IMPLEMENTASI DAN ANALISA HASIL DATA MINING UNTUK KLASIFIKASI SERANGAN PADA *INTRUSION DETECTION SYSTEM* (IDS) DENGAN ALGORITMA C4.5

Izza Khaerani[1], Lekso Budi Handoko[2]
[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Nakula 5 – 11, Semarang 50131, 024-3517261
E-mail : zzakhaerani@gmail.com[1], handoko@dosen.dinus.ac.id[2]

Abstrak

Intrusion Detection System (IDS) merupakan sebuah kemampuan yang dimiliki oleh sebuah sistem atau perangkat untuk dapat melakukan deteksi terhadap serangan yang mungkin terjadi dalam jaringan baik lokal maupun yang terhubung dengan internet. Masalah dimulai ketika paket data yang datang sangat banyak dan harus di analisa di kemudian hari. Teknik Data Mining merupakan teknik yang tepat untuk melakukan analisa terhadap sebuah data. Beberapa penelitian telah menggunakan teknik data mining untuk mengatasi masalah serangan IDS seperti analisis frequent itemset, analisis clustering, analisis klasifikasi dan analisis asosiasi. Tujuan dari penelitian ini adalah untuk mengklasifikasikan serangan pada data-data yang diujikan dengan menggunakan metode klasifikasi dan algoritma klasifikasi C4.5. Penelitian ini menggunakan koleksi data dari KDD'99 dan memiliki 41 atribut dimana atribut ini dilakukan fitur seleksi untuk menghapus atribut yang tidak relevan dengan menggunakan teknik evolusi. Hasil yang didapatkan dari fitur seleksi ini adalah 16 atribut dengan akurasi tinggi mencapai 98,67% dari 41 atribut yang ada. Kemudian hasilnya dilakukan pemodelan dengan menggunakan algoritma C4.5 dan menghasilkan sebuah aturan untuk digunakan dalam implementasi sistem analisa klasifikasi data. Aturan yang dihasilkan dapat digunakan dalam sistem untuk mengklasifikasikan data serangan seperti dos, u2r, r2l dan probe serta aktifitas jaringan normal.

Kata Kunci: *Klasifikasi, Algoritma C4.5, Fitur Seleksi, Evolusi, Intrution Detection System, IDS.*

Abstract

Intrusion Detection System is the device ability to detect attack posibility by the local network or the internet. The problem begin when many package data come and someday need to be analyze. Data Mining Method is the right technique to analyze the data. Some research have been using data mining method to handle an IDS attack problems like frequent itemset analysis, clustering analysis, classification analysis and association analysis. The objective of this research is to classified on some data testing using classification method and C4.5 algorithm classification. This research is using data set from KDD'99 and has 41 attribute, where the attribute have feature selection to delete irrelevant attribute using evolutionary technique. The result from this feature selection is 16 attributes with high accuracy 98,67% of 41 attributes. Then the result are modeled using C4.5 algorithm and producing some rules to apply in the implementation of classification data analysis system. The rules result could applied in the system to classify the data attack like dos,u2r, r2l and probe also normal activity networking.

Keywords: *Classification, C4.5 Algorithm, Feature Selection, Evolutionary, Intrution Detection System, IDS.*

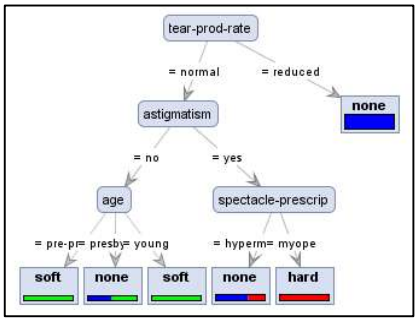
1. PENDAHULUAN

IDS adalah kemampuan yang dimiliki oleh perangkat keras atau peranti lunak yang berfungsi untuk mendeteksi aktivitas yang mencurigakan pada jaringan dan menganalisis serta mencari bukti percobaan intrusi (penyusupan).

Pada umumnya, IDS dibagi menjadi dua bentuk yang digunakan saat ini dan keduanya mempunyai perbedaan dalam hal mendeteksi dan menangguhkan kegiatan yang jahat. Keduanya harus dikembangkan, sehingga hasilnya lebih efektif mendeteksi setiap penyusupan dan menyiapkan strategi yang tepat. Berikut adalah dua bentuk IDS yang dipaparkan di buku Tom Thomas [1].

1. *Network-Based Intrusion Detection System* (NIDS)
Menempati secara langsung pada jaringan dan melihat semua aliran yang melewati jaringan. NIDS merupakan strategi yang efektif untuk melihat trafik keluar atau masuk maupun trafik di antara host atau diantara segmen jaringan lokal. NIDS biasanya dikembangkan di depan dan di belakang firewall dan VPN gateway untuk mengukur keefektifan peranti-peranti keamanan tersebut dan berinteraksi dengan mereka untuk memperkuat keamanan jaringan.

2. *Host-Based Intrusion Detection System* (HIDS)
Merupakan aplikasi perangkat lunak khusus yang diinstal pada komputer (biasanya server) untuk melihat semua aliran komunikasi masuk dan keluar ke dan dari server tersebut dan untuk memonitor sistem file jika ada perubahan. HIDS sangat efektif untuk server aplikasi Internet-accessible, seperti web atau e-mail server karena mereka dapat melihat aplikasi pada source-nya untuk melindungi mereka. Algoritma C4.5 merupakan salah satu algoritma klasifikasi, algoritma ini berfungsi untuk membuat *decision tree* (pohon keputusan). Selain menggunakan algoritma C4.5, ID3 dan CART merupakan algoritma yang dipakai dalam pembuatan *decision tree*. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 [2]. *Decision tree* berguna untuk mengekplorasi data dengan menemukan hubungan yang tersembunyi antara variabel input dengan variabel target. Data (input) pada algoritma C4.5 berupa tabel dan menghasilkan output berupa pohon.



Gambar 1. Pohon keputusan algoritma C4.5

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagi berikut [3] :

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus seperti tertera dalam persamaan 1 sebagai berikut :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \left(\frac{|S_i|}{|S|} * Entropy(S_i)\right) \quad (1)$$

Keterangan :
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

Sementara itu, perhitungan nilai entropy dapat dilihat pada persamaan 2 sebagai berikut :

$$Entropy(S.A) = \sum_{i=1}^{n} - pi * log_2\ pi \qquad (2)$$

Keterangan :
S : himpunan kasus
A : fitur
n : jumlah partisi S
pi : proporsi dari Si terhadap S

Metode *decision tree* tidak memilih atribut yang tidak relevan dan tidak membantu dalam pembuatan tree. Solusinya adalah dengan menghilangkan atribut kecuali atribut yang relevan dengan proses pembelajaran karena penghapusan atribut yang tidak relevan dapat meningkatkan performa algoritma pembelajaran [4]. Pada penelitian ini menggunakan *feature selection* dengan optimasi *evolutionary* untuk menghilangkan atribut yang tidak relevan.

2. METODE PENELITIAN

Metode pengembangan sistem yang digunakan pada penelitian ini adalah menggunakan model Waterfall, berikut ini adalah langkah-langkahnya [5] yang digunakan dalam penelitian ini :

2.1 Tahap Perencanaan

Tahap ini dapat dikatakan sebagai identifikasi masalah dan kebutuhan sistem untuk dilakukan pemodelan dan merancang aplikasi yang akan dibuat. Tahap ini juga dapat dikatakan project definition sebagai tujuan yang akan dicapai. Tujuan akhir penelitian ini adalah merancang aplikasi data mining untuk mengklasifikasi jenis serangan pada *Intrusion Detection System* (IDS) dengan menerapkan algoritma C4.5 sebagai salah satu jenis metode klasifikasi pada data mining.

2.2 Tahap Analisis Sistem

Pada tahapan analisis, melibatkan metode lain untuk proses standar Data Mining yaitu CRISP-DM. CRISP-DM mempunyai 6 tahap yaitu pemahaman bisnis, pemahaman data, pengolahan data, pemodelan, evaluasi, dan penyebaran. Berikut penjelasan masing-masing tahap [6].

2.2.1 Pemahaman Bisnis

Menurut data dari Indonesia Security Incident Responses Team on Internet Infrastructure (ID-SIRTII) [7], Pada tahun 2011 terdapat 1,25 juta serangan setiap hari. Data lain menunjukkan bahwa pada tahun 2013 terdapat peningkatan 27,4% kasus cyber crime dari tahun 2012, tahun 2012 terdapat 816 kasus dan meningkat menjadi 1.237 kasus pada tahun 2013 [8]. Deteksi intrusi merupakan proses monitoring dan menganalisis kejadian yang terjadi pada sistem komputer dalam mendeteksi tanda-tanda masalah keamanan. IDS adalah bagian terpenting pada infrastruktur jaringan yang terkoneksi internet, karena banyak cara untuk membahayakan stabilitas dan keamanan jaringan [9], namun terkadang IDS memberikan banyak alert yang salah (false positive) atau menampilkan banyak alert (alert flood) [10]. Pada IDS memiliki database besar dan harus mengenali pola dan harus dianalisis pola tersebut [1], sehingga harus menggunakan teknik analisis data yang tepat untuk mendeteksi serangan tersebut dan teknik data mining merupakan salah satu solusi untuk analisis data.

2.2.2 Pemahaman Data

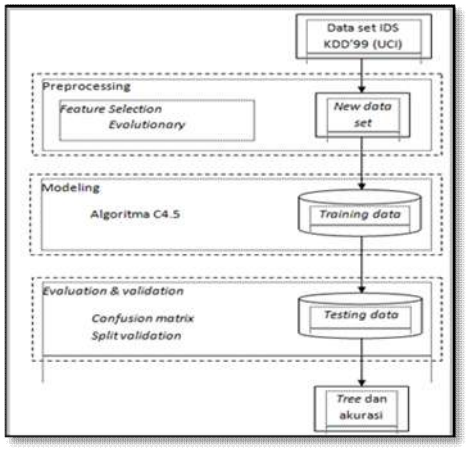
Data yang digunakan pada penelitian ini yaitu dengan menggunakan data IDS

yang dikelola oleh MIT Lincoln Labs yang juga dikompetisikan pada KDD Cup 1999 [11]. Dataset KDD'99 digunakan karena terdapat beberapa penelitian yang menggunakan data tersebut, sehingga membuktikan bahwa data KDD'99 layak unttuk digunakan [9][12][13][14]. Data yang digunakan yaitu 500 data dari KDD'99 10%.

2.2.3 Pengolahan Data

Karena atribut terlalu banyak yaitu 41 atribut, maka dilakukan proses feature selection dengan menggunakan evoluationary. Sehingga menjadi 16 atribut, berikut atribut data set yang baru.

2.2.4 Pemodelan



Gambar 2. Model C4.5 yang diusulkan

2.2.5 Evaluasi

Pada tahap ini, framework rapidminer 5.3 menyediakan cara untuk menghitung tingkat akurasi model dengan menggunakan confusion matrix dan split validation untuk validasi.

2.2.6 Penggunaan

Hasil dari penelitian ini adalah analisa yang mengarah ke Decision Support System (DSS) untuk perusahaan yang bergelut dibidang jaringan sebagai bahan pertimbangan untuk mengklasifikasikan jenis-jenis serangan sehingga dapat dianalisa dan ditangani dengan tepat.

2.3 Tahap Perancangan

Tahap ini merupakan proses menerjemahkan kebutuhan ke dalam representasi perangkat lunak untuk melakukan perancangan sebelum dilakukannya pengkodean. Tahap ini dibagi menjadi 3 yaitu sebagai berikut :

a. Merancang alur sistem dengan diagram menggunakan *Use Case Diagram* dan *Activity Diagram*.
b. Perancangan database untuk menampung semua data.
c. Perancangan *interface* untuk masukan dan keluaran sistem.

2.4 Tahap Implementasi

Implementasi merupakan tahap menerjemahkan perancangan sistem berupa diagram ke dalam bahasa pemrograman. Pembuatan aplikasi menggunakan bahasa pemrograman web dengan php dan SQLyog untuk memanipulasi database.

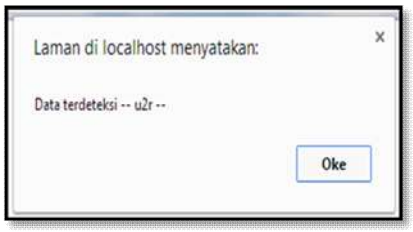
2.5 Tahap Pengujian

Tahap pengujian ini berfungsi untuk mengoreksi dan memastikan bahwa sistem tidak terjadi *error* ataupun kesalahan. Pengujian pada penelitian ini menggunakan pengujian *blackbox*.

3. HASIL DAN PEMBAHASAN

Gambar 3. Tampilan halaman input data testing

Gambar diatas menunjukkan tampilan halaman yang berfungsi sebagai inputan dataset dan memiliki 16 atribut data yang akan diinput dan terdapat dua button yaitu tombol input dan batal. Data tersebut akan disimpan ke database dan memberikan kotak pesan data testing dapat langsung terdeteksi pada kotak pesan, berikut tampilan kotak pesan seperti pada gambar 4 diawah ini.

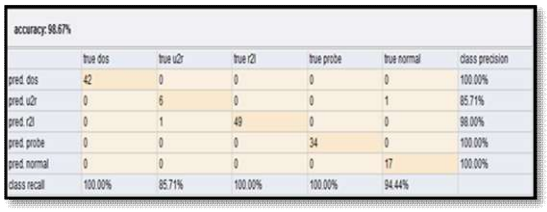


Gambar 4. Tampilan kotak pesan

Gambar 5. Tampilan halaman hasil mining

Gambar 6. Tampilan halaman pencarian

Hasil dari pemodelan C4.5 menggunakan *feature selection (evoluationary)* dapat dilihat pada gambar 7 berikut.



Gambar 7. Hasil akurasi model C4.5 dengan *feature selection*

Pemodelan ini jga menghasilkan akurasi dan nilai bobot pada setiap atribut, akurasi diukur dengan menggunakan *confusion matrix*.

Penjelasan atribut yang dihasilkan, waktu dan akurasi dapat dilihat pada tabel 1 dibawah ini :

Tabel 1: Tabel Statistik Pembentukan Atribut

Jumlah record	500 record
Waktu eksekusi	1 Menit 33 Detik
Akurasi	98.67 %
Memori yang digunakan	520 MB
Atribut yang terbentuk	16 Atribut

Data yang digunakan yaitu 500 *record* dan menghasilkan akurasi 98,67%, dengan waktu 1 menit 33 detik dan atribut yang terbentuk berjumlah 16 atribut. Berikut penjelasan atribut yang dipilih dari pengolahan *feature selection*.

Tabel 2: Tabel atribut berbobot "0" dan "1"

Atribut berbobot "0"	Atribut berbobot "1"
Duration	flag
protocol_type	src_bytes
Service	dst_bytes
wrong_fragment	land
Urgent	num_failed_logins
Hot	num_compromised
logged_in	su_attempted
diff_srv_rate	num_root
srv_diff_host_rate	num_access_files
dst_host_srv_count	num_outbond_cmds
dst_host_same_srv_rate	is_guess_login
dst_host_srv_rerror_rate	same_srv_rate
root_shell	dst_host_count
num_file_creation	dst_host_diff_srv_rate
num_shells	dst_host_srv_serror_rate
is_host_login	dst_host_rerror_rate
Count	
srv_count	
serror_rate	
srv_serror_rate	
rerror_rate	
srv_rerror_rate	
dst_host_same_src_port_rate	
dst_host_srv_diff_host_rate	
dst_host_serror_rate	

Data penelitian ini mempunyai 41 atribut, dari tabel diatas atribut yang mempunyai bobot "1" terdapat 16 atribut dan 25 atribut yang mempunyai bobot "0", sehingga atribut yang dipakai hanya 16 karena memiliki bobot "1". Model *feature selection* hanya digunakan untuk penghilangan nilai atribut yang tidak dipakai, sehingga untuk membuat pohon keputusan yang digunakan untuk rule dan diimplementasikan ke sistem, akan dibentuk model kembali dengan atribut yang sudah dihilangkan yaitu 16 atribut.

Penelitian ini menggunakan algoritma C4.5 dengan *feature selection*. Pada dataset ini, dari 41 atribut, hanya 16 atribut yang mempunyai bobot "1" dan 25 atribut mempunyai bobot "0". Bobot "1" ini yang dipakai untuk pembentukan pohon keputusan dan menghasilkan rules. Akurasi yang dihasilkan adalah 98.67%, didapat dari pengukuran *confusion matrix*. Penelitian ini juga menghasilkan pohon keputusan, dan dari pohon keputusan tersebut maka terbentuknya *rules* yang dapat diimplementasikan ke dalam sistem untuk digunakan pengujian data testing.

Sistem yang dibangun digunakan untuk mengklasifikasi serangan. Sistem ini mempunyai beberapa fungsi seperti fungsi input data testing, menampilkan hasil input dan fungsi melakukan pencarian, fungsi tersebut dapat berjalan dengan baik. Sistem ini berhasil mengklasifikasikan data serangan berdasarkan inputan data dari pengguna sesuai tujuan penelitian ini. Data dapat tersimpan di *database* dan adanya fungsi pencarian untuk memudahkan pengguna mencari data dari waktu lampau.

4. KESIMPULAN

Penelitian ini telah dilakukan dengan menggunakan model C4.5 dengan feature selection dan rules yang didapat untuk diimplementasikan pada sistem analisa data testing, sehingga dapat disimpulkan hasil dari percobaan antara lain :

1. Penelitian ini menggunakan metode klasifikasi dengan algoritma C4.5 untuk pembentukan rule, sebelum melakukan pemodelan C4.5 dilakukan preprocessing data menggunakan feature selection yaitu evoluationary. Awal jumlah atribut pada data penelitian ini yang berasal dari KDD'99 adalah 41 atribut, setelah dilakukan feature selection jumlah atribut menjadi 16 atribut dengan akurasi 98.67%.
2. Terbentuknya rules atau model untuk klasifikasi data serangan pada IDS, sehingga rules ini dapat diterapkan pada implementasi sistem.
3. Sistem yang dibuat hanya untuk pengujian data testing dan menampilkan data hasil klasifikasi.

DAFTAR PUSTAKA

[1] T. Thomas, 2005. *Network Security first-step*, Yogyakarta : Penerbit Andi.

[2] D. T. Larose, 2005. *Discovering Knowledge In Data : Introduction to Data Mining*, Canada : Wiley.

[3] Kusrini and E. T. Luthfi, 2009. *Algoritma Data Mining*, Yogyakarta, Indonesia: Penerbit Andi.

[4] I. H. Witten, E. Frank, and M. A. Hall, 2011. *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. USA : Morgan Kaufmann Publishers.

[5] R. S. Pressman, 2002. *Rekayasa Perangkat Lunak Pendekatan Praktisi*, 1st ed., L. Harnaningrum, Ed. Yogyakarta, Indonesia : Penerbit Andi.

[6] D. T. Larose, 2005. *Discovering Knowledge In Data : Introduction to Data Mining*, Canada : Wiley.

[7] Rizagana, Juni 2011. *Polisi Galakkan Patroli Cyber, Tiap Hari 1,25 Juta Serangan Cyber*. [Online] : http://www.investor.co.id/home/polisi-galakkan-patroli-cyber-tiap-hari-125-juta-serangan-cyber/14882

[8] R. Kurniawan, 2013. *Selama 2013 Tingkat Nasional, Cyber Crime Meningkat 27,4 Persen*. [Online] : http://www.itoday.co.id/metro/kriminal/selama-2013-tingkat-nasional-cyber-crime-meningkat-274-persen

[9] A. M. Chandrasekhar and K. Raghuveer, Jan 04-06, 2013. *Intrusion Detection Technique by using K-Means, Fuzzy Neural Network, and SVM Classifiers*, Coimbatore, India : International Conference on Computer Communication and Informatics (ICCCI-2013).

[10] R. S. Hakim, I. Winarno, and W. Yuwono, *Verifikasi Alert Berdasarkan klasifikasi serangan pada deteksi intrusi kolaboratif*.

[11] [Online] : http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data.

[12] M. Kumar, D. M. Hanumanthappa, and D. T. V. Suresh Kumar, 2012. *Intrusion Detection System Using Decision Tree Algorithm*, India.

[13] J.-H. Leet, J.-H. Leet, S.-G. Sohn, J.-H. Ryu, and T.-M. Chung, 2008. *Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System*, Korea : ICACT.

[14] G. Zhai and C. Liu, 2010. *Research and Improvement on ID3 Algorithm in Intrusion Detection System*, China : Sixth International Conference on Natural Computation (ICNC 2010).