應用資料探勘偵測電信資料異常之研究

Using Data Mining to Detect Abnormality in Telecommunication

鄭富山
輔仁大學資訊管理研究所
fushan@im.fju.edu.tw

翁頌舜
輔仁大學資訊管理研究所
im1032@mails.fju.edu.tw

摘　要

資料探勘（Data Mining）是近年來資料庫應用領域中相當熱門的議題。而資料探勘一般是指在資料庫中，利用各種方法和技術，將過去所累積的大量歷史資料，去進行分析、歸納、整合等工作，以找出有興趣的樣式（Interesting Patterns），粹取出有用的資訊，以提供管理階層作為訂定決策的依據。然而上面所提到的都是傾向於從過去大量的歷史資料中去做分析，但是在現實生活的應用上，有些資訊是需要即時地告知管理者，例如：電話盜打（Fraud）、網路干擾、信用卡盜刷…等，以將損失降至最少；而且這些異常的情況可能會經常的改變，因此要如何應用資料探勘的技術，來完成一個具有即時性、適應性（Adaptive）的系統，便成為本論文主要的目標。本論文將以電信資料為主，應用熱力學中的熵函數（Entropy）來作為評估資料庫資訊含量的重要指標，並將其標示出的正常與異常資料當作類神經網路（Neural Network）的輸入，經由類神經網路不斷地訓練、學習後，希望能夠準確地找出各種異常的情況，以幫助管理者做出最佳的決策，為企業謀得最大的利潤。

關鍵詞：資料探勘、盜打、類神經網路、熵函數

1、緒論

在資料庫的發展過程中，資料探勘可以說是一個剛興起的研究領域，其主要的目的就是從大量繁雜的資料中，找出有興趣且具有代表性的樣式（Pattern），以提供有用的資訊給管理者。就目前而言，大多數企業的資料庫已經充斥著各式各樣的資料（例如：交易資料、電信通話記錄、競爭對手資料、未來趨勢…等），但是這些企業卻無法有效的去分析、管理及使用這些資料，甚至對這些資料束手無策；而資料探勘卻能夠過濾出這些繁雜的資料，找出有意義的資訊。因此許多大企業已經體認到這些資料的重要性，進而開始從事資料探勘的工作；而尚未投入資料探勘的企業，也積極地著手準備，希望能儘快的開始進行，以為企業創造出真正的價值。

近年來由於網路的盛行與普及，再加上電信自由化，使得無線通訊已經成為現今最熱門的議題，而行動電話也成為無線通訊中競爭最激烈的一項服務。過去電信業在行動電話的服務上是屬於類比式通訊，因此僅有少數的資料儲存於資料庫中。但自從國內引進泛歐數位行動電話系統（Global System for Mobile Communication，簡稱 GSM）後，便開始從原本的類比式通訊轉變成數位式通訊，因此電信業者可以將用戶所有的通話資料詳細的記錄下來，這也造成了電信業資料呈現爆炸性的成長。

雖然電信業的通話服務已經由類比式通訊轉變成數位式通訊，這種數位技術在行動通訊上的確可以提供比類比技術還要高的通話品質及通話安全，並且可以避免干擾上的問題，但仍然無法有效地杜絕盜打的情況。而除了盜打的問題之外，電信業者如何有效的訂定其行銷策略，或者如何設計出讓顧客更滿意的計價方案…等，也都是電信業者相當關注的問題；由於一個好的行銷策略或計價方案，不但可以為企業吸引更多的新用戶加入，也可以保留住舊用戶，所以電信業者無不費盡心思在此方面。

GSM 系統的確幫助電信業者提供了更高品質的服務，但是在盜打方面而言，卻仍然是防不勝防；再加

上電信業者對於客戶群的使用行為無法充分的了解，所以不能訂定出一個好的行銷策略。因此電信業者要如何有效地應用這些龐大的通話資料，進而從其中挖掘出有用的資訊(即挖掘出異常的狀態)，便成為本論文最主要的研究動機。

1.1、問題描述

隨著電信市場的蓬勃發展，各家電信經營業者無不卯盡全力爭取最多的客戶及業績，卻也得要時時小心保護自己的利潤，因為已有一群不肖之徒，使用各種方式進行電話盜撥，進而獲取暴利，而成為全球電信業者的最痛（賴德謙，1998)。根據估計，全球的無線通訊盜打（Fraud）的規模大約為美金二十億到三十億元之間，從這個台幣大約將近一千億元的數字來看，可見得此一問題的嚴重性。以科技發達的美國為例，該國在歷經多年來與盜打份子周旋之後，如今也只能把這類無線電犯罪控制在總通話時數（airtime）的百分之零點五左右。然而環顧其他國家，像這類盜撥電話的情形卻是有增無減，特別是在市場還處於開發中的亞太地區以及拉丁美洲（柳林緯，1999)。

根據我們對這些盜打行為的分析，此類問題不僅會造成電信業者鉅大的金額損失，而且將造成許多有形的成本浪費及無形的衝擊。我們大致可以發現下列幾個重要的問題：

（一） 近年來在電信業中有許多的硬體防盜設備相繼地被發展出來，這些硬體設備雖然可以有效地減少盜打行為的發生，但這些硬體設備都相當的昂貴，因此並不是所有的電信業者都能夠即時的採用。

（二） 由於盜打行為的氾濫，會造成基地台及設備佔用的問題，而會影響到正常用戶的使用權利，因此電信業者必須增加成本，擴增基地台及購買新的設備。

（三） 若盜打情況無法有效地解決，將會嚴重影響到電信業者的聲譽，這樣不僅會造成舊用戶的流失，更會影響到開發新客戶的業績。

（四） 電信業者必須加派許多的人力來處理這類的問題，而客服中心也必須要接受這些被盜撥客戶的抱怨。

（五） 硬體防盜設備其功能完全專注於偵測盜打，而無法針對某些特定的用戶（族群）去做行為上的分析。

除了上面所描述的問題之外，我們常常可以從電視上、廣播媒體或者平面廣告看到各家電信公司五花八門的廣告，而這些廣告最主要的目的當然是吸引新用戶的加入。但是在競爭如此激烈的情況之下，如何訂定出一個好的行銷策略將成為勝利的關鍵。因此，本研究希望能夠設計出一個具有適應性、即時性的系統，藉由分析用戶的通話記錄來有效地解決這些問題，並且期望對於電信業能夠有所貢獻。

1.2、研究目的

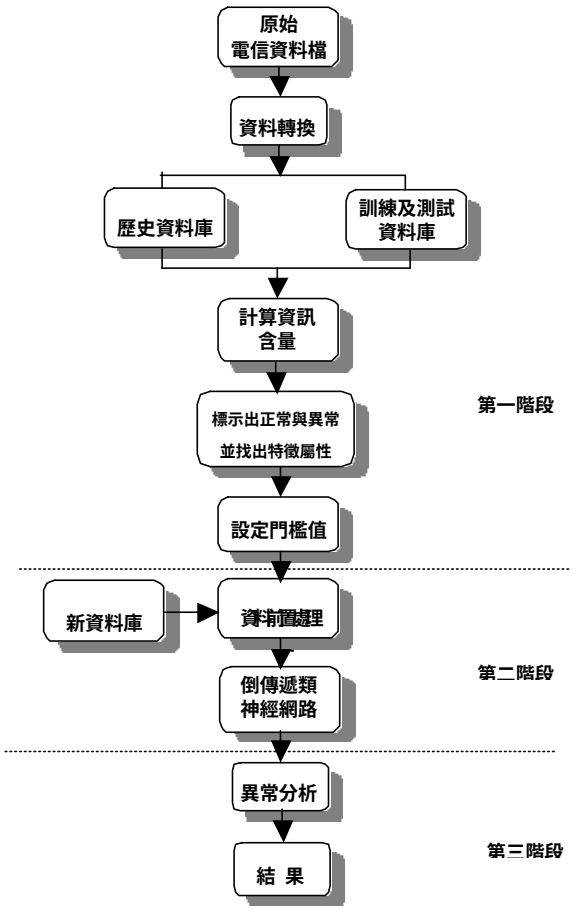
在現實生活中，資料探勘在某些特定領域上的應用，是需要即時地提供有用的資訊給管理階層，讓管理者能夠隨時得知那些資料產生了重要的改變，接下來才能針對這些具有代表性的樣式（Pattern）去進行監控、分析，以找出真正想要的資訊。本論文是以電信資料為主，希望能夠藉由資料探勘的技術，從這些通話記錄中偵測出具有異常行為的資料，以幫助管理者有目標地針對特定的樣式去做分析，進而能夠判斷此異常是否為一盜打的行為，或者是因為某些促銷策略所造成的影響（例如：若電信公司的費率下降，可能會導致用戶的使用率增加)。然而就盜打方面來說，盜撥者（Bandit）會想盡各種方式來破解電信業者的防盜系統，而讓電信業者無法輕易的偵測出其盜打的行為；換句話說，即盜撥者的盜打行為是會隨時改變的；另外，在這些通話記錄中，有些用戶的使用行為可能會隨著時間而改變（例如：職業變動)，或者隨時可能會有新用戶的加入。所以本系統除了即時性之外，還必須具有適應性（adaptive)，如此才能更精確的偵測出各種異常的情況。

由於本研究是採用熱力學中的熵函數作為評估指標以找出異常的區間，並使用類神經網路的技術來幫助我們達到分類的效果，因此我們可以每隔一段期間將這些新進的通話記錄，經由類神經網路的重新訓練、學習，讓系統能夠更準確地找出異常的情況，進而提供有效的資訊給管理者進行分析，希望能夠幫助電信業者找出盜打的情況及訂定出更好的行銷策略。因此本論文主要的目的就是希望能夠設計出一個具有即時性、適應性的系統，將所挖掘出的異常資訊提供

給管理者進行分析，讓管理者能夠清楚的了解異常的原因，並可監控特定用戶的使用行為，這樣不但可以有效地降低電信業者的損失，更可以幫助電信業者在行銷策略上做適當的調整，以獲得最大的利益。

2、研究方法

從過去的許多研究中，我們可以發現人工智慧（Artificial Intelligence，簡稱AI）的技術已經扮演了相當重要的角色，無論在醫學、工業工程、影像辨識…等領域都少不了人工智慧技術的應用，因此本論文將藉由人工智慧強大的能力，來幫助我們偵測出電信資料的異常。而我們在這裡所指的異常並不完全就是盜打的行為；也有可能是因為費率調降，而造成用戶通話時間增加；或者在某些地區的用戶其使用率突然降低…等，這些都屬於異常的範圍。在此章節，我們將開始介紹本論文的系統架構與流程，我們也將深入地探討如何有效地運用類神經網路，來作為本系統的核心架構，並且將說明如何去分析及確認異常的情況，以便能夠更精確地發出異常的警告給管理者，讓管理者可以更有目標的針對問題點加以解決。



2－1：系統架構與流程

在本研究中的系統架構與流程（如圖2－1所示），大致上可歸納為三個階段，而每個階段的執行步驟分別如下所示：

（一） 計算歷史資料庫的資訊含量，並且標示出訓練及測試資料庫的正常與異常區段。其詳細的執行步驟如下：

1. 收集足夠的電信資料。
2. 將這些電信資料經由SS7的轉換軟體進行轉換的工作，把這些原始資料轉成可以了解的通話記錄。
3. 將通話記錄分成歷史資料庫、訓練資料庫、以及測試資料庫。
4. 計算出歷史資料庫資訊含量的平均值與標準差。
5. 以歷史資料庫的資訊含量為評估指標，標示出訓練及測試資料庫的正常與異常區間，並且找出其特徵屬性。
6. 經由門檻值的設定，增加或減少異常區間的個數。

（二） 將所標示出的正常與異常區段做適當的轉換，並輸入倒傳遞類神經網路做訓練及測試。其詳細的執行步驟如下：

1. 從第一階段中所找出的正常與異常區間，選擇出適當的資料來作為訓練資料檔與測試資料檔。
2. 將訓練資料檔與測試資料檔做適當的轉換（即正規化）。
3. 執行倒傳遞類神經網路的訓練與測試。
4. 當類神經網路的模組訓練及測試完成之後，便可以正確地將新進的通話記錄做分類，以找出異常的情況。

（三） 將倒傳遞類神經網路所找出的異常結果做進一步的分析，以歸納出造成異常的主要因素。其詳細的執行步驟如下：

1. 將異常區段的特徵屬性與原始的通話記錄做比對，並找出其造成異常的原因。
2. 將這些異常的訊息提供給管理者做處理，以幫助電信業者能夠從大量的通話記錄中找出其中所隱含的資訊，並希望能夠對電信業者有所幫助。

3、完整的處理流程

由於本論文最主要的研究目的，就是要找出電信資料中發生異常的區間，因此如何去定義何謂正常？何謂異常？便成為我們最主要的工作。在這裡我們將採用前面所提到的方式－資訊含量（Entropy），來作為我們評估正常或異常的重要指標。有了這個重要的評估指標，接下來我們就可以使用它從大量的電信資料庫中，標示出哪些是異常的區段。

3.1、資料來源

在資料收集方面，我們很幸運地取得了二十萬筆的電信資料，在此我們再次誠心地感謝某電信公司的提供。但是由於這些資料可能會牽涉到個人的隱私權問題，因此我們將會對這些資料做適當的轉換，以避免不必要的麻煩。在主叫號碼與被叫號碼中，若是行動電話及呼叫器，其前四碼我們將分別以 0980-0999 這二十組目前尚未使用的號碼來做轉換，而後面六碼我們將隨機給定。若是室內電話，其區域號碼不變，而後面七碼我們將以 0000000-0000040 來做轉換（台北縣、市則以八碼來轉換，00000000-00000015）。

在電信資料中，由於其訊號格式複雜且繁多，因此我們將從其中選擇六個最重要的屬性來進行我們的實驗：

1、 主叫號碼（Calling）：即撥號者之電話號碼。
2、 被叫號碼（Called）：即接收者之電話號碼。
3、 主叫區域號碼（OPC）：即撥號者之區域代碼。
4、 被叫區域號碼（DPC）：即接收者之區域代碼。
5、 時間（Time）：為一通電話之起始時間。
6、 通話時間（Length）：為此次通話的時間長度。

這些通話資料是直接由基地台上的硬體設備所轉出的檔案，再經由解碼軟體進行處理，最後再從這些解碼完成的資料中選出上面所述之六個屬性，將之轉換並儲存於 SQL 資料庫中。如表 3－1 所示，即為我們原始通信資料的型式。

表 3－1：原始資料之樣式（Length 的單位：秒）

ID	Time	Called	Calling	OPC	DPC	Length
1	2000/11/12 09:13:02	0986123456	0985325812	2000	3000	13
2	2000/11/12 20:32:16	0991876543	0200000001	9000	2000	78
3	2000/11/12 18:11:49	030000001	030000002	4000	2000	157
4	2000/11/12 07:50:05	0995789432	0996190783	2000	3000	53
5	2000/11/12 11:59:59	070000003	0983394857	8000	2000	71
6	2000/11/12 12:45:01	070000004	0200000002	8000	2000	66
7	2000/11/12 06:20:11	0996197368	0200000003	3000	2000	19
8	2000/11/12 23:34:45	0998222890	0986090135	8000	3000	183
9	2000/11/12 09:44:42	0998134765	0993187349	2000	3000	344
10	2000/11/12 19:12:36	060000005	0200000004	7000	2000	98
….	….	….	….	….	….	….

3.2、取得資訊含量

從過去的文獻中（陳志安，2000），我們可以了解到資訊含量（Entropy）最主要就是用來評估混亂程度，其公式如下所示：

$$H(P(x)) = -\int P(x)\log_2 P(x)dx$$

如果我們把它應用於資料庫中，便可以說是用來評估整個資料庫的混亂程度。當資訊含量的值越高時，也代表了此資料庫越混亂；換句話說，此資料庫所包含的資訊也會越多。反之，當資訊含量的值越低時，代表了此資料庫較不混亂，也可以說此資料庫所包含的資訊較少。

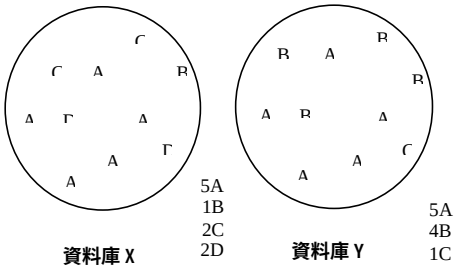


圖 3－1：兩個簡單的資料庫
（資料來源：陳志安，2000）

如圖 3-1 所示，在資料庫 X 裡面總共有十筆資料，其中包含了資料 A 五筆、資料 B 一筆、資料 C 二筆、及資料 D 二筆。接下來我們便可藉由 Entropy 函數，來計算資料庫 X 的資訊含量：

$$H(P(x))$$
$$= H(5/10, 1/10, 2/10, 2/10)$$
$$= -(5/10*\log_2(5/10) + 1/10*\log_2(1/10) + 2/10*\log_2(2/10) + 2/10*\log_2(2/10))$$
$$= -(-0.5 + (-0.3321) + (-0.4643) + (-0.4643))$$
$$= 1.7607$$

在資料庫 Y 中，一樣包含了十筆資料，其中分別為資料 A 五筆、資料 B 四筆、及資料 C 一筆。接下來我們一樣可以藉由 Entropy 函數，來計算資料庫 Y 的資訊含量：

H(P(x))

=H(5/10, 4/10, 1/10)

=－(5/10*$\log_2$(5/10) + 4/10*$\log_2$(4/10) +

1/10*$\log_2$(1/10))

=－(－0.5 + (－0.5287) + (－0.3321))

= 1.3608

由上面兩個資料庫的資訊含量，我們可以清楚的看出，資料庫 X 的資訊含量高於資料庫 Y 的資訊含量；也就是說，在相同的資料量之下（兩個資料庫都是十筆資料），資料庫 X 所包含的資訊是比較多的，也可以說它必定隱含了某種特殊的資訊。

3.3、Entropy 演算法

在 Entropy 演算法中，我們首先將探討過去文獻中所提到的資訊含量計算方式，並說明此方式若用在現實的電信資料庫中，會產生什麼樣的錯誤，接著我們再針對其演算法之缺失加以改良，並提出新的方式，讓 Entropy 函數能夠適當地運用在現實的電信資料庫中，且能夠有效地表現出何謂異常的區段。

3.3.1、過去的方式

在過去的文獻中（陳志安，2000），曾經提到如何使用 Entropy 函數來衡量電信資料庫的資訊含量，其作法如下所示。首先，假設表 3－2 為某一電信資料庫的內容，在過去的方式中，為了能夠方便統計出相同的樣式（Pattern），其做法必須先將這些原始的電信資料歸納到概念階層的最底層，如表 3－3 所示。

表 3－2：電信資料庫 A 之內容

ID	Time	Called	Calling	OPC	DPC	Length
1	2000/11/12 09:13:02	0986123456	0985325812	2000	3000	13
2	2000/11/12 20:32:16	0991876543	0200000005	9000	2000	78
3	2000/11/12 18:11:49	030000006	030000007	4000	2000	157
4	2000/11/12 07:50:05	0995789432	0996190783	2000	3000	53
5	2000/11/12 11:59:59	070000008	0983394857	8000	2000	71
……	……	……	……	……	……	……
101	2000/11/12 12:45:01	070000009	0200000006	8000	2000	66
102	2000/11/12 06:20:11	0996197368	0200000007	3000	2000	19
103	2000/11/12 23:34:45	0998222890	0986090135	8000	3000	183
104	2000/11/12 09:44:42	0998134765	0993187349	2000	3000	344
105	2000/11/12 19:12:36	060000010	0200000008	7000	2000	98
……	……	……	……	……	……	……

表 3－3：電信資料庫 A 之概念階層

ID	Time	Called	Calling	OPC	DPC	Length
1	白天	行動電話	行動電話	台北	桃園	短
2	晚上	行動電話	室內電話	屏東	台北	短
3	晚上	室內電話	室內電話	新竹	台北	中
4	白天	行動電話	行動電話	台北	桃園	短
5	白天	室內電話	行動電話	高雄	台北	短
……	……	……	……	……	……	……
101	白天	室內電話	室內電話	高雄	台北	短
102	白天	行動電話	室內電話	桃園	台北	短
103	晚上	行動電話	行動電話	高雄	桃園	中
104	白天	行動電話	行動電話	台北	桃園	長
105	晚上	室內電話	室內電話	台南	台北	短
……	……	……	……	……	……	……

當這些電信資料經過歸納之後，我們可以找出何謂單一樣式（如表 3－4 所示），因此我們只要找出此樣式在資料庫中出現了幾次，便可利用 Entropy 函數來計算出這個資料庫的資訊含量。

表 3－4：電信資料庫 A 之某一單一樣式

Time	Called	Calling	OPC	DPC	Length
白天	行動電話	行動電話	台北	桃園	短

經過我們仔細地分析過電信資料庫後，我們發現過去文獻中所描述的方式，並無法實際運用於現實的電信資料庫中，在表 3－2 裡所列舉之資料應該是呈現平均分布的一個情況。但是，在實際的電信資料庫中（如表 3－5 所示），其通話記錄都是按照時間來作排列的，而且往往在同一時間內就會有好幾筆通話記錄，再加上我們所取得的通話記錄，只限定在台中某些基地台的記錄，所以其主叫地域代碼（OPC）及被叫地域代碼（DPC）大多數都是當地的用戶所使用的，只有當某些用戶打到其他縣市，或者當其他縣市有人打電話到此地區，才會出現不同的區域代碼，但是這些記錄在資料庫裡畢竟是屬於少數的資料，因此若使用過去文獻中所提出的方法，可能會造成一些問題，以下我們便詳細地討論之。

表 3－5：電信資料庫 B 之內容

ID	Time	Called	Calling	OPC	DPC	Length
1	2000/11/12 09:13:02	0986123456	0985325812	4800	9160	13
2	2000/11/12 09:13:02	0991876543	0987890123	4800	9160	28
3	2000/11/12 09:13:02	0986123456	0994689234	4800	9160	17
4	2000/11/12 09:13:04	040000011	0996190783	4300	9160	53
……	……	……	……	……	……	……
331	2000/11/12 17:59:59	040000012	0200000009	9160	2000	66
332	2000/11/12 17:59:59	0996197368	040000013	4800	4300	19
333	2000/11/12 18:00:01	0998222890	0986090135	4800	9160	183
334	2000/11/12 18:00:01	0998134765	0993187349	4800	9160	344
……	……	……	……	……	……	……
751	2000/11/12 05:58:02	0985745934	0996981043	4800	4800	46
752	2000/11/12 05:59:47	060000014	0988683444	6000	4300	239
753	2000/11/12 06:07:23	0991100832	040000015	9160	4800	29
754	2000/11/12 06:13:39	0997237845	0983394857	4800	9160	31

.....	.....	.....	.....	.....	.....	.....

首先，若我們按照過去文獻中所提到的方式，我們必須先將表 3－5 的電信資料歸納到其概念階層的最底層（如表 3－6 所示），我們可以發現相同樣式的資料幾乎都聚集在一起，這樣便產生了問題，如果我們使用 Entropy 函數來計算，我們會發現每個區段的資訊含量大都為零或者接近零，而較容易出現不同樣式的區段，將變成時間的交界處（如白天跟晚上）。因此，若使用此方式我們可以發現異常區間大都會落在時間交界處，所以說使用這個方式在現實的資料庫中並無法真正的表現出每個區段的資訊含量的特性。

表 3－6：電信資料庫 B 之概念階層

ID	Time	Called	Calling	OPC	DPC	Length
1	白天	行動電話	行動電話	豐原	烏日	短
2	白天	行動電話	行動電話	豐原	烏日	短
3	白天	行動電話	行動電話	豐原	烏日	短
4	白天	室內電話	行動電話	清水	烏日	短
……	……	……	……	……	……	……
331	白天	室內電話	室內電話	烏日	台北	短
332	白天	行動電話	室內電話	豐原	清水	短
333	晚上	行動電話	行動電話	豐原	烏日	中
334	晚上	行動電話	行動電話	豐原	烏日	長
……	……	……	……	……	……	……
751	晚上	行動電話	行動電話	豐原	豐原	短
752	晚上	室內電話	行動電話	台南	清水	中
753	白天	行動電話	室內電話	烏日	豐原	短
754	白天	行動電話	行動電話	豐原	烏日	短
……	……	……	……	……	……	……

3.3.2、改良後的方式

本論文將針對過去文獻中的缺失提出改良，讓 Entropy 函數能夠真正地在電信資料庫中發揮其效用，而且能夠有效地標示出那些為異常的區段。接著我們便開始逐步說明本論文的做法。

首先，假設我們擁有一個電信資料庫 Z（如圖 3－2 所示），接著再將資料庫裡面的各個屬性都視為個別獨立的資料庫，然後我們便可藉由 Entropy 函數，計算出各個獨立資料庫各個區段所擁有的資訊含量，再把各個區段所計算出來的資訊含量總和加總起來，便可以得到整個資料庫的資訊含量。由於我們是針對原始的電信資料型式去做計算，並沒有將它歸納至所謂的概念階層，因此我們的方式將更能夠表現出其資訊含量的意義，我們也相信用此方式所標示出來的異常區間，一定隱含了某些我們所感興趣的訊息。由於我們是分別算出各個屬性所擁有的資訊含量之後再做加總，因此我們還可以找出究竟是哪個屬性造成資訊含量的增加或減少，並且把這個屬性標示出來當作這個區段的特徵屬性，以幫助後面階段能夠做更進一步的分析。

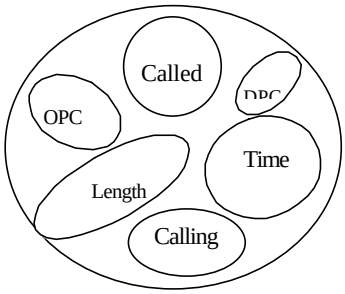


圖 3－2：電信資料庫 Z

我們來看個簡單的例子，假設表 3－7 是我們目前所擁有的電信資料庫，我們把它視為是某月份的通話記錄，若我們以 10 筆記錄為一個區間，則此資料庫第一個區段之資訊含量的計算方式如下：

TimeEntropy = H (2 / 10, 1 / 10, 3 / 10, 2 / 10, 1 / 10, 1 / 10) = 2.4459

CallingEntropy = H (3 / 10, 1 / 10, 4 / 10, 1 / 10, 1 / 10) = 2.0460

CalledEntropy = H (5 / 10, 1 / 10, 2 / 10, 2 / 10) = 1.7607

LengthEntropy = H (5 / 10, 2 / 10, 1 / 10, 1 / 10, 1 / 10) = 1.9579

OPCEntropy = H (5 / 10, 3 / 10, 2 / 10) = 1.4853

DPCEntropy = H (5 / 10, 2 / 10, 2 / 10, 1 / 10) = 1.7607

表 3－7：某一真實的電信資料庫

ID	Time	Called	Calling	OPC	DPC	Length
1	2000/11/12 09:00:02	0989821770	0986093845	4800	9160	13
2	2000/11/12 09:00:02	0991982001	040000016	4800	4300	11
3	2000/11/12 09:00:04	0996148244	0995092338	9160	4800	7
4	2000/11/12 09:00:07	0984803569	0991831323	4300	9160	378
5	2000/11/12 09:00:07	0992090743	0990172394	4800	4800	4
6	2000/11/12 09:00:07	040000017	0200000010	9160	2000	301
7	2000/11/12 09:00:13	0998092122	0989990274	4800	4300	38
8	2000/11/12 09:00:13	040000018	0995092338	9160	9160	32
9	2000/11/12 09:00:17	0991982001	0986093845	4800	9160	11
10	2000/11/12 09:00:31	0989821770	0986093845	4300	9160	249
.....	.....	.....	.....	.....	.....	.....

在計算各個屬性的資訊含量時，有幾點我們必須要特別注意，在通話長度（Length）這個屬性中，我們計算其出現次數的方式與其他屬性的計算方式並不相同。經過我們仔細地分析過整個資料庫後，我們發現在通話長度這個屬性中，每個區段要出現通話長度完全相同的機率相當的低，如果我們直接用原方式來計算其資訊含量，將無法有效地表現出其代表的含義，所以我們以每 30 秒為一個單位（即介於 1－30 秒都視為相同的資料），用來計算通話長度中所出現的

次數，如此便可更準確的表現出其資訊含量的意義。我們來看個簡單的例子，如表 3－8 所示，假設這是某兩個不同區段的通話長度（Length）值，我們若使用直接計算相同樣式的方法，則區段 A 與區段 B 它們的資訊含量都是 H (1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6)，但是我們若以每 30 秒為一個單位來計算，則兩個區段的資訊含量就會產生區別，區段 A 的資訊含量為 H (3 / 6, 2 / 6, 1 / 6)，而區段 B 的資訊含量為 H (5 / 6, 1 / 6)，所以我們使用此方式比較能夠表現出其資訊含量的差異。

表 3－8：兩個不同區段的通話長度值

Length
2
19
30
212
235
62

區段 A

Length
7
23
12
29
5
47

區段 B

而在主叫號碼（Calling）與被叫號碼（Called）這兩個屬性中，我們計算其出現次數的方式與其他屬性也不相同。經過我們仔細地分析過整個資料庫後，我們可以發現在一分鐘內往往就會有幾十筆的通話記錄出現，若在通話的尖峰時間甚至會出現上百通的情況，而這些通話記錄中的主叫號碼與被叫號碼要出現完全相同的機率相當的低，也就是說在一分鐘內，同一用戶會撥 2 通電話以上或者同一用戶會接 2 通電話以上的機率相當的低，如果我們直接用此方式來計算其資訊含量，將無法有效地表現出其代表的含義，所以我們以“門號”為單位來做區分（即行動電話將分成遠傳電信、中華電信、和信電信…等；室內電話將分成北部、中部、南部），如此便可更準確的表現出其資訊含量的意義。我們來看個簡單的例子，如表 3－9 所示，假設這是兩個不同區段的主叫號碼（或被叫號碼），我們若使用直接計算相同樣式的方法，則區段 A 與區段 B 它們的資訊含量都是 H (1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6)，但是我們若以“門號”為單位來計算，則兩個區段的資訊含量就會產生區別，區段 A 的資訊含量為 H (3 / 6, 2 / 6, 1 / 6)，而區段 B 的資訊含量為 H (4 / 6, 2 / 6)，所以我們使用此方式比較能夠表現出其資訊含量的差異。

表 3－9：兩個不同區段的主叫號碼（被叫號碼）

Calling（Called）
0989027247（遠傳）
0992433001（遠傳）
0986743587（台灣大）
0983926750（台灣大）
0995021211（遠傳）
0996987418（中華電）

區段 A

Calling（Called）
0982912030（遠傳）
0988278973（和信）
0990646985（和信）
0989238233（遠傳）
0991502388（遠傳）
0992840828（遠傳）

區段 B

當我們求得各個屬性的資訊含量之後，接著我們將上面各個屬性所計算出來的資訊含量加總起來，便可以得到整個資料庫第一個區段的資料含量；依此類推，我們便可求得整個資料庫中各個區段資訊含量的數值。

TotalDBEntropy = TimeEntropy + CallingEntropy + CalledEntropy + LengthEntropy + OPCEntropy + DPCEntropy

接著我們將詳細說明本論文之演算法，其主要可以分成三個部分，分別說明如下：

(一) 計算歷史資料庫資訊含量的平均值及標準差

由於我們的研究最主要的目的是要去分析電信資料的異常情況，因此如何有效地找出異常區間便成為我們最主要的工作之一，我們必須找出異常區間後，才能繼續進行下一步的分析。而我們要如何去評估何謂正常區間？何謂異常區間呢？我們可以利用過去的歷史資料，計算出其資訊含量的平均值及標準差，並以此當作一個正常區段該有的資訊含量。接著我們便可把這個資訊含量當作評估指標，用來找出現有資料庫中，那些區間是異常的情況。

(二) 標示出異常區間及特徵屬性

當我們有了歷史資料庫資訊含量的平均值與標準差後，我們便可利用它來標示現有資料庫的異常區間。假設歷史資料庫的平均值為 3、標準差為 0.5，當現有資料庫其資訊含量若介於 2.5~3.5 之間，則是一個正常的區間，反之，若其資訊含量小於 2.5 或者大於 3.5，則是一個異常的區間。

當我們找出異常區間之後，我們就可以進一步去分析造成此區間異常的原因，並將影響此區間為異常的屬性找出，並把它標示成為特徵屬性，以利於後面

的分析。在我們的研究中，由於整個資料庫的資訊含量是由六個屬性的資訊含量加總而來的，因此若現在資料庫的資訊含量大於 3.5，則表示這六個屬性必定有某幾個資訊含量值偏高，而影響了整個資料庫；反之，若現在資料庫的資訊含量小於 2.5，則表示這六個屬性必定有某幾個資訊含量值偏低。而我們要如何找出影響整個資料庫的特徵屬性呢？我們一樣利用平均值與標準差的概念，當我們在求取整個歷史資料庫的平均值與標準差時，我們也一併找出各個屬性資訊含量的平均值與標準差，接著我們便可利用它來找出特徵屬性。

(三) 設定門檻值，增加或刪除異常區間

由於每個不同的資料庫或者不同的分析方式，所找出來的異常區間個數都不同，因此經由門檻值的設定，將可以依照每個使用者的需求（容忍範圍）來增加或減少異常區間的個數。而在這裡門檻值所代表的意義，也可以說是去調整標準差的大小，例如在前面所提到的歷史資料庫其平均值及標準差分別為 3 及 0.5，而其預設的門檻值就是 16.67%（0.5 / 3 * 100%），因此我們可以依照各個使用者的需求來作調整，以增加或減少其異常區間的個數。如表 3－10 所示，我們舉一個簡單的例子來說明，假設在原門檻值 16.67%（即標準差為 0.5）的情況下，我們可以在資料庫中找出 4 個異常區間（分別為第 2、3、5、8 個區段），然而我們若把它的門檻值提高至 25%（即標準差為 0.75），也就是擴大其容忍範圍，此時我們所找出的異常區間將減少至 2 個（分別為第 2、5 個區段）。有了設定門檻值這個功能，我們便可以視異常區間個數的多寡來做適當的調整，如果我們所找出的異常區間個數較多，我們就可以提高它的門檻值，以幫助我們找出較顯著的異常區間；相反的，如果我們所找出的異常區間個數較少，我們就可以降低它的門檻值，以幫助我們找出一些影響程度較小的異常區間。我們相信增加這個功能對於整個系統必定有所幫助的。

表 3－10：不同門檻值所找出的異常區間

編號	區間	資訊含量	正常或異常
1	001-100	3.323	正常
2	101-200	3.769	異常
3	201-300	3.648	異常
4	301-400	3.341	正常
5	401-500	2.197	異常
6	501-600	3.294	正常
7	601-700	3.289	正常
8	701-800	2.477	異常

（A）門檻值為 16.67%

編號	區間	資訊含量	正常或異常
1	001-100	3.323	正常
2	101-200	3.769	異常
3	201-300	3.648	正常
4	301-400	3.341	正常
5	401-500	2.197	異常
6	501-600	3.294	正常
7	601-700	3.289	正常
8	701-800	2.477	正常

（B）門檻值為 25%

4、倒傳遞類神經網路

在倒傳遞類神經網路中，我們將說明為何我們要使用倒傳遞類神經網路來做作為我們的分類工具，並說明如何將電信資料庫中的資料，轉換成倒傳遞類神經網路的輸入，以及如何調整類神經網路以得到最佳的結果。

（一）為何使用倒傳遞類神經網路

在前面我們已經提過類神經網路無論在影像辨識、語音辨識、醫學、工業工程、資訊……等領域，都已經成為不可或缺的重要工具。而在本研究中，選用類神經網路來作為我們的分類工具，其最主要的原因有兩個：自動學習與重新訓練。接下來我們便詳細說明之：

1． 自動學習

在電信資料庫中，其通話記錄往往在短短的幾分鐘內就出現了幾百通電話，而在通話的尖峰時間更可能出現上千通的情形，再加上每個用戶的使用特性都不相同；因此，我們要如何對這些多而繁雜的資料進行處理、分析呢？在我們的研究中，由於我們希望在正常與異常的分類過程中，能夠含有過去的經驗法則，並且能夠自動地去學習在何種情況下出現的通話記錄是屬於異常的情況，因此我們決定使用類神經網路來作為我們分類的主要工具。

2． 重新訓練

由於電信資料庫裡面的通話記錄會經常不斷地變化（即使用者的行為可能會隨著時間改變），而且可能隨時會有新用戶的加入，或者舊用戶的退出，所以我們不能夠僅以過去的歷史資料當作評估的資訊，我們每隔一段期間就必須重新訓練，讓系統能夠不斷地學習新的規則，如此才能夠更正確地找出異常的情況，讓管理者能夠更有效地去分析結果。

（二）如何將電信資料轉換成類神經網路的輸入

由於電信資料庫裡面的資料屬性是屬於離散型的資料，它不同於財務報表或者股市分析這種數值型的資料，因此無法直接透過值域的變換來做正規化，所以其前置處理就變得更加重要，因為它將影響到輸出結果的準確性。經過我們使用不同的方式測試後，我們選擇了一個最適當的方式來作為本系統中類神經網路的輸入，其轉換方式如下所示：

1．Time

在 Time 這個屬性中，我們以每三小時為單位共把它分成八個區間，並且分別以八個輸入節點（Node）來表示。若某通話記錄其時間屬性為“07:12:48”，則表示它落在“06-09”這個區間內，此時僅有代表此區間的節點（Node3）之輸入值為 1，其餘節點的輸入值皆為 0。

2．Called 與 Calling

在 Calling 與 Called 這兩個屬性中, 若是行動電話我們將以它的門號所屬之電信公司來作為分類，若是室內電話我們將以它的地區來作為分類，而免付費電話與呼叫器號碼我們則把它歸類於其他，在這裡總共分成十個類別，並分別以十個輸入節點來代表。

3．Length

在 Length 這個屬性中, 我們依通話時間的長短將它分成三類，並分別以三個輸入節點來表示。

4．OPC 與 DPC

在 OPC 與 DPC 這兩個屬性中，由於我們所取得的電信資料庫是屬於台中某些基地台的通話記錄，因此其主叫地域與被叫地域幾乎都台中境內，雖然這些資料仍然包含了其他地區的主叫或被叫地域，但這些畢竟是屬於少數的資料。經過我們詳細地分析過這些資料庫後，我們將依照其主叫與被叫地域的不同共分成四類，並分別以四個輸入節點來表示。

（三）如何標示出異常區段中的異常記錄

當我們找出異常區段後，我們要如何標示出異常的記錄呢？為了要能夠增加訓練的正確性，我們不能直接將整個異常區段中的記錄都標示成異常，我們必須要有技巧的來標示。在前面的部分我們已經詳細敘述過如何找出異常區間以及標示出其特徵屬性，接著我們就利用這些特徵屬性來幫助我們標示出異常的記錄，假設造成某區段異常的原因是其資訊含量大於它的容忍範圍，此時我們就必須找出「出現次數少及出現頻率高」的記錄將它標示為異常。相反的，假設造成某區段異常的原因是其資訊含量小於它的容忍範圍，此時我們就必須找出「出現次數多及出現頻率低」的記錄將它標示成異常。為了更清楚的說明何謂「出現次數少及出現頻率高」與「出現次數多及出現頻率低」的做法，我們來看下面這個簡單的例子：假設某一資料庫其異常的原因是它的資訊含量大於歷史資料庫資訊含量的容忍範圍，且它的特徵屬性為 Time 及 Calling, 接下來我們就針對記錄屬性 Time 及 Calling 出現次數的陣列來做判斷，我們從表 4－11 及表 4－12 中可以找出其出現次數最少的是 1，且在整個陣列中出現頻率最高的也是 1，因此我們就針對這些資料把它標示成異常。

表 4－11：記錄 Time 屬性出現次數的陣列

樣式	2000/11/12 03:13:01	2000/11/12 03:13:25	2000/11/12 03:13:39
出現次數	3	1	1

2000/11/12 03:13:58	2000/11/12 03:14:27	2000/11/12 03:14:46
1	3	1

表 4－12：記錄 Calling 屬性出現次數的陣列

樣式	0996087113	0984139877	0995123441	0995872091
出現次數	2	1	1	1

0989489157	040000024	0983335309
3	1	1

反之，假設某一資料庫其異常的原因是它的資訊含量小於歷史資料庫資訊含量的容忍範圍，且它的特徵屬性為 OPC，接下來我們就針對記錄屬性 OPC 出現次數的陣列來做判斷，我們從表 4－13 中可以找出其出現次數最多的是 7，且在整個陣列中出現頻率最低的也是 7，因此我們就針對這些資料把它標示成異常。

表 4－13：記錄 OPC 屬性出現次數的陣列

樣式	4080	4400	9156	4300
出現次數	7	1	1	1

5、不同樣式的結果分析

在實驗的部分，我們最主要可分為三大部分：全部資料庫分析、特定族群分析、及單一樣式分析，我們希望能夠讓不同的使用者，根據他們不同的需求，藉由不同的分析方式，以找出隱藏在資料庫中的資訊，讓使用者能夠針對這些資訊做進一步的分析。以下我們就分別介紹三種不同的分析方式：

5.1、全部資料庫之分析

在全部資料庫的實驗中，我們選擇了六個屬性（即所有的屬性）來作為類神經網路的輸入資料，分別為

Time、Called、Calling、Length、DPC 及 OPC，而類神經網路的輸出將分成兩類：正常或異常，至於隱藏層節點個數的選擇，經過我們對不同節點個數的隱藏層訓練之後，我們可以發現當隱藏層節點個數為 39 個時有較好的收斂結果，接著我們再針對不同的學習因子來做訓練，當學習因子為 0.5 時會有最好的收斂結果。當測試完成之後，我們發現在全部資料庫的實驗中，我們的確可以將正常與異常的資料區分出來，而且能夠達到 66.8%的正確率。接著我們可以針對所找出的異常記錄加以分析，根據原始資料的顯示，我們可以發現在某些區段中，其 Called 與 Calling 屬性所出現的電話號碼幾乎都是同一家電信公司的門號，因此我們可以看出在這些異常區段中隱含了「網內互打增加」的資訊，而造成此異常的原因可能是此家電信公司推出了網內互打半價或者網內互打免費的行銷策略。我們也可以發現在某些區段中，其 Time 屬性的資訊含量突然地減少，而造成這些區段異常的主要原因是，在同一個時間中連續出現了 2~3 筆通話記錄，有時甚至出現了 6~7 筆，這個異常可以告知管理者在這些區段中是屬於通話的尖峰時間，此時管理者可以根據這些通話量的多寡，考慮是否必須增加硬體設備，以應付尖峰時間的通話量，才不會造成用戶的抱怨。

5.2、特定族群之分析

在 Time = “ 半夜 ” 與 Length = “ 全部 ” 這兩個特定族群的實驗中，我們選擇了五個屬性來作為類神經網路的輸入資料，分別為 Called、Calling、Length、DPC 及 OPC（由於其時間都是在半夜，因此對於訓練並無幫助，所以將時間屬性去除），而類神經網路的輸出將分成兩類：正常或異常，至於隱藏層節點個數的選擇，經過我們對不同節點個數的隱藏層訓練之後，我們可以發現當隱藏層節點個數為 17 個時有較好的收斂結果，接著我們再針對不同的學習因子來做訓練，當學習因子為 0.5 時會有最好的收斂結果。當測試完成之後，我們發現在特定族群的實驗中，我們的確可以將正常與異常的資料區分出來，而且能夠達到 74.62%的正確率。接著我們可以針對所找出的異常記錄加以分析，根據原始資料的顯示，我們可以發現在某些區段中，其 Length 屬性的資訊含量突然地增加，而造成這些區段異常的原因是在 Length 屬性中增加了一些通話時間較長的記錄，此時我們可能要針對這些門號做個別的監控，因為這有可能是一個盜打的情況。

在 Called = “ 行動電話 ” 與 Calling = “ 行動電話 ” 這兩個特定族群的實驗中，我們選擇了六個屬性來作為類神經網路的輸入資料（即所有的屬性），分別為 Time、Called、Calling、Length、DPC 及 OPC，而類神經網路的輸出將分成兩類：正常或異常，至於隱藏層節點個數的選擇，經過我們對不同節點個數的隱藏層訓練之後，我們可以發現當隱藏層節點個數為 19 個時有較好的收斂結果，接著我們再針對不同的學習因子來做訓練，當學習因子為 0.5 時會有最好的收斂結果當測試完成之後，我們發現在特定族群的實驗中，我們的確可以將正常與異常的資料區分出來，而且能夠達到 79%的正確率。接著我們可以針對所找出的異常記錄加以分析，根據原始資料的顯示，我們可以發現在某些區段中，其 Called 與 Calling 屬性所出現的電話號碼幾乎都是屬於同一家電信公司的門號，因此我們可以看出在這兩個異常區段中隱含了「網內互打增加」的資訊，而造成此異常的原因可能是此家電信公司推出了網內互打半價或者網內互打免費等行銷策略。而在某些區段中，其 Called 屬性的資訊含量突然地降低，經過我們的分析後可以發現這些造成此區段異常的通話記錄幾乎都是 A 電信公司的門號，因此我們可以發現本公司的用戶常常與 A 電信公司的用戶有通話的往來，所以我們也許可以找 A 電信公司共同推出一些新的方案，以吸引更多的用戶加入。

5.3、單一樣式之分析

在單一樣式的實驗中，我們選擇了五個屬性來作為類神經網路的輸入資料，分別為 Time、Length、Called、DPC 及 OPC（由於 Calling 是針對某一特定用戶，因此對於訓練並無幫助，所以將 Calling 屬性去除），而類神經網路的輸出將分成兩類：正常或異常，至於隱藏層節點個數的選擇，經過我們對不同節點個數的隱藏層訓練之後，我們可以發現當隱藏層節點個數為 14 個時有較好的收斂結果，接著我們再針對不同的學習因子來做訓練，當學習因子為 0.1 時會有最好的收斂結果當測試完成之後，我們發現在單一樣式的

實驗中，我們的確可以將正常與異常的資料區分出來，而且能夠達到79.92%的正確率。接著我們可以針對所找出的異常記錄加以分析，根據原始資料的顯示，我們可以發現在某些區段中，其 Called 屬性所出現的電話號碼幾乎都是同一位使用者，而且這個電話號碼並非本電信公司的門號，因此我們可以針對這個用戶做一些特別的行銷，譬如：若介紹一位新用戶加入，即可獲得網內互打免費，或者贈送100小時的免費通話…等，而這些行銷可以經由寄發帳單時附加在其中，以達到個別行銷的目的。我們也可以發現在某些區段中，其 OPC 與 DPC 突然出現了與平常不一樣的情況，此時我們就必須要特別注意了，因為這些通話可能是一個盜打的情況，或者是因為此用戶離開了台中地區所造成的結果。

6、結論與建議

在網路的盛行與普及之下，再加上電信自由化的影響，我們知道無線通訊已經成為我們身邊不可缺少的東西。而電信業者面對每天所累積下來的龐大通話記錄，要如何有效地去處理與應用呢？我們都知道電信業已經成為現今競爭最激烈的行業之一，各家電信業者也都紛紛提出許多不同的行銷方案，以吸引更多的新用戶加入，但是要如何有效的訂定這些策略也成為一個主要的問題。

在國內將資料探勘的技術運用在電信資料方面的相關研究尚未普及，因此本篇論文希望能夠藉由分析這些龐大的通話記錄進而找出其異常的部分，再經由特徵屬性來幫助我們分析這些異常的原因，以幫助電信業者有效地處理這些龐大的通話記錄，甚至對於如何訂定其行銷策略能有所貢獻。在這篇論文中，我們經由不同的分析方式來驗證本研究的可行性，我們發現本系統的確能夠找出資料庫中異常的區間，並提供有效的資訊給予使用者，相信對於異常方面的相關研究應該會有所幫助。

由於資料探勘應用於電信資料方面的研究尚未普及，再加上我們所取得的資料有限，因此尚有許多值得研究與改善的空間。分別如下所述：

1. 增加用戶的詳細資料

   由於我們所取得的電信資料是直接由基地台轉換出來的通話記錄，因此並沒有所謂的用戶詳細資料（例如：姓名、性別、年齡、職業、學歷…等資料），這也造成了我們研究上的限制，而無法做出更完整的分析；如果我們可以增加用戶的詳細資料，相信對於異常方面的分析一定會有更大的幫助（例如：我們可以針對不同年齡層的使用行為加以分析、根據職業的不同來分析他們的使用率、或者針對不同計費方案的用戶來分析其行為模式…等）。

2. 增加電信資料的相關屬性

   在我們的研究中，我們只針對 Time、Called、Calling、OPC、DPC、及 Length 這六個屬性來加以分析，但是在實際的電信資料中其擁有的屬性是相當多的，因此應該還是有很多屬性可以加以利用，如果我們能夠增加更多有用的屬性來做分析，相信對於電信方面的研究一定會有更多的貢獻。

3. 增加國際電話的通話記錄

   在我們所取得的電信資料中並沒有包含國際電話，而我們知道大多數盜打的情況都是發生在盜撥國際電話，因此如果可以增加國際電話的通話記錄，相信對於盜打方面的異常偵測一定會更有幫助的。

4. 將系統改善至 Multi-Tier 的架構

   由於本系統是開發在單機使用的架構，但是為求以後能夠利用網路來達到分散式處理或者多人共同存取的目的，我們希望能夠將整個系統架構由單機改善成為 Multi-Tier 架構，以提高本系統的可用性。

5. 改善演算法的執行效率

   本系統在計算資料庫的資訊含量時，必須要針對六個不同的屬性分別計算，因此會多次的掃描資料庫，希望未來能夠有發展出更好的演算法來改善其執行效率。

參考文獻

[1] 江天池，“全球行動電話發展趨勢”，台灣通訊雜誌，1999 年 8 月，頁 122-125。

[2] 柳林緯，“淺談 GSM 行動電話標準”，台灣通訊雜誌，1998 年 12 月，頁 118-123。

[3] 柳林緯，“淺談行動電話盜打之現況與因應對

策”，台灣通訊雜誌，1999 年 10 月，頁 128-132。
[4] 劉青儒，“GSM 數位行動電話的現況與展望”，新電子期刊，1997 年 2 月，頁 101-108。
[5] 賴德謙，“電信經營業者的痛－電話盜撥”，台灣通訊雜誌，1998 年 11 月，頁 92-97。
[6] 陳志安，“以屬性導向歸納法挖掘資料異常之研究”，中央大學資訊管理研究所碩士論文，2000。
[7] Azzedine Boukerche and Mirela Sechi Moretti Annoni Notare, “Neural Fraud Detection in Mobile Phone Operations”, IPDPS 2000 Workshops, 2000, pp.636-644.
[8] F. Bonchi, F. Giannotte, G. Mainetto, D. Pedreschi, “A Classification-Based Methodology for Planning Audit Strategies in Fraud Detection”, KDD-99 San Diego, CA, USA, 1999, pp.175-184.
[9] Gediminas Adomavicius, Alexander Tuzhilin, “User Profiling in Personalization Applications through Rule Discovery and Validation”, KDD-99 San Diego, CA, USA, 1999, pp.377-381.
[10] Jong Soo Park, Ming-Syan Chen, Philip S. Yu, “Data Mining：An overview from Database Perspective”, IEEE Trans. on Knowledge and Data Engineering. December 1996, Vol. 8, No. 6, pp. 866-883.
[11] John Shawe-Taylor, Keith Howker and Peter Burge, “Detection of Fraud in Mobile Telecommunications”, Information Security Technical Report, Vol. 4, No. 1, 1999, pp.16-28.
[12] J. Ross Quinlan, “Induction of Decision Trees”, Machine Learning, 1986, pp.81-106.
[13] J. Ross Quinlan, “Simplifying Decision Trees”, Man-Machine Studies, 1987, pp.221-234.
[14] P Burge, J Shawe-Taylor, C Cooke, Y Moreau, B Preneel, C Stoermann, “Fraud Detection and Management in Mobile Telecommunications Networks”, European Conference on Security and Detection, April 1997, Conference Publication No. 437, pp. 91-96.
[15] Saharon Rosset, Uzi Murad, Einat Neumann, Yizhak Idan, Gadi Pinkas, “Discovery of Fraud Rules for Telecommunications－Challenges and Solutions”, KDD-99 San Diego, CA, USA, 1999, pp.409-413.
[16] Tom Fawcett, Foster Provost, “Combining Data Mining and Machine Learning for Effective User Profiling”, NYNEX Science and Technology, 1994.
[17] Tom Fawcett, Foster Provost, “Activity Monitoring : Noticing interesting changes in behavior”, KDD-99 San Diego, CA, USA, 1999, pp.53-62.
[18] Tom Fawcett, Foster Provost, “Adaptive Fraud Detection”, Data Mining and Knowledge Discovery, 1997, pp.1-28.
[19] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, COMMUNICATIONS OF THE ACM. November 1996, Vol. 39, No. 11, pp. 27-34.
[20] Usama Fayyad, “Mining Databases : Towards Algorithms for Knowledge Discovery”, IEEE Computer Society Techinical Committee on Data Engineering, 1998, pp.1-10.
[21] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, “From Data Mining to Knowledge Discovery in Databases”, American Association for Artificial Intelligence, Fall, 1996, pp.37-54.