東海大學資訊工程學系研究所

碩士論文

指導教授：許玟斌 博士

植基於規則推導法探勘潛在企業客戶流失之研究

A study in mining potential lose of enterprise customers based on rule induction

研 究 生：林志欣

中華民國一百零二年六月

摘要

台灣的電信業自從民國 85 年政府啟動電信自由化後，使得原本由中華電信獨佔的電信市場變成了競爭激烈的寡佔市場，在電信市場中，電信費用平均的支出上企業客戶明顯的比個人用戶高出許多。因此，當電信業者進入市場後想要增加營收最快的方法就是從企業客戶著手。然而國內各電信業者能夠掌握的客戶基本資料相當有限，除了證號、地址、服務合約等資料外，其他諸如客戶背景、人口統計變項等資料付之闕如。所以在實務的做法上，國內的電信業消費行為分析大多採用營收資料庫中的通聯記錄或是帳單消費金額的變化，發展企業客戶流失分析模型，尋找企業客戶流失的徵兆。由於企業客戶可能剛開始申裝電信業務時，同時也申請他家業者的服務，如果依據企業客戶帳單消費金額的增減變化設計客戶流失分析模型，該客戶則有很大有可能成為潛在的流失客戶。本研究以電信業者之中小企業客戶為基礎對象，收集實際資料後運用資料探勘的技巧加以分析。本研究以既有的客戶群加以分類，嘗試以規則推導方法，採用客戶行業別及電信費用等資料屬性，進行潛在流失客戶之分析診斷，協助企業客戶服務團隊有效的掌握客戶動向，進而增裕公司營收。

關鍵字: 電信費用 、資料探勘、規則推導

Abstract

After our government released the licenses of telecommunication industry in 1996, the oligopoly market dominated by Chunghwa Telecom has changed into the competitive monopoly market. Generally average amount of transmission cost of enterprises is much higher than individual customers. Therefore, the fastest way to increase the profit is to target enterprise customers. However, expect the open information such as identification, address and contracts, the other data like customer background and geographic variation is very limited. Practically, most telecommunication providers develop a lost analysis model to search for the lost signal. The model is based on transactions variation of the bills from the consumption behavior data. To discover the potential lost of enterprise customers, the design of analysis model can be based on the variation of telecommunication cost. This is because a customer can apply services from two different suppliers in the beginning. In our research, we focused on small-to-medium enterprise customers, collected relevant data and then analyzed data by data mining technique. Specifically, we classified enterprise customers using rule induction based on attributes such as biz type and transmission cost. The model is suitable in diagnosing the potentially lost customer, In this way the provider service team can understand the customer movements effectively and then increase the profits.

Keyword: Telecommunication cost , Data Mining , Rule Induction

致　謝

在碩士求學的過程雖然辛苦也收穫很多。本篇論文能夠順利的完成，承蒙恩師 許玟斌教授，在論文撰寫中的細心指導與教誨，讓我受益良多，師恩浩蕩，銘感於心，感激之情，難以言表！也感謝實驗室的學長、學姐以及同學們在求學中的相互鼓勵和幫忙。謝謝大家!

林志欣 謹誌於
私立東海大學資訊工程學系所
中華民國一百零二年六月

目錄

圖目錄

表目錄

第一章　前言

我國的電信市場自從開放自由化以來，受到價格及各種促銷活動的影響，使得客戶流失，轉換服務的情況日益嚴重。相對於其他產業來說，電信業是一個典型的前期固定投資巨大且在一定範圍內投資資金多少不受用戶量影響的行業。因此，電信公司擁有的客戶越多，作為主要成本的前期投資就會攤得越薄其利潤就越大。客戶資源對電信公司來說其意義不言而喻，電信業之間的競爭實際上就是對客戶資源的競爭。而在電信市場上，平均每間公司行號花在電信費用上的支出明顯比個人用戶高出許多；因此，當電信業者進入市場後想要增加營收最快的方法就是從企業客戶著手。

在電信業中，企業客戶市場與個人消費市場明顯不同，所以各家電信公司在客戶管理上，通常將客戶區分為大企業客戶、中小企業及一般消費客群；透過不同的行銷管道，達到分眾行銷的目的。企業客戶市場的經營上針對大企業客戶都有指派專門的客服團隊對於其電信使用狀態都相對清楚能夠掌握，而相對於為數眾多的中小企業客戶群的市場經營上，慣例是針對其電信業消費行為上做分析，大多採用營收資料庫中的帳單消費金額的增減變化，發展企業客戶流失分析模型，尋找企業客戶流失的徵兆。而流失的客戶在本文為業務員上的定義指的是客戶同時採用 2 家或以上之通信服務則為流失客戶，如果依據企業客戶帳單消費金額區間的增減變化設計客戶流失分析模型，有可能因其採樣區間的時間上選擇或企業客戶可能剛開始申裝電信業務時，同時也申請他家業者的服務，而產生潛在的流失客戶。

資料探勘對於電信業這個接近飽合的市場裡更為重要。我們希望能夠透過客戶的屬性來幫助我們做分類，並採用規則推導方法中的決策樹推導，透過學習的方式，找出規則，然後對中小企業客戶做預測，透過預測我們可以找出中小企業客戶中的可能潛在流失客戶。

相關研究成果，預計可以提供未來電信業在針對中小企業之潛在流失客戶透過資料探勘方式分析診斷發現以做為業務經營上的相關策略參考，是本研究最大目的。

第二章 相關文獻探討

當電信業者面臨顧客流失問題時，除了強化其現有的顧客關係管理之外，同時也會執行各種可行的顧客流失管理程序，而為協助電信業者解決顧客流失的問題，國內外已有許多的研究相繼的提出 （邱義堂，2001）[1]（毛慧雯，2008）[2]（Wei et al., 2002）[5]。

Wei（Wei et al., 2002）[5]在其研究中提出可利用顧客的合約資訊參照顧客通話行為的改變來分析出可能流失的顧客。此研究提出一多重分類器整合(multi-classifier class- combiner)的分類方法來決定可能的流失顧客與非流失顧客。

Xia（Xia et al., 2008）的研究中提出利用支援向量機來架構顧客流失預測模型，並將其與多種資料探勘技術所架構的顧客流失預測模型進行比較，包括有類神經網路、決策樹、羅吉斯回歸以及貝氏分類器，其實驗證實使用支援向量機所架構出的顧客流失預測模型，其預測準確度比以其他資料探勘技術的預測準確度較佳。

Tsai（Tsai et al., 2009）的研究中提出利用複合式的資料探勘技術來找出可能流失的顧客。其所提出的方法是結合自我組織圖以及倒傳導類神經網路等兩種不同的類神經網路技術來協助解決顧客流失問題。此研究也以實驗證實，結合兩種不同類神經網路技術的複合式顧客流失預測模型的準確性優於只使用單一神經網路技術的預測模型。

2.1 知識發現

知識發現是一種由下而上的方式，從分析原始資料開始，找出我們所不知道的事實。1995年，在加拿大召開第一屆Knowledge Discovery in Databases和Data Mining國際學術會議。會議對KDD做了確切的定義。資料庫的知識發現是一個在知識中識別新穎有效模式的重複過程，這些模式具有潛在的可用性，並且最終可以被理解。資料探勘

作為知識發現的一個特定步驟，如圖1所示，它是一系列技術及應用，或者說是對大容量資料及資料間關係進行考察和建模的方法集，它的目標是將大容量資料轉換為有用的知識和資訊。

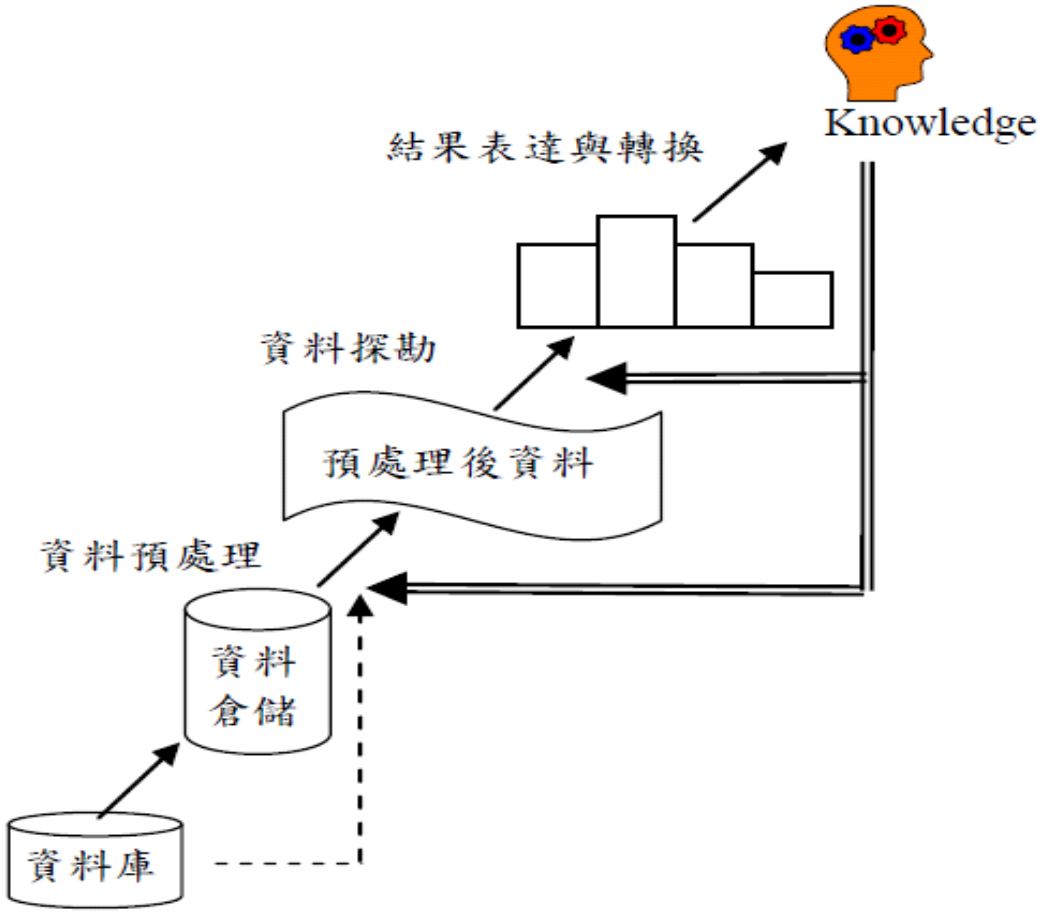

圖 1 知識發現處理流程

在資料採礦分析之前，所有資料的問題，諸如：不完整(Incomplete)，某些屬性值有遺缺，缺少某些分析時需要用到的屬性，只包含統計整合過的資料等；或有雜訊(Noisy)，資料有錯誤或特例(Outlier)；不一致(Inconsistent)，編碼方式不一致或命名方式不一致等問題。都需要在分析前處理完成。

2.2 資料探勘

「資料探勘」一詞是泛指從巨大的資料庫中粹取，綜合出未知資訊為主軸的複雜活動之通俗的講法，它是資料庫知識探索所有處理程序中的一個步驟，另一方面也是指有關於為了現實生活問題所存在的大量資料所作的各種領域的研究或發展的演算法及開發的軟體環境[6][7]，與KDD 相同的DM 的處理通常亦可以分為下列五個步驟(如圖2)：

(1) 資料選擇

由資料探勘處理的選擇目標和工具所組合的，辨識所要採掘的資料，然後選擇適合的輸入項屬性和輸出項的資訊來呈現給交付之工作。

(2) 資料轉換

包括下列這些操作：

a. 以想要的方法來組織資料。

b. 轉換資料的形式(如將符號轉為數字)。

c. 定義新的屬性縮減資料的幅員。

d. 消除雜訊(去除不必要的部份)與主題無關的部份。

e. 標準化。

f. 如果適當的話，決定如何處理遺失的資料的策略。

(3) 資料的探勘

就步驟的本身而論，接下來是對這些轉換後的資料進行採掘，使用一種或多種的技術來粹取感興趣的樣式。

(4) 結果詮釋與驗證（Result Interpretation and Validation）。對所探勘出的結果進一步的解釋建立模型，並應用已知的評估方法及未使用的資料庫中資料來測試它的正確度。

(5) 組織所探索到的知識（Incorporation of the Discovered Knowledge）。將結果呈現給決策者，做選擇或決定與目前的認知所潛在的衝突，將被粹取的知識應用於新探勘到的模型。

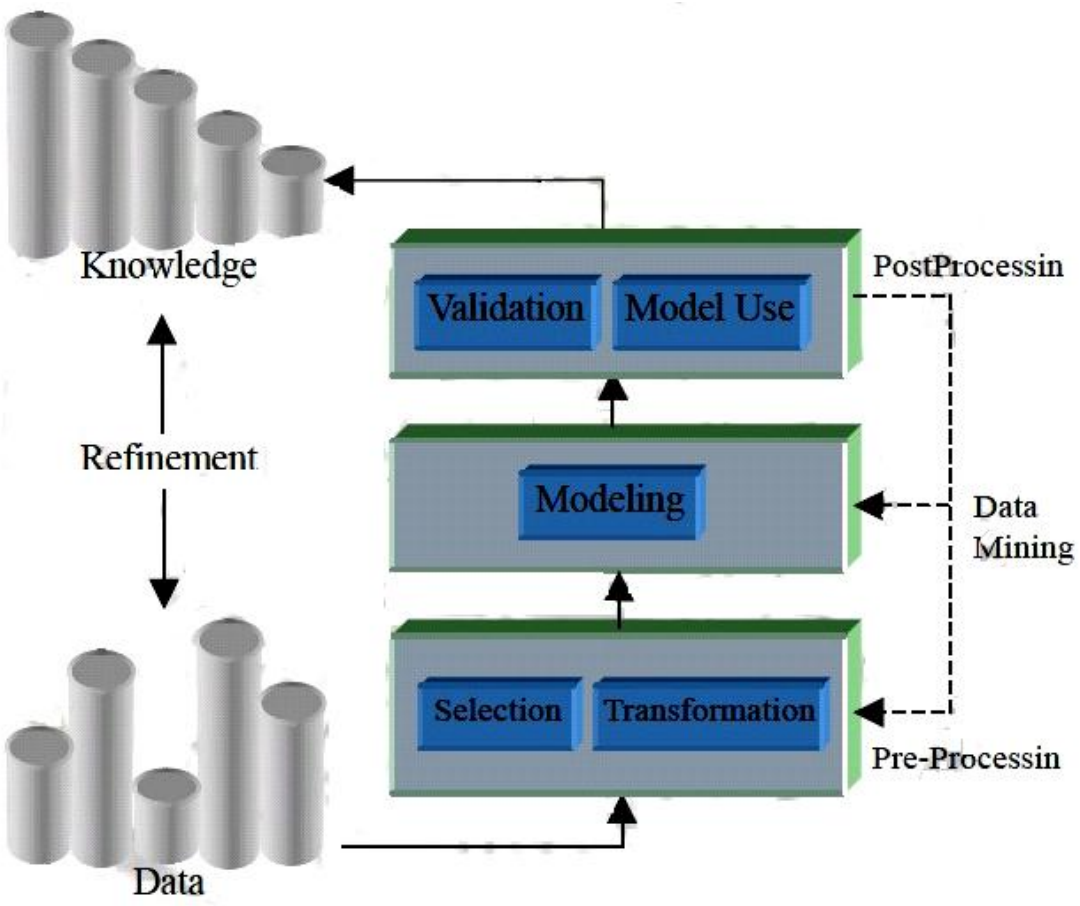

圖 2 資料探勘的處理步驟

在進行資料探勘時，依據所要探勘的目的及資料的性質，通常會進行下列幾種工作[7] [8]：

(1) 摘要（Summarization）

主要為對給予的資料產生特質或結構上的概略性描述。它可採取多種形式：數字(如簡單的統計描述語：平均值、標準差等)，圖表形式(如直方圖，散佈圖(Scatter Plots))，或者是 ”if-then” 形式的規則。它可能在整個資料庫裡或者在選擇的子集裡關於所推測的對象提供描述。

(2) 分群 (Clustering)

根據未知歸類特性資料的相似性，將相似的事物分群，讓群組內的資料相似度最高，讓群組跟群組間的資料相似度最低，找出彼此特性相似的群組，其目的是要將組與組之間的差異找出來，同時也要將一個組之中的成員的相似性找出來。

(3) 分類 (Classification)

是根據一些變數的數值做計算，再依照結果作分類。(計算的結果最後會被分類為幾個少數的離散數值，例如將一組資料分為“可能會回應” 或是 “可能不會回應” 兩類)。分類就是檢視、分析新物件的所有特性，然後將其指派到一個現有預先定義好的類別集群中，後續動作包含更新資料、標上類別編號。因為這些分類的事物通常是一組資料庫的交易資料，而賦予每一筆資料用以區別群集的辨識碼也是必須的，方能達到方便作業的功能。這些我們用來尋找特徵的已分類資料可能是來自我們的現有的歷史性資料，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個大的郵寄對象資料庫的部份取樣來建立一個分類模型 (Classification Model)，以後再利用這個模型來對資料庫的其他資料或是新的資料作預測。

(4) 迴歸分析 (Regression)。

迴歸分析 (Regression) 屬於建造一個些許透通的模型的監督式 (Supervised) 學習問題，其主要目的是做預測，目標是使用一系列的現有數值發展一種能以一個或多個預測變數的數值做為應變數，以預測一個連續數值的可能值的方法，可透過古典或更先進統計方法和以經常在分類任務過程中使用的符號式 (Symbolic) 方法來進行。

(5) 變動及偏移偵測(Change and Deviation Detection)。這項工作主要是在發現資料的實際內容與被預期的內容 (先前所預估的) 或標準值間是否有顯著的變動、誤差或偏移，這些變動可以包含時間上的偏差或群組間的差異。

(6) 依賴度 (從屬性) 模型 (Dependency Modeling)。依賴度模型的問題在於發現一個模型以描述屬性間顯著的的依賴或從屬關係，這些依賴度通常被以“if antecedent is true then consequent is true” 的“if-then”規則形式表示。

(7) 時序問題 (Temporal Problems)。

Time-Series Forcasting 與 Regression 很像，只是它是用現有的數值來預測未來的數值。Time-Series Forcasting 的不同點在於它所分析的數值都與時間有關。Time-Series Forcasting 的工具可以處理有關時間的一些特性，譬如時間的階層性(例如每個星期五個或六個工作天）、季節性、節日、以及其他的一些特別因素如過去與未來的關連性有多少。

(8) 因果關係 (Causation Modeling)。

這是一個在資料的屬性中發現因果關係的問題，使用一個 ”if-then” 形式的因果規則，表明條件（前項）和規則的當然結果（後項）之間有相互關係。

(9) 關聯規則 (Association Rule)

是要找出在某一事件或是資料中會同時出現的東西。Association 主要是要找出下面這樣的資訊：如果項目 A 是某一事件的一部份，則項目B也出現在該事件中的機率有 n %。（例如：如果一個顧客買了低脂乳酪以及低脂優酪乳，那麼這個顧客同時也買低脂牛奶的機率是 85%。）

(10) 屬性導向歸納法 (Attribute Oriented Induction)

屬性導向歸納法是一種以歸納屬性為基礎的資料分析技術，其技術核心為線上資料歸納方法，將相關式表格(Relational Dataset)資料集合中的每一個屬性，檢查其資料的分佈，判斷應歸納到那個相關的抽象層級[3]。

(11) 樣式導向相似性搜尋 (Pattern-Based Similarity Search)在時間或時間-空間資料庫搜索相似的樣式，經常會應用到兩種查詢類型：

a. 物件關聯相似度查詢（Object-relative Similarity Query），亦即相似度查詢（similarity query）或範圍查詢（range query）在所收集到的物件中，尋找使用者指定的範圍或距離中，符合的物件。

b. 完全關聯相似度查詢(All-pair similarity query)，亦即空間聯合(Spatial Join)目標是找到彼此都是在一段使用者指定的範圍或距離內的全部相符的要素。

(12) 資料方塊法 (Data Cube)

資料方塊法一般概念為將經常被要求的高成本計算具體化，尤其是計數(count)、總計(sum)、求平均數(average)、取最大值(max)等的歸納函數，將歸納後的具體化景觀儲存在一個多重維度資料庫(資料方塊)，可供決策支援、知識發現及其他應用做參考[3]。

(13) 序列樣式探勘 (Sequence Pattern Mining)。

在包含時序關係的資料庫中尋找一定數量所支持的序列樣式，主要是找出關聯順序進行行為模式上的預測，例如若A 事件發生，則B 事件可能接著會發生。

2.3. 機器學習

機器學習(Machine Learning, ML)通常與分類(Classification)畫上等號，是根據一些變數的數值做計算，再依照結果作分類，而這個計算與分類的工作是透過機器利用演算法經由自動學習的過程，由機器尋找出分類的結果。機器學習是用來協助知識獲取的工作，要從龐大的資料中作知識粹取的工作,光靠人力是很難達到的，唯有借助機器可快速執行重複運算的能力，經過適當的演算法和理論,才有辦法協助我們來探索前所未知的知識，也因此成熟的機器學習理論研究，也成了發展人工智慧系統裡，重要的一環。Peter Clark（1990）闡述機器學習並不是為了解決外部的問題，而是改善對知識本身的陳述[9]，Langley（1996）定義機器學習的演算法改善本身的執行效率是根據演算法本身的經驗[10]。ML 常見的演算法如下[11]：

(1) 古典的統計方法（Classical Statistical Methods），例如線性區別分析（Linear Discrinant Analyses）、二元區別分析（Quadratic DiscriminantAnalyses）、邏輯區別分析（Logistic Discriminant Analyses）。

(2) 現代的統計技術（Modern Statistical Techniques），例如投影追蹤分類法（ Projection Pursuit Classification, PPC ）、密度推估法(Density Estimation)、k 個最近鄰居分類法（k-Nearest Neighbor）, 因果網路（Casual Networks）、貝氏理論（Bayes theorem）。

(3) 類神經網路(Neural Networks)，例如倒傳遞網路(Back-Propagation Network)、

Kohonen 網路模型、機率神經網路(Probabilistic Neural Network，PNN)、霍普菲爾網路(Hopfield Neural Network, HNN) 與適應共振理論網路(Adaptive Resonance Theory Network, ART)、雙向聯想記憶網路(Bidirectional Associative Memory Network, BAM)、放射函數網路 (Radial Function Networks)。

(4) 支持向量機 (Support Vector Machine, SVM)。

(5) 決策樹方法 (Decision Tree Methods)，例如 ID3、CN2、C4.5、T2、Lazy decision trees、OODG、OC1、AC、BayTree、CAL5、CART、ID5R、IDL、TDIDT、PROSM 等。

(6) 決策規則演算法 (Decision Rule Algorithms) 例如AQ 系列、LERS 等。

(7) 分類學習系統 (Learning Classifier Systems) 例如GOFFER-1 、MonaLysa 、XCS 等。

(8) 關聯規則演算法 (Association Rule Algorithms)，例如APPIORI。

2.4. 規則推導

在機器學習眾多的理論當中，規則推導(Rule Induction)的學習理論是最被廣泛討論及應用之一，規則推導的主要涵義是從群訓練案例中尋找出最佳的、正確的、可了解的分類方法的規則[12]。規則推導是一種由一連串的「如果.../則... (If / Then)」之邏輯規則對資料進行細分的技術，在實際運用時如何界定規則為有效是最大的問題，通常需先將資料中發生數太少的項目先剔除，以避免產生無意義的邏輯規則。較常見的規則推導方法大致上有：以樹狀推導(Tree Induction)的表達方式，例如，Classification and Regression Trees 演算法；以類神經網路(Neural Network)的各連結(Link)權重(Weight)表達方式；以及J-Measure等方式。

(1) 決策樹推導

決策樹模型的推導(Decision tree induction)是一種使用樹狀架構的方法來做分類，節點代表不同的feature，樹枝為feature 的值，而樹葉則是不同的分類類別(class label)。這種方式是先找一個最佳的特徵作為根節點，所有的資料以此根節點為判斷根

據，進行分類，分類在每一個分支的資料再選出最佳的特徵作為根節點，再進行分類，形成一棵子樹，如此的過程一直重複，直到在一個分支內的所有資料都屬於同一個類別，推導過程才算結束，這個最終的分支就會形式樹葉，裡面記載著該樹葉內的資料所屬的類別，這樣就會形式一棵決策樹，如圖 3所示。

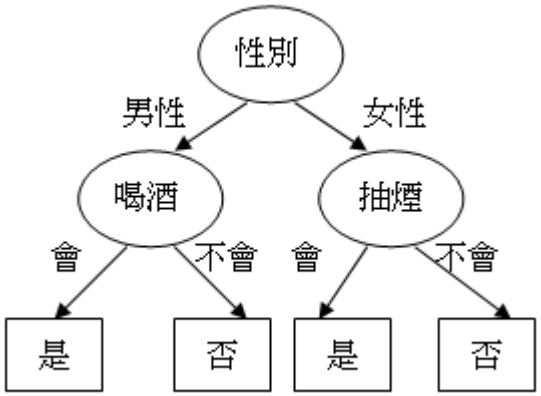

圖 3 決策樹示意圖

決策樹目前最被廣泛使用的決策樹演算法包括分類與迴歸樹(Classification and Regression Trees , CART)、卡方自動互動偵測(Chi-Square Automatic Interaction Detector , CHAID)、C4.5 等等三種演算法，這三種演算法主要是將顧客做分群，讓企業公司更瞭解顧客的喜好及需求，以幫助企業公司做決策。決策樹[4] ，歸納出決策樹三種常用的演算法為CART、CHAID、C4.5 之分析比較表如表 1所示。由下表的比較，本研究使用決策樹演算法為CART。

表 1 Decision Tree 分群分析比較表

演算法	CART	CHAID	C4.5
效能	高	低	中
節枝	少	多	中
分群	優	劣	中

(2) 類神經網路推導

類神經網路（Artificial Neural Network）類似人類神經結構的一個平行計算模式，是「一種基於腦與神經系統研究，所啟發的資訊處理技術」，通常也被稱為平行分散式處理模式（Parallel Distributed Processing Model）或連結模式（Connectionist Model）。類神經網路它可以利用一組範例，即系統輸入與輸出所組成的資料，建立系統模型（輸入與輸出間的關係）。有了這樣的系統模型便可用於推估、預測、決策、診斷，

而常見的迴歸分析統計技術也是一個可利用的範例，因此類神經網路也可以視為一種特殊形式的統計技術。

類神經網路顧名思義，其網路架構是模仿生物神經網路，整個網路可大致分為三個部分：神經元(又稱處理單元，Processing Element, PE)、層(Layer)、網路(Network)。神經元其輸入輸出的關係式，可用以下函數式子(1)表示：

$$Y_j = f\left(\sum_i W_{ij} X_i - \theta_j\right) \quad (1)$$

其中Yj=模仿生物神經元模型的輸出訊號。

f=模仿生物神經元模型的轉換函數(Transfer Function)，是一個將輸入值乘上權重加總後再經轉換成人工神經元輸出值的數學式。

W_{ij}=模仿生物神經元模型的神經節強度，又稱連結權重(Weight)值。

X_i=模仿生物神經元模型的輸入訊號。

θ_j=模仿生物神經元模型的偏權值。

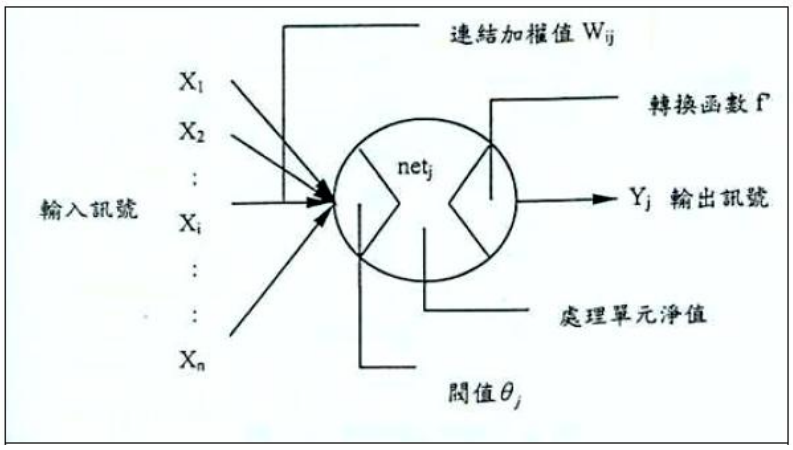

圖 4 神經元之示意圖

將上述的神經元組合起來就成為一個類神經網路。一個神經網路的架構分為輸入層、隱藏層、輸出層，由下方做說明：

輸入層：用來表現網路的輸入變數，其神經元數目依問題而定。

隱藏層：用來表現輸入神經元間的交互影響，其處理神經元數目並沒有標準的方法以做決定，通常要以試驗的方式來決定其最佳數目，網路可以不只一層隱藏層，也可以沒有隱藏層。

輸出層：用來表現網路的輸出變數，其神經元數目依問題而定。

如同在生物神經網路之中，神經元的強度可視為生物神經網路儲存資訊的所在，神經網路的學習即在調整神經結的強度。類神經網路各處理單元之間則以連接鍵互相連結，整個類神經網路的記憶就存放於這些連接鍵之中，以權重(Weight)來表示。圖 5為整體類神經網路之示意圖：

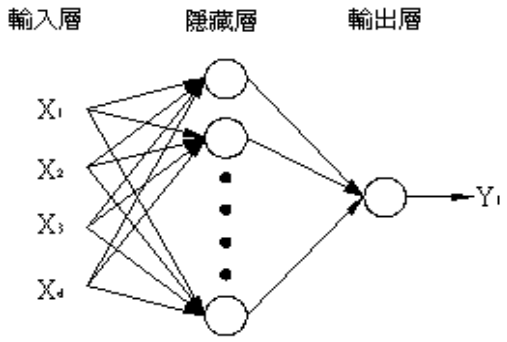

圖 5 類神經網路示意圖

目前為止，許多的學者針對欲解決問題的不同，提出許多的類神經網路模型，每一種類神經網路的演算法並不相同。常見的網路有：倒傳遞網路（Back-propagation Network)、霍普菲爾網路(Hopfield Network)、半徑式函數網路(Radial Basis Function Network），這些類神經網路並非適用所有的問題，我們必須針對欲解決問題的不同選擇適當的類神經網路。

(3) J-Measure 推導

利用互斥資訊(Mutual Information)的原理，計算某資訊對於問題的不確定性(Uncertainty，或稱熵(Entropy))能夠降低多少。但此方法較決策樹優越的地方在於：J-Measure 針對資料集N 中資料區分為數個類別(Class)，再以各類別中的區域(Region)進行計算，而不是單純將N 視為一個類別(Class)；因此，J-Measure 可以計算單一規則(某一類別中的某區域)所獲得的資訊，獲得更佳的推導結果。

第三章 研究方法

本研究包含:定義研究範圍、相關文獻探討、建立模型、收集資料、前置處理、實例研究、結論等7個步驟，研究流程如圖6所示:

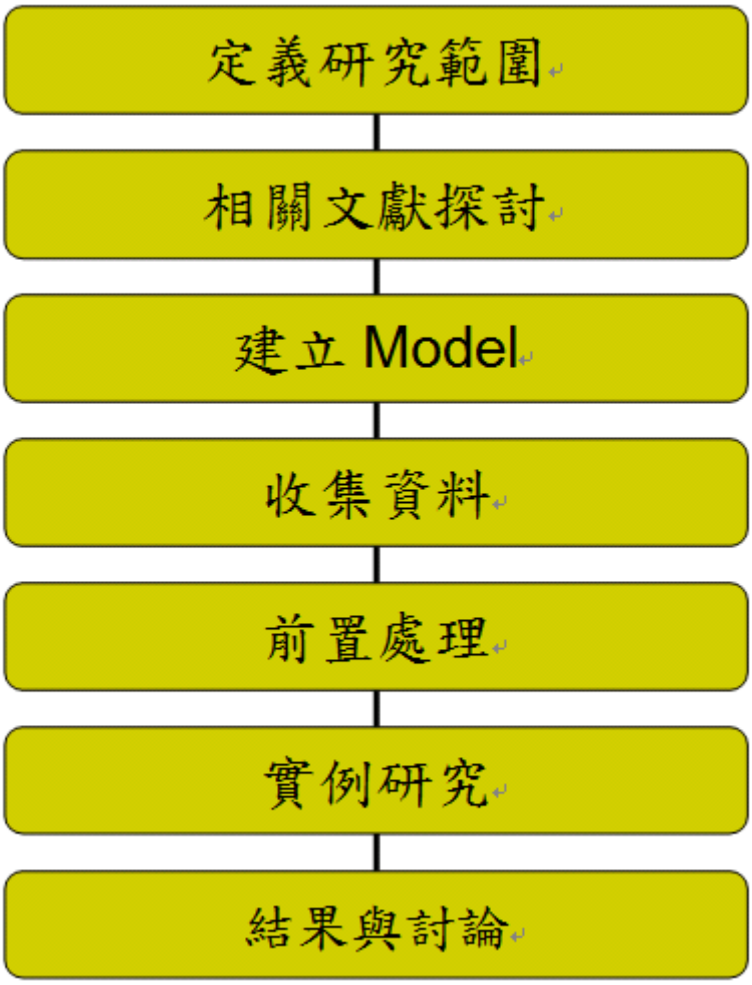

圖 6 研究步驟

3.1 定義研究範圍

筆者於國內某電信公司服務,故研究資料數據以所服務之中小企業客戶相關資料為研究範圍。首先，我們透過和專家討論的方式(包含了市場分析人員、資料探勘人員及企業客戶服務團隊)以及過去的經驗值從已流失的客戶裡找出特徵來挑選客戶資料的屬性，然後對資料進行預處理來建立潛在流失客戶的預測模型，並對模型做評估與比較。

3.2 資料的蒐集及預處理

資料的預處理的主要目的是避免垃圾進、垃圾出的情況。本研究主要應用資料探勘技術於潛在流失客戶之研究。主要的流程如圖 7；在確定要解決的問題或研究方向後，(1)進行相關資料(Data)的蒐集。(2) 將原始資料分組(Grouping)轉換成較有意義的資訊(Information)。(3)在所有資訊中擷取出對我們有用的知識(Knowledge)。

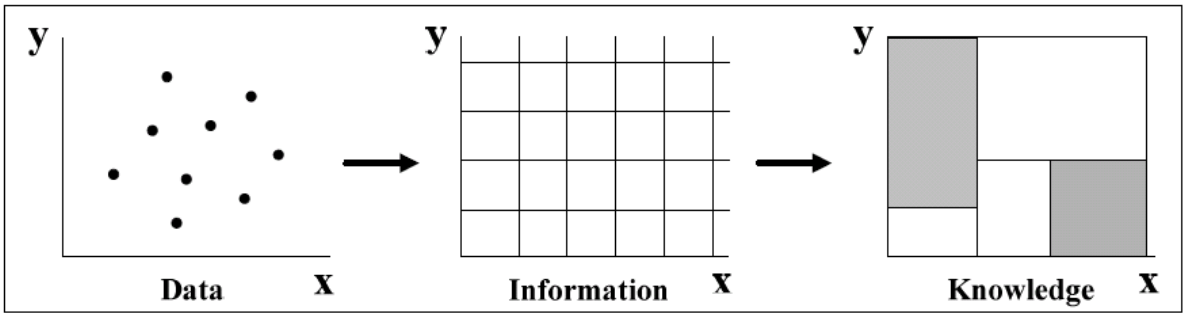

圖 7 資料探勘流程圖

遵行上述探勘過程，我們將實驗設計流程分成3階段：(1)資料蒐集部分，透過電信客戶關係資料庫抽樣取得客戶營收資料；(2)將取得之資料分別藉由標準差及相關係數計算模組、Recode重新編碼分組模組做資料預處理；(3)將預處理後之資料透過規則推導方法進一步運算，提取出對我們有用的知識。

(1)以標準差及相關係數篩選客戶的屬性

對於客戶的屬性挑選方式是透過和專家討論的方式(包含了市場分析人員、資料探勘人員及企業客戶服務團隊)以及過去的經驗值及過去流失客戶裡找出特徵來挑選客戶的屬性。由中小企業客戶中資料庫中依行業類別機械類的挑選8個類別資料分別為資本額/元(CA)、總營收/元(TC)、市話打市話費用/元(CC)、市話打長話費用/元(CL)、市話打行動費用/元(CM)、市話打國際費用/元(CN)，國內電路費/元(CE)及是否流失(LOST)等客戶屬性資料如表2。

表 2 客戶屬性資料表

CA	TC	CC	CL	CM	CN	CE	LOST
5000000	3091	237	85	19	380	402	是
16000000	8901	979	0	66	0	539	是
15000000	12206	5547	495	2719	0	1584	否
17500000	8377	321	188	40	0	483	是
20000000	14636	4744	2366	1821	70	539	否

標準差:(Standard Deviation)，在機率統計中最常使用作為統計分佈程度(Statistical Dispersion)上的測量。標準差定義為變異數的算術平方根，反映組內個體間的離散程度。標準差的公式(2)如下:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \qquad (2)$$

相關係數: 主要是告訴我們變項間的相關程度高或低，並沒有檢定「自變項」對「依變項」影響，因此得到的相關係數(r值)只能說明這兩個變項間是正相關、負相關，或者是無關。不能解讀為自變項對依變項的影響。

相關係數的公式(3)如下:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \qquad (3)$$

在相關係數解讀上，正負表示的是相關的方向，而非相關的程度。

R 值相關程度之高低,在正負0.3之間(即0.3至-0.3之間)稱為低度相關;在正負0.3-0.6之間(即指介於0.3至0.6，-0.3至-0.6之間)稱為中度相關；而在正負0.6至0.9之間(即指在0.6至0.9，-0.6至-0.9之間)則稱為高度相關；若是R值為正負1，即表示完全相關。將各屬性資料經標準差及相關係數計算如表3，篩選客戶屬性標準差值小及相關係數相關程度低之變數資料進行資料刪減的動作，將國內電路費資料屬性刪除。

表 3標準差及相關係數計算表

	總營收	市打市	市打長	市打行	市打國	國內電路費
標準差	5623.461	2028.259	899.657	1835.438	1271.612	424.6359548
相關係數(流失)	-0.12929	-0.39154	-0.49053	-0.66009	-0.32819	0.1575773

(2) 以Recode重新編碼資料分組

除了以Compute的方式轉換資料，亦可由Recode的方式重新編碼，為能建立決策樹狀客戶屬性資料表結構，以進行決策樹狀結構分析故把表2資料等級化之類別資料，採用五等份的劃分法將屬性資料依費用排序轉換為名義變數如表4。

表 4轉化後客戶屬性資料表

CA	TC	CC	CL	CM	CN	LOST
1	1	1	1	1	1	1
1	1	1	1	2	1	1
1	1	1	1	2	1	1
4	1	4	4	3	1	0
4	3	5	2	4	1	0
3	1	2	4	5	1	0

3.3 決策樹－CART規則推導

分類與迴歸樹(Classification and Regression Trees ,CART)是由Breiman、Friedman、Olshen and Stone 所開發的預測演算法，此演算法是一個二元(Binary)分割的方法，藉著一個單一輸入變數函數，在每一個節點將資料分為兩個子集合，以建構

一個二元式決策樹,其利用訓練組資料先建構一個完整的樹,再進行修剪樹的工作,CART演算法是依據整體節點的錯誤率(Error Rate)來做為修剪樹時的根據，如圖8。

假設資料Y 中含有n個類別為$X_1, X_2, X_3, \ldots, Xn$,依Gini值s將Y資料分割為兩個部分$\{Y_m, Y_n\}$,K_j和R_j為表示在Y_m及Y_n 節點的集合中屬於X_j類別的個數，以X_j為Y 資料中最大的類別，則Gini 值計算式如(4)，(5)，(6)所示：

$$Gini(s)=P_1(1-P_1)+P_2(1-P_2) \tag{4}$$

$$P_1=\frac{K_j}{Y_m} \tag{5}$$

$$P_2=\frac{R_j}{Y_n} \tag{6}$$

在分類與迴歸樹(Classification And Regression Trees,CART)演算法用Gini 條件建立完成一棵完整的決策樹(Full Tree)後，就會進行修剪決策樹的作業，這方法分為修剪法(Pruning Technique)及盆栽法(Bonsai Technique)， CART 演 算 法 是 用修剪法。一般而言，初長成的完整的決策樹，對於這些資料的分類尚未做最佳劃分類。CART 演算法在修剪決策樹是依據整體的誤差率來修剪的，以得到最佳的分類。首先要計算出節點的誤差率(Node Error Rate)，其公式如(7)所示：

$$\text{node}_{errorrate}=\frac{X_i}{\sum_{i=1} X_i}$$

X_i=節點中分類錯誤的資料數

$\sum_{i=1} X_i$=節點中資料總數 (7)

CART 演算法是從最底層的樹葉往上修剪，先計算決策樹最底層節點的上一層節點錯誤率，從錯誤率最小的節點的子節點開始修剪，然後再計算整體的錯誤率，如果整體錯誤率大於使用者定義的錯誤率，就停止修剪作業。

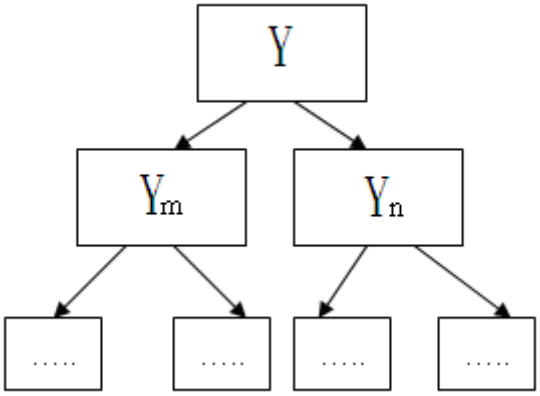

圖 8 CART 二元式決策樹

第四章 實証與分析

本研究使用 IBM 公司 SPSS 18 透過預處理後 3.2 節的 7 個客戶屬性(Attributes)包含了 CA、TC、CC、CL、CM、CN、LOST 等去建立我們的預測模型。首先將電信客戶關係資料庫中抽樣取得的 500 筆資料依上述 7 個客戶屬性，預處理分類後用 Recode 重新編碼方式依費用金額遞增方式採 5 等份等量畫分法將其分類。如表 5 所示。

表 5 客戶群屬性分類表

CA	TC	CC	CL	CM	CN	LOST
4	3	4	1	1	3	1
5	5	5	4	5	4	0
1	1	3	2	3	4	0
5	5	4	5	4	1	0
3	3	4	3	3	5	0
2	1	1	1	1	3	1
1	1	1	1	1	3	1
2	5	4	2	4	4	0
5	2	2	3	4	3	0

CA、TC、CC、CL、CM、CN 的分類:依費用金額遞增方式分類 1 表示費用金額介於 1~100 區間，2 表示費用金額介於 101~200 區間，3 表示費用金額介於 201~300 區間，4 表示費用金額介於 301~400 區間，5 表示費用金額介於 401~500 區間。

LOST: 0 表示否，1 表示是。

（1）將 7 個屬性經決策樹模型規則推導出潛在流失客戶之規則。我們找出預測的模型如圖 9，模型自變數的重要性如圖 10，模型數據如表 6。

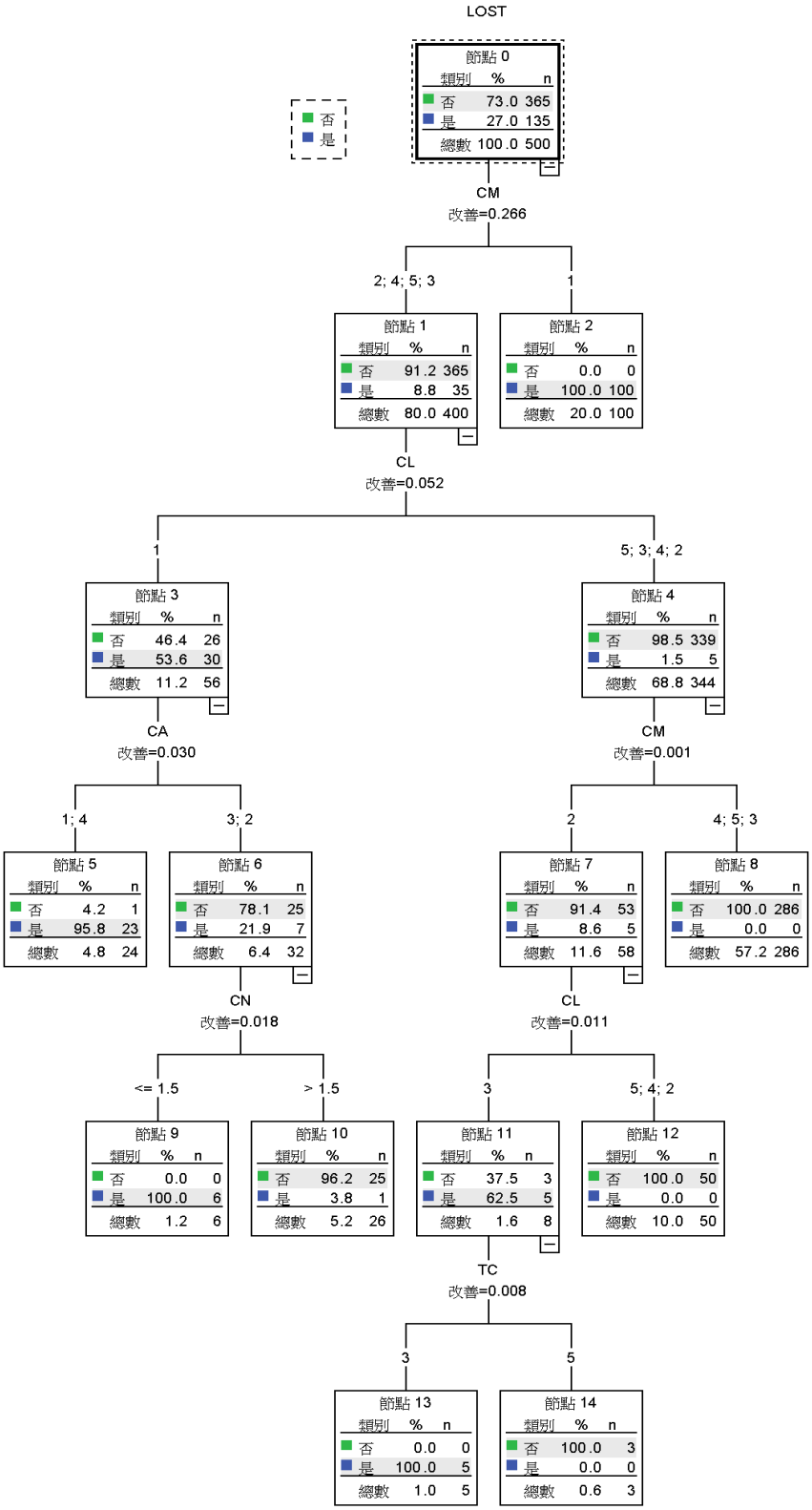

圖 9 決策樹所有參數規則推導流失因素

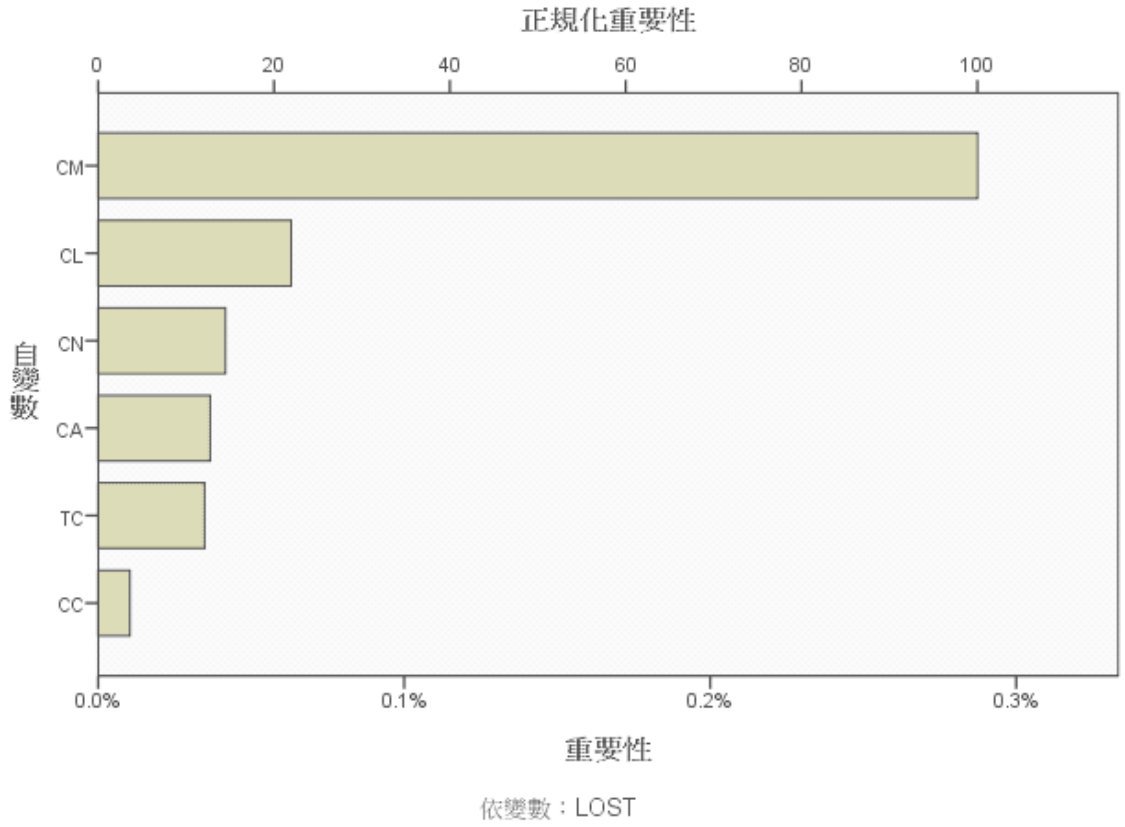

圖 10 決策樹自變數重要性比例圖

表 6 決策樹所有參數規則推導分類表

分類

觀察次數	預測次數		
	否	是	百分比修正
否	364	1	99.7%
是	1	134	99.3%
概要百分比	73.0%	27.0%	99.6%

成長方法:CART

依變數: LOST

產生規則如下:

```
/* Node 1 */.
IF (CM != "1") THEN  Node = 1  Prediction = 0 Probability = 0.912500
```

```
/* Node 3 */.
IF (CM != "1") AND (CL = "1") THEN Node = 3
Prediction = 1 Probability = 0.535714

/* Node 5 */.
IF (CM != "1") AND (CL = "1") AND (((CA = "1" OR CA = "4") OR (CA != "3"
AND CA != "2") AND ((TC = "1" OR TC = "3") OR (TC != "2" AND TC != "4")
AND ((CC NOT MISSING AND (CC <= 1.5)) OR CC IS MISSING AND ((CM = "2"
OR CM = "5") OR (CM != "4" AND CM != "3") AND (CN NOT MISSING AND (CN
<= 1.5)))))))
THEN Node = 5 Prediction = 1 Probability = 0.958333

/* Node 6 */.
IF (CM != "1") AND (CL = "1") AND (((CA = "3" OR CA = "2") OR (CA != "1"
AND CA != "4") AND ((TC = "2" OR TC = "4") OR (TC != "1" AND TC != "3")
AND ((CC NOT MISSING AND (CC > 1.5)) OR CC IS MISSING AND ((CM = "4" OR
CM = "3") OR (CM != "2" AND CM != "5") AND (CN IS MISSING OR (CN >
1.5)))))))
THEN Node = 6 Prediction = 0 Probability = 0.781250

/* Node 9 */.
IF (CM != "1") AND (CL = "1") AND (((CA = "3" OR CA = "2") OR (CA != "1"
AND CA != "4") AND ((TC = "2" OR TC = "4") OR (TC != "1" AND TC != "3")
AND ((CC NOT MISSING AND (CC > 1.5)) OR CC IS MISSING AND ((CM = "4" OR
CM = "3") OR (CM != "2" AND CM != "5") AND (CN IS MISSING OR (CN > 1.5)))))))
```

```
AND  (((CN NOT MISSING  AND  (CN <= 1.5)) OR CN IS MISSING  AND  (TC = "1")))
THEN Node = 9 Prediction = 1 Probability = 1.000000

/* Node 10 */.
IF (CM != "1")  AND  (CL = "1")  AND  (((CA = "3" OR CA = "2") OR (CA != "1"
AND  CA != "4")  AND  ((TC = "2" OR TC = "4") OR (TC != "1"  AND  TC != "3")
AND  ((CC NOT MISSING  AND  (CC > 1.5)) OR CC IS MISSING  AND  ((CM = "4" OR
CM = "3") OR (CM != "2"  AND  CM != "5")  AND  (CN IS MISSING  OR (CN > 1.5)))))))
AND  (((CN NOT MISSING  AND  (CN > 1.5)) OR CN IS MISSING  AND  (TC != "1")))
THEN Node = 10 Prediction = 0 Probability = 0.961538

/* Node 4 */.
IF (CM != "1")  AND  (CL != "1")
THEN Node = 4 Prediction = 0 Probability = 0.985465

/* Node 7 */.
IF (CM != "1")  AND  (CL != "1")  AND  (CM = "2")
THEN Node = 7 Prediction = 0 Probability = 0.913793

/* Node 11 */.
IF (CM != "1")  AND  (CL != "1")  AND  (CM = "2")  AND  (((CL = "3") OR (CL !=
"5"  AND  CL != "4"  AND  CL != "2")  AND  (CA = "4")))
THEN Node = 11 Prediction = 1 Probability = 0.625000

/* Node 13 */.
IF (CM != "1")  AND  (CL != "1")  AND  (CM = "2")  AND  (((CL = "3") OR (CL !=
```

"5" AND CL != "4" AND CL != "2") AND (CA = "4"))) AND (((TC = "3") OR (TC != "5") AND ((CC NOT MISSING AND (CC <= 3.5)) OR CC IS MISSING AND (CN IS MISSING OR (CN <= "4")))))

THEN Node = 13 Prediction = 1 Probability = 1.000000

/* Node 14 */.

IF (CM != "1") AND (CL != "1") AND (CM = "2") AND (((CL = "3") OR (CL != "5" AND CL != "4" AND CL != "2") AND (CA = "4"))) AND (((TC = "5") OR (TC != "3") AND ((CC NOT MISSING AND (CC > 3.5)) OR CC IS MISSING AND (CN NOT MISSING AND (CN > "4")))))

THEN Node = 14 Prediction = 0 Probability = 1.000000

/* Node 12 */.

IF (CM != "1") AND (CL != "1") AND (CM = "2") AND (((CL = "5" OR CL = "4" OR CL = "2") OR (CL != "3") AND (CA != "4")))

THEN Node = 12 Prediction = 0 Probability = 1.000000

/* Node 8 */.

IF (CM != "1") AND (CL != "1") AND (CM != "2")

THEN Node = 8 Prediction = 0 Probability = 1.000000

/* Node 2 */.

IF (CM = "1") THEN Node = 2 Prediction = 1 Probability = 1.000000

（2）使用類神經網路模型推導出各屬性影響潛在流失中的重要性，我們找出預測的模型如圖 11，各屬性重要性對照表如表 7，比例如圖 12。

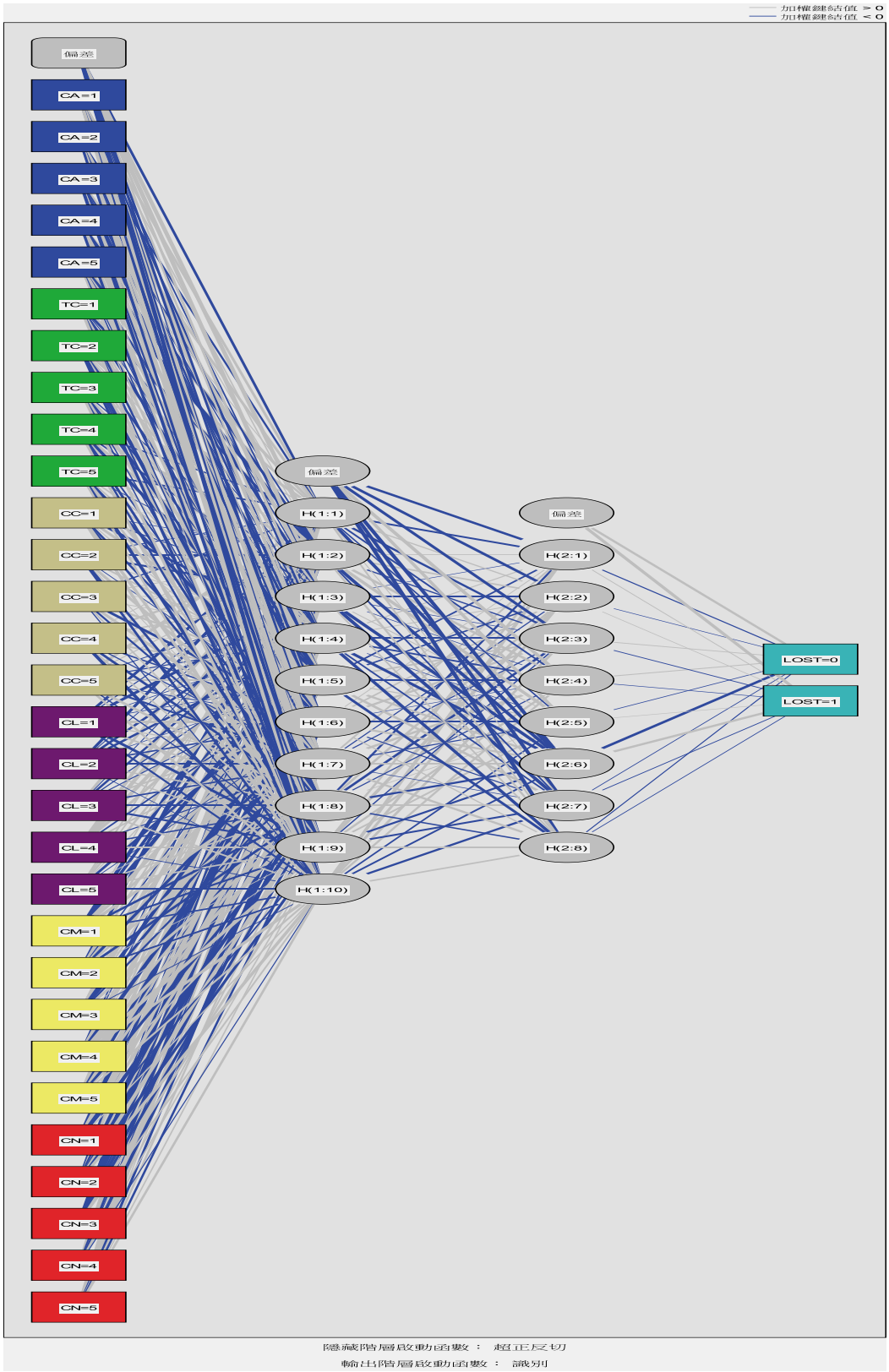

圖 11 自變數重要性之類神經網路

表 7 自變數重要性對照表

自變數的重要性

	重要性	正規化重要性
CA	.094	20.9%
TC	.026	5.8%
CC	.058	12.9%
CL	.203	44.9%
CM	.452	100.0%
CN	.144	31.8%
CE	.023	5.0%

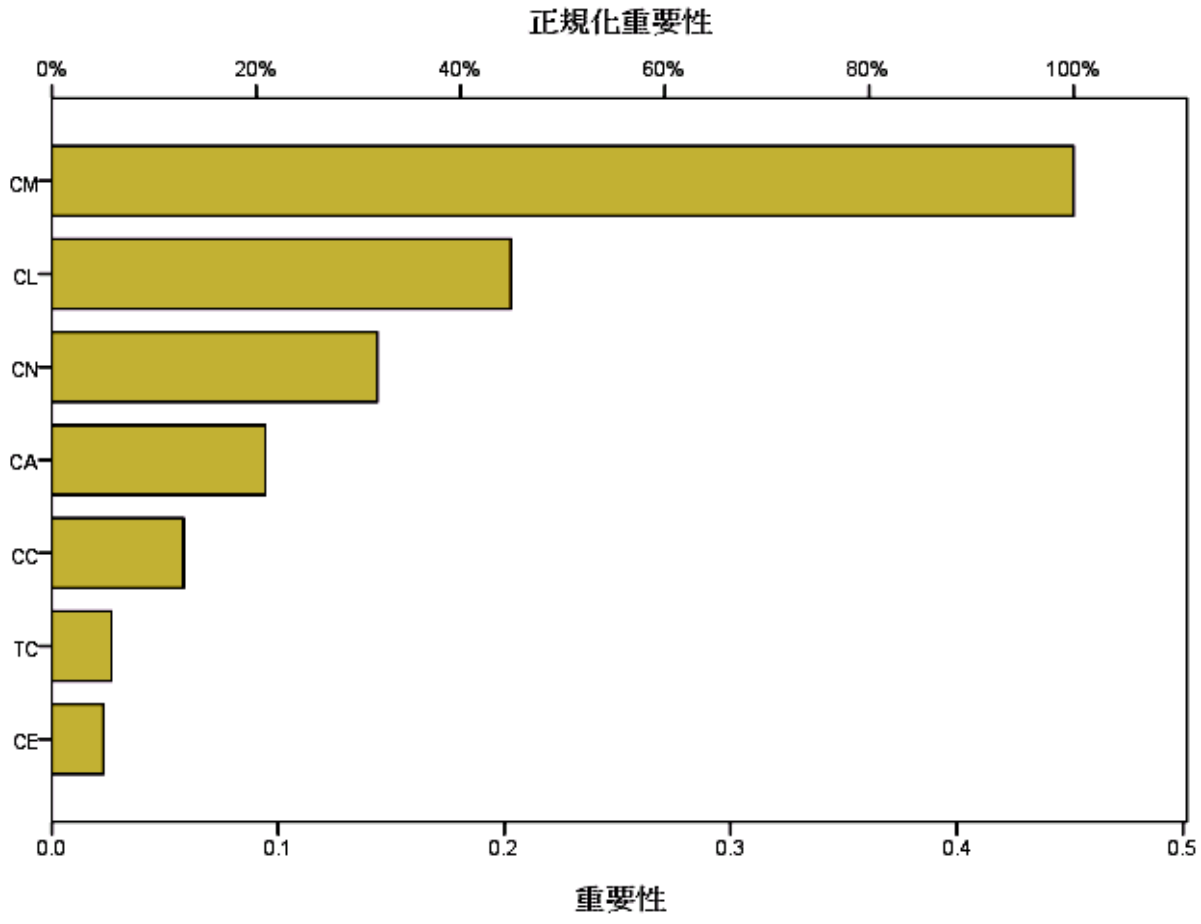

圖 12 類神經網路自變數重要性比例圖

（3）將所有屬性經決策樹模型及類神經網路推導出的前 2 個重要屬性 CM、CL，經決策樹模型規則推導出潛在流失客戶之規則，我們找出預測的模型如圖 13，模型數據如表 8。

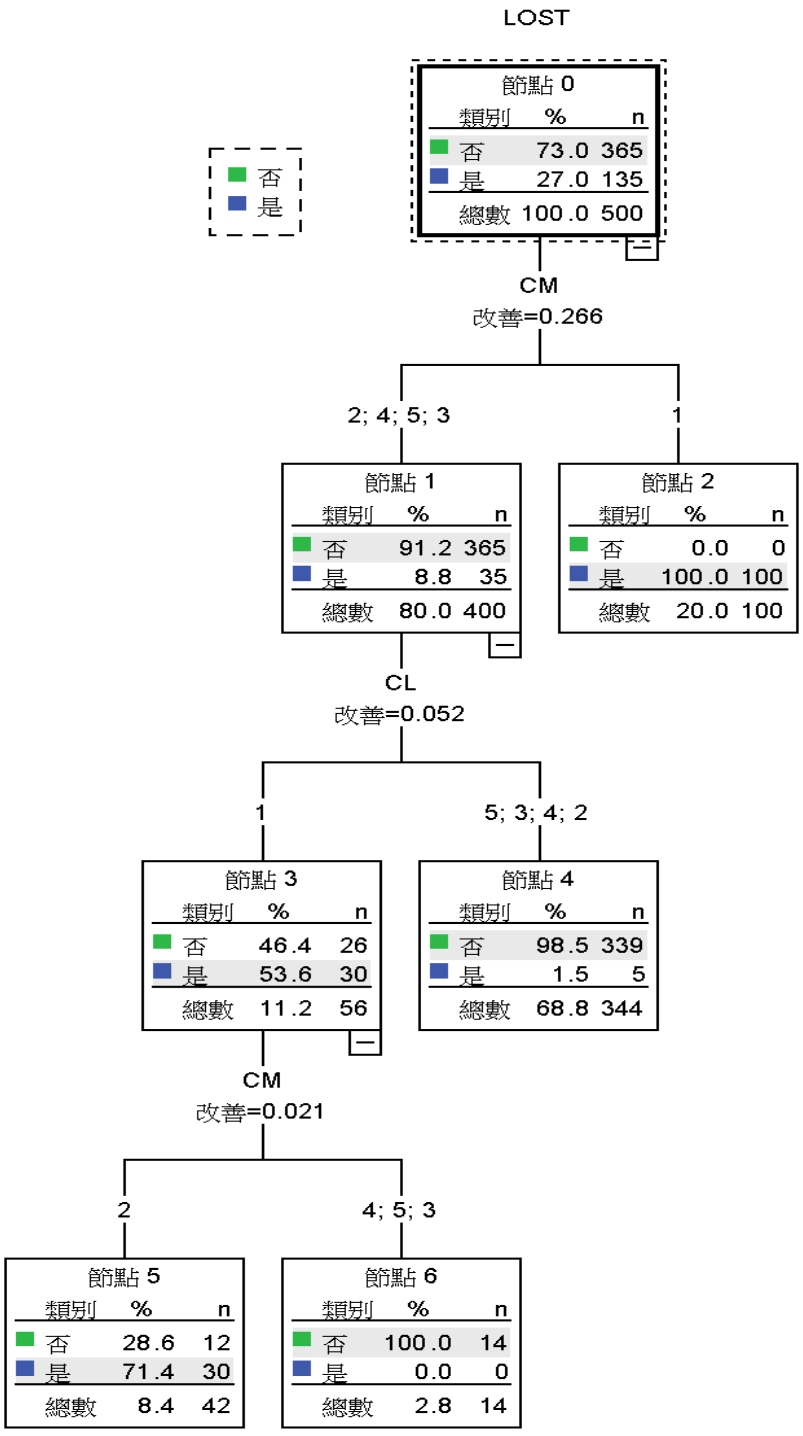

圖 13 決策樹 CM、CL 參數規則推導流失因素

表 8 決策樹 CM、CL 規則推導分類表

分類

觀察次數	預測次數		
	否	是	百分比修正
否	353	12	96.7%
是	5	130	96.3%
概要百分比	71.6%	28.4%	96.6%

依變數: LOST

自變數: CL, CM

產生規則如下:

```
/* Node 5 */.
IF (CM != "1") AND (CL = "1") AND (CM != "4" AND CM != "5" AND CM != "3")
THEN Node = 5 Prediction = 1 Probability = 0.714286

/* Node 6 */.
IF (CM != "1") AND (CL = "1") AND (CM = "4" OR CM = "5" OR CM = "3")
THEN Node = 6 Prediction = 0 Probability = 1.000000

/* Node 4 */.
IF (CM != "1") AND (CL != "1")
THEN Node = 4 Prediction = 0 Probability = 0.985465

/* Node 2 */.
IF (CM = "1")
THEN
Node = 2 Prediction = 1 Probability = 1.000000
```

運用決策樹模型及類神經網路模型分析的結論來對中小企業中潛在流失客戶的具體情況進行分析，將所有屬性代入決策樹模型萃取規則，所產生的規則有 14 條，可分類客戶是流失客戶的機率為 99.3%，模型分類的預測正確性為 99.6%，所有屬性的重要性前 2 名屬性為 CM、CL 與類神經網路分析跑出的自變數重要性前 2 名 CM、CL 相符合，

因此 CM、CL 此 2 屬性為判斷客戶流失的主要屬性，再將 2 屬性代入決策樹模型萃取規則，所產生的規則有 4 條，可分類客戶是流失客戶的機率為 96.3%，模型分類的預測正確性為 96.6%。

第五章 結論

客戶資源是電信公司的生命，企業客戶市場是各家電信業者的必爭之地，各家業者無不卯足全力，保留並鞏固客戶資源對電信公司來說意義重大。決策樹是資料挖掘中一個常用的分類方法其理論清晰，方法簡單，學習能力強，適於處理大規模的學習問題，是一種知識獲取的有用工具。本研究利用資料探勘技術中的決策樹演算法以中小企業客戶為對象目標做潛在客戶流失分析，能夠幫助電信公司瞭解發現潛在的流失客戶，提供企業客戶服務團隊有效的掌握客戶動向，改進客戶服務進而挽回該企業客戶以增裕公司營收。此外，決策樹演算法在電信的其他方面也有著廣泛的應用：如客戶細分、電話欺詐等，同樣決策樹在其他金融領域也有著廣闊的應用前景。

參考文獻

中文部分

[1]邱義堂，「通信資料庫之資料探勘：客戶流失預測之研究」，國立中山大學資訊管理學系碩士論文，2001。

[2]毛慧雯，「使用決策樹預測電信業優質及可能流失之客戶」， 輔仁大學資訊工程學系碩士論文，2008。

[3]賴志東、楊文超，「資料挖掘」，高等資料庫報告，2001。

[4]周佩蓁，「以決策樹分析顧客滿意度之研究」，育達商業技術學院資訊管理系碩士論文，2005。

西文部分

[5] Wei et al “Turning telecommunications call details to churn prediction: a data mining approach” . Expert Systems with Applications vol. 23 issue 2 , pp. 103-112, 2002.

[6] Hegland M., “Data Mining - Challenges, Models, Methods and Algorithms,” Publications of ANU Data Mining group, Draft, 2003.

[7] Olaru, C. and Wehenkel, L., “Data Mining,” IEEE Computer Applications in Power, Vol. 12, no. 3, pp. 19-25, 1999.

[8] Chen, M.S., Han, J., and Yu, P.S., “Data Mining: An Overview from Database Perspective,” IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No.6, pp. 866-883, 1996.

[9] Clark, P., “Machine learning: Techniques and recent developments,” In A. R.Mirzai, editor, Artificial Intelligence: Concepts and Applications in Engineering,pp. 65 - 93,1990.

[10] Langley, P., “Elements of machine learning,” San Francisco: Morgan Kaufman,1996.

[11] Kusiak, A., “Feature Transformation Methods in Data Mining,” IEEE

Transactionson Electronics Packaging Manufacturing, Vol. 24, No. 3, pp. 214-221, 2001.

[12] Bramer, M.A., “Induction of Classification Rules from Examples: A Critical Review,” Proceedings of Data Mining 96, London, April 1996, Unicom Conferences, pp. 140-166, 1996.