Algoritma Fuzzy dan *Reinforcement Learning* dalam Pengambilan Keputusan

Fuzzy Algorithm and Reinforcement Learning for Decision Making

[1]Tia Dianti Hajizah, [2]Yudha Purwanto, [3]Casi Setianingsih

[1,2,3] Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom

[1]tiadianti@student.telkomuniversity.ac.id, [2]omyudha@telkomuniversity.ac.id, [3]setiacasi@telkomuniversity.ac.id

Abstrak

Banyaknya pengguna internet saat ini dapat menyebabkan banyak fenoma-fenoma aneh, yang menjadi salah satu fenoma ialah adanya anomali trafik jaringan internet. Salah satu contoh dari fenomena yang terjadi adalah *flash crowd*, dimana peningkatan akses / trafik ke suatu *server* karena kejadian tertentu. *Denial of Service (DoS)* dan *Distributed Denial of Service (DDoS)* merupakan contoh serangan yang dapat merugikan pengguna ataupun penyelenggara pihak *provider* dengan cara membanjiri trafik jaringan dengan permintaan akses ke suatu host yang dilakukan secara terus menerus, sehingga para pengguna yang sah tidak dapat mengakses host tersebut. Oleh sebab itu diperlukan sebuah sistem yang dapat mencegah sekaligus mengatasi anomali agar anomali tersebut tidak membanjiri arus lalu lintas jaringan.

Pada penelitian tugas akhir ini teknik *Reinforcement Learning* (RL) dan algoritma fuzzy dijadikan cara untuk kasus diatas. RL merupakan sebuah area dari *machine learning* dengan mengandalkan kemampuan sebuah *agent* terhadap lingkungan dengan mengandalkan beberapa gagasan tentang *point* penghargaan *(reward)*. Penggunaan RL dilakukan untuk proses learning terhadap *service* yang dijadikan *agent* untuk terus dikontrol kenaikan anomalinya. Sedangkan algoritma fuzzy digunakan untuk menentukan berapa banyak *service* yang akan dikontrol dalam proses RL

Hasil dari penelitian tugas akhir ini sistem memiliki perfomansi yang baik dalam menangani setiap anomali, agar pada trafik jaringan selanjutnya ada penurunan angka anomali. Kemudian algoritma fuzzy dapat menentukan jumlah *service* yang dikontrol pada prose RL.

Kata kunci: *Reinforcement Learning, Anomaly Traffic, Pre-prcessing*, Algoritma Fuzzy

Abstract

The large number of internet users today can cause many strange phenomas, which became one of the phenoma is the anomaly of Internet network traffic. One example of the phenomenon that occurs is a flash crowd, where increased access / traffic to a server due to certain events. Denial of Service and Distributed Denial of Service are examples of attacks that can harm the users or providers of the provider by flooding the network traffic with requests for access to a host that is done continuously, so that legitimate users can not access the host. Therefore we need a system that can prevent as well as overcome the anomaly so that the anomaly does not flood the flow of network traffic.

In this final project, Reinforcement Learning (RL) technique and fuzzy algorithm are used as the way for the above case. RL is an area of machine learning by relying on an agent's ability on the environment by relying on some idea of reward points. The use of RL is done for the learning process of the service that is used as an agent to keep control of the anomaly increase. While the fuzzy algorithm is used to determine how many services will be controlled in the RL process

The result of this thesis research system has a good perfomansi in handling any anomaly, so that on subsequent network traffic there is decreasing of anomaly number. Then the fuzzy algorithm can determine the number of services controlled on the RL prose.

Keyword: *Reinforcement Learning, Anomaly Traffic, Pre-prcessing*, Algoritma Fuzzy

1. Pendahuluan

Saat ini perkembangan dunia komunikasi semakin pesat, terutama teknologi jaringan internet. Setiap pertukaran informasi yang dilakukan hampir sepenuhnya menggunakan fasilitas jaringan internet. Banyak sekali kemunculan aplikasi seperti *online shopping*, jaringan sosial, dan *cloud computing* yang membuat hubungan antar komputer dan jaringan menjadi hal yang utama. Oleh karena itu internet telah memberikan kemudahan untuk mengakses segala informasi yang beragam dalam kehidupan sehari-hari masyarakat *modern*. Perkembangan internet menyebabkan kenaikan jumlah penggunannya yang dapat memicu terjadinya kepadatan trafik atau

anomali trafik pada jaringan internet. Anomali trafik bisa terjadi karena dua hal, yaitu adanya ancaman atau serangan *flooding traffic* dan serangan *flash crowd.*

Ada banyak jenis serangan ataupun ancaman, khusus penelitian ini hanya melihat dari dua jenis serangan yang dapat menyebabkan anomali trafik. Flash crowd bentuk peningkatan akses yang cukup tinggi ke suatu server yang disebabkan karena kejadian tertentu. Kejadian seperti flooding traffic diakibat oleh *serangan Distributed Denial of Service (DDoS).* Serangan ini terjadi dengan membanjiri lalu lintas jaringan internet dengan banyak data atau dengan meminta request ke suatu *host* atau *service*, sehingga *user* yang berhak tidak dapat mengakses permintaannya karena sudah dibanjiri oleh *flooding traffic*. Jika keadaan anomali terus dibiarkan, akan mengakibatkan kerugian diberbagai pihak, baik dari sisi *user* ataupun penyedia layanan internet itu sendiri.

Dalam mengatasi anomali trafik dibutuhkan suatu sistem untuk melakukan proses penurunan angka anomali trafik menjadi normal. Salah satu teknik yang dipakai untuk melakukan penurunan angka anomali yaitu dengan proses *learning* (pembelajaran). Teknik pembelajaran yang dipakai adalah *Reinforcement Learning* (RL), yang merupakan bagian pembelajaran dari *Machine Learning* (ML). Proses pembelajaran yang terjadi dalam RL ada pada *agent* yang berinteraksi secara langsung terhadap suatu lingkungan. *Agent* akan berinteraksi dengan memilih dan mengeksekusi sebuah aksi. Aksi yang dipilih berupa *service* mana yang mempunyai angka anomali terkecil atau efisien agar tidak membanjiri kepadatan lalu lintas jaringan. Pembelajaran terjadi saat data DARPA (*capture traffic*) difilter setiap 100 data dengan orientasi pemilihian *service* yang angka anomalinya terkecil untuk diloloskan kedalam sebuah jaringan.

2. Kajian Pustaka

2.1. Deteksi Anomali Trafik

Suatu keadaan yang terjadi pada sebuah lalu lintas jaringan yang dapat menyebabkan kondisi menjadi tidak normal disebut sebagai anomali trafik. Mengetahui anomali yang terjadi dengan melihat kenaikan jumlah pengguna internet yang melonjak. Lonjakan yang terjadi bisa secara tidak sengaja ataupun serangan pada trafik jaringan tersebut. Perlu disadarin kenaikan anomali trafik menyebabkan dampak tidak baik bagi beberapa pihak. Kenaikan yang terjadi bisa mengurangi performansi jaringan internet. Pendekatan yang dilakukan untuk masalah deteksi serangan pada jaringan komputer dikenal sebagai *Intrusion Detection System (IDS).*

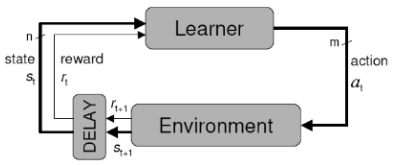
IDS adalah sebuah sistem keamanan jaringan yang dirancang untuk melayani ketersediaan layanan dan menjaga tingkat intergritas bagi seluruh pengguna jaringan internet. Ada dua metode dalam IDS/IPS untuk melindungi *host* dan *service* dari serangan DoS dan DDoS, yaitu *intrusion-signature* based dan *traffic anomaly based*. Metode *intrusion-signature based* mendeteksi adanya intrusi sesuai dengan database yang telah ada, karena intrusi sebelumnya sudah tersimpan dalam *database.*

Ketika ada intrusi datang akan dicocokan dengan database intrusi sebelumnya. Sedangkan metode *traffic anomaly based* mendeteksi intrusi berdasarkan keadaan anomali trafik yang tidak biasanya. Tentunya kedua metode ini memiliki kekurangan masing-masing, seperti metode *intrusion-signature based*, jika ada model serangan baru kemungkinan IDS tidak bisa mengenali serangan tersebut. Sedangkan metode *traffic anomaly based* mempunyai kekurangan, jika keadaan anomali yang terjadi memang dikarenakan adanya kenaikan jumlah *user* yang mengakses pada suatu *destination.*

2.2 Reinforcement Learning

Reinforcement Learning (RL) adalah salah satu jenis proses *learning* dari *Machine Learning (ML)*, yang melakukan pendeketan belajar dengan trial and error untuk mencapai tujuan. Oleh karena itu RL memerlukan *reward* dari lingkungannya sebagai pengganti data respon *input* dan *output*nya. *Reward* digunakan untuk menguji *state* lingkungan, pengumpulan jumlah *reward* secara maksimal sangat penting karena *reward* menjadi *signal feedback* dalam proses *learning*.

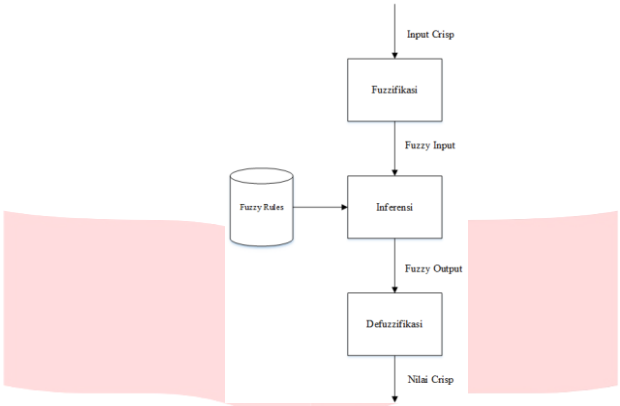
Ada beberapa elemen utama yang muncul pada RL, yaitu *learner, environment, dan reward*. *Learner* disebut sebagai algoritma RL, karena *learne*r akan berinteraksi dengan lingkungan *(environment). Learner* akan mempelajari segala kondisi *state* lingkungannya dan segera berinteraksi dengan lingkungannya. Cara learner berinteraksi dengan lingkungannya adalah dengan memberian aksi kepada lingkungannya *(environment).* Kemudian *environment* akan berinteraksi pada aksi yang diberikan *learner* dengan state baru dan *reward. Reward* adalah suatu nilai yang dibangkitkan/diberikan oleh fungsi *reinforcement* yang menguji *current state* dan aksi terakhir. Berikut adalah gambar diagram interaksi sistem dari RL:



Gambar 2.1 Diagram interaksi antara *Learner* dan *Environment* [12]

2.3 Algoritma Fuzzy

Logika fuzzy adalah suatu metodologi yang dapat menyelesaikan permasalahan dalam pengendalian sebuah sistem dengan menyediakan kesimpulan pasti dari infromasi yang tidak pasti, ambiguitas, samar-samar, atau tidak tepat. Oleh sebab itu, logika fuzzy digunakan karena konsepnya yang mudah dimengerti, sangat fleksibel, dan memiliki toleransi terhadap data yang tidak pasti/tepat. Logika fuzzy menjembatani bahasa mesin yang presisi dengan bahasa manusia yang menekankan arti secara signifikan. Sistem Berbasis Pengetahuan (SBP) Fuzzy memiliki 3 tahapan dalam penyelesaian masalah. Pada gambar 2.2 memperlihatkan alur yang ada pada SBP Fuzzy:



Gambar 2.2 SBP Fuzzy [4]

2.3.1 Fuzzyfikasi

pada fuzzyfikasi adalah tahap yang mengubah suatu nilai input tegas (crisp set) menjadi nilai fuzzy (*variable linguistic)* yang mempunyai fungsi kenggotaannya masing-masing. Fungsi keanggotaan adalah fungsi yang digunakan untuk memetakan nilai inputan *crisp* menjadi derajat keanggotaan pada variabel *linguistic* yang telah ditentukan. Interval derajat kenaggotaan bernilai [0,1].

2.3.2 Inferensi

Dalam tahap ini akan dilakukan simulasi pengambilan keputusan manusia berdasarkan konsep fuzzy menggunakan *rules of knowledge*. *Input* yang diproses sesuai dengan *rules* yang sudah dibuat, akan mengeluarkan *output* fuzzy. Rules yang digunakan berbentuk *IF-THEN*. Kondisi *IF-THEN* disesuaikan dengan *input* yang diberikan.

2.3.3 Defuzzyfikasi

Merupakan tahap terakhir yang dilakukan dalam proses Sistem Berbasis Pengetahuan (SBP). Pada defuzzyfikasi terjadi proses pengembalian nilai dari himpunan fuzzy pada bagian konklusi menjadi nilai tegas (crisp). Ada beberapa metode yang digunakan pada proses defuzzyfikasi, seperti *Centroid Method (Centre of Gravity), Height Method, first (or last) of Maxima, Mean-Max Method, dan Weighted Average*. Untuk fuzzy inferensi digunakan model Mamdani, sehingga metode yang dipakai untuk defuzzyfikasi adalah *Centroid Method*. Berikut adalah persamaan yang dirumuskan secara umum [5]:

$$\mu(x) = \frac{\sum_{i=1}^{n} x_i \mu(x_i)}{\sum_{i=1}^{n} \mu(x_i)} \quad \ldots 2.1$$

2.1 Rumus *Centroid Defuzzification* [5]

3. Perancangan Sistem

3.1. Gambaran Umum Sistem

Pada tugas akhir ini, berfokus pada dua hal yaitu *Reinforcement Learning* (RL) dan algoritma fuzzy untuk membangun sebuah sistem yang dapat mengontrol dan mengurangi anomali pada trafik jaringan. Gambaran umum sistem dapat dilihat pada gambar 3.1:

Gambar 3.1 Alur Skenario Sistem

3.2 Dataset Initialization

Sebelum masuk kedalam sistem learning dan algoritma fuzzy, hal pertama yang dilakukan adalah initialisasi dataset. Dataset yang digunakan adalah DARPA'98 berupa dataset *capture real time traffic* normal yang diduga didalamnnya terdapat anomali. Keadaan anomali tersebut yang menjadi perhatian khusus dalam tugas akhir ini, dengan menggunakan dataset tersebut, dapat menentukan suatu prediksi anomali pada trafik selanjutnya, kemudian hasil prediksi tersebut akan diproses kedalam sistem *learning* dengan teknik RL dan algoritma fuzzy. Agar pada trafik-trafik yang akan datang, jumlah ataupun laju dari anomali dapat berkurang sehingga keadaan arus lalu lintas jaringan menjadi stabil.

3.3. Preprocessing

Pre-processing dilakukan agar data mentah dari dataset DARPA'98 menjadi data *input* yang baik untuk proses *learning* dengan *reinforcement learning*. Pre-proses yang dilakukan adalah dengan pemilihan beberapa *service* dengan suspek yang laju anomalinya tinggi. Berikut adalah tabel fitur yang ada pada dataset DARPA'98:

Tabel 3.1 Fitur-Fitur Dataset DARPA 1998

No	Nama Fitur	Keterangan
1	No	Nomor urut data trafik
2	*Start Date*	Tanggal mulai akses
3	*Start Time*	Waktu mulai akses
4	*Duration*	Durasi akases antar request
5	*Service*	Layanan yang diminta
6	*Src Port*	Port sumber/asal
7	*Dest Port*	Port tujuan
8	*Src IP_Address*	IP address sumber/asal
9	*Dest IP_Address*	IP address tujuan
10	*Attack Score*	0= Normal 1= Attack
11	*Attack Name*	Nama / Jenis Serangan

Dataset DARPA'98 yang dipilih berupa data *testing*, data t*esting* tersebut berupa data *capture* trafik secara *real time* pada pertengahan tahun 1998. *Capture* data trafik dilakukan selama 2 minggu, mulai dari hari senin sampai dengan hari jumat. Jumlah data trafik yang ter-*capture* dalam setiap harinya berjumlah tak menentu, oleh sebab itu data trafik per hari difilter berdasarkan 5 *service* dengan laju anomali tertinggi.

3.3 Perancangan Sistem Reinforcement Learning

Tugas akhir ini menggunakan salah satu algoritma dari RL, yaitu *Q-Learning. Q-learning* adalah salah satu *model-free learning* dalam Teknik RL. Penggunaan *Q-learning* untuk mencari nilai optimal dari *Q-value (action value function*). Selama proses *learning* nilai Q akan terus ter*update*, dari nilai Q yang lama menjadi nilai Q yang baru. Setiap perubahan nilai Q bergantung pada pemilihan aksi yang ada pada *service*. Untuk meng*update* nilai Q terdapat persamaan, yaitu sebagai berikut:

$$\boldsymbol{Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Max_a Q(s',a) - Q(s,a)] \ldots 3.1}$$

Gambar 3.2 Fungsi Q-Value

Q-learning mempunyai cara kerja dengan melakukan evaluasi pada setiap episode, proses dalam satu episode dikatakan berakhir jika *agent* sudah mencapai titik *goal state*, dan setiap melakukan *action* akan mempengaruhi nilai Q. Nilai Q dijadikan *"brain"* oleh *agent* selama proses *learning*. Untuk lebih jelasnya berikut adalah gambar mengenai algoritma *Q-Learning*:

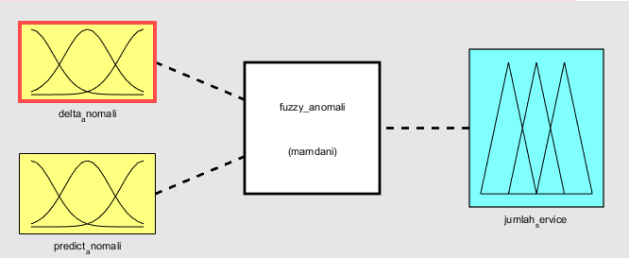
1. *Set* parameter *reward* +/-, jumlah service, *threshold*, α (diseparasi trafik)
2. Set Q-value awal (zero atau 100)
3. For each episode, (1 episode = 100 trafik)
a. Open service

b. *Do, for every traffic* i *Open service, traffic passed* ii *Received reward, depend on traffic passed* iii *Next traffic* iv *Update Q-value* $Service_{point} \leftarrow Service_{point} + \alpha[r + \gamma Max_a Q(s', a) - service_{point}]$
c. *End Do*
4. *End For*

Gambar 3.4 Algoritma Q-learning [9]

3.2 Perancangan Sistem Fuzzy

Tahap pre-proses dengan filtering 5 service pada laju anomali tertinggi akan berpengaruh pada hasil yang dikeluarkan oleh sistem fuzzy. Sistem fuzzy dirancang melalui tiga proses untuk mengetahui banyaknya service yang akan dikontrol secara otomatis pada proses learning didalam R, yaitu fuzzyfikasi, inferensi, dan defuzzyfikasi.



Gambar 3.5 Perancanngan sistem fuzzy

4. Pengujian dan Analisis

Pengujian dibagi menjadi 2 tahap, tahap pertama melakukan training dan tahap kedua adalah testing. Masing-maisng menggunakan dataset training DARPA 1998 dan dataset testing DARPA 1998. Selama proses training sistem RL hanya melakukan learning dalam setiap harinya, hasil dari setiap training akan menjadi acuan saat testing dilakukan.

4.1 Pengujian Training dan Testing

Training adalah tahap untuk mengetahui segala jenis kondisi lingkungan trafik pada setiap harinya. Dimulai dari week 5 dan week 6 untuk melakukan training. Ada 10 hari training, tetapi tidak semuanya mengandung anomali , oleh sebab itu hanya beberapa hari saja yang memiliki kandungan anomali. Hasil training akan menjadi acuan saat proses testing. berikut adalah salah satu hasil training yang melakukan penuruan anomali trafik dihari tertentu:

Tabel 4.1 Hasil Training Week 5 - Thursday

hasil thursday 5				
Status		ftp	ftp-data	telnet
before	normal	145	1625	89
	anomali	614	1041	849
lolos	normal	124	1603	84
	anomali	246	444	259
selisih/terbuang	normal	21	22	5
	anomali	368	597	590
persentase	normal	14%	1%	6%
	anomali	60%	57%	69%

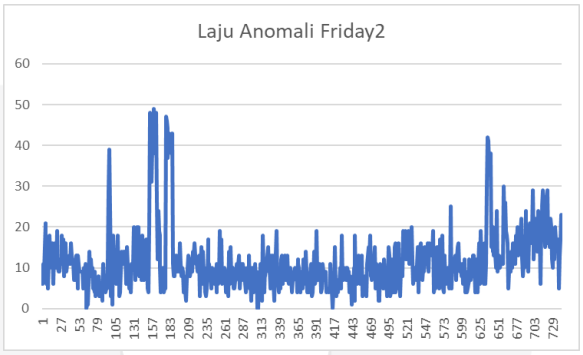
pada gambar 4.1 menampilkan hasil training yang dilakukan dalam satu hari yaitu, hari kamis week 5. Terjadi penurunan anomali dari masing-maisng service, sebesar 60% untuk service ftp, 57% untuk service ftp-data, dan 69% untuk service telnet. Sesuai dengan scenario pengujian, akan dilakukan testing pada

dengan kondisi lingkungan trafik yang serupa, dan melihat bagaimana hasil dari proses testing. berikut adalah hasil dari proses testing:

Tabel 4.2 Hasil Testing Week 2 - Friday

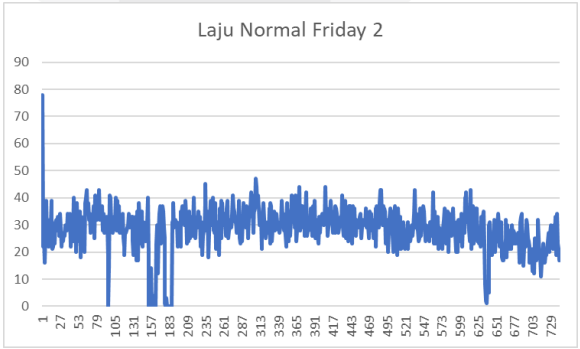
hasil testing Friday 2				
Status		ftp	ftp-data	telnet
before	normal	116	983	199
	anomali	88	152	1016
lolos	normal	40	396	83
	anomali	25	59	428
selisih/terbuang	normal	76	587	116
	anomali	63	93	588
persentase	normal	66%	60%	58%
	anomali	72%	61%	58%

Pada gambar 4.2 menampilkan hasil testing yang dilakukan pada hari jumat. Trafik anomali unggul dari service telnet, sehingga terjadi penuruan sebesar 58% dari trafik anomali service telnet. Pengambilan data testing, dilihat berdasarkan data training pada hari jumat di week 5 dan jumat week 6 yang memiliki kondisi anomali yang serupa, yaitu service telnet lebih banyak membawa anomali dibandingkan dengan service lainnya. Untuk melihat laju trafik anomali yang lolos pada data testing dihari jumat adalah sebagai berikut:



Gambar 4.1 Laju Anomali Lolos

Gambar 4.1 merepresentasikan grafik laju anomali yang lolos kedalam sebuah jaringan, terlihat dibeberapa episode tertentu terdapat kanaikan jumlah anomali, hal tersebut dikarenakan adanya anomali dari service lainnya selain ke 3 service yang dipantau yaitu ftp, ftp-data, telnet. Seperti service snmp/u yang memeloloskan anomali sejumlah 5719 anomali.



Gambar 4.2 Laju Normal Lolos

Gambar 4.2 menunjukan trafik normal yang berhasil diloloskan oleh RL selama proses learning. Trafik yang diloloskan baik dari ke 3 service ataupun service lainnya yang tidak dipantau oleh RL. Pada awal-awal proses testing dilakukan, jumlaha trafik normal melonjak hingga 80, dan seterusnya lebih menstabilkan kondisi trafik. Antara laju anomali dan laju normal yang lolos kedalam sebuah jaringan

memiliki rata-rata sebesar, laju anomali 11,47 dan laju normal sebesar 28,46. Dengan rata-rata penuruan anomali sebesar 64%.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari pene;itian tugas akhir ini adalah sebagao berikut:

1. Penurunan anomali trafik dapat dilakukan dengan menggunakan Teknik Reinforcement Learning (RL).Perubahan *learning rate* dapat berpengaruh terhadap hasil penurunan anomali.
2. Algoritma fuzzy dapat membantu sebagai pengontrol service pada proses learning selanjutnya
3. Sistem yang dibuat melakukan proses RL dan fuzzy secara terpisah. Proses fuzzy hanya dilakukan sebagai pengontrol antar satu episode ke episode berikutnya.
4. Penentuan reward baik positif ataupun negative dapat berpengaruh dalam hasil learning, semakin besar nilai reward negative terdapat potensi besar untuk cepat menutup service. Jika nilai reward besar akan berpotensi untuk membuka service

5.2 Saran

Adapun beberapa saran yang nantinya akan menjadi perbaikan pada penelitian selanjutnya, dikarenakan masih adanya kekurangan dalam penelitian ini:

1. Ada baiknya jika proses *learning* dilakukan secara *real time*, atau dengan *router* asli.
2. *Reinforcement learning* sepertinya akan lebih baik juka digabungkan dengan algoritma lainnya seperti fuzzy *genetic* ataupun fuzzy *Q-Learning*.
3. Agent dibuat *multi-agent*, sehingga proses learning menjadi semakin baik dengan banyak jumlah *agent* yang bekerja dalam penuruan anomali trafik.

Daftar Pustaka

[1] K. H. B. R. Yudha Purwanto, "Traffic Anomaly Detection in DDos Flooding Attack," in THE 8TH INTERNATIONAL CONFERENCE ON TELECOMMUNICATION SYSTEM,SERVICES,AND APPLICATION, 2014.

[2] H. O. S.-H. K. Jae-Hyum jun, "DDoS Flooding Attack Detection Through a Step-By-Step Investigation," in IEEE, Daegu, South Korea, 2011.

[3] M. TECHNOLOGY, "LINCOLN LABORATORY," [Online]. Available: http://www.ll.mit.edu/ideval/data/. [Accessed 9 May 2017].

[4] S. G. Hans Bandemer, Fuzzy Sets Fuzzy Logic Fuzzy Methods with Applications, Berlin: John Wiley & Sons, 1993.

[5] R. Munir, "Pengantar Logika Fuzzy," Teknik Informatika-STEI ITB, Bandung.

[6] K. Malialis, Distributed Reinforcement Learning for Network Intrusion Response, UNIVERSITY OF YORK COMPUTER SCIENCE, 2014.

[7] H. G. Kayacik and A. N. Zincir-Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion De tection Datasets," in Proceedings of the IEEE, Atlanta, USA, 2005.

[8] E. Alpaydin, Introduction to Machine Learning, Second Edition, London: The MIT Press Cambridge, Massachusetts, 2010.

[9] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, London: The MIT Press Cambridge, 2012.

[10] C. Szepesvari, Algorithms for Reinforcement Learning, Morgan & claypool Publisher, 2009.

[11] M. H. Tom Mitchell, Machine Learning, 1997.

[12] M. Carreras, J. Yuh and B. J, "A Neural-Q learning approach for online robot behavior learning," in The Autonomous Systems Laboratory, Hawaii (USA), 2003.

[13] M. V. Otterlo and M. Wiering, Reinforcement Learning and Markov Decision Processes.

[14] R. J. Williams, Reinforcement LEarning and Markov Decision Processes, Spring, 2007.