使用改良型基因演算法於網路入侵偵測系統之特徵選取

蘇民揚[1]，張凱棊[2]，魏華甫[3]，林俊淵[4]，甘懷誠[5]
[1,3,4,5]銘傳大學資訊工程學系
[1,2]銘傳大學資訊工程研究所
minysu@mcu.edu.tw

摘要

異常行為偵測型(abnormal detection)的網路入侵偵測系統，其成功與否之重要關鍵在於所選取以供判斷的特徵是否恰當。另一方面，在強調即時處理的入侵偵測系統上，時間因素也至為重要；愈多的特徵也意謂處理時間愈長。如何選擇適量且有效的特徵實為網路入侵偵測系統設計中最重要的一環。本論文使用基因演算法結合 KNN (K-Nearest Neighbor)做為特徵選取的工具，並以大量阻斷服務攻擊(DoS attacks)做實測。實驗顯示，使用少數 12 個特徵就可以達到 95.17%的偵測率；足以證明我們方法所選取之特徵的重要性。

關鍵詞：網路入侵偵測、網路安全、阻斷服務攻擊(DoS)、基因演算法、特徵選取、KNN

Abstract

The most significant factor for the abnormal-based NIDS is how to select important features for judgment by the detection engine. On the other hand, since most NIDS emphasize their real-time capabilities, time expense for judgment is very important. That means the number of features can not be large because more features selected, more time for processing is needed. The selection of features is a key point for the success of an NIDS. The paper presents a genetic algorithm combined with KNN for feature selection. We also applied a lot of DoS attacks to evaluate the performance of our algorithm. According to experiment result, using only 12 selected features we can get 95.17% detection rate. This shows that the features selected by our algorithm are really significant to the DoS attacks.

Keywords：Network Intrusion Detection System (NIDS)，Network Security，DoS Attacks，Genetic Algorithm，Feature Selection，KNN

1. 前言

在這廣大無疆界的網際網路裡，有人立志於建設也有人矢志破壞。 隨著資訊科技日新月異，帶動著電子商務的興起，網路安全也逐漸受到人們的重視。當人們都在享受著網路所提供便利以及功能時，網路攻擊也慢慢成為電子商務的隱憂；於是網路安全的重要性與日俱增，網路入侵偵測系統(Network Intrusion Detection System, NIDS)也因應而生。然而在強調可以即時偵測的 NIDS 裡，如何選取有效而又少量的特徵以降低處理時間同時提高偵測率，將關乎 NIDS 之成敗，也是本論文所要解決的問題。由於阻斷服務攻擊(DoS/DDoS)製作技術門檻最低，以致網路上隨處可以取得，也是現今攻擊事件的主流之一，故本論文所提之方法也將以 DoS/DDoS 攻擊作為實驗的標的。

就入侵偵測引擎主要有兩大類：分別為誤用偵測(Misuse Detection)與異常行為偵測(Anomaly Detection)。誤用偵測在偵測是否發生攻擊事件，是藉由系統管理者去定義出會造成電腦系統毀損或者是危害網路環境的型樣(Patterns)，然後藉由型樣比對(Pattern Matching)的方式去偵測出可能對系統或網路造成危害的入侵行為及攻擊的方式。異常行為偵測是利用統計上的特性，將使用者在電腦上的操作習慣或者是將整個內部網路的資料流動趨勢作一個統計並記錄起來，作為日後偵測電腦系統或者是是否被攻擊的基準。其特色在於大多分為訓練與測試兩階段，藉由訓練階段餵食大量的攻擊程式與樣本，使得異常偵測系統具有偵測未知攻擊的能力。

一般異常行為偵測型網路入侵偵測系統由兩大元件構成：偵測引擎與特徵選取。表一所示為部分常見的入侵偵測引擎與特徵選取的技術。在偵測引擎上，我們另有文獻討論[5, 12]。本篇論文主要是針對特徵選取的部份，提出一些改良。我們使用基因演算法結合 KNN (K-Nearest Neighbor)做為特徵選取的工具，演進過程，賦予每個特徵不同的權重，用以區別出每個特徵的重要性。藉由這樣的方

式，演進結束時可以找出最重要的特徵，再將權重較小的特徵摒棄，以縮減特徵的數目時同時又保有一定的偵測率。

本文於第二節介紹基因演算法與 KNN 演算法，第三節提出研究架構與實驗方法，第四節說明實驗數據與分析，結論與未來展望則放在第五節。

表一 入侵偵測系統兩大因素表

入侵偵測系統兩大因素	
偵測引擎	模糊關聯法則、情節法則、有限狀態機...等
特徵選取	基因演算法、向前特徵選取、排列個別特徵效能、線性識別分析準則特徵選取...等

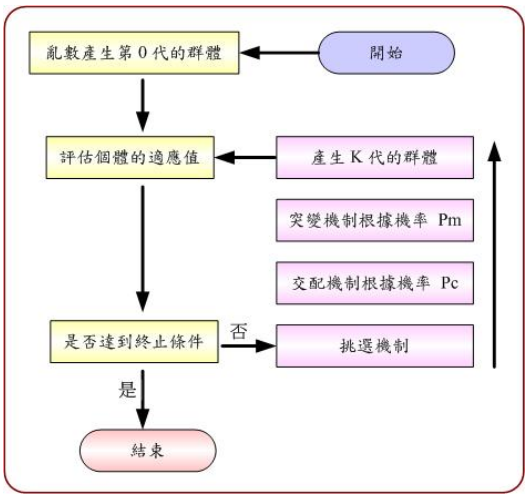


圖 1 基因演算法流程圖

2. 文獻探討

舊有的特徵選取方式，大多採用基因演算法(Genetic Algorithm)，藉由基因演算法的繁衍、突變的特性找到最好的特徵，然而基因演算法中的適應性函數(Fitness Function)，往往擁有著決定性的影響，於是我們藉由 KNN 演算法的分類特性，用以改良舊有的適應性函數。

2.1 基因演算法

Holland 認為，生物的演化主要發生在染色體的基因中，然而每種生物其特徵主要源於該生物母代的基因序列，演化指的是每一代基因所發生的變化情形。所謂適者生存是指這一代的基因排列優於母代的基因排列，而產生比上一代更能適應環境生存的世代(Holland, 1975)。

GA 是強調基因型的轉變，將欲求解問題的參數經過編碼成為基因格式，利用遺傳運算進行演化來找到問題的最佳解。這些遺傳運算是模擬自然界的演化程序，包括有複製(reproduction)、交配(crossover)與突變(mutation)等等。

我們由圖 1 得知，第 0 代的群組是由亂數產生於是將第 0 代帶入適應性函數，並計算出適應值，若不符合終止條件，並開始挑選第 0 代優秀的染色體，進行突變及交配。然而突變與交配與否，視其機率而定(是否大於機率 Pm、Pc)，如此反覆便可挑選出最好的染色體。其中，在交配這程序還細分為單點、兩點、多點交配…等。我們以兩點交配為例，如圖 2 所示。

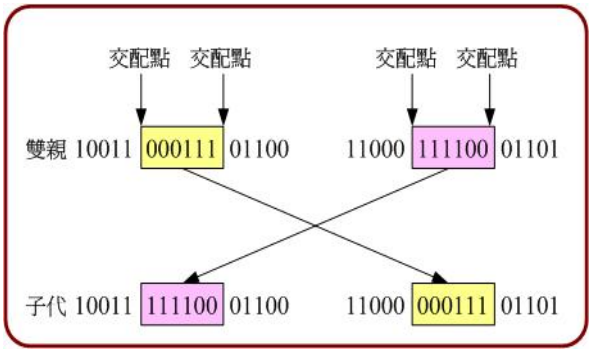


圖 2 基因交配示意圖

圖 2 中，四個交配點分別位於『000111』前後與『111100』前後，當雙親(母代)染色體決定交配點時便會與子代染色體部分基因交換，以達到交配的動作，除了保留的基因外剩餘的基因都不變，進而達到交配的動作。除此之外還有突變。突變是指在染色體上，決定一個、兩個或多個基因，利用亂數的機制進行突變，進而達到突變的效果[8, 9]。

藉由上述方式反覆行進，便可透過交配、突變…等機制，找出最佳的特徵，由上述流程得知，適應性函數擁有著決定性的因素，因為有效的適應性函數可以降低基因繁衍其子代的數目，進而提升系統效能。於是本論文提出利用 KNN 與賦予特徵權重的機制作為我們優化後適應性函數。

2.2 KNN (K-Nearest Neighbor)

K-Nearest Neighbor [1, 2]，屬於案例學習(Instance-based learning)中的一種方法，KNN 假設訓練案例 (training samples)中的每個例子 (instance)皆可由 n 維空間之一個空間座標點來描述。換句話

說，所有的訓練案例都可以儲存在這個 n 維的樣本空間(pattern space)中。當我們給予一個未分類的案例『y』時，KNN 分類器找出樣本空間中與『y』前 k 個最接近（k-closest）的訓練案例出來，這 k 個訓練案例就是『y』的 『k-Nearest Neighbors』然而我們判斷是否為『closest』是根據歐幾里得距離定理。由圖 3 得知，黃色網底的方形為『尚未分類的文件』，而裡面共有兩個類別，第一類:紅色直紋、第二類:藍色橫紋。所圈起來的灰色區塊，表示已選取了 K=3 個與測試文件最近的點，從中可以看出，測試文件應被分為第一類。(因為紅色直紋的數目較多)[10]。

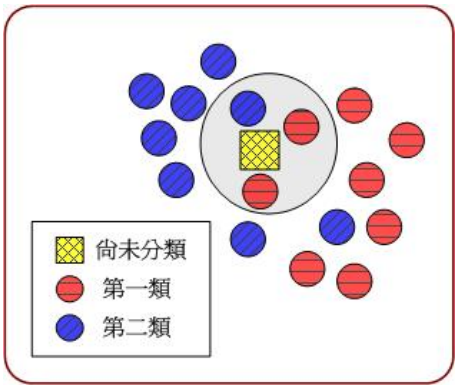


圖 3　KNN 示意圖; k=3

假設樣本空間有兩個點，$X=(X_1,X_2,X_3...X_n)$、$Y=(Y_1,Y_2,Y_3...Y_n)$ 其中的 X_1、X_2、$X_3...X_n$, Y_1、Y_2、$Y_3...Y_n$ 均為 X 與 Y 的屬性值則 X、Y 的歐幾里得距離 *dist(X, Y)*的表示如下：

$$dist(X,Y)=\sqrt{(X_1-Y_1)^2+(X_2-Y_2)^2...+(X_n-Y_n)^2}$$

在得到 k 個最相近的結果下，統計各個類別在 k 中所佔的個數，擁有最多個數的類別就表示該測試文件屬於此類別[3]。

3.　本文所提之方法與實驗架構

本實驗，以基因演算法為主軸逐步進行，在編碼的部分我們採用以實數編碼的方式進行，在初始化階段，便會隨機產生第 0 代的十條染色體，每個染色體代表一種可能的權重向量，表示為 W = [w_1, w_2, …, w_{26}]。染色體的適應性函數值計算如下：

假設空間有兩個案例 $X=(x_1,x_2,...,x_n)$ 與 $Y=(y_1,y_2,...,y_n)$，本文中它們的距離算法如下：

$$dist(X,Y)=\sqrt{w_1^2(x_1-y_1)^2+w_2^2(x_2-y_2)^2...+w_n^2(x_n-y_n)^2}$$

，我們選擇利用 w_i^2 來加權特徵 i 的重要性。接著透過 KNN，就可以利用訓練資料集中已標示類別的案例來計算每條染色體的適應函數值(fitness value)，我們的適應函數定義如下：

$$\text{Fitnesss}=\frac{Total-FP-2FN}{Total}$$

，其中 Total 為訓練案例的總數，FP 為將正常誤判為攻擊(假警報)的個數，FN 為將攻擊誤判為正常(漏報)的個數，其中 2FN 主要是為了加重懲罰漏報的情形，因 FN 的錯誤比 FP 的錯誤還嚴重。圖四是程式第 0 代所呈現的其中 3 條染色體, i.e.3 個權重向量,與其適應函數值。每代我們挑選最好與次好的染色體，進行交配、突變，進而產生下一代。

圖 4　第 0 代染色體(以三條為例)

所有實驗的案例(instances)都是統計單位時間(本文設兩秒)網路流量中這 26 個特徵的封包數做為其特徵值。接著我們對案例中的特徵值做正規化如下，如此可以避免單一特徵值過大所造成的影響。經過正規化處理後，所有案例中的特徵值都介於 0 至 1 之間。

$$\text{Normalize } f_{i,j}=\frac{f_{i,j}-\min(F_j)}{\max(F_j)-\min(F_j)}$$

，其中 $f_{i,j}$ 是指第 i 個案例中的特徵 j, F_j 是所有案例

特徵 j 的集合，而 $\min(F_j)$和 $\max(F_j)$分別代表這個集合中的最小值與最大值。

在交配產生子代的過程中，我們將產生後的子代與母代做比較，倘若子代其適應值沒有大於母代中最差的染色體，我們便將子代丟棄，反之便將其保留。反覆上述動作，我們已經擁有訓練好的權重，這些權重分別代表特徵的重要與否。

在突變的階段，我們隨機選取一條染色體，並且隨機突變任一個基因，藉由這種方式，可以增加基因程式的複雜度，進而產生最理想的權重。

本實驗，主要針對阻斷服務攻擊(DDOS/DOS)，分為訓練階段和測試階段這兩大部分，其中訓練階段用以統計特徵數目，我們的特徵遍及 TCP/IP 模型中的，網路層、傳輸層六個通訊協定包含(網路層：IP、IGMP、ICMP、ARP，傳輸層：TCP、UDP)，特徵如表二與表三所示。這些特徵，主要是為了防範 DDOS/DOS 攻擊所準備的，其中有不少特徵，是需要同時檢查多個欄位，如表二中的『S.IP + ACK Flag + ACK number』與『Port (20) 和 length(>1400)』。『S.IP + ACK Flag + ACK number』主要用於檢查，在相同的 IP 位置下，標記 ACK Flag=1，但標頭內沒有 ACK number 的個數。由於大多數的阻斷服務攻擊程式，只有標記 ACK Flag=1 或只出現 ACK number，藉由這種方式可以統計出這些異常封包的數量。『Port (20)和 length(>1400)』，這部分用來區分出正常的 FTP 下載與 DDOS/DOS 攻擊，在我們觀察眾多的攻擊程式之中，我們發現其實一般 DDOS/DOS 攻擊的封包數量很多，可是負載不大，然而正常的 FTP 下載，傳送資料是使用 TCP Port 20 同時會填滿整個負載。

表二 IP/ICMP/IGMP/ARP 所選取的特徵表

IP	Source Address
IP	Destination Address
IP	Header length
IP	Total length>1400 || <40 &&TTL = = 64
ICMP	Length
ICMP	Type
IGMP	Length
IGMP	Type
ARP	Length
ARP	S.IP + ARP count

表三 TCP/IP 所選取的特徵表

TCP	S.IP + S.port number
TCP	S.IP + D.port number
TCP	S.IP + SYN count
TCP	S.IP + URG Flag + URG data
TCP	S.IP + ACK Flag + ACK number
TCP	Length
TCP	Port (20)&& length(>1400)
TCP	S.port number
TCP	D.port number
TCP	SYN number
TCP	URG Flag + URG data
TCP	ACK Flag + ACK number
UDP	S.port number
UDP	D.port number
UDP	Length
	封包總數

4. 實驗數據與分析

這次實驗我們使用的作業系統為 Windows XP Professional 、程式語言為 C++，開發工具為 Microsoft Visual Studio 2005、攻擊程式的部分：我們準備了一些常見的 DOS/DDOS 攻擊程式，總數逾七十隻攻擊程式。本節所有實驗使用之攻擊程式及樣本詳列於 http://163.25.149.115/TANET2007。

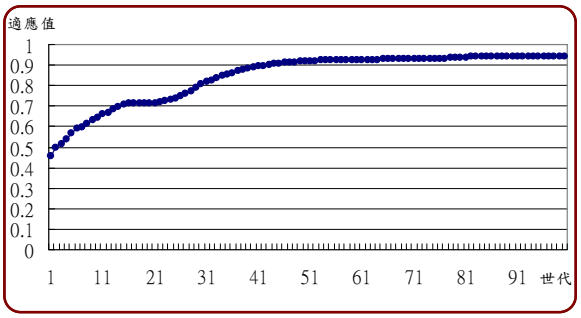
在採樣階段，正常封包樣本共 100 筆案例、異常封包樣本共 100 筆案例，其內容如表四所示。在訓練階段，訓練案例其總數為 900 筆。在實驗的同時我們發現，很多 DDOS/DOS 攻擊程式如果夾雜在背景流量之中，其實不易被發覺，為了要克服這個問題。於是我們在訓練階段，所使用的正常案例就需要使用擬真的環境，同時為了避免訓練的案例差異性太小，我們便將訓練的正常與異常的案例，分成四個類型。如表四所示。

表四 採樣、訓練階段裡正常與異常案例表

正常案例	異常案例
FTP 下載	DDOS/DOS
Web 下載	DDOS/DOS+FTP 下載
靜滯狀態	DDOS/DOS+Web 下載
FTP 下載和 Web 下載	混和眾多 DDOS/DOS

基於基因演算法其流程，於是我們產生二十次，

每一次均產生一百個世代，每一個世代均產生十條染色體，於是我們平均每一世代的適應值繪製而成圖五。由圖五上得知經過五十世代的繁衍後，染色體的適應值便逐漸趨於穩定，在第一百世代時其適應值更高達 0.944444。



圖五　訓練階段適應值的變化

於是我們挑選出一組適應值最高的權重，進行測試，在測試階段分為三大類，分別為已知攻擊、未知攻擊、混合攻擊。每一類均使用 1800 筆案例，其內容如表五所示。

表五 測試階段異常案例表

已知攻擊	使用訓練階段的二十支攻擊程式
未知攻擊	使用訓練階段未曾出現過的四十五支攻擊程式
混合攻擊	使用所有蒐集而來的六十五支攻擊程式

於是我們將訓練階段適應值最高的一組權重固定，並重新且計算測試階段的適應值 如表六上所示的『二十六個特徵』。緊接著篩選出權重較大的前 12 名進行第二部分的測試。如表六上所示的『十二個特徵』。這部分主要用以縮減特徵的數目並證明篩選出的特徵為有效特徵，Top12 特徵細項如表七所示。

由實驗結果發現，在相同的測試樣本中。過多的特徵並不會有效的提升偵測率反而會浪費系統資源與時間，同時過多的特徵還會降低偵測率。

表六 測試階段正常與異常案例表

測試名稱	二十六個特徵	十二個特徵
已知攻擊	0.950556	0.951667
未知攻擊	0.6738	0.672778
混合攻擊	0.7077	0.7083

表七 Top12 特徵表

特徵名稱	**權重值**	**權重排名**
TCP len_port 20	0.9352	1
TCP SYN	0.9277	2
ARP len_err	0.9046	3
S.IP_SYN	0.887	4
IGMP big_len	0.8485	5
封包總數	0.7611	6
S.IP_URG_err	0.6889	7
IP Hdr_len_err	0.5966	8
IP D_Address	0.5341	9
IGMP type_err	0.3701	10
TCP Dport	0.3683	11
ICMP type_err	0.317	12

5. 結論

經由反覆的實驗證明，我們發現在眾多特徵之中，只有部分的特徵對於攻擊有決定性的影響，同時我們藉由這一次的實驗，更可以突顯出特徵選取的重要性。我們下一個目標，便致力於找出蠕蟲(Worm)、特洛伊木馬程式(Trojan)、後門程式(back door)…等。攻擊程式的有效特徵，希望藉由這種方式可以改進入侵偵測系統的效率，並對網路安全進一份心力。

誌謝:感謝行政院國科會專題研究計畫之補助(NSC 95-2221-E-130-003)，使本論文得以順利完成。

參考文獻

[1] Fabrizio S., “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, 2002.

[2] Li B., Yu S. Lu Q., “An Improved K-Nearest Neighbor Algorithm for Text Categorization”, Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, 2003.

[3] William W. Cohen and Yoram Singer, “Context-Sensitive Learning Methods for Text Categorization”, Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 307-315.

[4] Melanie J. Middlemiss and Grant Dick, “Weighted Feature Extraction using a Genetic Algorithm for Intrusion Detection”,Evolutionary Computation, 2003. CEC '03,8-12 Dec. 2003 pp.

1669- 1675, Vol.3.

[5] Ming-Yang Su, Kai-Chi Chang and Hua-Fu Wei, "A Fast Algorithm for Generating Fuzzy Rules - Online for Network Intrusion Detection Systems", 2007 第十七屆資訊安全會議

[6] 李駿偉, “入侵偵測系統分析方法效能之定量評估(A Quantitative Performance Evaluation on Intrusion Detection Analysis Methods)”, 中原大學資訊工程研究所,碩士論文, 2002

[7] 李維漢, “網際網路惡意程式之活動調查－以某企業對外網路連線為例(On Investigation of Malicious Software’s Activities－A Case Study on a Company’s Internet Connections)”, 中央大學資訊管理研究所, 碩士論文, 2005

[8] 林豐澤, “演化式計算上篇：演化式演算法的三種理論模式, “智慧科技與應用統計學報,第三卷,第一期, pp. 1-28, 2005 年 6 月。

[9] 林豐澤, “演化式計算下篇：基因演算法以及三種應用實例, “智慧科技與應用統計學報,第三卷,第一期, pp. 29-56, 2005 年 6 月。

[10] 林卓彥, “自動分類方法之比較(Comparison of Automatic Classification Methods) “, 國立中正大學資訊工程研究所, 2005。

[11] 陳明佐, “植基於階層式規則之入侵偵測系統(Intrusion Detection System Based on Hierarchical Rules) “, 逢甲大學資訊工程研究所, 2002。

[12] 蘇民揚, 葉生正, 林呈俞, 張瑞德, “植基於模糊關聯規則的網路入侵偵測系統,” Journal of Internet Technology, Vol. 8, No. 2, pp. 221-228, 2007. (TSCI, EI)