یک رویکرد عصبی فازی برای شناسایی فیشینگ

چکیده

همراه با رشد اینترنت، معاملات تجارت الکترونیک نقش مهمی در جامعه مدرن ایفا می کند. در نتیجه، فیشینگ اقدامی عمدی توسط فرد یا گروهی از افراد برای سرقت اطلاعات شخصی مانند کلمه عبور، اطلاعات حساب بانکی، کارت اعتباری و غیره می باشد. اکثر این صفحات وب فیشینگ از لحاظ رابط و آدرس وب یا مکان یکنواخت منبع (URL) همانند صفحات اصلی هستند. تکنیک های بسیاری برای شناسایی صفحات وب پیشنهاد شده است، مانند تکنیک های مبتنی بر فهرست سیاه (Blacklist)، تکنیک های مبتنی بر اکتشاف، و غیره. با این وجود، به دلیل تکنیک های حفاظت ناکارآمد، تعداد قربانیان در حال افزایش می باشد. شبکه های عصبی و سیستم های فازی می توانند ترکیب شوند تا مزایای مشترکی داشته باشند و مشکلات مجزایشان را برطرف کنند. این مقاله مدل عصبی فازی جدید بدون استفاده از مجموعه قوانین برای شناسایی فیشینگ را معرفی می کند. بویژه، تکنیک پیشنهادی ارزش اکتشافات از توابع عضویت را محاسبه می کند. سپس، وزن ها توسط شبکه عصبی آموزش داده می شوند. تکنیک پیشنهادی بدون مجموعه داده از 11660 سایت فیشینگ و 10000 سایت قانونی ارزیابی می شود. نتایج نشان می دهد که تکنیک پیشنهادی می تواند بیش از 99٪ از سایت های فیشینگ را شناسایی کند.

۱. مقدمه

" فیشینگ" از کلمه "ماهی گیری" (fishing) ایجاد می شود. فیشرها، که سایت های فیشینگ ایجاد می کنند از تکنیک هایی برای فریب دادن قربانیان استفاده می کنند که شامل پیام های ایمیل، پیام های فوری، پست های انجمن، تماس های تلفنی و شبکه اجتماعی می شود. این فعالیت های فیشینگ منجر به خسارت اقتصادی شدید در سراسر جهان می گردد. برطبق مطالعه ای توسط Gartner [1]، 57 میلیون کاربر اینترنت آمریکا رسید ایمیل مرتبط به کلاهبرداری های فیشیگ را شناسایی کرده اند و 2 میلیون نفر از بین آنان برای دادن اطلاعات حساس شان فریب خوردند. در همین حال، سایت های فیشینگ به سرعت از لحاظ کیفیت و تعداد در حال افزایش هستند. بنابراین، خطر سرقت اطلاعات کاربر بسیار بالا است. به این دلایل، شناسایی مشکلات فیشینگ در جامعه مدرن

بسیار فوری، پیچیده و مهم است. اخیرا، مطالعات بسیاری وجود داشته است که برخلاف فیشینگ بر اساس ویژگی های سایت مانند URL وب سایت، محتوای وب سایت، URL وب سایت و محتوا، کد منبع وب سایت یا رابط وب سایت را ادغام می کند. با این حال، هر یک از این مطالعات نقاط ضعف و قدرت مختص خود را دارند. هنوز روش کافی ای وجود ندارد. در این مقاله، رویکرد جدیدی پیشنهاد شده است تا سایت های فیشینگ را شناسایی کند که بر ویژگی های URL (دامنه اصلی[1]، زیردامنه[2]، دامنه مسیر[3]) و رتبه بندی سایت(PageRank, AlexaRank, AlexaReputation) متمرکز هستند. پس، یک شبکه عصبی فازی سیستمی است که خطا را کاهش و عملکرد را افزایش می دهد. مدل عصبی فازی از مدل های محاسباتی استفاده می کند تا بدون مجموعه قوانین اجرا شود. راه حل پیشنهادی برای به درستی شناسایی بالای 99٪ با سیگنال کاذب پاین دست یافت.

باقی مقاله به شرح زیر می باشد: بخش II کارهای مرتبط را ارائه می کند. طرح سیستم در بخش III نشان داده شده است. بخش IV درستی مدل 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 مرتبط

تکنیک های شناسایی فیشینگ به سه دسته مانند فهرست سیاه، اکتشافی و یادگیری ماشین تقسیم می شوند. در رویکرد نخست، تکنیک شناسایی فیشینگ [5][4][3][2] فهرستی از وب سایت فیشینگ به نام بلک لیست (صفحه سیاه) بدست می آورد. تکنیک بلک لیست به دلیل رشد سریع تعداد سایت های فیشینگ تکنیکی ناکارآمد است. بنابراین، رویکردهای اکتشافی و یادگیری ماشین توجه محققان بیشتری را به سمت خود جلب کرده است.

Cantina [6] الگوریتم TF-IDF را بر مبنای 27 ویژگی صفحات وب ارائه کرده است.این تکنیک می تواند 97٪ از سایت های فیشینگ با 6٪ مثبت کاذب را شناسایی کند. اگرچه این تکنیک کارآمد است، اما زمان استخراج 27 ویژگی صفحه وب بسیار طولانی است تا تقاضای زمان واقعی را برآورده کند و برخی از ویژگی های برای بهبود دقت شناسایی فیشینگ ضروری نیستند. . به همین نحو،+Cantina [7] از تکنیک های یادگیری ماشین بر اساس 15

[1] PrimaryDomain
[2] SubDomain
[3] PathDomain

ویژگی صفحه وب استفاده کرد و تنها شش مورد از 15 ویژگی برای شناسایی فیشینگ مانند فرم بد، فیلد فعالیت بد، 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 [8]، 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 (54.3٪). تکنیک [9] 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 [10] مورد استفاده قرار گرفته است.

نویسندگان در [11] تکنیک فازی مبتنی بر 27 ویژگی صفحه وب پیشنهاد کرده اند، که به 3 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 86.2٪ است. اما نقص های بسیاری در [11] 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 [12] 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 95٪ است. با این حال، نقص هایی در [12] 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 ها به اندازه کافی بزرگ نیستند.

در تکنیک های پیشین، URL نقش کمی در شناسایی صفحات وب فیشینگ ایفا می کند. در این مقاله، بر ویژگی های URL متمرکز هستیم و مدل عصبی فازی جدید برای شناسایی سایت های فیشینگ طراحی می کنیم. کار ما چهار جنبه ی جدید را توسعه می دهد: i) اکتشافات جدید پیشنهاد شه است تا صفحات وب فیشینگ را به صورت

موثرتر و سریع تر شناسایی شود. ii) ارزش های پارامترهای مورد استفاده در توابع عضویت از مجموعه داده های بزرگ نشات می گیرد، بنابراین مدل همچنان برابر با مجموعه داده های جدید است.iii) وزن ها توسط شبکه عصبی آموزش داده شده اند بنابراین کارآمد تر بودند.iv) از مجموعه قوانین استفاده نمی شود.بنابراین نتیجه دقیق و عینی تر خواهد بود.

III. طرح سیستم

A. URL

یک URL (مکان یکنواخت منبع) برای مکانیابی منابع استفاده می شود [13].

ساختار URL به صورت زیر می باشد:

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

برای مثال، با URL: http://www.paypal.abc.net/login/web/index.html، شش مولفه به صورت زیر وجود دارد: پرتکل http است، زیردامنه paypal می باشد، دامنه اصلی abcاست، TLD شبکه است، دامنه abc.net است، دامنه مسیر login/web/index.html می باشد.

B. ویژگی های URL

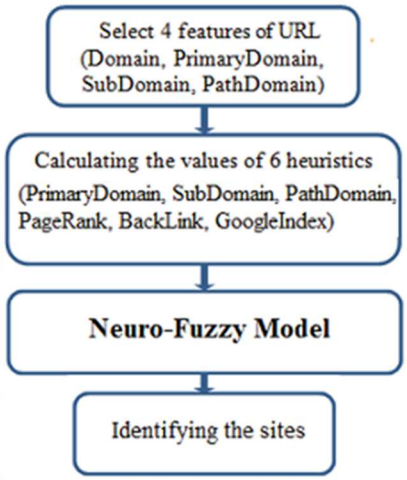
فیشرها معمولا سعی دارند که آدرس اینترنتی (URL) سایت های فیشینگ را مشابه سایت های قانونی بسازند تا کاربران آنلاین را فریب دهند. آن ها نمی توانند از URL دقیق سایت قانونی استفاده کنند، آنها خطاهای املایی بیشتری از ویژگی های URL ایجاد می کنند مانند PrimaryDomain, SubDomain, PathDomain. برای مثال، URL www.applle.com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

C. ویژگی های رتبه بندی دامنه ها

بدیهی است که سایت های فیشینگ نه در دسترس کاربران قرار دارد نه توسط صفحات دیگر مرتبط است. بنابراین، رتبه بندی سایت مانند PageRangk, AlexaRank, AlexaReputaion می تواند به شناسایی سایت فیشینگ کمک کند. فیشرها معمولا سایت های جعلی ای از سایت معروف ایجاد می کنند اما رتبه بندی سایت جعلی بالا نیست. همچنین می توانیم از رتبه بندی ها برای این استفاده کنیم که شناسایی کنیم آیا یک سایت از نوع فیشینگ است یا خیر.

D. طرح مدل سیستم

مدل می تواند در شکل 1 نشان داده شود.



(ترجمه شکل:انتخاب 4 ویژگی URL- محاسبه مقادیر 6 اکتشاف-مدل عصبی فازی-شناسایی سایت ها)

شکل1. مدل سیستم

1) **فاز I**– انتخاب چها ویژگی URL: چهار ویژگی از URL استخراج شده است مانند دامنه، دامنه اصلی، زیردامنه و دامنه مسیر.
2) **فاز II**– محاسبه ی شش مقدار اکتشافی: شش مقدار اکتشافی محاسبه می شوند و 6 مورد اکتشافی 6 گره ورودی شبکه عصبی فازی هستند.
3) **فاز III**– شبکه عصبی فازی: شبکه عصبی فازی برای محاسبه ی مقدار گره خروجی اجرا می شود.
4) **فاز IV**- شناسایی سایت ها: بر اساس مقدار گره خروجی تصمیم میگیریم که آیا یک سایت، سایت فیشینگ است.

E. مدل شبکه عصبی فازی

1) مدل: مدل شبکه عصبی فازی به صورت شکل 2 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، PageRank، AlexaRank، AlexaReputation می باشد.
- لایه دوم شامل 12 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 (1) و (2) 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)[4] و NP (فیشینگ نرمال شده) می باشد. NL و NP توسط (3) و (4) محاسبه می شوند.

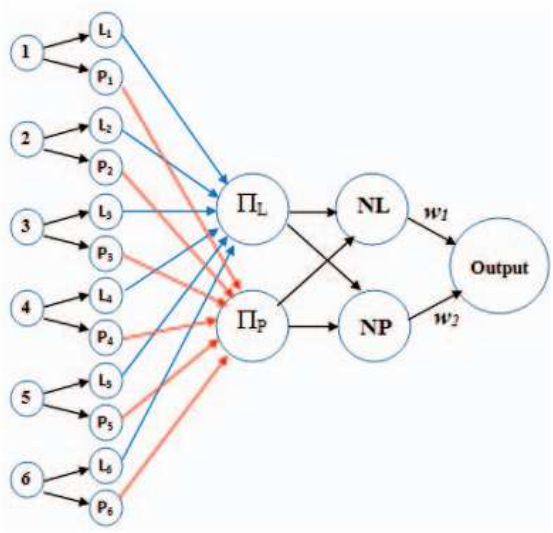
$$NL = \frac{\pi_L}{\pi_L + \pi_P} \qquad (3)$$

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

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

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 0 تا 1 است. مدل پیشنهادی به دو طبقه دسته بندی می شود تا اگر ارزش گره خروجی کمتر از 0.5 باشد سایت فیشینگ است و اگر ارزش بیشتر یا برابر با 0.5 باشد سایت قانونی است.

[4] Normalization Legitimate



شکل 2. مدل شبکه عصبی فازی

2) ارزش شش گره ورودی: بر اساس نتایج تجربی و آمار از مجموعه داده های 11660 سایت فیشنگ می باشد. در یافتیم که:

- سایت فیشینگ فاصله Levenshtein [14] بین دامنه اصلی، زیردامنه، دامنه مسیر و نتیجه ی پیشنهاد املایی موتور جستجوی گوگل دارد که کمتر از 4 است.
- ارزش PageRank از 0 تا 10 است. سایت فیشینک ارزش PageRank کمتر از 6 دارد.
- سایت فیشینگ ارزش AlexaRank دارد که بیشتر از 300000 است.
- سایت فیشینگ ارزش AlexaReputation دارد که کمتر از 20 است.

شش ارزش اکتشافی به صورت زیر محاسبه می شود:

- محاسبه ی ارزش اکتشافی دامنه اصلی:. الگوریتم در الگوریتم 1 نشان داده شده است.

Data: PrimaryDomain
Result: The value of heuristic "PrimaryDomain"
if *PrimaryDomain is IP* **then**
  $value = 0$; //doubt phishing
else
  $Result = Suggestion\_Google(PrimaryDomain)$;
  if *Result is NULL* **then**
    $value = 100$; //No doubt phishing
  else
    $value =$
    $Levenshtein(Result, PrimaryDomain)$;
  end
end

الگوریتم 1. محاسبه ارزش دامنه اصلی

- محاسبه ی ارزش اکتشافی زیردامنه و دامنه مسیر: الگوریتم مربوطه در الگوریتم 2 نشان داده شده است.

Data: m //m is SubDomain or PathDomain
Result: The value of heuristic m
if *m is Null* **then**
  $value = 100$; //No doubt phishing
else
  $Result = Suggestion\_Google(m)$;
  if *Result is NULL* **then**
    $value = 100$; //No doubt phishing
  else
    $value = Levenshtein(Result, m)$;
  end
end

الگوریتم 2: محاسبه ارزش زیردامنه/دامنه مسیر

- محاسبه ی ارزش اکتشافی PageRank:ارزش Google PageRank می تواند از [15] بدست آید. ارزش PageRank از 0 تا 10 است.
- محاسبه ارزش اکتشافی AlexaRank و AlexaReputation: ارزش این دو می تواند از [16] بدست آید.

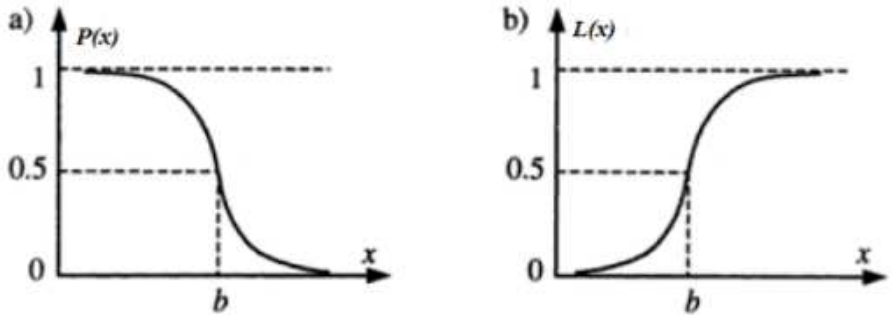
3) ارزش 12 گره در لایه دوم: طبقه بندی اکتشافی در دو عنوان زبانی و تخصیص توابع عضویت مانند سیگموئید چپ و راست برای هر ارزش زبانی. هر یک از این اکتشاف ها به عناوین زبانی مانند فیشینگ و قانونی طبقه بندی می شوند. بر اساس نتایج و آمار تجربی از مجموعه داده های 11660 سایت فیشینگ، توابع عضویت به صورت زیر محاسبه می شوند:

- توابع عضویت برای دامنه اصلی، زیردامنه، دامنه مسیر، PageRank و AlexaReputation: معادله های 5 و 6 دو تابع عضویت هستند که برای محاسبه ی ارزش های فازی ایجاد شده اند و نمودار توابع عضویت در شکل 3 نشان داده شده است.

$$L(x) = \frac{1}{1 + e^{-(x-b)}} \tag{5}$$

$$P(x) = \frac{e^{-(x-b)}}{1 + e^{-(x-b)}} \tag{6}$$

که پارامتر b برای دامنه اصلی، زیردامنه، دامنه مسیر، Pagerank و AlexaREputation به ترتیب برابر با 4، 4، 4، 6 و 20 است.

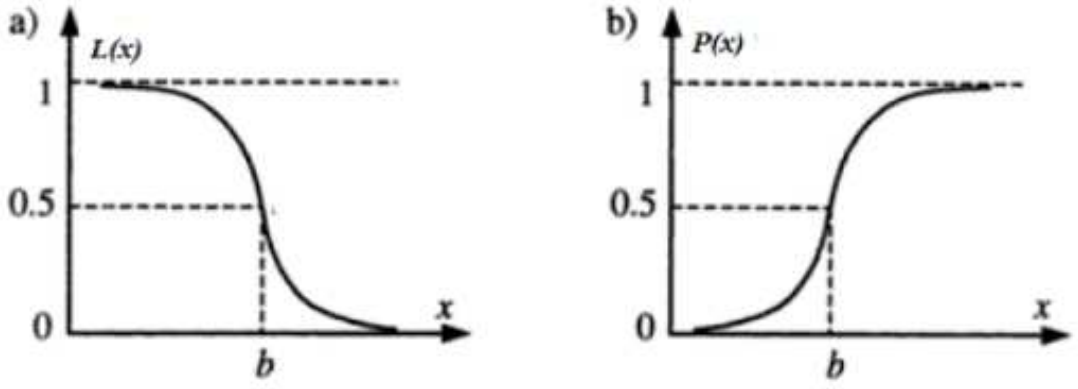


شکل 3. نمودار تابع عضویت

- توابع عضویت برای AlexaRank: معادله (7) و (8) دو توابع عضویت هستند که برای محاسبه ی ارزش های فازی با پارامتر b برابر با 300000 ایجاد شده است و نمودار توابع عضویت در شکل 4 نشان داده شده است.

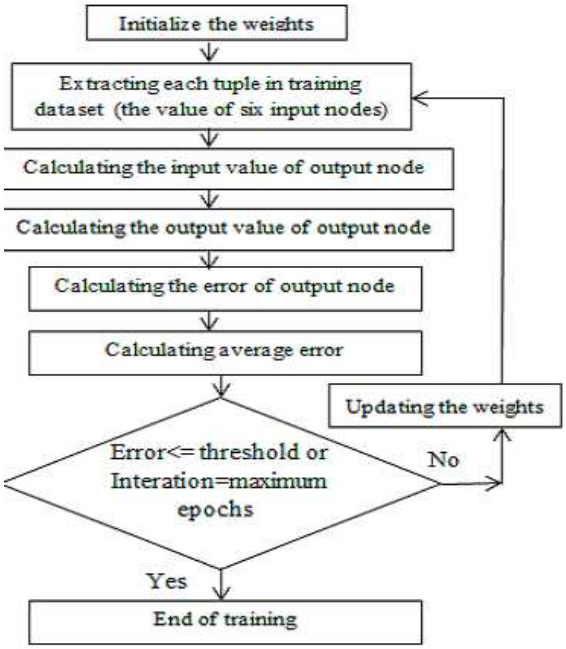
$$P(x) = \frac{1}{1 + e^{-(x-b)}} \qquad (7)$$

$$L(x) = \frac{e^{-(x-b)}}{1 + e^{-(x-b)}} \qquad (8)$$



شکل 4. نمودار تابع عضویت برای AlexaRank

4) الگوریتم آموزش شبکه عصبی: الگوریتم پیشنهادی در شکل 5 نشان داده شده است. الگوریتم دو فاز را به صورت زیر اجرا می کند:



شکل 5. الگوریتم آموزش شبکه عصبی

- فاز انتشار ارزش ورودی گره خروجی و ارزش خروجی گره خروجی را محاسبه می کند. ارزش ورودی گره خروجی توسط (9) محاسبه شده است

$$O_I = \sum_{i=1}^{6} W_i * I_i \qquad (9)$$

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

ارزش خروجی گره خروجی توسط (10) محاسبه می شود

$$O_O = \frac{1}{1 + e^{-O_I}} \qquad (10)$$

که O_O and O_I ارزش خروجی گره خروجی و ارزش ورودی گره خروجی می باشند.

- 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 (11) محاسبه می شود

$$Err = O_O * (1 - O_O) * (T - O_O) \qquad (11)$$

که T ارزش واقعی نمونه در مجموعه داده آموزش است. وزن ها توسط (12) به روز رسانی می شود

$$W_i = W_i + R * Err * O_O \qquad (12)$$

که R و W_i نرخ یادگیری و وزن اامین گره ورودی می باشد.

IV. ارزیابی

11660 سایت فیشینگ از PhishTank [2] و 10000 سایت قانونی از DMOZ جمع آوری کردیم [17]. مجموعه داده های آموزش شامل 6660 سایت فیشینگ از PhishTank و 5000 سایت قانونی از DMOZ بود. 2 مجموعه داده آزمون ایجاد کردیم، هریک از آن ها شامل 5000 سایت فیشینگ یا 5000 سایت قانونی بود. روند تجربی از طریق PHP و MYSQL به 2 فاز (آموزش و آزمون) تقسیم شد.

A. فاز آموزش

1) وارد کردن مجموعه داده آموزش: مجموعه داده آموزش به MYSQL وارد شد. نتیجه در شکل 6 نشان داده شده است.

phish_id	url	phish_detail_url	submission_time	verified	verification_time
2111050	http://www.montenegrodrive.me/components/googledoc...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 09:12:02	yes	2013-11-17 14:21:40
2111010	http://itunesconnect.apple.com.jooltec.com.br/upda...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 09:08:17	yes	2013-11-17 13:58:52
2111001	http://kuznyanova.org.ua/deal/googledocss/googledo...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 09:07:32	yes	2013-11-17 14:07:39
2110997	http://parnasseweb.tn/wp-includes/js/my.screenname...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 09:07:09	yes	2013-11-17 14:08:15
2110988	http://paypal.com-inc-security-account-45453612358...	http://www.phishtank.com/phish_detail.php?phish_id...	2013-11-17 09:06:17	yes	2013-11-17 14:01:12

شکل 6. مجموعه داده های آموزش از 11660 سایت در MYSQL

2) استخراج چهار ویژگی URL: چهار ویژگی (دامنه اصلی، زیردامنه، دامنه مسیر و دامنه) استخراج شدند. شکل 7 نتیجه بدست آمده را نشان می دهد.

phish_id	domain	primarydomain	subdomain	pathname
2111050	montenegrodrive.me	montenegrodrive		components,googledoc,index.htm
2111010	jooltec.com.br	jooltec	itunesconnect,apple,com	updats,
2111001	kuznyanova.org.ua	kuznyanova		deal,googledocss,googledocss,sss
2110997	parnasseweb.tn	parnasseweb		wp,includes,js,my.screenname.aol.com,my.screenname...
2110988	sorpi.fr	sorpi	paypal,com,inc,security,account	cmd,home&dispatch,2f643150d63de9bd3e4d110f71b5...

شکل 7. چهار ویژگی استخراج شده است

3) محاسبه ی ارزش شش گره ورودی: پیشنهاد املایی موتور جستجوی گوگل و Alexa.com برای محاسبه ی ارزش گره های ورودی مورد استفاده قرار می گیرند. نتیجه در شکل 8 نشان داده شده است.

phish_id	primarydomain	subdomain	pathdomain	pagerank	alexarank	alexareputation
2111050	100	100	2	0	6274104	2
2111010	100	0	100	0	6274104	2
2111001	100	100	0	1	6274104	2
2110997	23	100	0	0	160379	18
2110988	5	0	100	0	7104259	1

شکل 8. ارزش اکتشافات

4) محاسبه ارزش فازی 12 گره در دومین لایه: دو تابع هضویت سیگموئید چپ و راست برای محاسبه ی ارزش گره ها در دومین لایه مورد استفاده قرار می گیرند. نتیجه در شکل 9 نشان داده شده است.

phish_id	P1	P2	P3	P4	P5	P6	L1	L2	L3	L4	L5	L6
2111050	0.00	0.00	0.88	1.00	1.00	1.00	1.00	1.00	0.12	0.00	0.00	0.00
2111010	0.00	0.98	0.00	1.00	1.00	1.00	1.00	0.02	1.00	0.00	0.00	0.00
2111001	0.00	0.00	0.98	0.99	1.00	1.00	1.00	1.00	0.02	0.01	0.00	0.00
2110997	0.00	0.00	0.98	1.00	0.00	0.88	1.00	1.00	0.02	0.00	1.00	0.12
2110988	0.27	0.98	0.00	1.00	1.00	1.00	0.73	0.02	1.00	0.00	0.00	0.00

شکل 9. ارزش های فازی در دومین لایه

5) فاز آموزش شبکه: آموزش شبکه را با 9 ارزش نرخ یادگیری اجرا کردیم. در فاز آموزش، پارمترها به صورت زیر می باشد:

- نرخ یادگیری: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
- ارزش آستان میانگین خطا: 1٪
- تعداد Epoch: 10000
- وزن ها: وزن های اولیه ارزش های تصادفی از 0 تا 1

B. فاز آزمون

در این فاز، تکنیک پیشنهادی با 2 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, 0.9 آزمون می شود. RMSE (خطای جذر میانگین مربعات) مقیاس خوبی برای شناسایی دقت است.RMSE توسط معادله (13) محاسبه می شود

$$RMSE = \sqrt{\frac{\sum(A_i - I_i)^2}{N}} \qquad (13)$$

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=100-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 99.31٪ را با نرخ یادگیری 0.7 و بدیترین نسبت 98.32٪ با نرخ یادگیری 0.2 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	4,925
0.1	No.2	5,000	4,917
0.2	No.1	5,000	4,911
0.2	No.2	5,000	4,914
0.3	No.1	5,000	4,926
0.3	No.2	5,000	4,933
0.4	No.1	5,000	4,946
0.4	No.2	5,000	4,935
0.5	No.1	5,000	4,933
0.5	No.2	5,000	4,927
0.6	No.1	5,000	4,925
0.6	No.2	5,000	4,927
0.7	No.1	5,000	4,963
0.7	No.2	5,000	4,959
0.8	No.1	5,000	4,914
0.8	No.2	5,000	4,915
0.9	No.1	5,000	4,920
0.9	No.2	5,000	4,912

جدول II - RMSE و درستی با تکنیک پیشنهادی

Learing rate	RMSE	Accuracy
0.1	1.58	98.42%
0.2	1.75	98.25%
0.3	1.41	98.59%
0.4	1.20	98.80%
0.5	1.40	98.60%
0.6	1.48	98.52%
0.7	0.78	99.22%
0.8	1.71	98.29%
0.9	1.68	98.32%

C. مقایسه با تکنیک [11]

با تکنیک (11) آزمایش کردیم و با نتیجه ی تکنیک پیشنهادی مان مقایسه کردیم. نخست، 10 مجموعه داده آزمون را جمع آوری کردیم که هریک شامل 1000 سایت فیشینگ یا 1000 سایت قانونی است. دوم، تکنیک (11) را آزمایش کردیم و نتایج در جدول III نشان داده شده اند.از نتیجه بدست آمده و استفاده از RMSE، دریافتیم که تکنیک (11) درستی 86.06٪ دارد.

جدول III نتایج آزمون با تکنیک (11)

Testing dataset	(1)	(2)	(3)
No.1	867	82	51
No.2	865	76	59
No.3	847	90	63
No.4	902	172	26
No.5	841	109	50
No.6	64	873	63
No.7	50	911	39
No.8	39	895	66
No.9	97	871	32
No.10	85	863	52

D. مقایسه با تکنیک [12]

با تکنیک (12) با استفاده از 8 گره پنهان و تابع فعال سازی مماس هذلولی آزمایش کردیم. نخست، 2 مجموعه داده آزمون جمع آوری کردیم که هر یک شامل 5000 سایت فیشینگ یا 5000 سایت قانونی بود. دوم، تکنیک (12) 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 (12)، بهترین درستی 94.68٪ را بدست آوردیم.

جدول IV نتیجه آزمون با تکنیک [12]

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 [12]

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%

۷. نتیجه گیری و کارهای آتی

تکنیک جدیدی را برای شناسایی موثر سایت فیشینگ پیشنهاد کرده ایم. در این تکنیک، مدل سیستم برای شناسایی سایت فیشینگ با استفاده از شبکه عصبی فازی و شش اکتشافی (Primarydomain, subdomain, pathdomain, pagerank, alexarank, alexareputation) ایجاد شده است. تکنیک با مجموعه داده های آموزش شامل 11660 سایت و 2 مجموعه داده آزمون آزمایش شد که هر مجموعه داده شامل 5000 سایت فیشینگ یا 5000 سایت قانونی بود. بهترین نتایج نشان داد که 99.22٪ از وب سایت های فیشینگ با استفاده از مدل سیستم شناسایی می شوند. کار ما با نتایج در [12] ,[11] مقایسه شد و دریافت که کارآمد تر است. در آینده، مدل عصبی فازی ما برای افزایش نسبت شناسایی بهبود خواهد یافت. بعلاوه، سیستم می تواند با استفاده از مجموعه داده های بزرگ تر و پارامترهای اکتشافی بیشتر افزایش یابد.

REFERENCES

[1] Ollman, G. (2004) The Phishing Guide —Understanding and Preventing. White Paper, Next Generation Security Software Ltd.

[2] PhishTank. (2013, Nov.) Statistics about phishing activity and phishtank usage. [Online]. Available: http://www.phishtank.com/stats/2013/01/

[3] D. Goodin. (2012) Google bots detect 9,500 new malicious websites every day. [Online]. Available: http://arstechnica.com/security/2012/06/

[4] 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

[5] McAfee. (2011, July) Mcafee site advisor. [Online]. Available: http://www.siteadvisor.com

[6] 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

[7] 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.

[8] 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.

[9] A. Sunil and A. Sardana, A pagerank based detection technique for phishing web sites, in IEEE Symposium on Computers & Informatics, 2012, pp. 58—63.

[10] 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

[11] 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.

[12] N. Zhang and Y. Yuan, Phishing Detection Using Neural Network, CS229 lecture notes, http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf, 2012

[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] Alexa. [Online]. Available (2014) : http://data.alexa.com/data?cli=10&dat=snbamz&url=

[17] DMOZ. [Online]. Available (2014) : http://rdf.dmoz.org/rdf/