模糊關聯與情境法則探勘於入侵偵測

曹偉駿*　柯文元　林明孝
大葉大學資訊管理系
*E-mail: wjtsaur@yahoo.com.tw

摘要

本論文針對入侵偵測，提出基於模糊關聯法則與情境法則之探勘技術。首先以模糊群集技術將網路封包分群，產出正、異常群集供法則探勘使用，接著以模糊關聯法則探勘技術於各群集中找出其間的關聯，以挖掘出可能的關聯法則，從而找出單一攻擊事件。此外，本論文亦將使用模糊情境法則演算法，於各群集中找出多重序列間的相互關係，以發掘出攻擊事件的組成及發生次序，並將上述這些法則分別建構於異常及正常的法則資料庫中。綜合上述，本論文所提出之機制主要貢獻在於以加入模糊資料探勘技術，使得能更精確地偵測出單一攻擊或攻擊組成及發生次序，進而提高偵測率。本論文亦實際開發一套系統，以驗證提出之機制的成效。

關鍵字：入侵偵測系統、模糊理論、群集技術、關聯法則、情境法則

壹、緒論

近年來，隨著網際網路的日漸普及，網路應用的快速發展，開發出許多網路相關的應用與服務，使得人們能享受到網路帶來的便利性及其相關效益，網路儼然成為現代人日常生活不可或缺的一部份。然而，在這些效益背後所隱藏的安全問題是急需我們正視及解決的，這些問題如病毒的威脅、非授權的存取、阻絕服務攻擊(DOS Attack)或是其他形式的入侵。不論公司、政府部門，亦或是個人都遭受到這些安全性問題的嚴重威脅。

在 2002 年份的 CSI/FBI Computer Crime and Security Survey [4]調查中顯示出 98%回覆調查的企業有建置網站，其中 52%已透過網站開始進行電子商務交易；調查中的有 38%的網站在過去的十二個月曾遭受非法入侵，而遭受入侵者，內部資料約有 70%被破壞；在各式的電腦犯罪下，於 2002 年部分損失就高達四億美金；而 2003 年初剛開始，各地紛紛更是傳出不明原因的網路攻擊事件，導致各區網路一度因此而癱瘓或無法提供正常運作，究其原因是 SQL Slammer 蠕蟲病毒所造成的，SQL Slammer 感染了 20 萬台沒有安裝修補程式的微軟 SQL Server。該蠕蟲估計在病發十分鐘內便感染 90%有漏洞的伺服器，損失相當嚴重。在賽門鐵克公司網際網路安全趨勢報告[11]中指出，在 2002 年到 2003 年間逐漸增加的漏洞，比起 2001 年與 2002 年的 60%有大幅的增加，主要是因為公開發佈的攻擊程式(Exploit)有近五成的成長，而在這些有攻擊程式的漏洞當中，又有絕大部分的漏洞是屬於高度與中度嚴重性。因此，在面臨日漸增加的各種不同的駭客入侵方式，資訊安全的挑戰也日益加劇。

在傳統的資訊系統安全中，多仰賴防火牆的封包過濾、認證功能或是加、解密密碼系統來保護資訊資產，但在遭遇現今日益複雜的駭客攻擊手法下，仍無法杜絕這些攻擊事件的發生，因此入侵偵測系統(Intrusion Detection System)無疑的成為不可或缺的一道防線，以彌補防火牆無法解決的一些問題。一般來說，入侵偵測系統可以視為是一種監控工具，利用解讀網路封包或系統記錄檔的方式偵測可疑之入侵動作，判斷系統所遭受的入侵程度及損害程度，並據以發出警訊，使網路管理人員可依照警訊做及時的反應或修護工作。然而，目前大多數的入侵偵測系統仍存在產生過多錯誤警報之問題(如表 1)，即是常會

發生誤報率(False alarm rate)過高的問題，而這些因誤報產生的警訊(Alert)會使得系統管理人員疲於奔命；若是誤判率(False positive)過高則容易讓駭客入侵成功，影響各系統及網路之安全。因此，如何降低入侵偵測系統的錯誤率來提高防範駭客攻擊的防禦能力，便是急需探討的主題。

隨著網路使用量越來越高，在系統紀錄檔中所記錄的資料量也相對變多。資料庫中雖然有很多的資料，但是我們無法從表面上看出其隱藏的資訊，更無法以人力去分析。雖然簡單的統計方法、電腦報表及資料庫查詢工具可以用來幫助我們分析資料，但是它並不像新的智慧型分析工具功能如此的強大，可以快速且自動找出隱藏及有用的資料。因此，使用何種方式能正確且有效率的從各種資料庫中，找出使用者的行為樣式，也是入侵偵測研究的一個方向。

表 1 IDS 警訊之分類

	True	False
Positive	有入侵行為且有發出警訊	無入侵行為但卻發出警訊
Negative	無入侵行為且沒發出警訊	有入侵行為但沒發出警訊

目前入侵偵測系統仍存在之問題如下:

- 誤用偵測之入侵偵測系統以特徵比對方式偵測入侵，一般來說較能準確發現攻擊，卻無法偵測新型態之攻擊手法；相較之下，以異常偵測技術之入侵偵測系統雖能發現新型態之攻擊，但常會產生錯誤警報。
- 入侵偵測系統會產生大量警訊，但不少是錯誤警訊，會使得系統管理人員疲於奔命，而產生不信賴感。

基於上述，因此本研究目的如下：

一、 利用模糊群集技術建立正、異常使用者行為樣式

二、設計模糊關聯及情境法則探勘機制，提高入侵偵測之偵測率

貳、文獻回顧

本節主要介紹入侵偵測系統、模糊理論的概念，探討資料探勘方法，包含了群集技術、關聯法則、情境法則及整合模糊理論之相關演算法。

一、入侵偵測系統

一般來說，入侵偵測系統包含四大元件，即資料來源、資料前處理器、偵測分析引擎、回應機制(如圖 2-1)。依入侵偵測系統型態的不同資料來源可以分為系統記錄檔(Log Files)及網路封包兩種;資料前處理器將收集到的資料轉換為可供偵測分析引擎使用的格式化數據；根據格式化的資料，比對正常或異常特徵庫，若相同便是偵測出入侵跡象或是一般正常行為；以偵測分析引擎的分析結果，若為可疑攻擊行為,則產生回應給予管理人員做相關處理。

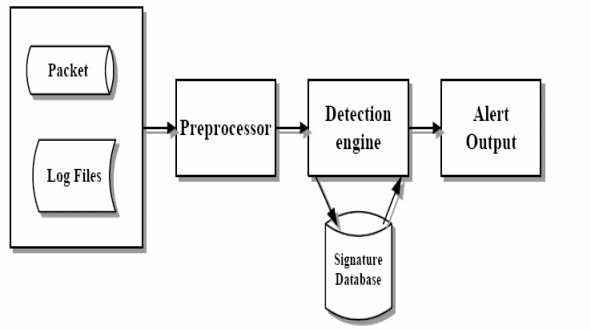

圖 2-1、入侵偵測元件

根據 Axelsson [1]的研究指出,入侵偵測系統可以依據資料來源(Audit Source Location)、偵測方法(Detection Method)、回應偵測行為(Behavior on Detection)及使用頻率(Usage Frequency)來加以分類，其詳細說明可參閱 Axelsson [1]的技術報告。

二、模糊理論

模糊理論(Fuzzy)是由美國加州柏克萊大學教授 Zadeh [14]教授於 1965 年第一次所提出來的，其主要目的在於使自然語言中的數量與推理名詞在數學或計算上具有模糊不清的意思，並嘗試以人類的思維方式去簡化問題的複雜度。在本研究中所需的網路資料記錄檔包含了許多數值的特徵，如來源 IP、時間性的變數等，若將模糊理論應用於資料探勘中的資料前處理步驟中，將可有著極大的幫助，對於駭客入侵行為的分析，所建立出的資料將會更準確。

模糊理論應用於資料探勘領域中不是新的概念，然而在資料探勘的發展過程中，不論是商業市場或學術領域裡都沒有一套有用的模糊資料探勘系統，這個現象直到 1998 年 Rubin 所建構出的系統才真正產生了完整的架構。在此系統中，提出的方法是利用模糊理論的概念進行法則的萃取與推論，並且將此結果應用在決策系統裡。在此決策系統裡，他使用模糊邏輯去解決隨機數值資料的問題。經由實驗的結果顯示，此一機制所萃取出來的法則對於一般人來說，容易理解且有很高的可讀性，其結論表示了模糊理論應用於資料探勘可以提供另一種形式去描述一般的資料。

在分析各個使用者行為時，本研究導入了模糊理論的觀念，以更正確的判定駭客入侵之行為。綜觀市面上有許多 IDS，其中最為人所詬病的是誤報率及誤判率，誤報過多會造成假警報的出現太過頻繁，而誤判過多則會把異常的使用行為判定為正常的。因此，若將模糊理論應用於網路使用者行為之分類，不但可以很容易的結合各種輸入值，還可準確的定義出許多入侵型態。在接下來的章節中，將更詳細探討一般資料探勘及模糊資料探勘應用於入侵偵測的文獻。

三、資料探勘

資料探勘（Data mining）是近年來被廣為利用的一項技術，其主要目標是要從龐大資料庫中，挖掘出對使用者來說有用或隱藏的資訊。而另外也有學者認為它是屬於 KDD (Knowledge Discovery in Database)的一部分。整個知識發掘的流程步驟如下[15]：先理解資料與所要進行的工作，並取得相關知識、技術及資料；再將原始資料做前置處理作業，除去錯誤或不一致的資料；對所想要的結果建立模式與假設；進行實際的資料探勘工作；對所發掘的知識進行評估與驗證，解釋該知識的可行性並使用。這些程序是一個循環的關係，一直重複的步驟，最後才得到一些有用的知識。從這些步驟中我們可以發現到，其實資料探勘在整個知識挖掘的過程中，只是其中的一小部分，大部分的時間是花在準備資料的前置動作，因為資料的來源不同，則資料的格式、資料型態、欄位長度等亦會隨之不同，所以資料轉換的動作也許會花掉整個過程的大量時間。而目前資料探勘的技術應用已非常普遍，除了在下述幾個資料探勘模式中所提的應用外，凡是只要有涉及到大量的資料處理，幾乎都會運用資料探勘技術。

因此，由於入侵偵測之資料來源的資料量，常常是相當大的資料量，因此便可透過資料探勘挖掘一致性、有用的系統特徵樣式，找尋可能隱藏的行為模式，並分析是否存在攻擊行為，將這些攻擊行為利用知識工程的技術擷取出來，配合使用者行為模式的分析判斷攻擊行為形式的可能性。在 Lee [17]研究中說明三種資料探勘技術最適用於挖掘主機型入侵偵測系統之稽核資料：

(1)分類(Classification)：

將資料項目映射到一些被預先定義的種類之內。舉例來說，這些運算法則以決策樹的形式或分類規則產出類別。應用到入侵偵測系統主要是要將『正常』和『不正常』的行為充份的聚集。也就是應用分類演算法規則對稽核資料決定當屬於正常類別或不正常類別。

(2)連接分析(Link analysis)：

判定資料欄與資料庫之間的關係，找出稽核資料的相互關係，將提供入侵偵測系統選擇系統特徵的洞察力。

(3)序列分析(Sequence analysis)：

也就是序列型樣模型化，這些運算法則能幫助我們了解稽核事件發生的頻率序列。這些頻率事件型樣是使用者或計畫的行為表示的重要元素。

以下將介紹本論文中所使用的資料探勘技術。

- 群集技術(Clustering Technology)

群集技術在資料探勘領域中，是一項非常重要的技術，它可以在大量的資料中，找出資料的分布狀況並找到其隱藏的意義，例如當使用者面臨要分析處理龐大的資料時，往往無法輕易的獲知這些資料所代表的意義，而利用群集技術可以先將這些資料分成若干個群集，再針對不同的群集加以分析。如此，便可以簡化使用者分析資料時的複雜性。

群集技術的目的是在分析資料的內容，將性質相似的資料歸類在一起，讓同一個群集中的資料相似性大，而不同的群集與群集間資料的相似性卻很小。群集技術與傳統分類(Classification)最大的不同是，群集技術不預先設定分類所代表的意義，而把資料先以群集技術將性質內容相近的資料聚集成一群群的群集後，再分析並定義各群集的意義；而傳統的分類是先將每個群集的的意義定義出來後，再將資料依特性歸類到某一個群集。

目前在模糊群集技術的演算法中廣泛運用的是 Bezdek [2] 提出之 Fuzzy C-Means (以下簡稱 FCM)。「模糊分群矩陣」的含義為：模糊群集可根據模糊分群矩陣，來判定資料屬於哪一群組。模糊分群矩陣 U，是一個 $c \times n$ 的矩陣，其中 c 代表群組的數目，n 是分群資料的總數，而矩陣中的成員 U_{ik} 代表資料點 k 在群集 i 中的隸屬函數值。而 FCM 模糊分群矩陣必須滿足下列性質：

$$\begin{aligned} &1.\ U_{ik} \in [0,1] && 1 \le i \le c,\ 1 \le k \le n \\ &2.\ \sum_{i=1}^{c} U_{ik} = 1 && 1 \le k \le n \\ &3.\ 0 < \sum_{k=1}^{n} U_{ik} < n && 1 \le i \le c \end{aligned} \qquad (2.1)$$

例如：設 $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ 共 7 個資料，將其分成 3 個群組 $C = \{C_1, C_2, C_3\}$，則可能的 FCM 模糊分群矩陣如下：

$$U = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array} \\ \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.5 & 0.6 & 0.7 & 0.8 \\ 0.0 & 0.2 & 0.8 & 0.2 & 0.3 & 0.2 & 0.2 \\ 0.7 & 0.4 & 0.0 & 0.3 & 0.1 & 0.1 & 0.0 \end{bmatrix} \end{array}$$

而在分群的過程中，模糊分群矩陣中的每個成員 U_{ik} 的值會隨著參數的變化，重複計算出 U_{ik} 的新值，直到 U_{ik} 不再變動為止。

【FCM 演算法】

假設 $X = \{x_1, x_2, \ldots, x_k, \ldots, x_n\}$ 為 n 個資料物件的集合，V 代表所有 c 個群組中心點之集合，$V = \{v_1, v_2, \ldots, v_i, \ldots, v_c\}$ 其中 $2 \le c \le n$。模糊分群矩陣 $U = (U_{ik})_{n \times c}$，$d_{ik} = \| v_i - x_k \|$ 代表資料點 k 到本身群集中心點 i 之距離，$d_{jk} = \| v_j - x_k \|$ 代表資料點 k 到其它群集中心點 j 之距離。而 m 稱為模糊程度，當 m 值越大，代表資料的模糊程度越大。FCM 的評估函數如下：

$$\min J(U,V) = \min \sum_{k=1}^{n} \sum_{i=1}^{n} (u_{ik})^m (d_{ik})^2 \qquad (2.2)$$

FCM 評估函數的作用在於尋求最佳的群集中心點與群集成員，而要達到此目標就是必須要使群集函數能達到最小值。整個步驟如下：

步驟 1：決定分群個數 c 與模糊程度 m 之後，任意選定模糊分群矩陣 $U^L=(U_{ik})_{n\times c}$，L 代表經過幾次的運算，先令 L=0。

步驟 2：由模糊分群矩陣 $U^L=(U_{ik})_{n\times c}$，透過公式(2.4)計算群集中心 $V^{(L)}$。

步驟 3：根據 $V^{(L)}$且令 $L=L+1$，透過公式(2.3)將 U^L 更新為 U^{L+1}。

步驟 4：定一個合適的門檻值 ε 來比較 $U^{(L)}$ 和 $U^{(L+1)}$，若滿足條件

$$\| U^{(L+1)} - U^{(L)} \| \le \varepsilon,$$

則停止運算，否則重複第二步驟。

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \qquad (2.3)$$

$$V_i = \frac{\sum_{k=1}^{n}\left[(u_{ik})^m x_k\right]}{\sum_{k=1}^{n}(u_{ik})^m} \quad i = 1, 2, 3, ..., c \qquad (2.4)$$

在上述演算法中 ε 介於 0~1，而 c 與 m 必須由使用者依據經驗自行決定的，目前尚無理論依據可循，但大部分 m 值都取 2。

在模糊群集技術應用於入侵偵測方面，Dickerson [3]等人以 FCM 為基礎，建構一個在網路式環境 下異常偵測的模糊入侵辨視引擎(Fuzzy Intrusion Recognition Engine)，主要目的是要偵測出 DoS 及主機與連接埠掃描的攻擊。在 Tsaur [12]等人發表的論文中，則是以 FCM 為基礎發展之 mrFCM (Multistage Random Fuzzy C-Means)，應用於網站記錄檔，從中分析出正常及異常的群集。

- 關聯法則(Association Rules)

在入侵偵測方法之中，除了利用群集演算法分出正、異常行為外，另外也有許多的研究[16]是採用關聯法則演算法，分析使用者行為，在各種行為特徵中找出其間的關聯，以發掘出可疑事件的關聯性。如Lee [16]等學者利用關聯法則演算法從記錄系統各個特徵的系統記錄檔中發掘出有用的樣式，這些樣式包括了許多程式和使用者的各項行為，並計算這一系列相關的系統特徵以產生分類器(Classifier)，而這分類器所存的資料就是關聯法則庫，其主要目的是辨別出各種異常和已知之入侵行為。Hossain [6]學者是利用關聯法則的方式，從資料庫中找出多個特徵值之間的關聯性，之後將關聯法則建構成決策樹型式的分類器，每個定義好的分類器就是使用者的行為模式，用以判定是正常或異常的行為。

接下來即介紹關聯法則的觀念，關聯法則主要概念是在於尋找資料庫中項目或屬性之間共同發生的關係，比如可以藉由關聯法則找出賣場中某些商品所具有共同被購買的關係，像80%購買牛奶的人,也會同時購買麵包[19]，當決策者取得這些資訊後，就可以考慮針對這兩項商品的銷售做改良，也許是將牛奶和麵包的展售櫃檯擺在一起，或是針對麵包和牛奶搭配做促銷活動。所以利用關聯法則挖掘的目的，是希望能從資料相對發生次數的分析著手，找出未知的關係，作為決策時的參考。

因此，Tsaur [13]等人提出改進的 FGBRMA (Fuzzy Grids Based Rules Mining Algorithm)演算法，是利用Boolean (AND、OR及XOR)的方式獲得法則。此演算法只需要對資料庫掃瞄一次，以降低系統從磁碟讀取資料的時間，但此演算法原先在利用XOR運算時，會產生一些無效的法則，如 $X_{lk1} \Rightarrow X_{2k1} * X_{3k1}$。再經由Tsaur等人改進後，便減少產生無效的法則。FGBRMA 演算法流程如圖2-2所示，另外，在此演算法中，會用到的符號如表2.1所示，*FG*所表示為每個欄位代表的是語意值(Linguistic Value)，每個列代表的是k-維

的模糊項目；在*TT*中每個欄位代表的是每一筆紀錄對於每一個k-維模糊項目的隸屬程度。在*FS*中儲存的是每個k-維模糊項目的模糊支持度。X_{1k1}表示在X_1屬性中的k1區間，如A_H就代表在A屬性中的High區間。

表 2.1 FGBRMA 符號定義

符號	定義
FG	包含所有模糊項目與每個語意變數間的Boolean值
TT	包含所有紀錄對於每個k-維模糊項目的隸屬度
FS	包含所有k-維模糊項目的模糊支持度
X_i	第i 個模糊項目
X_{ikj}	第i 個模糊項目中的第j 個區間
t_n	所有資料中的第n 筆紀錄
u_{ij}	對於第*i* 個模糊項目中，第*j* 個區間的隸屬度
f_s	每個k-維模糊項目的模糊支持度
supp	使用者定義的最小支持度
conf	使用者定義的最小信賴度
FC(R)	計算後法則R之隸屬度

- 情境法則(Episode Rules)

為了要找出更好的法則，資料探勘應用於入侵偵測領域在 1997 年由 Mannila 等人 [9] 發展出頻繁情境法則 (Frequent Episode Rules)，頻繁出現的情境可能來自於稽核網際網路的紀錄，一段情境就是表示一個網路的連接序列事件，而在網路傳輸的異常行為紀錄中去發掘一個情境(Episode)，目的是要發現行為事件之間的關係，這些關係能用在行為資訊流的分析，可降低多餘的行為分析，並預知事件的發生，也就是透過情境法則可以了解事件與事件之間的關聯性。

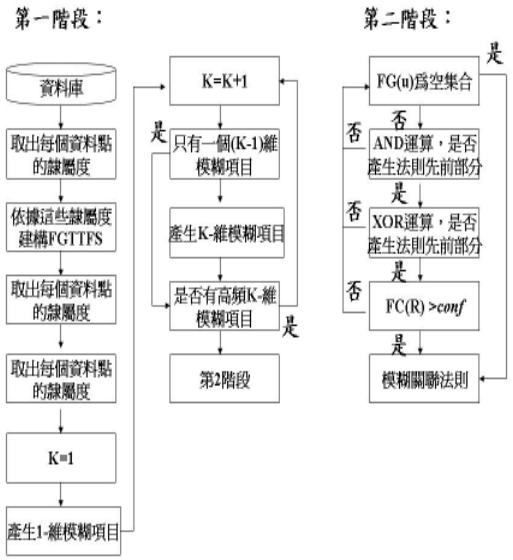

圖 2-2、模糊關聯法則-FGBRMA 流程

大多數的資料探勘和機器學習的技術，都趨向無序性資料收集的分析，然而，有些重要的應用領域中，所分析的資料是需要事件的序列所組成[20]，如網路傳輸中的異常行為、現實社會上的犯罪行為、醫療上疾病的復發情況、氣象中的降雨情況等等。以下便以簡單例子說明之如圖2-3，這裡A、B、C、D、E和F可能是一個網路傳輸的異常行為事件類同在一條時間軸線上。由這事件序列得知，在一段時間內事件E發生後總是接著發生事件F，另一現象，當事件A或B發生，緊接著會出現事件C。

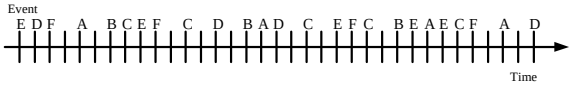

圖2-3、事件的序列

入侵行為資料流可認為是一種時間序列。情境法則便可從行為資料流的資料庫中搜索出所有支援度大於最小支持度的情

節，並輸出信賴度大於最小信賴度的規則。下面定義演算法的相關概念：

一、 事件(Event)：事件(A, t)，表示某個時間點 t 發生了類型是 A 的事件。

二、 事件類型(Type)：事件序列(E)中的事件類別集合。

三、 事件序列(Event Sequence)：對於一個滿足事件類型 E 範圍的事件序列 $s=(s, T_s, T_e)=\langle (A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n)\rangle$，其中 $A_i \in E(i=1, \ldots, n)$；$t_i \leqq t_i+1(i=1, \ldots, n-1)$，$t_i$ 表示事件發生的時間；T_s 表示事件序列開始時間，T_e 表示事件序列結束時間，T_s，T_e 都是整數，且 $T_e \leqq t_i < T_e$，$(i = 1, \ldots, n)$。

四、 情境(Episode)：情境就是若干事件類型的組合。L 個事件組成的情境，其長度是 L。

五、 視窗(Window)：視窗即相關事件的最大時間間隔。入侵行為是一時間序列的資料流，在考察事件關聯性時還要充分考慮事件之間的時間間隔，因此必須定義一個滑動視窗，在資料庫搜索的過程中，這個視窗從第一條記錄滑向最後一條記錄，只有在視窗裡的事件才被觀察。對於一個滿足事件類型 E 範圍的事件序列 $s = (s, T_s, T_e)$，有一個視窗 $w = (w, t_s, t_e)$，其中 t_s 表示視窗事件序列開始時間，t_e 表示視窗事件序列結束時間，且 $t_s < T_e$, $t_e > T_s$，w 是由介於 t_s，t_e 之間的事件所構成，(A, t) S, $t_s \leqq t < t_e$。

六、 視窗寬度(Window Width)：對一個視窗 $w = (w, t_s, t_e)$，則視窗寬度為 $win = t_e-t_s$，而 w(s, win)表示在事件序列 s 中所有長度為 win 的視窗事件序列。

Mannila 所提之情境法則基本上是個不斷反覆搜索的過程。即從最簡單的情境開始，即單個事件(單個入侵行為)的情境，組成初始候選集合 C_1，通過對資料庫的搜索，計算出每個情境發生的頻率，從而找出 C_1 中頻繁發生的情節集合 C'_1，然後一方面通過與其子集的頻率比較輸出規則，另一方面通過把 C'_1 的情境兩兩結合組成候選集 C_2，再重新搜索資料庫，找出 C_2 中頻繁發生的情境集合 C'_2，如此不斷反覆，直到再也找不到頻繁集合為止。在找出頻繁發生的情境集合，也就是頻繁集合，是針對一個頻率最低門限值 *min_fr* (即最小支持度公式 2.5)來比較的，如果 $fr(a,s,win) \geq min\_fr$，α 就是頻繁的。這裏把對於事件序列 *s*，視窗寬度 *win* 和最小支持度 *min_fr* 的頻繁集寫作 $F(s, win, \min\_ fr)$，找出頻繁集合，並輸出在事件序列 *s* 中，視窗寬度為 *win*、最小支持度 *min_fr* 和最小信賴度 *min_conf* 的情境法則。

$$fr(\alpha, s, win) = \frac{\left|\{w \in W(s,win) | \alpha \text{ occurs in } w\}\right|}{|W(s,win)|} \quad (2.5)$$

以下便為情境法則的法則形式：

$L_1, L_2, \ldots, Ln \rightarrow R_1, \ldots, R_m$ (c, s, window)

$L_i(1 \leq i \leq n)$和 $R_j(1 \leq j \leq m)$表示在序列中的有序性項目集合，所有項目集合是頻繁次序的發生，在法則中我們稱 $L_i(1 \leq i \leq n)$ 為 LHS 情節(Left Hand Side)，$R_j(1 \leq j \leq m)$ 為 RHS 情節(Right Hand Side)。

情境法則的支持度：
$S= Support(L_1 \cup L_2 \cup ....R_1..... \cup R_m) \geq S_0$，

而情境法則信賴度：

$$C = \frac{Support(L_1 \cup L_2 ..... \cup R_1 \cup .... \cup R_m)}{Support(L_1 \cup L_2 ..... \cup L_m)} \geq C_0$$

以下為一個網路事件序列中的情境法則：

(service = Authentication) → (service = SMTP)(service = SMTP)(60%, 10%, 2sec)，

此法則紀錄了一個認證的情境，表示這三個循序事件出現在整個網路事件序列有10%的可能，而系統的認證服務時間在兩秒之間，會跟隨二個 SMTP 連接，有 60%的可能。支持度 s 是網路事件序列的總稽查記錄中，情節發生的百分比，信賴度 c 是從 LHS 情節連接 RHS 情節的最小的發生可能性。

參、模糊關聯及情境法則探勘機制

在模糊關聯法則中是利用隸屬度的運算來得到法則的支持度及信賴度，這隸屬度的決定最重要的是找出群集的重心，一旦群集中心求出，則整個資料的模糊隸屬度也可以藉由群集中心建構出來。在模糊資料探勘應用於入侵偵測方面，Luo [7]利用 Kuok 等學者所提出之模糊關聯法則，從網路的流量及系統記錄檔中萃取出有用的資訊，以代表正常使用者的行為樣式。Florez [5]等學者基於 Kuok 等學者所提出之模糊關聯法則，改善了支持度及信賴度門檻值的設定，以產生比 Luo 等學者更有用的關聯法則，使得從資料庫中挖掘之使用者行為樣式能更正確。

情境法則一開始雖被發展於資料探勘，但因其特長描述網路交通的封包，所以最常被應用在入侵偵測系統中。在 Mannila [9]所提出之情境法則研究中，會產生法則數量過多且法則過長等問題。因此，在後續情境法則研究中，為解決在此之前的情境法則常有無效及過長的法則產生的問題，Luo [8]等人便以模糊理論來解決此問題。於 2003 年 Dobrowiecki [8]研究中應用情境法則來探勘警訊使得能有效的建構出攻擊的警訊模組。Qin 與 Hwang [10]在 2004 年更提出了其改進的演算法，削減大量多餘的情境法則的產生。然而，此一演算法所產出的法則對攻擊的偵測方面仍嫌不足，尤其是偵測率不足 50%，是以本研究基於其演算法加入模糊理論概念，提高所產出法則的偵測率，並與 Qin 等學者[10]所提之機制作一比較分析。

基於上述研究，可知入侵偵測之資料來源的資料量，常常是相當大的資料量，而所謂的攻擊行為常是這些資料集的一小部份，因此若能有效利用資料探勘技術，不但能提升效率更能提高偵測能力，於是本研究整理出一適合入侵偵測的資料探勘技術流程，即是在龐大的資料中先以分群技術將這些資料分成正、異常群集，再配合關聯法則與情境法則挖掘出法則來偵測攻擊。

資料來源為封包，以 DARPA 記錄檔為主，此機制是基於模糊法則探勘之技術，分析 DARPA 記錄檔，首先以模糊群集技術分出正、異常兩群集，經由分群演算法得出各個群集後，可得出每個群集的群集中心，而這群集中心將會被記錄下來，以做為新進資料被歸類為正常群集或異常群集的依據。接著便由模糊關聯法則將先前分群過的正常及異常群集，導出正常及異常行為之關聯法則，之後同樣以正常及異常群集，藉由模糊情境法則導出具攻擊步驟特徵的規則及一般使用者行為步驟，建立出這些法則後，便將這些法則再存入法則資料庫。這些法則可以提供給專家，做為建構入侵偵測系統之依據。因此，以下便詳述本論文之研究方法。

本研究先以 Tsaur [12]等學者提出之 Modified mrFCM 為基礎，較精確地將異常及正常的資料分群出來。資料經由群集技術分為各個群集後，由專家判斷每個群集內的資料為正常使用者行為，或是異常使用者行為，並將群集定義為正常群集或異常群集，經由分群演算法得出各個群集之後，可得出每個群集的群集中心，而這群集中心將會被記錄下來，以做為新進資料被歸類為正常群集或異常群集的依據。

在模糊關聯法則運算的部分採用 Tsaur [13]等人所提之改良式 FGBRMA (Fuzzy Grids Based Rules Mining Algorithm)演算法，利用此種方式只需要對資料庫掃描一次，以降低系統從磁碟讀取資料的時間。而模糊關聯法則中是利用隸屬度的運算來得到法則的支持度及信賴度，這隸屬度的決定最重要的是找出群集的重心，一旦群集中心求出則整個資料的

模糊隸屬度也可以藉由群集中心建構出來。本研究是利用前述之 Modified mrFCM 所計算出的群集中心，做為 FGBRMA 獲得隸屬度之依據，導出正常及異常行為之關聯法則，建立出這些法則後，可將這些法則再存入法則資料庫。

在 Tsaur [12]等人的研究之中，雖然以 mrFCM 與模糊關聯法則機制來提高分群樣本與關聯法則的效率與精確率，但對提升偵測率來說，仍然稍嫌不足。為了提升更高的偵測率，本研究加入模糊情境法則探勘，以期能藉由與關聯法則不同的特性，在網路傳輸的異常行為紀錄中去發掘一個情境(Episode)，發現行為事件之間的關係。在分析行為資訊流後，便可降低多餘的行為分析，與預知事件的發生，如此用於偵測多步驟網路攻擊行為，便可提高偵測率。然而，在情境法則研究中，自從 1997 年 Mannila [9]提出以來，雖有助於發現可疑事件的發生頻率以及次序，但由於演算法效率不彰，常導致大量無效或過長情境法則產生，有時反而降低入侵偵測系統效率，因此 Qin 與 Hwang [10]在 2004 年初更提出了修改演算法，削減大量多餘的情境法則的產生，解決在此之前的無效及過長的法則產生的問題。但在 Luo [8]研究中指出：(1)網路資料包含了許多數值的特徵，如來源 IP、時間性的變數等，有利於模糊理論的導入；(2)太過明確的定義容易將正常行為歸類為異常，因此導入模糊理論技術將有助於入侵偵測準確率的提昇。是故，基於上述原因，本論文將基於 Qin 所提之演算法修改成模糊情境法則演算法，以提升入侵偵測系統準確度。以下及為本文所提之模糊情境法則演算法：

在輸入部分主要為門檻值 f_0、所有的關鍵屬性與包含所有網路連線的集合 $\boldsymbol{T}$。

輸出部分則為模糊情境法則並將其加入法則庫中。

步驟 1：以改進的FGBRMA (Fuzzy Grids Based Rules Mining Algorithm)演算法，依據每個在T中的項目集X與Y計算其隸屬度。

步驟 2：根據表2.1 FGBRMA符號定義，將資料轉換為模糊集合。

步驟 3：根據每個關鍵屬性產生可能的模糊項目集合並且計算出其 Support $S(X) = Support(X \cup Y)$

步驟 4：掃描整個網路連線T集合並且依據 $L = \{Itemset.Y \mid f(Y) \geq f_0\}$ 以及 $f(Y) = \dfrac{Support(Y)}{S(Y)}$ 產生可能的項目集合。

步驟 5：將這些法則集合依照 $Support(I_1, I_2, ...I_n) \geq f_0 \times Min(S(I_i))$ 找出符合之項目集合。

步驟 6：進行信賴度計算依照 $C = \dfrac{Support(L_1 \cup L_2 ..... \cup R_1 \cup .... \cup R_m)}{Support(L_1 \cup L_2 ..... \cup L_m)} \geq C_0$，以確認是否產生法則，若 $C \geq C_0$ 則法則成立存入法則庫，否則回到步驟1。

因此，由上述可知本論文之模糊關聯與情境法則機制架構便是先以 Modified mrFCM 為基礎，較精確地將異常及正常的資料分群出來，接著便由模糊關聯法則將先前分群過的正常及異常群集，導出正常及異常行為之關聯法則，之後同樣以正常及異常群集，藉由模糊情境法則導出具攻擊步驟特徵的規則及一般使用者行為步驟，建立出這些法則後，便將這些法則再存入法則資料庫。

肆、實驗與分析

我們以實驗證明所提之模糊關聯與情境法則探勘機制是可以提高偵測率。而在實驗環境方面，作業系統為 Linux

SUSE9.0，系統內軟體為預設值，實驗利用麻省理工學院林肯實驗室(MIT Lincoln Laboratory)在 1999 年所錄製的網路流量資料，藉由 Tcpreplay 程式在獨立的區網中直接把流量送往受測主機，以確保偵測到的流量完全來自播放所得，不致於被其他網路流量所干擾，而實驗的流程如圖 4-1 所示。其中資料集分為 Training 階段 與 Testing 階段二部份，Training 階段是將第一至第三星期的資料經前處理存進資料庫。在 Testing 階段，則是以第四及第五星期作為測試資料集，以 Training 階段所產出的法則來判斷出攻擊。之後本研究程式以法則庫內法則與資料庫裡的 DARPA 資料集作比對，找出攻擊所在，再比照 DARPA 所公佈的攻擊以計算其偵測率與誤報率。而其攻擊的分類如下：

(1) Denial of Service 的攻擊模式為攻擊者利用本身的機器或經由操控其他大量存有攻擊程式的跳板主機發動攻擊(DDoS)使受害主機的 CPU、Memory 等資源耗盡讓服務無法正常運作。

(2) User to Root (U2R)的攻擊模式為攻擊者經由正常管道或竊聽網路上帳號密碼，取得一般使用者權限之後，再利用系統上某些涉及管理者權限的軟體弱點如緩衝區溢位，以取得系統管理者權限。

(3) Remote to Local (R2L)的攻擊模式為攻擊者無需要在受害主機上有任何的帳號，只需透過發送封包至受害主機即有可能達到取得敏感資料或對主機造成破壞，也可利用受害主機在網路提供的服務程式進行攻擊。

(4) Surveillance/Probing 通常是攻擊者在發動攻擊的前置作業，可依據向偵測主機發出封包所得到的回應得知其作業系統或其他軟體版本弱點等。

圖 4-2 即是根據前述所探討的入侵偵測系統的架構及偵測技術、模糊理論、資料探勘中的群集及關聯法則，開發出模糊關聯與情境法則探勘機制的系統畫面。我們亦實作出 Qin 等學者[10]所提之情境法則演算法，而圖 4-3 為本機制與 Qin 等學者所提之機制作比較，其中 R2L 及 U2R 攻擊比較偏向對主機攻擊，因此針對網路流量分析的情境法則便無法有效產生法則，導致偵測率偏低。但本研究因結合模糊關聯與情境法則，使得能找出單一的攻擊行為及攻擊事件的組成與發生次序，得以大大提高偵測能力，經實驗證明確實能對偵測高度發揮效用且誤報率也不會過高。

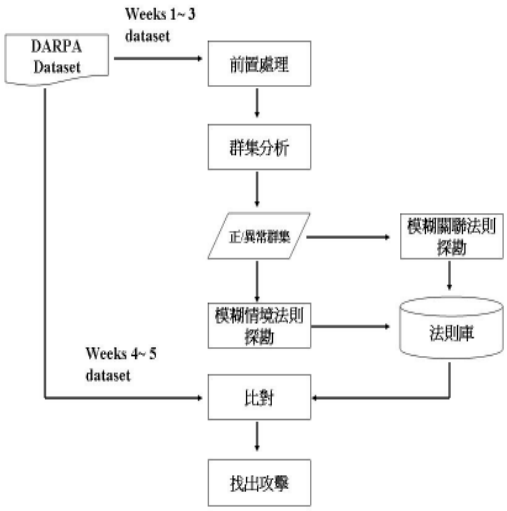

圖 4-1 實驗流程

圖 4-2 系統畫面

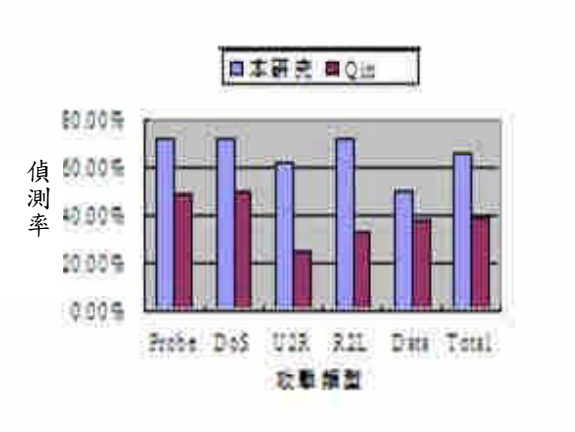

圖 4-3 系統偵測率比較

伍、結論與建議

入侵偵測之資料來源的資料量是相當龐大的，而攻擊的資料卻只佔其中的一小部分，因此透過資料探勘技術來挖掘一致性及有用的系統特徵樣式，可以找尋資料中隱藏的行為模式，並分析是否存在攻擊行為。因此，本研究提出一適用於入侵偵測之模糊關聯與情境法則機制，其特色在於加入模糊理論使得能提高準確度外，也以群集分析將資料分出正、異常群集，之後在法則探勘時也會提升法則正確度，進而應用模糊關聯法則探勘技術於各群集中找出其間的關聯，發掘出可能的關聯法則，找出單一的攻擊行為。此外，本論文亦使用模糊情境法則演算法，於各資料集中找出多重序列間的相互關係，以發掘出攻擊步驟的組成及其發生次序，並將這些法則建構於正常及異常的法則資料庫中。

藉由此兩種演算法產生的法則，可更準確的偵測攻擊行為特徵，讓網路管理者對駭客入侵攻擊手法有更深入的了解，以預防相同的攻擊事件再發生，提高入侵偵測系統偵測率。另外本論文也開發出模糊關聯與情境法則系統，利用 DARPA Dataset 來進行實驗，並與現有最新的法則探勘技術作一比較分析，而經過實驗測試驗證本論文所提出之機制確實有提高一定的偵測能力。

誌謝

本計畫接受國防部中山科學研究院研究計畫案 XC94E22P、國科會研究計畫案 NSC 92-2623-7-212-005 資助，特此致謝。

參考文獻

[1] S. Axelsson, “Intrusion Detection Systems: A Taxomomy and Survey”, Technical Report, Dept. of Computer Engineering, Chalmers University of Technology, Goteborg, Sweden, pp. 99-15, 2000.

[2] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms,” Plenum, New York, 1981.

[3] J.E. Dickerson, J. Juslin, O. Koukousoula and J.A. Dickerson, “Fuzzy intrusion detection,” IFSA World Congress and 20t h NAFIPS International Conference, Vol. 3, pp. 1506-1510, 2001.

[4] FBI/CSI, http://www.gocsi.com/press/20020407.html, 2002

[5] G. Florez, S.A. Bridges and R.B. Vaughn, “An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection,” Proceedings of the North American Fuzzy Information Processing Society Conference (NAFIPS- 2002), pp. 457-462, 2002.

[6] M. Hossain, “Integrating Association Rule Mining and Decision Tree Learning for Network Intrusion Detection: A Preliminary Investigation,” International Conference on Information Systems, Analysis and Synthesis, Vol. 11, pp. 65-70, 2002

[7] J. Luo and Susan M. Bridges, “Mining Fuzzy Association Rules and Fuzzy Frequency Episodes for Intrusion Detection,” *International Journal of*

Intelligent Systems, Vol. 15, pp. 687-703, 2000.

[8] T. Dobrowiecki, "Episode Mining to Automatically Filter False Alarms," *Proceedings of the 10th PhD Mini-Symposium on IEEE Hungary Section*, pp. 44-45, 2003.

[9] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences", Data Mining and Knowledge Discovery, Vol.1, No. 3, pp. 259-289, 1997.

[10] M. Qin, and K. Hwang," Frequent Episode Rules for Internet Anomaly Detection" *Proceedings of The Third IEEE International Symposium on Network Computing and Applications*, Cambridge,2004

[11] Symantec Internet Security Threat Report , https://enterprisesecurity.symantec.com/Content/displaypdf.cfm?SSL=YES&EID=0&PDFID=665&promocode=ITR

[12] W.J. Tsaur and I.M. Fan, "Anomaly Detection Mechanisms for Web Severs in Linux Environments," Communications of the CCISA, Vol. 8, No. 4, 2002.

[13] W.J. Tsaur and Y.C. Shieh, "Constructing Fuzzy Association Rules for Intrusion Detection Systems," Proceedings of the 2003 National Computer Symposium, pp. 1256-1263, 2003.

[14] L.A. Zadeh, "Fuzzy Sets", *Information Control*, Vol. 8, pp. 338-353, 1965.

[15] O.R. Zaiane, M. Xin and J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," *Proceedings of Advances in Digital Libraries Conference (ADL- 98)*, pp. 19-29, 1998.

[16] W. Lee, S. J. Stolfo, and K. W. Mok, "Mining audit data to build intrusion detection models," *In 4th International Conference on Knowledge Discovery and Data Mining*, pp. 66-72, 1998.

[17] W. Lee, S. J. Stolfo and K. W. Mok, "A data mining framework for building intrusion detection models," *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, pp. 120-132, 1999.

[18] W. Lee, S. J. Stolfo, P. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop and J. Zhang, "Real Time Data Mining-based Intrusion Detection," *Proceedings of the 2001 DARPA Information Survivability Conference and Exposition (DISCEX II)*, pp. 89-100, 2001.

[19] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Database," *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 207-216, 1993.

[20] H. Han, X. L. Lu, J. Lu, C. Bo and R. L. Yong, "Data mining aided signature discovery in network-based intrusion detection system," *Source ACM SIGOPS Operating Systems Review*, Vol. 36 , Issue 4, pp. 7-13, 2002