PERBANDINGAN METODE *SAMPLING* DAN *DIMENSIONALITY REDUCTION* UNTUK MEREDUKSI KOMPLEKSITAS ALGORITMA DETEKSI PADA DDOS

COMPARISON OF SAMPLING AND DIMENSIONALITY REDUCTION METHOD FOR DETECTION ALGORITHM COMPLEXITY REDUCTION IN DDOS

Fahmy Rezqi Pramudhito[1], Yudha Purwanto[2], Astri Novianty [3]
[1,2,3]Prodi S1 Sistem Komputer, Fakultas Teknik, Universitas Telkom
[1]fahmyrezqi@students.telkomuniversity.ac.id, [2]omyudha@telkomuniversity.ac.id, [3]astrinov@telkomuniversity.ac.id

Abstrak

Pada kenyataannya data yang berukuran besar tidak akurat, tidak komplit dan tidak konsisten. Sebuah data yang tidak berkualitas akan menghasilkan hasil proses yang tidak berkualitas. Yang menyebabkan data tidak akurat ,tidak komplit dan tidak konsisten diantaranya adalah kesalahan dari manusia dan *computer error* pada saat memasukan data. Selain itu yang menyebabkan data tidak komplit diantaranya tidak konsisten dalam kaidah penamaan dan tidak konsisten dalam format untuk pengisian. Dengan adanya data yang tidak konsisten akan membuat data yang relevan kemungkinan tidak terekam dan menjadi sulit untuk dimengerti. Apabila data tidak konsisten terekam maka data tersebut akan dihapus secara otomatis.Pada Tugas Akhir ini,melakukan pengolahan data mentah ke tahap preprocessing dengan menggunakan teknik data *reduction* yaitu sampling dan *dimensionality reduction*. Tujuannya untuk mereduksi kompleksitas data yang diteliti dan hasil dari data preprocessing yang diperoleh dapat diklasifikasikan berdasarkan kebutuhan algoritma yang diteliti.Pada proses sampling data yang besar akan diolah menjadi data baru secara acak dari data sample yang ada. Sementara pada proses selanjutnya *dimensionality reduction*, data yang mempunyai *high dimensionality* akan direduksi menjadi *lower dimensionality* sehingga akan mendapatkan ouput berupa *new feature*.Data yang akan diteliti berupa *raw* data hasil streaming yang dilakukan oleh NS-3. Data streaming yang dilakukan oleh NS-3 terdiri dari serangan normal dan anomali.Data tersebut akan diolah ke tahap preprocessing, sehingga akan memperoleh relevansi fitur trafik baru. Hasil dari penilitian ini memperoleh kompleksitas dari masing-masing algoritma. Dengan hasil kompleksitas tersebut maka kompleksitas skenario 1 lebih baik dengan skenario 2. Dengan adanya penggabungan antara Sampling + PCA maka diperoleh nilai big-O dengan notasi O(n,p) . Dengan n sebagai jumlah data analisis sampling dan p sebagai jumlah kolom dari analisis PCA.

Kata kunci: *DDoS,Sampling,Dimensionality Reduction,Time Complexity,NS-3*

Abstract

In fact, large-sized data is inaccurate, incomplete and inconsistent. A bad quality data is not going to produce the results of a process that is not qualified. Why data is inaccurate, incomplete and inconsistent they are from human error and computer error during data entry. Besides that cause data to complete them are not consistent in naming rules and inconsistent in a format for charging. With the inconsistent data will make the data relevant to the possibility of unrecorded and difficult to understand. If the recorded data is inconsistent then the data will be automatically deleted.In this final project, do the processing of raw data to the preprocessing stage using data reduction techniques that sampling and dimensionality reduction. The goal is to reduce the complexity of the data examined and the results of preprocessing the data obtained can be classified based on the algorithm needs vitality.Pada sampling process large amounts of data to be processed into new data randomly from the existing sample data. While in the process further dimensionality reduction, data that have a high dimensionality will be reduced to lower dimensionality that will get the output in the form of new features.Data that will be examined in the form of raw data stream results conducted by NS-3. Data streaming is done by the NS-3 consists of normal attacks and anomalies.Data will be processed to the preprocessing stage, so it will gain new relevance traffic features. The results of this research obtained the complexity of each algorithm. With the results of such complexity, the complexity of a better scenario 1 with scenario 2. With the merger between Sampling + PCA, the obtained value of big-O notation O (n, p). N is the number of sampling and analysis of data p as the number of columns from PCA analysis.
Keyword: DDoS,Sampling,Dimensionality Reduction,Time Complexity,NS-3

1. Pendahuluan

Sekarang pada kenyataannya, banyak sekali sebuah data yang berukuran besar tidak akurat, tidak komplit dan tidak konsisten. Sebuah data yang tidak berkualitas akan menghasilkan hasil proses yang tidak berkualitas. Yang menyebabkan data tidak akurat ,tidak komplit dan tidak konsisten diantaranya adalah kesalahan dari manusia dan *computer error* pada saat memasukan data. Selain itu yang menyebabkan data tidak komplit diantaranya tidak konsisten dalam kaidah penamaan dan tidak konsisten dalam format untuk pengisian. Dengan adanya data yang tidak konsisten akan membuat data yang relevan kemungkinan tidak terekam dan menjadi sulit untuk dimengerti. Apabila data tidak konsisten terekam maka data tersebut akan dihapus secara otomatis.*Preprocessing* merupakan teknik dari data *mining* untuk mengolah suatu data mentah menjadi data yang berkualitas. Dalam proses *Preprocessing* terdapat beberapa teknik yang digunakan terdiri dari Data *Cleaning*,Data *Integration*, Data *Reduction*, dan Data *Transformation and Data Discretization* [1]. Proses *preprocessing* sendiri akan menghasilkan tujuan diantaranya data tersebut dapat dikelompokkan berdasarkan objek dan attribute. Dengan adanya proses *Preprocessing* maka sebuah kesalahan dalam data akan berhasil dikurangi. Pada penilitian sebelumnya pada tahap *preprocessing*, algoritma yang digunakan adalah *Principal Component Analysis* yang memiliki keuntungan pendeteksian serangan tanpa ada *error* pada pengklasifikasian serangan dan data yang digunakan menggunakan KDD'99 [2][5] , sementara untuk *sampling* , algoritma yang digunakan adalah *Strafied Random Sampling* memiliki keuntungan dengan pengambilan sampel secara strata[14][15]. Hasil yang disimpulkan adalah *Principal Component Analysis* salah satu metode *Dimensionality Reduction* untuk mengekstrak fitur dari sebuah dimensional *vector* yang tinggi dari sebuah data [3][4]. Sementara metode *sampling* merupakan salah satu metode untuk mereduksi jumlah data

2. Dasar Teori

2.1 *Distributed Denial of Service*

Distributed Denial Of Service merupakan sebuah penyerangan yang dilakukan beberapa orang atau atau bahkan server bisa mati dengan sendirinya. Proses penyerangan DDOS dapat dilakukan melalui beberapa kombinasi penyerangan dan apabila satu aliran pada trafik DDoS harus bahkan banyak orang yang bertujuan untuk menghancurkan sebuah jaringan komputer atau server . Dampak yang ditimbulkan oleh DDOS adalah jaringan pada sebuah server akan mengalami down dalam keadaan stabil dan konstan pada *high rate traffic* dan akurasi akan berkurang dengan adanya peningkatan traffic[7].Sehingga paket traffic yang dihasilkan akan sangat tinggi dan user tidak dapat mengakses service tertentu. Menurut [8] pendeteksian DDoS diperoleh informasi yang berasal dari *IP packet* seperti, *IP address*,*time-to-live*(TTL),protocol type (tipe protokol). Pendeteksi DDoS dapat dibedakan oleh trafik normal yang berasal dari abnormal trafik yang bisa dideteksi sebagai serangan. Salah satu beberapa teknik untuk pendeteksin DdoS dengan teknik memonitor peningkatan jumlah *IP Source* yang masuk kedalam sebuah jaringan [13][11].

2.2 *Network Simulator – 3 (NS-3)*

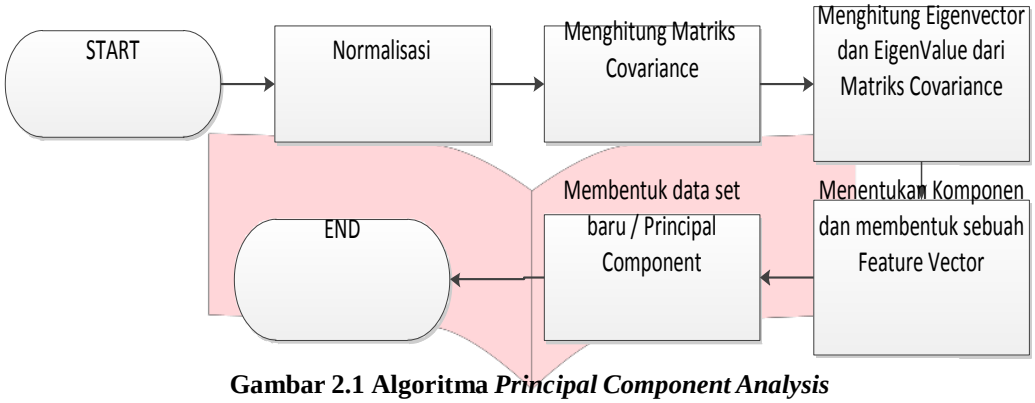
Network simulasi merupakan sebuah perangkat lunak yang didesain untuk digunakan dalam mempelajari struktur dinamik dari jaringan komunikasi riset. Beberapa simulator jaringan beberapa diantaranya *Network Simulator* 3. *Network Simulator* 3 merupakan sebuah pengembangan NS-2[9][10], NS-3 *simulator* jaringan, yang berguna untuk keperluan pengembangan ,percobaan dan penilitan riset. NS-3 berjalan dengan menggunakan bahasa C++ dan python.NS-3 dapat dijalankan di Sistem Operasi Linux / seluruh varian Linux dan windows dengan cygwin.Tujuan dari project NS-3 adalah untuk mengembangkan simulation yang berbasis open source untuk riset dunia *networking*. Dengan adanya perkembangan tersebut harus selaras dengan perkembangan riset *modern networking* dan harus didukung kontribusi komunitas,para *reviewer* dan validasi *software*.

2.4 *Preprocessing*

Preprocessing merupakan tahap penyiapan data sebelum masuk kedalam tahapan *processing*. Pada proses *preprocessing* data yang diolah merupakan data mentah yang mempunyai banyak kesalahan diantaranya adalah tidak akurasi, tidak konsisten dan memiliki noise. Dengan diprosesnya suatu data mentah pada tahapan yang terdapat didalam *preprocessing* akan menghasilkan sebuah data yang akurat sehingga data yang akurat tersebut akan memperoleh nilai akurasi yang tinggi. Pada tahap *preprocessing* terdapat beberapa tahap diantaranya dengan mengekstraksi ciri menggunakan *Principal Component Analysis* (PCA) dan mengambil beberapa sampel dari data menggunakan metode *Sampling*. Pada penilitian ini dilakukan proses *preprocessing data transformation*, yaitu mengubah suatu data agar diperoleh data yang lebih berkualitas. Tujuan *preprocessing* menghasilkan dataset baru yang akan menjadi data inputan algoritma deteksi [12].

2.5 *Principal Component Analysis*

Secara umum *Principal Component Analysis* merupakan suatu metode untuk mereduksi dimensi variabel pada banyak variable, kompresi data, *patern recognition* dan analisis *statistic*. *Principal Component Analysis* sendiri memiliki fungsi untuk mereduksi dimensi variable input menjadi komponen utama yang mempunyai dimensi yang lebih kecil dengan meminimalisir kehilangan informasi tetapi akan mempertahankan *variability* dalam data,dimana komponen utama yang terbentuk tidak berkorelasi komponen satu dengan yang lainnya. Algoritma *Principal Component Analysis* [4] pada umumnya mempunyai alur seperti diagram berikut :



Gambar 2.1 Algoritma *Principal Component Analysis*

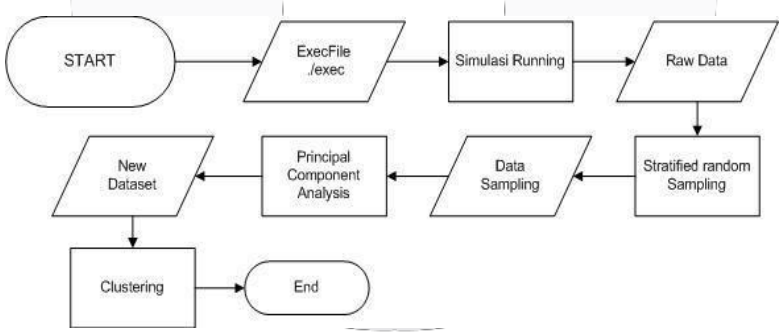
2.6 *Sampling*

Sampling merupakan salah satu dari teknik data *reduction*. Dimana apabila terdapat data yang besar akan diolah menjadi data kecil secara acak dari data sampel yang ada. *Sampling* merupakan teknik pengambilan sampel dari populasi. Sampel yang diambil adalah sampel yang dapat mewakili populasi. Sampling Metode sampling dibagi menjadi dua, yaitu: [14]

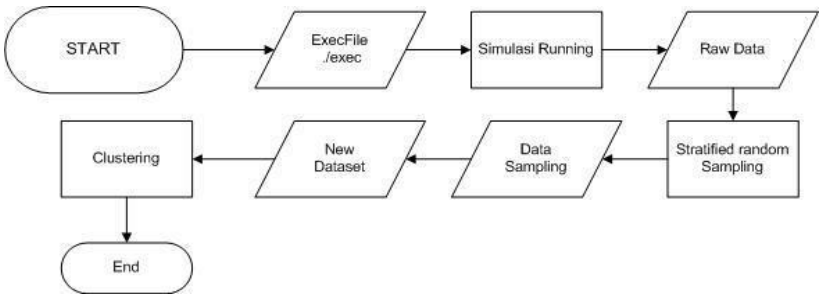
3. Pembahasan

3.1 Deskripsi Sistem

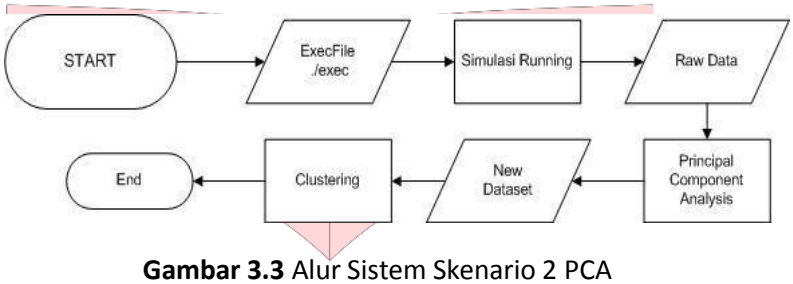
Pada penelitian tugas akhir mengikuti alur sistem seperti gambar berikut:



Gambar 3.1 Alur Sistem Skenario 1



Gambar 3.2 Alur Sistem Skenario 2 Sampling



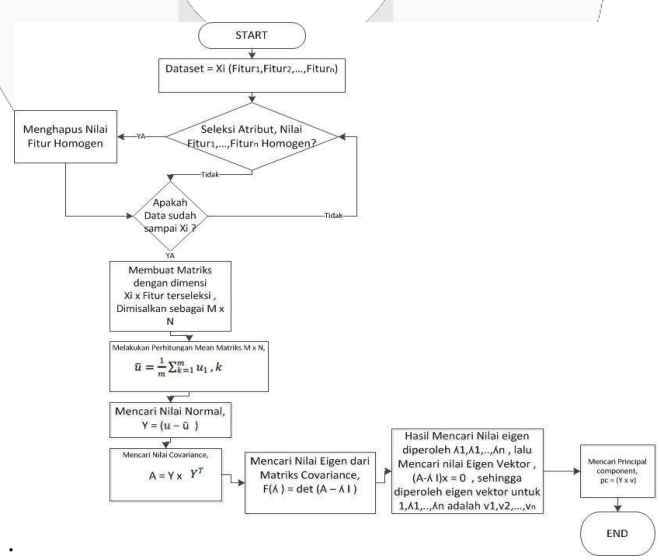
Gambar 3.3 Alur Sistem Skenario 2 PCA

3.2 Dataset Ns3

Dalam penilitian tugas akhir ini membutuhkan dataset sebagai data yang akan diolah pada algoritma deteksi. Dataset NS3 adalah data yang bersifat *real time traffic* yang diduga terdapat beberapa serangan didalamnya. Berdasarkan hasil raw data diatas akan dilakukan tahap *preprocessing*. Tahap *preprocessing* ini dilakukan untuk mendapatkan fitur – fitur yang relevan dan sudah tereduksi pada datanya. Hasil dari tahap *preprocessing* digunakan untuk inputan dari skema.

3.3 *Principal Component Analysis*

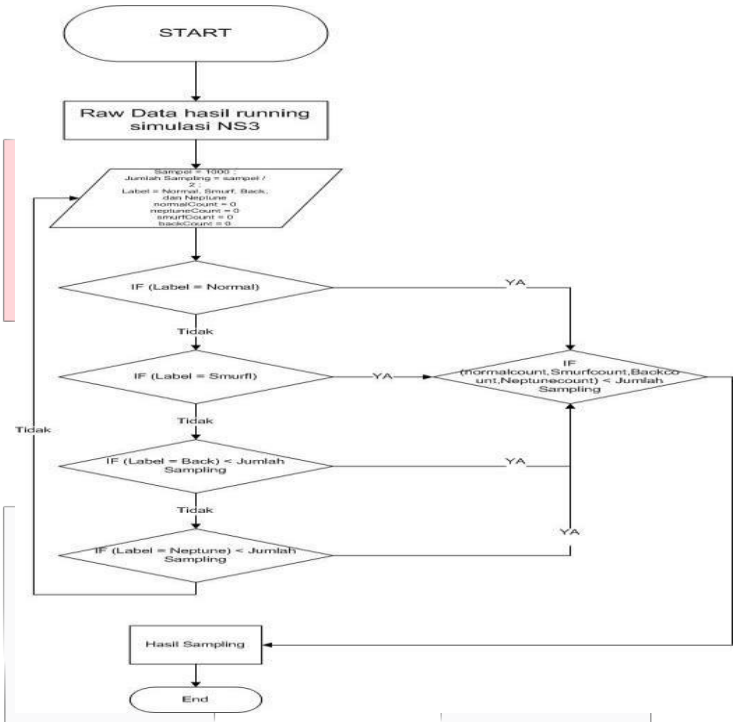
Pada tahap ini , hasil data dari pengolahan algoritma *sampling* akan diolah kembali kedalam algoritma PCA Sebelumnya fitur yang diambil terdapat 8 fitur lalu akan diperoses pereduksian fitur sebanyak 4 fitur. Diambilnya 4 fitur disebabkan oleh pengambilan *eigenvalue* tertinggi dan telah tentukan oleh penilitinya[4]. Dengan adanya proses ini *output* yang dihasilkan akan berkurang dari dimensi aslinya.



Gambar 3.2 Flowchart PCA

3.4 *Strafied Random Sampling*

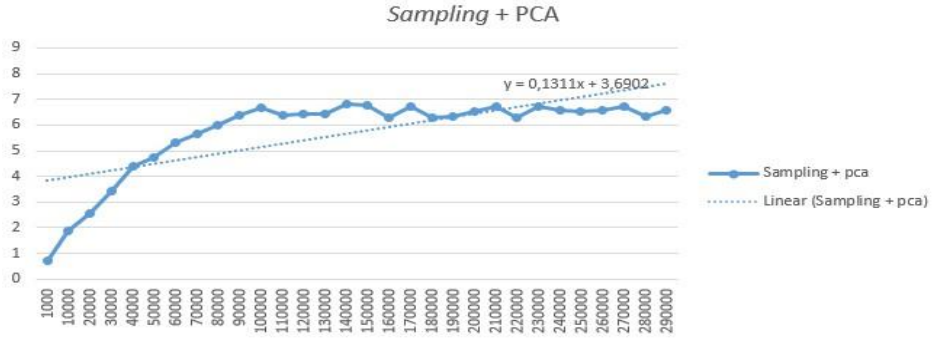
Pada penilitian ini menggunakan *Strafied Random Sampling* ,Proses pengambilan sampel dilakukan dengan memberi kesempatan yang sama pada setiap anggota populasi untuk menjadi anggota sampel. Jumlah sampel yang diambil terdapat minimum 30 sub sampel dari masing-masing serangan dan normal proses pemilihan tersebut dilakukan secara *random*[14][17]. Berikut *flowchart* dari algoritma *Stratified Random Sampling*.



Gambar 3.4 Flowchart PCA

4. Analisis

Dibawah ini merupakan skenario 1

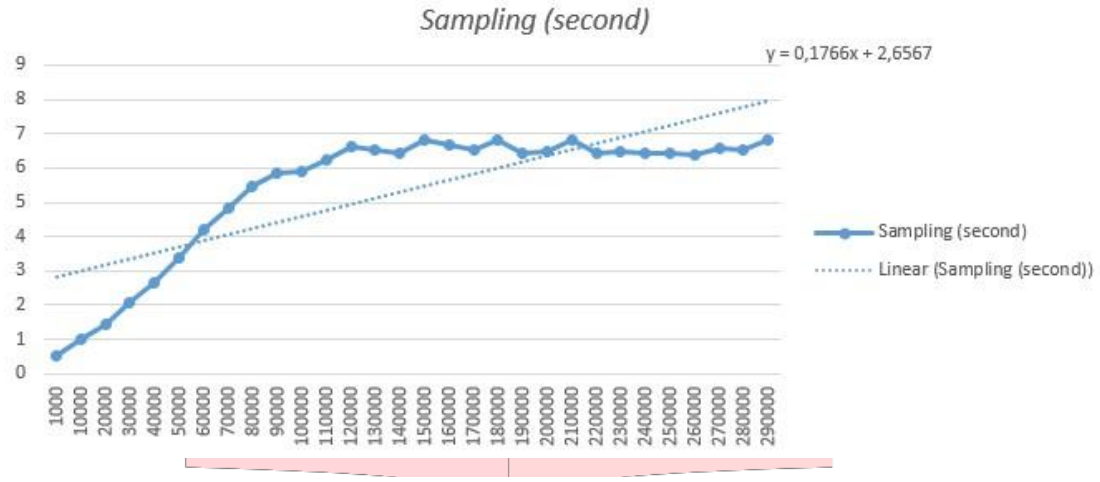


Gambar 4.2 Grafik waktu komputasi Sampling + PCA

Pada gambar dibawah dijelaskan bahwa waktu komputasi sistem akan terus bertumbuh seiring dengan pertumbuhan data dari jumlah data yang diberikan. Dengan hal tersebut dapat disimpulkan bahwa laju pertumbuhan dengan karakteristik yang linier dengan notasi O(n). Dengan adanya penggabungan antara

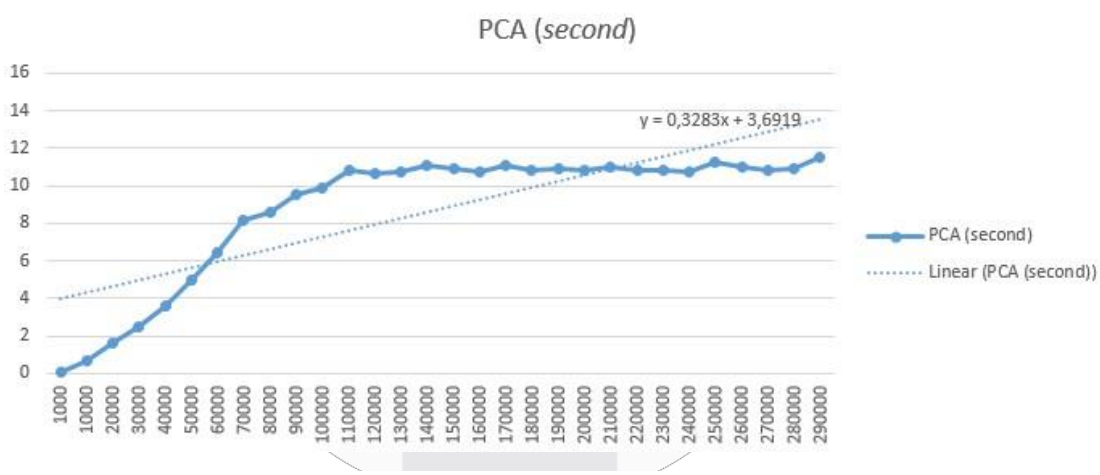
Sampling + PCA maka diperoleh nilai big-O dengan notasi O(n,p) . Dengan n sebagai jumlah data analisis sampling dan p sebagai jumlah kolom atau relevansi fitur dari analisis PCA.

Dibawah ini merupakan skenario 2



Gambar 4.3 Grafik waktu komputasi Sampling

Pada gambar dibawah dijelaskan bahwa waktu komputasi sistem akan terus bertumbuh seiring dengan pertumbuhan data dari jumlah data yang diberikan. Dengan hal tersebut dapat disimpulkan bahwa laju pertumbuhan dengan karakteristik yang linier dengan notasi O(n). Dengan n sebagai jumlah data analisis sampling.



Gambar 4.3 Grafik waktu komputasi PCA

Pada gambar dibawah dijelaskan bahwa waktu komputasi sistem akan terus bertumbuh seiring dengan pertumbuhan data dari jumlah data yang diberikan. Dengan hal tersebut dapat disimpulkan bahwa laju pertumbuhan dengan karakteristik yang linier dengan notasi O(n). Dengan n sebagai jumlah data analisis sampling.

5. Kesimpulan dan saran

5.1 Kesimpulan

Dari hasil yang didapatkan pada penelitian ini, dapat ditarik beberapa kesimpulan sebagai berikut :

1. Berdasarkan pengujian dataset yang dihasilkan bersifat stream sehingga data yang dihasilkan memiliki berbagai macam serangan.
2. Berdasarkan pengujian dari *time complexity* dihasilkan bahwa dari skenario 1 lebih baik daripada skenario 2.
3. Berdasarkan pengujian performansi sistem yang telah dibangun memiliki fungsi kompleksitas yang efisien pada masing-masing algoritma yang diuji.

5.2 Saran

Saran untuk penelitian selanjutnya adalah :

1. Untuk proses menghasilkan data ada lebih baiknya untuk mendapatkan lebih dari 8 fitur untuk menguji sistem dari PCA dan sampling sehingga dapat menghasilkan data yang lebih relevan.
2. Proses eksekusi dapat ditingkatkan dengan menambah beberapa aksi pada setiap serangan.
3. Untuk skema penyerangannya dilakukan dengan cara menambahkan beberapa fitur serangan.

DAFTAR PUSTAKA

[1] Jiawei Han, Micheline Kamber, and Jian Pei,Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA, ISBN 978-0- 12-381479-1

[2] Annie George A. George and A. V. Vidyapeetham, "Anomaly Detection based on Machine Learning : Dimensionalitty Reduction using PCA and Classification using SVM," International Journal of Computer Application, vol. 47, 2012.

[3] CHEN Bo, Ma Wu, “Research of Intrusion Detectionbased on Principal Components Analysis”, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[4] Smith,Lindsay I, “A tutorial on Principal Components Analysis”,2002.

[5] Trinita S.P, Yudha Purwanto, Tito Waluyo Purboyo, "A Sliding Window Technique for Covariance Matrix to Detect Anomalies on Stream Traffic",Electrical Engineering Faculty ,Telkom University,Bandung, Indonesia, International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) 2015

[6]Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol,Data Mining Methods for Recommender Systems, pp 39-71, Copyright 2011 ,DOI 10.1007/978-0- 387-85820- 3_2 ,Print ISBN 978-0- 387-85819-7.

[7]Yudha Purwanto, Kuspriyanto, Hendrawan, dan Budi Rahardjo,"Traffic Anomaly Detection in DDoS Flooding Attack",THE 8TH INTERNATIONAL CONFERENCE ON TELECOMMUNICATION SYSTEM,SERVICES,AND APPLICATION, 2014

[8]Theerasak Thapngam,Shui Yu,Wanlei Zhou & S. Kami Makki,"Distributed Denial of Service (DDoS) detection by traffic pattern analysis",Springer Science+Business Media New York 2012

[9]T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, and J. B.Kopena. Network simulations with the ns-3 simulator. SIGCOMM demonstration, 2008.

[10]S. McCanne and S. Floyd. The LBNL network,simulator. Software on-line: http://www.isi.edu/nsnam,1997. Lawrence Berkeley Laboratory.

[11]I Wayan Krismawan Putra, Yudha Purwanto, Fiky Yosef Suratman,"Perancangan dan Analisis Deteksi Anomaly Berbasis Clustering Menggunakan Algoritma Modified K-Means dengan Timestamp Initialization pada Sliding Window",Universitas Telkom,Bandung.

[12]Riski Pristi Ananto,Yudha Purwnato, dan Astri Novianty,"Deteksi Jenis Serangan Pada Distributed Denial Of Service Berbasis Clustering dan Classification Menggunakan Algoritma Minkowski Weighted K-Means dan Decision Tree",Telkom University.

[13]Tao Peng, Christopher Leckie, Kotagiri Ramamohanarao, “Proactively Detecting Distributed Denial of Service Attacks Using Source IP Address Monitoring,” dalam Networking 2004, Springer Berlin Heidelberg, 2004, pp. 771-782.

[14]Nurdinintya Athari S,S.Si,M.T. ,"Modul Materi Statistika Industri Dan Penilitian Operasional",Laboratorium Sipo Telkom Univeristy

[15]William G. Cochran, "Teknik Penarikan Sampel",1991 Jakarta: UI-Press.

[16]Roscoe, 1975, dikutip dari Uma Sekaran, 2006, "Metode Penelitian Bisnis",Salemba Empat, Jakarta.

[17]Krejcie, R.V. dan Morgan, D.W., 1970, Determining Sample Size for Research Activities,Educational and Psychological Measurements, Vol. 30, pp. 607-610.

[18]Ikhsanul F.L,"Remote Router Untuk Mengatasi Serangan Dalam Trafik Jaringan",2016,Telkom University

[19]Subandijo. (2011). Efisiensi Algoritma dan Notasi O-Besar.ComTech,Vol.2.