TEKNIK MESIN SOKONGAN VEKTOR UNTUK SISTEM PENGESANAN PENCEROBOHAN BERASASKAN RANGKAIAN

NURSYAZWANI SALAMAN
AZIZI ABDULLAH

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Sistem Pengesanan Pencerobohan (SPP) adalah salah satu sistem yang terkini yang berkebolehan untuk membantu tembok api bagi mengukuhkan ketahanan serta menghalang serangan daripada mencerobohi sistem komputer. Kaedah tradisional seperti tembok api tersebut tidak dapat mengesan serangan baru yang lebih sukar berbanding SPP berasaskan anomali. Anomali ini mampu membandingkan aktiviti normal dan anomali. Bagi menyelesaikan permasalahan mengenai pengelasan dalam sistem ini, terdapat dua kaedah Pembelajaran Mesin (PM) seperti Pembelajaran Mesin Berpenyelia (PMB) dan Pembelajaran Mesin Tanpa Berpenyelia (PMTB). Bagi mencapai maksud tersebut, set data KDD Cup 99 telah dipilih untuk digunakan di dalam ujikaji ini di mana set data ini terdiri daripada lima kategori seperti Normal, DOS, Probe, U2R, dan R2L yang mengandungi sekitar 5 juta rekod. Saiz data yang besar ini memerlukan masa pemprosesan yang amat lama dan sumber yang banyak terutamanya semasa proses latihan. Oleh hal demikian, teknik kluster telah digunakan bagi mengecilkan saiz data iaitu salah satu kaedah PMTP. Cara yang digunakan dalam teknik kluster ini adalah menyatukan rekod yang mempunyai kemiripan yang sama ke dalam satu kumpulan yang dikenali sebagai kluster. Terdapat dua cara yang digunakan antaranya sejagat dan setempat. Dalam kajian ini, cara setempat telah digunakan di mana rekod telah dipisahkan kepada lima kategori berlainan dan proses kluster dijalankan mengikut kategori tersebut. Selepas serangan telah dipisahkan kepada lima kategori, hanya kategori yang berasaskan rangkaian sahaja telah di ambil bagi menjalankan uji kaji ini. Antara serangan yang berasaskan rangkaian adalah seperti Normal, DOS dan Probe. Tujuan utama dalam kajian ini adalah untuk membina model ramalan untuk mengesan serangan yang hadir di dalam sistem komputer. Bagi mencapai objektif tersebut, maka pendekatan umpukan lembut telah digunakan iaitu mencari nilai jarak yang minimum bagi setiap kategori serangan dan Mesin Sokongan Vektor (SVM) untuk menjalankan pengelasan. Hasil keputusan kajian ini mendapati bahawa pengelasan terhadap serangan menggunakan teknik kluster adalah sebanyak 11% manakala pengelasan tanpa menggunakan teknik kluster mendapat keputusan sebanyak 56.4236%.

PENGENALAN

Seiring dengan perubahan masa, isu keselamatan rangkaian semakin meningkat sejajar dengan kemajuan teknologi rangkaian. Pada setiap tahun, dunia teknologi maklumat tidak pernah sunyi daripada kes-kes pencerobohan dan serangan rangkaian ke atas organisasi persendirian mahupun pihak awam.

Oleh hal demikian, Sistem Pengesanan Pencerobohan (SPP) adalah merupakan salah satu sistem pertahanan yang dapat mengesan aktiviti pelanggaran yang terdapat di dalam sistem rangkaian. SPP ini juga dapat menghalang aktiviti-aktiviti yang berbahaya antaranya ialah aktiviti yang menjejaskan keselamatan sistem ataupun cubaan penggodaman. SPP ini dianggap sebagai suatu alat yang dapat digunakan bersama-sama dengan produk keselamatan seperti tembok api. Terdapat dua cara serta kaedah SPP untuk mengesan penyalahgunaan

rangkaian seperti Sistem Pengesanan berasaskan anomali dan Sistem Pengesanan berasaskan tandatangan.

Kaedah pengesanan berasaskan anomali adalah dengan membandingkan aktiviti-aktiviti yang kelakuannya diketahui normal sebelumnya dengan aktiviti yang sedang diperhatikan pada sistem komputer. Manakala kaedah pengesanan berasaskan tandatangan adalah dengan membandingkan aktiviti-aktiviti yang diperhatikan dengan ciri-ciri unik yang terdapat pada sesuatu aktiviti yang terdahulu yang telah dikenalpasti sebagai aktiviti tidak sah.

Kajian ini menfokuskan kepada Sistem Pengesanan Pencerobohan berasaskan Rangkaian (SPPR). SPPR ini mempunyai pendekatan yang berbeza iaitu SPPR ini dapat mengumpulkan maklumat tersendiri daripada rangkaiannya dan tidak datang daripada hos yang berbeza. Sistem rangkaian ini akan memantau trafik yang ada pada rangkaian supaya dapat mencari aktiviti yang mencurigakan atau yang boleh menjadi serangan. Sistem ini juga mampu mengimbas tembok api tempatan atau pelayan rangkaian bagi mengekploitasi potensi atau mengimbas trafik secara langsung untuk melihat keadaan yang berlaku.

PENYATAAN MASALAH

Berdasarkan kepada kajian literasi, terdapat dua teknik yang boleh digunakan antaranya ialah teknik pengesanan secara anomali dan tandatangan. Berdasarkan kepada kedua-dua teknik tersebut, hasil yang diperolehi adalah teknik pengesanan secara tandatangan tidak berkesan bagi mendapatkan serangan-serangan baru atau yang tidak dikenali muncul di dalam sistem komputer bagi setiap kategori serangan. Oleh hal demikian, bagi mendapat senarai serangan yang baru, teknik berasaskan anomali telah dipilih. Selain daripada itu, proses untuk teknik pengesanan tandatangan ini memakan masa yang lama bagi menganalisis dan penampalan lubang-lubang keselamatan diperlukan untuk penyelenggaraan Sistem Pengesanan Pencerobohan ini.

Selain itu, peratusan ketepatan bagi setiap serangan sukar diramal dan dikesan kerana setiap serangan mempunyai ciri-ciri yang tertentu. Berdasarkan kepada penyelidik terdahulu, pelbagai kaedah yang digunakan sama ada Teknik Pembelajaran Mesin Berpenyelia seperti Mesin Sokongan Vektor atau Teknik Pembelajaran Mesin Berpenyelia seperti kluster ataupun

menggunakan kedua-dua teknik tersebut di dalam kajian. Setiap teknik yang digunakan mempunyai pelbagai keputusan sama ada mendapat hasil keputusan yang tinggi mahupun rendah. Oleh hal demikian, bagi mengkaji teknik tersebut, kajian ini telah menggunakan teknik Mesin Sokongan Vektor untuk tujuan pengelasan.

OBJEKTIF KAJIAN

Projek ini bertujuan bagi memperkenalkan Sistem Pengesanan berasaskan Rangkaian kepada semua masyarakat bahawasanya terdapat serangan-serangan yang berbahaya kepada pengguna sama ada serangan yang boleh dikenal pasti ataupun serangan tidak dapat dikesan oleh sistem komputer. Secara umumnya, objektif kajian ini adalah mengkaji atribut-atribut yang terdapat di dalam set data KDD Cup 99 yang terdiri daripada 41 atribut.

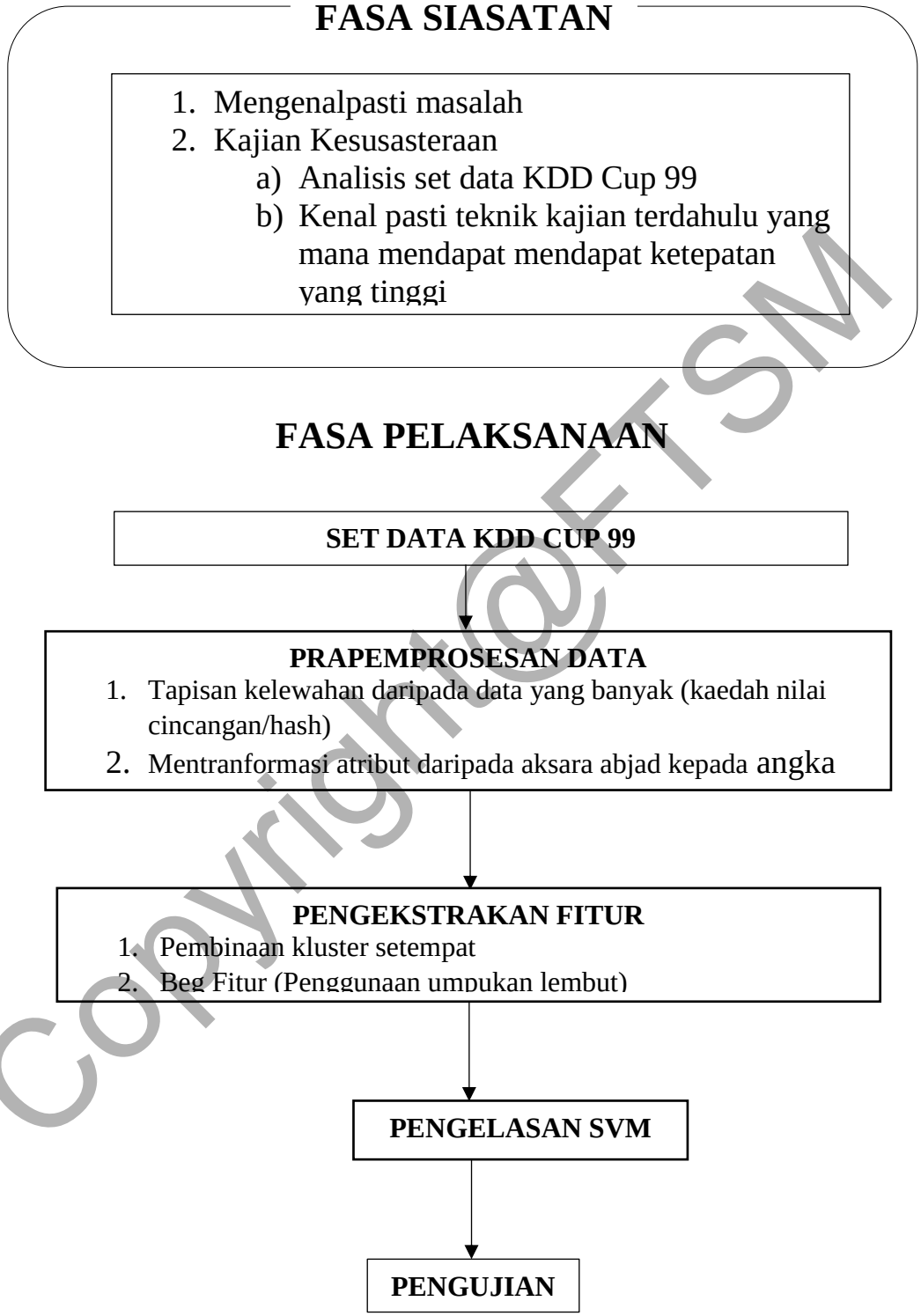
Selain itu, objektif lain pula adalah membangunkan model ramalan untuk Sitem Pengesanan Pencerobohan berasaskan Rangkaian sahaja. Tujuannya adalah dapat memfokuskan kepada serangan untuk rangkaian supaya ketepatan peratusan akan mendapat keputusan yang lebih tepat dan tinggi.

Objektif yang terakhir adalah supaya dapat menguji keberkesanan model ramalan berdasarkan set data KDD Cup 99. Bagi melaksanakan pengujian ini, data latihan dan data latihan akan diuji dan dinilai berdasarkan perbandingan kedua-dua data tersebut. Data latihan di ambil berdasarkan kepada 10% set data manakala bagi tujuan pengujian pula akan mengambil peratusan data yang selebihnya untuk di ujikaji.

METOD KAJIAN

Penggunaan modal pembangunan yang sesuai amat penting bagi memastikan perjalanan projek berjalan dengan lancar serta dapat menjamin hasil kerja yang berkualiti. Model Sistem Pengesanan Pencerobohan berasaskan rangkaian melibatkan dua fasa antaranya iala fasa siasatan dan fasa perlaksanaan. Fasa siasatan terdiri daripada pengenalpastian masalah dan kajian kesusateraan manakala fasa perlaksanaan terdiri daripada analisis, pra pemprosesan data, pengekstrakan fitur, pengkelasan dan pengujian. Rajah 1 menunjukkan model pelaksaan

yang digunakan untuk menjalankan ujikaji terhadap Sistem Pengesanan Pencerobohan berasaskan rangkaian.



Rajah 1 : Model Pelaksanaan Sistem Pengesanan Pencerobohan berasaskan rangkaian

Analisis Data

Langkah pertama adalah fasa siasatan di mana permasalahan berlaku perlu dikenalpasti dahulu sebelum meneruskan fasa pelaksaaan. Pemilihan set data KDD Cup 99 (Darpa 1999) telah dipilih sebagai bahan uji kaji dalam kajian ini di mana analisis terhadap set data diperlukan bagi mengetahui ciri-ciri fitur yang terkandung di dalam set data KDD Cup 99. Analisis ini diperlukan bagi menentukan pemilihan set data bagi tujuan latihan dan ujian semasa membuat pengelasan. Bedasarkan kepada kajian terdahulu, 10% set data KDD Cup 99 telah dipilih bagi menjalankan set data latihan manakala data yang selebihnya digunakan sebagai set data ujian.

Pra Pemprosesan Data

Pada tahap ini, setelah mengenalpasti semua atribut yang terdapat di dalam set data KDD Cup 99 yang terdiri daripada 5 juta data, maka langkah pertama pra pemprosesan data adalah menggunakan kaedah tapisan kelewahan (filtering redundancies). Kaedah ini digunakan untuk menyaring atau menapis data yang berlebihan dan data yang bertindan di dalam set data tersebut. Bagi menjalankan tapisan kelewahan ini, penggunaan java LinkedHashSet telah digunakan dimana set data baru akan dihasilkan yang terdiri daripada rekod yang unik sahaja dipilih. Setiap baris rekod diberikan nilai cincangan yang tertentu dan ianya akan dibandingkan dengan baris rekod yang seterusnya. Jika terdapat baris rekod yang unik, maka ianya akan dimasukkkan ke dalam senarai data set yang baru. Proses ini dilakukan sehingga semua baris rekod telah diproses sehingga habis.

Hasil daripada tapisan kelewahan tersebut, data menjadi semakin berkurang, maka langkah seterusnya adalah dengan mentransformasikan nilai atribut yang berbentuk aksara abjad kepada nilai aksara angka iaitu digunakan semasa membuat Kluster Purata-K dan semasa melakukan pengelasan menggunakan SVM. Di dalam set data tersebut, didapati bahawa hanya atribut *protocol, services dan flag* sahaja mengandungi nilai atribut berbeza. Julat angka akan ditentukan terlebih dahulu bagi mewakili setiap nilai atribut yang berbeza

tersebut. Mentransformasikan atribut ini amat penting kerana format data untuk algoritma pengesanan seperti SVM memerlukan format yang menggunakan nilai aksara angka sahaja.

Pengekstrakan Fitur

Setelah hasil daripada pra pemprosesan data dilakukan, maka pengekstrakan fitur dilakukan. Pemilihan fitur dilaksanakan bagi menjana fitur yang optimum. Tujuan pengekstrakan fitur ini adalah bagi memilih beberapa atribut untuk digunakan di dalam kluster Purata-K bagi mencari nilai sentroid dan digunakn di dalam algoritma pengesanan Atribut yang tidak digunakan akan dibuang Atribut yang dipilih ini adalah berdasarkan kepada kajian terdahulu. Berkurangnya attribute tersebut maka proses akan berjalan dengan lebih mudah

Seterusnya, pengkuantitian vektor turut berlaku bagi memampatkan data dan seterusnya digunakan untuk menghasilkan model beg fitur. Contoh pengkuantitian vektor seperti kaedah Kluster Purata-K iaitu memecahkan set data yang besar kepada beberapa kumpulan kecil di mana setiap kumpulan tersebut yang mempunyai sampel yang berada dekat dengan nilai purata akan dikumpulkan bersama yang juga dikenali sebagai sentroid.

Selain daripada itu, beg fitur diperlukan bagi mendapatkan rajah histogram untuk menggambarkan taburan corak bagi setiap serangan. Beg fitur ini dapat memetakan setiap fitur kepada sentroid yang diperolehi semasa pengkuantitian vektor. Terdapat 2 kaedah pemetaan antanranya umpukan lembut dan umpukan kasar. Dalam kajian ini telah menggunkan umpukan lembut bagi memilih jarak minimum. Selain itu, bagi mencari jarak minimum, jarak Euclidean telah dipilih.

Pengkelasan

Pada fasa ini pula, rekod yang terdapat di dalam set data ujian akan dikelaskan kepada rekod yang berada di dalam set data latihan. Pengelasan adalah salah satu kaedah Pembelajaran Mesin Berpenyelia (PMB) dan kaedah yang akan digunakan di dalam kajian ini adalah berdasarkan kepada teknik Sokongan Vektor Mesin (SVM)

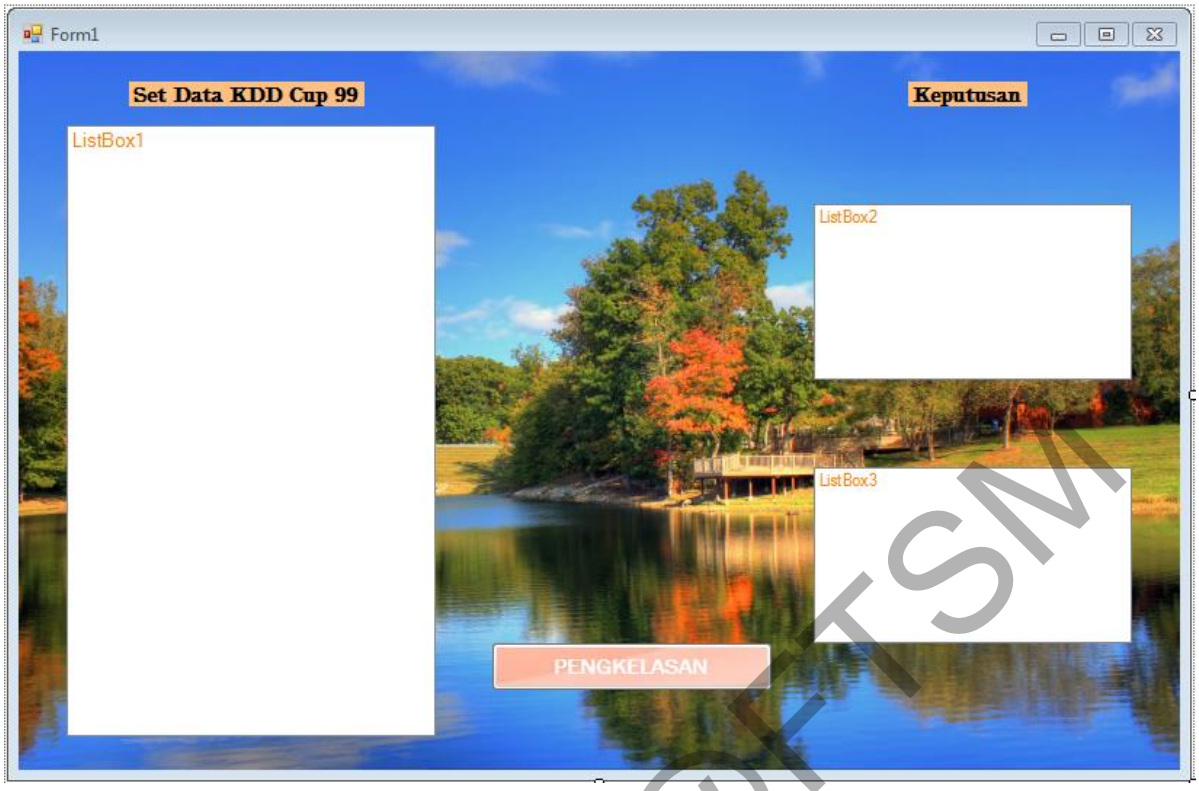
Pengujian

Pada fasa ini, data latihan dan data ujian akan diuji supaya dapat menentukan sama ada objektif pada Bab 1 dapat tercapai atau pun tidak. Selain itu, penilaian akhir ini dapat menentukan peratusan ketepatan penggunaan teknik yang mana dapat hasilkan peratusan yang tinggi bagi setiap jenis serangan.

HASIL KAJIAN

Pada bahagian ini menerangkan mengenai proses pembangunan Sistem Pengesanan Pencerobohan berasaskan Rangkaian. Berdasarkan hasil kajian yang diperolehi, fasa reka bentuk amat penting dalam pembangunan projek. Dalam projek ini tidak terdapat banyak antara muka yang dapat dilihat oleh kerana hanya model ramalan sahaja yang terdapat di dalam kajian ini.

Rajah 2 di bawah menunjukkan bentuk antara muka bagi model ramalan untuk Sistem Pengesanan Pencerobohan. Di bahagian kiri merupakan input data ujian yang akan di masukkan kedalam bahagian set data KDD Cup 99. Apabila salah satu daripada data ujian dipilih maka selepas itu butang pengkelasan akan ditekan dan keputusan akan dipaparkan di sebelah kanan bahagian atas sama ada set data tersebut dalam kategori serangan Normal atau DoS ataupun Probe berdasarkan keputusan yang didapati daripada algoritma pengkelasan iaitu SVM. Jika kategori itu adalah salah maka keputusan yang betul akan dipaparkan di sebelah kanan juga tetapi di bahagian bawah.



Rajah 1: Reka bentuk antara muka

Penggunaan set data KDD cup 99 telah dikaji dan di analisis bagi mencari atribut yang sesuai utuk rangkaian sahaja. Maka terdapat 13 atribut sahaja digunakan berbanding 41 atribut bagi atribut penuh yang terdapat di dalam set data. Selepas mengasingkan atribut tersebut, atribut yang selebihnya akan atau yang tidak digunakan dibuang. Selepas mendapatkan data yang mempunyai 13 atribut untuk setiap data latihan dan ujian, maka menjalankan kaedah Kluster Purata-K iaitu mencari sentroid atau titik tengah bagi setiap serangan. Selain itu, ianya juga dapat mengkelaskan serangan tersebut kepada 50 kluster untuk setiap serangan. Setelah mendapatkan sentroid untuk titik tengah dan kluster maka proses seterusnya ialah mencari nilai jarak yang minimum menggunakan jarak Euclidean.

Seterusnya, selepas mendapatkan jarak minimum tersebut maka penambahan nilai pada setiap serangan untuk mewakili angka seperti 1 untuk Normal, 2 untuk Dos dan 3 untuk Probe. Apabila telah mendapatkan hasil maka ketiga-tiga fail tersebut akan disatukan menjadi satu fail. Langkah semua di atas akan di ulang sama untuk set data ujian. Set data latihan dan set data ujian diuji di dalam algoritma pengesanan iaitu SVM dan hasilnya seperti pada Rajah 2 dan Rajah 3.

train 1000 test 10000.txt

```
13 attribut train1000 test 10000 (cen 50 n dis all centroid )

Accuracy train1000 test10000

C:\libsvm-3.22\tools>python easy.py all_combine(train1000).svm ALL_combine(test1
0000).svm
Scaling training data...
Cross validation...
Best c=32768.0, g=8.0 CV rate=75.817
Training...
Output model: all_combine(train1000).svm.model
Scaling testing data...
Testing...
Accuracy = 11.8244% (2682/22682) (classification)
Output prediction: ALL_combine(test10000).svm.predict
```

Rajah 2: Hasil daripada membuat kluster sebanyak 50 kluster

Pada rajah 1 adalah hasil keputusan selepas membuat kluser yang terdiri daripda 50 kluster. Berdasarkan pada rajah di atas didapati hanya 11.8244% ketepatan sahaja diperolehi. Keputusan ini rendah adalah disebabkan data latihan mengambil data sebanyak 1000 untuk Normal daripada 87832, 1000 untuk Dos daripada 54572 dan mengambil semua 2131 untuk Probe. Bagi data ujian pula mengambil sebanyak 10000 untuk Normal daripada 47913, 10000 untuk DoS daripada 23568 dan mengambil semua 2682 untuk Probe. Bagi mendapat keputusan yang lebih tinggi dan tepat maka diperlukan penggunaan semua data berbanding menggunakan sedikit data.

train 10000 test 10000.txt

```
13 attribut(tanpa cen n dis)

Accuracy train10000 test10000

C:\libsvm-3.22\tools>python easy.py ALL_train.svm ALL_test.svm
Scaling training data...
Cross validation...
Best c=2048.0, g=2.0 CV rate=99.7152
Training...
Output model: ALL_train.svm.model
Scaling testing data...
Testing...
Accuracy = 56.4236% (12798/22682) (classification)
Output prediction: ALL_test.svm.predict
```

Rajah 3: Hasil daripada tanpa melakukan kluster

Seterusnya, pada rajah 2 pula adalah keputusan kepada data yang tanpa menggunakan kluster. Hasil daripada penggunaan algoritma SVM ini mendapat keputusan sebanyak 56.4236% untuk tanpa kluster. Keputusan bagi algoritma pengesanan tanpa melakukan kluster mendapati lebih tinggi berbanding menggunakan kluster. Bagi data latihan sebanyak 10000 untuk Normal daripada 87832, 10000 untuk DoS daripada 54572 dan menggunakan semua data Probe iaitu sebanyak 2131. Selain daripada itu, bagi data ujian pula telah diambil data sebanyak 10000 untuk Normal daripada 47913, 10000 untuk DoS daripada 23568 dan mengambil semua data Probe iaitu sebanyak 2682.

Kesimpulannya, bagi mendapatkan hasil keputusan yang tinggi, pertama ialah menggunakan semua data iaitu yang terdapat di dalam set data untuk Normal, DoS dan Probe untuk kedua-dua data set iaitu data latihan dan ujian. Selain daripada itu, penggunaan kluster yang berulang-ulang seperti 50, 100, 150 dan seterusnya sehingga mendapatkan hasil keputusan yang tinggi. Akhir sekali, penggunaan komputer yang berkuasa tinggi diperlukan bagi menjalankan kajian ini.

KESIMPULAN

Sistem Pengesanan Pencerobohan ini adalah satu cara bagi mengesan serangan dan dapat membantu tembok api untuk menghalang serangan tersebut memasuki sistem komputer pengguna. Selain daripada itu, Pencerobohan merupakan suatu usaha yang disengajakan sama ada berjaya diakses atau tidak berjaya diakses serta disalahgunakan data yang privasi atau sensitif dalam sistem yang sepenuhnya dikawal oleh komputer atau rangkaian. Tidak dinafikan bahawasanya tiada rangkaian yang seratus peratus selamat daripada kesan serangan atau pencerobohan daripada pengguna hasad. Oleh sebab itu, adalah pentingnya mengesan pencerobohan dengan menghadkan kesan terhadap rangkaian sebanyak mungkin.

Secara keseluruhannya, kajian ini dapat memberikan penyelesaian kepada masalah terhadap sistem komputer dan dapat mengesan serangan yang berasaskan rangkaian seperti Normal, DoS dan Probe. Penggunaan SPP amat berguna kepada semua pengguna yang menggunakan sistem komputer supaya dapat menyelamat komputer pengguna daripada serangan yang berbahaya

RUJUKAN

Balu, R. 2015. "Design and Development of Automatic Appendicitis Detection System Using Sonographic Image Mining." *Shodhganga : A Reservoir of Indian Theses @ INFLIBNET*, 167. http://shodhganga.inflibnet.ac.in/handle/10603/33597.

"Computer Security." n.d. http://www.contrib.andrew.cmu.edu/~aishah/Sec.html.

"How to Choose Machine Learning Algorithms | Microsoft Docs." n.d. https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice.

"IDS: Signature versus Anomaly Detection." n.d. http://searchsecurity.techtarget.com/tip/IDS-Signature-versus-anomaly-detection.

"KDD-CUP-99 Task Description." n.d. http://kdd.ics.uci.edu/databases/kddcup99/task.html.

"Mind: How to Build a Neural Network (Part One)." 2017. Accessed March 25. https://stevenmiller888.github.io/mind-how-to-build-a-neural-network/.

S, Prof Dighe Mohit, Kharde Gayatri B, Mahadik Vrushali G, Gade Archana L, and Bondre Namrata R. 2015. "Classification and Invention of Intrusion in KEYWORDS :," 1102–8.

Scherer, Peter, Martin Vicher, and Jan Martinovič. 2011. "Using SVM and Clustering Algorithms Using SVM and Clustering Algorithms in IDS Systems in IDS Systems," 108–19.

Siddiqui, Mohammad Khubeb, and Shams Naahid. 2013. "Analysis of KDD CUP 99 Dataset Using Clustering Based Data Mining." *International Journal of Database Theory and Application* 6 (5): 23–34. doi:10.14257/ijdta.2013.6.5.03.

"Signature-Based or Anomaly-Based Intrusion Detection: The Practice and Pitfalls." n.d. https://www.scmagazine.com/signature-based-or-anomaly-based-intrusion-detection-the-practice-and-pitfalls/article/548733/.

"UMPUKAN LEMBUT KLUSTER SEJAGAT DAN SETEMPAT UNTUK SISTEM PENGESANAN PENCEROBOHAN : SATU KAJIAN PERBANDINGAN." n.d.

"Weka 3 - Data Mining with Open Source Machine Learning Software in Java." n.d. http://www.cs.waikato.ac.nz/ml/weka/.

"What Is Network Traffic? - Definition from Techopedia." n.d. https://www.techopedia.com/definition/29917/network-traffic.

"Why Artificial Neural Networks (ANN) Technology Offers a Promising Future in IDS/IPS." 2017. Accessed March 24. http://resources.infosecinstitute.com/why-artificial-neural-networks-ann-technology-offers-a-promising-future-in-idsips/#gref.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14 (1): 1–37. doi:10.1007/s10115-007-0114-2.