

<http://www.ftsm.ukm.my/apjitm>

Asia-Pacific Journal of Information Technology and Multimedia

Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik

Vol. 6 No. 1, June 2017: 57 - 69

e-ISSN: 2289-2192

GLOBAL AND LOCAL CLUSTERING SOFT ASSIGNMENT FOR INTRUSION DETECTION SYSTEM: A COMPARATIVE STUDY

MOHD RIZAL KADIS
AZIZI ABDULLAH

ABSTRACT

Intrusion Detection System (IDS) plays an important role in computer network defence mechanism against malicious objects. The ability of IDS to detect new sophisticated attacks compared to traditional method such as firewall is important to secure the network. Machine Learning algorithm such as unsupervised learning and supervised learning is capable to solve the problem of classification in IDS. To achieve that, KDD Cup 99 dataset is used in experiments. This dataset contains 5 million instances with 5 different categories which are Normal, DOS, U2R, R2L and Probe. With such a large dataset, the learning process consumes a lot of processing times and resources. Clustering is unsupervised learning method that can be used for organizing data by grouping similar features into same group. In literature, many researchers used global clustering approach whereby all input will be combined and clustered to construct a codebook. However, there is an alternative technique namely local clustering approach whereby the input will be split into 5 different categories and clustered independently to construct 5 different codebooks. The main objective of this research is to compare the classification performance between the global and local clustering approaches. For this purpose, the soft assignment approach is used for indexing on KDD input and SVM for classification. In the soft assignment approach, the smallest distance values are used for attack description and RBF kernel for SVM to classify attack. The results show that the global clustering approach outperforms the local clustering approach for binary classification. It gives 83.0% of the KDD Cup 99 dataset. However, the local clustering approach outperforms the global clustering approach on multi-class classification problem. It gives 60.6% of the KDD Cup 99 dataset.

Keywords: Intrusion Detection System, Soft Assignment, Global and Local Clustering Approaches, Codebook, KDD Cup 99 Dataset.

UMPUKAN LEMBUT KLUSTER SEJAGAT DAN SETEMPAT BAGI SISTEM PENGESANAN PENCEROBOHAN: SATU PERBANDINGAN

ABSTRAK

Sistem Pengesanan Pencerobohan (SPP) berperanan penting sebagai satu mekanisme pertahanan sistem komputer daripada serangan pengguna hasad. Berbanding dengan kaedah tradisional iaitu tembok api, SPP yang berasaskan anomali berkebolehan mengesan serangan yang sofistikated dengan membanding aktiviti normal dan anomali. Kaedah Pembelajaran Mesin (PM) seperti Pembelajaran Mesin Tanpa Penyelia (PMTP) dan Pembelajaran Mesin Berpenyelia (PMB) mampu menyelesaikan masalah pengkelasan dalam SPP berasaskan anomali. Set data KDD Cup 99 dipilih sebagai bahan ujikaji dengan set data mengandungi 5 juta rekod dengan pecahan 5 kategori iaitu Normal, DOS, U2R, R2L dan Probe. Saiz data yang besar memerlukan masa proses yang lama dan sumber yang banyak terutamanya semasa proses latihan. Teknik kluster diguna bagi mengecil saiz data dengan menyatu rekod yang mempunyai kemiripan fitur yang sama ke dalam satu kumpulan, dikenali sebagai kluster. Kebanyakan pengkaji mengguna pendekatan kluster sejagat dengan proses kluster dijalankan ke atas semua rekod. Sebagai alternatif, satu pendekatan yang dinamai kluster setempat, memisah rekod kepada 5 kategori yang berlainan. Proses kluster dijalankan mengikut kategori tersebut. Kajian ini bertujuan membanding prestasi model pengkelasan yang terhasil daripada pendekatan kluster sejagat dan setempat. Bagi mencapai objektif tersebut, pendekatan umpukan lembut diguna untuk pengindeksan data KDD dan Mesin Sokongan Vektor (SVM). Dalam pendekatan umpukan lembut, nilai jarak terendah diguna untuk penghuraian serangan dan kernel RBF untuk SVM diguna sebagai pengkelas. Kajian mendapati pengkelasan perdua melalui pendekatan kluster sejagat adalah lebih baik berbanding pendekatan kluster setempat dengan 83.0% ketepatan daripada set data KDD Cup 99. Bagi

pengkelasan berbilang kelas pula, pendekatan kluster setempat adalah lebih baik berbanding pendekatan kluster sejagat dengan memberi 60.6% ketepatan purata daripada set data KDD Cup 99.

Kata Kunci: Sistem Pengesanan Pencerobohan, Umpukan Lembut, Kluster Sejagat dan Setempat, Beg-Fitur, Set Data KDD Cup 99

PENGENALAN

Sistem Pengesanan Pencerobohan (SPP) memainkan peranan penting dalam membendung dan menghentikan penyalahgunaan rangkaian oleh entiti yang tidak bertanggungjawab seperti pengguna hasad. SPP mempunyai dua kaedah bagi mengesan penyalahgunaan rangkaian iaitu pengesanan berasaskan anomali dan tandatangan. Kaedah pengesanan berasaskan anomali dilaksanakan dengan membanding aktiviti yang kelakuannya diketahui normal sebelumnya dengan aktiviti yang sedang diperhati pada sistem komputer (Vigna et al., 2003). Manakala kaedah pengesanan berasaskan tandatangan dilaksanakan dengan membanding aktiviti yang sedang diperhati dengan ciri-ciri unik yang terdapat pada sesuatu aktiviti terdahulu yang dikenal pasti sebagai aktiviti tidak sah. Kaedah pengesanan berasaskan tandatangan mempunyai kelemahan mengesan aktiviti serangan yang baharu kerana ciri-ciri uniknya belum terkandung dalam pangkalan data SPP. Manakala kelemahan dalam kaedah serangan berasaskan anomali adalah tersalah jangkaan di antara aktiviti yang sah atau sebaliknya.

Penyelidikan SPP berasaskan anomali julung kalinya dipertanding dalam *The Third International Knowledge Discovery and Data Mining Tools* pada tahun 1999. Set data KDD Cup 99 (KDD) diguna secara rasmi sebagai set data piawai ujikaji keberkesanan algoritma pengesanan yang dibangun oleh peserta. Set data KDD mengandungi 4,898,431 sampel dengan setiap rekod mempunyai 42 atribut yang mewakili 5 kategori iaitu Normal, DOS, URL, R2L dan Probe. Daripada keseluruhan sampel, 494,021 daripadanya dipilih secara rawak bagi tujuan pembelajaran dan 311,029 sampel dipilih sebagai sampel ujian (Sahu, Sarangi & Jena, 2014). Proses pengkelasan memerlukan masa dan sumber komputer yang banyak (Horng et al., 2011) kerana jumlah rekod yang besar. Kluster merupakan satu kaedah dalam pembelajaran mesin tanpa penyelia yang sesuai bagi mengatasi masalah tersebut. Kaedah kluster beroperasi dengan setiap rekod yang mempunyai persamaan dari segi ciri-cirinya ditempatkan ke dalam kumpulan yang sama. Sentroid bagi setiap kumpulan dijana bagi menyulahi keseluruhan rekod. Terdapat dua pendekatan kluster yang diguna oleh pengkaji terdahulu iaitu pendekatan kluster sejagat oleh Am et al. (2013), Mohammad Khubeb & Shams (2013), Ravale, Marathe & Padiya (2015), Warusia et al. (2013) dan Wathiq et al. (2015). Manakala pendekatan kluster setempat diguna oleh Horng et al. (2011). Kaedah menggabung semua rekod dan melaksana proses kluster dikenali sebagai pendekatan kluster sejagat manakala kaedah mengasing rekod mengikut kategori dan melaksana proses kluster mengikut setiap kategori dikenali sebagai pendekatan kluster setempat.

Objektif kajian ini ialah membanding prestasi pengkelasan perdua (*binary classification*) dan berbilang kelas (*multi-class classification*) di antara kedua-kedua pendekatan kluster dengan teknik umpukan lembut diguna bagi proses pembentukan fitur baharu set data KDD. Kaedah pengkelasan kontemporari iaitu Mesin Sokongan Vektor (SVM) dengan Kernel RBF diguna kerana kebolehannya mengklas pada pelbagai aplikasi dalam bidang pengecaman corak. Prestasi model diukur berdasarkan kepada pengkelasan berbilang kelas iaitu kebolehan menentu kategori serangan yang terlibat dan pengkelasan perdua iaitu kebolehan menentu aktiviti normal dan anomali.

LATAR BELAKANG

SET DATA KDD

Setiap rekod dalam set data KDD mengandungi 42 atribut dan 41 atribut daripadanya (iaitu dari atribut ke-1 hingga ke-41) diguna sebagai input. Atribut ke-42 pula diguna sebagai label. Label berfungsi menyatakan sama ada rekod tersebut adalah normal atau anomali. Anomali mempunyai 21 jenis serangan yang dipecah kepada empat kategori iaitu DOS, U2R, R2L dan Probe. Jenis serangan bagi setiap kategori dipapar seperti dalam Jadual 1, 2 dan 3. Terdapat lima kategori dalam set data latihan iaitu Normal, DOS, U2R, R2L dan Probe. Keterangan tentang setiap kategori adalah seperti berikut:

- **Normal** – aktiviti yang dikategori sebagai Normal dianggap bukan satu serangan atau ancaman kepada keselamatan maklumat. Aktiviti tersebut adalah daripada sumber yang sah.
- **DOS** – merupakan satu serangan yang bersifat halangan perkhidmatan. Pengguna yang sah tidak dapat membuat capaian atau mengguna perkhidmatan.
- **R2L** – merupakan cubaan membuat capaian atau memasuki sistem dengan cara yang tidak sah.
- **U2R** – merupakan serangan yang berjaya membuat capaian ke dalam sistem tetapi dengan keistimewaan yang terhad dan cuba mendapat keistimewaan yang tinggi seperti akaun Super User.
- **Probe** – merupakan satu tindakan tinjauan atau intipan iaitu teknik dengan penyerang cuba mendapat sebanyak mungkin maklumat tentang sistem yang bakal diserang seperti kemudahterancaman (*vulnerability*).

JADUAL 1. Jumlah sampel dalam set data latihan mengikut jenis serangan atau label bagi setiap kategori

Bil	Kategori	Jenis Serangan / Label	Bilangan Jenis	Jumlah
1	Normal	normal	1	97,278
2	DOS	back (2,203), land (21), neptune (107,201), pod (264), smurf (280,790), teardrop (979)	6	391,458
3	U2R	buffer_overflow (30), loadmodule (9), perl (3), rootkit (10)	4	52
4	R2L	ftp_write (8), guess_passwd (53), imap (12), multihop (7), phf (4), spy (2), warezclient (1,020), warezmaster (20)	8	1,126
5	Probe	ipsweep (1,247), nmap (231), portsweep (1,040), satan (1,589)	4	4,107
Jumlah			23	494,021

JADUAL 2. Jumlah sampel dalam set data pengujian mengikut jenis serangan atau label bagi setiap kategori

Bil	Kategori	Jenis Serangan / Label	Bilangan Jenis	Jumlah
1	Normal	normal	1	60,593
2	DOS	back (1,098), land (9), neptune (58,001), pod (87), smurf (164,091), teardrop (12)	6	223,298
3	U2R	buffer_overflow (22), loadmodule (2), perl (2), rootkit (13)	4	39
4	R2L	ftp_write (3), guess_passwd (4,367), imap (1), multihop (18), phf (2), warezmaster (1,602)	6	5,993
5	Probe	ipsweep (306), nmap (84), portsweep (354), satan (1,633)	4	2,377
Jumlah			21	292,300

JADUAL 3. Jumlah sampel dalam set data pengujian mengikut jenis serangan baharu bagi setiap kategori

Bil	Kategori	Jenis Serangan / Label	Bilangan Serangan	Jumlah
1	DOS	apache2 (794), mailbomb (5,000), processtable (759), udpstorm (2)	3	6,555
2	U2R	httptunnel (158), ps (16), sqlattack (2), xterm (13)	4	189
3	R2L	sendmail (17), named (17), snmpgetattack(7,741), snmpguess (2,406), xlock (9), xsnoop (4), worm (2)	8	10,196
4	Probe	mscan (1,053), saint (736)	2	1,789
		Jumlah	17	18,729

KLUSTER PURATA-K

Algoritma Kluster Purata-K diguna bagi mengagih setiap data supaya berada dalam kelompok tersendiri (K) mengikut kriteria jarak terdekat di antara data dan sentroid. Sentroid adalah titik tengah atau nilai purata bagi sekelompok data (K). Algoritma Kluster Purata-K adalah mudah dan pantas berbanding dengan algoritma kluster yang lain. Namun begitu, terdapat cabaran dalam menentu jumlah bilangan kluster (K) yang perlu dicipta (Pham, Domov & Nguyen, 2004). Kaedah Kluster Purata-K Terubah Suai (Wathiq et al., 2015) adalah kaedah alternatif daripada algoritma sedia ada yang mengguna nilai ambang (*threshold*) bagi menentu jumlah K seperti yang ditunjukkan dalam pseudokod berikut:

Algoritma Kluster Purata-K Terubah Suai
<ol style="list-style-type: none"> 1. Mula. 2. Baca rekod dalam set data dan nilai ambang. 3. Permulaan <ol style="list-style-type: none"> 3.1. Tetapkan $K=1$, $c_1 = x_1$ di mana $x \in X$, X adalah set data latihan dan x adalah sampel. 3.2. Setiap sampel $x_i \in X$ dan $i \neq 1$ <ol style="list-style-type: none"> 3.2.1. Jika $\ x_i - c_s\ > t, \forall s = 1, \dots, k, t = \text{ambang}$ maka 3.2.2. $k = k + 1, c_k = x_i$ di mana k adalah jumlah kluster dan c adalah sentroid. 3.3. Setiap sampel $x_i \in X$ <ol style="list-style-type: none"> 3.3.1. Jika $\ x_i - c_s\ < \ x_i - c_j\ , \forall s = 1, \dots, k, \forall j = 1, \dots, k, s \neq j$, maka umpuk $x_i \in C_s$ 3.3.2. Ulang proses 3.3 sehingga sampel terakhir. 3.4. Kira sentroid $c_i = \frac{1}{n} \sum_{j=1}^n x_j, n = \sum C_i, i = 1, \dots, k, x \in C_i$ 3.5. Jika sentroid tidak berubah, hentikan proses. Jika sebaliknya, ulang langkah 3.3. 4. Output sentroid bagi K kluster. 5. Tamat

METODOLOGI

Kajian mengguna pendekatan kuantitatif dengan penglibatan pengiraan aritmetik dan analisis statistik ke atas set data KDD yang menjadi kayu ukur kepada keberkesanan model pengesanan pencerobohan yang dibangun. Perkara utama yang perlu dilakukan adalah dengan melaksana pemprosesan awal ke atas set data KDD. Pemprosesan awal data adalah proses transformasi dari data mentah kepada bentuk yang bersesuaian supaya analisis secara aritmetik dapat dilakukan (Sahu, Sarangi & Jena, 2014). Transformasi data perlu dilakukan kerana set data mentah latihan dan pengujian mengandungi sampel yang banyak, lewah, mengandungi nilai aksara abjad, serta tidak normal. Bagi mengelak berlakunya kecenderungan atau berat sebelah dalam membuat keputusan pada sesuatu model pengkelasan, maka set data latihan perlu bebas dari masalah kelewahan (iaitu terdapat rekod yang sama secara berulang-ulang) (Tavallaee et al., 2009). Kaedah Hasil Tambah Semak (*checksum*) diguna untuk membanding keunikan setiap rekod. Sekiranya terdapat rekod yang mempunyai maklumat atribut yang sama (*identical*), maka nilai Hasil Tambah Semak bagi setiap rekod tersebut juga adalah sama.

Justeru, bagi rekod yang mempunyai nilai Hasil Tambah Semak yang sama, hanya satu sahaja rekod yang dikekalkan manakala rekod selebihnya dihapus bagi memasti setiap rekod yang tinggal adalah unik. Proses ini dilakukan sehingga semua baris rekod digelintar. Setelah melalui proses tapisan kelewahan, jumlah rekod akhir yang diguna dalam set data latihan adalah 145,586 berbanding jumlah rekod asal iaitu 494,021, dengan pengurangan sebanyak 29.5%. Semua atribut yang bernilai aksara abjad seperti atribut *protocol*, *services* dan *flag* ditransformasi kepada bentuk aksara angka. Pada setiap atribut tersebut, nilai atribut yang berbeza perlu ditukar kepada bentuk angka dengan cara menggabungkan kedua-dua set data latihan dan pengujian dan seterusnya julat angka ditentukan untuk mewakili setiap nilai atribut berbeza. Sebagai contoh, setiap aksara abjad bagi atribut *protocol* seperti *icmp*, *tcp* dan *udp* masing-masing ditukar kepada nilai 0, 1 dan 2.

Setelah proses transformasi data dilaksanakan, semua nilai atribut adalah berbentuk angka, $x_i \in \mathbb{R}^n$. Konsep pengkuantitan vektor diguna bagi memampatkan data dan seterusnya diguna bagi menghasilkan model beg-fitur. Kaedah Kluster Purata-K merupakan sebahagian daripada algoritma pengkuantitan vektor (Michael, Yishay & Ng, 2013) yang diguna memecahkan set data yang besar kepada kumpulan yang kecil (K) dengan setiap sampel yang berada dekat dengan nilai purata dikumpulkan bersama. Kajian ini menggunakan kaedah Kluster Purata-K Terubah Suai kerana pemilihan bilangan K dijalankan secara sistematik berdasarkan jarak nilai ambang dan rekod berbanding dengan pemilihan K secara rawak dalam kaedah Kluster Purata-K asal. Pengukuran jarak memainkan peranan penting dalam proses kluster kerana bertindak mencari persamaan di antara rekod. Jarak Euclidean diguna mengikut formula berikut:

$$D(a, b) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Bagi memenuhi objektif kajian, dua set kluster iaitu kluster sejagat dan setempat dihasilkan secara berasingan. Hasil kluster yang terbaik adalah berdasarkan kepada faktor kepadatan dan pengasingan (Kovács et al., 2006). Justeru, pemilihan nilai ambang bermula daripada nilai 50 sehingga 10,000 diguna semasa fasa prosesan kluster dan setiap kluster yang terhasil diukur dengan indeks keesahan kluster iaitu Indeks Dunn (Bezdek & Pal, 1995) berdasarkan formula berikut:

$$Dunn_{indeks} = \min_{i=1, \dots, n_c} \left\{ \min_{j=i+1, \dots, n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1, \dots, n_c} (diam(c_k))} \right) \right\} \quad (2)$$

Kluster yang menghasilkan Indeks Dunn terbesar dipilih untuk membentuk beg-fitur seperti dalam Jadual 4.

JADUAL 4. Kluster dengan Indeks Dunn terbesar dalam pendekatan kluster sejagat dan setempat

Bil	Pendekatan	Ambang	Indeks Dunn	Bilangan Kluster (K)
1.	Sejagat	50	1.0002583471905588	11,256
2.	Setempat (Normal)	50	1.0070379112887158	10,884
3.	Setempat (DOS)	2000	1.0000193962887198	26
4.	Setempat (U2R)	50	15.161484945614609	43
5.	Setempat (R2L)	2000	1.0513817044430447	21
6.	Setempat (Probe)	300	7032.3370991652570	41

Algoritma Kluster Purata-K Terubah Suai yang dilaksanakan menghasilkan satu kumpulan sentroid K kluster daripada pendekatan kluster sejagat dan lima kumpulan sentroid K kluster yang berlainan daripada pendekatan kluster setempat. Setiap sentroid bagi kumpulan K kluster

ini dikenali sebagai beg-fitur berdasarkan konsep beg-perkataan yang diguna pakai dalam teknik pengkelasan abjad dan konsep beg *Visual Codebook* yang diguna pakai dalam teknik pengecaman imej (Azizi, 2010; O'hara & Draper, 2010). Setiap beg-fitur diguna bagi mewakili K kluster dalam pendekatan kluster sejagat dan setiap kategori dalam pendekatan kluster setempat untuk membina perihalan serangan dalam bentuk histogram dengan memeta setiap rekod dalam set data KDD kepada beg-fitur bagi membentuk tatasusunan atribut baharu. Tatasusunan atribut baharu dibentuk kerana jumlah atribut asal set data KDD yang berjumlah 41 atribut adalah besar dan berdimensi tinggi sehingga proses pembelajaran dan pengkelasan menjadi sukar dan memerlukan masa pemprosesan yang lama. Terdapat dua kaedah pemetaan iaitu melalui pendekatan umpukan kasar (Azizi, 2010; Sivic & Zisserman, 2003) dan lembut (Azizi, 2010; Gemert et al., 2010). Kedua-dua pendekatan mengguna jarak Euclidean paling minimum di antara rekod dan sentroid. Namun pendekatan umpukan kasar memilih kedudukan sentroid dalam beg-fitur atau dikenali sebagai “*Winner-Take-All*” berbanding pendekatan umpukan lembut yang mengguna jarak bagi mewakili fitur. Pendekatan umpukan lembut mempunyai kelebihan kerana meskipun rekod yang mempunyai sifat yang sama dipeta kepada sentroid berlainan, namun jarak minimumnya tidak menunjukkan perbezaan yang ketara. Justeru, pendekatan umpukan lembut diguna-dalam kajian ini kerana kualiti fitur yang dihasil adalah baik.

Bagi kaedah kluster setempat, lima beg-fitur dibentuk bagi mewakili lima kategori iaitu Normal, DOS, U2R, R2L dan Probe. Jarak euclidean diukur di antara setiap data dan sentroid dalam setiap beg-fitur. Nilai jarak yang diperolehi untuk setiap beg-fitur selanjutnya disusun mengikut tertib menaik bagi memudahkan pemilihan nilai jarak ke- n paling minimum. Empat set atribut baharu dibentuk dengan setiap satunya terdiri daripada 5, 10, 15 dan 20 atribut bersumberkan kepada nilai jarak ke- n paling minimum dalam setiap beg-fitur. Jadual 5 menunjukkan hasil tatasusunan atribut mengikut bilangan atribut 5, 10, 15 dan 20.

JADUAL 5. Tatasusunan atribut mengikut bilangan atribut bagi pendekatan kluster setempat

Bil Atribut	Kriteria	Tatasusunan (a – Normal, b – DOS, c – U2R, d – R2L, e – Probe)
5	Jarak paling minimum pada setiap kluster	a_1, b_1, c_1, d_1, e_1
10	Jarak 2 terendah pada setiap kluster	$a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, e_1, e_2$
15	Jarak 3 terendah pada setiap kluster	$a_1, \dots, a_3, b_1, \dots, b_3, c_1, \dots, c_3, d_1, \dots, d_3, e_1, \dots, e_3$
20	Jarak 4 terendah pada setiap kluster	$a_1, \dots, a_4, b_1, \dots, b_4, c_1, \dots, c_4, d_1, \dots, d_4, e_1, \dots, e_4$

Bagi kaedah kluster sejagat, hanya satu beg-fitur dibentuk bagi mewakili kesemua kategori iaitu normal, DOS, U2R, R2L dan Probe. Mengguna kaedah yang sama seperti kluster setempat, empat set atribut baharu dibentuk dengan setiap satunya terdiri daripada 5, 10, 15 dan 20 atribut bersumberkan kepada nilai jarak ke- n paling minimum dalam sub-kluster. Jadual 6 menunjukkan hasil tatasusunan atribut mengikut bilangan atribut 5, 10, 15 dan 20 daripada kaedah kluster sejagat.

JADUAL 6. Tatasusunan atribut mengikut bilangan atribut bagi pendekatan kluster sejagat.

Bil Atribut	Kriteria	Tatasusunan (a – Normal, DOS, U2R, R2L, Probe)
5	Jarak 5 terendah pada kluster sejagat	a_1, \dots, a_5
10	Jarak 10 terendah pada kluster sejagat	a_1, \dots, a_{10}
15	Jarak 15 terendah pada kluster sejagat	a_1, \dots, a_{15}
20	Jarak 20 terendah pada kluster sejagat	a_1, \dots, a_{20}

Pada peringkat ini, penormalan dilaksanakan iaitu satu proses transformasi sekelompok fitur kepada julat yang tertentu. Tujuan penormalan adalah bagi mengelak fitur yang mempunyai nilai angka yang besar mendominasi fitur yang mempunyai nilai angka yang kecil (Azizi, 2010; Hsu, Chang & Lin, 2010). Selain daripada itu, penormalan memudahkan proses aritmetik dilaksanakan dalam algoritma SVM (Hsu, Chang & Lin, 2010) dan menghasilkan keputusan yang baik dalam masalah pengkelasan (Azizi, 2010). Penormalan perlu dilaksanakan pada kedua-dua set data latihan dan ujian. Nilai julat angka yang diguna ialah $[-1, +1]$ dengan menggunakan formula berikut:

$$x' = \frac{2(x - \min)}{(\max - \min)} - 1 \quad (3)$$

dengan nilai min adalah nilai fitur paling minimum dan maks adalah nilai fitur terbesar dalam set data latihan.

EKSPERIMEN DAN KEPUTUSAN

Kaedah pengkelasan menggunakan algoritma SVM bersama trik kernel RBF diguna untuk membuat pengkelasan. Sebanyak 52 rekod daripada setiap kategori dalam set data latihan diguna bagi membentuk model pengkelasan. Jumlah tersebut dipilih kerana jumlah rekod bagi kategori R2L yang sedikit iaitu berjumlah 52 rekod sahaja. Sehubungan itu, bagi menyeimbangkan rekod bagi tujuan latihan, maka hanya 52 rekod dipilih daripada setiap kategori. Kernel RBF menggunakan dua paramater iaitu c dan γ untuk mendapat margin hipersatah yang memuaskan. Teknik pengesanan bersilang dengan menggunakan 5, 10, 15 dan 20 lipatan dijalankan untuk mencari nilai parameter c dan γ bagi menghasilkan model pengkelasan yang terbaik semasa fasa latihan dengan menggunakan alat bantu LIBSVM.

Tiga peringkat ujian dijalankan iaitu pada peringkat pertama, hanya 100 rekod pertama daripada sampel ujian diguna bagi mengenal pasti pola dan prestasi awal pengkelas. Pada peringkat kedua pula, keseluruhan rekod dalam sampel ujian sebanyak 311,029 diguna bagi menentu prestasi keseluruhan pengkelas. Peringkat terakhir adalah ujian bagi mengenal pasti serangan baharu yang tidak terdapat dalam set data latihan dengan jumlah sampel sebanyak 18,727.

Bagi menilai kebolehan pengkelasan perduaan, beberapa kriteria digaris oleh para pengkaji seperti Am (2013); Singh, Kumar & Singla (2015) dan al-Twajry dan al-Garny (2012) bagi menguji prestasi pengkelas. Kriteria tersebut ialah kadar Positif Benar (PB) bererti sistem berupaya mengesan dengan betul paket serangan sebagai serangan, Positif Salah (PS) bererti sistem tersilap mengesan dengan betul paket serangan sebagai normal, Negatif Salah (NS) bererti sistem tersilap mengesan dengan betul paket normal sebagai serangan dan Negatif Benar (NB) bererti sistem berjaya meneka paket normal dengan betul. Semua kriteria tersebut digabung menjadi parameter bagi mencari kadar ketepatan (K), pengesanan (P) dan amaran palsu (AP) berdasarkan formula berikut (Chandolikor & Nandavadekar, 2012; Warusia et al., 2013; Wathiq et al., 2015):

$$K = \frac{PB + NB}{PB + PS + NS + NB} \quad (4)$$

$$P = \frac{PB}{PB + NB} \quad (5)$$

$$AP = \frac{PS}{PS + NB} \quad (6)$$

Ketepatan Purata (KP) adalah kaedah kedua yang diguna dalam kajian ini untuk menilai prestasi pengkelasan berbilang kelas. Jumlah ketepatan purata pengkelas memberi-gambaran

secara menyeluruh tentang ketepatan pengkelas secara umum. Formula bagi pengukuran tersebut adalah seperti berikut:

$$\text{Purata Ketepatan (KP)} = \frac{1}{n} \sum_{i=1}^n x_i, \begin{cases} 1 \text{ jika } x_i = C(x_i) \\ 0 \text{ jika } x_i \neq C(x_i) \end{cases} \quad (7)$$

$$\text{Peratus Ketepatan Purata (PKP)} = \text{PK} \times 100 \quad (8)$$

$$\text{PKP} = \frac{\sum_{i=1}^C K_i}{\sum C} \quad (9)$$

KEPUTUSAN UJIAN PERTAMA

Pada peringkat ujian pertama, kedua-dua pendekatan kluster menunjukkan set data ujian dengan bilangan atribut 15 mempunyai nilai KP yang tinggi dengan pendekatan kluster sejagat mencatat nilai KP tertinggi iaitu 0.492 pendekatan kluster setempat mencatat nilai KP tertinggi iaitu 0.652.

KEPUTUSAN UJIAN KEDUA

Pada pengujian peringkat kedua keseluruhan set data pengujian berjumlah 311,029 diguna dengan bilangan atribut 15 bagi pendekatan kluster sejagat dan setempat. Pendekatan kluster sejagat menunjukkan prestasi pengkelasan perdua yang baik iaitu dengan kadar Ketepatan (K) tertinggi **0.935252** berbanding dengan pendekatan kluster setempat iaitu **0.829753**. Namun, pendekatan kluster setempat mengatasi pendekatan kluster sejagat dalam pengkelasan berbilang kelas dengan nilai Ketepatan Purata (KP) masing-masing sebanyak **0.605632** dan **0.408246**. Keputusan ujian peringkat ini ditunjuk dalam Jadual 7.

JADUAL 7. Keputusan prestasi pengkelasan perdua ujian peringkat kedua mengikut pengukuran kadar ketepatan, pengesanan dan amaran palsu bagi pendekatan kluster sejagat dan setempat

Pendekatan Kluster	Lipatan Pengesanan Bersilang	Parameter SVM		Ketepatan (K)	Pengesanan (P)	Amaran Palsu (AP)
		c	γ			
Sejagat (15 atribut)	5	32768.0	0.125	0.935252	0.474931	0.810195
	10	32768.0	0.125	0.935252	0.474931	0.810195
	15	32768.0	0.125	0.935252	0.474931	0.810195
	20	2048.0	2.0	0.930712	0.360553	0.830276
Setempat (15 atribut)	5	2048	0.5	0.829753	0.580609	0.755727
	10	128	8	0.824143	0.568125	0.762861
	15	32768	8	0.795566	0.562175	0.761000
	20	32768	8	0.795566	0.562175	0.761000

JADUAL 8. Keputusan prestasi pengkelasan berbilang kelas ujian peringkat kedua mengikut pengukuran ketepatan purata bagi pendekatan kluster sejagat dan setempat

Pendekatan Kluster	Lipatan Pengesanan Bersilang	Parameter SVM		Ketepatan Purata (KP)
		c	γ	
Sejagat (15 atribut)	5	32768.0	0.125	0.396470
	10	32768.0	0.125	0.396470
	15	32768.0	0.125	0.396470
	20	2048.0	2.0	0.408246
Setempat (15 atribut)	5	2048	0.5	0.506340
	10	128	8	0.605632
	15	32768	8	0.482322
	20	32768	8	0.482322

Perbandingan dengan kajian lepas menunjukkan keputusan yang diperoleh tidak dapat mengatasi prestasi pengelasan perdua dan berbilang kelas. Namun, bilangan sampel yang diguna oleh beberapa pengkaji seperti Chandolika & Nandavadekar (2012) dan Pfahringer (2000) tidak mengguna keseluruhan sampel dalam set data ujian atau tidak diketahui jumlah sebenarnya. Selain daripada itu, Ketepatan Purata bagi serangan kategori R2L yang sememangnya diketahui sukar dikesan menunjukkan keputusan yang tinggi berbanding dengan kajian lain.

JADUAL 9. Perbandingan keputusan pengelasan perdua dengan kajian lain.

Sumber	Ketepatan	Kepersisan	Perolehan Balik	Skor-F
Kajian ini	0.830276	0.959714	0.849142	0.901048
Shah & Trivedi (2015)	0.910000	0.996699	0.900509	0.946150
Chandolika & Nandavadekar (2012)	0.997400	0.998000	0.999000	0.998000

JADUAL 10. Perbandingan keputusan pengelasan berbilang kelas dengan kajian lain

Sumber	Kategori					% Ketepatan Purata (PKP)
	Normal	DOS	U2R	R2L	Probe	
Kajian ini	68.86	47.55	64.91	75.86	45.63	60.56
Canedo, Marono & Betan (2011)	96.63	24.12	90.20	29.98	89.94	66.17
Hornig et al. (2011)	99.30	99.50	19.70	28.80	97.50	68.96
Am et al. (2013)	Tiada	99.93	0.005	6.4	6.56	28.22
Pfahringer (2000)	99.45	97.12	13.16	8.40	83.32	60.29

KEPUTUSAN UJIAN KETIGA

Pengujian peringkat ketiga melibatkan sebanyak 18,727 jenis serangan baharu yang tidak terdapat dalam set data latihan. Hasil ujian menunjukkan pendekatan kluster setempat mengatasi pendekatan kluster sejagat dalam pengelasan berbilang kelas dengan nilai KP tertinggi masing-masing adalah 0.462829 dan 0.25549 seperti dalam Jadual 11.

JADUAL 11. Keputusan prestasi pengelasan berbilang kelas ujian peringkat ketiga bagi pendekatan kluster sejagat dan setempat.

Pendekatan Kluster	Lipatan Pengesanan Bersilang	Parameter SVM		Ketepatan Purata (KP)
		c	γ	
Sejagat (15 atribut)	5	32768.0	0.125	0.25549
	10	32768.0	0.125	0.25549
	15	32768.0	0.125	0.25549
	20	2048.0	2.0	0.24338
Setempat (15 atribut)	5	2048	0.5	0.337111
	10	128	8	0.462829
	15	32768	8	0.276788
	20	32768	8	0.276788

Berbanding dengan kajian lepas, keputusan yang dicatat dalam kajian ini menunjukkan peningkatan secara keseluruhan iaitu dengan penambahan sebanyak 7.55% seperti dalam Jadual 12.

JADUAL 12. Perbandingan keputusan pengkelasan berbilang bagi serangan baharu dengan kajian lain.

Sumber	Normal	DOS	Kategori U2R	R2L	Probe	% Ketepatan Purata (PKP)
Kajian ini	Tiada	0.21	65.08	76.85	42.98	46.28
Horng et al. (2011)	Tiada	85.05	13.23	0.08	95.3	38.73

KESIMPULAN

Kedua-dua pendekatan kluster sejagat dan setempat mempunyai kelebihan tersendiri dalam pengkelasan perduaan dan berbilang kelas. Dalam pengkelasan perduaan, prestasi pendekatan kluster sejagat adalah lebih baik berbanding pendekatan kluster setempat. Namun begitu, pendekatan kluster setempat mengatasi pendekatan kluster sejagat dalam pengkelasan berbilang kelas. Meskipun keputusan yang diperoleh dalam kajian ini tidak dapat mengatasi prestasi kajian lepas, namun ketepatan bagi mengesan serangan R2L dapat ditingkat. Bagi serangan baharu pula, kajian berjaya meningkat kadar ketepatan.

RUJUKAN

- Am Riad, Ibrahim el-Henawy, Ahmed Hassan & Nancy Awadallah. 2013. Visualize Network Anomaly Detection by Using K-Means Clustering Algorithm. *International Journal of Computer Networks & Communications*, 5(5): 195.
- Azizi Abdullah. 2010. Supervised Learning Algorithms for Visual Object Categorization. Tesis Phd, Universiteit Utrecht.
- Bezdek, J. C. & Pal, N. R. 1995. Cluster Validation with Generalized Dunn's Indices. *Proceedings of the Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, Dunedin.
- Bolón-Canedo, V., Sánchez-Marono, N. & Alonso-Betanzos, A. 2011. Feature Selection and Classification in Multiple Class Datasets: An Application to Kdd Cup 99 Dataset. *Expert Systems with Applications*, 38(5):5947-5957.
- Chandolika, N. S. & Nandavadekar, V. D. 2012. Efficient Algorithm for Intrusion Attack Classification by Analyzing Kdd Cup 99. *Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, 20-22 Sept., Indore, India.
- Gemert, J. C. V., Veenman, C. J., Smeulders, A. W. M. & Geusebroek, J. M. 2010. Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271-1283.
- Horng, S.-J., Su, M.-Y., Chen, Y.-H., Kao, T.-W., Chen, R.-J., Lai, J.-L. & Perkasa, C. D. 2011. A Novel Intrusion Detection System Based on Hierarchical Clustering and Support Vector Machines. *Expert Systems with Applications*, 38(1):306-313.
- Hsu, C.-W., Chang, C.-C. & Lin, C.-J. 2010. A Practical Guide to Support Vector Classification. Taiwan: Department of Computer Science, National Taiwan University.
- Michael, K., Yishay, M. & Y., Ng. A. 1998. An Information-Theoretic Analysis of Hard and Soft Assignment Methods for Clustering. Dordrecht: Springer.
- Mohammad Khube S. & Shams N.. 2013. Analysis of Kdd Cup 99 Dataset Using Clustering Based Data Mining. *International Journal of Database Theory and Application*, 6(5):23-34.
- O'hara, S. & Draper, B. A. 2011. Introduction to the Bag of Features Paradigm for Image Classification and Retrieval. Computer Research Repository, vol. arXiv:1101.3354v1
- Pfahring, B. 2000. Winning the Kdd99 Classification Cup: Bagged Boosting. *SIGKDD Explor. Newsl.*, 1(2): 65-66.
- Pham, D. T., Dimov, S. S. & Nguyen, C. D. 2004. Selection of K in K-Means Clustering. *Proceedings of the Institution of Mechanical Engineers Part C: Journal of Mechanical Engineering Science*, 219(1), 103-119.

- Ravale, U., Marathe, N. & Padiya, P. 2015. Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and Rbf Kernel Function. *Procedia Computer Science*, 45:428-435.
- Sahu, S. K., Sarangi, S. & Jena, S. K. 2014. A Detail Analysis on Intrusion Detection Datasets. *IEEE International Advance Computing Conference (IACC)*, 21-22 Feb., Gurgaon, India.
- Shah, B. & Trivedi, B. H. 2015. Reducing Features of Kdd Cup 1999 Dataset for Anomaly Detection Using Back Propagation Neural Network. *Fifth International Conference on Advanced Computing & Communication Technologies*, , 21-22 Feb., Haryana, India.
- Singh, R., Kumar, H. & Singla, R. K. 2015. An Intrusion Detection System Using Network Traffic Profiling and Online Sequential Extreme Learning Machine. *Expert Systems with Applications*, 42(22): 8609-8624.
- Sivic, J. & Zisserman, A. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, 13-16 Oct, Nice, France.
- Tavallae, M., Bagheri, E., Lu, W. & Ghorbani, A. A. 2009. A Detailed Analysis of the Kdd Cup 99 Data Set. *2nd IEEE International Conference on Computational Intelligence for Security and Defense Applications*, 8 Jul, Ottawa, ON, Canada.
- al-Twaijri & al-Garny. 2012. Bayesian Based Intrusion Detection System. *Journal of King Saud University-Computer and Information Sciences*, 24(1):1-6
- Vigna, G., Robertson, W., Kher, V. & Kemmerer, R. A. 2003. A Stateful Intrusion Detection System For World-Wide Web Servers. *19th Annual Computer Security Applications Conference, Proceedings* 8-12 Dec, Las Vegas: NV.
- Warusia Yassin, Nur Izura Udzir, Zaiton Muda & Md. Nasir Sulaiman. 2013. Anomaly-Based Intrusion Detection through K-Means Clustering and Naives Bayes Classification, *Proceedings of the 4th International Conference on Computing and Informatics*, ICOCI No 49. 28 - 30 August, Sarawak.
- Wathiq Laftah al-Yaseen, Zulaiha Ali Othman & Mohd Zakree Ahmad Nazri. 2015. Hybrid Modified K-Means with C4.5 for Intrusion Detection Systems in Multiagent Systems. *The Scientific World Journal*, vol. 2015.

Mohd. Rizal Kadis
 Jabatan Kemajuan Islam Malaysia
 Bahagian Pengurusan Maklumat,
 Aras 5, Blok A,
 Kompleks Islam Putrajaya,
 No 23, Jalan Tunku Abdul Rahman, Presint 3,
 62100 Putrajaya, Wilayah Persekutuan, Malaysia
 rizalkadis@gmail.com

Prof. Madya Dr. Azizi Abdullah
 Fakulti Teknologi dan Sains Maklumat,
 Universiti Kebangsaan Malaysia,
 436000 Bangi, Selangor, Malaysia.
 azizia@ukm.edu.my

Received: 13 February 2017
 Accepted: 24 June 2017