p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek

IMPLEMENTASI SELEKSI FITUR MENGGUNAKAN ALGORITMA FVBRM UNTUK KLASIFIKASI SERANGAN PADA *INTRUSION DETECTION SYSTEM* (IDS)

Jupriyadi
Program Studi Teknologi Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Bandarlampung, Jl. H. Zaenal Abidin, Pagaralam, No. 9-11, Labuhan Ratu, 35142
jupriyadi@teknokrat.ac.id

Abstrak

Seleksi fitur merupakan salah satu fokus penelitian pada data mining untuk dataset yang memiliki atribut yang relatif banyak. Dengan menghilangkan beberapa atribut yang tidak relevan terhadap kelas label akan dapat meningkatkan kinerja algoritma klasifikasi. Algoritma FVBRM merupakan salah satu algoritma untuk mencari fitur yang tidak relevan terhadap kelas label. Algoritma ini menggunakan teknik *wrapper* untuk menghilangkan atribut yang tidak relevan. Penelitian ini bertujuan untuk mengimplementasikan seleksi fitur menggunakan algoritma FVBRM terhadap dataset deteksi intrusi NSL KDD yang memiliki jumlah atribut relative banyak. Dataset dengan atribut terpilih akan diuji menggunakan algoritma klasifikasi naive bayes. Evaluasi dilakukan dengan melihat tingkat akurasi klasifikasi yang dihasilkan tanpa seleksi fitur dengan akurasi klasifikasi yang dihasilkan setelah implementasi seleksi fitur. Eksperimen klasifikasi dilakukan dengan dua cara yaitu *binary classification* (serangan atau bukan serangan) dan lima kelas klasifikasi yaitu *dos, r2l, u2r, probe* dan normal. Eksperimen dilakukan menggunakan *library* data mining dilingkungan Weka menggunakan 10 *fold validation*. Hasil eksperimen menunjukkan bahwa dengan implementasi seleksi fitur menggunakan algoritma FVBRM dapat meningkatkan akurasi klasifikasi menjadi 90,81% untuk dataset dengan 2 kelas label dan 86,55% untuk dataset dengan 5 kelas label.

Kata kunci: FVBRM, *Intrusion Detection System* (IDS), seleksi fitur

Abstract

Feature selection is one of the research focuses on data mining for dataset that have relatively many attributes. Removing some attributes that are not relevant to the class label will improve the performance of the classification algorithm. The FVBRM algorithm is one of the algorithms for finding features that are not relevant to the class label. This algorithm uses a wrapper technique to eliminate irrelevant attributes. This study aims to implement feature selection using the FVBRM algorithm on the NSL KDD dataset for intrusion detection which has a relatively large number of attributes. The dataset with the selected attributes will be tested using the Naive Bayes classification algorithm. Evaluation is done by looking at the level of classification accuracy produced without feature selection with classification accuracy that is generated after the implementation of feature selection. Classification experiments are carried out in two ways, namely binary classification and five classification classes, namely dos, r2l, u2r, probes and normal. Experiments carried out using data mining libraries in Weka environment using 10 fold validation. The experimental results show that the implementation of feature selection using FVBRM algorithm can improve the classification accuracy to 90.81% for datasets with 2 label classes and 86.55% for datasets with 5 label classes.

Keywords : FVBRM, *Intrusion Detection System* (IDS), *feature selection*

p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek

PENDAHULUAN

Seleksi fitur merupakan salah satu bidang penelitian dalam data mining untuk dataset yang memiliki atribut yang relatif banyak. Hal ini bertujuan untuk mereduksi data yang ada sehingga dapat mempercepat waktu komputasi. Tidak hanya mempercepat waktu komputasi, seleksi fitur juga dapat meningkatkan akurasi klasifikiasi dari algoritma yang digunakan (Dony dkk, 2017). Sistem deteksi intrusi adalah salah satu area penelitian yang penting dalam jaringan dan keamanan komputer. Beberapa penelitian telah dilakukan untuk deteksi intrusi menggunakan data mining berbasis klasifikasi untuk menentukan paket data yang terekam apakah masuk dalam kategori serangan atau bukan. *Framework* sistem deteksi intrusi jaringan berbasis naïve bayes *classifier* telah diusulkan dan dibandingkan dengan pendekatan jaringan saraf tiruan, hasil menunjukkan bahwa metode yang diusulkan menghasilkan tingkat deteksi yang lebih tinggi namun menghasilkan *false positive* yang cukup tinggi (Mrutyunjaya, 2007). Algoritma klasifikasi Naïve bayes dan *decision tree* telah dibandingkan dan menunjukkan bahwa naïve bayes umumnya 7 kali lebih cepat dari metode *decision tree* untuk proses training dan testing, tetapi memiliki akurasi yang lebih rendah (Nehla, 2004).

Dataset NSL-KDD merupakan dataset yang dapat digunakan sebagai pembanding berbagai metode klasifikasi untuk deteksi intrusi. Dataset ini memiliki dimensi yang cukup tinggi dengan 41 fitur. Seleksi fitur merupakan salah satu preproses penting untuk mereduksi dataset dengan menghilangkan fitur yang tidak penting dari dataset NSL-KDD. Tidak semua fitur yang ada dalam dataset tersebut memiliki pengaruh terhadap kelas label. Oleh karena itu menghilangkan atribut yang tidak penting dengan kelas label merupakan hal penting yang harus dilakukan untuk meningkatkan kinerja *classifier*. Tujuan utama dari seleksi fitur adalah memilih fitur yang penting dan menghapus fitur yang tidak penting dan kurang penting terhadap kelas label guna meningkatkan kinerja *classifier* yaitu meningkatkan akurasi dan mengurangi waktu komputasi.

Deteksi Intrusi

Sistem deteksi intrusi merupakan proses memonitor trafik jaringan dalam sebuah sistem untuk mendeteksi adanya pola data yang mencurigakan yang memungkinkan adanya serangan dalam sistem tersebut. Terdapat dua kategori teknik yang digunakan untuk deteksi intrusi yaitu deteksi intrusi berbasis *anomaly* dan deteksi intrusi berbasis *signature* (Nehla, 2004). Deteksi intrusi berbasis *anomaly* menggunakan pendekatan terhadap pola data normal untuk mendeteksi intrusi. Jika data yang terekam menyimpang dari data normal maka mengindikasikan adanya serangan. Deteksi intrusi berbasis *signature* menggunakan pola data serangan yang sudah diketahui sebelumnya untuk mendeteksi intrusi. Paket data yang terekam dicocokkan dengan data serangan yang ada dalam database. Terdapat tiga jenis IDS jika dilihat dari lokasi atau area yaitu *host-based* IDS (HIDS), *network-based* IDS (NIDS) dan *hybrid* IDS. HIDS ditempatkan pada sebuah device seperti server atau workstation dimana data yang dianalisa berada pada mesin lokal. Sedangkan NIDS menganalisa semua trafik data dalam sebuah jaringan dalam mendeteksi serangan.

Dataset NSL KDD merupakan dataset yang digunakan sebagai pembanding untuk penelitian dibidang deteksi intrusi. Dataset dibangun berdasarkan hasil simulasi dalam sebuah LAN selama 9 minggu. Dalam dataset tersebut terdapat 22 jenis serangan yang dikelompokkan ke dalam 4 kategori serangan yaitu:

1. *Denial of Service* (DoS) yaitu jenis serangan pada komputer yang bertujuan untuk mematikan layanan yang disediakan (contohnya Neptune, smurf, back, dan lain-lain).
2. *Probing* yaitu jenis serangan yang bertujuan untuk mendapatkan informasi dari sebuah sistem komputer atau jaringan.
3. *User to Root* (U2R) yaitu jenis serangan yang bertujuan untuk meningkatkan akses user menjadi super user dimana penyerang sudah memilik akun dalam sistem pada level user.
4. *Remote to Login* (R2L) yaitu jenis serangan yang bertujuan untuk mendapat akses terhadap sistem tujuan menggunakan komputer lain dalam jaringan.

Algoritma FVBRM

Berikut ini adalah algoritma FVBRM dalam memilih fitur yang penting atau berpengaruh terhadap kelas label.

Input:

F = Full set of all features of NSL-KDD dataset
ac = classifier accuracy
err = RMSE

p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek

avg_fpr = average FPR
//value of ac, err, and avg_fpr resulting from //invocation of NBC full dataset is used as a //threshold values for feature selection
//Modified FVBRM algorithm
Begin
Initialize: S = {F}
For each feature {f} form
(1) T = S – {f}
(2) Invoke NBC on dataset with T features
// Original
(3) If CA >= ac and RMSE <= err and A_TPR<= avg_TPR then
S = S – {f}
F = S // The selected feature set
End

Kriteria dan Model Evaluasi

Ukuran kinerja yang digunakan pada penelitian ini adalah jumlah fitur yang terpilih, akurasi *classifier* (CA), *false positive rate* (FPR) dan *detection rate* (DTR). Pengukuran ini didasarkan pada *confusion matrix* seperti tampak pada tabel 1 berikut ini dalam mendeteksi serangan.

Tabel 1. Tabel *confusin matrix*

	Klasifikasi Kelas	
Kelas	**Normal**	**Serangan**
Normal	TN	FP
Serangan	FN	TP

TN (*True Negative*) adalah jumlah kelas normal yang berhasil diklasifikasi kedalam kelas normal.
FP (*False Positive*) adalah jumlah kelas normal yang masuk dalam klasifikasi serangan (dideteksi sebagai serangan).
TP (*True Positive*) adalah jumlah serangan yang berhasil dideteksi sebagai serangan.
FN (*False Negative*) adalah jumlah serangan yang terdeteksi sebagai data normal oleh sistem.

CA (akurasi *classifier*) =
(TN + TP) / (TN + FP + FN + TP) (1)
DTR (*detection rate*) =
TP / (TP + FN) (2)
FPR (*false positive rate*) =
FP / (FP +TN) (3)

METODE

Pada bagian ini penulis uraikan pelaksanaan penelitian yang penulis lakukan.

A. **Pre-proses**

Dataset yang digunakan dalam eksperimen ini adalah NSL-KDD yang sudah diberi label yang merupakan perbaikan dari dataset KDD-99 dengan beberapa kelebihan yaitu:

1. Tidak terdapat duplikat record dalam data training sehingga kinerja pembelajaran *classifier* tidak bias kearah record yang sering muncul.
2. Tidak terdapat duplikat record dalam data testing sehingga kinerja *classifier* tidak bias ketika menggunakan metode yang memiliki *detection rate* yang baik pada data yang sering muncul.
3. Jumlah record yang dipilih sesuai dengan proporsi dataset original. Sehingga membuat pembelajaran menjadi lebih efisien dalam melakukan evaluasi menggunakan metode pembelajaran yang berbeda.
4. Jumlah record yang dilih sebagai data training dan testing cukup layak, sehingga dalam melakukan eksperimen tidak perlu lagi membagi data kedalam ukuran yang lebih kecil. Hal ini akan menjadikan evaluasi dari hasil berbagai penelitian akan konsisten dan dapat dibandingkan.

Dataset NSL-KDD tersedia pada dengan total data 125.973 record secara detail ditunjukkan pada Tabel 2 berikut ini.

Tabel 2. Distribusi dataset NSL-KDD

Serangan	**Jumlah record**	**Kelas**	**Jumlah record tiap kelas**
Back	956	DOS	45927
Land	18	DOS	
Neptune	41214	DOS	
Pod	201	DOS	
Smurf	2646	DOS	
Teardrop	892	DOS	

p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek

Satan	3633	PROBE	11656
Ipsweep	3599	PROBE	
Nmap	1493	PROBE	
Portsweep	2931	PROBE	
Normal	67343	NORMAL	67343
Guess_passwd	53	R2L	995
ftp_write	8	R2L	
Imap	11	R2L	
Phf	4	R2L	
Multihop	7	R2L	
Warezmaster	20	R2L	
Warezclient	890	R2L	
Spy	2	R2L	
Buffer_overflow	30	U2R	52
Loadmodule	9	U2R	
Perl	3	U2R	
Rootkit	10	U2R	

Pada dataset NSL-KDD terdapat 1 data normal data dan 22 jenis serangan yang dikelompokkan kedalam 4 kategori serangan yaitu DOS, Probe, R2L, dan U2R. Dalam penelitian ini penulis hanya menggunakan 62.986 record dari 125.973. Pemilihan record dilakukan dengan cara menerapkan *remove percentage* 50 pada weka. Distribusi data yang terpilih ditunjukkan pada tabel 3 berikut ini.

Tabel 3. Distribusi dataset yang digunakan

Label kelas	Jumlah record	Persentase record
Normal	33896	53,82%
DOS	22817	36,23%
PROBE	5781	9,18%
R2L	467	0,74%
U2R	25	0,04%
Total	**62.986**	**100%**

Dalam penelitian ini dataset juga diuji dengan mengelompokkan menjadi dua kelas yaitu kelas normal dan serangan dengan distribusi data seperti pada tabel 4 sebagai berikut.

Tabel 4. Distribusi dataset 2 kelas label

Label kelas	Jumlah record	Persentase record
Normal	33896	53,82%
Serangan	29090	46,18%
Total	**62.986**	**100%**

Eksperimen dilakukan menggunakan library WEKA 3.7 pada editor *Netbeans* 7.1 untuk seleksi fitur menggunakan algoritma FVBRM serta untuk mengukur kinerja dari *classifier* terhadap set fitur yang terpilih. Penulis menggunakan naive bayes *classifier* terhadap dataset untuk mengidentifikasi fitur yang penting dengan pengujian 10 *cross fold validation.* Hal ini juga dilakukan untuk mengukur kinerja *classifier*. Pengujian menggunakan 10 *fold cross validation* membagi dataset ke dalam 10 subset dengan ukuran yang relative sama. Satu subset digunakan sebagai data testing dan 9 subset yang lain digunakan sebagai data training untuk membangun model dari *classifier.* Hal ini diulang sebanyak 10 kali untuk setiap data training dan data testing. Kinerja yang diperoleh berdasarkan rata-rata hasil dari setiap *classifier*. *Cross validation* telah diuji dan umumnya bekerja lebih baik ketika dataset yang digunakan mencukupi.

HASIL DAN PEMBAHASAN

Pada bagian ini akan diuraikan hasil eksperimen yang telah dilakukan. Hasil pemilihan fitur yang relevan/penting terhadap kelas label ditunjukkan pada tabel 5 berikut ini.

p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek

Tabel 5. Hasil pemilihan fitur

Jumlah kelas	Jumlah fitur terpilih	Nomor fitur
2	21	2, 3, 5, 6, 7, 8, 18, 20, 23, 24, 25, 26, 29, 32, 33, 34, 35, 36, 37, 38, 39
5	25	2, 3, 4, 8, 10, 11, 12, 14, 15, 17, 18, 19, 22, 23, 24, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39

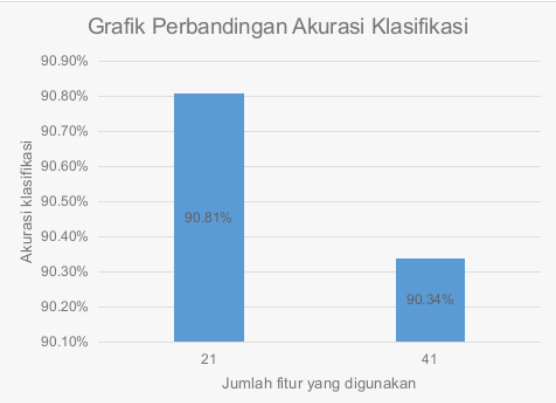
Berdasarkan tabel diatas dapat dilihat bahwa ditemukan 21 fitur penting untuk dataset menggunakan 2 kelas label dan 25 fitur penting untuk dataset dengan 5 kelas label. Hal ini artinya hampir 50% dari fitur dataset tidak berpengaruh terhadap kelas label.

Berikut ini adalah tabel hasil pengujian akurasi klasifikasi.

Tabel 6. Hasil pengujian akurasi klasifikasi

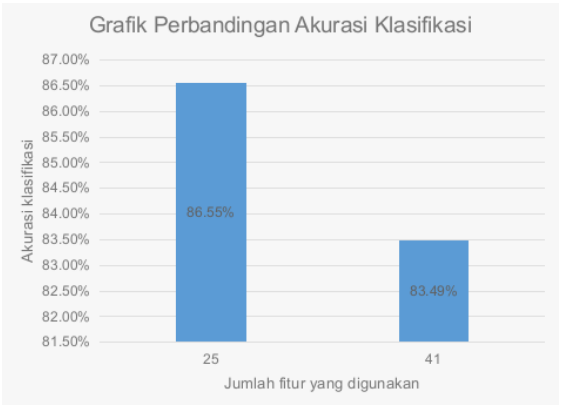
Jumlah kelas	Fitur yang digunakan	Akurasi
2	21	90,81 %
2	41	90,34%
5	25	86.55%
5	41	83,49%

Grafik hasil pengujian akurasi klasifikasi terhadap 2 dan 5 kelas label dapat dilihat pada gambar 1 dan gambar 2 berikut ini.



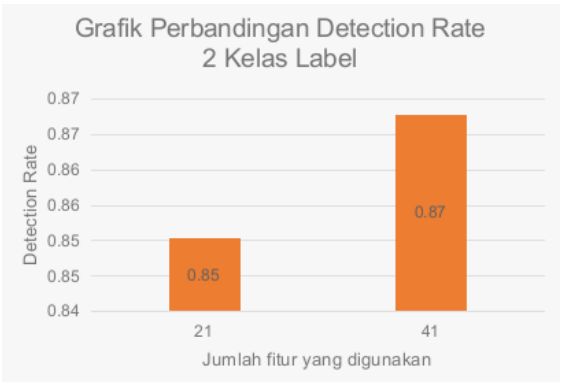
Gambar 1. Perbandingan akurasi klasifikasi 2 kelas label

Pada gambar 1 dapat dilihat bahwa akurasi klasifikasi terhadap dataset yang telah terpilih memberikan nilai akurasi yang lebih tinggi jika dibandingkan dengan akurasi menggunakan seluruh fitur.

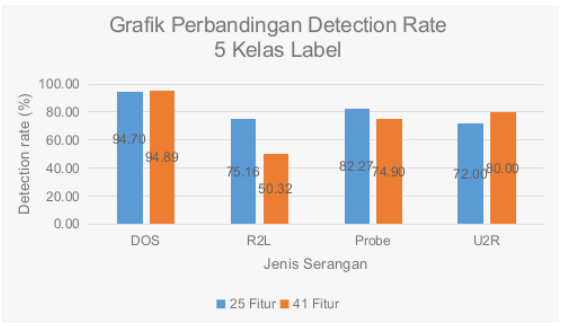


Gambar 2. Grafik Perbandingan akurasi klasifikasi

Berikut ini adalah detection rate yang dihasilkan terhadap dataset 2 dan 5 kelas label.



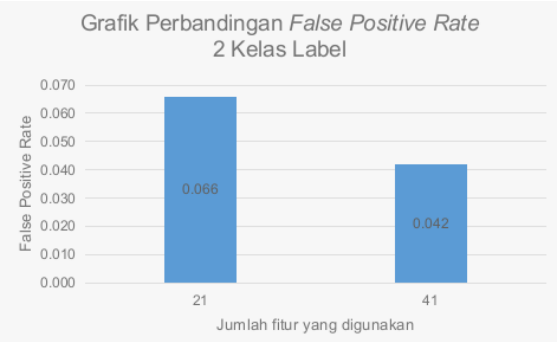
Gambar 3. Grafik Detection Rate 2 Kelas label



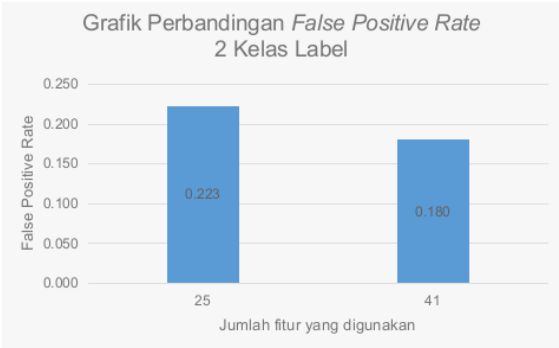
Gambar 4. Grafik Detection rate 5 kelas label

Pada gambar 4 dapat dilihat bahwa rata-rata *detectin rate* dengan fitur terpilih lebih besar daripada *detection rate* menggunakan seluruh fitur.

p- ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.umj.ac.id/index.php/semnastek



Gambar 5. Grafik *false positive rate* 2 kelas label



Gambar 6. Grafik *flase positive rate* 5 kelas label

SIMPULAN DAN SARAN

Berdasarkan hasil eksperimen yang telah dilakukan dapat disimpulkan bahwa terdapat peningkatan akurasi klasifikasi terhadap terhadap dataset 2 dan 5 kelas label sebesar 90,81% dan 86,55%. Hasil *Detection rate* terbaik pada eksperimen terhadap 5 kelas label namun menghasilkan nilai *false positive* yang lebih tinggi daripada menggunakan fitur secara keseluruhan.

DAFTAR PUSTAKA

Mrutyunjaya Panda dan Manas Ranjan Patra, 2007, *Network Intrusion Detection Using Naïve Bayes.* International Journal of Computer Scienece and Network Security, vol. 7, no 12

Saurabh Mukherjee dan Neelman Sharmaa, 2012, *Intrusion Detection using Naïve Bayes Classifier with Feature Reduction*. Procedia Technology 4 (2012) 119 – 128

Dataset NSL KDD, http://www.unb.ca/cic/datasets/nsl.html (Diakses 14 September 2018)

Weka machine learning tool available on http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Wirawan dan Eksistyanto, 2015, *Penerapan Naive Bayes Pada Intrusion Detection System Dengan Diskritisasi Variabel.* JUTI - Volume 13, Nomor 2

Dony, Noor, dan Adhistya, 2017, *Implementasi Data Mining dengan Seleksi Fitur untuk Klasifikasi Serangan pada Intrusion Detection System (IDS).* Prosiding CITEE Hal. 314-321

Nahla Ben Amor, et. al., 2004, *Naïve Bayes vs Decision Trees in Intrusion Detection System.* ACM Symposium on Applied Computing

Aulia, Rahmadani, Safriadi, 2016, *Analisys Information Gain Attribut Evaluation Untuk Klasifikasi Serangan Intrusi*, Jurnal ISD Vol 2 no. 2