ANOMALY DETECTION PADA INTRUSION DETECTION SYSTEM MENGGUNAKAN METODE BAYESIAN NETWORK

Oktavia Ari Marlita[1], Adiwijaya[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

. Intrusion Detection System (IDS) merupakan sistem yang dapat mendeteksi adanya intrusi atau gangguan pada suatu jaringan atau sistem informasi. Salah satu jenis IDS adalah anomaly detection dimana suatu data trafik jaringan akan dikatakan intrusi apabila mempunyai karakteristik yang berbeda dari kebanyakan data lainnya. Anomaly detection dapat mendeteksi serangan dalam host atau network yang menyimpang dari aktivitas normal berdasarkan probabilitas statistika. Statistical anomaly detection tidak memiliki model intelligent learning yang mungkin menyebabkan false alarm memiliki tingkat deteksi tinggi. Metode yang digunakan sistem adalah Bayesian Network TAN Classifier. CI Test merupakan salah satu algoritma yang handal untuk membangun Model TAN Classifier untuk klasifikasi. Dengan representasi grafis gabungan dari probabilitas fungsi distribusi lebih dari satu set variabel. Struktur jaringan direpresentasikan sebagai Directed Acyclic Graph (DAG) di mana setiap node berkorespondensi dengan variabel acak dan setiap edge menunjukkan hubungan ketergantungan antar variabel-variabel yang terhubung. Data trafik jaringan melalui proses training untuk membentuk model TAN Classifier, kemudian dilakukan proses testing. Pengujian dilakukan dengan beberapa skenario untuk mengetahui akurasi sistem dilihat dari nilai detection rate (DR) dan false positive rate (FPR), pengaruh proporsi data training dan data testing, serta pengaruh proporsi data normal dan data intrusi pada masing-masing dataset. Bayesian Network dapat mendeteksi intrusi, dengan DR sebesar 97,88%dan FPR 6,11%.

Kata Kunci : intrusi, Intrusion Detection System, anomaly detection, Bayesian Network, TAN Classifier

Abstract

Intrusion Detection System (IDS) is a system that can detect any intrusion or interference on a network or information systems. One type of IDS is anomaly detection in which a data network traffic if the intrusion would be said to have characteristics different from most other data. Anomaly detection can detect the attack on host or network who deviate from normal activities based on statistical probability. Statistical anomaly detection does not have a model of intelligent learning that may lead to false alarms have a high detection rate. The method used is a system of TAN Bayesian Network Classifier. CI Test is one of a reliable algorithm to build classification models for the TAN Classifier. With a graphical representation of the joint probability distribution function over a set of variables. The network structure is represented as a Directed acyclic Graph (DAG) where each node corresponds to a random variable and each edge shows the dependency relationships among the variables are connected. Data network traffic through the training process to form a model of TAN Classifier, then do the testing. Tests carried out with several scenarios to determine the accuracy of the detection system is seen rate (DR) and false positive rate (FPR), the influence of the proportion of training data and testing the data, as well as the influence of the proportion of normal data and data intrusion on each dataset. Bayesian Network can detect the intrusion, the DR of 97.88% and 6.11% FPR.

Keywords : intrusi, Intrusion Detection System, anomaly detection, Bayesian Network, TAN Classifier

1. Pendahuluan

1.1 Latar belakang

Data dan informasi sensitif dalam skala besar diproses dalam jaringan komputer. Sehingga diperlukan suatu sistem keamanan dalam jaringan komputer yang tahan dan toleran terhadap intrusi jaringan. Intrusi jaringan merupakan upaya untuk mendapatkan akses ilegal ke *resource* jaringan atau mem-*by-pass* sistem keamanan yang ada. Upaya itu dapat dilakukan dengan mengakses sistem melalui Internet, maupun melalui celah keamanan jaringan lainnya. Oleh karena itu, *Intrusion Detection System* (IDS) diperlukan untuk mengatasi permasalahan tersebut. *Intrusion detection* adalah suatu proses *monitoring* kejadian yang terjadi pada sistem komputer atau jaringan serta menganalisisnya untuk mengetahui aktivitas tersebut termasuk normal atau intrusi.

Dalam model proses IDS terdapat tiga fungsi dasar. Pertama, pengambilan data dari berbagai *level* dari sistem seperti *network, host,*dan *application.* Kedua, analisis data yang diperoleh untuk mengenali intrusi. Pendekatan yang sering digunakan untuk mengenali intrusi yaitu *anomaly detection* dan *misuse detection/signature analysis*. Ketiga, respon terhadap serangan. Ada dua macam: respon aktif dalam hal ini berarti melakukan beberapa aksi secara otomatis untuk mengintervensi sistem yang ada, sedangkan pasif adalah memberikan *report* pada administrator yang akan melakukan respon terhadap sistem.

Anomaly detection dapat mendeteksi serangan dalam host atau network yang menyimpang dari aktivitas normal berdasarkan probabilitas statistika. *Statistical anomaly detection* tidak memiliki model *intelligent learning* yang mungkin menyebabkan *false alarm* memiliki tingkat deteksi tinggi. *Anomaly detection* menggunakan pendekatan *unsupervised learning(*tidak ada *intelligent learning)*, yang mampu mendeteksi intrusi tanpa harus mempelajari data sebelumnya[9]. *Misuse detection/signature analysis* membutuhkan akses ke database besar dari *signature intrusion* yang diketahui. Detektor melakukan analisis terhadap aktivitas sistem, mencari *event* atau set *event* yang cocok dengan pola perilaku yang dikenali sebagai serangan.

Dalam tugas akhir ini digunakan pendekatan metode *anomaly detection*. *Anomaly detection* yang merupakan salah satu tugas dari proses data *mining*, diharapkan dapat menemukan objek yang berbeda dari kebanyakan objek yang ada. Seringkali objek anomali tersebut dikatakan sebagai *outlier* atau pencilan. Metode deteksi anomali dapat mendeteksi serangan berdasarkan statistik probabilitas, yang memungkinkan untuk generalisasi dan membantu dalam pendeteksian instrusi baru. Namun, statistik deteksi anomali tidak didasarkan pada *adaptive intelligent model* dan tidak bisa dipelajari dari pola aktivitas normal maupun *malicious traffic pattern (unsupervised learning)* [5].

Bayesian Network (BN) adalah representasi grafis dari gabungan dari probabilitas fungsi distribusi lebih dari satu set variabel. Struktur jaringan direpresentasikan sebagai *Directed Acyclic Graph* (DAG) di mana setiap *node* berkorespondensi dengan variabel acak dan setiap *edge* menunjukkan hubungan ketergantunga antar variabel-variabel yang terhubung [5]. Setiap variabel memiliki satu finite set *mutually exclusive states.*

Dalam Tugas Akhir ini metode *Bayesian Network* yang digunakan yaitu *Tree Augmented Naive Bayes* (TAN) *classifier*. Klasifikasi dengan TAN memiliki beberapa kelebihan dalam pemodelan data. Pertama, TAN memodelkan semua *dependencies* antar variabel, tujuannya agar mudah dalam menangani kasus beberapa entri data yang hilang. Kedua, TAN dapat digunakan untuk mempelajari hubungan *causal*, sehingga dapat digunakan untuk memperoleh pemahaman tentang domain masalah dan memprediksi konsekuensi dari intervensi. Ketiga, karena dimodelkan dengan *causal* dan semantik probabilistic maka akan membentuk representasi ideal untuk menggabungkan *prior knowledge* (yang sering datang dalam bentuk *causal*) dan data. Metode statistik Bayesian dan *Bayesian Network* merupakan pendekatan yang efisien dan berprinsip untuk menghindari overfitting data [14]. Sistem Bayesian mempunyai dasar matematika yang kuat dalam menangani implementasi IDS. *Bayesian Network* IDS harus membedakan antara intrusi/serangan dan aktivitas normal jaringan dengan membandingkan *metric* dari setiap *network traffic sample*.

Adaptive network IDS dengan medote BN mengambil data *offline* dari dataset KDD Cup tahun 1999 untuk mengukur kelayakan dan efektivitas sistem. Tipe datanya ada kontinu dan nominal dengan mengambil 9 dari 41 *feature*s termasuk *unbalanced* data[12]. Karena setiap variabel (*node*) dalam BN dikaitkan dengan *Conditional Probability Tabel* (CPT), yang menyebutkan probabilitas bersyarat untuk variabel yang memberikan semua kombinasi induknya. Akurasi deteksi intrusi dapat dilihat dari seberapa dekat kah data klasifikasi dengan data asli. Dengan BN sebagai model yang ideal untuk menggabungkan *prior knowledge* sebelumnya dengan data baru dan menyimpulkan menjadi *posterior knowledge*. Oleh karena itu, *Bayesian Network* dengan TAN *classifier* diharapkan dapat meningkatkan tingkat akurasi IDS.

1.2 Perumusan masalah

Dengan mengacu pada latar belakang masalah, maka permasalahan yang dibahas dan diteliti adalah :

1. Bagaimana mengimplementasikan metode *Bayesian Network* untuk deteksi anomali pada IDS.
2. Bagaimana mengevaluasi performansi dari metode *Bayesian Network* untuk mendeteksi terjadinya anomali dengan kasus data intrusi.

1.3 Batasan masalah

Batasan masalah yang digunakan dalam penelitian ini antara lain:

1. Pendekatan yang digunakan adalah pendekatan analisis deteksi anomali, dengan menggunakan metode *Bayesian Network* untuk menganalisa.
2. Data yang digunakan adalah *network connection record*.
3. Menggunakan data *offline*, yaitu data yang digunakan pada KDD Cup tahun 1999.
4. Menggunakan tools Microsoft Excel untuk *preprocessing* data.
5. Tidak dapat digunakan untuk mendeteksi intrusi yang dilakukan oleh pihak yang mempunyai akses dalam sistem komputer tersebut, dan intrusi yang dilakukan dengan menyerupai data normal.
6. Data set yang digunakan untuk dianalisis tidak boleh didominasi oleh intrusi.

1.4 Tujuan

Berdasarkan rumusan masalah, maka tujuan dari tugas akhir ini adalah:

1. Mengimplementasikan metode *Bayesian Network* dengan TAN *classifier* untuk mendeteksi anomali yang terjadi dalam suatu jaringan.
2. Menganalisis performansi deteksi anomali dengan metode *Bayesian Network* berdasarkan prediksi nomal (*true negative rate* dan *false negative rate)* dan prediksi intrusi *(true positive rate* dan *false positive rate).*

1.5 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Studi Literatur :
   a. Pencarian referensi, dan sumber-sumber lain yang berhubungan dengan data *mining, Bayesian Network, TAN classifier, network connection record, intrusion,* jaringan komputer dan pengukuran evaluasi pada IDS(*Intrusion Detection System*).
   b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir.
2. Melakukan analisis dan desain perangkat lunak yang dibangun. Perancangan akan dimulai dengan membangun skema / *flowchart* tentang alur sistem keseluruhan yang akan dibangun. Dilanjutkan dengan mencari kebutuhan-kebutuhan yang diperlukan oleh perangkat lunak dan sistem. Kemudian, mempersiapkan data yang akan diolah termasuk data *prepocessing* didalamnya.
3. Melakukan implementasi perangkat lunak mengunakan Matlab R2008a.
4. Melakukan analisis performansi dari hasil pengujian perangkat lunak yang telah diciptakan dengan melihat tingkat *detection rate* dan *false positive rate*.
5. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. Penutup

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada Tugas Akhir ini, diperoleh beberapa kesimpulan sebagai berikut:

1. Faktor yang mempengaruhi pembentukan model TAN *Classifier* yaitu jumlah *record*, jumlah atribut, dan jumlah *value* tiap atribut pada data *training*. Karena pada TAN *Classifier* melibatkan ketergantungan antar atribut yang dibangun menggunakan Algoritma *CI Test*.
2. Proporsi data *training* dan data *testing* yang optimal untuk membangun model TAN *Classifier* dengan hasil performansi yang optimal yaitu dengan proporsi 90% data *training* dan 10% data *testing*. lebih besar dari 50%. Semakin banyak data training semakin optimal Bayesian Network yang terbentuk.
3. Pengambilan sample untuk proporsi data normal dan data intrusi pada data *training* sangat berpengaruh terhadap pembentukan model TAN *Classifier*. Secara umum dapat diambil kesimpulan bahwa proporsi data normal pada *Anomaly detection* ini haruslah lebih besar dari data intrusi. Dari hasil pengujian proporsi data yang menghasilkan performansi optimal yaitu 80% data normal dan 20% data intrusi. Performansi sistemnya yaitu nilai DR 100% dan FPR 0%.
4. Algoritma *Bayesian Network* TAN Classifier bisa diimplementasikan untuk *anomaly detection* pada IDS dengan performansi yang baik, yaitu dari detection rate 100% dan false alarm rate 2%. Hal ini dikarenakan model klasifikasi pada TAN dibangun menggunakan *CI Test Based Algorithms* menghasilkan nilai akurasi yang rata-rata lebih tinggi di bandingkan dengan *Bayesian Network* biasa (misalnya *Naïve Bayes*). Hal ini menunjukan pengaruh ketergantungan diantara atribut pada TAN dapat menaikkan nilai akurasi jika di bandingkan dengan *Bayesian Network* lain yang hanya mempunyai ketergantungan setiap atribut dengan kelasnya saja.

5.2 Saran

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada Tugas Akhir ini, ada beberapa saran diantaranya:

1. Mencoba mencari metode konversi data *symbolic* ke *numeric* untuk metode yang statistical based. Karena pada sistem Tugas Akhir ini tidak dilakukan pengkonversian data karena dirasa belum perlu. Namun ada baiknya dicoba untuk penelitian yang lain. Misalnya metode *conditional probabilities* atau *SSV (separability split value) critetion.*
2. Dapat dikembangkan dengan pembangunan model klasifikasi menggunakan algoritma yang lain seperti algoritma *Scoring and Searching* pada saat penentuan model graf untuk TAN *classifier.*
3. Mencoba mencari metode *feature selection* selain dengan *ranking Information Gain*, misalnya *Chi Square* atau *PCA.*

Daftar Pustaka

[1]	Afianti Mira, 2011, "Implementasi Algoritma Y-Means sebagai *Anomaly Detection* (Studi Kasus:*Intrusion Detection System*)", IT Telkom Bandung.
[2]	Amanda Delamer,2002,"*Intrusion Detection with* Data *Mining*" Donau-Universität Krems, Dublin.
[3]	Baesens, B., M. Egmont Petersen., R. Castelo., J. Vanthienen. "Learning *Bayesian Network Classifier*s for Credit Scoring using Markov Chain Monte Carlo Search". K.U.Leuven Dept. of Applied Economic Sciences Naamsestraat, Leuven, Belgium. *www.cs.uu.nl/re**search**/techreps/repo/CS-2001/2001-58.pdf.*
[4]	Bringas, Pablo G. dan Igor Santos. *Bayesian Networks for Network Intrusion Detection,*
[5]	Cemerlic Alma, Li Yang, Joseph M. Kizza. *Network Intrusion Detection Based on Bayesian Networks.* Diakses pada 21 Maret 2011 di http://www.utc.edu/Faculty/Li-Yang/MyPaper/SEKE08-Cemerlic-Yang.pdf
[6]	Charles River Analytics, Inc, 2004, "About Bayesian Belief Networks", Cambridge. www.cra.com. Page 2. https://www.cra.com/pdf/BNetBuilderBackground.pdf.
[7]	Cheng, Jie, dkk, "An Algorithms for Bayesian Belief Network Construction from Data". School of Information and Software Engineering University Ulster. Northern Ireland.
[8]	Chia-Ping Chen, "Entropy and Mutual Information Notes on Information Theory", Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan ROC.
[9]	Firmansyah, Ivan Suci. IP *Network-Packet Shared Media* pada Mesin Cluster *Intrusion Detection System*. diakses pada 24 Maret 2011 http://budi.insan.co.id/courses/el695/projects2002-2003/ivan-report.pdf
[10]	Fradhany Yustiar, 2008, "Learning Klasifikasi *Bayesian Network* Menggunakan Algoritma Conditional Independence *Test*", IT Telkom Bandung.
[11]	Ghorbani A., Guan Yu, dkk. 2003. "Y-Means: A *Clustering* Method for Intrusion Detection", *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*. Montreal, Canada. pp 87-99.
[12]	H. Güneş Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood "Selecting *Feature*s for Intrusion Detection:A *Feature* Relevance Analysis on KDD 99 Intrusion Detection Datasets", Dalhousie University.
[13]	Heckerman, David, 1995, "A Tutorial on Learning With *Bayesian Network*s", Advanced Technology Division. Microsoft Corporation.
[14]	Heckerman, David. *Bayesian Networks for* Data *Mining*. 1997 diakses pada 21 Maret 2011 di http://www.springerlink.com
[15]	Hernández- Pereira E.,Suárez-RomeroJ. A., dkk. 2009. "Conversion methods for symbolic *features*: A comparison applied to an intrusion detection problem", *Expert System With Applications*, Vol. 36(2009) 10612-10617.

[16]	Jiang, Liangxiao, Harry Zhang, Jiang Su, "Learning Tree Augmented Naïve Bayes for Ranking", Department of Computer Science, China University of Geosciences. Wuhan, China. www.ai.mit.edu/projects/jmlr/papers/volume3/ling02a/top.pdf
[17]	Jiawei Han, Micheline Kamber, 2001, "Data Mining : Concepts and Techniques", Simon Fraser University.
[18]	Kannan, Sivanadiyan Sabari. 2005.*Y-Means Clustering Vs N-CP Clustering With Canopies for Intrusion Detection.* Thesis. Oklahoma State University.
[19]	Leung, Kingsly & Christopher Leckie. *Unsupervised Anomaly Detection in Network Intrusion Detection Using Cluster.*
[20]	M. Tavallaee, E. Bagheri, W. Lu, dan A. Ghorbani. 2009. "A Detailed Analysis of the KDD CUP 99 Data Set". *Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA).*
[21]	Pang-Ning Tan, Vipin Kumar, Michael Steinbach, 2004, "Introduction to Data Mining", Michigan State University, University of Minnesota.
[22]	Munir, Rinaldi. 2008. Diktat Kuliah IF2091 Struktur Diskrit. Program Studi Teknik Informatika, Sekolah Tinggi Teknik Elektro dan Informatika, Institut Teknologi Bandung.
[23]	Tran, D., Wanli Ma, Sharma, D. 2008. "Automated network *feature* weighting-based anomaly detection," *Intelligence and Security Informatics, IEEE International Conference on* , pp.162-166.