IMPLEMENTASI ALGORITMA Y-MEANS SEBAGAI ANOMALY DETECTION (STUDI KASUS: INTRUSION DETECTION SYSTEM)

Mira Afianti[1], Deni Saepudin[2], Tribroto Harsono.[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. Pada anomaly detection terdapat pendekatan clustering based dimana data akan dikelompokkan menjadi beberapa cluster. Cluster intrusi adalah cluster dengan jumlah anggota yang sedikit. Algoritma clustering yang cukup dikenal adalah K-Means karena mudah diimplementasikan dengan kompleksitas yang rendah. Akan tetapi terdapat beberapa kekurangan pada algoritma tersebut. Algoritma Y-Means adalah algoritma clustering yang dibangun untuk memperbaiki kekurangan K-Means. Data trafik jaringan akan dikelompokkan ke beberapa cluster, lalu dilihat cluster mana yang merupakan cluster intrusi berdasarkan threshold. Pengujian dilakukan dengan beberapa skenario untuk mengetahui akurasi sistem dilihat dari nilai detection rate dan false positive rate, pengaruh besar konstanta merging terhadap jumlah cluster akhir, dan juga pengaruh konstanta merging dan threshold terhadap nilai akurasi. Y-Means dapat mendeteksi intrusi dengan tingkat akurasi yang cukup baik dilihat dari nilai detection rate sebesar 92.46%. Untuk nilai false positive rate Y-Means menunjukkan akurasi yang tidak terlalu buruk yaitu sebesar 9.69%.

Kata Kunci : intrusi, clustering, anomaly detection, Y-Means

Abstract

Intrusion detection systems (IDS) is a system that can detect any intrusion or attack on a network or information systems. Anomaly detection is one of the method in IDS. In anomaly detection, network traffic data is detected as intrusion if it has different characteristics from most of data. There are clustering based approach in anomaly detection where data will be grouped into several clusters. Intrusion cluster is the cluster with a small number of members. One of well-known clustering algorithm is K-Means because it is easy to implement and has low complexity. However there are some shortcoming in that algorithm. Y-Means is a clustering algorithm that is built to solve the shortcoming of K-Means. Network traffic data will be grouped into several clusters, and there will be intrusion cluster which can be seen by the threshold. Test carried out with several scenarios to determine the accuracy of the system based on the value of detection rate and false positive rate, the influence of the merging variable on the number of final cluster and also the influence of merging variable and threshold value on accuracy. Y-Means can detect intrusions with fairly good accuracy based on detection rate (92.46%). From false alarm value, Y-Means accuracy is not too bad (9.69%).

Keywords : intrusion, clustering, anomaly detection, Y-Means

1. Pendahuluan

1.1 Latar Belakang

Intrusi adalah akses yang tidak sah atau tindakan yang membahayakan pada sebuah komputer ataupun sistem informasi [3]. Ketika internet menyebar ke segala penjuru dunia seperti sekarang ini, komputer akan menjadi sangat rentan dari berbagai intrusi dari *World Wide Web*. Untuk itu sangat diperlukan *intrusion detection system (IDS)* yang handal untuk melindungi komputer dari aktivitas-aktivitas yang membahayakan tersebut.

Terdapat dua paradigma utama pada IDS berbasiskan *data mining* yaitu *misuse detection* dan *anomaly detection*. Pada *misuse detection*, setiap data pada sebuah dataset mempunyai label normal atau intrusi, dan algoritma *learning* akan dilatih dengan menggunakan data berlabel [2]. Keuntungan utama dari teknik *misuse detection* adalah tingkat akurasi yang tinggi dalam mendeteksi serangan yang telah dikenal dan berbagai variasinya. Kelemahannya adalah ketidakmampuan untuk mendeteksi serangan yang belum diamati [9][2].

Metode kedua adalah *Anomaly Detection*. Pada metode ini, intrusi akan dideteksi dengan cara mengenali aktivitas-aktivitas normal yang biasanya berlangsung. Kelakuan normal dari *user* dan aktivitas sistem akan dirangkum ke dalam sebuah profil normal yang kemudian akan dijadikan sebuah tolok ukur. Peringatan terjadinya intrusi akan muncul ketika timbul suatu aktivitas yang berbeda dengan profil normal. Kelebihan metode ini adalah dapat mendeteksi bentuk serangan yang baru dengan asumsi bahwa serangan akan selalu menyimpang dari aktivitas normal [2].

Pada *anomaly detection* terdapat beberapa teknik, salah satunya adalah *outlier detection scheme* yaitu mendeteksi anomali dengan cara mengelompokkan data dan mengidentifikasikan data tersebut apakah merupakan deviasi dari sebuah kelakuan normal. Salah satu pendekatan dalam *outlier detection scheme* adalah *clustering based approach* [2]. *Clustering* merupakan suatu metode pengelompokan objek-objek ke dalam beberapa sub-kelas berbeda yang mempunyai karakteristik tersendiri sehingga anggota dari tiap sub-kelas memiliki kemiripan, dan anggota dari sub-kelas yang berbeda juga berbeda satu sama lainnya. Oleh karena itu, metode *clustering* bisa digunakan untuk mengklasifikasi log data dan mendeteksi intrusi. Jumlah data pada aktivitas normal biasanya jauh lebih besar daripada jumlah intrusi, maka dari itu populasi dari *cluster* normal juga lebih besar daripada *cluster* intrusi [3].

Algoritma *clustering* yang cukup banyak dipakai adalah K-Means, yaitu membagi data menjadi *k* buah *cluster* dimana nilai *k* ditentukan di awal. K-Means terkenal atas kesederhanaannya dan kompleksitas waktu yang cukup rendah [9]. Akan tetapi terdapat tiga kekurangan pada algoritma ini. Yang pertama adalah ketergantungan terhadap inisialisasi awal *centroid* yaitu hasil akhir *clustering* sangat dipengaruhi oleh pemilihan *centroid* di awal [9]. Yang kedua adalah ketergantungan terhadap jumlah *cluster*, dimana nilai *k* pada inisialisasi awal akan sangat mempengaruhi hasil *clustering*. Kekurangan terakhir adalah *degeneracy*, dimana hasil akhir *clustering* mungkin saja menghasilkan *cluster* kosong[9][3].

Y-Means merupakan metode *clustering* yang merupakan pengembangan dari K-Means. Perbedaan yang mendasar dari kedua algoritma ini adalah kemampuan Y-Means untuk secara otomatis memilih jumlah *cluster* berdasarkan dari keadaan statistik data. Sehingga inisialisasi nilai *k* di awal tidak bersifat mutlak ke hasil akhir *cluster* karena nilai *k* akan berubah seiring iterasi algoritma Y-Means.

Untuk mengukur keakuratan dari kedua algoritma ini digunakan dua buah parameter yaitu *detection rate* (DR) dan *false positive rate* (FPR). DR adalah jumlah *instance* intrusi yang berhasil dideteksi oleh sistem dibagi dengan total jumlah *instance* intrusi yang ada pada dataset. FPR adalah total jumlah *instance* normal yang salah diklasifikasikan sebagai intrusi dibagi dengan total jumlah *instance* normal [9][3]. Pada sistem deteksi intrusi yang baik, diharapkan diperoleh nilai DR yang setinggi-tingginya dan FPR yang sekecil-kecilnya.

Sudah banyak penelitian yang menjadikan *anomaly detection* sebagai objek penelitiannya. Misalnya pada penelitian Ali A. Ghorbani[9] yang mengimplementasikan algoritma Y-Means dengan mendapatkan DR sekitar 71% dan FPR sekitar 9%. Pada penelitian Y-Means lainnya yang dilakukan oleh Sivanadiyan S.Kannan [9] diperoleh nilai DR 86.63% dan FPR 2,72%.

1.2 Perumusan Masalah

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

1. Bagaimana mengimplementasikan metode *clustering* Y-Means untuk menentukan *cluster-cluster* anomali?
2. Bagaimana mengevaluasi keakuratan hasil prediksi dari sistem ini jika dilihat dari *detection rate* dan *false positive rate*?

Sedangkan yang menjadi batasan masalah dalam tugas akhir ini adalah :

1. Data trafik jaringan yang digunakan bersifat *offline*, yaitu data KDD Cup 1999 Data yang diambil dari http://kdd.ics.uci.edu/.
2. Asumsi pada dataset, data intrusi jumlahnya jauh lebih sedikit dari data normal. Dimana persentase data intrusi tidak lebih dari 2%.

1.3 Tujuan

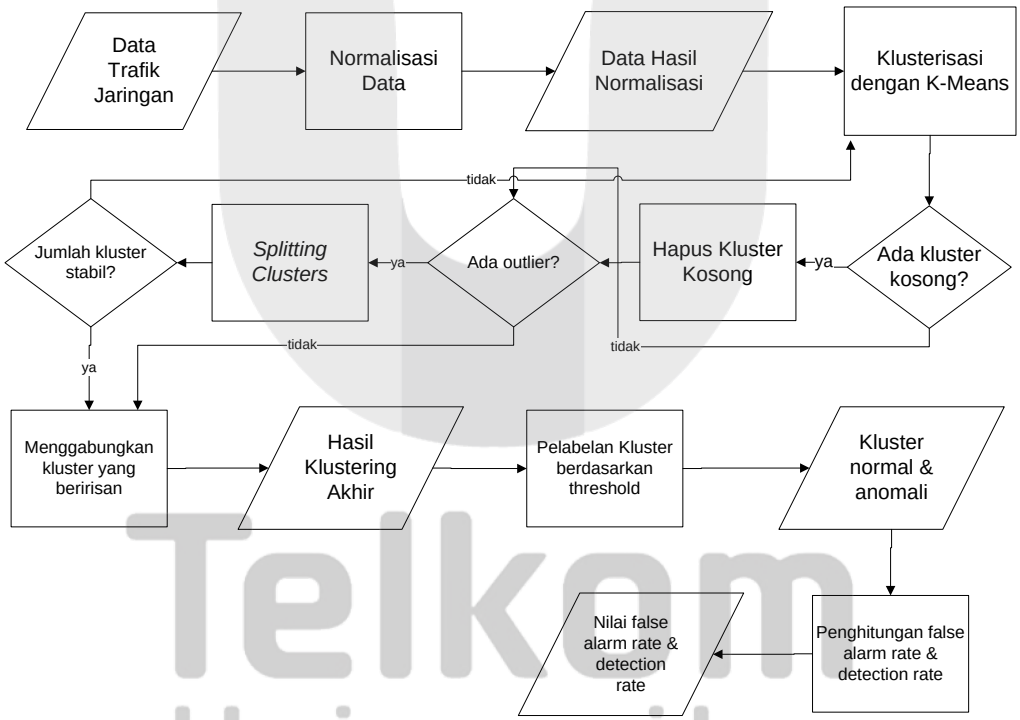
Tujuan yang ingin dicapai dalam tugas akhir ini yaitu :

1. Mengimplementasikan algoritma Y-Means pada metode *outlier detection scheme* dalam mendeteksi anomali yang terjadi pada jaringan.
2. Menentukan tingkat keakuratan dari nilai *detection rate* dan *false positive rate* yang ada.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi literatur
   Pencarian referensi dan sumber-sumber yang berhubungan dengan *intrusion detection system*, *outlier detection scheme*, *anomaly detection*, dan metode algoritma Y-Means dan K-Means dalam menyelesaikan tugas akhir ini.
2. Analisis dan perancangan sistem
   Melakukan analisis dan perancangan terhadap sistem yang dibangun, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan, termasuk menentukan bahasa pemrograman yang digunakan, arsitektur, fungsionalitas, dan antarmuka sistem. Input sistem berupa data uji yaitu data trafik jaringan. Output dari sistem adalah *cluster* normal dan *cluster* anomali serta hasil analisis dari nilai *detection rate* dan *false positive rate* yang didapatkan. Alur sistem dapat dilihat pada Gambar 1.1 berikut:



Gambar 1-1: Alur Algoritma Y-Means

3. Implementasi dan pembangunan sistem
   Melakukan implementasi dari hasil analisis dan perancangan sistem terhadap metode yang digunakan serta mengevaluasi apakah sistem ini dapat mendeteksi intrusi dengan baik.
4. Pengujian dan analisis
   Menguji sistem berdasarkan beberapa skenario yang dibuat kemudian menganalisis hasil pengujian tersebut.
5. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. Penutup

5.1 Kesimpulan

1. Semakin besar nilai konstanta *merging*, maka jumlah *cluster* final, nilai *Detection Rate* (DR) dan *False positive rate* (FPR) yang dihasilkan akan semakin kecil
2. Semakin kecil nilai *threshold*, maka nilai DR dan FPR juga akan semakin kecil.
3. Proses konversi data *symbolic* ke *numeric* dengan menggunakan *indicator variable* mempunyai performansi yang lebih baik dibandingkan dengan dataset yang data *symbolic*nya dibuang, dilihat dari segi besar nilai DR.

5.2 Saran

1. Mencoba algoritma *anomaly detection* lain yang dapat menghasilkan performansi yang lebih baik dan akurat dengan komputasi yang rendah misalnya algoritma *clustering squeezer*.
2. Untuk penelitian selanjutnya coba dilakukan proses pendeteksian intrusi terhadap data log jaringan secara *realtime*.
3. Mencoba mencari metode konversi data *symbolic* ke *numeric* dengan cara selain menggunakan *indicator variable*, misalnya metode *conditional probabilities* atau *SSV (separability split value) critetion.*

Referensi

[1] Arnold Andrew, Eskin Eleazar, Portnoy Leonid, dkk. 2002. "A Geometric Framework For Unsupervised Anomaly Detection: Detecting Intrusions In Unlabeled Data", *Data Mining for Security Applications*, Kluwer Academic, Boston, Mass, USA.

[2] Eleazar Eskin, Portnoy Leonid, dkk. 2001. "Intrusion Detection with Unlabeled Data Using *Clustering*", *Proceedings of ACM CSS Workshop on Data Mining Applied to Security*, Columbia. pp 5-8.

[3] 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.

[4] Ghorbani A., Onut Iosif-Viorel. 2010. "Y-Means: An Autonomous *Clustering* Algorithm", *Lecture Notes in Computer Science*, Volume 6076/2010, 1-13.

[5] Han, Kamber. 2001. *Data Mining: Concept and Techniques.* Morgan Kaufmann Publishers.

[6] Hernández- Pereira E., Suárez-Romero J. 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.

[7] Kannan, Sivanadiyan Sabari. 2005. *Y-Means Clustering Vs N-CP Clustering With Canopies for Intrusion Detection.* Thesis. Oklahoma State University.

[8] Lakhina Shilpa, Joseph Sini, Verma Bhunderpa. 2010. "Feature Reduction using Principal Component Analysis for Effective Anomaly-Based Intrusion Detection on NSL-KDD". *International Journal of Engineering Science and Technology*, Vol. 2(6), pp 1790-1799.

[9] Lazaveric Aleksandar, Jaideep Srivastava, dkk. 2002. "Data Mining for Network Intrusion Detection". *Proc. NSF Workshop on Next Generation Data Mining*, pp 21-30.

[10] 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).*

[11] Tan. Pang-Ning, Steincbach. Michael, Kumar. Vipin, 2006, *Introduction to Data Mining*, Pearson Education Inc.

[12] 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.