應用資料探勘技術於防火牆最適化策略訂定之先期研究

林俊男	林克偉	林俊成	謝昆霖
中正大學資管所	中正大學資管所	亞洲大學資工系	台東大學資管系
a103p2@yahoo.com.tw	kwlin@yahoo.com.tw	Smart0133@hotmail.com	klhsieh@nttu.edu.tw

摘要

在有限的資源下，如何將效益發揮到最大是企業日前一個非常重要的課題，特別是關於資訊安全的相關議題，如何在現存的防禦設備下，發揮最大的防禦偵測能力，便是企業面臨的難題。本研究嘗試利用智慧型的資料探勘技術，從防禦設備—防火牆的日常紀錄檔中做群聚分析，以挖掘觀察出各種連線活動的特徵，經由特徵的判斷，找出適合一般性及各種特殊性的連線活動其適切的防禦偵測規則，以及其規則最適切的判斷順序組合。

關鍵詞：資訊安全、資料探勘、防火牆、群聚分析。

Abstract

It had become an important issue for making the efficiently performance under the limited resources to most enterprises, especial for the topic of information security. In this study, we intend to apply the data mining technique to mine the optimum strategy setting by using the clustering analysis for the log file. The characteristics kept behind the log file can be mined via the data mining technique, and the combination and the sequence of the strategy for the firewall can be obtained.

Keywords: information security, data mining, firewall, clustering analyrsis.

1. 緒論

隨著資訊科技與網路的快速成長，使人類正式步入數位化時代，造成電腦與網路已和我們日常生活有著密不可分的關係。透過電腦與網路雖然帶給人們極大的便利性，但也相對的造成了許多過去所沒有的問題，而「資訊安全」(Information Security)便是在此環境下所衍生出來的問題。

關於資訊安全問題的重要性，我們可以透過各行業逐年增加資訊安全需求的預算，來得知資訊安全問題有逐漸受到重視的趨勢(圖 1)[1]，許多的企業為了有效的防治資訊安全的問題，便使用了許多的防範措施，其中，最常見的一種方式，便是建置防火牆設備(Firewall)。PC 安全幽靈 Steve Gibson 表示防火牆無法主動的偵測到資訊攻擊，只能被動式的採取一些事先定義好的過濾規則來阻擋一些非法的使用者以及未經過允許的應用程式。因此，防火牆過濾規則設計的好壞，便成為了一個關鍵的課題。

有關防火牆的研究，已經有很多的專家學者探討過，但經由文獻的探討後，我們可以發現這些相關的研究，幾乎都在討論如何制定或加強防火牆的安全策略[2,5]、提出一個新穎的架構[3,6]、混合其他技術加強防禦能力[4]等議題，但在現實環境中，企業不可能一直隨著這些新的研究成果，不斷的改善自身的防火牆，因為這將造成企業組織一個龐大的成本花費。因此，在現實環境限制底下，我們應該去思考如何將現有防火牆設備，其功能發揮到極致，以平衡成本與效益間的關係。所以，本研究嘗試應用資料探勘(Data mining)的技術，從防火牆的 log 檔案中，利用模糊群聚解析(Fuzzy clustering analysis)的方式，試圖找出各種連線類型的特徵，以及適用的安全策略規則，並且找出一組最適切的安全策略規則順序，使在現有受限的資源下，防火牆設備能達到最佳的效益。

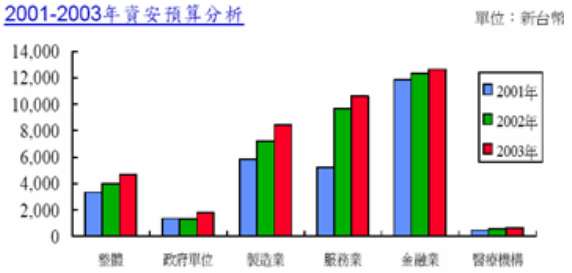

圖 1. 我國 2001 年至 2003 年各行業資訊安全預算

2. 文獻整理

2.1 防火牆(Firewall)

「防火牆」一詞，依照經濟部網路商業應用資源中心的定義為：位於 Gateway 上的一種安全措施，以保護企業內部或私人網路資源免於外來侵入。防火牆系統可為一路由器、個人電腦或是主機，同時包括安全政策，以規定可被允許的服務與連線許可。防火牆通常建置於獨立的電腦中，如此外來的請求無法直接取得私人網路內的資源，其過濾所有的網路封包，並決定是否將其傳送至目的地。一種簡單的篩選外來請求方式，為確認其係來自於先前定義可接受的網域名稱及 IP 位址。對機動使用者，防火牆則使用安全登錄程度及授權認證，來允許遠端存取[7]。 防火牆的基本功能應包含[8]：

1.執行保全策略所禁止之資料存取管制。

2.紀錄所有可疑的資料存取。

3.當有入侵行為發生時，必須能夠對網管人員提出警告。
4.有些防火牆甚至還可以提供相關的統計資料，供網管人員參考。

2.2 資料探勘(Data mining)

資料探勘可說是資料庫技術的一個邏輯上的演進。最早的資料庫不過是被作為紙張紀錄的一種替代品，是個僅能夠作概述和報告的資料儲藏室。但是，隨著 SQL 這樣的查詢工具的不斷發展，資料庫管理員可以較靈活地查詢資料。而資料探勘技術的出現使得資料庫管理員能更加有效率的找出隱藏的樣本和知識。最純粹的資料探勘其實並不包括尋找特定的資訊，資料探勘僅只是找出資料裡已有的樣本，而不是先給定一個問題或是假設再來進行探勘的動作。Fayyad(1996)[9,10]定義的將資料探勘與資料庫中之知識發掘（Knowledge Discovery and Data Mining）嚴格的區分開來。資料庫中的知識發掘是從資料中選取合適資料後，再進行資料處理、轉換、資料探勘以及結果評估，資料探勘不過是知識發掘的步驟之一。Michael 和 Gordon(2000)[16]認為資料探勘是經由自動或半自動的方法探勘及分析大量的資料，以建立有效模型及規則。Kleissner(1998)[11]則認為資料探勘是一種新的且不斷循環的決策支援分析過程，它能夠從資料中發現隱藏價值的知識，以提供專業企業人員參考。綜合以上所述，本研究認為可將資料探勘定義為：結合資料視覺化（Data Visualization）、機器學習（Machine Learning，ML）、統計方法（Statistics）和資料庫（Database）等多種技術以期在大型資料庫中搜尋、擷取以及歸納出隱藏的知識規則或是模型，然後將這些規則或是模型提供給專業人員作為其決策時的重要依據。Fayyad[9,10]認為資料探勘之步驟應包含：

1.開發一個已知的應用領域，和有意義的/有關的先備知識（prior knowledge），並且從顧客觀點確認知識發掘（KDD）的最終目標。
2.建立一目標資料集合： 選擇一個資料集合，或是變數子集，或是資料樣本來進行資料探勘。
3.資料淨化（Cleansing）以及其他前置作業：至少須包括消除雜訊，收集建構模型的必要資訊，以決定如何處理遺失的資料檔案，並列出時間序列的資訊和變動。
4.資料轉換（transfer）和減量（Reduction）：根據任務目標中所描述的資料，找出其中有用的特點。利用維度換算或轉換的方法，使得有效變數的數量減少，或發現資料中的不變數。
5.選擇適合進行知識發掘（KDD）流程（步驟一）的一種資料探勘類型。例如：彙總（summarization）、分類（classification）、分群（clustering）等。
6 選擇演算法：依知識發掘（KDD）的資料屬性、型態，選擇適合的資料探勘演算法尋找資料關係模型（pattern）。
7.進行資料探勘：找出有意義的資料關係模型。
8.解釋步驟七所建立的資料模式，或是再次執行步驟一到七之間任一步驟。
9. 解釋探勘結果及評估所得關係類型（Interpretation/Evaluation）：包含知識的呈現（Knowledge Presentation）。

2.3 模糊群聚(Fuzzy Clustering)

群聚演算法(Clustering Algorithm)的理論基礎是在資料集之中，劃分別出許多各有共同特色的群組，並進行特色分析，它影響所呈現的現象為如何，則可找出對策略有益的參考資訊，以增強、改善或了解策略的規劃與實施。群聚技術又可以分為模糊群聚(Fuzzy Clustering)與硬式群聚(Hard Clustering)兩大類，其差別在資料點與群組之間的關聯度不同。例如其各別關聯度呈現之值域[11]：

1.Hard Clustering: 0, 1 (0 or 1)
2.Fuzzy Clustering: [0, 1] (0~1 之間所有可能之值)

Hard Clustering [14] 是在資料點被分配到某一群組後，就與其它群組沒有關聯了；而 Fuzzy clustering 之資料點則會有與各群組的隸屬值，才將其歸類到隸屬程度高的群組，所以當資料點介於兩群組間之較模糊地帶時，尚可明確地知其資料點對於兩群各別之隸屬程度如何。所以能較 Hard Clustering 方法，得知資料在其它群組的方面，它所擁有的隸屬度。所以 Fuzzy Clustering 應用在市場區隔[13]、影像切割或需要較詳細的分群功能時，則可以有更進一步之隸屬資訊，如圖 2 所示。

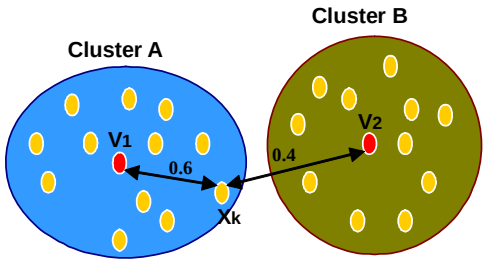

圖 2. 模糊群聚圖[11]

本研究採用 Fuzzy C-means algorithm (FCM)之模糊聚群演算法，由 Bezdek(1981)[15]所提出之目標函數：

$$\text{Min } J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m d^2(x_k-v_i) \quad (1)$$

而 c: 群組數
n: 資料數
m: (m≧1) 模糊度之參數

$$d^2(x_k-v_i)=\sqrt{\sum_{j=1}^{p}(x_{kj}-v_{ij})^2} \quad \text{(p 為資料之維度)} \quad (2)$$

μ_{ik}：為資料點對中心點之隸屬度

$$\mu_{ik}=\frac{1}{\sum_{h=1}^{c}\left[d^2(x_k-v_i)/d^2(x_k-v_h)\right]^{\frac{1}{m-1}}} \quad (3)$$

$1 \leqq i \leqq c, 1 \leqq k \leqq n$

c 個群與各資料點之隸屬度:

$$U_{c\times n} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & & & \\ . & & & \\ \mu_{c1} & . & \cdots & \mu_{cn} \end{bmatrix} \quad (4)$$

為 c 個群之中心點距陣:

$$V_c = \begin{bmatrix} v_1 \\ v_2 \\ . \\ v_c \end{bmatrix} \quad (5)$$

$$v_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m} \qquad 1 \leqq i \leqq c \quad (6)$$

群組的分割好壞，可由群組內的相似程度高低，與群組與群組之間的差異多寡來判別。群組內的相似程度愈高，則其群組分得愈好，而群組與群組之間的差異，則是愈大愈好。

3. 研究方法與實驗

接下來我們即將描述本研究的資料處理程序，如圖 3 所示，

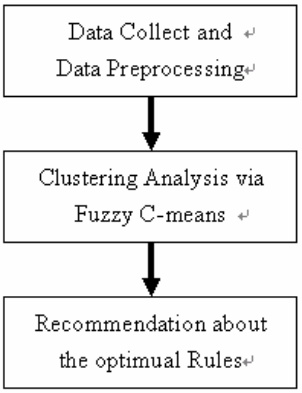

圖 3.本研究資料處理程序

3.1 資料收集與前處理

我們從某地方政府機關的資訊部門之防火牆設備中，將一日內的 log 紀錄檔隨機抽出 1000 筆，且將各種類別型資料加以編碼，並彙整成 15 項屬性(如圖 4、圖 5)。

圖4. 原始log紀錄檔

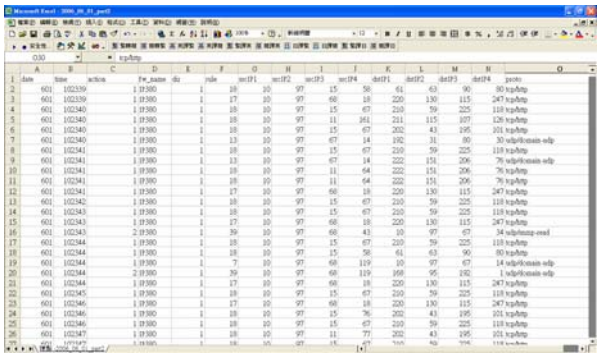

圖5. 經前處理過後的資料

3.2 模糊分群

為了簡化分析操作，我們利用 Matlab 工具軟體來實現 FCM 的模糊分群方法，首先便是要進行 FCM 相關的參數設定與操作，其過程請參見圖 6(以分兩群為例)。

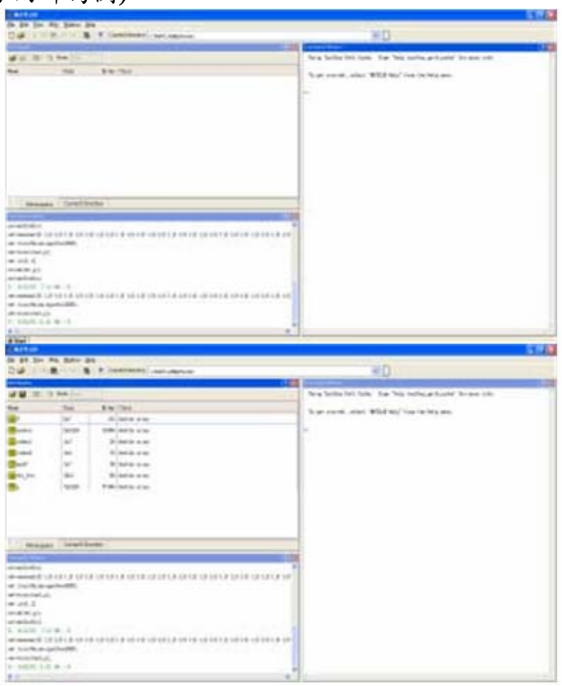

圖 6. FCM 相關參數設定與操作過程

接著我們利用試誤法(try-and-error)來獲得決定分群數的資訊。由於決定群數時常是決策者在進行

群聚分析中，最感到困惑的地方，因此我們設計了一個指標來輔助決策者選擇分群群數，這指標是取自於 FCM 的特色—模糊歸屬函數(fuzzy membership degree)；歸屬函數值越大，代表該資料所屬於該群的程度越大，歸屬函數值越小，則反之；我們也可以利用此特性推導出：如果在某一分群的結構下，其歸屬函數值之最大值與最小值相減後的差值越大，表示該分群結構越明顯，否則反之。利用這個原理將可以輔助決策者在進行群聚分析時的參考依據。我們所設計的此指標步驟如下：

步驟一、將每一個分群過程的歸屬函數值統整紀錄，繪製成一個比較表。

首先將每一個分群過程的歸屬函數值統整紀錄，目前我們僅針對群數為 2 到 10 的情形進行群聚分析。以劃分成 2 群的情形為例，其第 1 筆至第 10 筆資料的歸屬度值如表 1 所示。

表 1. 歸屬度值

Data items	1	2	3	4	5	6	7	8	9	10
Cluster1	0.000146	0.000146	0.000145	0.000145	0.000145	0.000142	0.000145	0.000142	0.000145	0.000142
Cluster2	0.999854	0.999854	0.999855	0.999855	0.999855	0.999858	0.999855	0.999858	0.999855	0.999855
$DV_{i,k}$	0.999708	0.999708	0.99971	0.99971	0.99971	0.999715	0.99971	0.999717	0.99971	0.99971

步驟二、接著利用公式(7)計算不同分群結構時，各筆資料對應所屬群的歸屬函數差值，公式如下：

$$DV_{j,k} = \underset{all\ \ i}{MAX}\left(MD_{i,j,k}\right) - \underset{all\ \ i}{MIN}\left(MD_{i,j,k}\right) \tag{7}$$

其中，$MD_{i,j,k}$ 意指在第 k 個分群過程中，第 j 個資料對於第 i 個分群所構成的歸屬函數值；$DV_{j,k}$ 意指在第 k 個分群過程中，第 j 個資料點所構成的歸屬度差值(最大歸屬度值 - 最小歸屬度值)；i 表示在 k 個分群過程中之群數；j 表示在進行群聚的資料點；k 表示被用來執行的可能分群過程。同樣以劃分成 2 群的情形為例，其第 1 筆至第 10 筆資料的歸屬度差值如表 1 裡的黃色部分所示。

步驟三、針對每個分群的處理過程，找出最小的歸屬函數差值，其公式說明如下：

$$MDV_k = \underset{all\ \ j}{MIN}\left(DV_{j,k}\right) \tag{8}$$

其中，MDV_k 表示在第 k 的分群處理過程中的最小歸屬函數差值。所以其各別分群之最小歸屬度值的差值如表 2 所示。

表 2.最小歸屬度差值

Clusters	2	3	4	5	6	7	8	9	10
MDV	0.88947	0.522761	0.518352	0.443006	0.438319	0.45923	0.428386	0.449165	0.311245

步驟四、找出所以分群過程中最小歸屬函數值差值之最大值以決定最適分群數，所採用的判斷式如下：

$$ODV = \underset{all\ \ k}{MAX}\left(MDV_k\right) \tag{9}$$

其中，ODV 表示在所有分群過程中之最小歸屬函數差值之最大值，其所對應的分群數即為最適分數。由表 4 我們可以發現，在分成兩群的結構時，依據我們所設計的指標所得到之 ODV 值最大，因此在本研究中，我們可以選擇兩群的結構是最適切的群數結果。

3.3 群聚解析

根據 ODV 值，我們選擇 2 群的結構進行群聚分析，並且有幾點有趣的發現：

(1)如表 3 所示，落在第 1 群的 10 筆資料，皆受到規則第 39 條的判斷而被阻擋在防火牆外，此 10 筆資料有一個共同的特徵，便是皆從同一個來源端所發送出來的封包(10.97.68.106)，因此我們可以建議管理者針對該 IP 位置做更進一步的瞭解與處理。或是針對該 IP 將規則第 39 條調整至較為前面的檢查順序，以優化防火牆設備的效能。

表 3. 劃分 2 群時，歸屬於第 1 群的資料

item	time	action	dir	rule	srcIP	srcIP	srcIP	srcIP	dstIP	dstIP	dstIP	dstIP	protocol
718	102624	2	1	39	10	97	68	106	59	120	208	174	111851
719	102624	2	1	39	10	97	68	106	211	22	216	142	111167
723	102624	2	1	39	10	97	68	106	218	166	154	65	119384
724	102624	2	1	39	10	97	68	106	219	79	84	77	110448
766	102624	2	1	39	10	97	68	106	61	230	204	115	111334
768	102624	2	1	39	10	97	68	106	218	103	196	159	123331
769	102624	2	1	39	10	97	68	106	218	103	168	126	117657
770	102624	2	1	39	10	97	68	106	202	86	186	44	118133
771	102624	2	1	39	10	97	68	106	59	115	75	174	111959
773	102624	2	1	39	10	97	68	106	218	255	242	129	110538

(2)針對剩餘的 990 筆資料，我們以防禦規則為基礎進行簡單的統計後可知，在這些資料中，最常被使用來判斷的規則依序是規則 18(555 次)、規則 17(148 次)和規則 39(126 次)，因此我們可以建議管理者可以將此三條規則其判斷的優先順序調整至較前面的優先順序，並且將未使用的規則如規則 1、規則 2...等調整至較後面的優先順序，以使防火牆設備的效率更佳。

表 4.劃分 2 群時，第 2 群資料使用之判斷規則統計

Rule	1	2	3	4	5	6	7	8	9
Number	0	0	0	0	10	0	14	0	0
Rule	10	11	12	13	14	15	16	17	18
Number	3	6	1	26	18	9	0	148	555
Rule	19	20	21	22	23	24	25	26	27
Number	0	62	0	0	0	0	0	0	0
Rule	28	29	30	31	32	33	34	35	36
Number	0	0	0	0	0	0	0	0	0
Rule	37	38	39	40	41				
Number	0	0	126	0	12				

4.討論與結論

網路的安全性已成為一個廣泛討論的話題，該如何在如此複雜的環境裡找到安全的通道，想必永遠是一個重要的議題。有鑑於目前防火牆的昂貴，專業人員的不足以及經費的缺乏，希望透過本研究提出的方法，可以幫助企業、學校等組織在有限的

經費、設備下，將其防禦設備之功能發揮到極致。但是在時間、人力、經費與設備等種種的限制下，尚有一些議題可以藉由後續的討論加以改進：

(1)資料的取樣：由於本研究目前著重於此方法的可行性與合理性，因此在實驗的階段只利用一小段的時間內資料來進行實驗，後續應該完整的針對單一組織來蒐集一個長時間的資料，並進行分析的程序，且應採取其他資料進行驗證，以發揮本方法真正的功效。

(2)多維度的分群解析：目前本研究只針對防火牆設備已制定好的防禦規則來進行討論其設定的規則之判斷優先順序是否適切，後續研究可以針對其他的特徵、或是搭配多種不同的特徵去做混合式的群聚解析，相信會有意想不到的收穫。

(3)防禦規則的特徵：本研究初步只討論其防禦規則順序排列的適不適切，後續研究可以針對防禦規則裡的特徵，進行相關的討論，以期望制定出更為適切的規則。

(4)其他的資訊設備：本研究目前只針對防火牆設備來進行討論，但其相關資訊安全設備事實上種類非常繁多，所以未來可以針對其他的資訊安全設備來做進一步的解析。

(5)Fuzzy C-means：本研究主要利用 FCM 演算法來進行群聚解析的動作，主要理由是想藉助 FCM 裡的模糊歸屬函數特性，制定出一個劃分群數的依據，但是事實上，還可以針對模糊歸屬函數，進而探討其單一資料在群與群間的隸屬程度，依照隸屬程度的強弱，應可搭配出更加適切的制定規則。

(6)其他的分析技術：未來研究可以嘗試使用不同的群聚技術(例如：統計方法、類神經網路等)，或是採用其他的思維方式(非群聚的觀念)來進行探討，或許會出現出有趣的結果。

參考文獻

[1]資訊安全發展現況與趨勢 http://www.nbl.org.tw/nbl_old/service/report/s003.pdf

[2]黃志琦(民86)。經濟型企業網路防火牆。中原大學資訊工程研究所碩士論文，桃園縣。

[3]王振茂(民90)。電子商務交易課稅應用防火牆之架構探討。中華大學科技管理研究所碩士論文，新竹市。

[4]黃新琪(民84)。建構結合IP路由器及代理伺服器之多層防火牆。元智大學電機與資訊工程研究所碩士論文，桃園縣。

[5]陳世洋(民92)。以IPTABLES設計高可用性叢集式防火牆。南台科技大學資訊管理研究所碩士論文，台南縣。

[6]游啟勝(民91)。合作式防火牆之設計與應用。國立中央大學資訊管理研究所碩士論文，桃園縣。

[7]經濟部網路商業應用資源中心 http://www.ec.org.tw/knowledge/glossary_detail.asp?keyword=19

[8]網際家數位科技 http://www.e-ipro.com/products/security/firewall/index.php

[9] Fayyad, U. M., Data mining and knowledge discovery: making sense out of data, IEEE Expert, Vol. 11, No. 5, 1996, pp.20-25.

[10] Fayyad, U. M., G.. Piatetsky-shapiro, and P. Smyth, From Data Mining to Knowledge Discovery in Databases, 1996, AI Magazine, pp.37-54.

[11] Kleissner, C., Data mining for the enterprise, Proceedings of the Thirty-First Hawaii International Conference, Vol. 7, 1998, pp. 295-304.

[12]簡祥全(民91)。知識經濟國家時間群聚分析。朝陽科技大學資訊管理研究所碩士論文，台中縣。

[13] Hsu, Tsuen-Ho(1999). *An Application of Fuzzy Clustering in Group-Positioning Analysis*. Department of Business Administration, I-Shou University,.

[14] Zimmermann, J. J., Fuzzy Set Theory and Its Applications, 1991, Kluwer Academi, Boston.

[15] Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms, 1981, Plenum Press, New York, NY,

[16] Michael J. A. Berry, Gordon S. Linoff., Mastering Data Mining, 2001, John Wiley & Sons, Inc.