ISBN: 978-602-61268-5-6

Optimasi Algoritma C4.5 Menggunakan Genetic Algoritma Dalam Memprediksi Website Phishing

Aswan Supriyadi Sunge
Sekolah Tinggi Teknologi Pelita Bangsa
e-mail: aswan.sunge@pelitabangsa.ac.id

Abstrak – Salah satu isu terpenting saat ini dalam dunia online yaitu keamanan. Masalah keamanan terbesar salah satunya adalah *Phishing* yang melibatkan duplikat situs yang sah atau asli untuk menipu dengan mencuri informasi pengguna online. Memang diakui sangat sukar untuk membedakan situs asli dengan palsu. Oleh sebab itu dibutuhkan klasifikasi dalam memprediksi website yang terindikasi *Phishing*. Dengan klasifikasi dalam Algoritma C4.5, permasalahan tersebut dapat diselesaikan dengan menghasilkan rule dari pohon keputusan. Untuk dapat meningkatkan akurasi dari prediksi algoritma C4.5 dapat digunakan fitur seleksi dengan menggunakan *algoritma genetika*. Berdasarkan penerapan algorima C4.5 dihasilkan akurasi sebesar 83,25% untuk memprediksi website Phishing dan dengan seleksi fitur menggunakan algoritma genetika meningkatkan akurasi sebesar 3,22% menjadi 86,47. Dari penelitian ini algoritma genetika terbukti dapat meningkatkan akurasi, *precission* dan *recall* untuk prediksi website *phishing*.

Kata Kunci: phishing, prediksi, algoritma C4.5, algoritma genetika

PENDAHULUAN

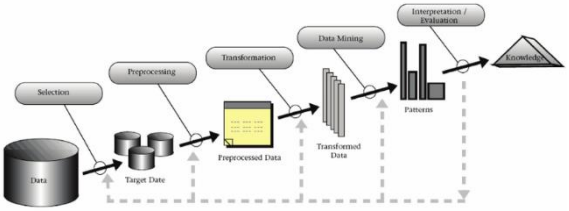
Perkembangan internet begitu signifikan, jika dilihat pertumbuhan internet di dunia sudah lebih dari 4 milyar dan di Indonesia lebih dari 143 juta pengguna (internetworldstats.com/stats,2018) dari total populasi lebih dari 266 juta penduduk. Hal ini disebabkan berkembangan teknologi informasi yang sangat cepat dan berbagai macam media dan fungsi yang salah satunya dalam hal transaksi keuangan maupun *e-commerce*. Hal tersebut memudahkan pelanggan tanpa harus bersusah payah dan tanpa perlu keluar rumah. Tetapi di dalam kemudahan bertransaksi muncul salah satu masalah terbesar yaitu keamanan bertransaksi. Ini menjadi momok menakutkan bagi pengguna online, apalagi sudah merambah dalam pengguna social media (Wibowo, Mia & Fatiman, 2017). Satu hal dari keamanan dari ketidaktahuan dari segi pengguna yang akibatnya terjerumus ke dunia *Cybercrime.* Juga banyak pengguna online tidak bisa membedakan antara situs asli maupun situs palsu atau *Phishing* , maka dari itu penelitian ini bertujuan untuk bisa memprediksi akan situs yang terindakasi *Phishing*

TINJAUAN PUSTAKA

Phishing merupakan metode ataupun cara dalam pengelabui pengguna online dan paling umum dalam serangannya dengan memberikan link atau pesan email ke situs yang tampaknya asli (Junaind, Shafique, Robert, 2016). Teknik pun semakin beragam bukan hanya membuat situs asli atau memberikan link tapi menggunakan mobile (Belal, Amro, 2018) , inilah sebagai celah dalam ketidak tahun pengguna online. Memang diakui metode maupun cara sulit dalam mendeteksi apalagi seorang pengguna yang tidak tahu akan keamanan. Maka dari itu dibutuhkan prediksi dalam mendeteksi terindikasi Phishing, untuk itu dibutuhkan klasifikasi dalam data mining (Mofleh, Al-diabat, 2016) dalam melihat data maupun parameter yang ada yang dijadikan patokan dalam pendekteksian Phishing.

Data mining merupakan asal kata dari mining yang berarti tambang, dikembangkan menjadi konsep dalam melihat informasi maupun pengetahuan, dari data lampau maupun masa lalu yang tersimpan dalam database (Larose, 2005) dan penggunaan data mining digunakan untuk menganalisis suatu prilaku maupun prediksi, juga bukan hanya digunakan dalam ilmu computer saja tetapi bidang lain seperti bisnis maupun industri (Giudici & Figini. 2009)

Istilah data mining maupun Knowledge Discovery in Databases (KDD) tidak lepas dari keduanya dikarenakan menggali data yang tersimpan dalam data yang sangat besar (Fayyad, U.; Piatetsky-Shapiro, G; Smyth, 1996). Skema tersebut digambarkan dibawah ini,



Sumber (Fayyad, U.; Piatetsky-Shapiro, G; Smyth, 1996)

Gambar 1. Proses Skema KDD

Tahapan proses KDD dalam data mining, sebagai berikut :

1. *Data Selection*, data akan diseleksi berdasarkan kecocokan data yang akan diambil keputusannya.
2. *Data Preprocessing/Cleaning* tahap ini

ISBN: 978-602-61268-5-6

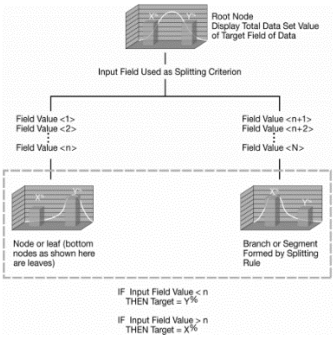
dilakukan pembersihan data yang kosong, penggandaan atau yang tidak sesuai dari hasil yang diputuskan.

3. *Transformation*, melihat data sudah dipilih dan dipresentasikan dari hasil yang diinginkan
4. *Data Mining*, melihat pola yang ingin ditampilkan dari metode maupun teknik yang dipilih sebelumnya misalnya klasifikasi, clustering, regresi dan lain sebagainya.
5. *Interpertation/Evaluation*, pada proses ini sebagai peterjemah dari data yang telah ditampilkan dan melihat hasil dari teknik atau metode yang digunakan.

Algoritma dalam klasifikasi yang banyak digunakan ialah *Decision Tree*. Dikarenakan sangat mudah dimengerti dan dijabarkan oleh banyak pengguna juga mudah dipahami dimana cabang pohon disimpulkan dalam bentuk klasifikasi (Gorunescu.2011).
Pohon keputusan mempunyai tiga pendekatan klasik ;

1. Classification (Klasifikasi), melihat hasil prediksi berdasarkan kelas atau label (misalnya, Ya atau Tidak, Lulus dan TIdak Lulus)
2. Regression (Regresi) melihat hasil prediksi belum tahu akan hasilnya (misalnya : Pemberian Kredit, Pencapain Target Pasar, Hasil Medis)
3. CART (Classification & Regresi Tree) yaitu berdasarkan susunan pertanyaan yang saling berkaitan dan berurutan dan hasil jawaban tersebut menjabarkan pertanyaan berikutnya. Namun jika pertanyaan tidak sesuai maka akan berhenti dan tidak lanjutkan pertanyaan.

Dari setiap pohon keputusan menghasilkan simpul yang merupakan hasil prediksi atau solusi untuk menghasilkan solusi dari pertanyaan yang saling berkaitan (Seemam Rathi, Mamta, 2012)



Sumber (Seemam Rathi, Mamta, 2012)
Gambar 2. Ilustrasi *Decision Tree*

Algoritma dalam *Decision Tree* banyak sekali (Wu, Xindong, 2007) namun yang banyak digunakan yaitu ID3 dan Algoritma C4.5. Kedua mempunyai kesamaan dikarenakan Algoritma C4.5 merupakan pengembangan dari ID3 namun ada perbedaan yang utama yaitu :

- Ketika data(atribut) yang berkelanjutan atau putus-putus terutama berhubungan data traninig maka Algoritma C4.5 dapat memperbaikinya.
- Hasil yang didapat dari Algoritma C4.5 dapat dipangkas ketika terbentuk.
- Penyeleksian variabel dilakukan dengan *Gain Ratio*

Perubahan dari ID3 ke C4.5 dalam Gain Ratio untuk diperbaharui information gain maka dengan rumus :

$$GainRatio(S.A) = \frac{Gain(S.A)}{SplitInfo(S.A)} \text{.....(1)}$$

Dimana;
S = Ruang/Data Sample yang dipergunakan untuk data training
A = Atribut
Gain(S,A) = information gain pada atribut A
SplitInfo(S,A) = split information pada atribut A

Pemilihan atribut dari Gain Ratio yang tertinggi dijadikan sebagai atribut test untuk simpul. Pendekatan ini menerapkan konsep normalisasi pada information gain yang disebut dengan split information dengan rumus dibawah ini :

$$SplitInfo(S.A) = -\sum_{i=1}^{i} \frac{S_i}{S} log_2 \frac{S_i}{S} \text{.....(2)}$$

Dimana:
S = Ruang (data) sample yang digunakan untuk training.
A = Atribut.
Si = Jumlah sample untuk atribut i

Pada tahun 1970 *Algortima Genetica* (GA) diperkenalkan oleh John Holland di Universitas Michigan (J.H. Holland, 1975), bahwa dari bagian masalah merupakan bentuk dari adaptasi dari alam maupun buatan yang dapat diformulakan mejadi bagian genetika(Suryanto, 2007). GA merupakan bagian optimasi dan pencarian yang didasarkan pada seleksi alam dan seleksi makluk hidup secara apa adanya. Pada akhirnya, mengembalikan satu bagian yang terbaik yang dijadikan solusi dari masalah yang akan dipecahkan sebagai kromosom (. Desiani, A., & Muhammad, A, 2006) Ada tiga aspek dalam dalam menggunakan GA :

a) Definisi Fungsi Objektif / Definisi
b) Definisi dan Implementasi Representasi Genetika
c) Definisi dan implementasi dari operator genetik

ISBN: 978-602-61268-5-6

METODOLOGI PENELITIAN

Sampel dalam penelitian ini merupakan data sekunder yang didapat dari hasil komputasi digital pada UCI Neda Abdelhamid Auckland Institute of Studies. Data yang didapat terdiri dari Variabel Rendah (0), Sedang (-1) dan Tinggi (1). Untuk paramaternya terdapat 9 yaitu *SFH, popUpWindow , SSLfinal_State, Request_URL, URL_of_Anchor, web_traffic, URL_Length, age_of_domain, having_IP_Address*. Dari data yang dihasilkan yang dijadikan data training maka akan diperoleh *decision tree* untuk hasil klasifikasi dan data training untuk melihat akurasi dari klasifikasi tersebut. Data yang digunakan untuk melakukan penelitian adalah data primer dan data sekunder. Untuk mengukur tingkat akurasi dari prediksi menggunakan Rapid Miner Studio.

Tahap dalam penelitian ini adalah sebagai berikut:

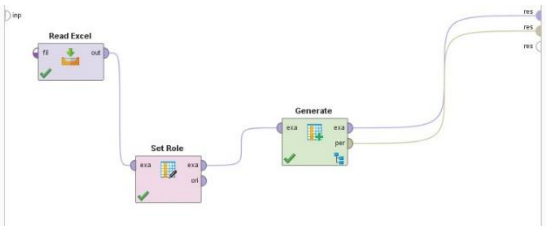
1) Pengumpulan (Pengambilan) data
   Pada tahap ini mencari data yang tersedia, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.
2) Pengolahan data awal
   Ditahap ini dilakukan penyeleksian, pembersihan termasuk melihat data yang kosong kemudian merubah data yang dinginkan.
3) Metode yang diusulkan
   Pada tahap ini penganalisisan data kemudian pengelompokan variabel yang saling berhubungan dengan yang lain, kemudian penerapan model yang sesuai data yang telah dibentuk.
4) Eksperimen dan pengujian metode
   Pada tahap ini penentuan model yang diusulkan ketika akan diuji dan melihat hasil rules yang dijadikan pengambilan keputusan.
5) Evaluasi dan validasi
   Pada tahap ini melakukan hasil evaluasi yang didapat dari model yang ditetapkan sebelum dan melihat hasil akurasi dengan pengujian aplikasi terhadap metode yang digunakan. Dibawah ini gambar skema dalan tahapan penelitian yang dilakukan :

Gambar 3. Kerangka Pemiikiran

HASIL DAN PEMBAHASAN

Penelitian ini menggunkan tools Rapidminer untuk melakukan olah data, berikut adalah tahapan dalam penggunaan Rapid miner:

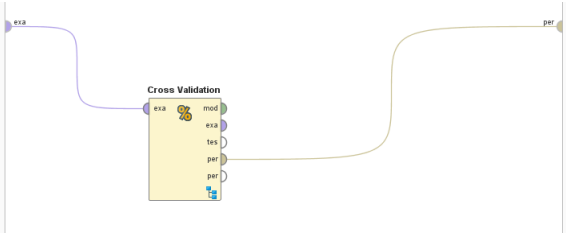
1. Pengujian menggunakan *Decision Tree* dengan fitur seleksi *Algoritma Genetika*



Sumber ; Rapidminer Studio

Gambar 4. Optimize by Generation

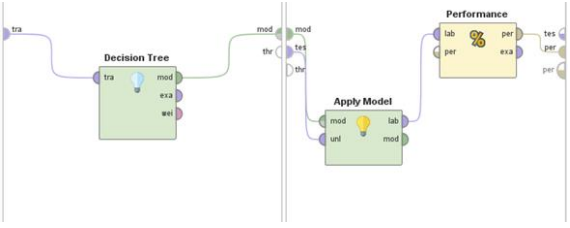
Database *website phishing* dihubungkan dengan *Set Role Attribute* untuk menentukan atribut yang menjadi label dan dalam dataset dijadikan label dari dataset *website phishing* adalah atribut *result*. Selanjutnya Set Role dihubungkan dengan *Feature Generation* yaitu *Optimize by Generation* (Evolutionary) untuk dilakukan pemilihan atribut-atribut terindikasi website phishing. Fitur Optimize by Generation (Evolutionary) terdapat Cross Validation.



Sumber ; Rapidminer Studio

Gambar 5. *Cross-Validation (CV)*

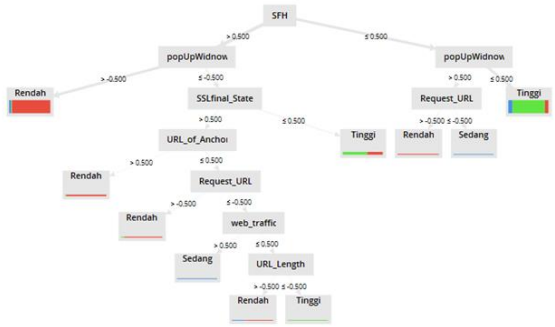
Dalam penggunaan CV dalam prediksi website phishing terdiri 10-fold validation berjumlah 1353 data yang terbagi 20 atribut dipecah menjadi 10 atribut setiap bagian dibagi secara acak. Intinya adalah 1:9, dimana 1 bagian merupakan data testing dan 9 bagian dijadikan data traning, sehingga 10 menjadi bisa dijadikan menjadi data testing setelah itu dilihat tingkat akurasinya. Pada cross validation terdapat tahap dalam penetapan algoritma *decision tree*.



Sumber: Rapidminer Studio

Gambar 6. *Model Decision Tree*

ISBN: 978-602-61268-5-6



Gambar 7. *Decision Tree*

2. Klasifikasi dapat mengevaluasi berdasarkan kriteria lain seperti akurasi, keakuratan, stabilitas maupun interpretabilitas (Vercellis, Carlo, 2009). Maka diuji tingkat hasil akurasi dari penggujian menggunakan *Decision Tree* dengan fitur seleksi *Algoritma Genetika* dilakukan tingkat akurasi *confussion matrix*.

Tabel 1. Hasil Pengujian

	True Sedang	**True Tinggi**	**True Rendah**	**Class Precision**
Pred Sedang	31	2	1	91,18%
Pred Tinggi	46	524	86	79,88%
Pred Rendah	26	22	615	92,76%
Class Recall	30,10%	95,62%	87,61%	
Accuracy	**86,47%**			

Dari tabel diatas dapat diambil kesimpulan bahwa hasil prediksi menggunakan Decision Tree dengan seleksi fitur algoritma genetika tingkat akurasinya adalah 86, 47%. Hasil akurasi ini meningkat dari penelitian sebelumnya dengan menggunakan metode yang sama tetapi tidak menggunakan seleksi fitur sebesar 83,%.

accuracy: 83.25%

	true Sedang	true Tinggi	true Rendah	class precision
pred. Sedang	3	4	0	42.86%
pred. Tinggi	14	153	29	78.06%
pred. Rendah	14	7	182	89.66%
class recall	9.68%	93.29%	86.26%	

Gambar 8. *Accuracy Decision Tree*

accuracy: 86.47% +/- 2.94% (mikro: 86.47%)

	true Sedang	true Tinggi	true Rendah	class precision
pred. Sedang	31	2	1	91.18%
pred. Tinggi	46	524	86	79.88%
pred. Rendah	26	22	615	92.76%
class recall	30.10%	95.62%	87.61%	

Gambar 9. *Accuracy Decision Tree Optimasi GA*

KESIMPULAN

Dari hasil pembahasan diambil kesimpulan sebagai berikut :

1. Algortima C4.5 dengan Optimasi Algoritma Genetika maka indikasi website *Phishing* dapat diprediksi dan dapat dijadikan konstribusi terhadap proses pengambil keputusan ke pengguna online.
2. Evaluasi dalam menguji hasil prediksi dari Decision Tree algoritma C4.5 dengan seleksi fitur algoritma genetika, dah hasil prediksi yang didapatkan dalam pengujian ini adalah 86,47% hasil ini meningkat dari penelitian yang sebelumnya menggunakan data yang sama dan algoritma yang sama yaitu algoritma decision tree hasil prediksinya adalah 83,25%, sehingg dapat disimpulkan bahwa tingkat dengan penggunaan seleksi fitur algoritma genetika mendapatkan hasil yang lebih baik dengan tingkat akurasi yang meningkat.

Berdasarkan hasil penelitian memberikan saran sebagai berikut :

1. Perlu adanya penelitian lebih lanjut dengan melakukan pengujian dengan motede lain maupun dikomparasi seperti SVM, k-NN, Neural Network, Naïve Bayes dan lain-lain agar melihat hasil perbandingan dengan akurasi yang tertinggi dalam prediksi yang terindikasi website *Phishing.*
2. Perlu diterapkan lebih lanjut optimasi menggunakan metode lain seperti Adaboost, atau PSO untuk mengetahui peningkatan akurasi dengan seleksi fitur.

REFERENSI

Amro, Belal, (2018). Phishing Techniques in Mobile Devices, Journal of Computer and Communications, , 6, 27-35

Al-diabat Mofleh, (2016). Detection and Prediction of Phishing Websites using Classification Mining Techniques. International Journal of Computer Applications (0975 – 8887) Volume 147 – No.5, August

Chaudhry Junaid, Chaudhry Shafique, Rittenhouse Robert, (2016). Phishing Attacks and Defenses, Internasional Journal of Security and its Applications," V ol. 10, No. 1 (2016), pp.247-256

Desiani, A., & Muhammad, A., (2006). Konsep Kecerdasan Buatan. Yogyakarta: Cv. Andi Offset.

Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P, (1996). From Data Mining to Knowledge Discovery: An overview in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass.. pp. 1-36

Giudici & Figini. (2009). Applied Data Mining for Business and Industry, 2nd Edition

ISBN: 978-602-61268-5-6

Gorunescu.(2011). Data Mining Concepts, Models and Techniques. Romania. Springer-Verlag Berlin Heidelberg.

J.H. Holland, (1975). Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, MI

Larose, (2005), "Discovering Knowledge in Data: An Introduction to Data Mining", John Willey & Sons, Inc

Mia Haryati Wibowo dan Nur Fatimah, (2007). Ancaman Phishing Terhadap Pengguna Sosial Media Dalam Dunai Cyber Crime" Volume 1 Nomor 1 : 1 – 5

Seema, Rathi Monika, Mamta, (2013). Decision Tree: Data Mining Techniques, International Journal of Latest Trends in Engineering and Technology (IJLTET)

Suryanto. (2007). Artificial Intelligent, Searching, Reasoning Planning dan Learning. Bandung: Informatika Bandung.

Wu, Xindong, (2007) "Top 10 Algorithms in Data Mining", Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007 Published online: 4 December 2007

Vercellis, Carlo. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. United Kingdom: John Willey & Son

http;//www.internetworldstats.com/stats (2018)

Aswan Supriyadi Sunge, M.Kom. Lahir di Jakarta, 26 Januari 1980. Penulis adalah Staff Pengajar di STT Pelita Bangsa sejak tahun 2014-sekarang. Menyelesaikan Studi S2 di Pascasarjana STMIK Nusa Mandiri Jakarta program studi Ilmu Komputer. Penelitian yang pernah dilakukan seperti : (1) Komparasi Menggunakan Algoritma C4.5, Neural Network dan Naïve Bayes Dalam Prediksi Ujian Kompentensi SMK Mahadhika 4 Jakata, Terbit di Seminar Nasional Ilmu Pengetahuan dan Teknologi Komputer 2 (1), 391-397 Vol. 2014. (2) Prediksi Ujian Kompetensi Dengan Menggunakan Klasifikasi Algoritma C4. 5 Di SMK Mahadhika 4 Jakarta, terbit di Bina Insani ICT Journal 1 (2), 136-150 Vol. , 2014. (3) Prediksi Kompetensi Karyawan Menggunakan Algoritma C4.5 (Studi Kasus : PT Hankook Tire Indonesia) terbit di Seminar Nasional Teknologi Informasi dan Komunikasi Universitas Atmajaya Jogyakarta tanggal 23 -24 Maret 2018. Nomor ISSN Publikasi Online Sentika : 2337-3377