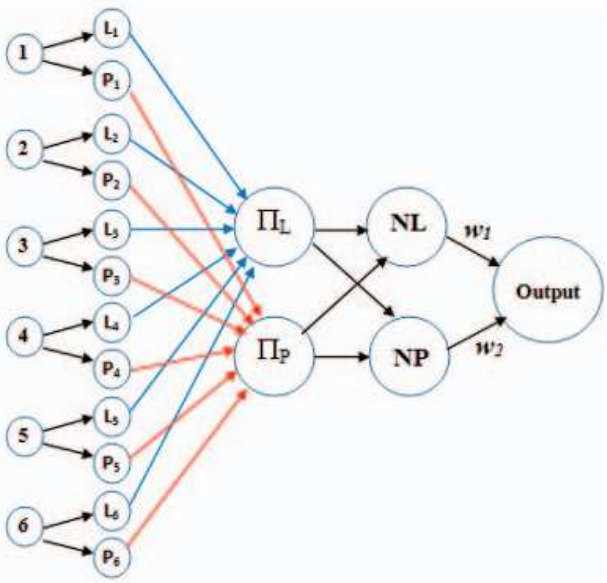
- چهارمین لایه شامل دو گره است که NL (قانونی نرمال شده)[۶] و NP (فیشینگ نرمال شده) می باشد. NL و NP توسط (۳) و (۴) محاسبه می شوند.

$$NL = \frac{\pi_L}{\pi_L + \pi_P} \qquad (3)$$

$$NP = \frac{\pi_P}{\pi_L + \pi_P} \qquad (4)$$

- پنجمین لایه به نام لایه خروجی، یک گره خروجی دارد.

[6] Normalization Legitimate



شکل ۲. مدل شبکه عصبی فازی

شبکه عصبی از چهارمین لایه به لایه خروجی عمل می کند. وزن ها توسط الگوریتم آموزش آموزش می بینند و تابع فعال سازی سیگموئید در مدل پیشنهادی مورد استفاده قرار می گیرد، بنابراین ارزش خروجی گره خروجی از ۰ تا ۱ است. مدل پیشنهادی به دو طبقه دسته بندی می شود تا اگر ارزش گره خروجی کمتر از ۰,۵ باشد سایت فیشینگ است و اگر ارزش بیشتر یا برابر با ۰,۵ باشد سایت قانونی است.

۲) ارزش شش گره ورودی: هر ارزش ورودی از ۰ تا ۱ متفاوت است. اگر ارزش ورودی نزدیک به صفر باشد، سایت مشکوک به سایت فیشینگ است. اگر ارزش ورودی نزدیک به ۱ باشد، سایت یک سایت قانونی است. شش ارزش گره ورودی به صورت زیر محاسبه می شود:

- محاسبه ی ارزش دامنه اصلی: ارزش دامنه اصلی بر اساس فاصله ی Levenshtein (۱۴) بین دامنه اصلی و نتیجه ی پیشنهادات موتور جستجوی گوگل می باشد. الگوریتم در الگوریتم ۱ نشان داده شده است.
- محاسبه ی ارزش زیردامنه و دامنه مسیر: مانند دامنه اصلی، الگوریتم برای محاسبه ی ارزش زیردامنه و دامنه مسیر در الگوریتم ۲ نشان داده شده است.

- محاسبه ی ارزش PageRank، BackLink، GoogleIndex: ارزش ها از (۱۵) بدست می آیند. الگوریتم ها برای محاسبه ی ارزش ها در الگوریتم ۳، ۴ و ۵ نشان داده شده اند.

۳) تبدیل ارزش شش گره ورودی: در دومین لایه، تابع عضویت سیگموئید برای محاسبه مقادیر فازی مورد استفاده هستند. متغیر X از توابع عضویت از 10- تا ۱۰ است. بنابراین، ارزش گره ورودی باید توسط معادله (۵) تبدیل شود.

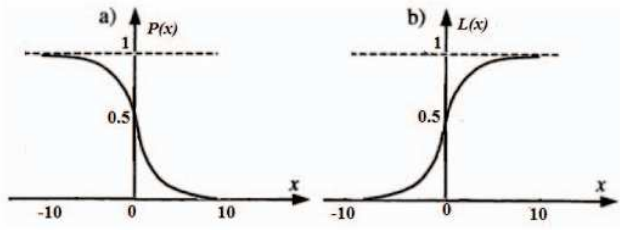
$$Value_{new} = Value_{old} * (Max - Min) + Min \qquad (5)$$

که ماکسیمم ۱۰ و مینیمم 10- است. $Value_{old}$ از ۰ تا ۱ است. $Value_{new}$از 10- تا ۱۰ است.

۴) ارزش ۱۲ گره در لایه دوم: هر یک از این اکتشافی ها به متغیرهای زبانی مانند فیشینگ و قانونی طبقه بندی شده اند. معادله (۶) و (۷) دو تابع عضویت هستند که ایجاد شده اند تا ارزش های فازی را محاسبه کنند و نمودار توابع عضویت در شکل ۳ نشان داده شده است.

$$L(x) = \frac{1}{1 + e^{-x}} \qquad (6)$$

$$P(x) = \frac{e^{-x}}{1 + e^{-x}} \qquad (7)$$



شکل ۳. نمودار توابع عضویت

داده: دامنه اصلی

نتیجه: ارزش اکتشافی دامنه اصلی

```
if PrimaryDomain is IP then
    value = 0; //doubt phishing
else
    Result = Suggestion_Google(PrimaryDomain);
    if Result is NULL then
        value = 1; //No doubt phishing
    else
        if Levenshtein(Result, PrimaryDomain)=0 then
            value = 1; //No doubt phishing
        else
            value = 1 −
            (1/Levenshtein(Result, PrimaryDomain));
        end
    end
end
```

الگوریتم ۱: محاسبه ی ارزش دامنه اصلی

```
Data: m //m is SubDomain or PathDomain
Result: The value of heuristic m
if m is NULL then
    value = 1; //No doubt phishing
else
    Result = Suggestion_Google(m);
    if Result is NULL then
        value = 1; //No doubt phishing
    else
        if Levenshtein(Result, PrimaryDomain)=0 then
            value = 1; //No doubt phishing
        else
            value = 1 − (1/Levenshtein(Result, m));
        end
    end
end
```

الگوریتم ۲: محاسبه ی ارزش زیردامنه/دامنه مسیر

```
Data: URL
Result: The value of heuristic PageRank
value = Google_PageRank(URL);
if value <= 0 then
    value = 0; //doubt phishing
else
    value = 1 − (1/value);
end
```

الگوریتم ۳: محاسبه ی ارزشPageRank

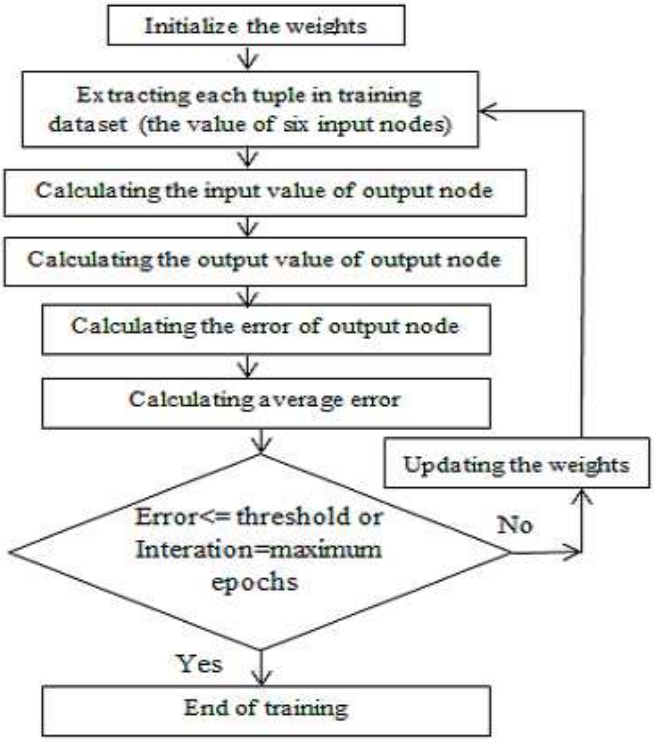
```
Data: URL
Result: The value of heuristic BackLink
value = Google_BackLink(URL);
if value <= 0 then
    value = 0; //doubt phishing
else
    value = 1 − (1/value);
end
```

الگوریتم۴: محاسبه ی ارزش Backlink

Data: URL
Result: The value of heuristic GoogleIndex
$value = Google\_Index(URL)$;
if $value <= 0$ **then**
| $value = 0$; //doubt phishing
else
| $value = 1 - (1/value)$;
end

الگوریتم ۵: محاسبه ی ارزش GoogleIndex.

۵) الگوریتم آموزش شبکه عصبی: الگوریتم پیشنهادی در شکل ۴ نشان داده شده است. الگوریتم دو فاز به صورت زیر اجرا می کند:



شکل ۴. الگوریتم آموزش شبکه عصبی

- فاز انتشار ارزش ورودی گره خروجی و ارزش خروجی گره خروجی را محاسبه می کند. ارزش ورودی گره خروجی توسط (۸) محاسبه می شود

$$O_I = \sum_i W_i * I_i \qquad (8)$$

که O_I, I_i and W_i به ترتیب ارزش ورودی گره خروجی، ارزش گره ورودی iام، و وزن گره ورودی iام می باشد.

ارزش خروجی گره خروجی توسط (۹) محاسبه می شود

$$O_O = \frac{1}{1 + e^{-O_I}} \tag{9}$$

که O_O و O_I ارزش خروجی گره خروجی و ارزش ورودی گره خروجی هستند.

- فاز به روز رسانی وزن خطای گره خروجی را محاسبه می کند و وزن ها را به روز رسانی می کند. خطای گره خروجی توسط (۱۰) محاسبه می شود

$$Err = O_O * (1 - O_O) * (T - O_O) \tag{10}$$

که T ارزش واقعی نمونه در مجموعه داده های آموزش است. وزن ها توسط (۱۱) به روز رسانی می شوند

$$W_i = W_i + R * Err * O_O \tag{11}$$

که R و W_i به ترتیب نرخ یادگیری و وزن iامین گره ورودی می باشند.

IV. ارزیابی

۱۱۶۶۰ سایت فیشینگ از PhishTank (۱) و ۱۰۰۰۰ سایت قانونی از DMOZ جمع آوری کردیم (۱۶). مجموعه داده های آموزش شامل ۶۶۶۰ سایت فیشینگ از PhishTank و ۵۰۰۰ سایت قانونی از DMOZ بود. ۲ مجموعه داده آزمون ایجاد کردیم، هریک از آن ها شامل ۵۰۰۰ سایت فیشینگ یا ۵۰۰۰ سایت قانونی بود. روند تجربی از طریق PHP و MYSQL به ۲ فاز (آموزش و آزمون) تقسیم شد.

A. فاز آموزش

۱) وارد کردن مجموعه داده آموزش: مجموعه داده آموزش به MYSQL وارد شد. نتیجه در شکل ۵ نشان داده شده است.

phish_id	url	phish_detail_url	submission_time	verified	verification_time
2110838	http://www.paypal.com.uk.webapp.filippotteau.be/.d...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 00:41:35	yes	2013-11-17 13:04:59
2110837	http://klapkasuli.hu/wp-admin/includes/remax	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 00:41:25	yes	2013-11-17 07:14:39
2110836	http://www.umbrellacreative.com.au/wp-content/plug...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 00:41:16	yes	2013-11-17 13:50:14
2110835	http://www.livingabroadmagazine.co.uk/googledocss/...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 00:41:10	yes	2013-11-17 13:04:59
2110831	http://www.paypal.com.login.account.eecp.org/5cc46...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 00:40:26	yes	2013-11-17 13:50:14

شکل ۵. مجموعه داده آموزش از ۱۱۶۶۰ سایت در MYSQL

۲) استخراج چهار ویژگی از URL: چهار ویژگی (دامنه اصلی، زیردامنه، دامنه مسیر و دامنه) استخراج شدند. شکل ۶ نتیجه بدست آمده را نشان می دهد.

phish_id	domain	primarydomain	subdomain	pathname
2110838	filippotteau.be	filippotteau	paypal,com,uk,webapp	.dd
2110837	klapkasuli.hu	klapkasuli		wp,admin,includes,remax
2110836	umbrellacreative.com.au	umbrellacreative		wp,content,plugins,2013gdocs,index.htm
2110835	livingabroadmagazine.co.uk	livingabroadmagazine		googledocss,googledocss,sss
2110831	eecp.org	eecp	paypal,com,login,account	webscr,cmd%3D,account.php

شکل ۶. چهار ویژگی استخراج شده اند

۳) محاسبه ی ارزش شش گره ورودی: پیشنهاد املایی موتور جستجوی گوگل و Google API برای محاسبه ی ارزش گره های ورودی مورد استفاده قرار گرفتند. سپس، ارزش ها از 10- به ۱۰ تبدیل می شود. نتیجه در شکل ۷ نشان داده شده است.

phish_id	PrimaryDomain	SubDomain	PathDomain	PageRank	BackLink	GoogleIndex
2110838	5	10	5	-10	-10	-10
2110837	5	5	10	-10	-10	-10
2110836	7.1	5	5	-10	-10	-10
2110835	5	5	10	-10	-10	-10
2110831	10	10	10	-10	-10	-10

شکل ۷.ارزش گره های ورودی از 10- تا ۱۰

۴) محاسبه ی ارزش فازی از ۱۲ گره در لایه دوم: دو تابع عضویت سیگموئید چپ و راست برای محاسبه ی ارزش گره ها در لایه دوم مورد استفاده قرار گرفتند. نتیجه در شکل ۸ نشان داده شده است.

phish_id	P1	P2	P3	P4	P5	P6	L1	L2	L3	L4	L5	L6
2110838	0.00699	0.00005	0.00699	0.99995	0.99995	0.99995	0.99331	0.99995	0.99331	0.00005	0.00005	0.00005
2110837	0.00669	0.00669	0.00005	0.99995	0.99995	0.99995	0.99331	0.99331	0.99995	0.00005	0.00005	0.00005
2110836	0.00079	0.00669	0.00669	0.99995	0.99995	0.99995	0.99921	0.99331	0.99331	0.00005	0.00005	0.00005
2110835	0.00669	0.00669	0.00005	0.99995	0.99995	0.99995	0.99331	0.99331	0.99995	0.00005	0.00005	0.00005
2110831	0.00005	0.00005	0.00005	0.99995	0.99995	0.99995	0.99995	0.99995	0.99995	0.00005	0.00005	0.00005

شکل ۸. ارزش های فازی در لایه دوم

۵) فاز آموزش شبکه: آموزش شبکه با ۹ نرخ ارزش یادگیری را اجرا کردیم. در فاز آموزش، پارامترها به شرح زیر می باشد:

- نرخ یادگیری: $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \text{ and } 0.9$
- تعداد Epochs: ۱۰۰۰۰
- وزن ها: ارزش های تصادفی وزن اولیه از ۰ تا ۱

B. فاز آزمون

در این فاز، تکنیک پیشنهادی با ۲ مجموعه داده آزمون بر اساس وزن های آموزش شبکه با نرخ یادگیری $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \text{ and } 0.9$ آزمون می شود. RMSE (خطای جذر میانگین مربعات) مقیاس خوبی برای شناسایی دقت است.RMSE توسط معادله (۱۲) محاسبه می شود

$$RMSE = \sqrt{\frac{\sum (A_i - I_i)^2}{N}} \qquad (12)$$

که I_i تعداد سایت های شناسایی کننده، A_i تعداد سایت های واقعی و N تعداد نمونه ها در مجموعه داده های آزمون است. نسبت درستی به صورت زیر محاسبه می شود: درستی-نسبت=۱۰۰-RMSE. نتایج آزمون با نرخ یادگیری $0.1, 0.2, 0.3, 0.4, 0.5, 0.5, 0.6, 0.7, 0.8, 0.9$ در جدول I نشان داده خواهد شد. از نتایج بدست آمده، RMSE و درستی در جدول II نشان داده می شوند. دریافتیم که بهترین نسبت ۹۹,۳۱٪ را با نرخ یادگیری ۰,۷ و بدیترین نسبت ۹۸,۳۲٪ با نرخ یادگیری ۰,۲ را نشان می دهد.

جدول I - نتایج آزمون با تکنیک پیشنهادی

Learning Rate	Testing dataset	A_i	I_i
0.1	No.1	5,000	4,928
0.1	No.2	5,000	4,921
0.2	No.1	5,000	4,914
0.2	No.2	5,000	4,918
0.3	No.1	5,000	4,932
0.3	No.2	5,000	4,935
0.4	No.1	5,000	4,949
0.4	No.2	5,000	4,939
0.5	No.1	5,000	4,938
0.5	No.2	5,000	4,930
0.6	No.1	5,000	4,935
0.6	No.2	5,000	4,935
0.7	No.1	5,000	4,969
0.7	No.2	5,000	4,962
0.8	No.1	5,000	4,917
0.8	No.2	5,000	4,917
0.9	No.1	5,000	4,922
0.9	No.2	5,000	4,917

جدول II - RMSE و درستی با تکنیک پیشنهادی

Learning rate	RMSE	Accuracy
0.1	1.51	98.49%
0.2	1.68	98.32%
0.3	1.33	98.67%
0.4	1.12	98.88%
0.5	1.32	98.68%
0.6	1.30	98.70%
0.7	0.69	99.31%
0.8	1.66	98.34%
0.9	1.61	98.39%

C. مقایسه با تکنیک (۱۰)

با تکنیک (۱۰) آزمایش کردیم و با نتیجه ی تکنیک پیشنهادی مان مقایسه کردیم. نخست، ۱۰ مجموعه داده آزمون را جمع آوری کردیم که هریک شامل ۱۰۰۰ سایت فیشینگ یا ۱۰۰۰ سایت قانونی است. دوم، تکنیک (۱۰) ا آزمایش کردیم و نتایج در جدول III نشان داده شده اند.از نتیجه بدست آمده و استفاده از RMSE، دریافتیم که تکنیک (۱۰) درستی ۸۶,۰۶٪ دارد.

جدول III - نتایج آزمون با تکنیک (۱۰)

Testing dataset	(1)	(2)	(3)
No.1	867 (86.7%)	82 (8.2%)	51 (5.1%)
No.2	865 (86.5%)	76 (7.6%)	59 (5.9%)
No.3	847 (84.7%)	90 (9.0%)	63 (6.3%)
No.4	902 (90.2%)	172 (17.2%)	26 (2.6%)
No.5	841 (84.1%)	109 (10.9%)	50 (5.0%)
No.6	64 (6.4%)	873 (87.3%)	63 (6.3%)
No.7	50 (5.0%)	911 (91.1%)	39 (3.9%)
No.8	39 (3.9%)	895 (89.5%)	66 (6.6%)
No.9	97 (9.7%)	871 (87.1%)	32 (3.2%)
No.10	85 (8.5%)	863 (86.3%)	52 (5.2%)

D. مقایسه با تکنیک (۱۱)

با تکنیک (۱۱) با استفاده از ۸ گره پنهان و تابع فعال سازی مماس هذلولی آزمایش کردیم. نخست، ۲ مجموعه داده آزمون جمع آوری کردیم که هر یک شامل ۵۰۰۰ سایت فیشینگ یا ۵۰۰۰ سایت قانونی بود. دوم، تکنیک (۱۱) را آزمایش کردیم و نتایج در جدول IV نشان داده خواهند شد. سپس، نتایج بدست آمده از RMSE و درستی در جدول V نشان داده شده است. با استفاده از تکنیک در (۱۱)، بهترین درستی ۹۴,۶۸٪ را بدست آوردیم.

جدول IV - نتایج آزمون با تکنیک(۱۱)

Learning Rate	Testing dataset	A_i	I_i
0.1	No.1	5,000	4,612
0.1	No.2	5,000	4,520
0.2	No.1	5,000	4,624
0.2	No.2	5,000	4,478
0.3	No.1	5,000	4,689
0.3	No.2	5,000	4,735
0.4	No.1	5,000	4,456
0.4	No.2	5,000	4,792
0.5	No.1	5,000	4,732
0.5	No.2	5,000	4,736
0.6	No.1	5,000	4,721
0.6	No.2	5,000	4,678
0.7	No.1	5,000	4,599
0.7	No.2	5,000	4,725
0.8	No.1	5,000	4,772
0.8	No.2	5,000	4,697
0.9	No.1	5,000	4,719
0.9	No.2	5,000	4,699

جدول V - RMSE و درستی با تکنیک (۱۱)

Learing rate	RMSE	Accuracy
0.1	8.73	91.27%
0.2	9.10	90.90%
0.3	5.78	94.22%
0.4	8.24	91.76%
0.5	5.32	94.68%
0.6	6.03	93.97%
0.7	6.88	93.12%
0.8	5.36	94.64%
0.9	5.82	94.18%

V. نتیجه گیری

تکنیک جدیدی برای شناسایی موثر سایت های فیشینگ پیشنهاد کرده ایم. در این تکنیک، مدل سیستم برای شناسایی سایت های فیشینگ با استفاده از شبکه عصبی فازی با پنج لایه و شش اکتشافی (PrimaryDomain, SubDomain, Path- Domain, PageRank, BackLink, GoogleIndex) است. تکنیک با مجموعه داده های آموزش حاوی ۱۱۶۶۰ سایت و ۲ مجموعه داده آزمون آزمایش شد که هر مجموعه داده شامل ۵۰۰۰ سایت فیشینگ یا ۵۰۰۰ سایت قانونی است. بهترین نتایج درستی می تواند ۹۹,۳۱٪ بدست آید. مقایسه ای در مورد شناسایی درست با (۱۰) و (۱۱) انجام دادیم، کار ما نشان داد که موثر تر و دقیق تر است. در آینده، مدل عصبی قازی بهبود خواهد یاتف تا نسبت شناسایی را افزایش دهد. بعلاوه، سیستم می تواند با استفاده از مجموعه داده های بزرگ تر و پارامترهای اکتشافی بیشتر افزایش یابد (بهبود یابد).

REFERENCES

[1] PhishTank. (2013, Nov.) [Online]. Available: http://www.phishtank.com
[2] D. Goodin. (2012) Google bots detect 9,500 new malicious websites every day. [Online]. Available: http://arstechnica.com/security/2012/06/
[3] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. (2009) An empirical analysis of phishing blacklists. [Online]. Available: http://ceas.cc/2009/papers/ceas2009-paper-32.pdf
[4] McAfee. (2011, July) Mcafee site advisor. [Online]. Available: http://www.siteadvisor.com
[5] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in The 16th international conference on World Wide Web, 2007, pp. 639—648

[6] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: a feature-rich machine learning framework for detecting phishing web sites, "ACM Transactions on Information and System Security", vol.14, no.2 .pp. 1—28, Sept. 2011.
[7] M. E. Maurer and D. Herzner, "Using visual website similarity for phishing detection and reporting," in CHI '12 Extended Abstracts on Human Factors in Computing Systems, 2012, pp. 1625—1630.
[8] A. Sunil and A. Sardana, "A pagerank based detection technique for phishing web sites," in IEEE Symposium on Computers & Informatics, 2012, pp. 58—63.
[9] M. G. Alkhozae and O. A. Batarfi, "Phishing websites detected based on phishing characteristic in the webpage source code," in International Journal of Information and Communication Technology Research, vol. 1, no. 6, Oct. 2011, pp. 283—291
[10] M. Aburrous, M. Hossain, F. Thabatah, and K. Dahal, "Intelligent phishing website detection system using fuzzy techniques," in Third International Conference on Information and Communication Technologies: From Theory to Applications, 2008, pp. 1—6.
[11] N. Zhang and Y. Yuan, "Phishing Detection Using Neural Network", CS229 lecture notes, http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf, 2012
[12] L. A. T. Nguyen , B. L. To , H. K. Nguyen , C. Pham , C. S. Hong, "A novel neuro-fuzzy approach for phishing identification", 2014 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 188—193, Dec. 2014.
[13] Wikipedia. [Online]. Available (2014) : http://en.wikipedia.org/wiki/Uniformresourcelocator
[14] Levenshtein. [Online]. Available (2014) : http://en.wikipedia.org/wiki/Levenshteindistance
[15] G. Inc. [Online]. Available (2014) : http://toolbarqueries.google.com
[16] DMOZ. [Online]. Available (2014) : http://rdf.dmoz.org/rdf/