PROSIDING SEMINAR NASIONAL STATISTIKA
UNIVERSITAS DIPONEGORO 2013
ISBN: 978-602-14387-0-1

PLUG-IN CLASSIFIER DENGAN BAYESIAN STATISTICS UNTUK MENDETEKSI SITUS WEB PALSU

Anisah, Sapto W. Indratno
Jurusan Matematika FMIPA ITB

Abstrak

Meningkatnya penipuan melalui situs web palsu mendorong orang untuk menciptakan sistem pendeteksi situs web palsu. Melalui *Statistical Learning Theory*, penulis mengajukan sistem pendeteksi situs web palsu yakni metode *Plug-in Classifier* dengan *Bayesian Statistics*. Pada penelitian ini, penulis menggunakan training data yakni petunjuk kecurangan yang berupa internal *link*, level halaman web, dan *screenshoot* dari *header* suatu halaman web. Metode ini diaplikasikan untuk mendeteksi beberapa situs web. Simulasi digunakan untuk menunjukkan performa dari metode *Plug-in Classifier* ini.

Kata Kunci: *Statistical Learning Theory*, Klasifikasi, *Bayesian Statistics*, Situs Web Palsu.

1. Pendahuluan

Perkembangan teknologi membuat orang menggunakan internet sebagai salah satu media untuk berkomunikasi dengan orang lain. Melalui internet, transaksi jual beli pun dapat dilakukan sehingga keamanan berinternet merupakan salah satu faktor terpenting dalam menjalankan usaha maupun bisnis. Semakin menjamurnya aktifitas atau bisnis *online* maka semakin meningkat pula resiko penipuan yang terjadi dalam dunia maya. Pada tahun 2012, *id.cert* yang merupakan sebuah lembaga penelitian tentang kejahatan internet di Indonesia melaporkan bahwa peringkat pertama insiden internet adalah *Network Insident*, disusul dengan pelanggaran HaKI (Hak atas Kekayaan Intelektual), *malware*, *spam*, *spam komplain*, dan *spoofing/phishing*. Berbagai cara digunakan para pelaku kecurangan seperti mengirimkan email spam atau bahkan membuat situs web palsu. Ketidaktahuan pengguna internet akan situs web asli atau palsu dimanfaatkan oleh pelaku kecurangan untuk melakukan penipuan. Oleh karena itu, pada penulisan kali ini akan diusulkan salah satu model sistem deteksi situs web palsu yakni dengan menggunakan *Statistical Learning Theory* (SLT).

Sistem pendeteksi situs web palsu menggunakan mekanisme klasifikasi untuk mendeteksi situs web palsu. Klasifikasi adalah langkah awal untuk menemukan hubungan dari sekumpulan data berdasarkan karakteristik atau pola tertentu. Pemetaan $f: X \rightarrow Y$ dinamakan sebuah pengklasifikasi (*classifier*) (Luxburg and Scholkof, 2008).

ISBN: 978-602-14387-0-1

Klasifikasi terdiri dari klasifikasi biner dan multi-klasifikasi. Klasifikasi biner mengklasifikasikan objek kedalam label $Y = \{1, -1\}$ dari input X, sedangkan multi-klasifikasi merupakan perkembangan dari klasifikasi biner dimana objek diklasifikasikan kedalam lebih dari dua kelas.

Penulis mengajukan salah satu metode yang dipakai dalam klasifikasi yakni metode *Plug-in Classifier*. Dalam metode ini akan digunakan *Bayesian Statistics* dan fungsi logistik untuk membuat *decision rule*. Situs web yang akan dipakai adalah situs web yang sudah terlabel (diketahui asli dan palsu). Penelitian yang dilakukan adalah dengan mengambil 5 (lima) halaman web pada setiap satu situs web. Dari kelima halaman yang diperoleh, akan dibuat suatu aturan bahwa situs web yang asli akan memenuhi kriteria yang diajukan dan situs yang palsu tidak akan memenuhi kriteria tersebut.

2. Model Klasifikasi

Plug-in classifier adalah salah satu cara untuk mengkontruksi sebuah *classifier* menggunakan *training data* kemudian meletakkannya kedalam *Bayes Classifier*. Estimator ini berbentuk,

$$\hat{\eta}_n(x) = \eta\left(x; \{X_i, Y_i\}_{i=1,\dots,n}\right)$$

Definisi 1. [*Bayes Classifier*]

$$f^*(x) = \begin{cases} 1, & \eta(x) \geq \frac{1}{2} \\ -1, & \text{lainnya} \end{cases}$$

dimana $\eta(x) = E[Y|X = x]$

$$= P(Y = 1|X = x) - P(Y = -1|X = x)$$

Karena distribusi peluang dari *training data* tidak diketahui maka klasifikasi *Bayes* tidak dapat dilakukan. Oleh karena itu, akan digunakan pendekatan *Bayesian statistics* sehingga *decision rule* dapat ditemukan.

Pada *Bayesian Statistics* terdapat pdf *prior* dan pdf *posterior* dimana untuk menentukan pdf *prior* terdapat subjektifitas berdasarkan *prior knowledge* atau pengalaman peneliti. Sedangkan pdf posterior dapat ditulis sebagai berikut (Hogg, 2013),

$$k(\theta, y) \propto L(y|\theta).h(\theta)$$

ISBN: 978-602-14387-0-1

dengan $L(y|\theta)$ adalah pdf bersyarat bersama dari vektor acak y jika diberikan $\Theta = \theta$ dan $h(\theta)$ adalah pdf *prior*.

Untuk menemukan *decision rule* diperlukan *training data*. *Training data* berupa $\{x_i, y_i\}_{i=1,\dots,n}$ dimana $x_i \in R^m$ merupakan input yakni petunjuk kecurangan yang terdiri dari m buah. Pada penulisan ini, digunakan tiga buah petunjuk kecurangan yakni internal *link* (x_{i1}), level (x_{i2}), dan *screenshoot* (x_{i3}) sehingga $x_i \in R^3$ dan dapat ditulis $x_i = (x_{i1}\ x_{i2}\ x_{i3})^T$. Sedangkan y_i adalah output yang berpadanan dengan input x_i. Internal *link* adalah sebuah *hyperlink* yang merupakan elemen navigasi dalam sebuah halaman web ke halaman web yang lain dalam situs web yang sama atau domain internet yang sama. Situs web palsu cenderung mempunyai halaman yang sedikit dan akibatnya sedikit pula *link* diantara halaman-halamannya (Abbasi, 2010). Oleh karena itu,

$$x_{i1} = \frac{\sum h(k,l)}{P_2^5}$$

dimana $(k,l) \in \{(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5)\}$

$$h(k,l) = \begin{cases} 2, & \text{jika link halaman } k \text{ ada pada halaman } l \text{ dan sebaliknya} \\ 1, & \text{jika link halaman } k \text{ atau } l \text{ ada pada halaman } l \text{ atau } k \text{ dan tidak sebaliknya} \\ 0, & \text{jika lainnya} \end{cases}$$

Petunjuk kecurangan yang kedua yakni level halaman web dapat dilihat dari jumlah garis miring "/" yang terdapat pada URL *link* tersebut, selain itu situs web asli mempunyai ratusan halaman web, merentang 4-5 level (Abbasi dan Chen, 2009). Karena setiap halaman web mempunyai puluhan bahkan ratusan internal *link* dengan jumlah garis miring "/" yang berbeda, maka

$$x_{i2} = \frac{\sum_{j=1}^{5} E[C_j]}{5}$$

dimana $E[C_j] = \sum_{r=0}^{n} r.\frac{A_{r,j}}{Total\ A_{r,j}}$

$A_{r,j} = \#$ level r pada halaman ke $-\ j$

Selanjutnya digunakan fungsi logistik agar nilai x_{i2} berada diantara 0-1.

Selain internal *link* dan level dilakukan pula pembandingan visual dari kelima halaman web dalam satu situs web. Dengan bantuan *add-on* dari peramban *mozilla firefox* diperoleh *screenshoot* dari masing-masing halaman web yang kemudian

ISBN: 978-602-14387-0-1

dipisahkan untuk masing-masing *layer* dan "dipotong" dengan dimensi matriks yang sama (150,1000) untuk mendapatkan bagian *header*. Selanjutnya matriks yang diperoleh dibentuk menjadi vektor, katakanlah v_1, v_2, v_3, v_4, v_5. Maka untuk $i \neq j$, dengan $i, j = 1, \dots, 5$ diperoleh

$$A_{v_i,v_j} = \text{korelasi}\,(v_i, v_j) = \text{korelasi}\,(v_j, v_i)$$

$$\overline{A_{v_i,v_j}} = \frac{\sum A_{v_i,v_j}}{C_2^5}$$

dimana A adalah *layer* yakni *Red*, *Green* dan *Blue*. Sehingga

$$x_{i3} = \frac{\overline{A_{v_i,v_j}}(Red) + \overline{A_{v_i,v_j}}\,(Green) + \overline{A_{v_i,v_j}}(Blue)}{3}$$

Masalah kali ini adalah menentukan distribusi dari bobot $w \in R^m$. Karena menggunakan tiga buah petunjuk kecurangan maka $w \in R^3$. Jika diasumsikan distribusi prior dari w adalah $N_3(0, \Gamma)$, maka distribusi posterior dari w dapat ditentukan dengan hubungan

$$\begin{aligned} P(w|y, X) &\propto L(Y|w, X).P(w) \\ &= P(Y = y|w, X)\,.P(w) \end{aligned}$$

Model yang digunakan adalah model linier yang berbentuk

$$f(x) = w^T x$$

dengan output berbentuk

$$y_i = f(x_i) + \varepsilon_i\,, i = 1, 2, \dots, n$$

dimana ε_i adalah error yang memiliki mean 0 dan variansi konstan σ^2. Dengan mengasumsikan $\varepsilon_i \sim N(0, \sigma^2)$ diperoleh

$$\begin{aligned} P(Y = y|w, X) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|w, X) \\ &= P(\{Y_i - f(x_i) = y_i - f(x_i)\}_{i=1}^n|w, X) \\ &= P(\{\varepsilon_i = y_i - f(x_i)\}_{i=1}^n|w, X) \\ &= \prod_{i=1}^n P(\varepsilon_i = y_i - f(x_i)|w, X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y - Xw)^T \Sigma^{-1}(y - Xw)\right\} \sim N(Xw, \Sigma) \end{aligned}$$

dimana $\Sigma = \sigma^2.I_n$ dengan I_n matriks identitas berukuran $n \times n$

ISBN: 978-602-14387-0-1

dan $X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$

Sehingga

$$P(w|y,X) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y - Xw)^T \Sigma^{-1}(y - Xw)\right\} \frac{1}{(2\pi)^{3/2}|\gamma|^{1/2}} \exp\left\{-\frac{1}{2} w^T \Gamma^{-1} w\right\}$$

$$= C.\exp\left\{-\tfrac{1}{2}\left[y^T(\sigma^2 I_n)^{-1}y - y^T(\sigma^2 I_n)^{-1}Xw - w^T X^T(\sigma^2 I_n)^{-1}y + wTXT\sigma 2In - 1X + \Gamma - 1w\right.\right.$$

$$= C_1.\exp\left\{-\tfrac{1}{2}(w-\mu)^T A(w-\mu)\right\}$$

dimana $A = \left(X^T(\sigma^2 I_n)^{-1}X + \Gamma^{-1}\right)$ dan

$\mu = A^{-1}X^T y$

Jadi $P(w|y,X) \sim N(\mu, A^{-1})$.

Karena berdistribusi normal maka

$w_{max} = \mu = arg\max_w P(w|y,X)$.

Oleh karena itu, dengan fungsi logistik akan diperoleh peluang data baru, yakni

$$\hat{\eta}(x) = \frac{1}{1 + \exp\{-({w_{max}}^T x)\}},$$

yang nantinya kita sebut sebagai $\hat{\eta}_n(x) = \eta\left(x; \{x_i, y_i\}_{i=1,\dots,n}\right)$.

Maka *decision rule* yang diharapkan dapat diperoleh yakni berbentuk

$$\hat{f}(x) = \begin{cases} 1, & \hat{\eta}(x) \geq \frac{1}{2} \\ -1, & \text{lainnya} \end{cases}$$

Dalam penulisan kali ini σ diasumsikan berada diantara selang [0,1 ; 1] sehingga digunakan $\sigma = 0{,}5$ untuk menghitung w_{max}. Selain itu, kami juga menggunakan σ dengan algoritma sebagai berikut yang selanjutnya kami sebut sebagai $\sigma_{observasi}$,

1. Berikan n training data $\{x_i, y_i\}_{i=1,\dots,n}$ dengan $x_i \in R^m$ yang diklasifikasikan dalam skalar y_i, dimana $y_i = 1$ jika x_i memenuhi karakteristik tertentu dan $y_i = -1$ jika x_i tidak memenuhi karakteristik tersebut,

ISBN: 978-602-14387-0-1

2. Buat matriks $X \in R^{n \times m}$ dimana kolom-kolom X adalah $x_j, j = 1, \ldots, m$ dan matriks identitas Γ^{-1} dengan dimensi bersesuain dengan *m* petunjuk kecurangan yang digunakan,
3. Berikan $\sigma_{old} = 0{,}5\,, \epsilon, k = 1$ dan $\Delta f = \epsilon + 1$. Ketika $\Delta f \geq \epsilon$, maka lakukan perhitungan berikut:
   - Hitung error yakni $\varepsilon_i = y_i - w^T . x_i$
   - Hitung $\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}{n-1}}$
   - Update $\Delta f = |\sigma_{old} - \sigma_{new}|, \sigma_{old} = \sigma_{new}$
4. Misalkan $\sigma_{observasi}$ = arg $(\Delta f < \epsilon)$, hitung w_{max}
5. Hitung $\hat{\eta}(x) = \frac{1}{1+\exp\{-(w_{max}{}^T . x)\}}$, dimana *x* adalah data baru yang akan dites,
6. Buat *decision rule* yakni $\hat{f}(x) = \begin{cases} 1, & \hat{\eta}(x) \geq \frac{1}{2} \\ -1, & \text{lainnya} \end{cases}$

   sehingga diperoleh hasil prediksi.

3. Hasil Dan Simulasi

Sebuah situs web mengandung banyak halaman dengan masing-masing halaman terdiri dari banyak gambar, beserta teks, *source code*, *URLs* dan atribut yang terstruktur berdasarkan *link* suatu halaman yang satu dengan halaman yang lain [3]. Situs asli yang digunakan diambil dari situs-situs resmi bank umum di Indonesia dan data situs palsu diperoleh dari daftar situs www.phisthank.com. Sebanyak 28 situs web atau 140 halaman web asli dan palsu diunduh dari tanggal 17 Juni sampai 14 Juli 2013. Pada *training data* diperoleh bahwa situs-situs asli mempunyai karakteristik x_{i1}, x_{i2}, x_{i3} yang bernilai diatas 0.5 dan situs-situs palsu mempunyai karakteristik x_{i1}, x_{i2}, x_{i3} yang bernilai dibawah atau sama dengan 0.5. Namun demikian dari keseluruhan data yang diperoleh terdapat 1 data situs palsu katakanlah F_1 yang mempunyai karakteristik menyerupai data-data situs asli dan sebaliknya terdapat 1 data situs asli katakanlah L_1 yang mempunyai karakteristik menyerupai data-data situs palsu. Selanjutnya data-data tersebut kita katakan sebagai data salah tafsir.

Untuk metode *plug-in classifier* dengan $\sigma = 0{,}5$ dilakukan empat simulasi. Simulasi pertama terdiri dari 5 tahap dengan menggunakan *training data* berupa 7 data

ISBN: 978-602-14387-0-1

situs asli dan 7 data situs palsu dan tidak menggunakan data salah tafsir F_1 dan L_1. Tahap pertama dengan 14 *training data* dan $\sigma = 0{,}5$ diperoleh $w_{max} = (1.9411, -14838, 0.4139)$ sehingga ketika ada 3 data baru yang sebenarnya sudah terlabel (sudah diketahui palsu atau asli) dapat dihitung $\hat{\eta}(x)$ nya. Ketiga data baru ini sebenarnya adalah data situs asli dan ketika dihitung menggunakan fungsi logistik diperoleh hasil bahwa data ini termasuk kedalam data situs asli. Tahap kedua adalah dengan menambahkan 3 data asli tersebut kedalam *training data* sehingga *training data* menjadi 17 dan diperoleh w_{max} yang baru. Terdapat 3 data (situs palsu) yang akan dites, maka dengan w_{max} ini diperoleh hasil bahwa 3 data ini masuk kedalam data situs palsu. Kemudian tahap ketiga adalah dengan menambahkan 3 data situs palsu tersebut kedalam *training data* sehingga *training data* sekarang menjadi 20. Hasil ketika ada data baru yang akan dites ternyata masih sesuai sehingga dilakukan tahap keempat yakni menambahkan kembali data tes menjadi *training data* sehingga *training data* berjumlah 23. Ternyata hasil menunjukkan nilai prediksi yang masih sesuai sehingga tahap kelima *training data* menjadi 26. Pada tahap yang terakhir ini, tes dilakukan dengan menggunakan data salah tafsir. Ternyata hasil prediksi data L_1 adalah 0.4387 dan F_1 adalah 0.7555 yang berarti bahwa data tersebut masuk kedalam kategori palsu dan sebaliknya. Ketidaksesuaian ini terjadi karena data F_1 dan L_1 tidak dimasukkan kedalam *training data* awal.

Simulasi kedua dilakukan dengan menambahkan data salah tafsir kedalam *training data* awal sehingga *training data* berjumlah 16 yang terdiri dari 8 data situs asli dan 8 data situs palsu. Hal yang serupa dilakukan pada simulasi kedua ini dan hasil dari seluruh tahap menunjukkan bahwa penambahan data salah tafsir masih memberikan nilai prediksi yang sesuai. Begitu pula untuk simulasi ketiga dengan menambahkan data salah tafsir F_2 dan L_2 yang diperoleh dengan mengenerate data F_1 dan L_1, hasil prediksi masih sesuai. Hal ini tidak terjadi pada simulasi keempat. Pada tahap pertama simulasi keempat dengan penambahan data salah tafsir F_3 kedalam *training data* awal, menunjukkan hasil yang tidak sesuai yakni dua dari tiga tes data menghasilkan nilai prediksi 0.4952 dan 0.4872 yang berarti data diklasifikasikan sebagai situs palsu, padahal tes pada tahap pertama menggunakan data situs asli. Oleh karena itu, simulasi ini adalah simulasi terakhir bagi metode *Plug-in Classifier* dengan $\sigma = 0{,}5$ dan

ISBN: 978-602-14387-0-1

diperoleh bahwa proporsi data salah tafsir yang menjadi *training data* maksimal 28,57% dari data yang benar sehingga hasil prediksi tetap memberikan nilai yang sesuai.

Berikutnya, untuk metode *Plug-in classifier* dengan $\sigma_{observasi}$ terdapat enam simulasi dimana pada simulasi pertama terdapat kesalahan klasifikasi seperti pada simulasi pertama *plug-in classifier* dengan $\sigma_0 = 0{,}5$. Hal ini juga dikarenakan oleh alasan yang sama. Simulasi kedua sampai simulasi kelima masih menunjukkan hasil yang benar. Sedangkan simulasi terakhir ditunjukkan pada simulasi keenam dimana pada tahap pertama tes data, dua dari tiga tes data menghasilkan nilai prediksi 0.4912 dan 0.4855 yang berarti situs yang seharusnya situs asli diklasifikasikan sebagai situs palsu. Jadi diperoleh proporsi data salah tafsir yang menjadi *training data* maksimal 57,14% dari data yang benar.

4. Kesimpulan

Metode *Plug-in Classifier* dengan pemilihan $\sigma = 0.5$ menghasilkan proporsi data salah tafsir lebih kecil dibandingkan dengan $\sigma_{observasi}$ sehingga pemilihan $\sigma_{observasi}$ untuk mendeteksi situs web palsu lebih baik dibandingkan $\sigma = 0.5$ jika dilihat dari variasi data yang dapat dijadikan *training data*.

DAFTAR PUSTAKA

Luxburg, U. dan Scholkopf, B. 2008. *Statistical Learning Theory: Models, Concepts, and Results.*

Hogg, dkk. 2013. *Introduction to Mathematical Statistics*. Pearson Education, Inc. USA.

Abbasi, A. 2010. *Detecting Fake Website: The Contribution of Statistical Learning Theory.*

Abbasi, A dan Chen, H. 2009. *A Comparison of Fraud Cues and Classification Methods for Fake Escrow Website Detection*. Springer Science+Business Media, LLC 2009.