ANALISIS DAN IMPLEMENTASI RANDOM FOREST DAN CLASSIFICATION DAN REGRESSION TREE (CART) UNTUK KLASIFIKASI PADA MISUSE INTRUSION DETECTION SYSTEM

Fransiska Amalia Kurniawan[1], Adiwijawa[2], Angelina Prima Kurniati[3]

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

Abstrak

Intrusion Detection System (IDS) merupakan alat yang memonitor event yang terjadi pada sistem atau jaringan komputer dan memberikan peringatan jika ada aktivitas yang berbahaya. Cara mengenali aktivitas yang berbahaya membutuhkan proses mengenali dari aktivitas sebelumnya yang disebut Misuse IDS. Misuse IDS ini menggunakan teknik klasifikasi.

Pada Tugas Akhir ini penulis mengimplementasikan gabungan metode Random Forest (RF) dan Classification and Regression Tree (CART) untuk membangun model klasifikasi yang digunakan dalam Misuse IDS. Melihat karakteristik dataset KDD Cup 1999 yang digunakan dalam Tugas Akhir ini merupakan unbalanced dataset yang memiliki banyak features, dengan penggabungan kedua metode diharapkan dapat memberikan solusi atas permasalahan pada Dataset KDD Cup 1999 agar meningkatkan akurasi deteksi intrusi.

Pengujian dilakukan dengan mengevaluasi matrix confusion dan menentukan nilai precison, recall, F Measure. Berdasarkan hasil evaluasi terhadap ketiga parameter tersebut, disimpulkan bahwa implementasi gabungan metode RF dan CART mampu mengklasifikasikan kelas minor pada dataset KDD Cup 1999 dengan sedikit modifikasi pada metode RF yaitu mengimplementasikan teknik Balanced Random Forest (BRF) dengan cara menyeimbangkan jumlah kelas mayor dan kelas minor. Akurasi yang dicapai oleh model belum mencapai nilai maksimal, karena keterbatasan jumlah record pada data training setelah diimplementasikan BRF yang tidak dapat menggambarkan karakteristik kelaskelas yang ada.

Kata Kunci : Klasifikasi, Ranfom Forest, Classification and Regression Tree,

Abstract

Intrusion Detection System (IDS) is a tools or application to monitor events that happened in computer system and give alert when a danger activity occurs. The process to recognize the danger activity is learning from the activity before. This process is called misuse Intrusion Detection System which use classification techique.

In this Final Project the author implements Random Forest (RF) and Classification and Regression Tree (CART) methods to build classification models that used in misuse IDS. The characteristics of KDD Cup 1999 dataset which used in this final project are unbalanced dataset, and there are many features in it. The excellences in that method are wiiling to give solutions of the problem in the dataset to improve the accuracy of intrusion detection.

Testing is done by evaluating confusion matrix and calculating the value of precion, recall and F Measure. The result showed that implementation of RF and CART could classify the minority classes in KDD Cup 1999 dataset with a little modification in RF method that is balancing the number of records of majority and minority classes. This modification named Balanced Random Forest (BRF). The accuracy is not quite high because of the limited number of records used in training phase after BRF was implemented. Limited number of records couldn't give detail characteristics of each class in the dataset.

Keywords : Classification, Ranfom Forest, Classification and Regression Tree,

1. Pendahuluan

1.1 Latar Belakang Masalah

Jaringan komputer sering menjadi target para *hacker* dan *intruder*. Dalam melakukan serangan, para *hacker* atau *intruder* melakukan aktivitas diluar aktivitas normal jaringan. Aktivitas inilah yang disebut sebagai intrusi.

Intrusion detection adalah proses untuk memonitor *event* yang terjadi pada jaringan komputer dan melakukan analisis data tersebut untuk mengetahui adanya intrusi. *Intrusion Detection System* (IDS) memainkan peranan penting dalam menjaga *integrity, confidentiality dan availability* dari sumber daya jaringan komputer [2]. Tujuan utama IDS adalah mengklasifikasikan aktivitas jaringan apakah termasuk aktivitas normal atau intrusi.

Ada 2 pendekatan dalam mengenali intrusi dalam IDS, yaitu *anomaly* dan *misuse detection*. *Anomaly detection* adalah mengidentifikasi perilaku tak lazim yang terjadi (*anomaly*) dalam *host* atau *network*. Detektor berfungsi dengan asumsi bahwa intrusi itu berbeda dengan aktivitas normal. Oleh karena itu dalam *anomaly detection* menggunakan *clustering* (pengelompokan) supaya jika ada aktivitas yang berbeda, segera dicurigai sebagai intrusi. *Anomaly detection* menggunakan pendekatan *unsupervised clustering*, yang mampu mendeteksi intrusi tanpa harus mempelajari data terlebih dahulu . Sedangkan *misuse detection* termasuk *supervised learning*, dimana dibutuhkan data sebelumnya sebagai pembelajar. Bedasarkan data yang sudah dipelajari tersebut akan dibentuk *pattern* (pola) untuk masing-masing kelas. Dengan demikian, metode klasifikasi dapat menangani *misuse detection*.

Yang menjadi fokus dalam penelitian tugas akhir ini adalah *misuse detection* yaitu menggunakan metode klasifikasi. *Misuse detection* adalah menganalisis aktivitas sistem, mencari *event* yang cocok dengan pola perilaku yang dikenali sebagai serangan . Salah satu kelebihan dari *misuse detector* adalah mampu dengan cepat dan handal mendiagnosa penggunaan teknik serangan tertentu. Kekurangan dari *misuse detection* ini adalah banyak *false positive* (terdeteksi ada serangan padahal tidak terjadi serangan). Telah banyak metode yang digunakan untuk *misuse detection* antara lain: ANFIS [21], Hidden Markov Model [19], SVM, Naive Bayes [16], C4.5 [14].

Proses klasifikasi *misuse detection* menggunakan *dataset* KDD Cup 1999. Karakteristik dataset ini adalah *dataset* yang memiliki banyak *features* (42 features) yang terdiri dari *features* kategoris dan numerik. Selain itu *dataset* KDD Cup 1999 tergolong *unbalanced data*, dimana ada kelas yang terdistribusi tidak seimbang diantara kelas yang berbeda [13].

Melihat karakteristik dataset KDD Cup 1999, diperlukan metode yang mampu menangani masalah *unbalanced data* dan menangani banyak *features* yang berbeda-beda jenisnya (kategoris dan numerik).

Random Forest (RF) merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi. RF dapat meningkatkan akurasi karena adanya pemilihan secara acak dalam membangkitkan simpul anak untuk setiap *node* (simpul diatasnya) dan diakumulasikan hasil klasifikasi dari

setiap pohon, kemudian dipilih hasil klasifikasi yang paling banyak muncul [22] . Banyaknya pohon yang akan dibentuk sangat berpengaruh terhadap tingkat akurasi hasil klasifikasi. Semakin banyak pohon, semakin akurat hasil klasifikasinya. Selain itu juga RF dapat menangani input variabel yang besar, menyeimbangkan *error* dalam *unbalanced dataset*. Untuk menangani *unbalanced data*, algoritma RF mengalami sedikit modifikasi pada pemilihan data *training*, yaitu dengan menyeimbangkan jumlah *record* pada kelas mayor dan minor. Teknik ini disebut *Balanced Random Forest* (BRF).

Dalam algoritma RF, diperlukan algoritma untuk membangun *tree*. Salah satu algoritma yang dapat digunakan adalah algoritma *Classification and Regression Tree* (CART). CART membagi pohon keputusan dengan teknik *binary tree* dan dapat diterapkan pada variabel numerik dan kategoris sekaligus [2]. Telah dibuktikan pada [3], bahwa CART cocok diterapkan untuk data dengan variabel yang banyak dan kompleks.

Dalam penelitian ini dilakukan penggabungan metode RF dan CART. Dengan kekuatan metode – metode yang digunakan dalam penelitian ini, diharapkan dapat menjawab permasalahan *misuse detection* menggunakan dataset KDD Cup 1999. Permasalahan *unbalanced data* ditangani oleh metode RF dan permasalahan perbedaan jenis *features* ditangani oleh CART.

1.2 Perumusan Masalah

Permasalahan yang menjadi objek dari penelitian tugas akhir ini, terdiri atas :

a. Bagaimana cara menerapkan algoritma Random Forest dan CART untuk membangun model klasifikasi pada *misuse* IDS?
b. Bagaimana performansi (bedasarkan *precison, recall, F Measure*) sistem yang dibangun dengan algoritma RF dan CART dalam mendeteksi intrusi (diklasifikasikan menjadi Normal, Probe, DoS, U2R, R2L)?
c. Berapa banyak jumlah pohon optimum yang dibentuk agar dapat menghasilkan tingkat akurasi yang lebih baik dalam mendeteksi intrusi (diklasifikasikan menjadi normal, probe, DoS, U2R, R2L)?

Hipotesa awal dari penelitian ini adalah gabungan metode RF dan CART dapat menghasilkan akurasi yang lebih baik daripada *single classifier* untuk klasifikasi dalam mendeteksi intrusi.

Batasan masalah untuk penelitian ini adalah:

1. *Preprocessing* (Normalisasi, *remove outlier*, *feature selection*) dilakukan menggunakan tools Weka 3.6.1
2. *Dataset* yang digunakan berasal dari data KDD Cup 1999
3. Sistem yang dibangun tidak *real-time*

1.3 Tujuan

Tujuan dengan dilakukannya penelitian ini adalah:

1. Menganalisis dan mengimplementasikan algoritma RF dan CART untuk membangun model yang dapat digunakan untuk klasifikasi dalam *misuse detection* IDS bedasarkan *dataset* KDD Cup 1999
2. Menganalisis performansi sistem yang dibangun (berdasarkan *precison, recall, F Measure*) serta faktor yang mempengaruhi keakuratan dari sistem yang dibangun
3. Menganalisis pengaruh banyaknya pohon yang dibentuk terhadap tingkat akurasi dalam mendeteksi intrusi (diklasifikasikan menjadi Normal, Probe, DoS, U2R, R2L)

1.4 Metodologi Penyelesaian Masalah

Metodologi yang dilakukan untuk menyelesaikan permasalahan adalah sebagai berikut:

1. Studi literatur

   Melakukan pencarian serta mempelajari informasi dan pembelajaran tentang *misuse* IDS, khususnya mengenai konsep dan cara kerja metode RF dan CART untuk klasifikasi *misuse* IDS. Selain itu juga mempelajari karakteristik dan cara *preprocessing data* yang dapat diterapkan pada *dataset* KDD Cup 1999.
2. Pengumpulan data-data

   Melakukan pencarian data yang digunakan untuk penelitian Tugas Akhir ini. Data yang dicari adalah *dataset* KDD Cup 1999, serta keterangan *features* didalamnya.
3. Analisis modifikasi CART

   Melakukan analisis kemungkinan modifikasi atau pengembangan terhadap metode CART yang akan digabungkan dengan metode RF, bedasarkan kelebihan dan kekurangan dari metode RF dan CART.
4. Analisis dan perancangan aplikasi

   Melakukan analisis dan perancangan aplikasi yang akan dibentuk menggunakan metode RF dan CART sehingga dapat digunakan untuk menghitung akurasi dari klasifikasi *misuse* IDS.
5. Implementasi aplikasi

   Melakukan implementasi aplikasi sesuai dengan hasil analisis dan perancangan metode RF dan CART sehingga dapat digunakan untuk menghitung akurasi dari klasifikasi *misuse* IDS.
6. Pengujian aplikasi

   Melakukan pengujian aplikasi dan menganalisis hasil keluaran aplikasi, sejauh mana keakuratan dari *detection model* yang dibangun dengan menggunakan metode RF dan CART dalam mengklasifikasikan *data testing* KDDCup 1999.
7. Pembuatan laporan Tugas Akhir

   Melakukan penyusunan laporan hasil penelitian yang telah dilakukan serta memberikan kesimpulan dari hasil penelitian tersebut.

5. Penutup

5.1 Kesimpulan

Bedasarkan analisis yang diperoleh dari hasil pengujian dalam penelitian Tugas Akhir ini, dapat ditarik kesimpulan sebagai berikut:

1. Gabungan Algoritma RF dan CART sudah mampu diimplementasikan untuk klasifikasi dalan sistem pendeteksi intrusi, namun berdasarkan parameter pengujian *precison*, *recall*, *F Measure*, sistem yang dibangun menggunakan RF dan CART belum dapat dikatakan akurat melihat nilai akurasi tertinggi yang dapat dicapai sebesar 70,72% saat jumlah *record* = 224 *records* dan banyaknya pohon = 60 pohon.
2. Akurasi dipengaruhi oleh pemilihan *features* secara acak, kandidat atribut pada setiap *node*, pemilihan data *training* secara acak, pengukuran *impurity measure* yang digunakan, dan jumlah *record* pada data *training*.
3. Teknik pemilihan *bootstrap sample* (data yang akan dijadikan data *training* pada tiap pohon) berpengaruh besar terhadap akurasi dan jumlah kelas yang dapat diprediksi. Memilih *bootstrap sample* menggunakan BRF mampu mendeteksi kelas-kelas minor dalan *unbalanced dataset*, sedangkan RF belum mampu mendeteksi kelas minor sebaik BRF.
4. Banyaknya pohon yang optimum untuk deteksi intrusi adalah 50 – 60 pohon. Banyaknya pohon optimum juga bergantung pada kasus karakterisktik intrusi. Banyaknya pohon optimum untuk deteksi intrusi yang digunakan sebagai data *training* dan intrusi dari luar data *training* adalah berbeda.

5.2 Saran

Berdasarkan hasil analisis dan kesimpulan, terdapat beberapa saran untuk perbaikan pada penelitian klasifikasi intrusi sebagai berikut:

1. Sebaiknya proses *preprocessing feature selection* lebih ditekankan, keterkaitan tiap-tiap kolom dengan karakteristik masing-masing kelas, sehingga didapatkan *features* yang mencerminkan keseluruhan kelas.
2. Mengganti algoritma *decision tree* yang diimplementasikan pada RF.
3. Menggunakan pengukuran *impurity measure* yang lain sebagai kriteria *split condition*.
4. Memperbesar nilai parameter F (jumlah kandidat atribut pada setiap *node*).
5. Untuk implementasi di Indonesia, sebaiknya mengambil data traffic dari jaringan *backbone* Indonesia, yang menggambarkan aktivitas umum jaringan di Indonesia.

Daftar Pustaka

[1] ___. Data Mining. http://lecturer.eepis-its.edu/~tessy/lecturenotes/db2/bab10.pdf

[2] ___. IP Network-Packet Shared Media pada Mesin Cluster Intrusion Detection System. http://budi.insan.co.id/courses/el695/projects2002-2003/ivan-report.pdf

[3] ___. Konsep Data Mining. http://bertalya.staff.gunadarma.ac.id/.../Klasifikasi-Pohon+Keputusan.pdf

[4] ___. Sistem Deteksi Intrusi. http://id.wikipedia.org/wiki/Sistem_deteksi_intrusi

[5] Adetunmbi, Adebayo O , Samuel O. Falaki, Olumide S. Adewale, Boniface K. Alese. *Network Intrusion Detection Based on Rough Set and K-Nearest Neighbour*. Department of Computer Science, Federal University of Technology. 2008

[6] Breiman, Leo, Jerome H Friedman, Richard A Oshlen, Charles J Stone. *Classification and Regression Trees*. University of California, Berkeley.

[7] Breiman, Leo. *Random Forest*. Statistics Department, University of California, Berkeley. 2001

[8] Chen, Chao, Andy Liaw, Leo Breiman. *Using Random Forest to Learn Imbalanced Data*. Department of Statictics, UC Berkeley.

[9] Degorski, Lukasz, Lukasz Kobylinski, Adam Przepiorkowski. *Definition Extraction: Improving Balanced Random Forest*. Institute of Computer Science, Warszawa, Poland. 2008

[10] Gagne, David john, Amy McGovern. *Using Multiple Machine Learning to Improve the Classification of a Storm Set*. University of Oklahoma, Norman - Oklahoma.

[11] Lee, Jin-Seon, Il-Seok Oh. *Binary Classification Trees for Multiclass Classification Problems*. Department of Computer Engineering, Woosuk University, Korea. 2003

[12] Loh, Wei – Yin. *Classification and Regression Tree Methods*. University of Wisconsin, Madison. 2008

[13] Rahmani, Luthfia. *Metode Feature Selection Dalam Menangani Data Imbalance pada Klasifikasi Dokumen Multi-Label*. Institut Teknologi Telkom. 2007

[14] Rajeswari, L Prema, Kannan Arputharaj. *An Active Rule Approach for Network Intrusion Detection with Enhanced C4.5 Algorithm*. College of Engineering, Guindy, Anna University, Chennai, India. 2008

[15] Tan, Pang-Ning, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining. Boston.* United States of America. 2006

[16] Tavallaee, Mahbod, Ebrahim Bagher, Wei Lu, Ali A. Ghorbani. *A Detailed Analysis of the KDD Cup 99 Data Set.* University of New Brunswick, Fredericton, NB, Canada. 2009

[17] Toosi, Adel Nadjaran, Mohsen Kahani, Reza Monsefi. *Network Intrusion Detection Based on Neuro-Fuzzy Classification*. University of Mashhad, Ferdowsi. 2006

[18] Wang, Bo, Kening Gao, Bin Zhang. *Algorithm of Feature Selection for Inconsistent Data Preprocessing Based Rough Set*. Institute for Scientific Computing and Information. 2005

[19] Wang, Liangjiang, Caiyan Huang, Jack Y Yang. *Predicting siRNA Potency with Random Forest and Support Vector Machines.* Purdue University, Indiana. 2010

[20] Widjanarko Otok, Bambang, Sumarmi. *Bagging CART pada Klasifikasi Anak Putus Sekolah*. FMIPA-ITS, Surabaya. 2009

[21] Yudibert Donny Pasaribu, Manaek. *Analisa dan Implementasi Metode Hidden Markov Model pada Intrusion Detection System (IDS)*. Institut Teknologi Telkom. 2007

[22] Zainal, Anazida, Mohd Aizaini Maarof ,Siti Mariyam Shamsuddin, Ajith Abraham. *Ensemble of One Class Classifier for Network Intrusion Detection System.* Faculty of Computer Science and Information System, Universiti Teknologi Malaysia.