輔仁管理評論 (專刊)
中華民國 91 年 9 月，第九卷第三期，77-110

應用資料採擷探究電信資料異常之研究

翁頌舜．鄭富山

輔仁大學資訊管理學系

(收稿日期：91 年 5 月 9 日；第一次修正：91 年 6 月 10 日；
接受刊登日期：91 年 7 月 30 日)

摘要

資料採擷 (Data Mining) 是近年來資料庫應用領域中相當熱門的議題。資料採擷一般是指在資料庫中，利用各種分析方法與技術，將過去企業所累積的大量歷史資料，進行分析、歸納與整合等工作，以粹取出有用的資訊，找出使用者有興趣的樣式 (Interesting Patterns)，提供企業管理階層作為訂定決策的依據。目前，無論是零售業、百貨業、電子商務公司、金融機構、電信業、網站管理或醫學診斷等，都已經逐漸體認到資料採擷的重要性，因此也開始積極從事資料採擷的工作，以為企業創造出真正的價值。然而上述都傾向於從過去大量的歷史資料中去作分析，在現實生活的應用上，有些資訊是需要即時告知管理者。例如：電話盜撥、網路干擾、信用卡盜刷等，藉由即時告知以將損失降至最低；而這些異常的情況可能會經常改變，因此要如何應用資料採擷的技術，來完成一個具有即時性與適應性 (Adaptive) 的系統，便成為本研究主要的目標。本研究以電信資料為實驗環境，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，並利用類神經網路 (Neural Networks) 的技術，將標示出的正常與異常資料當作輸入資料，經由不斷地訓練與學習後，期望能夠準確地找出各種異常的情況，以幫助電信企業管理者做出最佳的決策，為企業謀得最大的利潤。

關鍵詞彙：資料採擷，知識探索，類神經網路，電信欺詐

壹．前言

在資料庫的發展過程中，資料採擷 (Data Mining) 是一個興起不久的研究領域，其主要的目的就是從大量繁雜的資料中，找出有意義且具有代表性的樣式 (Patterns)，以提供有用的資訊給企業管理者。就目前而言，大多數企業的資料庫儲存著各式各樣的資料 (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分析、管理及使用這些資料，甚至對這些資料束手無策。傳統的資料庫雖然對於交易處理具有強大的功能，卻沒有智慧型的分析能力，使用者必須明確告訴資料庫其所要的是什麼。而資料採擷的技術可以過濾這些繁雜的資料，利用智慧型的技術，找出有意義的資訊。因此，許多大企業已經體認到儲存於企業本身大量資料的重要性，進而開始從事資料採擷的工作，而尚未投入資料採擷的企業，也積極地著手準備，希望能儘快的著手進行，以為企業創造出更多的利益與價值。

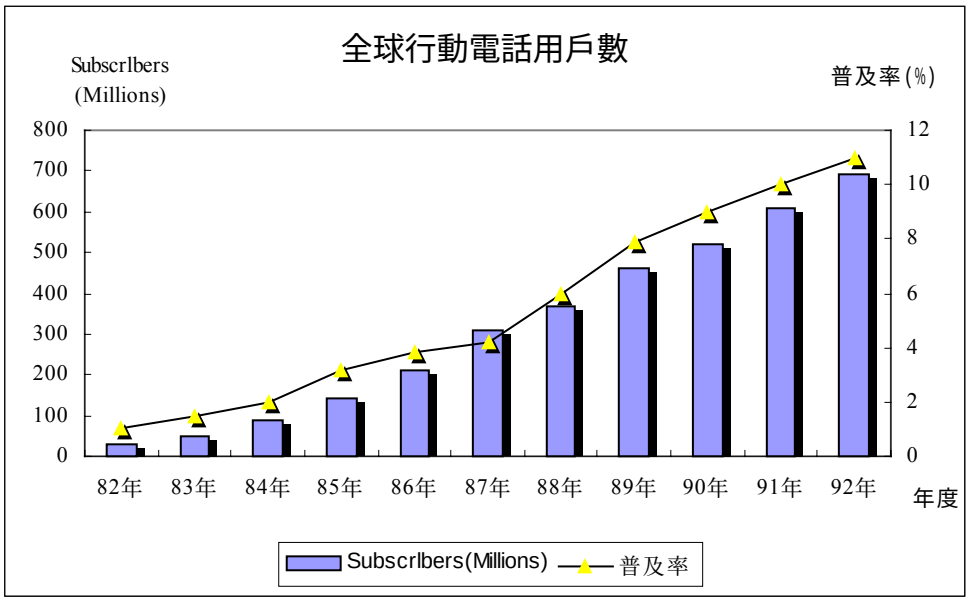
近年來，由於網路的盛行與普及，再加上電信自由化，使得無線通訊已經成為現今最熱門的議題之一，而行動電話也成為無線通訊中競爭最激烈的一項服務，由圖一及圖二 (資料來源：The Strategis Group，1998)，我們可以清楚的了解行動電話服務市場發展的趨勢。過去電信業在行動電話的服務上是屬於類比式通訊，因此僅有少數的資料儲存於資料庫中，但自從國內引進泛歐數位行動電話系統 (Global System for Mobile Communication，簡稱 GSM) 後，便開始從原本的類比式通訊轉變成數位式通訊，因此電信業者可以將用戶所有的通話資料詳細的記錄下來，也造成了電信業資料呈現爆炸性的成長。

過去在使用類比式行動電話時，常有遭人盜拷的情況，非常令人困擾。目前使用泛歐數位行動電話系統的用戶，電信業者發給用戶識別卡 (Subscriber Identity Module，簡稱 SIM)，因此用戶便可輕易攜帶在身上，只要有 SIM 卡隨時都能夠使用任何類型的 GSM 話機，且用戶可以隨時在手機上更改個人密碼 (Personal Identification Number，簡稱 PIN Code)，有效的降低了被攔截盜拷的危險 (劉青儒，1997)。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電話被盜撥的情況。

而除了行動電話被盜撥的問題之外，如何有效的訂定行銷策略，或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，但在電話盜撥方面，卻仍然讓其防不勝防；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或異常的狀態，便成為研究最主要的研究動機。

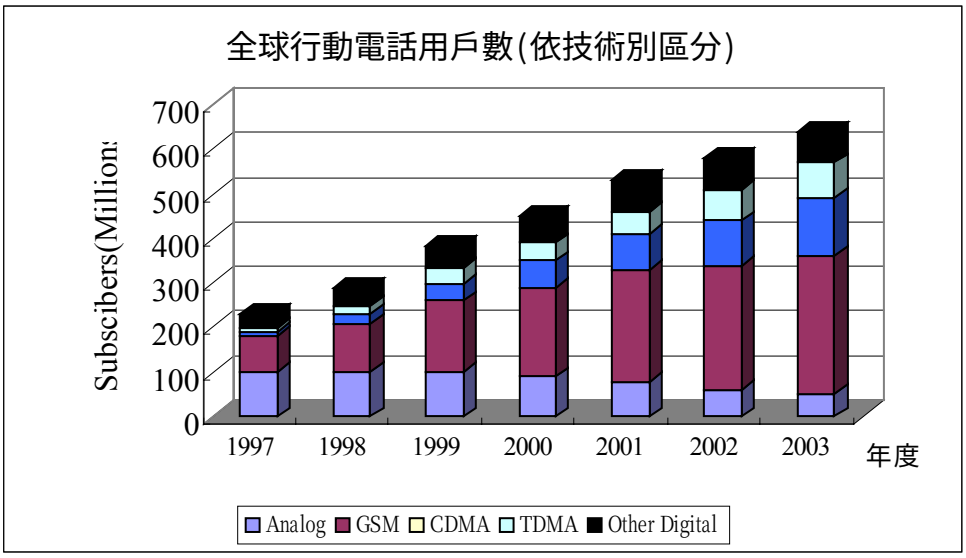
隨著電信市場的蓬勃發展，各家電信經營業者無不卯盡全力爭取最多的客戶與業績，卻也得時時小心保護自己的利潤，因為已有一群不肖之徒，使用各種方式進行電話盜撥 (Telephony Fraud)，進而獲取暴利，而成為全球電信業者的最痛 (賴德謙，1998) 根據估計，全球的無線通訊盜撥的規模大約為美金二十億到三十億元之間。從這個台幣大約將近一千億元的數字來看，可見此一問題的嚴重性。以科技發達的美國為例，該國在歷經多年來與盜撥份子周旋之後，如今也只能把這類無線電犯罪控制在總通話時數 (Airtime) 的百分之零

點五左右 (柳林緯，1999)。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。



(資料來源：The Strategis Group，1998)

圖一　全球行動電話用戶數



(資料來源：The Strategis Group，1998)

圖二　全球行動電話用戶數 (依技術別區分)

根據對這些盜撥行為的分析，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。經分析，大致可以發現下列幾個重要的問題：

1.近年來在電信業中有許多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撥行為的發生，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。

2.由於盜撥行為氾濫，因此會造成基地台及設備佔用的問題，而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。

3.若盜撥行為無法有效解決，將會嚴重影響電信業者的聲譽，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。

4.電信業者必須加派許多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。

5.由於這些硬體設備其功能完全專注於偵測盜撥，而無法針對某些特定的用戶族群去做行為上的分析，以幫助策略上的制定與調整。

除了上述所描述的問題之外，我們常常可以從電視、廣播媒體或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下，如何訂定出一個好的行銷策略，將成為企業致勝的關鍵。從上述第五個問題來看，我們可以知道，要真正去了解用戶的使用行為及特性，才能根據不同的用戶提出不同的行銷方案。因此，本研究期望能夠設計出一個具有適應性與即時性的系統，藉由分析用戶的通話記錄來有效解決這些問題，並期望對於電信業能夠有所貢獻。

在現實生活中，資料採擷在某些特定領域上的應用，是需要即時提供有用的資訊給管理階層，讓管理者能夠隨時得知哪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 (Patterns) 去進行監控與分析，以找出使用者真正想要的資訊。本研究是以電信資料為實驗環境，希望能夠藉由資料採擷的技術，從這些通話記錄中偵測出具有異常行為的資料，以幫助電信企業管理者有目標地針對特定樣式去做分析，進而能夠判斷此特定樣式是否為一異常行為，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)。然而就電話盜撥方面來說，盜撥電話者會使用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

其盜撥的行為；換句話說，即電話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。所以本研究的系統除了即時性之外，還必須具有適應性 (Adaptiveness)，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 (Neural Networks) 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與學習，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電話盜撥的情況及訂定出更好的行銷策略。因此，本研究的主要的目的就是希望能夠設計出一個具有即時性與適應性的系統，將所探索出的異常資訊提供給電信企業管理者進行分析，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。

貳．文獻探討

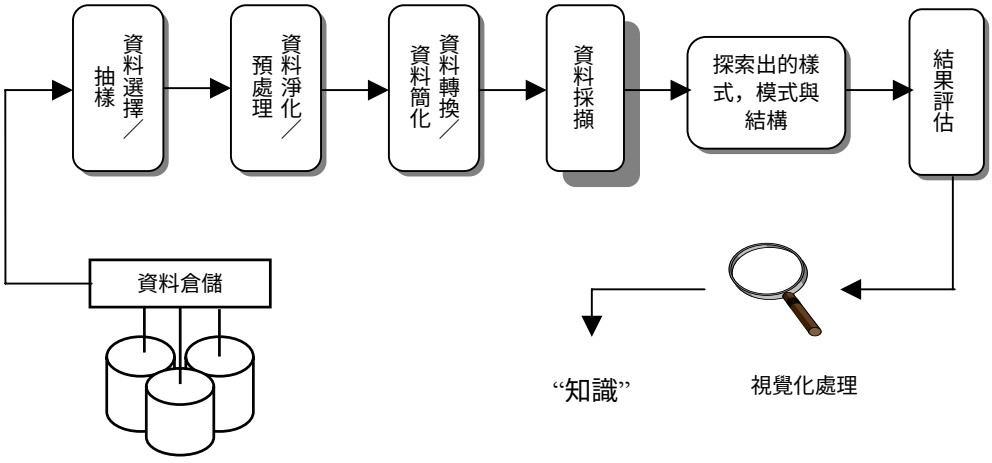
一、資料採擷 (Data Mining)

新的世紀，由於資訊與數位技術的進步，不但徹底改變資料儲存的方式，也造成了我們四周充滿了各式各樣的資料，但卻對它束手無策的窘境。由此可知，傳統人工的分析方式已經無法應付如此龐大的資料，因此我們需要一些新的技術或方法，來幫助我們從這些大量的資料中，粹取出有用的資訊，所以知識探索 (Knowledge Discovery in Databases，簡稱 KDD) 與資料採擷 (Data Mining) 便開始成為一新興的研究領域。

Fayyad et al. (1996) 對知識探索的定義為：「The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data」，其整個流程步驟是：先理解所要應用的領域，熟悉相關知識，接著選擇 (Select) 目標資料集，並專注於所選擇之資料子集；再從目標資料中作前置處理 (Preprocessing)，去除錯誤或不一致的資料；然後作資料簡化 (Reduction) 與轉換工作 (Transformation)；再經由資料採擷的技術與程序探索出資料的樣式 (Patterns) 或找出分類型態；最後經過解釋與評估

(Interpretation and Evaluation) 成為有用的知識。這些程序是一個反覆的過程，重複的進行之後才可以得到一些有用的知識。所以知識探索是一連串的程序，資料採擷是整個程序中的一個核心步驟。由圖三中，我們可更清楚的了解 KDD 與 Data Mining 的關係。

目前資料採擷已被許多研究人員視為結合資料庫系統與機器學習 (Machine Learning) 技術的重要領域，許多產業界人士也認為此領域是一項增加各企業潛能的重要指標。資料採擷既然可以增加企業智慧，提升企業競爭優勢，到底我們應該如何來進行呢?根據 Glymour et al. (1997) 的研究，提出一個參考的進行步驟，即：(一)理解資料與進行的工作，(二)獲取相關知識與技術 (Data Acquisition)，(三)整合與查核資料 (Integration and Checking)，(四)去除錯誤或不一致的資料 (Data Cleaning)，(五)發展模式與假設 (Model Hypothesis and Checking)，(六)實際資料採擷工作，(七)測試與檢核所探索出的資料 (Testing and Verification) 與(八)解釋與使用資料 (Interpretation and Use)。從這八個步驟來看，資料採擷牽涉大量的規劃與準備，而從文獻與實際從事資料倉儲建置或資料採擷工作專家的經驗得知，高達約 80%的過程花在資料準備與資料預處理階段，由此可知，資料採擷只是知識探索過程中的一個核心步驟，要到達資料採擷步驟之前還有許多的工作必須完成。



圖三　知識探索的過程 (資料來源：Fayyad et al.，1996)

由上述的步驟可以清楚地了解知識探索的進行流程，而資料採擷的建置過程則為：(謝邦昌與葉瑞鈴，2000)

1.目標 (Target) 設定：研究及分析現有的商業流程以確認資料採擷可以應用的領域，這些領域可能包括行銷、銷售及顧客服務等。在此階段從事使用者訪談、蒐集資料等工作。其次，將資料按可能使用的模型加以淨化 (Cleaning)、過濾 (Filtering) 與轉換 (Transformation)。在此階段應產生下列各項：(1)有關資料採擷先導計畫實施目的之說明，(2)評估計畫之評估準則，(3)資料整理結果及初步分析報告，(4)計畫時程與 (5)最後目標之大綱。

2.建立雛型 (Prototyping)：從第一階段所獲致結果，使用合適的軟硬體從事系統雛型 (Prototype) 模型之開發。在開發雛型模型的過程中，建立修正資料之整理方法的模型。在本階段應完成下列各項：(1)雛型模型開發系統，(2)資料採擷技術及工具之評估，(3)調整商業流程及資料採擷系統整合之計畫，(4)資料採擷環境設定及開發計畫

3.**系統建置：**在此階段應產生下列各項：(1)資料擷取及整理之程序及軟體，(2)資料採擷模型開發系統，(3)資料採擷上線之第一個版本，(4)解決方案及系統移植計畫 (Migration Plan) 的執行。

4.系統移植 (Migration)：資料採擷文化的建立及使用者的訓練，在此階段應產生下列項目：(1)資料採擷之上線環境，(2)使用者清單，(3)商業流程對資料採擷之回應，(4)系統改善計畫。

二、電信欺詐 (Telecomunication Fraud)

隨著行動電話的普及，再加上通話費率大幅的下降，促成了電信市場的蓬勃發展；各家電信經營業者為了全力爭取最多的客戶及業績，卻疏忽了一些身份查核的工作，而讓電話盜撥者有機可乘，造成企業重大的損失。一般來說，電話盜打最主要可以分為兩種類型，分別是：「技術性盜打 (Cloning)」以及「文件性盜打 (Subscription Fraud)」。所謂技術性盜打，就是利用破解技術或電子方式，傳送出一個可被系統認為有效的帳號 (這也可能是一般合法用戶的帳號)，然後無限制地進行「暢談」。而文件性盜打，則是利用人頭帳號 (可能是偷來的身份證或已經死去的人) 或無效、偽造的身份證件，來申請一個正式有效的帳號，從此坐享「無溝通障礙」的境界 (柳林緯，1999)。由於文件

性盜打很容易在先進國家或戶籍系統發達的地區被查獲 (如：美國)，即使不是如此的地區，也會在呆帳產生之後被業主停話。因此，在亞太地區的盜打情況中，不肖份子主要都是採取技術性盜打的手段。

根據保守估計，無線通訊的盜撥約占業者總體通話費收入的百分之二左右。因此，如果以平均每個用戶通話費八十美元 (大約合 2,560 元新台幣)，每個系統業者共有十萬名用戶來看，這將使系統業者每年短收一百九十二萬美金 (合新台幣大約 6,144 萬元) 的進帳。而這個條件還是建立在十萬戶的情形下，如果用戶更多，那流失的就更大了 (柳林緯，1999)。

由上述的探討中，我們可以清楚地知道盜撥氾濫的情況，因此如何有效的杜絕盜撥，將成為電信經營業者首要的目標。在過去的文獻中曾經提出偵測欺詐行為 (Fraud) 的方式，最主要就是先利用標準法則學習程式 (Standard Rule Learning Program)，從帳單資料中找出具有關聯性的法則，並且從中選擇出具有代表性的法則，其所找出的法則如下所示：(Tom Fawcett and Foster Provost，1997)

(Time-Of-Day = Night) AND (Location = Bronx) → Fraud

Certainty factor = 0.89

有了這些法則之後，便可藉由這些法則來建立出使用者的概要 (User Profiles)，最後再透過不同種類的監測系統 (Monitors) (如 Threshold Monitors 或 Standard Deviation Monitors 等) 來作過濾的動作，如此便能適時地對盜打的帳單資料發出警訊以告知管理者。在這方面的相關研究中，其所擁有的帳單資料或者通話記錄都已經事先知道哪些是屬於正常的情況，哪些是屬於盜撥的情況。而在本研究中所取得的通話記錄，並無法事先確定哪些是屬於正常或異常，因此本研究將提出一個新的方法，希望能夠從這些大量的通話記錄中找出其異常的情況。

三、資訊含量 (Entropy)

熵 (Entropy) 是物理學的一個概念，熵的定律就是所謂「熱力學第二定律」。簡單來說，熱力學有兩個重要的定律。(熱力學還有與低溫問題有關的第三定律。) 「熱力學第一定律」是指：宇宙中的物質與能量的總和是個常數 (即固定不變)，不會增加，也不會減少，只是形式上的改變而沒有本質上的變化。「熱力學第二定律」是指：物質與能量在形式上的變化是一種不可逆轉的

轉換, 由有效轉化到無效的狀態。換句話說, 在閉合系統中熱力增加時, 熵 (混亂度) 也會增加, 故熵總是隨時間增加而保持不變或增加, 閉合系統中熵不可減少 (熵增原理／不可逆原理)。當我們燒一根柴取暖時，其火和柴所產生熱能和光能是正轉向一種不可再轉用的能量發展。能量會轉化, 且向無用的狀態作「單程的轉化」，這大約就是「熵」的意思。(葛納文，2000)

當然熵函數的應用不只是在熱力學方面，另外在醫學、資訊理論、統計學和經濟學等研究領域也都佔有一席之地, 以下我們就來看幾個應用熵函數的例子。根據新科學家雜誌 (New Scientist) 的報導，包括英國劍橋大學 (University of Cambridge)、愛爾蘭的都柏林尖端研究學院 (Dublin Institute for Advanced Studies) 與瑞典 Telia 公司的研究團隊都嚐試以基礎的數學工具，企圖瞭解不知何時何地會出現的網路壅塞問題。他們所用的方法之一就是測量網路交通量的「溫度」變化，也就是它的「熵值」。網路上的伺服器有很多時間都是閒置的, 但是當它們突然接收到或必須送出大量的資料時, 高熵值的狀態使得它們的行為極為不穩定, 而且難以預測。而這些研究人員的目標就是希望能設計一套新的電腦網路運作機制, 避免這種無法控制情況所可能帶來的破壞 (Internet 快訊，1997)。在資料分類 (Classification) 的研究領域中，也有相關的文獻探討過資訊含量，其中最著名的就是如何應用熵函數來建立決策樹 (Decision Trees), 其原理就是利用熵函數來評估這些決策資料中哪一個資料所含的資訊含量最低, 若某個資料其資訊含量最低, 也就是代表它所含有的樣式最少，因此可以用它來作為決策樹的頂點 (J.Ross，1986、1987)。在統計學上的熵指數係由數學家 Shannon and Wiener 所建立，而 Theil 則最先將其應用於經濟分析及預測，故經濟上的熵指數又稱為 Theil 指數，熵指數愈小，代表著區域間的不均衡程度愈小。其熵指數的公式如下所示：

$$I = \sum (Yi/Y)\log[(Yi/Y)/(Fi/N)] \qquad \textbf{(1)}$$

其中 Yi 代表第 i 區國民所得，Y 代表全國的國民所得，Fi 代表第 i 區人口總數，N 代表全國的人口總數。(國立中山大學中國經濟企業研究所，1998)

在資訊理論 (Information Theory) 中，更把熵函數發揮的淋漓盡致，其中最著名的就是影像辨識。利用熵函數, 我們可以判斷一張模糊與一張清楚的影像是否相同。而在本研究中，我們也將熵函數作為演算法的一部份，其基本定義如下：

$$H(P(x)) = -\int P(x)\log_2 P(x)dx \qquad \textbf{(2)}$$

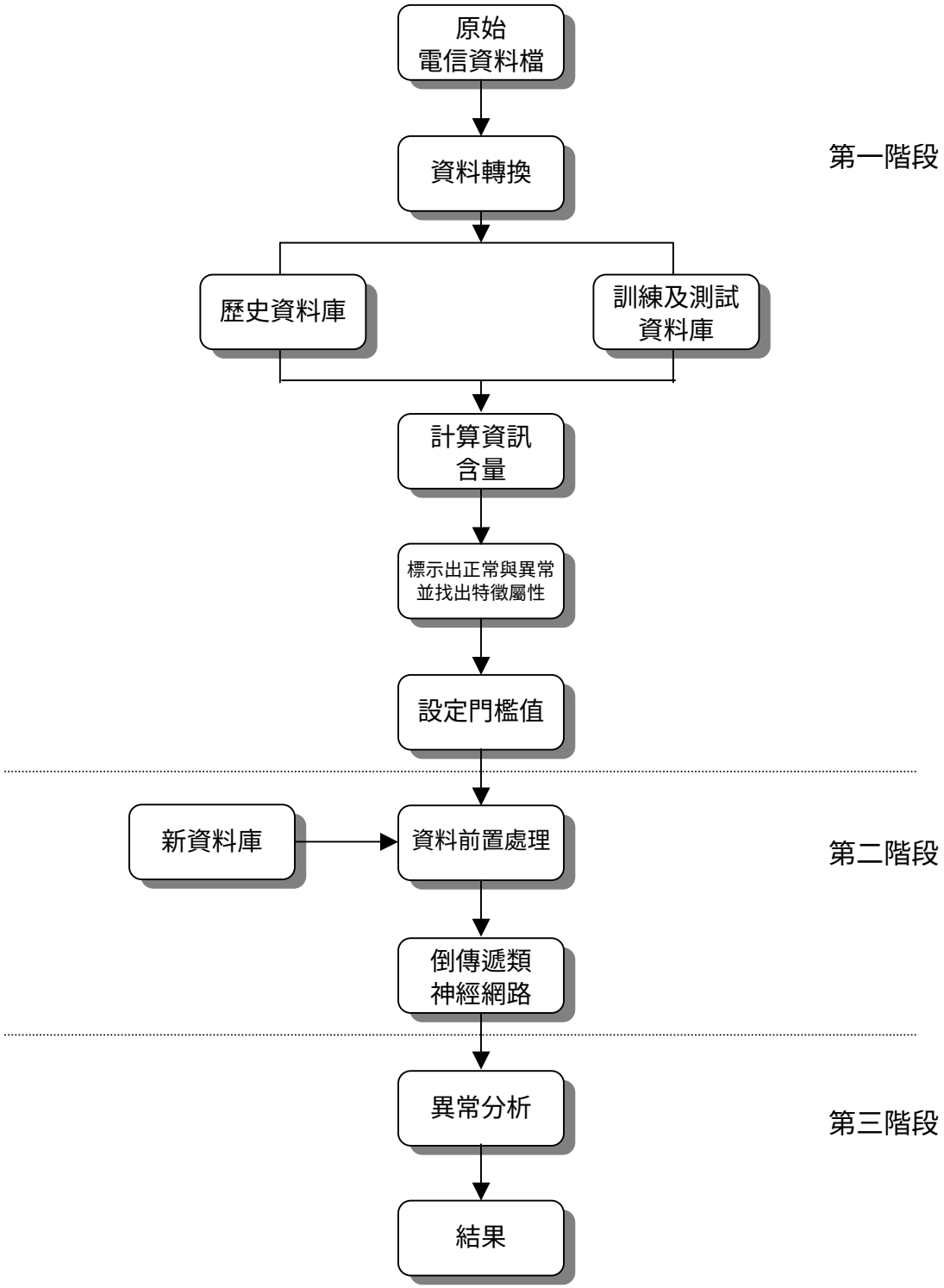
其中 x 代表資料庫中的某一樣式, P(x)代表 x 這個樣式在資料庫中出現的機率, 而 H(P(x))則代表 x 這個樣式在資料庫中的混亂程度。當 H(P(x))的值越大時, 則表示資料庫中的樣式越多, 也就是說此資料庫越混亂；反之, 當 H(P(x))的值越小時, 則表示資料庫中的樣式越少, 也就是說此資料庫越不混亂。所以在本研究中, 我們將把熵函數所計算出的數值稱之為資訊含量。

參．研究方法

從過去的許多研究中, 我們可以發現, 無論在醫學、工業工程或影像辨識等領域裡, 都少不了人工智慧 (Artificial Intelligence, 簡稱 AI) 技術的應用, 人工智慧的技術在許多研究領域中, 已經扮演一相當重要的角色。本研究也將藉由人工智慧強大的運算與分析能力, 來幫助偵測出電信資料的異常。而在這裡所指的異常並不完全就是電信盜撥的行為, 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。本文將探討如何有效地運用人工智慧中的類神經網路技術, 來作為本系統的核心架構, 並且將描述如何分析及確認異常行為的情況, 以便能夠更精確地發出異常的警告給電信企業管理者, 讓管理者可以更有目標的針對問題點加以解決。

本研究的實驗資料為某電信公司提供的二十萬筆電信資料, 由於這些資料可能會牽涉到個人的隱私權問題, 因此本研究已經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, 其前四碼本研究將分別以 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)。取得這些電信資料後, 本研究便著手分析資料的特性, 判斷其資料型態適用於何種分析工具, 並歸納整理出過去研究不足的地方, 進而提出本研究的理論基礎, 設計出本研究的系統架構與流程, 且選用適當的工具進行分析與探討。本研究的系統實作部份是使用 BCB (Borland C++ Builder) 為系統主要的開發工具, 並且設計出一個互動式的介面, 讓不同的管理者能夠依照其不同的需求, 找出其所需要的異常區間；接著本研究將針對此系統所產生的結果進行更深入的分析, 以判斷此結果是否能夠有效地達成目標, 並評估其正確性。

本研究最主要是利用資料採擷的精神，針對電信資料進行分析，藉以從這些電信資料中找出異常的樣式，以提供管理者作為訂定決策的依據。本研究的系統架構與流程，大致上可歸納為三個階段，如圖四所示



圖四　系統架構與流程

而本研究每個階段的執行步驟分別如下所述：

1.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。

2.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，便可以將新進的通話記錄做分類，以找出異常的情況。

3.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，期望對電信業者有所幫助。

由於 GSM (Global System for Mobile Communication) 系統的引進，我國的電信服務已經從傳統的類比式通訊轉變成數位式通訊，不但解決了社會大眾對行動電話的大量需求，也提供了更高品質的通話服務。由於 SS7 (第七號共通道信號系統) 已經成為電信網路的信號骨幹，而 GSM 系統為了訂定一套信號傳輸的通訊協定以供各國遵循，因此採用了 SS7 通訊協定系統作為電信信號傳輸的標準。在 SS7 訊號系統中，由於訊號格式複雜且繁多，因此本研究從其中選擇了六個最重要的屬性來進行實驗：(1)主叫號碼 (Calling Number)：即撥號者之電話號碼。(2)被叫號碼 (Called Number)：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 SS7 的解碼軟體進行處理，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。如表一所示，即為本研究原始通信資料的型式。

表一　原始資料之樣式 (Length 的單位：秒)

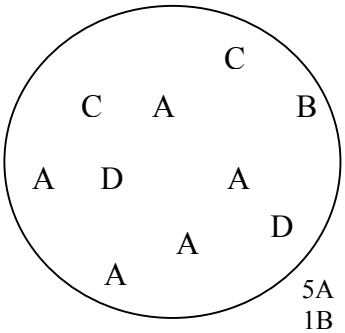
ID	Time	Called	Calling	OPC	DPC	Length
1	2000/11/12 09:13:02	0986123456	0985325812	2000	3000	13
2	2000/11/12 20:32:16	0991876543	0200000001	9000	2000	78
3	2000/11/12 18:11:49	030000001	030000002	4000	2000	157
4	2000/11/12 07:50:05	0995789432	0996190783	2000	3000	53
5	2000/11/12 11:59:59	070000003	0983394857	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
……	……	……	……	……	……	……

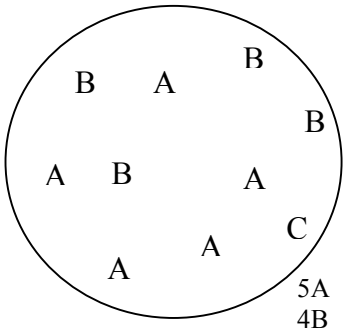
資訊含量 (Entropy) 最主要就是用來評估混亂程度；如果我們把它應用於資料庫中，便可以用來評估整個資料庫的混亂程度。當資訊含量的值越高時，代表了此資料庫越混亂；換句話說，此資料庫所包含的資訊也會越多。反之，當資訊含量的值越低時，代表了此資料庫較不混亂，也可以說此資料庫所包含的資訊較少。利用資訊含量這個函數，我們可以算出某個資料庫的資訊含量。如圖五所示，在資料庫 X 裡面總共有十筆資料，其中包含了資料 A 五筆、資料 B 一筆、資料 C 二筆、及資料 D 二筆。之後便可藉由資訊含量函數，來計算資料庫 X 的資訊含量：

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

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。我們一樣可以藉由 Entropy 函數，來計算資料庫 Y 的資訊含量：

$$H(P(x)) = H(5/10, 4/10, 1/10) = -(5/10 * \log2(5/10) + 4/10 * \log2(4/10) + 1/10 * \log2(1/10))$$
$$= -(-0.5 + (-0.5287) + (-0.3321)) = 1.3608$$





圖五　兩個簡單的資料庫 X 與 Y

由上面兩個資料庫的資訊含量，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。舉例來說，假設資料庫 X 代表了三月份儲存的資料，而資料庫 Y 代表了四月份儲存的資料，由資訊含量這個評估指標我們可以看出，在四月份中其資訊含量突然明顯的降低許多，因此它可能是一個異常的情況，也就是說在四月份的資料庫中，必定隱含了某種特殊的資訊。

本研究將利用資訊含量 (Entropy) 函數，在電信資料庫中發揮其效用，期望能夠有效地標示出那些資料區段為異常的區段。首先，假設本研究擁有一個電信資料庫 Z (如圖六所示)，接著，再將資料庫裡面的各個屬性都視為個別獨立的資料，之後，便可藉由資訊含量 (Entropy) 函數，計算出各個獨立資料之每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。由於本研究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，因此本研究的方式將更能夠表現出其資訊含量的意義。相信利用此方式所標示出來的異常區間，一定隱含了某些本研究所感興趣的訊息。由於本研究是分別算出各個屬性所擁有的資訊含量之後再做加總，因此本研究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，以幫助後續階段做更進一步的分析。

假設表二是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

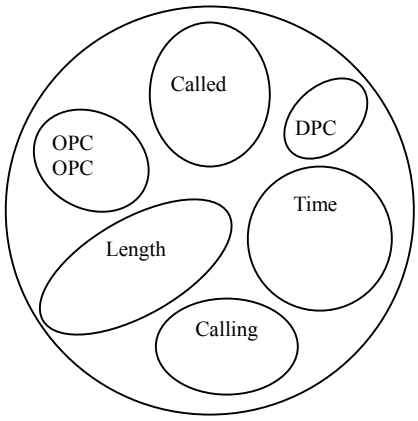
CallingEntropy = H (3 / 10, 1 / 10, 4 / 10, 1 / 10, 1 / 10) = 2.046

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



圖六　電信資料庫 Z

在計算各個屬性的資訊含量時，有幾點必須特別注意。在計算通話長度 (Length) 出現次數的方式與其他屬性的計算方式並不相同，經過仔細分析整個資料庫後，本研究發現，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低，如果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涵義，所以本研究以每 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。以表三為例，假設這是某兩個不同區段的通話長度 (Length) 值，若使用直接計算的方式，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 的資訊含量則為 H (5 / 6, 1 / 6), 所以使用後者比較能夠表現出其資訊含量的差異。

表二　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
……	……	……	……	……	……	……

表三　兩個不同區段的通話長度值

Length		Length
2		7
19		23
30		12
212		29
235		5
62		47
區段 A		區段 B

而在主叫號碼 (Calling Number) 與被叫號碼 (Called Number) 這兩個屬性中, 本研究計算其出現次數的方式與其他屬性也不相同。經過仔細地分析整個資料庫後, 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涵義, 所以本研究以"門號"為單位來做區分 (即行動電話將分成 Y 電信業者、C 電信業者、H 電信業者等；室內電話將分成北部、中部、南部等), 如此便可更準確的表現出其資訊含量的意義。以表四為例, 假設這是兩個不同區段的主叫號碼 (或被叫號碼), 若使用直接計算相同樣式的方式, 則區段 A 與區段 B 的資訊含量都是 H

(1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6, 1 / 6)，但是若以"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)，所以使用後者比較能夠表現出其資訊含量的差異。

表四　兩個不同區段的主叫號碼 (被叫號碼)

Calling (Called)		Calling (Called)
0989027247 (Y 業者)		0982912030 (Y 業者)
0992433001 (Y 業者)		0988278973 (H 業者)
0986743587 (T 業者)		0990646985 (H 業者)
0983926750 (T 業者)		0989238233 (Y 業者)
0995021211 (Y 業者)		0991502388 (Y 業者)
0996987418 (C 業者)		0992840828 (Y 業者)
區段 A		區段 B

求得各個屬性的資訊含量之後，接著將上述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，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

以下為本研究所提出的演算法詳細步驟的描述：在本研究架構的第一階段中，最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：

由於本研究最主要的目的是要分析電信資料的異常情況，因此如何有效找出異常區間便成為最主要工作之一，本研究必須找出異常區間後，才能繼續進行下一步驟的分析。而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。

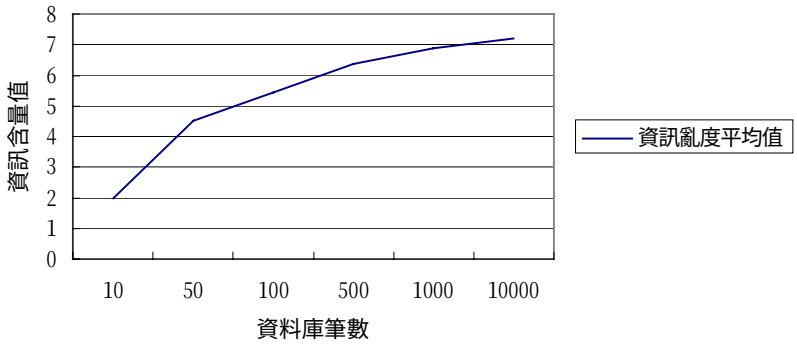
在計算資訊含量的過程中，本研究發現，當歷史資料庫的資料筆數越多時，所計算出來的資訊含量也會越大。如圖七所示，我們可以清楚的看出資訊含量與資料量的增加呈現正比的情況 (即當資料量增加時，資訊含量也會跟著增加)，不過當資料筆數多到一定的數量時，雖然資訊含量仍然會繼續成長，

但是其增加速度逐漸趨於緩和, 這也表示若使用歷史資料的資訊含量來當成評估指標，當歷史資料量大到某一程度時，其平均值也會呈現穩定的狀態。

本研究利用統計學的概念來計算整個歷史資料庫的平均值及標準差。假設在歷史資料庫中全部的資料量為 N 筆 (或 N 個區段), 而 X1、X2…Xn 分別代表每一筆 (或每個區段) 資料的資訊含量, 因此歷史資料庫資訊含量的平均值($\overline{X}$)及標準差(S)分別為：

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{N} \quad \textbf{(3)}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X - X_i)^2}{N}} \quad \textbf{(4)}$$



圖七　資訊含量與資料庫筆數之比較圖

而歷史資料庫將會不斷的增加，我們要如何將新增的資料加入歷史資料庫中呢？此時可利用統計學中加總的公式來計算, 以節省加入資料後重新計算資訊含量的時間。其平均值(μ)與變異數(σ^2)的加總公式如下所示，其中 N_i、X_i 及 σ_i^2 分別代表各個資料庫的數量、平均值及變異數。

$$\mu = \frac{\sum_{i=1}^{n} N_i \overline{X_i}}{\sum_{i=1}^{n} N_i} \quad \textbf{(5)}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} N_i[\sigma_i^2 + (\overline{X_i} - \mu)^2]}{\sum_{i=1}^{n} N_i} \quad \textbf{(6)}$$

(二)標示出異常區間及特徵屬性：

當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。

找出異常區間之後，本研究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。本研究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。而要如何找出影響整個資料庫的特徵屬性呢？本研究仍利用平均值與標準差的概念，當在求取整個歷史資料庫的平均值與標準差時，也一併找出各個屬性資訊含量的平均值與標準差，接著便可利用它來找出特徵屬性。假設六個屬性其資訊含量的平均值與標準差分別如表五所示，若某異常區段的資訊含量大於 3.5，我們即可以從六個屬性中，找出其個別的資訊含量是否也大於其平均值加上標準差，如果是，則表示此屬性就是影響資料庫為異常的特徵屬性；反之，若某異常區段的資訊含量小於 2.5，我們可從六個屬性中，找出其個別的資訊含量是否也小於其平均值減去標準差，如果是，則表示此屬性就是影響資料庫為異常的特徵屬性。

表五　各屬性的平均值與標準差

屬性名稱	平均值	標準差
Time	0.486	0.021
Called	0.668	0.036
Calling	0.734	0.042
OPC	0.268	0.016
DPC	0.302	0.018
Length	0.528	0.024

為了能夠更清楚的說明本研究的做法，以表六為例，其為某一現有的電信資料庫，本研究根據歷史資料庫的平均值及標準差標示出其異常的區間，接著就可以分析這些異常的原因是其資訊含量大於或小於它的容忍範圍 (即大於 3.5 或小於 2.5)。從表六中我們可以發現第 2、3 區段其異常的原因是它的資訊含量大於 3.5，因此我們可以知道在其中的六個屬性中，必定有某些屬性造成它的異常，在區段 2 中我們可以找出 Calling Number 及 DPC 這兩個屬性大於其容忍範圍，而在區段 3 中我們可以找出 Time 及 Length 這兩個屬性大於其容忍範圍 (與表五做比對)，因此分別把它們標示為特徵屬性，以作為後續類神經網路分析，設定正常與異常的依據。而其他的異常區段也可以依此類推，標示出其特徵屬性。

表六　某一電信資料庫異常區間的標示結果 (平均值為 3，標準差為 0.5)

編號	區間	各屬性的資訊含量						正常或異常	特徵屬性
		Time	Called	Calling	DPC	OPC	Length		
1	001-100	0.482	0.674	0.693	0.260	0.287	0.551	正常	NULL
2	101-200	0.502	0.694	0.983	0.487	0.316	0.549	異常	Calling DPC
3	201-300	0.746	0.682	0.774	0.282	0.309	0.793	異常	Time Length
4	301-400	0.467	0.682	0.754	0.283	0.301	0.543	正常	NULL
5	401-500	0.502	0.691	0.709	0.267	0.297	0.535	正常	NULL
6	501-600	0.469	0.429	0.503	0.263	0.290	0.515	異常	Calling Called
7	601-700	0.499	0.649	0.698	0.258	0.283	0.509	正常	NULL
8	701-800	0.471	0.642	0.497	0.257	0.293	0.239	異常	Calling Length
9	801-900	0.470	0.701	0.727	0.272	0.297	0.511	正常	NULL
10	901-1000	0.469	0.638	0.754	0.269	0.319	0.527	正常	NULL
……	……	……	……	……	……	……	……	……	……

(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 (Threshold) 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。而門檻值所代表的意義，也可以說是調整標準差的大小，例如前述例子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，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。

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，要如何對這些多而繁雜的資料進行處理與分析呢？本研究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，本研究以類神經網路作為資料分類的主要工具。

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，所以不能夠僅以過去的歷史資料當作評估的資訊，必須每隔一段期間就重新訓練，讓系統能夠不斷地學習新的規則，如此才能夠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。

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。經過本研究使用不同的方式測試後，本研究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**：在時間 (Time) 這個屬性中，本研究以每三小時為單位，將時間區段分成八個區間，並且分別以八個輸入節點 (Node) 來表示 (如表七所示)。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. **Calling 與 Called**：在主叫號碼 (Calling) 與被叫號碼 (Called) 這兩個屬性中，若是行動電話，本研究以它的門號所屬之電信公司來作為分類，若是室內電話本研究則以它的地區來作為分類，而免付費電話與呼叫器號碼，則把它歸類於其他，在此總共分成十個類別，並分別以十個輸入節點來代表 (如表八所示)。

表七　時間屬性的輸入轉換 (單位：小時)

時間	Node1	Node2	Node3	Node4	Node5	Node6	Node7	Node8
00-03	1	0	0	0	0	0	0	0
03-06	0	1	0	0	0	0	0	0
06-09	0	0	1	0	0	0	0	0
09-12	0	0	0	1	0	0	0	0
12-15	0	0	0	0	1	0	0	0
15-18	0	0	0	0	0	1	0	0
18-21	0	0	0	0	0	0	1	0
21-24	0	0	0	0	0	0	0	1

表八　主叫號碼與被叫號碼的輸入轉換

Called (Calling)	Node1	Node2	Node3	Node4	Node5	Node6	Node7	Node8	Node9	Node10
C 業者	1	0	0	0	0	0	0	0	0	0
H 業者	0	1	0	0	0	0	0	0	0	0
Y 業者	0	0	1	0	0	0	0	0	0	0
T 業者	0	0	0	1	0	0	0	0	0	0
TS 業者	0	0	0	0	1	0	0	0	0	0
F 業者	0	0	0	0	0	1	0	0	0	0
室內 (北部)	0	0	0	0	0	0	1	0	0	0
室內 (中部)	0	0	0	0	0	0	0	1	0	0
室內 (南部)	0	0	0	0	0	0	0	0	1	0
其他	0	0	0	0	0	0	0	0	0	1

3.Length：在通話長度 (Length) 這個屬性中，本研究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 (如表九所示)。

4.OPC 與 DPC：在主叫地域 (OPC) 與被叫地域 (DPC) 這兩個屬性中，由於本研究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。經過詳細分析這些資料後，本研究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 (如表十所示)。

表九　通話長度的輸入轉換 (單位：秒)

Length	Node1	Node2	Node3
0-120	1	0	0
120-300	0	1	0
300 以上	0	0	1

表十　主叫地域與被叫地域的輸入轉換

OPC (DPC)	Node1	Node2	Node3	Node4
4080	1	0	0	0
4300	0	1	0	0
9160	0	0	1	0
其他	0	0	0	1

表十一　某異常區段標示之異常記錄

ID	Time	Called	Calling	DPC	OPC	Length	正常或異常
1	2000/11/12 03:13:01	0986128432	0996087113	4300	9156	39	0 1
2	2000/11/12 03:13:01	040000019	0984139877	4080	4080	125	1 0
3	2000/11/12 03:13:01	0992974878	0995123441	4300	4080	67	1 0
4	2000/11/12 03:13:25	0989889210	0995872091	4300	9156	92	1 0
5	2000/11/12 03:13:39	040000020	0989489157	4080	4080	87	1 0
6	2000/11/12 03:13:58	0984190890	040000021	4080	4080	44	1 0
7	2000/11/12 03:14:27	030000022	0983335309	4080	4080	13	0 1
8	2000/11/12 03:14:27	0997145002	0996087113	4300	4080	390	0 1
9	2000/11/12 03:14:27	0999556909	0989489157	4300	9156	47	0 1
10	2000/11/12 03:14:46	060000023	0989489157	4080	4300	112	1 0

當找出異常區段後，要如何將之標示出異常的記錄呢？為了要能夠增加訓練的正確性，本研究不能直接將整個異常區段中的記錄都標示成異常，必須要技巧性的來標示。前述已經詳細描述如何找出異常區間以及標示出其特徵屬

性, 接著就利用這些特徵屬性來幫助標示出異常的記錄。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, 此時就必須找出「出現次數少以及出現頻率高」的記錄, 將它標示為異常。反之, 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, 就必須找出「出現次數多以及出現頻率低」的記錄, 將它標示成異常。以表十一為例, 其為某一個異常區段的詳細資料, 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, 於是對記錄屬性 Time 及 Calling 出現次數的陣列來做判斷, 我們從表十二及表十三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, 因此就針對這些資料, 把它標示成異常 (如表十一所示)。

表十二　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

表十三　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, 假設表十四為某一個異常區段的詳細資料, 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, 從表十五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 (如表十四所示)。

表十四　某異常區段標示之異常記錄

ID	Time	Called	Calling	DPC	OPC	Length	正常或異常
1	2000/11/12 03:13:01	0986128432	0996087113	4300	4080	39	1 0
2	2000/11/12 03:13:03	040000025	0984139877	4080	4080	125	1 0
3	2000/11/12 03:13:13	0992974878	0995123441	4300	4080	67	1 0
4	2000/11/12 03:13:25	0989889210	0995872091	4300	4400	92	0 1
5	2000/11/12 03:13:39	040000026	0989489157	4080	4080	87	1 0
6	2000/11/12 03:13:58	0984190890	040000027	4080	4080	44	1 0
7	2000/11/12 03:14:11	030000028	0983335309	4080	4080	13	1 0
8	2000/11/12 03:14:27	0997145002	0996087113	4300	4080	390	1 0
9	2000/11/12 03:14:32	0999556909	0989489157	4300	9156	47	0 1
10	2000/11/12 03:14:46	060000029	0989489157	4080	4300	112	0 1

表十五　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

肆．實證研究與結果分析

在實證研究的部分，本研究最主要可分為三大部分：全部資料庫分析、特定族群分析與及單一樣式分析。本研究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。以下分別介紹三種不同的分析方式：

一、全部資料庫之分析：

在全部資料庫的分析中，本研究先從目前所擁有的電信資料庫中取出 54000 筆通話記錄來作為實驗的資料。根據本研究的需求，我們將這些資料分成三部份：即歷史資料庫、訓練資料庫與測試資料庫。

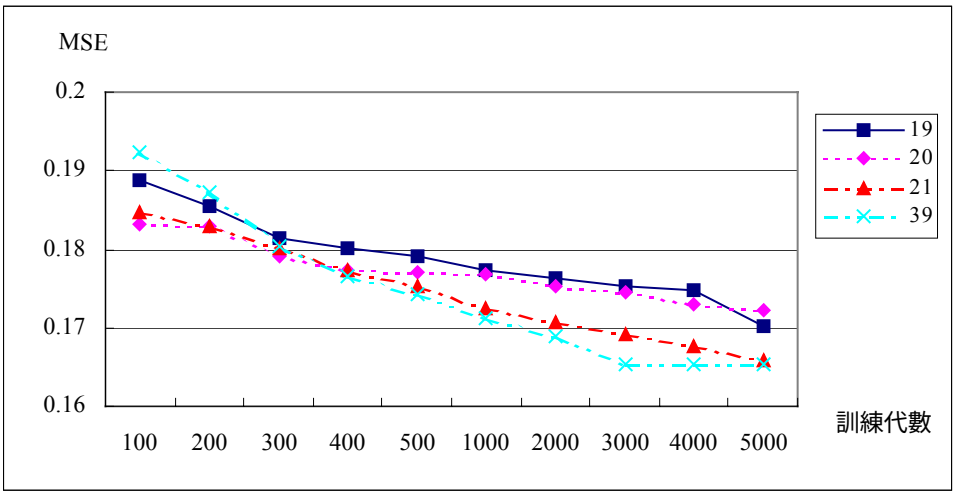
當準備好所需的資料後，本研究就可以利用歷史資料庫來計算其資訊含量的平均值及標準差，本研究以 50 筆記錄作為一個區間長度，計算出來的平均值為 21.7255，標準差為 2.0033。有了歷史資料庫資訊含量的平均值及標準差後，就可以利用它來評估訓練資料庫及測試資料庫的異常區間，如果其資訊

含量大於 23.7288 (21.7255 + 2.0033)，或者小於 19.7222 (21.7255 – 2.0033)，便把它視為異常，並且將造成此區段異常的屬性標示出來，當作其特徵屬性。經由本系統的計算後，在訓練資料庫中共找出了 10 個異常區間，而在測試資料庫中共找出了 5 個異常區間。

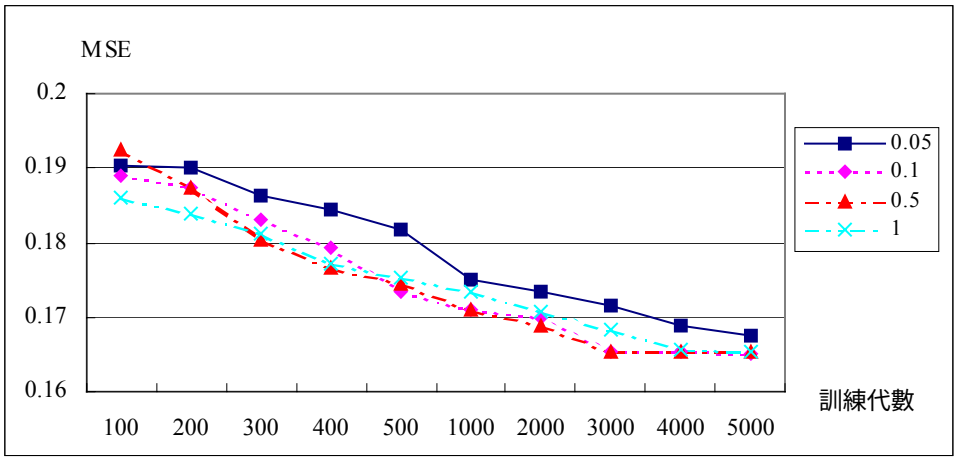
當有了正常與異常的區段後，接著就可以透過本系統，將其轉換成特定格式的訓練資料檔及測試資料檔，並標示出哪些記錄是正常或異常，然後便可繼續進行下一步驟，即類神經網路的訓練與測試。在訓練資料檔中，我們共選擇了 20 個區段的資料 (即 1000 筆記錄) 來作為訓練的樣本，其中包含了 10 個正常區段及 10 個異常區段；而在測試資料檔中，我們共選擇了 10 個區段的資料 (即 500 筆記錄) 來作為測試的樣本，其中包含了 5 個正常區段及 5 個異常區段。

在全部資料庫的實驗中，本研究選擇了六個屬性 (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：即正常或異常。至於隱藏層節點個數的選擇，經過本研究對不同節點個數的隱藏層訓練之後，可以發現當隱藏層節點個數為 39 時，有較好的收斂結果 (如圖八(a)所示)，接著本研究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 (如圖八(b)所示)，因此本研究選擇 39 個隱藏層節點與學習因子為 0.5 來作為本研究全部資料庫分析類神經網路的架構。

當全部資料庫分析的類神經網路架構設計且訓練完成之後，接著就可以將測試資料檔輸入進行測試的工作；當測試完成之後，本研究發現在全部資料庫的實驗中，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。接著本研究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，本研究發現在區段 2、區段 4 及區段 10 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。本研究也發現在區段 6 及區段 8 中，其 Time 屬性的資訊含量突然地減少，而造成這兩個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。



圖八　不同隱藏層個數與不同學習速率之比較圖
(a)不同隱藏層的測試



圖八　不同隱藏層個數與不同學習速率之比較圖
(b)不同學習速率的測試

二、特定族群之分析：

在特定族群的分析中，本研究將針對兩個不同的族群組合來做分析，首先先利用 SQL 的指令，將所選擇的特定族群屬性從資料庫中挑選出來，並把它儲存在一個新的資料表中。首先，本研究將針對 Time =“半夜”與 Length =“全部” (全部是指包含所有的通話長度) 這兩個特定族群來做分析，總共從資料庫中取出 4500 筆記錄。根據需求，本研究將這些資料分成三部份：即歷史資料庫、訓練資料庫與測試資料庫。

當準備好所需的資料後，本研究便可以利用歷史資料庫來算出其資訊含量的平均值及標準差，本研究以 50 筆記錄為一個區間長度，而計算出來的平均值為 10.5488，標準差為 0.1503。有了歷史資料庫資訊含量的平均值及標準差後，就可以利用它來評估訓練資料庫及測試資料庫的異常區間，如果其資訊含量大於 10.6991 (10.5488 + 0.1503)，或者小於 10.3985 (10.5488 – 0.1503)，便把它視為異常，並且將造成此區段異常的屬性標示出來當作其特徵屬性。經由本系統的計算後，在訓練資料庫中共找出了 6 個異常區間，而在測試資料庫中共找出了 3 個異常區間。

有了正常與異常的區段後，接著就可以透過本系統將其轉換成特定格式的訓練資料檔及測試資料檔，並標示出哪些記錄是正常或異常，然後便可繼續進行類神經網路的訓練與測試。在訓練資料檔中，本研究共選擇了 12 個區段的資料 (即 600 筆記錄) 來作為訓練的樣本，其中包含了 6 個正常區段及 6 個異常區段；而在測試資料檔中，本研究共選擇了 6 個區段的資料 (即 300 筆記錄) 來作為測試的樣本，其中包含了 3 個正常區段及 3 個異常區段。

在屬性 Called =“行動電話”與屬性 Calling =“行動電話”這兩個特定族群的實驗中，本研究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，經過對不同節點個數的隱藏層訓練之後，可以發現當隱藏層節點個數為 19 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，因此本研究選擇 19 個隱藏層與學習因子為 0.1 來作為類神經網路的架構。

當這兩個特定族群 (Called =“行動電話”及 Calling =“行動電話”) 的類神經網路架構設計且訓練完成之後，接著就可以將測試資料檔輸入進行測試的工作；當測試完成之後，本研究發現在特定族群的實驗中，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。接著本研究針對所找出的異常記錄加以分析，根據原始資料的顯示，可以發現在區段 1 及區段 3 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。而在區段 5 中，其 Called 屬性的資訊含量突然地降低，經過分析後，可以發現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，可以發現此公司的用戶常常與 A 電信公司的用戶有通話的往來，所以此公司也許可以找 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。

三、單一樣式之分析：

在單一樣式的分析中，與上述的實驗方法類似，根據結果顯示，本研究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該電信公司的門號，因此便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。本研究也發現在某區段中，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，因為這些通話可能是一個盜撥的情況，或者是因為此用戶離開了資料所在地區所造成的結果。

伍．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生活不可或缺的使用項目。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。

在國內將資料採擷的技術運用在電信資料方面的相關研究尚未普及，因此本研究希望能夠藉由分析龐大的電信通話記錄，進而找出其異常的部分，再經由特徵屬性來幫助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。本研究經由不同的分析方式來驗證本研究方法的可行性，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，相信對於偵測資料異常方面的相關研究應該會有所幫助。

由於資料採擷應用於電信資料方面的研究尚未普及，再加上本研究所取得的資料有限，因此尚有許多值得研究與改善的空間。分別如下所述：

一、增加用戶的詳細資料：

由於本研究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；如果可以增加用戶的詳細資料，相信對於資料異常方面的分析一定會有更大的幫助 (例如：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)。

二、增加電信資料的相關屬性：

在我們的研究中，本研究只針對 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，因此還有很多屬性必須加入分析，如果能夠增加更多有用的屬性來分析，相信對於電信方面的研究一定會有更多的貢獻。

三、增加國際電話的通話記錄：

本研究所取得的電信資料中並沒有包含國際電話，而我們知道大多數電話盜撥的情況都發生於盜撥國際電話，因此如果可以增加國際電話的通話記錄，相信對於盜撥方面的異常偵測一定會更有幫助的。

四、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，本研究期望未來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。

五、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，必須針對六個不同的屬性分別計算，因此會多次的掃描資料庫，期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

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，頁 128-132。

柳林緯，「淺談 GSM 行動電話標準」，*台灣通訊雜*誌，1998 年，頁 118-123。

劉青儒，「GSM 數位行動電話的現況與展望」，*新電子期刊*，1997 年，頁 101-108。

賴德謙，「電信經營業者的痛－電話盜撥」，*台灣通訊雜誌*，1998 年，頁 92-97。

謝邦昌、葉瑞鈴，「統計在資料探勘之應用」，*主計月報*，第 530 期，2000 年，頁 67-84。

Adomavicius, G.and A. Tuzhilin, "User Profiling in Personalization Applications through Rule Discovery and Validation", *KDD-99*, San Diego, CA, U.S.A., 1999, pp.377-381.

Berry, M. J. A. and G. Linoff, "Data Mining Techniques: for Marketing, Sales, and Customer Support", New York: John Wiley & Sons, 1997.

Bonchi, F., F. Giannotte, G. Mainetto and D. Pedreschi, "A Classification-Based Methodology for Planning Audit Strategies in Fraud Detection", KDD-99, San Diego, CA, U.S.A., 1999, pp.175-184.

Boukerche A., M. Sechi and M. A. Notare, "Neural Fraud Detection in Mobile Phone Operations", IPDPS 2000 Workshops, 2000, pp.636-644.

Burge, P., 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*, No.437, April 1997, pp.91-96.

Cabena, P. "Discovering Data Mining from Concept to Implementation", NJ: Prentice Hall, 1997.

Chen, M. S., J. Han and P. S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Trans. Knowledge and Data Engineering*, 8, 1996, pp.886-883.

Cleveland, W. "Visualizing Data", Summit, NJ: Hobart Press, 1993.

Devore, J. L. "Probability and Statistics for Engineering and the Science", 4th ed. New York: Duxbury Press, 1995.

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", *American Association for Artificial Intelligenc*e, Fall, 1996, pp.37-54.

Fayyad, U. M., G. Piatetsky-Shapiro, P.Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining", Cambridge, MA: AAAI/MIT Press, 1996.

Fayyad, U.M., "Mining Databases: Towards Algorithms for Knowledge Discovery", *IEEE Computer Society Technical Committee on Data Engineering*, pp.1-10, 1998.

Fawcett, T. and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behavior", KDD-99, San Diego, CA, U.S.A., 1999, pp.53-62.

Fawcett, T. and F. Provost, "Adaptive Fraud Detection", *Data Mining and Knowledge Discovery*, 1997, pp.1-28.

Glymour C., D. Madigan, D. Pregibon and P. Smyth, "Statistical Themes and Lessons for Data Mining", *Data Mining and Knowledge Discovery*, 1(1), 1997, pp.11-28.

Han, J. "Towards On-Line Analytical Mining in Large Databases", *SIGMOD Record*, 27(1), 1998, pp.97-107.

Han, J., Y. Cai and N. Cercone, "Data-driven Discovery of Quantitative rules in Relational Databases", *IEEE Trans. Knowledge and Data Engineering*, 5, 1993, pp.29-40.

Han, J. and Y. Fu, "Exploration of the Power of Attribute-Oriented Induction in Data Mining", In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, Cambridge, MA: AAAI/MIT Press, 1996, pp.399-421.

Han, J. and Y. Fu, "Mining Multiple-Level Association Rules in Large Databases", *TKDE*, 11(5), 1999, pp.798-804.

Kennedy, R. L, Y. Lee, B. Van Roy, C. D. Reed, and R. P. Lippman, "Solving Data Mining Problems Through Pattern Recognition", *Upper Saddle River*, NJ: Prentice Hall, 1998.

Piatetsky-Shapiro, G. and W. J. Frawley, "Knowledge Discovery in Databases", Cambridge, MA: AAAI/MIT Press, 1991.

Pyle, D. "Data Preparation for Data Mining", San Francisco: Morgan Kaufmann, 1999.

Quinlan, J. R., "Induction of Decision Trees", *Machine Learning*, 1986, pp.81-106.

Quinlan, J. R., “Simplifying Decision Trees”, Man-Machine Studies, pp.221-234, 1987.

Ross, S. M., "A Course in Simulation", Maxwell Macmillan, New York, 1990.

Rosset, S., U. Murad, E. Neumann, Y. Idan, and G. Pinkas, "Discovery of Fraud Rules for Telecommunications-Challenges and Solutions", *KDD-99*, San Diego, CA, U.S.A., 1999, pp.409-413.

Shawe-Taylor, J., K. Howker and P. Burge, "Detection of Fraud in Mobile Telecommunications", *Information Security Technical Report*, Vol.4, No.1, 1999, pp.16-28.

Using Data Mining to Analyze Abnormal Data in Telecommunications

SUNG-SHUN WENG, FU-SHAN CHENG

Department of Information Management, Fu Jen Catholic University

ABSTRACT

In recent years, data mining is one of the top issues in the field of database applications. Data mining generally means that it utilizes various kinds of methods and techniques to mine data. It analyzes, generalizes, and integrates the past, accumulated and large quantity of historical information to find out the interesting patterns and pick out useful information as the basis of decision making processes for business executives. No matter in categories of retailing, electronic commerce, finance, telecommunications, web management, medical diagnosis, or others, people have already recognized the importance of data mining gradually. Therefore, they begin to dedicate to data mining aggressively for creating the real values of the enterprises.

However, as stated above, data mining tends to analyze the large quantity of historical data. But in order to apply it in the real world, some information, such as telephone frauds, network interruption, credit fraud and so on, is needed to let the company know in time for minimizing the possible loss. But these abnormal situations may change frequently. How to apply data mining techniques to develop a real time and adaptive system is the main goal of this thesis.

This research is based on the telecommunication data and uses the “Entropy” theory of Thermodynamics as the main guide for appraising the information capacity in the databases. We use the marked normal and abnormal data as the input of neural networks. Through the iterative process of training and learning of neural networks, we wish to find out abnormal situations precisely in order to help the business executives making the best strategy to earn the maximum profits for enterprises.

Keywords: data mining, knowledge discovery, neural networks, telecommunication fraud