رویکرد کارآمد برای ردیابی فیشینگ با استفاده از مدل عصبی فازی

چکیده

امروزه، تعاملات آنلاین در جامعه مدرن معروف تر و معروف تر می شود. در نتیجه،فیشینگ تلاشی از جانب فرد یا گروهی از افراد برای سرقت اطلاعات شخصی مانند کلمه عبور، اطلاعات حساب بانکی، و کارت اعتباری و غیره می باشد. اکثر این صفحات وب فیشینگ مانند صفحات اصلی در قالب رابط وب سایت و ادرس وب (URL) هستند. تکنیک های بسیاری برای ردیابی وب سایت های فیشینگ پیشنهاد شده است مانند تکنیک های مبتنی بر بلک لیست، تکنیک مبتنی بر اکتشاف و غیره. با این حال، تعداد قربانیان به دلیل تکنیک محافظت ناکارآمد در حال افزایش است. شبکه های عصبی و سیستم های فازی می تواند برای داشتن مزایای مشترک و از بین بردن مشکلات مجزا با هم ترکیب شوند. این مقاله مدل عصبی فازی بدون استفاده از مجموعه قوانین برای ردیابی فیشینگ پیشنهاد می کند. بویژه، تکنیک پیشنهادی ارزش اکتشافی از توابع عضویت را محاسبه می کند. سپس، وزن ها توسط شبکه عصبی ایجاد می شوند. تکنیک پیشنهادی با مجموعه داده های ۱۱۶۶۰ سایت فیشینگ و ۱۰۰۰۰ سایت قانونی ارزیابی می شوند. نتایج نشان می دهد که تکنیک پیشنهادی می تواند بیش از ۹۹٪ از سایت فیشینگ را ردیابی کند.

واژگان کلیدی: فیشینگ، عصبی فازی، شبکه عصبی

I. مقدمه

فیشرها مجموعه ای از تکنیک ها برای فریب دادن قربانیان می باشند که شامل پیام های ایمیل، پیام های فوری، پست های انجمن، تماس های تلفنی، پیام های متنی می باشد. این فعالیت های فیشینگ منجر به خسارت اقتصادی شدید در سراسر جهان می شود. نیمه دوم گزارش APWG برای ۲۰۱۰ ادعا کرده است که حملات فیشینگ ۱۴۲٪ بیش از نیمه ی اول ۲۰۱۰ رشد کرده است. گزارش هدف ها را به این صورت طبقه بندی می کند که شامل ۳۷,۹٪ خدمات پرداخت، ۳۳,۱٪ نهادهای مالی، ۶,۶٪ طبقه بندی شده، ۴,۶٪ بازی، ۲,۸٪ شبکه های اجتماعی و باقی در

دسته های دیگر می باشد. در سال ۲۰۱۱، ۸۳٪ از آمریکایی ها و ۸۵٪ از اروپایی ها مرتبا به صورت آنلاین خرید می کنند (Fortune Magazine, 2011). در همین حین، سایت های فیشینگ نیز از لحاظ کیفیت و تعداد در حال افزایش هستند . بنابراین خطر سرقت اطلاعات کاربر بسیار بالا است. به این دلایل، ردیابی مشکل فیشینگ در جامعه مدرن فوری، پیچیده و بسیار مهم تلقی می شود. اخیرا، مطالعاتی وجود داشته است که برخلاف فیشینگ بر اساس ویژگی های سایت مانند URL وب سایت،محتوای وب سایت، هم URL وب سایت و هم محتوا، کد منبع وب سایت یا اسکرین شات وب سایت را ترکیب می کند. با این وجود، هر مطالعه نقاط ضعف و قوت مختص خود را داراست. هنوز هم روش کافی ای وجود ندارد. در این مقاله، رویکرد جدیدی پیشنهاد شده است تا سایت های فیشینگ را ردیابی کند که بر ویژگی های URL (دامنه اصلی، زیردامنه، دامنه مسیر) و سایت های رتبه بندی (PageRank, AlexaRank, AlexaReputation) متمرکز است. پس، یک شبکه عصبی فازی پیشنهادی سیستمی است که خطا را کاهش و عملکرد را افزایش می دهد. مدل عصبی فازی پیشنهادی از مدل های محاسباتی استفاده می کند تا بدون مجموعه قوانین اجرا شود. راه حل پیشنهادی دقت ردیابی بالای ۹۹٪ با سیگنال کاذب پایین 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 درستی 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 مرتبط

تکنیک های ردیابی فیشینگ به سه طبقه ی بلک لیست، اکتشافی، و یادگیری ماشین طبقه بندی شده اند. در رویکرد نخست، تکنیک ردیابی فیشینگ (۱)-(۴) فهرستی از وب سایت های فیشینگ به نام بلک لیست را بدست می آورد. با این حال، تکنیک بلک لیست به دلیل رشد سریع تعداد سایت های فیشینگ ناکارآمد است. بنابراین، رویکردهای اکتشفای و یادگیری ماشین توجه محققان را به خود معطوف ساخته است. Cantina (۵) الگوریتم TF-IDF را بر مبنای ۲۷ ویژگی صفحه وب ارائه کرده است. این تکنیک می تواند ۹۷٪ از سایت های فیشینگ با ۶٪

مثبت کاذب را ردیابی کند. اگرچه این تکنیک کارآمد است، زمان استخراج ۲۷ ویژگی وب سایت برای تقاضای زمان واقعی طولانی است و برخی از ویژگی ها برای بهبود درستی ردیابی فیشینگ ضروری نیستند. بعلاوه، مجموعه داده های ارزیابی بسیار کوچک است. به همین نحو، Cantina+ (6) از تکنیک یادگیری ماشین بر مبنای ۱۵ ویژگی برای صفحه وب استفاده کرد و تنها ۶ مورد از ۱۵ ویژگی برای ردیابی فیشینگ کارآمد هستند مانند فرم بد، فیلد فعالیت بد، URL غیرمنطبق، صفحه در بالای نتیجه جستجو، حق کپی رایت به اضافه دامنه و حق کپی رایت جستجو به اضافه نام میزبان. در (۷)، نویسنده از URL برای ردیابی سایت فیشینگ به صورت خودکار با استخراج و تایید عبارات متفاوت URL از طریق موتور جستجو استفاده کرد. اگرچه این مقاله تکنیک جالب و جدیدی را ارائه می دهد، نرخ ردیابی نسبتا پایین (۵۴,۳٪) است. تکنیک (۸) یک رویکرد مبتنی بر محتوا را برای ردیابی فیشینگ به نام CANTINA توسعه داد که ارزش صفحه ی PageRank گوگل را در نظر می گیرد، با این حال مجموعه داده ارزیابی نسبتا کوچک است.ویژگی کد منبع برای دریابی سایت های فیشینگ در(۹) مورد استفاده قرار گرفته است. نویسندگان در (۱۰) تکنیک های فازی بر اساس ۲۷ ویژگی صفحه وب پیشنهاد کردند که به ۳ لایه طبقه بندی می شود. هر ویژگی سه مقدار زبانی کم، متوسط بالا دارد. تکنیک فازی مجموعه قوانین، توابع عضویت مثلثی و ذوزنقه ای ایجاد کرده است. نرخ فیشینگ وب سایت بدست آمده از تکنیک ۸۶,۲٪ است. با این حال، نقص های بسیاری در (۱۰) وجود دارد. نخست، مجموعه قوانین عینی نیستند و بسیار بر سازنده بستگی دارند. دوم، وزن هر معیار اصلی بدون مشخص شدن مورد استفاده قرار گرفته است. در نهایت، اکتشاف پیشنهادی بهینه و موثر است. نویسندگان (۱۱) تکنیک شبکه عصبی را پیشنهاد کردند. تکنیک (۱۱) سه لایه شامل لایه ورودی، لایه پنهان و لایه خروجی ایجاد کرده است. بهترین نرخ بدست آمده از تکنیک ۹۵٪ است. با این وجود، 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 لایه پنهان استفاده می کنند. سوم، ارزش ویژگی ها بدون شفافیت محاسبه می شود. در نهایت، مجموعه داده ها به اندازه کافی بزرگ نیستند.

در تکنیک های پیشین، URL نقش جزیی ای در ردیابی وب سایت های فیشینگ داشت. در این مقاله، بر ویژگی های URL تمرکز می کنیم و از تکنیک عصبی فازی برای ردیابی سایت های فیشینگ استفاده می کنیم. سهم مقاله ما به شرح زیر می باشد: i) اکتشاف های جدید برای ردیابی موثر و سریع وب سایت فیشینگ پیشنهاد شده است. ii) ارزش های آستانه مورد استفاده دز توابع عضویت از مجموعه داده های بزرگ نشات گرفت تا اینکه مدل برابر با مجموعه داده های جدید باشد. iii) وزن های اکتشاف بهینه تر است که به این دلیل می باشد که وزن ها با استفاده از شبکه عصبی آموزش دیده شده اند. Iv) مجموعه قوانین استفاده نمی شوند. بنابراین، نتیجه دقیق و عینی خواهد بود.

III. طرح سیستم

A. شبکه عصبی فازی بدون مجموعه قوانین

شبکه های عصبی و منطق فازی که تکنیک های رایانشی نرم نامیده می شوند، ابزارهای ایجاد سیستم های هوشمند می باشند. سیستم رابط فازی (FIS) که از قوانین if-then فازی در کسب دانش از متخصصان انسانی استفاده می کند می تواند مشکلات مبهم و غیردقیق را بررسی کنند (۱۲). FIS به طور گسترده در کاربردهایی مانند بهینه سازی، کنترل، و شناسایی سیستم استفاده شده اند. سیستم های فازی معمولا خودشان یاد نمی گیرند و تنظیم نمی شوند (۱۳)، در خالیکه شبکه عصبی (NN) ظرفیت یادگیری از محیط، سازمان دهی خود، و تطبیق در روشی تعاملی را دارا هستند. به این دلایل، سیستم عصبی فازی که ترکیبی از سیستم فازی و عصبی است، برای تولید سیستم مبتنی بر قانون فازی معرفی شده است (۱۴)، (۱۵). با این وجود، مجموعه قوانین عینی نیستند و بر سازنده بستگی دارند، بنابراین مجموعه قوانین در مدل عصبی فازی پیشنهادی استفاده نشده است. بنابراین، نتیجه دقیق و عینی خواهد بود.

B. URL

یک URL (مکان یکنواخت منبع) برای مکانیابی منابع استفاده می شود(۱۶).ساختار URL به صورت زیر می باشد:

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

برای مثال، با URL:

http://www.paypal.abc.net/login/index.html

شش مولفه به صورت زیر وجود دارد: پروتکل http است، زیردامنه paypal می باشد، دامنه اصلی abc است، TLD شبکه است، دامنه abc.net است، دامنه مسیر login/index.html است.

C. ویژگی URL

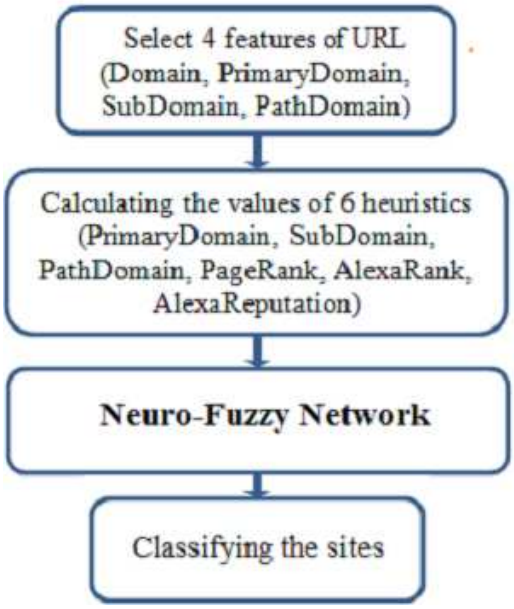
فیشرها معمولا سعی دارند که آدرس اینترنتی (URL) سایت های فیشینگ را مشابه سایت های قانونی بسازند تا کاربران آنلاین را فریب دهند. آن ها نمی توانند از URL دقیق سایت قانونی استفاده کنند، آنها خطاهای املایی بیشتری از ویژگی های URL ایجاد می کنند مانند PrimaryDomain, SubDomain, PathDomain. برای مثال، URL www.applle.com 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.

D. ویژگی رتبه بندی دامنه

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.

E. 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)

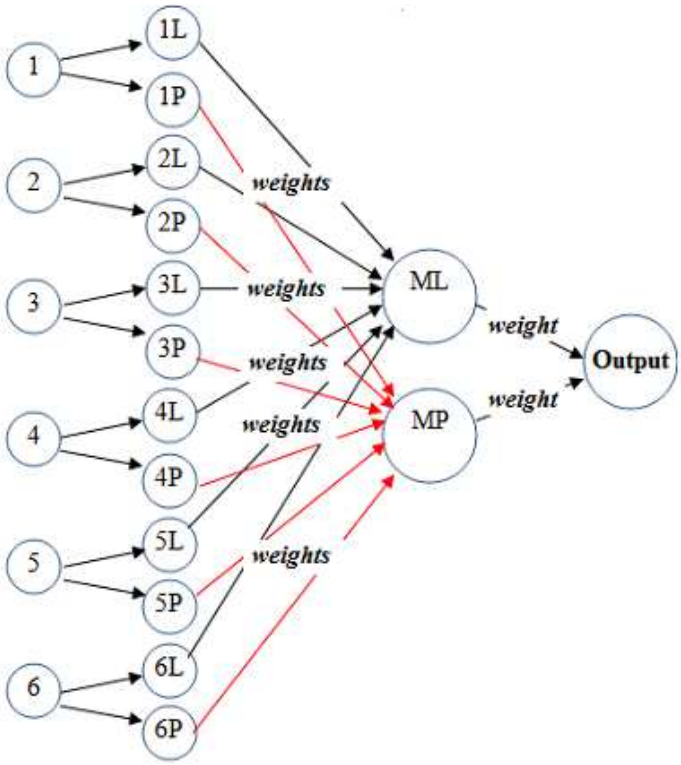
شکل ۱. مدل سیستم

- **فاز I**- انتخاب چها ویژگی URL: چهار ویژگی از URL استخراج شده است مانند دامنه، دامنه اصلی، زیردامنه و دامنه مسیر.
- **فاز II**- محاسبه ی شش مقدار اکتشافی: شش مقدار اکتشافی محاسبه می شوند و ۶ مورد اکتشافی ۶ گره ورودی شبکه عصبی فازی هستند.
- **فاز III**- شبکه عصبی فازی: شبکه عصبی فازی برای محاسبه ی مقدار گره خروجی اجرا می شود.
- **فاز IV**- طبقه بندی وب سایت ها: بر اساس مقدار گره خروجی تصمیم میگیریم که آیا یک سایت، سایت فیشینگ است.

F. مدل شبکه عصبی فازی

۱) مدل

مدل شبکه عصبی فازی به صورت شکل ۲ طراحی شده است.



شکل ۲. مدل شبکه عصبی فازی

مدل با چهار لایه به صورت زیر طراحی شد:

- لایه نخست به نام لایه ورودی شامل شش گره است که شش اکتشاف به نام دامنه اصلی، زیردامنه، دامنه مسیر، PageRank، AlexaRank، AlexaReputation است.
- لایه دوم شامل ۱۲ گره است. ارزش هر گره ، ارزش فازی است و از تابع عضویت Sشکل و zشکل محاسبه می شود.
- لایه سوم شامل دو گره ML و MP است. ML (میانگین قانونی) مجموع توزینی گره های L در لایه دوم است. MP (میانگین فیشینگ) مجموع توزینی گره های P در لایه دوم است.
- لایه چهارم به نام لایه خروجی تنها یک گره خروجی دارد.

تابع فعال سازی سیگموئید در شبکه عصبی پیشنهادی استفاده می شود و ارزش خروجی گره خروجی از ۰ تا ۱ است. مدل پیشنهادی به دو طبقه تقسیم می شود بنابراین اگر ارزش گره خروجی کمتر از ۰,۵ باشد سایت فیشینگ است و اگر بالاتر یا برابر با ۰,۵ باشد سایت قانونی است.

۲) ارزش شش گره ورودی

بر اساس نتایج تجربی و آمار از مجموعه داده های ۱۱۶۶۰ سایت فیشنگ می باشد. در یافتیم که:

- زمانی سایت فیشینگ است که فاصله Levenshtein (۱۷) بین دامنه اصلی، زیردامنه، دامنه مسیر و نتیجه ی پیشنهاد املایی موتور جستجوی گوگل باشد کمتر از ۴ باشد.
- ارزش PageRank از 1- تا ۱۰ است. زمانی سایت فیشینگ است که ارزش کمتر داشته باشد.
- زمانی سایت فیشینگ است که ارزش AlexaRank بیشتر از ۳۰۰۰۰۰ باشد.
- زمانی سایت فیشینگ است که ارزش AlexaReputation کمتر از ۳۰ باشد.

شش ارزش اکتشافی به صورت زیر محاسبه می شود:

- محاسبه ی ارزش اکتشافی دامنه اصلی: الگوریتم در شکل ۳ نشان داده شده است.

```
Input: PrimaryDomain
Output: Value of heuristic "PrimaryDomain"

If PrimaryDomain is IP then value= 0; // doubt phishing
If  PrimaryDomain is not IP then
        result = Suggestion_Google(PrimaryDomain);
        If result is null then
            value = 4; // no doubt phishing
        End if
        If result is not null then
            value=Levenshtein(result, PrimaryDomain);
        End if
End if
```

شکل ۳. محاسبه ی ارزش اکتشافی دامنه اصلی

- محاسبه ی ارزش اکتشافی زیردامنه و دامنه مسیر: الگوریتم در شکل ۴ نشان داده شده است.

```
Input: m     // m is SubDomain or PathDomain
Output: Value of heuristic "m"

If SubDomain is null then value=4; // no doubt phishing
If SubDomain is not null then
    result = Suggestion_Google(m);
    If result is null then
        value = 4; // no doubt phishing
    End if
    If result is not null then
        value= Levenshtein(result, m);
    End if
End if
```

شکل ۴. محاسبه ی ارزش اکتشاف زیردامنه و دامنه مسیر

- محاسبه ی ارزش اکتشافی PageRank: ارزش Google PageRank می تواند از (۱۸) بدست آید. PageRank از 1- تا ۱۰ است.
- محاسبه ی ارزش اکتشافی AlexaRank و AlexaReputation: این دو می تواند از (۱۹) محاسبه شوند.

۳) ارزش ۱۲ گره در دومین لایه

طبقه بندی اکتشاف در دو عنوان زبانی و تخصیص توابع عضویت مانند S شکل و Z شکل برای هر ارزش زبانی. هر یک از این اکتشاف ها به عنوان زبانی فیشینگ و قانونی طبقه بندی می شود. معادله ۱ و ۲دو تابع عضویت S شکل و Z شکل است. بر اساس نتایج تجربی و آمار از مجموعه داده ۱۱۶۶۰ سایت فیشینگ، توابع عضویت به صورت زیر محاسبه می شوند:

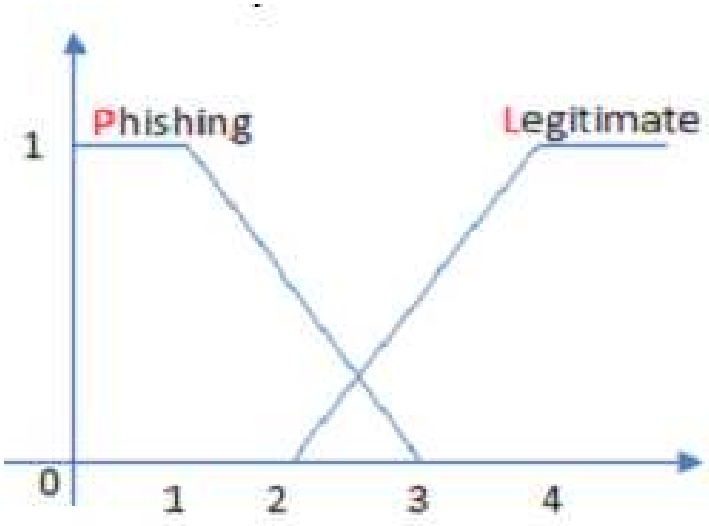
$$Z(x,a,b)=\begin{cases}1, & x\le a\\ 1-2\left(\frac{x-a}{b-a}\right)^2, & a<x\le\frac{a+b}{2}\\ 2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2}<x<b\\ 0, & x\ge b\end{cases} \qquad (1)$$

$$S(x,a,b)=\begin{cases}0, & x\le a\\ 2\left(\frac{x-a}{b-a}\right)^2, & a<x\le\frac{a+b}{2}\\ 1-2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2}<x<b\\ 1, & x\ge b\end{cases} \qquad (2)$$

- توابع عضویت برای دامنه اصلی، زیر دامنه و دامنه مسیر: معادله (۳) و (۴) دو تابع عضویت است که برای محاسبه ی ارزش های فازی ایجاد شده است و نمودار توابع عضویت در شکل ۵ نشان داده شده است.

$$P(x) = Z(x, 1, 3) \qquad (3)$$

$$L(x) = S(x, 2, 4) \qquad (4)$$

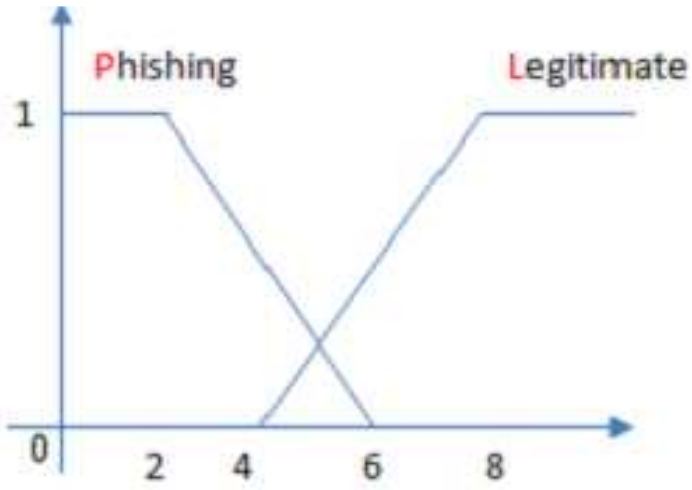


شکل ۵. نمودار تابع عضویت

- توابع عضویت برای PageRank: معادله ۵ و ۶، دو تابع عضویت است که برای محاسبه ی ارزش های فازی ایجاد شده است و نمودار توابع عضویت در شکل ۶ نشان داده شده است.

$$P(x) = Z(x, 2, 6) \qquad (5)$$

$$L(x) = S(x, 4, 8) \qquad (6)$$



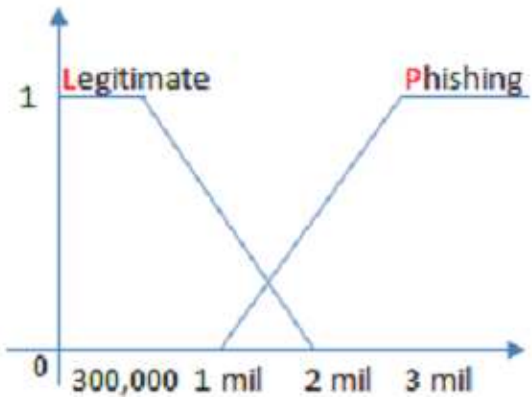
شکل ۶. نمودار تابع عضویت PageRank

- توابع عضویت برای AlexaRank: معادل ۷ و ۸ دو تابع عضویت است که برای ارزش های فازی ایجاد شده است و نمودار توابع عضویت در شکل ۷ نشان داده شده است.

$$P(x) = S(x, 1mil, 3mil) \quad (7)$$

$$L(x) = Z(x, 300k, 2mil) \quad (8)$$

که 300k و mil خلاصه ی ۳۰۰۰۰۰ و میلیون است.

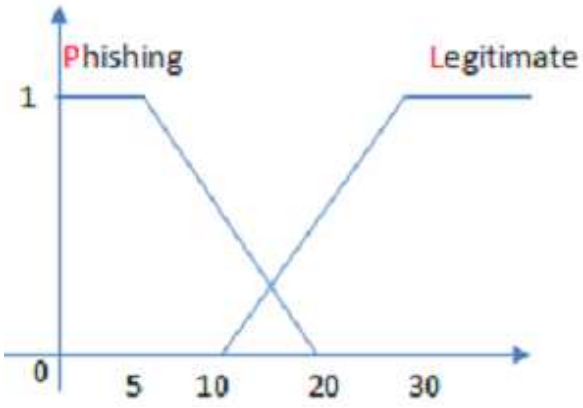


شکل ۷. نمودار تابع عضویت AlexaRank

- توابع عضویت برای AlexaReputation: معادله ۹ و ۱۰ دو تابع عضویت است که برای محاسبه ارزش های فازی ایجاد شده است و نمودار توباع عضویت در شکل ۸ نشان داده شده است

$$P(x) = Z(x, 5, 20) \quad (9)$$

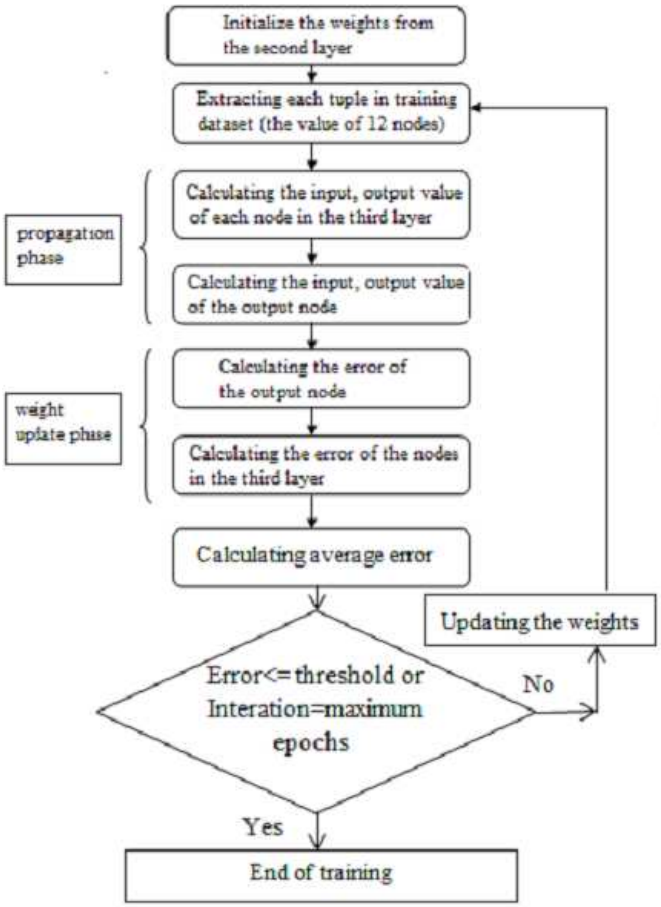
$$L(x) = S(x, 10, 30) \quad (10)$$



شکل ۸. نمودار تابع عضویت AlexaReputation

۴) الگوریتم آموزش شبکه

الگوریتم پیشنهادی در شکل ۹ نشان داده شده است. الگوریتم دو فاز را به صورت زیر ایفا می کند:



شکل ۹. الگوریتم آموزش شبکه

- فاز انتشار ارزش ورودی، ارزش خروجی هر گره در لایه سوم و لایه خروجی را محاسبه می کند. ارزش ورودی گره ها توسط (۱۱) محاسبه می شود.

$$I_j = \sum_i W_{ij} O_i \qquad (11)$$

که I_j, O_i and W_{ij} ارزش ورودی گره**j**ام در لایه کنونی، ارزش خروجی گره **i**ام در لایه پیشین و وزن از **h**امین گره لایه قبلی گره **j**ام در لایه کنونی است.

ارزش خروجی گره ها توسط (۱۲) محاسبه می شود.

$$O_j = \frac{1}{1+e^{-I_j}} \qquad (12)$$

که I_j, O_j ارزش ورودی، ارزش خروجی گره **j**ام می باشد.

- فاز به روز رسانی خطای گره ها در لایه سوم و لایه خروجی را محاسبه می کند و سپس وزن ها را به روز رسانی می کند. خطای گره خروجی توسط (۱۳) محاسبه می شود.

$$Err = O_o * (1 - O_o) * (T - O_o) \qquad (13)$$

که T, O_O ارزش واقعی نمونه در مجموعه داده ها، ارزش خروجی گره خروجی است.

خطای گره **j**ام در لایه سوم توسط (۱۴) محاسبه می شود

$$Err_j = O_j * (1 - O_j) * \sum Err * W_j \qquad (14)$$

که O_j, W_j and Err به ترتیب ارزش خروجی گره **j**ام، وزن ارتباط از گره **j**ام به گره خروجی و خطای گره خروجی می باشد.

وزن ها از لایه دوم به لایه سوم توسط (۱۵) به روز رسانی می شود

$$W_{ij} = W_{ij} + R * Err_j * O_i \qquad (15)$$

که R, Err_j, O_i نرخ یادگیری، خطای گره **j** ام در لایه سوم و ارزش خروجی گره **i** ام در لایه دوم می باشد.

وزن های متصل از لایه سوم به لایه خروجی توسط (۱۶) به روز رسانی می شود

$$W_i = W_i + R^* Err^* O_i \qquad (16)$$

که Err, O_i خطای گره خروجی و ارزش خروجی iامین گره در لایه سوم می باشد.

IV. ارزیابی

۱۱۶۶۰ سایت فیشینگ از PhishTank]۱[و ۱۰۰۰۰ سایت قانونی از DMOZ جمع آوری کردیم (۲۰). 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

- وارد کردن مجموعه داده های آموزش: مجموعه داده های آموزش به MYSQL وارد شده اند. نتیجه در شکل ۱۰ 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 ۱۰. وارد کردن MYSQL

- استخراج چهار ویژگی از URL: چهار ویژگی دامنه ی اصلی، زیردامنه، دامنه مسیر و دامنه استخراج شده اند. نتیجه در شکل ۱۱ نشان داده شده است.

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 ۱۱. انتخاب دامنه اصلی، زیردامنه، دامنه مسیر و مسیر

- محاسبه 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 ۱۲ نشان داده شده است.

phish_id	primarydomain	subdomain	pathdomain	pagerank	alexarank	alexareputation
2111050	4	4	2	0	6274104	2
2111010	4	0	4	0	6274104	2
2111001	4	4	0	1	6274104	2
2110997	23	4	0	-1	160379	18
2110988	5	0	4	0	7104259	1

شکل ۱۲. ارزش شش اکتشاف

- محاسبه ارزش ۱۲ گره در لایه دوم: دو تابع عضویت S شکل یا Z شکل برای محاسبه ی ارزش گره در دومین لایه مورد استفاده قرار می گیرد. نتیجه در شکل ۱۳ نشان داده شده است.

phish_id	P1	P2	P3	P4	P5	P6	L1	L2	L3	L4	L5	L6
2111050	0.00	0.00	0.50	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
2111010	0.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00
2111001	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
2110997	0.00	0.00	1.00	1.00	0.00	0.28	1.00	1.00	0.00	0.00	1.00	0.32
2110988	0.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00

شکل ۱۳. ارزش های فازی

- فاز آموزش شبکه: آموزش شبکه با ۹ ارزش نرخ یادگیری اجرا شد. در فاز آموزش، پارامترها به صورت زیر می باشد:
  - نرخ یادگیری: 0.1, 0.2, 0.3, 04, 0.5, 0.6, 0.7, 0.8, 0.9
  - ارزش آستانه خطای میانگین: ۱٪
  - تعداد Epoch: ۱۰۰۰۰
  - وزن ها: وزن های اولیه ارزش های تصادفی 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, 0.9 آزمون می شود. RMSE (خطای جذر میانگین مربعات) مقیاس خوبی برای شناسایی دقت است.RMSE توسط معادله (۱۷) محاسبه می شود

$$RMSE = \sqrt{\frac{\sum (A_i - D_i)^2}{N}} \qquad (17)$$

که Di تعداد سایت های ردیابی 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 ۹۹,۱۰٪ را با نرخ یادگیری ۰,۷ و بدترین نسبت ۹۸,۲۲٪ با نرخ یادگیری ۰,۲ و ۰,۸را نشان می دهد.

جدول I. نتیجه آزمون با تکنیک پیشنهادی

Learning Rate	Testing dataset	Actual Sites (A_i)	Detecting Sites (D_i)
0.1	No.1	5,000	4,918
	No.2	5,000	4,916
0.2	No.1	5,000	4,908
	No.2	5,000	4,914
0.3	No.1	5,000	4,914
	No.2	5,000	4,931
0.4	No.1	5,000	4,939
	No.2	5,000	4,924
0.5	No.1	5,000	4,933
	No.2	5,000	4,921
0.6	No.1	5,000	4,925
	No.2	5,000	4,919
0.7	No.1	5,000	4,955
	No.2	5,000	4,955
0.8	No.1	5,000	4,914
	No.2	5,000	4,908
0.9	No.1	5,000	4,920
	No.2	5,000	4,912

جدول II. RMSE و دقت با تکنیک پیشنهادی

Learning Rate	RMSE	Accuracy
0.1	1.66	98.34%
0.2	1.78	98.22%
0.3	1.56	98.45%
0.4	1.38	98.62%
0.5	1.46	98.54%
0.6	1.56	98.44%
0.7	0.90	99.10%
0.8	1.78	98.22%
0.9	1.68	98.32%

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) را 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. نتیجه آزمون با تکنیک (۱۰)

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 (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 نتیجه آزمون با تکنیک (۱۱)

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

جدول V. RMSE و دقت با تکنیک (۱۱)

Learning 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 و کارهای آتی

تکنیک جدیدی برای ردیابی موثر سایت های فیشینگ پیشنهاد کردیم. در تکینک پیشنهادی، مدل سیستم برای ردیابی سایت های فیشینگ با شبکه عصبی فازی و شش اکتشاف (دامنه اصلی، زیردامنه، دامنه مسیر، pagerank، alexarank, alexareputation) ایجاد شده است. تکنیک با مجموعه داده آموزش شامل ۱۱۶۶۰ سایت و ۲ مجموعه داده آزمون آزمایش شد که شامل ۵۰۰۰ سایت فیشینگ یا ۵۰۰۰ سایت قانونی است. بهترین نتیجه نشان داد که ۹۹,۱۰٪ وب سیات فیشینگ با استفاده از تکنیک پیشنهادی ردیابی شده است. تکنیک پیشنهادی با تکنیک (۱۰)، تکنیک (۱۱) مقایسه می شود و دریافت که کارآمد تر است. در آینده،مدل عصبی فازی برای افزایش نسبت ردیابی بهبود خواهد یافت. بعلاوه، سیستم می تواند با استفاده از مجموعه داده های بزرگ تر و پارامترهای اکتشافی افزایش یابد.

REFERENCES

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

[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. (July 2011). 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 *Proc. 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 *Proc. 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," *International Journal of Information and Communication Technology Research*, vol. 1, no. 6, pp. 283–291, Oct. 2011.

[10] M. Aburrous, M. Hossain, F. Thabatah, and K. Dahal, "Intelligent phishing website detection system using fuzzy techniques," in *Proc. 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. [Online]. Available: http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf, 2012

[12] Y. Norazah, B. A. Nor, S. O. Mohd, and C. N. Yeap, "A concise fuzzy rule base to reason student performance based on roughfuzzy approach," in *Fuzzy Inference System-Theory and Application*, InTech, 2010.

[13] R. Ata and Y. Kocyigit, "An adaptive neuro-fuzzy inference system approach for prediction of tip speed ratio in wind turbines," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5454–5460, 2010.

[14] J. S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[15] K. S. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*, New York: JohnWiley & Sons, 1999.

[16] Wikipedia. (2014). [Online]. Available: http://en.wikipedia.org/wiki/Uniformresourcelocator

[17] Levenshtein. (2014). [Online]. Available: http://en.wikipedia.org/wiki/Levenshtein_distance

[18] G. Inc. (2014). [Online]. Available: http://toolbarqueries.google.com

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

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