BINARY LOGISTIC REGRESSION UNTUK MENDETEKSI WEBSITE PHISING MENGGUNAKAN CORRELATION-BASED FEATURE SELECTION

Bekti Maryuni Susanto
Jurusan Teknologi Informasi, Politeknik Negeri Jember
email: bekti.polije@gmail.com

ABSTRACT

Internet provides the facility to find customers worldwide without limitation use e-commerce market is effective. As a result the number of customers that rely on the Internet in the purchase has increased dramatically. In the field of computer security, phishing is a criminal activity that is trying to get sensitive information illegally. Sensitive information could include usernames, passwords and credit card details. This study aims to select the features or attributes in order to obtain the most influential attributes in detecting phishing websites. Selection feature using the correlation-based feature selection. Some of the most important attributes will be selected using the CFS and is applied to the binary logistic regression algorithms. Based on the research results show that CFS is able to eliminate redundant attributes. The subset of attributes generated have this level of accuracy is not much different from the full attributes. This level of accuracy before the selection of attributes 93.99% and 93.20% after the selection attribute. Subsequent studies applying other methods of feature selection and compared the results with the study.

Keywords: *Binary Logistic Regression, Website Phising, Correlation-based Feature Selection*

I. PENDAHULUAN

Internet memberikan fasilitas untuk mencapai pelanggan di seluruh dunia tanpa batasan pasar menggunakan e-commerce yang efektif. Sebagai dampaknya jumlah pelanggan yang bergantung pada Internet dalam pembelian mengalami peningkatan secara dramatis. Ratusan juta dolar ditransfer melalui Internet setiap harinya. Peningkatan ini membuat penipu tergoda untuk melancarkan operasi penipuan melalui Internet. Menurut Aaron dan Manning dalam Mohammad, McCluskey, & Thabtah (2013) Phishing adalah bentuk ancaman web yang didefinisikan sebagai seni meniru website suatu perusahaan otentik bertujuan untuk memperoleh informasi pribadi.

Pada bidang keamanan computer, phising adalah aktivitas criminal yang berusaha untuk mendapatkan informasi sensitive secara tidak sah. Informasi sensitive tersebut bisa berupa username, password, dan detil kartu kredit. Phising dilakukan dengan menyamar menjadi entitas yang bisa dipercaya dalam komunikasi electronic (Dhanalakshmi, Prabhu, & Chellapan, 2011). Website phising secara luas melancarkan serangan social engineering untuk menipu orang pada informasi pribadi termauk nomor kartu kredit, informasi akun bank, nomor pin dan identitas pribadi untuk digunakan untuk menyerang mereka.

Berbagai penelitian telah dilakukan dalam mendeteksi website phising. Diantaranya Predicting Phishing Websites using Neural Network trained with Back-Propagation (Mohammad, McCluskey, & Thabtah, 2013) menunjukkan bahwa neural network merupakan teknik yang baik dalam mendeteksi website phising. Hasil terbaik dicapai saat hidden layer 2 dan learning rate 0,7 dengan MSE sebesar 0,022. Penelitian lain yang berjudul Intelligent Rule based Phishing Websites Classification (Mohammad, McCluskey, & Thabtah, Intelligent Rule based Phishing Websites Classification, 2014) menunjukkan bahwa tingkat akurasi

algoritma C4.5 dalam mendeteksi website phising mengungguli algoritma RIPPER, PRISM dan CBA. Namun demikian setelah dilakukan pemilihan atribut CBA memiliki tingkat error yang paling rendah yaitu 4,75%. Penelitian lain yang berjudul Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code (Alkhozae & Batarfi, 2011) menunjukkan bahwa website phising dapat ditentikan tingkat keamananya

dengan mengekstrak karakteristik phising melalui standard W3C. Hasilnya website phising memiliki tingkat kemanan yang rendah dibandingkan dengan website legitimate.

Penelitian ini bertujuan untuk menyeleksi feature atau atribut sehingga diperoleh atribut yang paling berpengaruh dalm mendeteksi website phising. Pemilihan feature menggunakan metode Correlation-based feature selection (Hall, 1999). Beberapa atribut terpenting akan dipilih menggunakan metode CFS dan diterapkan ke dalam algoritma binary logistic regression.

Manfaat dari penelitian ini adalah untuk meminimalkan waktu komputasi yang dibutuhkan dalam mendeteksi website phising. Waktu komputasi yang semakin kecil akan mengurangi jumlah sumber daya computer yang digunakan. Sehingga konsumsi energi listrik akan terkurangi.

II. METODE PENELITIAN

Pelaksanaan penelitian ini dilakukan dengan kegiatan sebagai berikut.

a. Studi Literatur
   Mempelajari definisi/istilah pada web phising, mempelajari literatur tentang identifikasi web phising.

b. Persiapan data
   Pada penelitian ini menggunakan dataset web phising yang didownload dari UCI Machine Learning Repository.

c. Seleksi atribut
   Dari dataset yang ada dilakukan seleksi atribut menggunakan metode correlation based feature selection.

d. Evaluasi
   Pada tahap ini, dataset yang sudah direduksi atributnya diterapkan pada algoritma binary logistic regression. Dataset dibagi menjadi dua data training dan data testing. Pembagian dataset ini menggunakan 10 X-Validation. Selanjutnya dihitung nilai akurasi berdasarkan pada eksperimen yang dilakukan.

III. HASIL DAN PEMBAHASAN

Hasil yang dicapai sesuai dengan metode yang telah dijelaskan, adalah sebagai berikut.

1. Studi Literatur

Phising adalah sebuah tindakan kriminal untuk mencuri informasi pribadi orang lain menggunakan entitas electronic, salah satunya adalah website. Sebuah website dikategorikan menjadi website phising apabila memenuhi karakteristik phising. Karakteristik phising tersebut digolongkan menjadi empat golongan utama yaitu, Address Bar based Feature, Abnormal based Feature, HTML and JavaScript based Feature dan Domain based Feature (Mohammad, McCluskey, & Thabtah, An Assessment of Features Related to Phishing Websites using an Automated Technique, 2012).

Pada Address Bar based Feature terdapat 12 feature, yang akan dijelaskan sebagai berikut:

a. Using the IP Address
   Jika sebuah IP address digunakan sebagai alternative nama domaian di dalam URL, seperti "http://125.98.3.123/fake.html", hal ini mengindikasikan seseorang berusaha untuk mencuri informasi pribadinya.

Rule:
IF
{(If The Domain Part has an IP Address → Phishing@Otherwise→ Legitimate)

b. Long URL to Hide the Suspicious Part

Rule:
IF

$$\begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases}$$

c. Using URL Shorthening Services "Tiny URL"

Rule:

$$\text{IF} \begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

d. URL's having "@" Symbol

$$\text{Rule: IF} \begin{cases} \text{Url Having @ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

e. Redirecting Using "//"

Rule: IF

{ThePosition of the Last Occurrence of "//" in the URL > 7 → Phishing
Otherwise → Legitimate

f. Adding Prefix or Suffix Separated by (-) to the domain

Rule: IF

{Domain Name Part Includes (−) Symbol → Phishing
Otherwise → Legitimate

g. Sub Domain dan Multi Sub Domain

Rule: IF

{Dots In Domain Part = 1 → Legitimate
Dots In Domain Part = 2 → Suspicious
Otherwise → Phishing

h. HTTPS

Rule:

IF

{Use https and Issuer Is Trusted *and Age of Certificate* ≥ 1 Years → Legitimate
Using https and Issuer Is Not Trusted → Suspicious
Otherwise → Phishing

i. Domain Registration Length

Rule:

IF{Domains Expires on ≤ 1 years → Phishing
Otherwise → Legitimate

j. Favicon

Rule:

IF

{Favicon Loaded From External Domain → Phishing
Otherwise → Legitimate

k. Using Non-Standard Port

Rule:

IF

{Port # is of the Preffered Status → Phishing
Otherwise → Legitimate

l. The Existence of "HTTPS" Token in the Domain Part of the URL

Rule:

IF

{Using HTTP Token in Domain Part of The URL → Phishing
Otherwise → Legitimate

Pada Abnormal based Fetaure terdapat enam feature, yaitu:

a. Request URL

Rule: IF

{% of Request URL < 22% → *Legitimate*
%of Request URL ≥ 22% and 61% → Suspicious
Otherwise → feature = Phishing

b. URL of Anchor

Rule:

IF

{% of URL Of Anchor < 31% → *Legitimate*
% of URL Of Anchor ≥ 31% And ≤ 67% → Suspiciqus
Otherwise → Phishing

c. Links in <Meta>, <Script> and <Link> tags

Rule: IF{(% of Links in "<Meta>","<Script>" and "<"Link>\""<17% → Legitimate
IF(% of Links in <Meta>","<Script>" and "<"Link>\"" ≥17% And≤81% → Suspicious)Otherwise→ Phishing)⊣

d. Server Form Handler (SFH)

Rule:

IF

{SFH is "about: blank" Or Is Empty → Phishing
SFH Refers To A Different Domain → Suspicious
Otherwise → Legitimate

e. Submitting Information to Email

Rule: IF{(Using ""mail()\" or \"mailto:\" Function to Submit User Information" → Phishing@Otherwise → Legitimate)⊣

f. Abnormal URL

Rule: IF

{The Host Name Is Not Included In URL → Phishing
Otherwise → Legitimate

Pada HTML and JavaScript based Feature terdapat lima feature, yaitu:

a. Website Forwarding

Rule: IF

{ofRedirect Page ≤ 1 → Legitimate
of Redirect Page ≥ 2 *And* < 4 → *Suspicious*
Otherwise → Phishing

b. Status Bar Cuztomization

Rule:

IF

{onMouseOver Changes Status Bar → Phishing
It Does't Change Status Bar → Legitimate

c. Dissabling Right Click

Rule:

IF{Right Click Disabled → Phishing
Otherwise → Legitimate

d. Using Pop-up Window

Rule: IF

{Popoup Window Contains Text Fields → Phishing
Otherwise → Legitimate

e. IFrame Redirection

Rule: IF {Using iframe → Phishing; Otherwise → Legitimate}

Pada Domain based Feature terdapat tujuh feature, yaitu;

a. Age of Domain

Rule: IF
{Age Of Domain ≥ 6 months → Legitimate
Otherwise → Phishing

b. DNS Record

Rule:
IF{no DNS Record For The Domain → Phishing
Otherwise → Legitimate

c. Website Traffic

Rule:

IF{Website Rank < 100,000 → *Legitimate*
Website Rank > 100,000 → *Suspicious*
Otherwise → Phish

d. Page Rank

Rule: IF{PageRank < 0.2 → *Phishing*
Otherwise → Legitimate

e. Google Index

Rule:
IF{Webpage Indexed by Google → Legitimate
Otherwise → Phishing

f. Number of Links Pointing to Page

Rule: IF{(Of Link Pointing to The Webpage=0 → Phishing
Of Link Pointing to The Webpage>0 and≤2→Suspicious
Otherwise → Legitimate)

g. Statistical-Reports based Feature

Rule: IF{(Host Belongs to Top Phishing IPs or Top Phishing Domains → Phishing
Otherwise → Legitimate)

2. Binary Logistic Regression

Model regresi logistik biner digunakan untuk melihat apakah variabel tak bebas yang berskala dikotomi (Y = 0 dan Y = 1) dipengaruhi oleh variabel bebas baik yang kategorik maupun numerik. Bentuk umum model peluang regresi logistik dengan k variabel diformulasikan sebagai berikut :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_i + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_i + \ldots + \beta_k x_k)}$$

Dimana $\pi(x)$ adalah peluang sukses probabilitas suatu peristiwa/kasus yang ditentukan oleh y = 1, β_i adalah nilai parameter.

Fungsi tersebut merupakan fungsi linier sehingga perlu dilakukan transformasi ke dalam bentuk logit agar dapat dilihat hubungan antar variabel respon dengan penjelas. Dengan melakukan transformasi logit dari phy (x) , didapat persamaan yang lebih sederhana yang merupakan fungsi linier data parameter-parameternya, yaitu:

$$g(x) = ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$$

Dimana $\{\pi(x) / (1- \pi(x)\}$ merupakan resiko dari y = 1 untuk x tertentu.

Apabila terdapat sebanyak p peubah bebas dan peubah ke-j merupakan merupakan peubah kategorik, maka akan terdapat peubah boneka sebanyak k-1, dengan dummy variabel k_j dinamakan D_{ju} dengan koefisien B_{ju}, u = 1,2,....., k_{j-1}. Sehingga model transformasi logit dapat dituliskan seperti persamaan berikut ini:

$$g(x) = \beta_0 + \beta_1 + \ldots + \sum_{u=1}^{kj-1} \beta_{ju} D_{ju} + \beta_k x_k$$

3. Correlation-based Feature Selection

Banyak faktor yang menentukan kesuksesan machine learning pada suatu tugas tertentu. Faktor yang paling menentukan adalah kualitas dan representasi dari data example. Secara

teori, memiliki lebih banyak atribut atau feature seharusnya menghasilkan kekuatan yang membedakan. Akan tetapi, pengalaman praktis dengan machine learning tidak semua kasus menunjukkan demikian. Banyak algoritma learning dapat dipandang sebagai penciptaan estimasi probabilitas label kelas yang diberikan seperangkat feature. Data ini kompleks dan mempunyai distribusi dimensi yang tinggi. Sayangnya, algoritma induksi hanya bisa diterapkan pada data yang terbatas. Hal ini membuat estimasi banyak parameter probabilitas menjadi sukit dilakukan(Hall, 1999). Feature selection atau seleksi atribut adalah proses mengidentifikasi dan menghapus informasi yang tidak relevan dan redundan sebanyak mungkin(Hall,1999).

Pengurangan dimensi data ini memungkinkan algoritma machine learning untuk bekerja lebih cepat dan lebih efektif. Pada beberapa kasus akurasi klasifikasi dapat ditingkatkan; lainya hasilnya lebih sederhana dan mudah dipelajari dan diintrepretasikan.

Algoritma feature selection menampilkan pencarian melalui seperangkat subset feature dan sebagai konsekuensinya, harus mengarah pada empat kriteria dasar pencarian (Langley, 1994):

1. Starting Point atau titik awal. Pemilihan titik awal untuk pencarian seperangkat subset akan mempengaruhi arah pencarian. Salah satu pilihan dengan memulai nol feature dan secara berurutan menambahkan atribut. Pada kasus ini, pencarian dikatakan bergerak maju di dalam ruang pencarian. Sebaliknya, pencarian dimulai dengan semua feature kemudian secara berurutan mengurangi feature sampai nol, ini dikatakan pencarian bergerak mundur. Alternatif lain dengan mencari dari titik tengah kemudian bergerak keluar.
2. Search Organization. Pencarian subset yang mendalam menjadi penghalang pencarian semua atribut. Misal terdapat N atribut maka ada 2N kemungkinan subset. Strategi heuristik lebih mungkin dari pada pencarian yang mendalam dan dapat memberikan hasil yang bagus, walaupun tidak menjamin menemukan subset yang optimal.
3. Evaluation strategy. Bagaimana seperangkat feature dievaluasi adalah faktor yang paling membedakan diantara algoritma seleksi atribut untuk machine learning. Salah satu paradigma disebut filter, yang beroperasi secara independen dari algoritma machine learning apapun. Pada metode filter ini feature yang tidak diinginkan dikeluarkan dari data sebelum dilakukan pembelajaran. Pendekatan lain adalah sebuah algoritma induksi tertentu diterapkan untuk memilih atribut, metode ini disebut wrapper.
4. Stoping criterion. Sebuah pemilih atribut harus memutuskan kapan untuk menghentikan pencarian pada sepernagkat feature. Tergatung dari strategi evaluasi yang digunakan, pemilih atribut bisa menghentikan atau menambahkan atribut ketika tidak ada lagi atribut alternatif yang meningkatkan merit subset feature saat ini.

Correlation-based feature selection yang selanjutnya disebut seleksi atribut berbasis korelasi atau CFS adalah sebuah algoritma filter sederhana yang meranking subset berdasarkan fungsi evaluasi heuristik berbasis korelasi (Hall, 1999). Berdasarkan hipotesis bahwa subset atribut yang bagus berisi atribut yang mempunyai korelasi tinggi terhadap kelas dan tidak saling berkorlasi satu sama lain. Korelasi yang tinggi satu sama lain atribut menandakan atribut tersebut redundan. Atribut yang berkorelasi rendah terhadap kelas adalah atribut yag tidak relevan. Atribut yang tidak relavan dan redundan harus dihapus. Rumus untuk pencarian subset atribut berdasarkan korelasi adalah (Hall, 1999)

$$r_{zc} = \frac{kr_{zi}}{\sqrt{k+k(k-1)rii}}$$

4. Hasil dan Pembahasan

TABEL I
HASIL PENGUJIAN PADA ALGORITMA LOGISTIC REGRESSION

	Sebelum Seleksi Atribut	Sesudah Seleksi Atribut
Correctly Classified Instances (%)	93,99	93,20

Incorrectly Classified Instances (%)	6,01	6,80
Kappa statistic	0,88	0,86
Mean absolute error	0,09	0,1
Root mean squared error	0,21	0,22
Relative absolute error (%)	17,50	19,34
Root relative squared error (%)	42,38	44,36
Total Number of Instances	11055.00	11055.00

TABEL 2. ATRIBUT SETELAH FEATURE REDUCTION

Prefix_Sufix
having_Sub_Domain
SSLfinal_State
Request_URL
URL_of_Anchor
Links_in_tags
SFH
web_traffic
Google_Index
Result

Berdasarkan hasil eksperimen menunjukkan bahwa setelah dilakukan seleksi atribut menggunakan metode correlation-based feature selection, terjadi penurunan jumlah true positif dan true negatif. Selain itu juga terjadi peningkatan false positif dan false negatif. Kedua hal ini yang mempengaruhi penurunan tingkat akurasi.

Jumlah atribut sebelum dilakukan seleksi atribut adalah 31, setelah dilakukan seleksi atribut jumlah atribut jauh berkurang menjadi 9 atribut. Walaupun terjadi penurunan jumlah atribut yang cukup signifikan namun berdasarkan hasil eksperimen menunjukkan bahwa tingkat akurasi tidak jauh berbeda, hanya selisih 0,8%. Hal ini menunjukkan bahwa seleksi atribut menggunakan metode correlation-based feature selection mampu menghilangkan atribut yang tidak relevan serta redundan.

IV. KESIMPULAN DAN SARAN

Berdasarkan uraian pada pada bagian-bagian sebelumnya dapat disimpulkan bahwa CFS mampu menghilangkan atribut redundan. Subset atribut yang dihasilkan mempunyai tingkat akurasi yang tidak jauh berbeda dengan atribut lengkap. Tingkat akurasi sebelum seleksi atribut 93,99% dan akurasi setelah seleksi atribut 93,20%.

Penelitian selanjutnya menerapkan metode feature selection lainya dan dibandingkan hasilnya dengan penelitian ini.

V. DAFTAR PUSTAKA

[1] Alkhozae, M., & Batarfi, O. (2011). Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code . *International Journal of Information and Communication Technology Research* , 283-291.

[2] Dhanalakshmi, R., Prabhu, C., & Chellapan, C. (2011). Detection Of Phishing Websites And Secure Transactions. *International Journal Communication & Network Security (IJCNS)*, 15-21.

[3] Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning.* Hamilton: Thesis University of Waikato.

[4] Langley, P. (1994). Selection of Relevant Feature in Machine Learning. *AAAI Symposium on Relevance .* Los Angeles: AAAI.

[5] Mohammad, R., McCluskey, T., & Thabtah, F. A. (2012). An Assessment of Features Related to Phishing Websites using an Automated Technique. *International Conferece For Internet Technology And Secured Transactions.* (ss. 492-497). London: ICITST 2012

[6] Mohammad, R., McCluskey, T., & Thabtah, F. A. (2013). Predicting Phishing Websites using Neural Network trained with Back-Propagation. *Proceedings of the 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing. WORLDCOMP 2013* (ss. 682-686). Las Vegas: World Congress in Computer Science, Computer Engineering, and Applied Computing.

[7] Mohammad, R., McCluskey, T., & Thabtah, F. A. (2014). Intelligent Rule based Phishing Websites Classification. *IET Information Security*, 153-160.