復興崗學報
民 97，92 期，177-212

以資料探勘技術改善國軍網路入侵偵測效能之研究

吳文進
國立政治大學資管系 博士研究生

摘 要

資訊及通訊科技的快速發展與普及化，已成為生活與工作中不可或缺的重要工具，不論政府機關（構）、企業組織乃至於個人用戶，對電腦與網路的依賴日趨緊密，但相對所引發的資通安全防護問題卻益顯嚴重。因應數位時代的來臨，政府部門自 1994 年起陸續推展「數位台灣」相關計畫，並於 2008 年 3 月首度發表「資通安全政策白皮書」，揭示資通信科技發展與資通安全政策要求應同步與時推移之決心。國軍屬政府機關一環，在資訊化、自動化建軍理念引領下，不容置外於資通安全政策規範，尤其因應戰備任務需求與國軍網路機敏特性，除屬行「實體隔離」政策以實現「資訊邊疆」保護概念之外，更應採行「保護－偵測－反應－復原」四個風險管控策略來有效防堵資通安全防護缺口。

在眾多資通安全防護機制中，入侵偵測系統可以有效偵測入侵滲透及人員濫（誤）用行為，並能提供適切的補償控制措施與建議，符合風險管控保護、偵測及反應之防護策略，可視為國軍落實資通安全政策要求與落實執行之重要機制。為結合現階段國軍需求，發展專用之入侵偵測系統以確保國軍網路安全，本研究深入分析入侵偵測系統運作之架構，從封包偵測效能瓶頸點，尋思結合資料探勘技術與前置封包分類器來改善偵測效能，經實驗證明，利用資料探勘技術實作前置封包表頭分類器的確可以改善封包比對效能不佳與正確率低等問題，而在多種分類器演算法中，又以倒傳遞類神經網路最佳，其正確率 92.704%稍低於約略集合理論 92.867%，但執行速度可增加約 112.95 倍，同時還兼具偵測未知攻擊的能力。本研究除能為國軍落實資通安全防護提供更多的選項之外，前置封包分類器的實作亦能使國軍實現資通安全無虞之目標更邁向前一步。

關鍵字：入侵偵測、倒傳遞類神經網路、資料探勘、網路安全、資訊安全

The Study on Improving Intrusion Detection Efficiency of the Military Internet by the Technique of Data Mining

Wen-Chin Wu
Doctoral Student
Department of Management Information System
National ChenChi University

Abstract

Intrusion detection system can effectively detect intrusion, penetration and the misuse behavior by staff as well as appropriately apply for compensation measure, control measure, and suggestion to fit the protection strategy of risk control and prevention, detection and reaction in the midst of the information and communication security mechanism. Intrusion detection system is an important mechanism to request and perform the information and communication security policy for military. The study is to analyze the operation framework of intrusion detection system for finding the bottleneck of intrusion detection system efficiency in order to help the military build a personalized intrusion detection system to protect military networks. Setting packet classifier and head classifier before using the technique of data mining could improve poor efficiency of compare package and low accuracy through the evidence from experiments.

Back Propagation Neural Network（BPN） is the best one of the multiple classifier algorithm, and its accuracy is up to 92.704% roughly lower than Rough Set Theory（RST） is up to 92.867%; however its executed velocity could increase some 112.95 times and detect the unforeseen the attacking ability. Therefore, the study is not only to apply for more choices to perform information and communication protection but also to help the military to reach the goal about the nest step in the future of not leaking the military information and communication.

keywords：intrusion detection、back propagation neural network、data mining、internet security、information security

吳文進

壹、緒論

一、國軍資通安全防護現況

資訊及通訊科技（Information and Communication Technology, ICT）的快速發展與普及化，已成為生活與工作中不可或缺的重要工具，不論政府機關（構）、企業組織乃至於個人用戶，對電腦與網路的依賴日趨緊密，但相對所引發的資通安全防護問題卻益顯嚴重。究其原因，係吾人所面對的數位威脅，其入侵本質及方法均有別於傳統的思維模式且不斷翻新，尤其高度資訊化、網路化之後，入侵工具可經由網路蒐尋輕易下載，攻擊模式新穎且具創意，尤其這些手法均兼具自動化、遠端遙控及全球化散播能力（Bruce Schneier 2001），這些將徹底顛覆往常我們所熟知的安全觀念與防護作為，畢竟與攻擊者相關之資訊包含人、事、時、地、物、如何及為何等因素，我們均無法確切地掌握。

因應數位時代的來臨，我國自 1994 年成立跨部會專案小組，推動「國家資訊通信基本建設」；2001 年成立行政院「國家資通安全會報」，下設六個工作組，並經過評估後納管政府機關共 3,713 單位；另依資訊能力、重要性、機敏性及保護標的等要素，將資通安全防護等級由高至低以 A、B、C、D 區分，及將事件由輕至重以 1~4 級加以區別；2005 年要求各政府機關（構）主管資通安全業務之副首長兼任資訊安全長；2006 年將納管單位擴及教育體系並增加為 6,797 個單位，俾能全面提升國家資通安全防護能力與水準。另 2008 年 3 月，行政院科技顧組首度正式發表「資通安全政策白皮書」，揭示「安全信賴的資訊化社會、安心優質的數位化生活」之願景（行政院科技顧問組 2008）。綜觀我國資通安全基礎建設相關安全機制之推動，政府機關扮演關鍵催化的角色，而我國軍單位為政府機關之一環，除納編「國家資通安全全會報」所轄標準規範、稽核服務、法規偵防、資訊服務及通報應變等工作組之外，更因國軍網路戰備任務需求及其機敏特性，整體資通安全防護等級訂為 A 級，而各單位之等級則由國防部通資部門通盤考量後統一律訂，足見國軍單位對國家整體資通安全工作之推展具有舉足輕重的地位。此外，國軍自陸續推展「精進案」、「精實案」之後，以「量小、質精、戰力強」為目標，大量引入資訊科技及發展資訊系統以填補因人員精簡所產生任務遂行上的間隙，已逐步走向資訊化、自動化的現代化國軍。因此可預見地，未來國軍除應賡續戮力於戰訓本務之外，資通安全防護之落實與推展亦應與時俱進並符合國家資通安全政策，畢竟「安全是一切的基礎，沒有安全就沒有一切」。

在資訊安全防護的領域中，除防毒軟體、防火牆、虛擬私人網路、加密系統、備援及認證等機制外，入侵偵測系統（Intrusion Detection System, IDS）已繼防火牆（firewall）之後成為落實企業組織資產保護的第二道防線，因為 IDS 可以有效的偵測外部駭客的入侵滲透及內部人員的濫（誤）用行為，並能提供適切的補償控制措施與建議。檢視當前國軍厲行「實體隔離」政策，在「資訊邊疆（information frontier）」的概念推展之下，雖有一定的成效，但國軍網路的普及與頻寬的增加，導致封包流量大增，相關也提高了遭受入侵或管控失當衍生機密資訊外洩的風險。因此如何更精確的掌握網路入侵警訊，並能兼顧「保護－偵測－反應－復原」四個風險管控策略環節（行政院科技顧問組 2008），為國軍單位落實行政院「資通安全政策白皮書」相關規範首應面對的挑戰。在資訊安全實務運用上，IDS 因同時兼具保護、偵測及反應等管控機制，深受業界推崇，國軍單位如能結合「實體隔離」政策，發展專屬的 IDS 應可有效防堵資通安全缺口，有效提升資通安全防護強度，進一步落實國軍整體安全。

二、IDS 的發展趨勢

入侵偵測的觀念始於 1980 年，迄今已逾廿餘年，其間雖不乏學者及商業投資行為遂行研究，惟相關技術仍未成熟，功能亦有所限制，如偵測效能不足、誤報率高、無法與網路頻寬匹配等問題。尤其自 2000 年之後，電腦與網路的蓬勃發展，使得網路使用者樣態日趨複雜，各式各樣新生的攻擊模式與手法不斷地出現，同時整體網路流量大幅增加，亦加深網路即時監偵的難度，這些問題將使入侵偵測技術的效能面臨更嚴峻的考驗。分析當前入侵偵測技術之發展與運用，不僅種類繁多，功能亦多所差異，若欲與之分類，約略可從分析的資料來源及分析的技術兩個面向來加以區分。就分析的資料來源視之，主要為網路型入侵偵測系統（Network based IDS, NIDS）及主機型入侵偵測系統（Host based IDS, HIDS），NIDS 是利用封包監聽的技術，在企業組織重要網段側錄所有進出的封包，並利用特有的演算法進行比對，這種防護架構的優點是防禦面廣，佈署、偵測、稽核及維護管理的成本較低，同時可以即時找出攻擊行為並提供適切之回應行動建議，但也由於封包比對的負荷相當重，如採較嚴格之稽核政策，將導致比對不及而有掉封包（drop packet）之現象，連帶也會影響整體偵測的正確率，此外因對加密的封包因無法解析，故無從檢測判定；而 HIDS 則是裝設於企業組織的重要主機上，優點是能提供該主機完整之保護措施，沒有加密封包無法解析的問題，但防禦面小，主機端資源之損耗及對攻擊反應效率較差是其主要的缺點（吳文進 2004）。

另從分析技術來看，主要分為誤用偵測（Misuse）及異常偵測（Anomaly）兩種，前者是使用特徵比對（signature based）的方式，依其事先定義的特徵規則（rule）做為入侵判斷的依據，正確率很高，但因軟體弱點與漏洞的數量日益增多，使得規則數量膨脹迅速，間接影響偵測效能，而對未知的攻擊模式無法偵測，是此種方法的主要限制；後者主要是利用統計的方法歸納及建立正常特徵（normal pattern），並利用啟發式（heuristic）的方法作為入侵判斷的準據，凡超過一定的門檻值（threshold）即視為異常，優點是偵測速度較快且能偵測未知的攻擊模式，但其先天的模糊性使其相對容易產生較高的誤判率。近年來入侵偵測之架構發展趨勢為網路型入侵偵測系統置於企業防火牆之後，並配合主機端設置輕量型之主機型入侵偵測系統，以期能同時獲致較高的正確率及偵測率，其核心技術則是利用誤用偵測技術先行過濾已知攻擊，並於主機端裝設異常偵測機制，俾能同時提升偵測率及降低誤判率，以確保企業組織重要網路區段安全無慮。

三、IDS 偵測技術所面臨的問題

近年來入侵偵測技術雖然有長足的進步，各式新的比對演算法不斷被提出且實際運用在 IDS 上，但入侵偵測技術仍有精進的空間，一般評斷 IDS 之良窳，可以從偵測正確率及效能兩個面向來討論，正確率包含誤報（false positive）及判錯（false negative），其高低取決於偵測引擎比對演算法的精確度，精確度愈高愈能找出更多的入侵行為，但效能必隨之降低，因此必須適當取捨；而偵測效能主要受網路流量及頻寬影響，此外入侵手法翻新及軟體潛藏缺陷，除使入侵特徵（pattern）或規則（rule）遽增外，入侵特徵的複雜度亦相對的提高，這些都意味著偵測核心必需採用更具效能的偵測技術，方能勝任日趨複雜的電腦與網路作業環境。

根據Gupta及McKeown在（2001）提出的「封包分類演算法」（Algorithms for packet classification）相關研究報告顯示，預先對封包進行系統化分類將有助於後續特徵比對效能之提升（P. Gupta & N. McKeown 2001, 24-32）；而Kruegel等人（2002）利用類似網路流量負載平衡（Slicing Approach）的方法，配合高速網路進行流量的分散來提升入侵偵測核心之比對效能，依其實驗結果，單一偵測點的網路流量超過130Mbps後將會使原有偵測率的穩定度驟降，同時規則（rules）數量的增加亦是造成偵測率驟降的原因（C. Kruegel et al. 2002, 285-294）；Charitakis（2003）則利用封包偵測兩階段處理的作法，第一階段稱為 “Early Filtering”，藉封包前處理程序直接篩檢過濾不需比對的封包，第二階段稱 “Locality Buffer”，利用記憶體快取的特性，將已預先分類之封包載入，再與記憶

體中預存之入侵規則子集合進行比對，不僅提高記憶體的利用率，也相對提升比對效能（I.Charitakis et al. 2003, 238－241）；至於Bolzoni 等人在2005年所發表“Poseidon~兩階段誤用入侵偵測系統”，其作法大致上也繼承類似的概念，惟採用自我組織映射圖網路（Self Organizing Map, SOM）做為前處理的核心（D. Bolzoni et al. 2006, 1-10），對偵測效能之提升亦達一定之效果。

四、本文研究方向、方法與限制

本研究將參考以往 IDS 有關效能提昇之研究，從入侵偵測的架構加以分析，找出封包前置處理是影響入侵偵測效能之關鍵因素後，採用 Snort IDS 為測試平台，從眾多資料探勘（Data Mining）技術中，選用約略集合理論、支援向量機及類神經網路等技術；另囿於測試資料集取得不易，仍使用 IDS 相關研究最廣泛採用之麻省理工學院林肯實驗室 KDD-Cup’99 的資料集以進行實作並比較其效能，以期能發展出偵測速度快、比對正確率高之偵測引擎，進而改善國軍網路入侵偵測效能。

貳、文獻探討

網路安全防護技術之精進，繫於不斷創新研究，而技術成熟與否，效能是一個重要的評估指標。本節首先介紹入侵偵測系統，再參考以往入侵偵測相關文獻所提之研究成果，分析其概念與作法，藉各種偵測技術運作架構之比較，瞭解其對封包流量之處理能力及對偵測效能的影響，從而說明本研究所採前置封包表頭分類器的原因。

一、入侵偵測系統簡介

在資訊安全領域中，一般所謂的「入侵」是指資訊系統內未經授權的存取或活動，而「入侵偵測」則針對這些已發生、進行中或即將進行的入侵意圖之一種確認程序（Paul E. Proctor 2002）；另依據美國國家標準暨技術機構（National Institute of Standards and Technology, NIST）的定義，入侵為任何企圖危及資訊資源的機密性、完整性及可用性的所有活動，而這些活動包括個人或代理人企圖闖入、誤用某一系統或網路，致違反其所建之安全政策，因此「入侵偵測系統」就是為監控這些資訊系統及網路活動程序，並能對入侵的行為與跡證進行分析、儲存及示警的一種系統（Rebecca Bace 2003）（Joseph S. sheriff & Rod Ayers 2003,

222-229）。

除前述定義之外，入侵偵測系統效能評估一般是以誤判率來加以衡量，誤判率可區分為「誤報」（false positive）及「判錯」（false negative）兩種，所謂 false positive 是指預警條件未成立但系統卻判定遭受入侵而發出警報，相對地 false negative 則指入侵條件已成立但卻未發出預警。圖 1 即以真假（true/false）、肯否定（positive/negative）兩種屬性組合來討論入侵偵測的四種狀況，其中 false positive 及 false negative 是入侵偵測領域相關研究人員亟欲克服的部分，太高的 false positive 將使假警報充斥而網管人員疲於奔命，更甚者不予理會警報訊息或將之關閉而致使系統形同虛設，至於 false negative 則要靠提高偵測的精確度來解決，兩者均為 IDS 效能評估上的重要指標。因此兩者之值愈小代表偵測正確率愈高，若就重要性而言，false negative 是較嚴重的錯誤，因為屬於入侵封包卻未被偵測出來，而 false positive 則是假警報，雖不影響系統運作，但過多的警訊將肇致使用者降低使用意願。

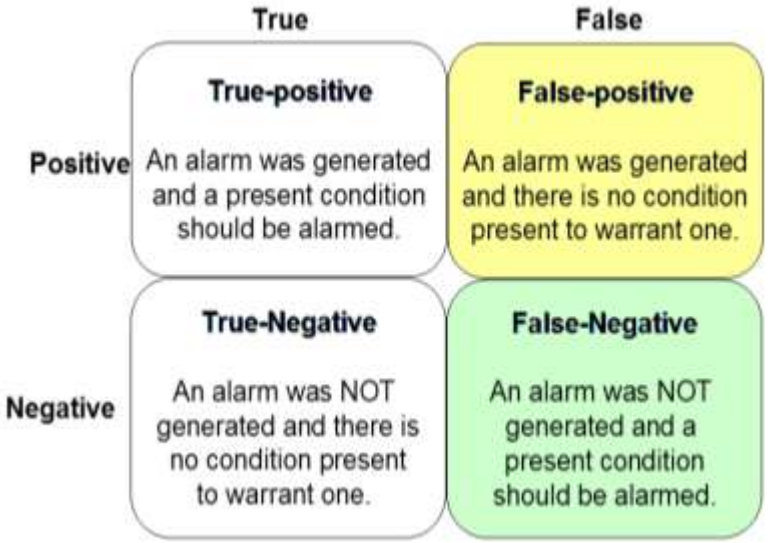

圖 1　偵測入侵之真偽分類圖

（資料來源：Paul E. Proctor）

入侵偵測核心技術的發展，雖可區分許多不同的面向，一般均由偵測核心的運作模式來加以探討。偵測模式又區分為分析的資料來源及偵測引擎的分析方法兩大類（李駿偉等人 2002, 21-37; Rebecca Gurley Bace 2001; Northcutt & Judy Novak 2002），相關分類概要彙整如表 1 所示。

表 1 入侵偵測系統分類概要

類別	種類	監控目標及方式
以資料來源區分	網路型 IDS	監控網路封包
	主機型 IDS	監控主機系統內部活動紀錄
	應用程式 IDS	監控應用程式所產生的紀錄檔
	目標式 IDS	監控特殊或機密的檔案資料
	混合型 IDS	結合網路封包及主機系統活動紀錄之監控
以分析方法區分	誤用偵測	採負面列舉，符合所比對規則即判定入侵
	異常偵測	採正面列舉，不符合行為輪廓即視為異常
	混合式偵測	結合兩種分析方法使優缺點互補

（資料來源：本研究整理）

入侵偵測技術因攻擊型態日趨複雜（CERT Statistics 2007），網路頻寬遽增及網站應用與內容物大量成長且多元的影響，單一形式之入侵偵測系統不論在偵測速度及正確率方面均不敷實需，已逐漸朝向混合模式之入侵偵測系統發展。這意味著偵測之資料來源將採混合型 IDS（Hybrid IDS），而依目前實務上多以網路型 IDS 負責路由器端（router）之流量監聽，配合主機型 IDS 或其他專屬種類之 IDS 進行主機端（host）之細部稽核資料監聽及應用程式端資料封包解密；而分析方法部分則採用混合式偵測核心，即以誤用偵測來比對已知攻擊型態，並輔以異常偵測來找出未知之攻擊形態。

二、現行偵測技術之發展與優缺點

（一）偵測技術之演進

入侵偵測之偵測技術種類繁多，早期為較單純之規則基礎（rule based）（Jiawei Han & Micheline Kamber 2001; W. Lee et al. 1999, 120-132; Yang Xiang-Rong et al. 2001, 19-23; Sang-Jun Han & Sung-Bae Cho 2003, 120-125）17,18,19,20，陸續相關研究採用糢糊理論之偵測模式，以提供較佳之未知攻擊偵測能力（Ming-Guang Ouyang et al. 2002, 281-284; Orfila, A et al. 2003, 1237-1242; Zhoujun Xu et al. 2004, 645 – 648）21,22,23。其後則衍生出混合偵測模式的作法，如利用基因演算法（Genetic Algorithm）配合規則基礎以產生較高之模糊性（Lina Wang et al. 2001,13-18; Wei Lu & Traore I. 2003, 2165-2172; Guan Jian et al. 2004, 4339-4342）、及利用隱藏式馬可夫鏈（Hidden Markov Model, HMM）降低資料維度（dimension）以配合模糊理論之偵測模式（Bo Gao et al. 2002, 381-385; Fei

Gao et al. 2003, 893-896; Al-Subaie M. & Zulkernine M. 2006, 325-332）等研究。

而貝氏演算法（Bayesian methodology）則是利用機率統計的特性，可大幅改善入侵偵測系統對於未知攻擊的偵測能力（Burroughs D. J. et al. 2002,329-334; Kruegel C. et al. 2003, 14-23; Anagnostopoulos T. et al. 2005, 425-428），然而其機率統計的核心模型（model）需仰賴適當的訓練資料（training data）以及合適的測試資料（test data）才能發揮顯著之功效，而訓練一個效能良好之核心模型往往需耗費大量的時間與成本，因此隨後出現 Pseudo-Bayes estimator、Naïve-Bayesian 等方法，均是以貝氏演算法為基礎所做的改良。

近年來相關研究則趨向於利用各種不同的演算法來實作核心偵測引擎以達成前述目的，如以類神經網路（Neural Network, NN）（Ryan Jake & Meng-Jang Lin 1998; Lei J.Z. & Ghorbani A. 2004, 190-197; Yufeng Kou et al. 2004, 749-754）、Self Organizing features Map（SOM）（Depren M. O. et al. 2004, 76-79）及基於 SOM 改良之決策支援向量機（Support Vector Machine, SVM）（Mill J. & Inoue A. 2004, 407-410; Faour A. et al. 2006, 3175-3180; Zonghua Zhang & Hong Shen. 2004, 568-573）等相關研究，均顯示入侵偵測技術領域發展之蓬勃與重要性。目前改善 IDS 效能相關解決方案除藉由不同的方式降低資料維度、提升偵測效能外，也有相關研究採用 Agent Based IDS 或 Multi-Agent Based IDS，藉由流量分散或多處理單元分工的方法，以降低偵測核心的運算負載，同樣可以達成提升效能之目的（Kussul N. et al. 2003, 120-122; Hegazy I. M. et al. 2005, 27-30）。

（二）網路流量對偵測效能之影響

經由前述 IDS 相關研究發現，各種偵測模式所採用的方法主要可歸納為減少資料維度、精簡規則數量或降低偵測核心運算負載等方向以進行改善。而參考 Kruegel 等人（2002）利用 Snort 針對單一主機入侵偵測系統進行網路流量及規則數量之於偵測率之實驗中顯示，當網路流量大於 150Mbps 時，偵測率即開始驟降，而在同樣的環境下當比對之規則數量超過 120 則時，偵測率也開始驟降，有關網路流量與特徵規則（rule）增加對 IDS 效能之影響如圖 2 所示。偵測率的下降的原因有許多，在硬體環境方面，如中央處理器之時脈、記憶體之大小；軟體部分則包括程式執行之平台、偵測核心所採演算法之比對速度等。深入分析此偵測率的驟降現象主要是因為偵測引擎本身架構的限制，因為隨著網路封包流量或規則之增加，當偵測引擎處理不及時會主動丟棄封包，而過高的掉包率（packet drop rates）將嚴重影響偵測正確率。

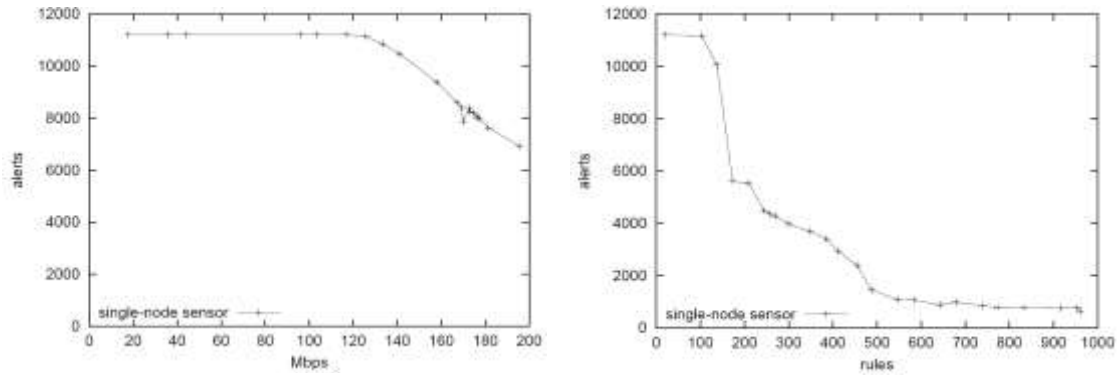

圖 2　網路流量與特徵規則（rule） IDS 效能之影響

（資料來源：修改自 Kruegel 等人）

三、Snort IDS

入侵偵測雖因所使用技術及分析資料來源的不同而發展出各類型的 IDS，惟就目前市場接受度較高是的網路型 rule-based 的 IDS 而言，主要考量是偵測範圍廣且建置成本較低，而同時規則比對有較高的精確度，不易產生 false positive 及 false negative，其中 Snort IDS 堪稱為 rule-based IDS 最具代表性的產品。Snort 是由 Marty Roesch 所設計開發，其特點是具備跨平台能力，採用相容於 TCPDump 的 Berkeley Packet Filter（BPF）過濾器以解析封包，且能將各種不同類型的封包以相同的格式呈現，尤其是開放原始碼的特性，使得核心程式十分穩健，因此效能足以媲美商業版 IDS；此外其屬 open source 之免費軟體，因此核心程式（kernel）依依使用需者自行改良，因此具備相當程度的穩健性及效能。Snort 的架構主要組成元件包括封包監聽（sniffer）、前置處理器（preprocessor）、偵測引擎（detection engine）及示警（alert）系統等四部分，許多研究均運用其架構來實作驗證，整體架構之運作如圖 3 所示。

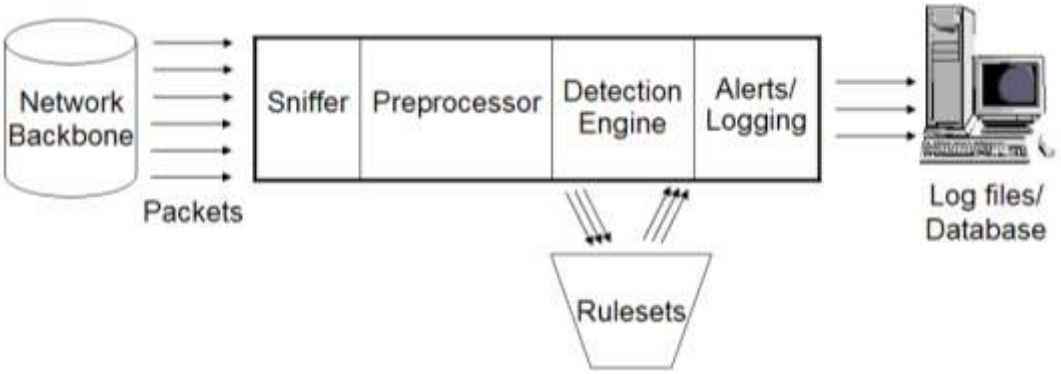

圖 3　Snort 架構示意圖

（資料來源：Jay Beale 等人（2003）所著“Snort 2.0 Intrusion Detection”）

Snort 的偵測作法是利用 libpcap 函式庫抓取網路封包，經解譯器解析及格式轉換等前置處理後，將封包送至偵測引擎並參照特徵資料庫之 rule set 以進行比對，如發現特徵吻合即判定為入侵，並示警及紀錄，如屬正常封包則逕予丟棄。Snort 主要監測 TCP、IP、UDP、ICMP 等四種通信協定之封包，各協定有其所屬的解析器及前處理器據以轉換封包之格式內容，以遂行比對工作，以 Snort 2.x 版為例，其 rule 之格式主要區分為表頭（header）及本體（body，部分研究稱為 content）兩部分：

1.header：紀錄封包的協定、來源及目的之位址與通信埠，以及符合該條規則所應採取的行動準則。

2.body：主要紀錄入侵的相關訊息，包括封包 payload 上的特徵字串、嚴重性之優先程度....等相關資訊計 40 餘個選項。

此外 Snort 採用二維鏈結串列實施封包比對，比對程序包含第一階段的 RTN（rule tree node）之 header 比對及第二階段的 OTN（option tree node）之 body 比對，亦即依封包解析順序比對封包 header，包括來源及目的地等與網路位址與通信埠相關資訊，如果比對吻合即依該 node 所屬的 option tree 繼續比對 body 部分，詳細的運作方式如圖 4 所示。因此一個封包可能在比對過第一階段的所有 RTN 之後，發現均未吻合而丟棄，亦可能在符合 RTN 中所屬的某一 node 之後，但在第二階段之封包 body 的比對時未能吻合而同樣丟棄；唯有兩者皆吻合才會判定為入侵並示警與紀錄，有關 RTN 及 OTN 之兩階段比對的詳細判定流程如圖 5 所示。

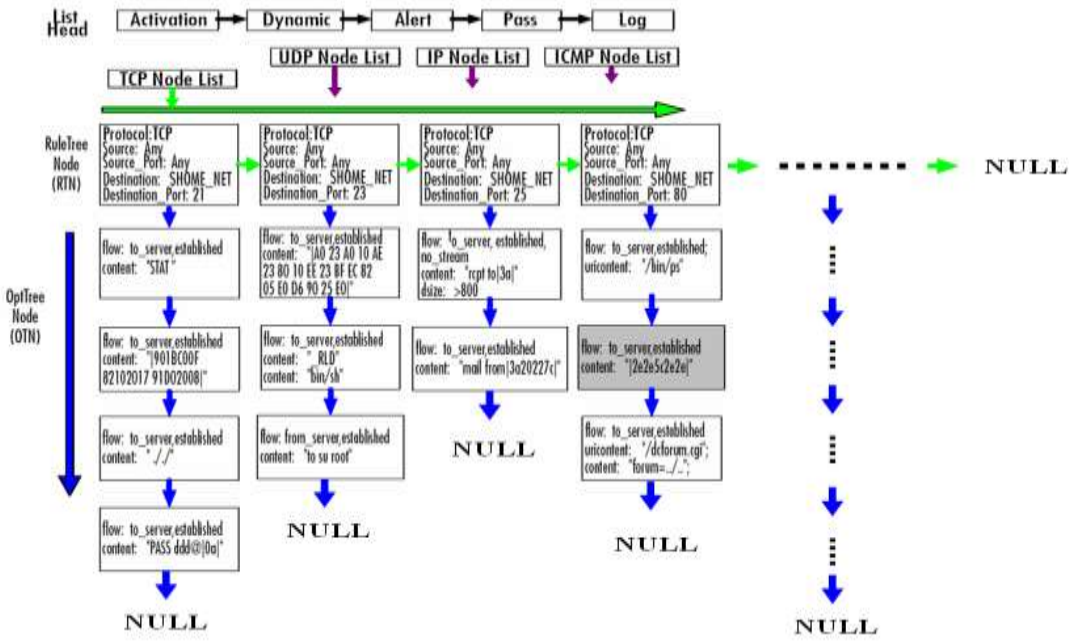

圖 4 Snort 封包比對的樹狀規則

（資料來源：修改自 Jay Beale 等人所著“Snort 2.0 Intrusion Detection” ）

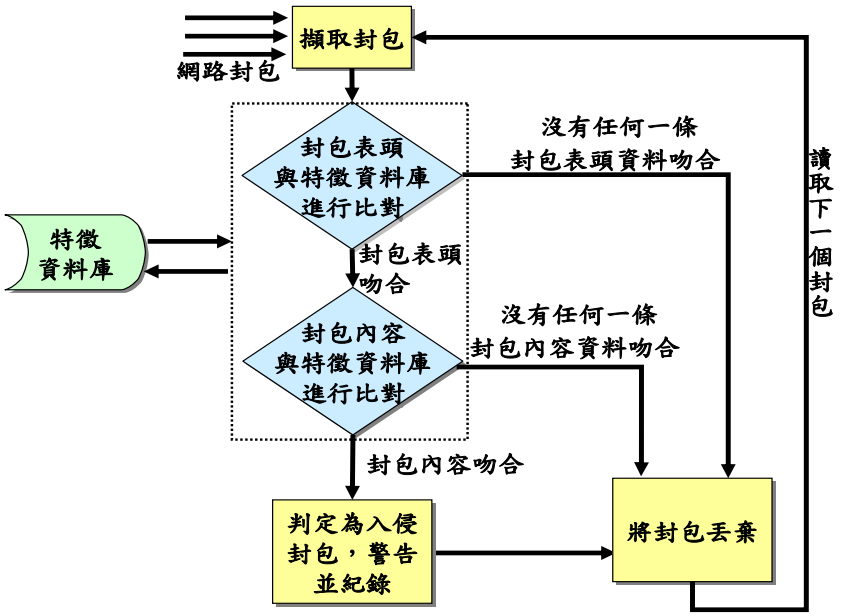

圖 5 Snort 封包特徵比對之流程

（資料來源：修改自 Jay Beale 等人所著“Snort 2.0 Intrusion Detection”）

四、資料探勘相關技術

資料探勘（Data Mining）是指將資料中隱藏的資訊挖掘出來，可視為 Knowledge Discovery 的一部份，Data Mining 使用許多統計分析與 Modeling 的方法，在資料中找尋有用的特徵（Patterns）及關連性（Relationships）。一般而言，Data Mining 功能含括分類（classification）、迴歸（Regression）、推估（estimation）、預測（prediction）、群聚（Clustering）、關聯（Association）、時間序列（Time Series）等，藉由探勘演算法來模擬真實世界以建立模式（Model），這些模式可充分描述資料中的特徵（Patterns）以及關係（Relations）。傳統技術主要利用統計分析方法，如敘述統計、機率論、迴歸分析、類別資料分析、變因分析（factor analysis）、區隔分析（discriminated analysis）、群集分析（cluster analysis）等；而後改良的方法包括類神經網路（artificial neural network）、決策樹（decision tree）、基因演算法（genetic algorithms）、規則推論法（rules induction）、模糊理論（fuzzy logic）等。本研究將參考 IDS 相關研究最普遍採用的資料探勘技術，包括約略集合理論、支持向量機及倒傳遞類神後網路，相關演算法理論說明分述如下：

（一）約略集合理論

約略集合理論（Rough Set Theory, RST）主要用於數字類型資料之探勘，透過 RST 之屬性化簡，可以找到資料之隱藏樣式，並能產生決策法則。RST 由於其準確性高，因此在最近十多年來已經被廣泛的使用在許多決策領域的相關探勘工程（Pawlak, Z. & Rough. 1991; Fayyad et al. 1993, 1002-1007; Pawlak, Z., and Slowinski, R. 1994, 443-459; 黃承龍及唐文政 2003）。

RST 在資料前處理上首先需要產生決策表（Decision table）或資訊表（Information table）用以表示資料的組成，一般係定義為 $S=<U, A, V, f>$，其中 U 為有限物件的集合（Universe），如 $U=\{x_1,x_2,..x_5\}$，A 為屬性（Attributes）的集合，V 是屬性值，而 f 為方法；如果 P 為一個屬性子集合，$P \subseteq A$ 且 $x,y \in U$，則 Ind（P）稱為「不可區別關係」；其次需定義資料所在之「上界」（Upper approximation）與「下界」（Lower approximation），上下邊界（Boundary）是 RST 中兩個極重要的觀念。當 X 為物件之子集合，而 P 為屬性子集合，如 $X \subset U$、$P \subseteq A$，則 X 在 P 中的下界定義為：$\underline{PX} = \{x_i \in U \mid [x_i]_{Ind(P)} \subset X\}$，亦即下界是由 X_i 物件所組成，而 X_i 物件是「不可區別關係」的群組包含於 X 者所組成的，這個群組稱為「P-Lower approximation」；而上界則定義為：$PX = \{x_i \in U \mid [x_i]_{Ind(P)} \cap X \neq 0\}$，用以表示 X_i 物件是屬於「不可區別關係」群組與 X

之交集不為空集合者所組成，稱為「P-Upper approximation」，而 X 在 U 的邊界定義為：$PNX = PX - \underline{PX}$。

在完成物件資料定義之後，可實施資料屬性化簡，因資料所包含之屬性不全然對實驗結果會產生重大影響，因此如能加以化簡則可有效降低複雜度，而 RST 的作法是當「不可區別屬性」=「不可區別屬性之子集合」時，則代表其中有部分屬性是多餘的，只要保留必要的屬性即具代表性，因此可將不要必的屬性抽離。此外 RST 主要處理名目類屬之資料，當資料庫的資料屬性若是連續型態的實數值，就必須將屬性作數值之離散化。在數值屬性離散化上有許多的離散方法，其主要是將連續的數值切割為 N 個區塊，以將屬性之性質作一個明確的區分，若資料切割後為一個區間則我們稱此區間為一個空的區間，此種方法是由 Fayyad 及 Irani 所提出，亦即利用 class-entropy 的概念找出所有屬性之最佳分割（Ryan Jake & Meng-Jang Lin 1998）。

在經過上下界、屬性化簡及資料離散化切割等處理過程後，我們就可得到所謂的決策表，而此決策表係由條件屬性和決策屬性所共同構成，利用決策表可以產生決策規則以作為分類之依據；如進一步將決策法則中的屬性利用聯集的觀念加以歸納（generalized），則許多單筆資料可能重複地符合於不同的法則中，而經歸納後法則數量雖未改變，惟因提高了支持度（Support）而使準確率增加，因此只要將新的資料代入這些最後產生的法則中，即能以實際的數據驗證是否分類正確。

（二）支持向量機

支持向量機（Support Vector Machines, SVM）是以統計學習理論（Statistical learning theory）為基礎所發展出來的機器學習系統，其應用領域相當的廣泛，包括文字分類（Text categorization）、影像識別（Image recognition）、手寫數字辨識（Hand-written digit recognition）、資料探勘（Data Mining）、生物資訊（Bioinformatics）等，極適合處理分類的問題（Vapnik V. 1998; Cristianini N.& Shawf-Taylor J. 2000; Xueqin Zhang et al. 2006, 2594-2598）。

SVM 的主要理論是將分類資料區分為兩部份，如將 X_1 到 X_n 當作訓練資料，而 X_{n+1} 到 X_{n+m} 為測試資料，當我們先將 X_1 到 X_n 的所有資料輸入 SVM 分類後得到反應輸出變數 Y，我們就可以利用其所建立的 model 將測試資料代入以適當分類。而 SVM 是由「最小誤差估計法」（Structural Risk Minimization, SRM）為基礎衍生而來，其目的是使學習分類器在待估測的誤差（expected risk）中能找最小值。

而一般線性支持向量機處理資料的方法首先對每筆不同類的訓練資料加上標註，其值為”+1”或”-1”，假設有一個超平面可以將標註為“+1”及標註為”-1”之

兩類資料區分，則此超平面稱為區分平面（separating hyperplane）；而落在此平面上的所有點必須滿足 $W*X+B=0$ 之超平面法向量（normal vector），接著定義此區分平面之邊界（margin）為 $d+ + d-$，$d+$、$(d-)$為所有標註為"+1"、"(-1)"之訓練資料和區分平面的最短距離。在處理可區分為兩類的資料時，線性支持向量機會找尋具有最大邊界的區分平面，而此類型資料必須符合以下兩個限制式：

1. $X_i * W+B \geq +1$
2. $X_i * W+B \leq -1$

由前兩式可得知 d+ = d- = $\|w\|/1$，而邊界為$\|w\|/2$，因此可求得具有最大邊界的區分平面。

（三）類神經網路

類神經網路運用於入侵偵測領域之相關研究為數甚多，範圍涵蓋誤用及異常偵測，如利用學習功能以 system call 建立入侵行為模式之 profile、利用類神經網路建立關鍵字群以加速字串比對等複雜度較高的問題，此外亦結合其他演算法，進而達成提高正確率、偵測新的入侵行為或提高效能等目的（Richard P. Lippmann & Robert K. Cunningham 2000, 597-603; Bonifacio J. M. et al. 1998. 205-210; SANS Institute 2001）。

眾多研究採用類神經網路模式的原因，主要係其模仿人類神經元的運作，可以處理大量平行分散的資料，並藉學習產生合理推論，極適合解決複雜的問題，其非線性模型的準確性高，具有可接受邏輯、數值、有序及無序分類等輸入的優點，不僅適應性強且應用十分廣泛；但其仍有缺點，如隱藏層數、神經元數、學習率及學習次數等網路參數設定複雜、費時，且只有原則沒有明確的標準；此外其無法解釋所產生的結果，也無法保證訓練的結果就是最好的網路模型，儘管網路不能提供明確規則，當模型運算的結果比瞭解模型是如何運作來得重要時，類神經網路是一個很好的選擇。目前著名的類神經網路模式不下數十種，一般主要的分類有以下四種：

1.監督式學習網路
2.非監督式學習網路
3.聯想式監督式學習網路
4.最適化學習網路

因此類神經網路的應用範圍十分地廣泛，包括工業上如化工廠製程故障診斷及製程監控；商業上如股票投資分析及期貨交易決策；管理上如排程策略選擇；科學上如醫學之疾病診斷及基因分類；資訊科技的應用如語音辨識.....等。而其中監督式學習網路是應用最多且最成功的網路模式，約佔現有應用的 95 %以上（葉怡成 2003）。

監督式的學習網路是依據系統過去輸入和輸出的資料樣本或訓練範例，經過網路學習程序建立系統模型（類聚規則），以推論出新的案例之類屬，因此極適合處理分類及預測等問題。有關倒傳遞類神經網路相關基本理論，包括網路架構及學習程序等原理簡述如次

（1）倒傳遞類神經網路的基本架構

倒傳遞（Back Propagation Neural Network, BPN）類神經網路模式屬於監督式學習網路，是目前類神經網路學習模式中最具代表性的一個，同時也是最廣為運用的模式。倒傳遞類神經網路基本原理是利用「最陡坡降法」（the gradient steepest descent method）的觀念，將實際輸出與期望輸出的誤差函數予以最小化，由於增加隱藏層及採用平滑可微分的轉換函數，使得網路應用最陡坡降法導出修正網路加權值的公式。因此只要給定足夠的隱藏單元，線性閥值函線的多層前饋類神經網路將可逼近任何函數。其網路架構如圖 6 所示。

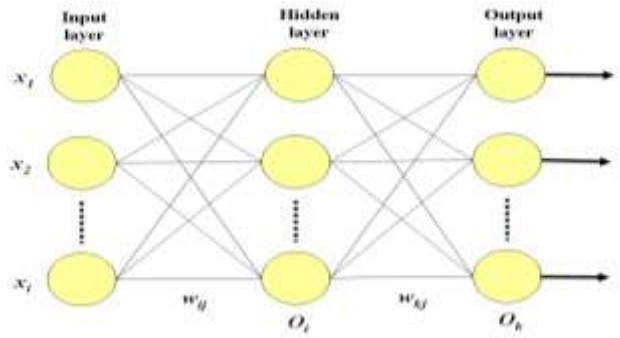

圖 6 倒傳遞網路架構圖

（資料來源：Jiawei, 2001）

在上圖中，訓練樣本 $X=(x_1, x_2, ..., x_i)$ 餵資料給輸入層，每層之間存在著加權連接，其中 w_{ij} 表示某層的單元 j 到前一層的單元 i 的權重（weight）。其各層的功能分述如下：

1.輸入層：用以代表網路的輸入變數，其處理單元數目依問題而定，並採用線性轉換函數，即 $f(x) = X$。

2.隱藏層：用以呈現輸入處理單元間的交互影響，其處理單元數目並無標準方法可以決定，通常需透由 trail-and-error 的過程來決定其最佳數目；此外因係採非線性轉換函數，故網路可以是一層以上的隱藏層，也可以設計不具隱藏層。

3.輸出層：用以表現網路的輸出變數，其處理單元數目依問題而定，亦使用非線性轉換函數。

（2）倒傳遞網路的學習程序

基於前述網路拓撲，網路學習程序概分為初始化、向前及向後等三個過程，分述如次：

1.初始化權重（initialize the weight）：

網路的權重被初始化成很小的隨機數，如介於-1 到 1，或-0.5 到 0.5 的區間，而每個區間均有一個偏差（bias），偏差也類似地被初始化為極小的隨機數。

2.向前傳播輸入（Propagate the input forward）：

此步驟在計算隱藏層及輸出層間每個單元的淨輸入和輸出。首先由訓練樣本提供資料給輸入層，對輸入層的單元 j 而言，它的輸出等於其輸入，亦即 $O_j = I_j$，然後隱藏層和輸出層的每個單元之淨輸入由其輸入的線性組合計算，圖 7 的模型即顯示一個人工神經元由輸入單元、集成函數和轉換函數等三個基本元件所組成。

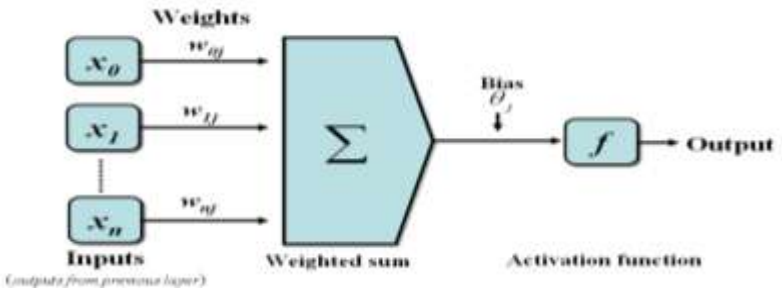

圖 7　非線性神經元的模型

（資料來源：Jiawei , 2001）

而所謂的輸入單元即是一組突觸的強度（以權重值之大小表示之），如輸入單元 x_1 經突觸 1 連接到神經元 j，則突觸的權重以 w_{1j} 表示，集成函數將所有輸入變數的值合併成一個加權總和，並加上偏差 θ_j，該偏差充當閥值（threshold），用以改變單元的活性。其計算式如下：

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

轉換函數 f 代表神經元的活性，依集成函數所得的結果轉換輸出值，如給定單元 j 的輸入 I_j，則單元 j 的輸出 O_j 用下式計算：

$$O_j = \frac{1}{1+e^{-I_j}}$$

此函數可將較大的輸入值域映射到介於 0 到 1 之間的較小區間，而達到收斂之目的。

3.向後傳遞誤差（backpropagate the error）：

透過更新權重和偏差來反映網路預測的誤差，再向後傳遞誤差。對於輸出層單元 j ，誤差 Err_j 計算：

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

其中 O_j 是單元 j 的實際輸出值，而 T_j 是 j 基於給定訓練的已知歸屬的真正輸出。而隱藏層單元 j 的誤差計算如下：

$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

其中，w_{jk} 是指由下一較高層中單元 k 到單元 j 的連接權重，Err_k 則是單元 k 的誤差。至於權重和偏差會被更新以反映傳遞的誤差，權重由下式更新

$$\Delta w_{ij} = (l)Err_j O_i$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

其中 Δw_{ij} 是 w_{ij} 的改變。（5）式中變數 l 是指學習率（learning rate），通常會取 0 到 1 之間的常數值，學習率可避免陷入決策空間的局部最小（也就是權值看似收斂，但卻不是最佳解），並有助於找到全域最小值。如果學習率太小，則學習效果會十分緩慢，太大則又造成數值振盪而無法收斂，依據經驗法則可設為 1/t，其中 t 代表已對訓練樣本資料集進行過的疊代（iteration）次數。偏差可由下式更新：

$$\Delta\theta_j = (l)Err_j$$

$$\theta_j = \theta_j + \Delta\theta_j$$

當整個網路達到以下三個條件之一即終止訓練：

（1）前一週期所有的 Δw_{ij} 都很小，且小於某個指定的閥值。

（2）前一週期未正確分類的樣本百分比小於某個閥值。

（3）超過預先指定的訓練次數。

五、以前置處理器提升 IDS 效能相關作法

有關 IDS 封包前置處理器（preprocessor），目前學者所提出的類型有多種，但如何降低偵測核心比對之負載，進而提升入侵偵測系統整體效能之想法卻顯一致。相關方法不外乎以下三類：（1）偵測演算法：利用多種演算法之先後組合，

先期過濾封包或對封包先期進行分類、（2）規則管理：針對規則或關聯法則之效能最佳化之改良、（3）網路流量：利用多種網路負載平衡之技術，先期將流量分散至不同之入侵偵測處理單元（如 sensor），並以分散式處理；以下將分別就這三類作法加以說明。

（一）偵測演算法之前置處理器

Charitakis（2003）提出之"Early Filtering（EF）"方法採用兩段式的設計，第一階段僅針對封包表頭(header)進行比對，經判定不需要比對封包內容(payload)者即予丟棄；第二階段採用"Locality Buffer（LB）"的封包預先分類快取設計，利用記憶體快取的特性盡量將同類型的封包集中於記憶體內與規則進行比對。僅第一階段 EF 的方法即可將封包減量 32% （效能上可提升約 8%），而和 LB 一併使用則可達到 20%之效能改善。然而以 EF 進行比對，其效能之瓶頸仍因 rule based 先天性之限制，尤其 rule 數量急遽增加時，其 EF 效能受到影響愈為明顯；LB 部分雖採用記憶體快取設計，理論上應可提升比對效能，惟在大量的規則完全載入記憶體後，僅剩餘少量可供封包比對之記憶體區塊，因此對效能之增加仍屬有限，此外 LB 採用何種封包分類演算法可使效能達到最佳化，亦是另一個值得深入研究之議題。

Damiano（2005）所提出的 Poseidon 兩階段異常偵測入侵系統，基於 PAYL 之入侵偵測架構並加上 SOM （1-layer, unsupervised） 做為封包之前置分類處理器，將原本 PAYL 系統之偵測率由 58.8% 提升為 73.2%，且判錯率皆小於 1%。然而該實驗之 SOM 模型訓練是採用 attack free 之資料進行訓練，不同之協定採用不同之 SOM 模型，總訓練筆數為 2,444,591 筆，極可能產生 over-fitting 之問題，此外其實驗設定僅針對四種 service 型態，與實務上各項 service 可能遭受攻擊之差異甚大。

（二）規則管理前置處理器

Sergei 等人在 2000 年提出之規則編譯器 SNORTRAN[52]，係在 Snort 之實驗環境下，利用 benchmarking setwise algorithms 將原本 Snort 之 rule 架構重新規劃，提高其運作時之資源利用率，經其實驗可提升 1.3 至 3.32 倍之效能（至少 800 則規則條件下）。然而此等作法將使得不同之入侵偵測系統之規則架構必須使用不同之規則編譯器，規則透通性及再利用率較低。

（三）網路流量前置處理器

直接改變網路架構藉分散流量來提升效能的概念來自於傳統流量負載平衡（Load Balance），常見的作法除了將骨幹網路升級為 Gigabit 或光纖外，採用前置之負載平衡橋接器（Switch）、路由器（Router）、後端處理器之多中央處理單元（SMP）、叢集伺服器（Cluster）或改良封包分類演算法等，皆為常見之解

決方案。Kruegel 等人以 Snort 為入侵偵測系統之企業及流量分散架構，試圖將流量分散處理後仍能保有網路串流交易之連慣性（statefulness），以解決較先進之 stateless 攻擊型態，並提升大網路流量下的承載能力。Gupta 等人（2001）之研究調查中，則介紹多種網路層等級之封包分類方法，其藉由較低階之封包路由調校，使適用性更廣，且後續封包比對效能亦獲致不錯的改善。然而網路流量之前置處理器所要求之硬體成本極高，若非於網路規劃時即考慮配合入侵偵測系統之流量分散擴充計畫（scalability），其後續建置或預算成本之追加，易成為此種方法在實行上的障礙。

本研究則採用前述提升偵測效能方法之前兩種，亦即從比對演算法及入侵規則化簡兩部分著手，有關整體模擬系統則三種方法兼用之，並透由實作驗證以尋求較佳的解決方案。

參、研究內容與架構

一、研究內容

為實現前置封包表頭分類器偵測模式，本研究將以 Snort 的架構為基礎施以小幅度修改。由於 Snort 主要監測 TCP、IP、UDP、ICMP 等四種通訊協定，利用 libpcap 函式庫抓取封包，經協定解析器解析及資料格式轉換等前處理程序後，封包會被送至偵測引擎依封包表頭（header）的 RTN 及本體（body）的 OTN 兩階段比對，如發現 header 及 body 的特徵均吻合，即判定入侵並示警及紀錄，如屬正常封包則與予丟棄。依此架構可以瞭解，RTN 的封包表頭比對是整個比對程序的第一步，如果能先行以分類器濾除屬於正常的封包，則可大幅減少封包表頭比對的次數；因此本研究仍參考 Snort 2.x 版的架構，惟一不同之處在 RTN 比對前加入前置封包表頭分類器，並分別利用倒傳遞類神經網路、約略集合理論、支持向量機等三種資料探勘演算法，來建立封包分類的模型，除測試分類器之正確率外，並與 Snort 內建的 Boyer Moore 分類演算法進行比較，期藉適度的訓練封包表頭分類器，可以有效節省 EvalHeader（佔 8.5%）及 CheckSrcIPNotEq（6.7%）兩個函式的執行成本。至於 Snort 架構的其它重要元件包括封包監聽、前置處理器及示警系統等部分如何運用配合，因與模擬實驗之核心無緊密關係，故不列入本研究之範圍。

二、系統架構之改良與模擬

本研究主要是改良 rule based 類型之封包處理程序，利用前置封包表頭分類器（pre-classifier），先期對封包表頭進行過濾，亦即透過演算法將分類結果屬於”正常”的封包丟棄，以使核心偵測引擎專注於第一階段被歸類為”攻擊”類型的封包，以進行第二階段的內容比對，因此只要封包表頭分類之正確率達到合理的範圍，則經前置封包表頭分類器先期過濾後仍需比對內容的封包將大幅減量，如此不僅可克服網路流量增大導致掉封包的情形，也可解決入侵 pattern 或 rule 的遽增的問題，而整體偵測效能將因而大幅提升，整體模擬架構如圖 8 所示。

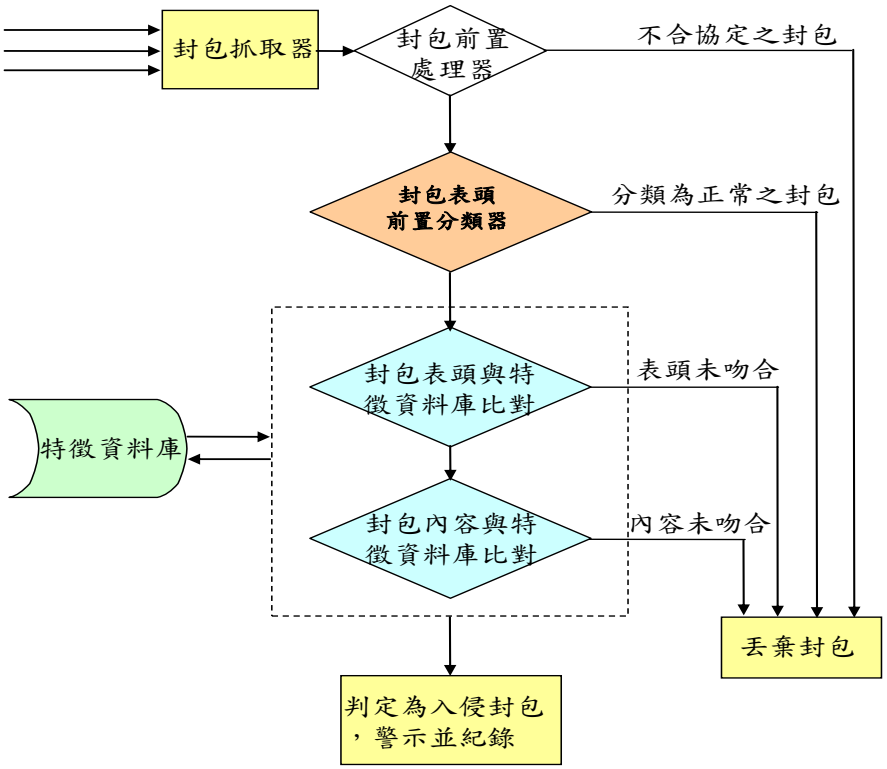

圖 8　模擬架構示意圖

（資料來源：本研究整理）

三、實驗模擬設計

本研究將利用以 KDD-Cup’99 資料集實驗模擬的方式來驗證及評估前節所提相關架構及方法，以下將說明模擬資料之來源、資料隨機處理，以及實驗模擬測試的重點要項。

（一）KDD 實驗資料集

KDD-Cup'99 是由麻省理工學院林肯實驗室於 1998 年所發布 DARPA 入侵偵測資料集，此資料集中每筆資料具有 42 個屬性（attribute, 或欄位），足以描述網路封包所隱含的行為模式，在許多類似的研究中，大部分都是選用此資料集的 10%的資料量（約 49 萬筆）進行實作測試。因此 KDD-Cup'99 資料集可視為客觀的測試資料來源，其實作結果亦具公信力；另分析 42 項屬性中僅有 9 項為標準封包之表頭資訊，其餘則是根據時間（Time-Based）、主機（Host-Based）建立之封包內容萃取後所得資訊，可免去一般封包擷取時可能引發之個人隱私洩漏問題。

KDD-Cup'99 資料集將資料分為五種類別，即 Normal：正常資料；U2R（User to Root）：異常取得管理者權限；R2L（Remote to Local）：遠端攻擊行為；DOS（Denial of Service）：阻斷攻擊；PRB（Probing）：偵測或掃描。除 Normal 外每種資料集都有其附屬之子類別（詳如表 2）：

表 2　選用 10%之 KDD-Cup'99 資料集

類別	子類別（攻擊名稱）	取樣筆數
Normal		95,278（19.3%）
U2R	Buffer_overflow, loadmodule, multihop, perl, rootkit	59（0.01%）
R2L	ftp_write, guess_passwd, imap, phf, spy, warezclient, warezmaster	1,119（0.23%）
DOS	Back, land, Neptune, pod, smurf, teardrop	391,458（79.5%）
PRB	Ipsweep, nmap, portsweep, satan	4,107（0.83%）

（資料來源：本研究整理）

本實驗採用 KDD-Cup'99 資料集所提供之 kddcup.data_10_percent 做為訓練樣本、kddcup.newtestdata_10_percent（corrected）作為測試樣本。為達隨機性，分別將兩個樣本的資料集經由程式亂數隨機選取各五萬筆的資料，並以 training_50k_hbbf_2label.dat 及 test_50k_hbbf_2label.dat 兩個檔案區分訓練及測試樣本，資料集內容如表 3 所示。而隨後將依不同的演算法訓練及測試封包表頭分類器，另為利識別，分別使用支援向量機（SVM）、倒傳遞類神經網路（PNN）、約略集合理論（RSTs）等簡稱。

表 3　入侵偵測資料集內容

類別	檔名	筆數	檔案大小
訓練樣本	training_50k_hbbf_2label.dat	50,000	1.189 KB
測試樣本	test_50k_hbbf_2label.dat	50,000	1.187 KB

（資料來源：本研究整理）

（二）實驗環境

Snort 封包表頭比對採用 rule-based 方法，其為了提升比對效能，亦使用 B-tTree 的方式來重組所有的 rule，並將此 rule tree 整個讀進記憶體中，待封包抓取器取得封包後，逐一與記憶體中的 rule tree 進行比對的工作。本實驗以 Microsoft Visual Studio .NET 2003 開發軟體以模擬封包表頭比對的行為，同樣使用 B-Tree 及記憶體運作的架構。而不同於原始 Snort 行為的，本實驗採用的程式可接受前端 KDD-Cup'99 資料集的輸入，並直接將正確率及執行時間結果輸出，藉由 Snort 封包表頭比對結果的呈現的數據，可以明顯的比較出各種分類器的績效，以利進一步的分析。本研究之實驗針對封包表頭分類器分別選用四種分類器來進行比對時間及正確率之效能評估，整體測試環境及相關軟體參數設定如表 4、5、6、7 及圖 9 所示：

表 4　電腦軟硬體環境（本研究整理）

作業系統	Microsoft Windows XP Professional SP2
中央處理器	AMD K7 3600+
記憶體	DDR-400 1G
硬碟	Hitachi 160G
主機板	Asus A7N8X v2.0
網路卡	Intel 82559 Management
資料庫系統	MySQL 2.6.0-pl2
軟體開發	Microsoft Visual Studio .NET 2003

（資料來源：本研究整理）

表 5　約略集合理論軟體（RSES2）參數設定（本研究整理）

使用軟體	Rough Set Exploration System ver 2.2.1 （RSES2）
參數	Calculate reducts Rules → Genetic algorithm（High speed）
訓練樣本	50000, rses_training_50k_hbbf_2label.dat
測試樣本	50,000, rses_test_50k_hbbf_2label.dat

（資料來源：本研究整理）

表 6　支援向量機軟體（LibSVM）參數設定（本研究整理）

使 用 軟 體	LibSVM 2.8
參　　數	C-SVC, Radial Basis, cost = 1
訓 練 樣 本	50,000, svm_training_50k_hbbf_2label.dat
測 試 樣 本	50,000, svm_test_50k_hbbf_2label.dat

（資料來源：本研究整理）

表 7　類神經網路軟體（QNet）參數設定（本研究整理）

使用軟體	QNet V2K build 721
參數	Layer:4, Input:9, Layer1:5, Layer2:2, Output:1, Sigmoid Max Iterations:10000 Learn Rate Control:10001 AutoSave Rate:500 Screen Update Rate:5 Learn Rate （ETA）:0.007000 Momentum （Alpha）:0.8 FAST-Prop Coefficient:0.000000 Training Patterns used per Weight Update:400 Toerance:0.000000 Quit at Training RMS Error:0.000000 Reset/Initialize Network Weights : True
訓練樣本	50000, nn_training_50k_hbbf_2label.dat
測試樣本	50000, nn_test_50k_hbbf_2label.dat

（資料來源：本研究整理）

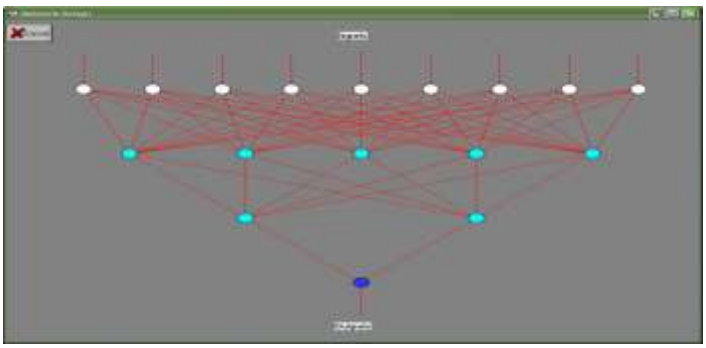

圖 9　QNet 類神經網路圖

（資料來源：QNet）

肆、實驗模擬與結果分析

一、實驗模擬

為加速封包比對機制之運作，本實驗先就 Snort 架構中偵測引擎處理封包程序的第一階段表頭比對部分，利用多種分類演算法分別進行相關的實驗，包括使用倒傳遞類神經網路、約略集合理論及支持向量機等三種演算法訓練封包表頭分類器，整體模擬架構及比對流程如圖 10 所示，實驗的重點置於分類器執行速度及分類正確率之評估，藉以找出最適合處理本研究資料集之封包表頭分類演算法。

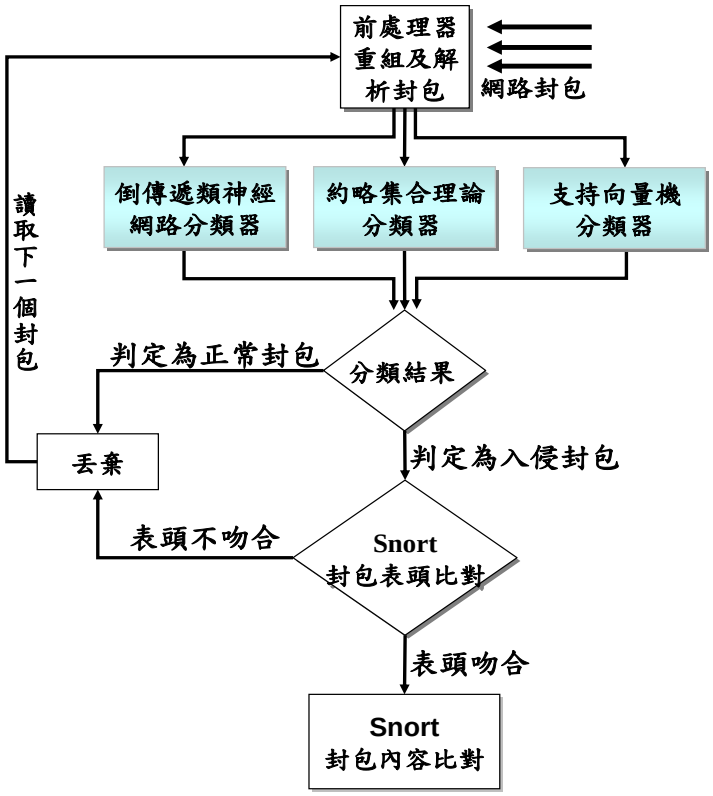

圖 10　分類器模擬架構及流程圖

（資料來源：本研究整理）

二、績效評量作法

本研究致力於降低封包表頭偵測時所耗用的時間，因為利用封包表頭分類器

可以判斷出此封包是屬於正常亦或異常（攻擊），惟其先決條件是分類器的正確率必需維持一定的水準以上，否則無法達成效能改善之目的。本研究所謂之正確率，其計算方式係將 dataset 先經亂數選取訓練及測試樣本各 50,000 筆，經四種演算法建立 model 後將測試樣本送入比對，其分類結果再比對原始封包之正常或異常類別據以計算之。而在圖 7 模擬架構中，當加入封包表頭分類器之後，判定屬於正常之封包將予濾除，而直接將可能屬於異常之封包導入偵測引擎以對封包內容，可以大幅減少封包表頭比對所耗費的時間。因此實驗績效之衡量指標包括（1）偵測時間；（2）偵測正確率兩部分。

三、實驗結果

表頭比對部分主要在測試本研究所提三種分類器與 Snort IDS 所採用以鏈結串列進行比對的 Boyer Moore 分類演算法在效能上的差異，經利用 Visual Studio .NET 2003 撰寫測試程式，以檔案型式直接讀取資料檔，有關 4 種演算法執行時間及正確率如表 8 所示；另表 9 及表 10 分別顯示不使用或使用分類器的情況下，其正常或攻擊不同比重的封包量下，其所耗用的比對時間的差異情形；圖 11 則是以長條圖顯示使用或不使用分類器在不同封包量下的差異情形。

表 8　四種演算法執行時間及正確率分析

實驗結果 / 使用方法	執行時間	正確率	其他資訊
支持向量機	1m25s9	45,097/50,000, 90.194% FP=12.03% ; FN=0.6%	MSE = 0.09806 Squared correlation coefficient = 0.581227
類神經網路	1s2	463,52/50,000, 92.704% FP=8.79%, FN=1.1%	
約略集合理論	2m18s8	46,433/50,000, 92.867% FP=0.181% ; FN=26.2%	Coverage=0.974
Snort 表頭比對	114m11s	39,463/50,000=78.926% FP=0.129% ; FN=13.394%	

說明：（1）FP: false positive; （2）FN: false negative（資料來源：本研究整理）

表 9　不使用封包表頭分類器執行時間

No. of dataset	Attack	Normal	TP	FP	classified packet	func. : EvalOpts	func. : mSearch	Total Time
1	25,000	0	25,000	2,125	1,675	1,450	7,750	13,000
2	22,500	2,500	25,000	2,125	1,675	1,450	7,750	13,000
3	20,000	5,000	25000	2,125	1,675	1,450	7,750	13,000
4	17,500	7,500	25,000	2,125	1,675	1,450	7,750	13,000
5	15,000	10,000	25,000	2,125	1,675	1,450	7,750	13,000
6	12,500	12,500	25,000	2,125	1,675	1,450	7,750	13,000
7	10,000	15,000	25,000	2,125	1,675	1,450	7,750	13,000
8	7,500	17,500	25,000	2,125	1,675	1,450	7,750	13,000
9	5,000	20,000	25,000	2,125	1,675	1,450	7,750	13,000
10	2,500	22,500	25,000	2,125	1,675	1,450	7,750	13,000
11	0	25,000	25,000	2,125	1,675	1,450	7,750	13,000

（資料來源：本研究整理）

表 10　使用封包表頭分類器執行時間

No. of dataset	Attack	Normal	TP	FP	classified packet	func. : EvalOpts	func : mSearch	Total Time
1	25,000	0	24,929.25	0	24,929.25	1,445.8965	7,728.07	9,173.96
2	22,500	2,500	22,436.325	31.275	22467.6	1,303.1208	6,964.96	8,268.08
3	20,000	5,000	19,943.4	62.55	20,005.95	1,160.3451	6,201.84	7,362.19
4	17,500	7,500	17,450.475	93.825	17,544.3	1,017.5694	5,438.73	6,456.30
5	15,000	10,000	14,957.55	125.1	15,082.65	874.7937	4,675.62	5,550.42
6	12,500	12,500	12,464.625	156.375	12,621	732.018	3,912.51	4,644.53
7	10,000	15,000	9,971.7	187.65	10,159.35	589.2423	3,149.40	3,738.64
8	7,500	17,500	7,478.775	218.925	7,697.7	446.4666	2,386.29	2,832.75
9	5,000	20,000	4985.85	250.2	5236.05	303.6909	1623.18	1926.87
10	2,500	22,500	2492.925	281.475	2774.4	160.9152	860.06	1020.98
11	0	25,000	0	312.75	312.75	18.1395	96.95	115.09

（資料來源：本研究整理）

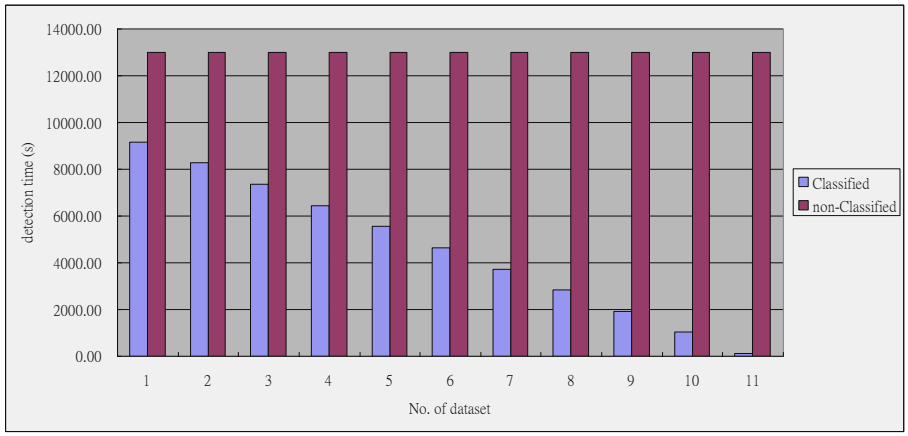

圖 11　使用/不使用封包表頭分類器耗用時間比較

（資料來源：本研究整理）

四、綜合分析

綜合比較以上四種演算法實驗結果，封包比對使用時間最少者為倒傳遞神經網路 1.2 秒，正確率最高者約略集合理論達 92.867%，次高者為倒傳遞類神經網路 92.704%，雖略低於約略集合理論，惟當異常（攻擊）封包量減少時，倒傳遞神經網路所花費的偵測時間遠比約略集合理論少，以本研究實作結果，當異常封包量達最高 2,5000 筆時，其因每筆封包均需比對，故速度僅提高 1.42 倍，然 2,5000 筆封包全為正常封包時，其執行速度可提升約 112.95 倍，兩者在不同封包量下執行速度之差異情形如圖 12 所示。基此，我們可以發現倒傳遞神經網路執行速度最快，而正確率僅於約略集合理論，因此以倒傳遞神經網路具有最佳的封包比對改善效能。此外與 Snort 原始的比對方法相比較，三種分類器的加入均能提升封包比對效能，而分類器演算法本身對未知的攻擊亦具偵測能力，這也是 Snort 採用 rule-Based 的比對方法無法偵測未知攻擊所不及之處。

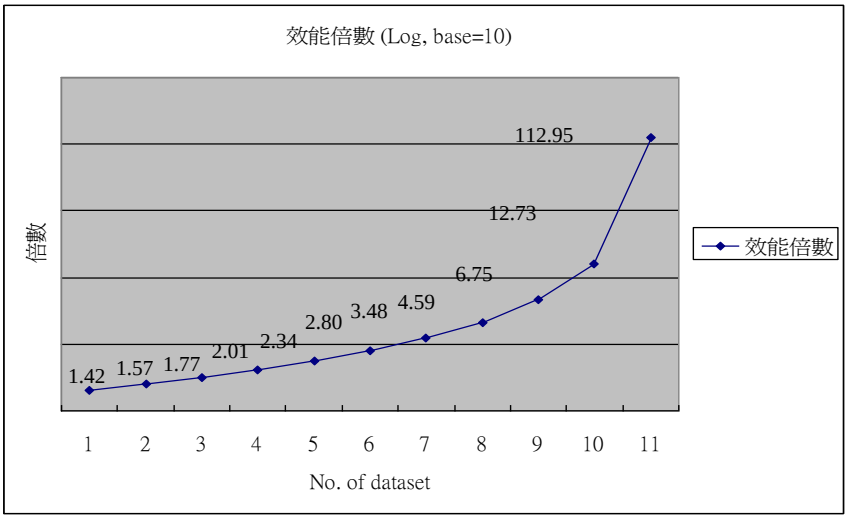

圖 12　封包表頭分類器效能增進倍率曲線圖

（資料來源：本研究整理）

伍、結論與後續研究

資訊化、自動化是現代化國軍追求的目標，但在藉由資訊化來提高國軍指管通資情監偵能力的同時，卻也因資訊科技的導入萌生各種負面的效應，直接或間接地影響執行成效。尤其當前國軍仍以「實體隔離」為最高之資通安全政策，而 IDS 的建置因具備保護、偵測及反應等功能，可以實現風險管控策略，為國軍網路落實邊界保護的重要機制是無庸置疑的。本研究因應行政院 2008 年「資通安全政策白皮書」，在眾多資通安全防護機制中，選擇最重要的 IDS 進行研究，並從其執行效能不佳、正確率低等問題，發現運用資料探勘技術結合表頭分類器確能改善封包比對的效能，而在多種分類器演算法中，又以倒傳遞類神經網路最佳，其正確率 92.704%稍低於約略集合理論 92.867%，但執行速度快約 112.95 倍，同時俱備偵測未知攻擊的能力。

絕對的安全是理想的標竿，而相對的安全才是實現的指標。資料探勘技術結合封包表頭分類器的作法雖可大幅提昇偵測效能，但其正確率僅達 92.704%，因此建議國軍單位採用本研究所提 IDS 加入分類器的作法時，最好能運用「深度防禦」的觀念採一前一後的配置，亦即將加入分類器的 IDS 部署於機敏網路外部防火牆的外側，如此可發揮其快速處理封包的優點，避免網路頻寬增加導致掉封包的問題，而部署於防火牆內側的 IDS 則能更專注於入侵特徵的比對，從而發揮入

侵阻絕的功效，達到符合資通安全防護等級 A 級單位的防護與要求。

入侵偵測技術迄今仍不斷發展，如何因應國軍需求，投入更多的人力研採效能更佳、正確率更高的偵測技術，為未來確保國軍網路安全重要的努力方向。正如國際安全訓練、認證與研究機構 SANS 協會研究總指揮 Alan Paller 所預見，未來資通安全將由制定政策、程式撰寫、教育訓練等軟性安全（soft security）技能，轉變為攻擊行為、入侵偵測、隔離、區段分割等硬性安全（hard security）技術，如果企業組織資訊人力仍以 80%培養軟性安全技能，只有 20%擁有硬性安全技術，而不把兩者配比顛倒過來，將無法找出組織的資通安全問題，更遑論及如何阻絕入侵及防止再生。本研究運用資料探勘技術以提升 IDS 的效能，雖可提供國軍單位解決網路流量日增及正確率低等問題之參考，但資通安全防護內容十分廣泛，各項技術亦不斷推陳出新，未來如何發展硬性安全技術，並運用在 IDS 的核心技術開發上，將使網路防禦裝置更趨完善，同時也為國軍落實資通安全防護提供更多的選項，有效達成國軍網路安全無虞之目標。

參考文獻

行政院科技顧問組。2008。《2008 資通安全政策白皮書》。台北：行政院科技顧問組。

李駿偉、田筱榮、黃世昆。2002。〈入侵偵測分析方法評估與比較〉。《資訊安全通訊》。第八卷第二期（16）：21-37。

吳文進。2004。〈利用排除的觀念改善入侵偵測特徵比對效能之研究〉。華梵大學資訊管理學系。碩士論文。

吳蔓玲譯。Bruce Schneier 著。2001。《祕密與謊言—如何建構網路安全防護系統》。台北：商周出版。

黃承龍、唐文政。2003。〈應用約略集合於醫學與信用卡資料之分類〉。「第九屆資訊管理暨實務研討會」論文。台北：中原大學。

陳正昌譯。Northcutt, Judy Novak 著。2002。《網路入侵偵測教戰手冊》。台北：培生教育出版。

葉怡成。2003。《類神經網路模式應用與實作》，第八版。台北：儒林圖書公司。

賴冠州編譯。Rebecca Gurley Bace 著。2001。《駭客入侵偵測專業手冊》。台北：旗標出版社。

Al-Subaie M., Zulkernine M. 2006 .“Efficacy of Hidden Markov Models Over Neural Networks in Anomaly Intrusion Detection.” Computer Software and Applications Conference, COMPSAC '06, 30th Annual International Volume 1. 8:325-332.

Anagnostopoulos T., Anagnostopoulos C., and Hadjiefthymiades S. 2005.“Enabling attack behavior prediction in ubiquitous environments.” Pervasive Services, CPS '05 Proceedings, International Conference. 4:425-428.

Burroughs D.J., Wilson L.F., and Cybenko G.V. 2002 .“Analysis of distributed intrusion detection systems using Bayesian methods” 21st IEEE International. 6:329-334.

Bonifacio J. M. et al. 1998. “neural networks applied in intrusion detection system”IEEE. 6:205-210.

Bo Gao, Hui-Ye Ma, and Yu-Hang Yang. 2002. “HMMs （Hidden Markov models）based on anomaly intrusion detection method” Machine Learning and Cybernetics, Proceedings, 2002 International Conference on Volume 1. 5:381-385.

C. Kruegel, F. Valeur, G. Vigna, and R. A. Kemmerer. 2002. “Stateful intrusion

detection for high-speed networks.” IEEE Symposium on Security and Privacy. 10:285-294.

Cristianini N., Shawf-Taylor J. 2000. “ An Introduction to Support Vector Machines and other kernel-based learning methods.” Cambridge University Press.

CERT Statistics. http://www.cert.org/stats/. Latest updatw 2007/10/8.

D. Bolzoni, S. Etalle, P. Hartel. 2006. “POSEIDON: a 2-tier anomaly-based network intrusion detection system.”IEEE, Information Assurance. 10: 1-10.

Depren M. O., Topallar M., Anarim E., Ciliz K. 2004. “Network-based anomaly intrusion detection system using SOMs.”Proceedings of the IEEE 12th. 4:76-79.

Faour A., Leray P., and Eter B. 2006. “A SOM and Bayesian Network Architecture for Alert Filtering in Network Intrusion Detection Systems.”Information and Communication Technologies, ICTTA '06, Volume 2. 6:3175-3180.

Fayyad, U. M., Irani, K.B. 1993.“Multi-interval discretization of continuous-valued attributes for classification learning.” In Proceeding. of the 13th International Joint Conference on Artificial Intelligence. 6:1002-1007.

Fei Gao, Jizhou Sun, and Zunce Wei. 2003. “The prediction role of hidden Markov model in intrusion detection” Electrical and Computer Engineering, IEEE CCECE, Canadian Conference on Volume 2. 4:893-896.

Guan Jian, Liu Da-Xin, and Cui Bin-Ge. 2004. “An induction learning approach for building intrusion detection models using genetic algorithms.”Intelligent Control and Automation, Fifth World Congress on Volume 5. 4:4339-4342.

Hegazy I. M. et al. 2005.“ Ahmed T.,Evaluating how well agent-based IDS perform.”IEEE, Volume 24. 4:27-30.

I.Charitakis, K.Anagnostakis, and E.Markatos. 2003. “An active traffic splitter architecture for intrusion detection”Proceedings of 11th IEEE/ACM International Symposium on Modeling “Analysis and Simulation of Computer and Telecommunication Systems（MASCOTS 2003）.”4:238–241.

Jay Beale et al. 2003. Snort 2.0 Intrusion Detection. Canada: Syngress.

Jiawei Han and Micheline Kamber. 2001. Date Mining: Concepts and Fechniques, San Francisoc: Morgan Kaufmann.

Joseph S. sheriff, Rod Ayers, “Intrusion detection: Methods and system. Part II”, Information Management and computer security, P222-229, 2003。

Kruegel C. et al. 2003. “Bayesian event classification for intrusion detection.” Proceedings for 19th Annual “Computer Security Applications Conference.”

10:14-23.

Kussul N. et al. 2003.“Intelligent multi-agent information security system.” Proceedings of the Second IEEE International Workshop. 3:120-122.

Lei J.Z., Ghorbani A. 2004. “Network intrusion detection using an improved competitive learning neural network.”Communication Networks and Services Research Conference. 8:190-197.

Lina Wang et al. 2001. “Method of evolutionary neural network-based intrusion detection, Info-tech and Info-net.”2001 International Conferences on Volume 5. 6:13-18.

Ming-Guang Ouyang, Wei-Nong Wang, Yun-Tao Zhang. 2002. “A fuzzy comprehensive evaluation based distributed intrusion detection. “Proceedings for Machine Learning and Cybernetics.” 2002 International Conference on Volume 1. 5:281-284.

Mill J., Inoue A. 2004. “Support vector classifiers and network intrusion detection.” 2004 IEEE International Conference on Proceedings, Volume 1. 4:407-410.

Orfila, A., Carbo, J., Ribagorda. 2003. “A Fuzzy logic on decision model for IDS, Fuzzy Systems.” The 12th IEEE International Conference on Volume 2. 6:1237-1242.

Paul E. Proctor, “The Practical Intrusion Detection Handbook”, Prentice Hall, 2000.

Pawlak, Z., and Slowinski, R. 1994. “Rough Set Approach to Multiattribute Decision Analysis.” European Journal of Operational Research, Vol. 72. 17:443-459.

Pawlak, Z., Rough. 1991. “Sets: Theoretical Aspects of Reasoning About Data.”Boston: Kluwer Academic Publishers.

P. Gupta, N. McKeown. 2001. “Algorithms for packet classification, Network.”IEEE, Volume 15, Issue 2. 9:24-32.

Rebecca Bace, Peter Mell, “intrusion detection system”, NIST Special Publication on Intrusion Detection Systems, 2003.

Richard P. Lippmann, Robert K. Cunningham. 2000. “Improving intrusion detection performance using keyword selection and neural networks.”Elsevier, computer network. 7:597-603.

Ryan Jake, Meng-Jang Lin. 1998. “Intrusion Detection with Neural Networks.” Advences in Neural information processing system 10. MIT press.

Sang-Jun Han, Sung-Bae Cho. 2003. “Rule-based integration of multiple measure-models for effective intrusion detection Systems.”IEEE International

Conference on Volume 1. 6:120-125.

SANS Institute. 2001. “Application of Neural Networks to Intrusion Detection.”http://www.sans.org/reading_room/whitepapers/detection/336.php. Latest update 2008/3/10.

Vapnik V. 1998. Statistical Learning Theory. Wiley,.

W. Lee, S. J. Stolfo, and K. W. Mok. 1999.“A data mining framework for building intrusion detection models.” Proceedings of the 1999 IEEE Symposium “Security and Privacy.”13:120-132.

Wei Lu, Traore I. 2003. “Detecting new forms of network intrusion using genetic programming, Evolutionary Computation.” The 2003 Congress on Volume 3. 8:2165-2172.

Xueqin Zhang, Chunhua Gu, and Jiajun Lin. 2006. “Support Vector Machines for Anomaly Detection” Intelligent Control and Automation, WCICA 2006, The Sixth World Congress on Volume 1. 6:2594-2598.

Yang Xiang-Rong, Song Qin-Bao, and Shen Jun-Yi. 2001. “Implementation of sequence patterns mining in network intrusion detection system.”Info-tech and Info-net Proceedings, International Conferences on Volume 5. 5:19-23.

Yufeng Kou et al. 2004. “Survey of fraud detection techniques.”IEEE International Conference, Volume 2. 6:749-754.

Zonghua Zhang, Hong Shen. 2004. ”Online training of SVMs for real-time intrusion detection.”AINA 2004, 18th International Conference on Volume 1. 6:568-573.

Zhoujun Xu, Jizhou Sun, and Wenjie Li. 2004. “Intrusion detection using fuzzy window Markov model,Electrical and Computer Engineering, Canadian Conference onVolume 2, , 645 – 648.

（投稿日期：97年4月17日；採用日期：97年7月23日）

附錄　KDD-Cup'99 封包分類及屬性列

資訊分類	KDD 欄位名稱	資料型態	註解
Basic Features	duration	Continuous	連線時間（秒數）
	protocol_type	Discrete	通訊協定種類（tcp、udp 等）
	service	Discrete	服務種類（http、telnet 等）
	flag	Discrete	連線狀態正常與否
	src_bytes	Continuous	來源端傳送的資料量（byte）
	dst_bytes	Continuous	目的端傳送的資料量（byte）
	land	Discrete	判斷來源和目的之主機位置/埠號是否相同（1 同;0 異）
	wrong_fragment	Continuous	分割錯誤之封包數量
	urgent	Continuous	具"緊急位元"封包之數量
Additional Features / Content Features	hot	Continuous	進入系統目錄後建立或執行檔案之次數
	num_failed_logins	Continuous	登入失敗之次數
	logged_in	Discrete	1 登入成功;0 登入失敗
	num_compromised	Continuous	不存之檔案或路徑被開啟、及非循正常路徑索取資料之次數
	root_shell	Discrete	1 取得管理者權限;0 其他權限
	su_attempted	Discrete	1 試圖取得管理者權限;0 無
	num_root	Continuous	管理者權限之使用者連線次數
	num_file_creations	Continuous	建立檔案之次數
	num_shells	Continuous	操作環境（shell）之使用數量
	num_access_files	Continuous	操作權限控制用檔案之次數
	num_outbound_cmds	Continuous	於 FTP 連線時對外執行其他命令之次數
	is_hot_login	Discrete	1 屬管理等級之登入;0 非管理等級登入
	is_guest_login	Discrete	1 訪客或匿名性質之登入;0 其他登入
Time Based Features	dst_host_same_srv_rate	Continuous	
	dst_host_diff_srv_rate	Continuous	
	dst_host_same_src_port_rate	Continuous	
	dst_host_srv_diff_host_rate	Continuous	
	dst_host_serror_rate	Continuous	
	dst_host_srv_serror_rate	Continuous	
	dst_host_rerror_rate	Continuous	
	dst_host_srv_rerror_rate	Continuous	
Traffic Features（2-second time window）	count	Continuous	本機端接受之連線數量
	serror_rate	Continuous	SYN 錯誤之發生率
	rerror_rate	Continuous	REJ 錯誤之發生率
	same_srv_rate	Continuous	同一服務之連線率
	diff_srv_rate	Continuous	不同服務之連線率
	srv_count	Continuous	單一服務接受之連線數量
	srv_serror_rate	Continuous	SYN 錯誤之發生率
	srv_rerror_rate	Continuous	REJ 錯誤之發生率
	srv_diff_host_rate	Continuous	不同主機之連線率