NAÏVE BAYES UNTUK MENDETEKSI GANGGUAN JARINGAN KOMPUTER DENGAN SELEKSI ATRIBUT BERBASIS KORELASI

Bekti Maryuni Susanto
Manajemen Informatika AMIK BSI Yogyakarta
Jalan Ring Road Barat Ambarketawang Gamping Sleman 55294
Telp. (0274)4342536
Email: bekti.bms@bsi.ac.id

Abstract

Internet increasing is also exponentially increasing intrusion or attacks by crackers exploit vulnerabilities in Internet protocols, operating systems and software applications. Intrusion or attacks against computer networks, *especially the Internet* has increased from year to year. Intrusion detection systems into the main stream in the information security. The main purpose of intrusion detection system is a computer system to help deal with the attack. This study presents a *correlation-based feature selection* to detect computer network intrusions. *Feature selection result applied on naïve bayes algorithm.* Performance is measured based on the level of accuracy, sensitivity, precision and spesificity. Dataset used in this study is a dataset KDD 99 intrusion detection system. Dataset is composed of two training data and testing data. From the experimental results obtained by the accuracy of naïve Bayes without feature selection 76,12 %, and the accuracy with feature selection 81,89 %. Correlaiton-based feature selection can improve naïve bayes accuration.

Keyword: naïve bayes, intrusion detection, correlation-based fetaure selection

Abstrak

Internet yang meningkat secara eksponensial meningkatkan juga gangguan atau serangan yang dilakukan oleh cracker mengeksploitasi kelemahan pada protokol internet, sistem operasi dan software aplikasi. Gangguan atau serangan terhadap jaringan komputer khususnya internet mengalami peningkatan dari tahun ke tahun. *Intrusion detection system* menjadi aliran utama di dalam keamanan informasi. Tujuan utama intrusion detection system adalah membantu sistem komputer untuk menangani serangan. Penelitian ini menyajikan seleksi atribut berbasis korelasi (*correlation-based feature selection*) dalam mendeteksi gangguan jaringan komputer. Hasil seleksi atribut diterapkan pada algoritma naïve bayes. Performa diukur berdasarkan tingkat accuracy, sensitivity, precision dan spesificity. Dataset yang digunakan dalam penelitian ini adalah dataset intrusion detection system KDD 99. Dataset terdiri dari dua yaitu data training dan data testing. Dari hasil eksperimen diperoleh tingkat akurasi algoritma naïve bayes tanpa seleksi atribut 76,12 %, sedangkan akurasi dengan seleksi atribut 81,89 %. Seleksi atribute berbasis korelasi mampu meningkatkan akurasi naïve bayes.

Kata kunci: naïve bayes, intrusion detection, correlation-based fetaure selection

A. Pendahuluan

Internet yang meningkat secara eksponensial meningkatkan juga gangguan atau serangan yang dilakukan oleh cracker mengeksploitasi kelemahan pada protokol internet, sistem operasi dan software aplikasi. Gangguan atau serangan terhadap jaringan komputer khususnya internet mengalami peningkatan dari tahun ke tahun. Berdasarkan laporan dari Kaspersky Lab jumlah serangan melalui browser internet sejumlah 23.680.646 pada tahun 2007, meningkat menjadi 73.619.767 pada tahun 2009 dan meningkat lagi menjadi 580.371.937 pada tahun 2010. Internet browser menjadi alat utama dalam menyebarkan program-program malicious diantara sebagian besar pengguna komputer pada tahun 2010. Algoritma Kaspersky Security Network (KSN) hanya mampu mendeteksi serangan web sebesar 60 % (Gostev & Namestnikov, 2011).

Intrusion detection adalah proses memonitor kejadian pada sistem komputer atau jaringan dan menganalisanya untuk memberikan tanda insiden yang mungkin, yang mana yang merupakan pelanggaran atau mendekati penlanggaran sebuah kebijakan keamanan komputer, kebijakan penggunaan yang disetujui atau praktik keamanan standar. Intrusion prevention adalah proses untuk menampilkan intrusion detection dan berusaha untuk menghentikan kejadian yang mungkin dideteksi. *Intrusion detection* dan *prevention system* adalah perhatian utama dalam mengidentifikasi kejadian,

mencatat informasinya, berusaha untuk menghentikanya dan melaporkanya kepada administrator keamanan. Sebagai tambahan organisasi menggunakan *intrusion detection* dan *prevention system* (IDPS) untuk tujuan lain, seperti mengidentifikasi permasalahan kebijakan keamanan, mendokumentasikan perlakuan yang ada, dan menghambat individu dalam melakukan pelanggaran kebijakan keamanan. IDPS menjadi tambahan yang perlu terhadap infrastruktur keamanan bagi setiap organisasi(Scarfone & Mell, Februari, 2007).

Intrusion detection system menjadi aliran utama di dalam keamanan informasi. Tujuan utama *intrusion detection system* adalah membantu sistem komputer untuk menangani serangan.

Ada dua tipe intrusion detection system berdasarkan tipe operasi yang digunakan untuk mendeteksi gangguan, anomaly detection system dan misuse detection. Anomaly detection system membuat database tingkah laku normal dan penyimpanganya dari tingkah laku normal yang terjadi, sebuah peringatan dipicu oleh sebuah adanya gangguan. Misuse detection system menyimpan pola serangan yang telah terdefinisi sebelumnya di dalam sebuah database jika situasi dan data yang mirip terjadi diklasifikasi sebagai serangan. Berdasarkan sumber data IDS diklasifikasi menjadi IDS *host based* dan *network based*. IDS *netwok based* menganalisa paket secara individual yang melalui jaringan. IDS *host based* menganalisa aktivitas pada sebuah komputer tunggal atau host(Neethu, 2012).

Sebagian besar IDS saat ini menggunakan sistem berbasis *rule* atau pakar. Kekuatanya sangat tergantung pada kemampuan personel keamanan yang mengembangkan IDS. Dahulunya, IDS hanya bisa mendeteksi tipe serangan yang diketahui dan sekarang cenderung membangkitkan alarm *false* positif. Hal ini menyebabkan penggunaan teknik inteligence yang dikenal sebagai data mining ataupun *machine learning* sebagai alternatif kemampuan manusia yang mahal dan berat. Teknik ini secara otomatis mempelajari data atau mengekstrak pola yang bermanfaat dari data sebagai referensi profil tingkah laku normal atau serangan dari data yang ada untuk klasifikasi trafik jaringan selanjutnya(Olusola et al., 2010). *Machine learning* adalah sebuah bidang studi yang menyediakan komputer dengan kemampuan pembelajaran dari pengalaman sebelumnya. *Machine learning* berdasarkan analisa data statistik yang sangat besar dan beberapa algoritma dapat menggunakan pola yang ditemukan pada data sebelumnya untuk membuat keputusan tentag data baru(Tavallaee et al., 2009).

Penemuan pola atau pengetahuan dari dataset menggunakan teknik data mining klasik terbukti sulit dilakukan karena besarnya dimensi dataset baik atribut maupun sampel(Biesiada & Duch, 2007). Kemampuan *machine learning* dalam mengenali pola atau menemukan pengetahuan sangat tergantung pada kualitas dataset(Yu & Huan, 2003). *Feature selection* merupakan salah satu permasalahan dalam machine learning. *Feature selection* akan memilih feature yang relevan serta mengabaikan feature yang tidak relevan serta redundan. Feature yang tidak relevan akan menurunkan performa machine learning. Sedangkan *feature* yang redundan akan membuat *machine learning* bekerja lebih lama(Marcelloni, 2003). Pada beberapa kasus *feature selection* mampu meningkatkan performa algoritma. Pemilihan subset feature yang relevan dari sekumpulan feature yang besar sangat penting untuk meningkatkan performa machine learning(Pedrycz & Vukovich, 2001). Apalagi jika dataset mempunyai jumlah feature yang besar, akan sulit memilih seperangkat feature yang paling relevan terhadap kelas(Leng et al., 2010).

Ada dua metode dalam feature selection yaitu, metode filter dan wrapper. Metode filter dilakukan dengan menyaring atribut relevan menggunakan formula khusus. Berbagai macam penelitian feature selection menggunakan metode filtering, misalnya *correlation based feature selection* (CFS)(Hall, 1999). Feature selection berdasarkan korelasi dilakukan dengan menghitung korelasi suatu atribut dengan kelas. Untuk menghitung korelasi dapat menggunakan berbagai macam formula misal *symetrical uncertainty*, *gain ratio*, *information gain* dan *minimum description length.* Setelah menghitung korelasi kemudian dicari sekelompok feature S dimana jumlah feature di dalam S lebih kecil dari jumlah N feature sesungguhnya. Pencarian bisa menggunakan forward selection atau backward selection. Pada forward selection pencarian dimulai dengan jumlah atribut 0 sampai semua atribut masuk ke dalam subset S. Pada backward selection pencarian dimulai dengan N feature kemudian dikurangi satu sampai 0. Masing-masing subset dihitung rata-rata nilai korelasi terhadap kelas dan masing-masing atribut. Hasil feature selection adalah subset yang mempunyai nilai rata-rata korelasi terhadap atribut paling tinggi dan nilai korelasi antar feature dalam subset S yang paling rendah. Namun, metode ini mempunyai kelemahan yaitu tidak bisa menangani atribut kelas numerik. Kelemahan lain dalam metode filter adalah performa feature selection lebih rendah bila dibandingkan dengan performa feature selection wrapper. Keuntungan menggunakan metode ini waktu eksekusi yang lebih singkat.

Metode lain dalam feature selection adalah metode wrapper. Pada wrapper diterapkan algoritma machine learning tertentu dalam memilih subset

feature sebuah dataset(Kohavi & John, 1997). Algoritma feature subset selection melakukan sebuah pencarian subset yang bagus menggunakan algoritma induksi sendiri sebagai bagian dari evaluasi feature subset. Metode feature selection wrapper mampu menghapus feature-feature yang tidak releven serta noisy sebelum menerapkan teknik data mining untuk menganalisa dataset(Leng et al., 2010). Keuntungan menggunakan metode ini memiliki performa yang lebih tinggi dibandingkan metode filter. Kelemahan metode ini memerlukan waktu eksekusi yang lebih lama dibandingkan filter.

Penelitian ini menyajikan penerapan naïve bayes dengan seleksi atribut berbasis korelasi (*correlation-based feature selection*) dalam mendeteksi gangguan jaringan komputer. Performa diukur berdasarkan tingkat accuracy, sensitivity, precision dan spesificity. Dataset yang digunakan dalam penelitian ini adalah dataset intrusion detection system KDD 99. Reduksi atribut atau seleksi atribut menggunakan *correlation-based feature selection.* Dataset terdiri dari dua yaitu data training dan data testing menggunakan 10 X-Cross Validation.

B. Tinjauan Pustaka

B.1. Intrusion Detection System

Intrusion detection adalah proses memonitor kejadian yang terjadi pada sistem komputer atau jaringan dan menganalisanya untuk menandai kejadian yang mungkin, yang mana yang merupakan pelanggaran atau mendekati penlanggaran sebuah kebijakan keamanan komputer, kebijakan penggunaan yang disetujui atau praktik keamanan standar. Insiden mempunyai banyak penyebab, seperti malware, attacker yang mendapatkan unauthorized access ke sistem melalui Internet, dan pengguna yang sah yang menyalahgunakan hak akses mereka atau usaha untuk mendapatkan hak akses tambahan dimana mereka tidak berhak. Meskipun banyak insiden yang secara alami adalah malicious, banyak kejadian yang lain yang bukan malicious, sebagai contoh, orang yang salah mengetikkan alamat komputer dan usaha yang tidak sengaja untuk menghubungkan ke sistem yang berbeda tanpa hak.

Intrusion detection system adalah sebuah software yang secara otomatis mendeteksi gangguan. Intrusion detection system biasanya digunakan bersama-sama dengan *intrusion prevention system*, yaitu sebuah software yang mempunyai kemampuan seperti *intrusion detection system* dan dapat berusaha menghentikan kejadian yang mungkin. Pada beberapa referensi istilah *intrusion detection and prevention system* (IDPS) digunakan untuk menggantikan keduanya. IDS utamanya fokus pada mengidentifikasi kejadian yang mungkin. Sebagai contoh, IDS dapat mendeteksi seorang attacker secara sukses mengganggu sebuah sistem dengan meneksploitasi kelemahan di dalam sistem. IDS kemudian membuat laporan kepada administrator keamanan, yang dapat secara cepat menginisiasi tindakan tanggapan kejadian untuk meminimalkan kerusakan yang disebabkan oleh kejadian tersebut. IDS dapat juga mencatat informasi yang dapat digunakan oleh administrator keamanan(Scarfone & Mell, Februari, 2007).

Teknologi IDS menggunakan berbagai macam metodologi dalam mendeteksi insiden, yaitu *signature-based detection, anomaly-based detetction* dan *stateful protocol analysis*. Signature-based detection sering juga disebut misuse detection. Signature adalah sebuah pola yang berhubungan dengan ancaman yang sudah diketahui. Signature based detection adalah sebuah proses membandingkan signature dengan kejadian yang diamati untuk mengidentifikasi insiden yang mungkin(Scarfone & Mell, Februari, 2007). Conoth signature adalah sebuah telent yang mencoba dengan username root yang melanggar kebijakan keamanan organisasi, sebuah email dengan subject free picture dan sebuah file lampiran freepics.exe yang tergolong sebagai malware, dan sebuah log entry sistem operasi dengan nilai kode status 645, yang mengindikasikan audit host dinonaktifkan. Signature-based detection sangat efektif mendeteksi ancaman yang sudah diketahui tetapi sangat tidak efektif untuk mendeteksi ancaman yang sebelumnya tidak diketahui, ancaman menyamar dengan teknik pengelabuan, dan banyak variasi ancaman yang sudah diketahui. Sebagai contoh, jika attacker memodifikasi malware pada contoh sebelumnya menggunakan nama file freepics2.exe, sebuah signature mencari freepics.exe tidak akan cocok.

Signature-based detection adalah metode paling sederhana karena metode ini hanya membandingkan unit aktivitas saat ini, seperti paket atau log entry, dengan daftar signature menggunakan operasi perbandingan string. Metodologi signature-based detection memiliki pemahaman yang sedikit tentang jaringan atau protokol aplikasi yang banyak dan tidak dapat melacak dan memahami kondisi komunikasi yang kompleks. Sebagai contoh, metode ini tidak dapat memasangkan permintaan dengan respon yang bersesuaian, seperti mengetahui bahwa permintaan ke web server untuk halaman tertentu dibangkitkan sebuah respon dengan nilai kode status 403, berarti bahwa server menolak memenuhi permintaan. Metode ini juga kurang mampu mengingat permintaan sebelumnya ketika memproses permintaan saat ini. Keterbatasan ini mencegah

metode signature-based detection dari mendeteksi serangan yang terdiri dari banyak kejadian jika tidak ada kejadian berisi indikasi yang jelas sebauh serangan.

Anomaly-based detection adalah proses membandingkan definisi aktivitas yang dikatakan normal dengan kejadian yang diamati untuk mengidentifikasi penyimpangan yang signifikan(Scarfone & Mell, Februari, 2007). IDS yang menggunakan anomaly-based detection mempunyai sebuah profil yang mewakili tingkah laku normal hal-hal seperti user, host, koneksi jaringan, atau aplikasi. Profil dikembangkan dengan memonitor karakteristik aktivitas khusus selama periode tertentu. Sebagai contoh, sebuah profil untuk jaringan mungkin menunjukkan bahwa aktivitas web terdiri atas rata-rata 13% bandwidth jaringan pada batas internet selama beberapa jam hari kerja khusus. IDS menggunakan metode statistik untuk membandingkan karakteristik aktivitas saat ini dengan ambang batas profil yang berhubungan, seperti mendeteksi ketika aktivitas web terdiri dari bandwidth yang lebih signifikan dari pada yang diharapkan dan memberi tanda alarm kepada administrator tentang anomaly. Profil dapat dikembangkan untuk banyak atribut tingkah laku, seperti jumlah email yang dikirimkan oleh seorang user, jumlah login gagal yang dilakukan oleh seorang user dan tingkat penggunaan prosesor untuk host selama periode waktu tertentu.

Keuntungan utama *anomaly-based detection* adalah bahwa metode ini sangat efektif untuk mendeteksi ancaman yang tidak diketahui sebelumnya. Sebagai contoh, andaikan misal sebuah komputer terinfeksi tipe malware baru. Malware bisa mengkonsumsi sumber daya komputer, mengirimkan banyak email, mengawali koneksi jaringan yang banyak, dan menampilkan tingkah laku lain yang sangat berbeda dari profil komputer yang sudah ada. Sebuah profil awal dibangkitkan selama periode waktu (biasanya beberapa hari atau minggu) kadang-kadang disebut *training period*. Profil untuk anomaly-based detection bisa statik atau dinamik. Setelah dibangkitkan, profil statik tidak dapat diubah kecuali IDS secara khusus diarahkan untuk membangkitkan profil baru. Profil dinamik diatur secara konstan ketika kejadian tambahan diamati. Karena sistem dan jaringan berubah sepanjang waktu, pengukuran yang sesuai tingkah laku normal juga berubah; sebuah profil statik bahkan bisa menjadi tidak akurat, sehingga perlu dibangkitkan kembali secara periodik. Profil dinamik tidak memiliki masalah ini, tetapi mereka rentan terhadap usaha pengelabuan dari para attacker. Sebagai contoh, seoreng attacker kadang-kadang menampilkan aktivitas malicious kecil, kemudian dengan perlahan meningkatkan frekuensi dan kuantitas aktivitas. Jika rata-rata perubahan cukup lambat IDS mungkin berpikir bahwa aktivitas malicious tersebut adalah aktivitas normal dan memasukkan ke dalam profil. Aktivitas malicious mungkin juga diamati selagi IDS membangun profil awal.

Metode yang ketiga yang sering digunakan dalam IDS adalah stateful protocol analysis yaitu membandingkan profil yang sudah ditentukan untuk masing-masing kondisi protocol dengan kejadian yang diamati untuk mengidentifikasi penyimpangan(Scarfone & Mell, Februari, 2007). Tidak seperti anomaly-based detetction, yang menggunakan profil khusus host atau jaringan, stateful protocol analysis bersandar pada profil umum yang dikembangkan oleh vendor yang menentukan bagaimana protokol tertentu seharusnya dan tidak seharusnya digunakan. Kata stateful di dalam stateful protocol analysis berarti bahwa IDS mampu memahami dan melacak kondisi jaringan, transport, dan protokol aplikasi yang mempunyai catatan kondisi. Sebagai contoh, ketika user memulai sesi File Transfer Protocol (FTP), sesi diawali pada kondisi unauthenticated. User unauthenticated hanya dapat menampilkan beberapa perintah pada kondisi ini, seperti melihat informasi help atau penyediaan username dan password. Bagian penting memahami kondisi adalah mempasangkan permintaan dan respon, sehingga ketika usaha authentication FTP terjadi, IDS dapat menentukan sukses jika menemukan kode status pada respon yang bersesuaian. Setelah user diautentikasi secara sukses, sesi ada pada kondisi autentikasi dan user bisa menampilkan banyak perintah. Menampilkan sebagian besar perintah ini pada sesi unauthentication dipertimbangkan sebagai mencurigakan, tetapi menampilkan perintah ini pada kondisi authentication dipertimbangkan sebagai jinak atau tidak berbahaya.

Kelamahan utama metode stateful protocol analysis adalah metode ini menggunakan sumber daya komputer yang sangat besar karena kompleksitas analisis dan menampilkan pelacakan kondisi untuk banyak sesi yang berurutan. Permasalahan serius lain adalah metode stateful protocol analiysis tidak dapat mendeteksi serangan yang tidak melanggar karakteristik tingkah laku protokol yang umum disetujui, seperti menampilkan banyak tindakan yang tidak berbahaya selama periode waktu tertentu bisa menyebabkan denial of service.

Ada dua tipe IDS yang paling umum digunakan yaitu *Network-based* dan *Host-based*. *Network-based* memonitor trafik jaringan untuk segmen atau perangkat jaringan tertentu dan menganalisa aktivitas protokol jaringan dan aplikasi untuk mengidentifikasi

aktivitas yang mencurigakan. Perangkat ini diterapkan pada batas antar jairngan, seperti dalam jarak untuk memagari firewall atau router, server VPN, server remote access, dan jaringan wireless. Tipe yang kedua adalah Host-based, yang memonitor karakteristik host tunggal dan kejadian yang terjadi di dalam host tersebut untuk aktivitas yang mencurigakan. Contoh tipe karakteristik IDS host-based bisa memonitor trafik jaringan (hanya untuk host tersebut), system logs, proses yang berjalan, aktivitas aplikasi, modifikasi dan akses file, dan perubahan konfigurasi sistem dan aplikasi. IDS host-based biasanya diterapkan pada host yang kritis seperti server yang bisa diakses publik dan server yang berisi informasi yang sensitif.

B.2. Algoritma Naïve Bayes

Algoritma naïve bayes merupakan algoritma yang menggunakan pendekatan statistik dalam mengambil keputusan. Algoritma naïve bayes berdasarkan teorema bayes bahwa semua atribut memberikan kontribusi yang sama penting dan saling bebas pada kelas tertentu(Witten et al., 2011). Walaupun teori ini bertentangan dengan kehidupan nyata bahwa atribut tidak sama penting atau independen, tetapi naïve bayes menunjukkan performa yang mampu bersaing dengan algoritma klasifikasi yang terkenal, decision tree dan neural network(Mitchell, 1997).

Algoritma naïve bayes menggunakan perhitungan probabilitas dalam menentukan kelas. Naïve bayes diterapkan pada machine learning dimana masing-masing instance dideskripsikan oleh konjungsi nilai atribut dan dimana fungsi target dapat mengambil nilai apapun dari beberapa perangkat kelas C. seperangkat training example fungsi target disediakan dan instance baru dihadirkan, dideskripsikan oleh tuple nilai atribut (a_1, a_2, …, a_n). Algoritma ditugaskan untuk memprediksi nilai target, atau klasifikasi untuk instance baru ini(Mitchell, 1997). Perhitungan probabilitas pada naïve bayes dirumuskan

$P(x,c)=P(c) \prod_{i=1}^{p} p(x_i|C)$ …… (1)

Persamaan (1) dapat ditulis (Kusumadewi, 2009)

$P(x_{1, \ldots, x_k})|C=P(x_1|C)$ x….x $P(x_k|C)$ ….. (2)

Table 1. Weather data dengan jumlah dan probabilitas

Outlook			Temperatur			Humidity			Windy			Play	
	yes	No		yes	no		yes	no		yes	no	yes	no
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cood	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cood	3/9	1/5								

Sebagai contoh tabel 1 berisi ringkasan data cuaca yang diperoleh dengan menghitung berapa banyak masing-masing nilai atribut berpasangan dengan masing-masing nilai (yes or no) kelas play pada dataset cuaca. Nilai-nilai pada tabel 1 didapatkan dari tabel 2 dengan menghitung jumlah masing-masing nilai atribut untuk masing-masing kelas. Sebagai contoh pada tabel 2 atribut outlook bernilai sunny ada lima instance, dimana dua instance menajdi kelas play=yes dan tiga instance menjadi kelas play=no. Pada baris pertama tabel 1 berisi jumlah kejadian untuk masing-masing nilai atribut yang mungkin, dan kolom play pada kolom terakhir tabel 1 jumlah total kejadian yes atau no. Bagian di bawahnya berisi bilangan pecahan atau probabilitas yang diamati. Sebagai contoh pada tabel 2 diiketahui sembilan instance menjadi kelas play=yes, nilai atribut outlook=sunny ada dua instance, menghasilkan bilangan pecahan 2/9 pada tabel 1.

Untuk atribut play nilai bilangan pecahanya berbeda, yaitu perbandingan antara instance atribut play bernilai yes atau no secara berurutan.

Table 2. Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Misalkan diberikan instance baru seperti pada tabel 3, apakah keputusan yang diambil bermain atau tidak? Apakah nilai atribut play, yes atau no? Yes berarti bermain no berarti tidak bermain. Ada lima atribut yang sama penting dan saling bebas seperti ditunjukkan pada tabel 2, yaitu outlook, temperature, humidity, windy dan play. Untuk mencari kemungkinan bernilai yes

Table 3. Instance baru

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

Kemungkinan yes = 2/9 x 3/9 x 3/9 x 3/9 x 9/14 = 0,0053

Bilangan pecahan tersebut diambil dari nilai kelas play=yes sesuai dengan nilai atribut instance yang baru, dan bilangan terakhir 9/14 adalah bilangan pecahan keseluruhan yang mewakili perbandingan instance dimana play=yes. Perhitungan yang mirip untuk hasil no adalah

Kemungkinan no = 3/5 x 1/5 x 4/5 x 3/5 x 5/14 = 0,0206

Hal ini mengindikasikan bahwa instance yang baru lebih bernilai no dari pada yes, empat kali lebih mungkin. Angka tersebut bisa diubah menjadi probabilitas:

Probabilitas yes = $\frac{0,0053}{0,0053+0,0206} = 20,5\%$

Probabilitas no = $\frac{0,0206}{0,0206+0,0053} = 79,5\%$

Perhitungan tersebut adalah metode sederhana naïve bayes yang terbukti mampu bersaing dengan algoritma klasifikasi decision tree dan neural network. Berdasarkan perhitungan di atas dapat dirumuskan probabilitas menggunakan algoritma naïve bayes

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]} \quad \text{.....(3)}$$

B.3. Seleksi Atribut Berbasis Korelasi

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 2^N 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)

$$r_{zc} = \frac{kr_{zi}}{\sqrt{k+k(k-1)rii}} \quad \text{.....(4)}$$

Dimana r_{zc} adalah korelasi antara jumlah komponen dan variabel luar (kelas), k adalah jumlah komponen atau atribut, r_{zi} adalah rata-rata korelasi antara komponen dan variabel luar (kelas), dan r_{ii} adalah rata-rata korelasi antar komponen.

Berdasarkan rumus (4) dapat diketahui korelasi yang lebih tinggi antara komponen-komponen dan variabel luar (kelas), korelasi yang lebih tinggi antara atribut subset dengan variabel luar (kelas). Korelasi yang lebih rendah antar komponen menunjukkan korelasi yang lebih tinggi antara subset atribut dengan variabel luar (kelas). Jumlah komponen di dalam subset atribut meningkat menunjukkan korelasi antara komponen di dalam subset dan variabel luar (kelas) meningkat dengan asumsi komponen tambahan sama dengan komponen orisinal dalam hal korelasi antar atribut dan variable luar.

Pengukuran korelasi bisa menggunakan symetrical uncertainty, relief, minimum description length ataupun pengukuran korelasi yang lain misal korelasi person. Pada penelitian ini menggunakan pengukuran korelasi menggunakan *symetrical uncertainty*. Entropy adalah ukuran uncertainty atau unpredictability pada sebuah system. Entropy Y dirumuskan

$H(Y) = -\sum_{y \in Y} p(y) log_2(p(y))$ .....(5)

Misalnya diberikan atribut X dan Y pada data training, hubungan antara feature Y dan X dirumuskan

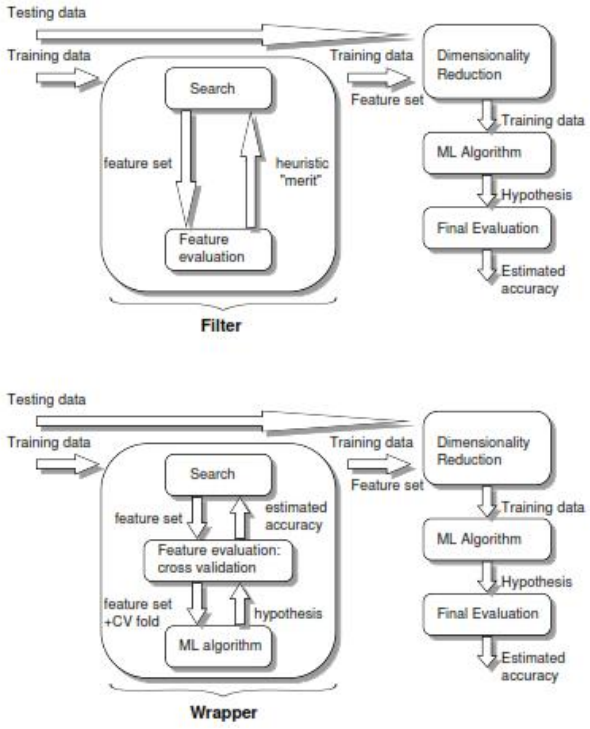
$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2(p(y|x))$ .....(6)

Jumlah entropi Y yang mana yang menurun merefleksikan informasi tambahan tentang Y yang disediakan oleh X dan disebut information gain, dirumuskan

Gain = H(Y) – H(Y|X) ……(7)
= H(X) – H(X|Y)
= H(Y) + H(X) – H(H,Y)

Information gain adalah ukuran simetris, yaitu jumlah informasi yang diperoleh Y setelah mengamatai X adalah sama dengan jumlah informasi yang diperoleh X setelah mengamati Y. Simetrical adalah properti yang diinginkan untuk mengukur feature-feature yang saling berkorelasi. Rumus untuk menghitung koefisien symetrical uncertainty

Symetrical uncertainty = $2\ x\ \left[\frac{gain}{H(Y)+H(X)}\right]$……(8)

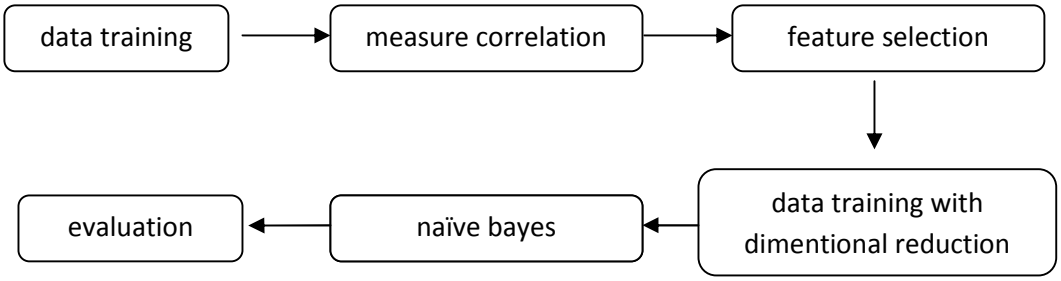


Gambar 1. Pemilihan atribut menggunakan filter dan wrapper (Hall, 1999)

C. Metode Penelitian

Penelitian ini adalah penelitian eksperimen. Penelitian ini menyajikan penerapan naïve bayes dengan seleksi atribut berbasis korelasi (*correlation-based feature selection*) dalam mendeteksi gangguan jaringan komputer. Performa diukur berdasarkan

tingkat accuracy, sensitivity, precision dan spesificity. Dataset yang digunakan dalam penelitian ini adalah dataset intrusion detection system KDD 99. Reduksi atribut atau seleksi atribut menggunakan *correlation-based feature selection.* Pengukuran korelasi menggunakan symetrical uncertainty. Dataset terdiri dari dua yaitu data training dan data testing menggunakan 10 X-Cross Validation, dimana data dibagi menjadi sepuluh bagian sama besar, sembilan bagian dijadikan data training dan satu bagian menjadi data testing. Begitu seterusnya sampai masing-masing bagian menjadi data testing. Nilai akursai di dapat dari rata-rata akurasi masing-masing bagian.



Gambar 2. Alur Penelitian

D. Hasil dan Pembahasan

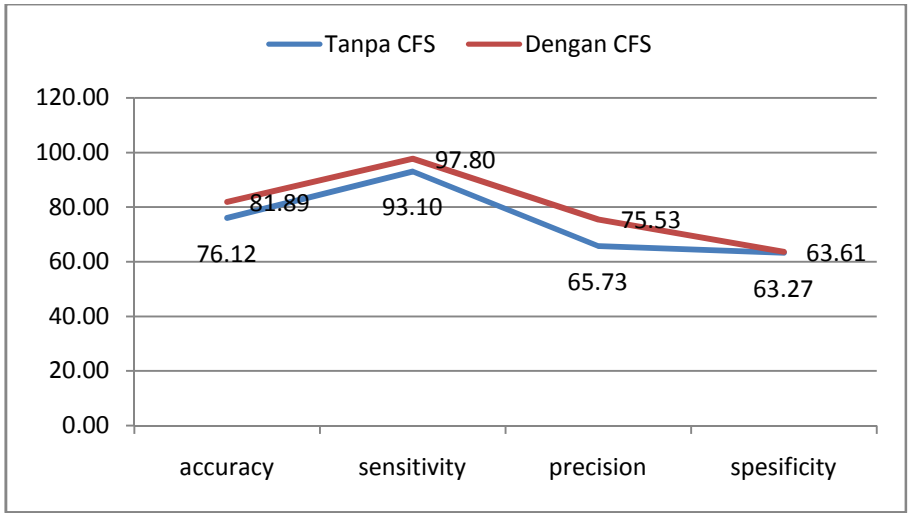
Eksperimen dilakukan dengan menerapakan correlation-based feature selection atau seleksi atribut berbasis korelasi yang selanjutnya disebut CFS pada dataset *intrusion detection system* (Bache & Lichman, 2013). Masing-masing atribut dihitung korelasinya terhadap kelas menggunakan persamaan 8. Selanjutnya dipilih subset atau kumpulan beberapa atribut yang memiliki merit tertinggi. Merit dihitung menggunakan persamaan 4. Sebelum menghitung merit terlebih dahulu menghitung rata-rata korelasi atribut dalam subset dengan atribut kelas dan rata-rata korelasi antar atribut. Atribut yang memiliki korelasi yang tinggi dengan atribut yang lain dalam satu subset, atribut tersebut harus duhilangkan karena atribut tersebut redundan. Atribut yang memiliki korelasi yang rendah terhadap kelas, atribut tersebut juga harus dihilangkan karena atribut tidak relevan. Untuk memilih subset atau sekumpuan atribut yang mempunyai merit tertinggi menggunakan algoritma *forward selection search* (fss) atau *backward selection search* (bfs). FSS mencari subset dengan mula-mula atribut 0 kemudian ditambahkan satu atribut. Kemudian menghitung merit, dan seterusnya sampai semua atribut diujicoba. Kemudian dipilih satu subset atau kumpulan atribut yang mempunyai merit tertinggi. Subset tersebut adalah hasil dari seleksi atribut berbasis korelasi (CFS). BFS mencari subset dengan mula-mula seluruh atribut kemudian satu persatu atribut dikurangi, dan dihitung meritnya. Begitu seterusnya sampai atribut 0. Kemudian dipilih subset dengan merit tertinggi. Berdasarkan eksperimen diperoleh subset dengan enam atribut yang merupakan subset dengan merit tertinggi. Subset tersebut terdiri dari atribut flag, src_bytes, dst_bytes, logged_in, srv_serror_rate, diff_srv_rate serta satu atribut kelas dari semual sejumlah 41 atribut dan satu atribut kelas. Selanjutnya dataset yang terdiri dari enam atribut hasil seleksi atribut ini diujicobakan pada algoritma naïve bayes untuk menentukan kelas gangguan jaringan komputer atau bukan gangguan jaringan komputer.

Berdasarkan hasil eksperimen seperti ditunjukkan pada tabel 4, menunjukkan bahwa performa algoritma naïve bayes dengan seleksi atribut berbasis korelasi (CFS) lebih baik dibandingakn dengan algoritma naïve bayes tanpa seleksi atribut. Jumlah atribut setelah diterapkan seleksi atribut CFS sebanyak enam atribut, yaitu. Akurasi naïve bayes tanpa seleksi atribut CFS sebesar 76,12 % sedangkan akurasi naïve bayes dengan seleksi atribut CFS sebesar 81,89%. Sensitivity naïve bayes tanpa seleksi atribut CFS sebesar 93,10% sedangkan naïve bayes dengan seleksi atribut CFS sebesar 97,80%. Precision naïve bayes tanpa seleksi atribut CFS 65,73%, dengan seleksi atribut CFS 75,53%. Spesificity naïve bayes tanpa seleksi atribut CFS sebesar 63,27%, naïve bayes dengan seleksi atribut CFS 63,61%. Dari keempat kriteria hanya spesificity yang mempunyai perbedaan nyang paling sedikit, lainya menunjukkan perbedaaan yang signifikan.

Table 4. Performa algoritma naïve bayes pada KDD dataset

	accuracy	sensitivity	precision	spesificity
Tanpa CFS	76.12	93.10	65.73	63.27
Dengan CFS	81.89	97.80	75.53	63.61

Berdasarkan tabel 4 di atas diketahui bahwa akurasi naïve bayes dengan CFS lebih besar dibandingkan akurasi naïve bayes tanpa CFS. Hal ini menunjukkan bahwa jumlah jumlah instance yang diprediksi dengan benar oleh algoritma naïve bayes dengan CFS lebih besar dibandingkan dengan naïve bayes tanpa CFS. Dengan kata lain, jumlah true posiitif dan true negatif algoritma NB dengan CFS lebih besar dibandingkan jumlah TP dan TN algoritma NB tanpa CFS. Sensitivity NB dengan CFS lebih besar dibandingkan sensitivity NB tanpa CFS, keduanya menunjukkan mendekati angka 100%. Hal ini menunjukkan jumlah true positif jauh lebih besar dibandingkan jumlah false negatif. Walaupun proporsi TP lebih besar dibandingkan FN secara signifikan namun kenyataanya jumlah instance yang secara aktual positif tetapi diprediksi negatif (FN) cukup besar karena dataset intrusion detection system ini memiliki jumlah instance yang sangat besar (ratusan ribu instance). Di dunia nyata false negatif ini artinya ada sebuah kejadian yang sebenarnya adalah sebuah gangguan jaringan komputer atau serangan tetapi diprediksi bukan serangan jaringan. Hal ini yang menyebabkan gagalnya IDS dalam mendeteksi serangan. Sedangkan false positif adalah kejadian yang sebenarnya bukan serangan (attack) tetapi oleh naïve bayes dianggap sebagai serangan. Jumlah ini cukup besar seperti terlihat pada tabel 4, pada kolom precision. Pada kriteria spesificity tidak terdapat perbedaan yang mencolok antara NB dengan CFS dan NB tanpa CFS. Hal ini menunjukkan proporsi antara true negatif dan false positif tidak berbeda secara signifikan antara NB dengan CFS dan NB tanpa CFS. Namun kedua memiliki nilai pada ngka 60-an, artinya perbedaan jumlah instance true negatif dan false positif tidak terlalu besar.



Gambar 3. Grafik perbandingan algoritma naïve bayes

Seleksi atribut berbasis korelasi atau CFS mampu meningkatkan performa algoritma naïve bayes pada dataset Intrusion detection System. Area di bawah kurva ROC atau AUC naïve bayes dengan seleksi atribut CFS yaitu sbesar 0,94 lebih besar dibandingkan dengan naïve bayes tanpa seleksi atribut CFS yaitu sebesar 0,91. Selain meningkatkan AUC, waktu eksekusi naïve bayes dengan seleksi atribut CFS lebih singkat dibanding waktu eksekusi naïve bayes tanpa seleksi atribut CFS, yaitu 1,03 detik berbanding 10,86 detik. Sehingga dikatakan wktu eksekusi naïve bayes dengan seleksi atribut CFS lebih cepat sepuluh kali lipat dibandingkan waktu eksekusi naïve bayes tanpa seleksi atribut CFS. Seleksi atribut berbasis korelasi (CFS) mampu menyederhanakan sistem yang tadinya suatu sistem dengan 41 atribut disederhanakan menjadi enam atribut tanpa mengurangi tingkat akurasinya.

E. Kesimpulan

Seleksi atribut berbasis korelasi atau CFS mampu meningkatkan performa algoritma naïve bayes pada dataset Intrusion detection System. Selain meningkatkan performa, seleksi atribut CFS juga mempersingkat waktu eksekusi sepuluh kali lipat lebih cepat dibandingkan tanpa seleksi atribut CFS. Penelitian selanjutnya bisa membandingkan hasil penelitian ini dengan dataset yang lain maupun dengan metode seleksi atribut yang lain misal relief, chi square ataupun dengan menggunakan metode wrapper feature selection.

F. Daftar Pustaka

Bache, K. & Lichman, M., 2013. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. [Online] Available at: http://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/ [Accessed 22 Juli 2013].

Biesiada, J. & Duch, W., 2007. Feature Selection for High-Dimentional Data - A Person Redundancy Based Filter. *Advances In Soft Computing*, 45, pp.242-49.

Gostev, A. & Namestnikov, Y., 2011. *Kaspersky Security Bulletin 2010. Statistics, 2010*. [Online] Available at: http://www.securelist.com/en/analysis/204792162/Kaspersky_Security_Bulletin_2010_Statistics_2010 [Accessed 6 Juni 2013].

Hall, M., 1999. *Correlation-based Feature Selection for Machine Learning*. Ph.D. dissertation. Hamilton, New Zealand: Ph.D Thesis University Of waikato.

Kohavi, R. & John, G.H., 1997. Wrapper for Subset Feature Selection. *Artificial Inteligence*, pp.273-324.

Kusumadewi, S., 2009. Klasifikasi Status Gizi Menggunakan Naive Bayesian Clasification. *Jurnal CommIT*, 3(01).

Langley, P., 1994. Selection of Relevant Feature in Machine Learning. In *AAAI Fall Symposium on Relevance*. Los Angeles, 1994. AAAI.

Leng, J., Valli, C. & Armstrong, L., 2010. A Wrapper-Based Feature Selection For Analysis Of Large Dataset. In *Proceeding Of 2010 3rd International Conference On Computer And Electrical Engineering*., 2010. ECU Publications.

Marcelloni, F., 2003. Feature Selection Based On A Modified Fuzzy C-Means Algorithm With Supervision. *Information Sciences*, 151, pp.201-26.

Mitchell, T., 1997. *Machine Learning*. New York: McGraw Hill.

Neethu, B., 2012. Classification of Intrusion Detection Dataset Using Machine Learning Approaches. *International Journal of Electronics and Computer Science Engineering*, pp.1044-51.

Olusola, A.A., Oladele, A.S. & Abosede, D.O., 2010. Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Feature. In *World Congress on Engineering and Computer Science 2010*. San Fransisco, 2010. WCECS 2010.

Pedrycz, W. & Vukovich, G., 2001. Feature Analysis Through Information Granulation and Fuzzy Sets. *Pattern Recognition*, 35, pp.825-34.

Scarfone, K. & Mell, P., Februari, 2007. *Special Publication 800-94: Guide To Intrusion Detection and Prevention Systems*. Gaithersburg, Maryland: National Institute Standard and Technology.

Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani, A.A., 2009. A Detailed Analysis Of The KDD Cup 99 Data Set. In *Proceedings Of The 2009 IEEE Symposium On Computational Inteligence in Security and Defense Application (CISDA)*. Otawa, 2009. IEEE Press Piscataway, NJ, USA.

Witten, I.H., Frank, E. & Hall, M.A., 2011. *Data Mining Practical Machine Learning Tools and Technique Third Edition*. New York: Morgan Kaufmann.

Yu, L. & Huan, L., 2003. Feature Selection for High Dimentional Data: A Fast Correlation-Based Filter Solution. In *Proceeding of the Twentieth International Conference on Machine Learning (ICML-2003)*. Washington DC, 2003.