# 資料挖掘與顧客保留/顧客利潤創造

張璋倫　　苑守慈

輔仁大學資訊管理研究所　輔仁大學資訊管理研究所

allen@im.fju.edu.tw　　yuans@tpts1.seed.net.tw

*摘要*

*近年來由於顧客對於公司的重要性逐漸超越傳統的地位，所扮演的角色也轉變為主動積極的。因此對於如何保留顧客與創造顧客利潤也成為當前最重要的議題。本研究從資料挖掘的相關應用技術來探討嘗試此議題解決之突破，包括了視覺化自我組織對應資料分群法(SOM)、階層式自動標記資料分群法(Automatic Labeling SOM)、以及分類決策樹(Decision Tree)與智慧型助理軟體(Intelligent knowledge-based Agent)，並結合四項技術形成新的整合方法，全自動地將所有資料分群並標記出特徵屬性，藉由分類找出可能產生偏差之群集，最後運用助理軟體與專家知識庫來找出原因並提供相關決策建議，以解決目前所遭遇之困難。*

*關鍵字：顧客保留/顧客利潤、資料挖掘、Self-Organization Maps、Automatic Labeling SOM*

## 一、緒 論

根據一份由 Customer Retention Practice 的 Newsletter 於 1998 年的報導中指出：『典型的企業中有百分之八十的利潤是百分之二十的顧客所創造出來的』[22]，顯示顧客於企業中所扮演的角色佔有程度，已經超越傳統的局勢，更說明了為何公司對於其主要顧客無不使用各種競爭手段與行銷方法來加以保留與開拓新群體。換言之，顧客(Customer)，對於各企業的影響力已經超越以往，從被動的角色轉變成為主動積極的地位，因此，如何確認顧客需求，以及如何保留並創造顧客已成為當前最重要的議題[23]。

另外，對於大型的企業來說，雖然大部份皆已擁有大量的顧客群，但是仍須不斷的找尋潛在的目標顧客群(Target Customer)，以期為公司帶來更多豐厚的利潤。而顧客忠誠度則是在潛在的無限商機中是相當重要的一個環節。因為除了可以利用現存顧客群刺激重覆購買，並可藉由分析其屬性來預測可能潛在的顧客區段。但是如何能夠真正的達到提升顧客忠誠度目標呢？我們認為，積極面的創造顧客利潤與消極面的保留顧客應該是首要的目標。目前有許多顧問公司提供完整的診斷服務，但是卻相當耗時間與人力，並且無法完全的找出正確目標顧客群。另一方面，由於限制於分析軟體以及統計分析方法，部份的仍需經由有經驗的專家來做，如此浪費太多的成本。因此本研究將從此問題著手，設計簡易的整合型系統改善目前所面臨的瓶頸。

資料挖掘(Data Mining)是目前在資料分析運用上較新穎的方法，目的是將一般資料庫中看似無用的資料(data)轉換成有用的資訊(information)，除此之外，並找出隱藏之關聯性。更進一步的定義：『資料挖掘之工具使用了演算法去探索與發掘資料，並搜尋出大量資料中難以偵測之模式』。它包含了多項相關技術如關聯式法則(association rule)、分類法(classification)、資料分群法(clustering)、循序模式(sequential pattern)、時間序列(time series)等[9, 20, 23]。所有的方法主要皆以挖掘潛在的資訊為目的。但是資料挖掘能做些什麼呢？它能夠避免風險(risk)與欺騙(fraud)、預測(predict)收入與消費行為、推銷產品與服務、管理與顧客之聯繫以及電子商務之推行。由此可知資料挖掘之應用範圍相當廣泛，並且可以幫助使用者找出潛在的隱藏資訊，以利行銷策略之推展。

但是，在這眾多的應用技術中，何種較適用於本研究呢？這需從問題的本質來探討，因為本研究主要的目的在於保留現有可能流失的顧客群與發掘潛在之顧客區段，故最重要的即是要將顧客分群區隔，以便更進一步的分析，分類過濾出可能有偏差之顧客或者是找出可能成為主要顧客群的區段，然後在過濾的同時標記出特徵屬性(如過去消費模式、購買產品等)。本研究嘗試將資料挖掘技術加以運用，形成一個視覺化與自動化分析處理的方法，針對顧客資料加以分析，並求以達到顧客保留/利潤為主要目標。除了應用資料挖掘中的分群與分類技術外，將加上智慧型助理軟體(intelligent agent)來相互配合。系統之前兩部份分別為資料分群法與分類法，目的是要將原始大量的資料分群並且自動標記出各群集之特徵屬性，接著將有可能產生偏差的群體分類過濾出來，如此便做到了真正的區隔出不同的顧客區段，以及後端的自動分析出偏差因素與提供可行之策略。

因此本研究將傳統應用技術形成新的結合- 階層式自動標記視覺化(hierarchical automatic labeling SOM)資料分群法與分類決策樹(decision tree 之結合 - 形成主要系統架構，後端並運用智慧型助理軟體來找出可能產生偏差之原因，並連結知識庫來提供因應策略。因此將大量原始資料輸入後，階層式自動標記視覺化資料分群法會不斷的減少集群與屬性個數，最後以圖形顯示分群結果與標記出來之特徵屬性。接下來的分類就可以更容易的找出偏差集群，並根據其特徵屬性由智慧型助理軟體去由專家經驗所建之知識庫中搜尋，如同現實生活中專家會診並提供意見一般，找出產生偏差的原因以及可行之方案。故將預期達成下列幾項目標：

1. 原始的顧客資料(Raw Customer Data)經由多層 SOM 的分析方法產生視覺化的圖形分群結果，並且將各集群的主要特徵屬性標記出來，因此可以透過各屬性組成圖來挖掘相互之間的關聯性。
2. 標記的屬性特徵可藉由分類決策樹的分類來找出偏差群集，可能偏差群集與正常群集的分佈。
3. 可經由公司相關部門的決策主管經驗，建立一個儲存所有建議的知識庫(Knowledge Base)，並且由智慧型助理軟體來找出問題與對應之建議。
4. 藉由和使用者互動與階層化自動標記屬性資料分群方法來減少群集個數，並過濾出重要特徵屬性。
5. 本研究之系統將提供預警系統；若經過使用者選

擇群集個數後的分析效果不顯著，則會提出警告告，使用者可重新設定群集個數或由電腦自動運作產生。

綜合上述幾項目標，可了解本研究之貢獻在於利用由電腦自動產生之各屬性組成圖(Component Plane)與分析結果，提供經使用者互動後所需之合適建議，並輸出各階段運作結果，如總集群分佈圖、可能偏差群集(deviation clusters)以及最後產生的建議。

本篇論文主要包括三部份。第一部份為本篇論文對於顧客保留/顧客利潤創造問題之資料挖掘研究方法架構。第二部份為本研究資料挖掘研究方法所使用三個技術之說明，包括 Automatic Labeling SOM (主要是將未處理過的原始資料標記分群，找出所有的群集分佈)。決策樹分類法 (目的是將所有的群集分類出偏差與正常之群集。) 以及系統知識庫與智慧型助理軟體之結合 (嘗試找出可能產生問題之原因外，亦提供相關因應策略做參考。)。 第三部份則提供模擬範例以說明本研究方法流程及未來系統效能評估之評估指標(metrics)。

## 二、研究方法架構

根據研究顯示，對於過去平均每年百分之二十的高顧客判離率，各企業無不紛紛提高警覺，並思考其所帶來的衝擊與影響。因為如此，許多相關的因應策略也相繼的被提出，然而許多公司卻無法有效的保留顧客，更不用提創造新的顧客群。因此，該如何解決上述所遭遇的問題呢？簡單的來說可以從三方面來看，也就是『針對正確的顧客群』、『提供合適的服務』與『有效的績效改善』。從圖 1 中可以看出其關係：

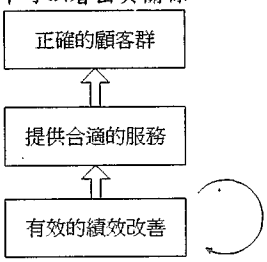


圖 1. 顧客、服務與績效關係

由圖 1 可以看出，公司內部的績效改善有助於未來提供合適之服務給予其顧客，從上圖中可以了解改善績效是首要的步驟，唯有如此才能夠進一步提供更好的產品與服務給正確的顧客群，產品與服務之提升便可以提高顧客滿意度，忠誠度相對的也會增加。本研究將先區隔出目標顧客群(包括可能產生判離與未來潛在集群)，然後分析其特徵屬性找出原因，最後提供有效的解決方案[1]。

對於上述三方面來說，目前所存在之解決方法相當有限，大多數的顧問公司都以有限的方法分析並提供專家意見，然而傳統的分析軟體並無法真正顯示出各群體間的關聯性與界限(boundary)，也無法從結果中找出具代表性的特徵屬性，只能看出各群集之分佈狀況。因此便無法做正確的分析，對於潛在的關聯性更是無法挖掘出來。除此之外，如何在找出叛離原因後做對應的措施，如何針對特定集群來分析，這都是需要靠電腦做進一步的分析，單是靠專家是存在相當高的風險的，由於人類是情感的動物，週遭環境的改變會影響人類的思維，故會產生結果之差異性。況且以人類來做分析所耗費的成本與時間是相當多的，也存在一定的風險，若能以電腦加以取代，只需要儲存專家意見於知識庫中，其餘便以電腦做自動化的分析步驟，而且可以提供更多的回饋資訊讓使用者互動，這些問題都是當前所極欲克服的，本研究也將從這些問題來探討並改進。

因此，本研究的目標在於對顧客資料分群並分類，並且將產生偏差之群集(clusters)與正常群集的屬性加以分析比較，藉以獲得判離原因並提供有效的建議。因此系統將分成三個部份，前端部份將利用非監督式學習(unsupervised learning)中的 automatic labeling SOM 以階層式方法來將資料分群，藉由產生多個屬性組成圖(component plane)找出各屬性間相互影響的程度。接著利用分類法中的分類決策樹來將可能有偏差(或潛在目標顧客群)的群集分類出來並過濾出特徵屬性，此部份的重點是將符號化(symbolic 後的屬性以分類決策樹的集群分類，最後得到可能產生偏差的集群屬性(attributes)。第三個部份則是將不同群集的屬性比較，分析出可能判離原因，並結合智慧型助理軟體(agent)與決策知識庫，搜尋出合適與正確的有效建議。

簡言之，本研究之系統架構主要可分為三個部份，前端為階層式視覺化資料分群法，並自動篩選特徵屬性。中端為決策樹分類法，將所有分群後的群集再加以分類，找出可能產生偏差或潛在目標群集(依據使用者定義)。後端則是簡易的智慧型助理軟體與決策知識庫，主要目的是要找出可能產生叛離的原因，並且提供相對應的策略給予決策管理者。

從本研究系統架構圖(圖 2)中可以清楚顯示整個實驗流程方向，將主要步驟歸納整理如下：

1. 整理原始顧客資料。
2. 將資料輸入 Hierarchical Automatic-Labeling SOM 的資料分群法中，並產生各屬性成份圖。
3. 將群集的特徵屬性標記出來。
4. 將經過標記後之群集分類，分成正常、偏差以及可能偏差三類群集。
5. 分析偏差集群與附近正常集群屬性之差異，推測顧客可能判離之原因。
6. 經由智慧型助理知識庫軟體 (intelligent knowledg -based agent)，找出最適合之建議與策略。

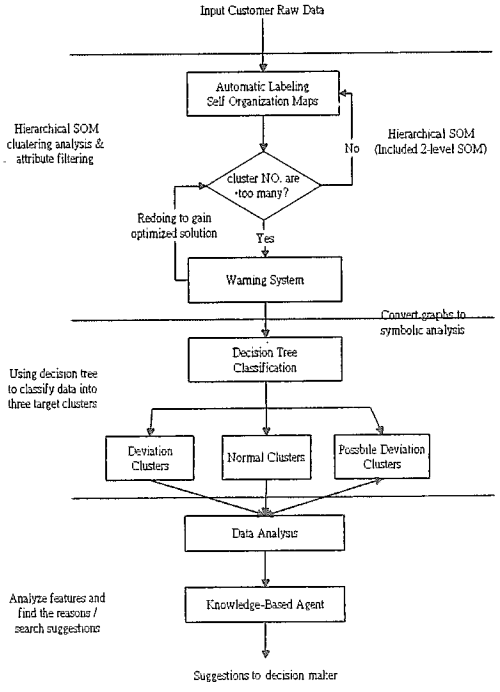


圖 2. 系統架構圖

本研究將結合不同的資料挖掘技術，並且與智慧型助理知識庫軟體配合使用，除了希望能夠將各技術之特性完全發揮之外，亦求能達到本研究之主要目標，茲將主要三種應用技術加以深入說明，並且加以探討人機界面之問題。

**(一)視覺化集群分析(Visualization Clustering Analysis)**

一般來說，公司所擁有之顧客資料一定是相當大量與複雜的，對於公司的高階主管來說，這些資料並無法立即產生有用的資訊，因此如何將看似無效用的資料轉換成有效用的資訊是相當重要的。本研究嘗試利用視覺化分群法來產生各屬性的組成圖，讓使用者可以很明確的經由圖形分析屬性間的差異性與關聯性。實驗中將會把輸入資料之各屬性用數值或符號來代替，如男與女可分別用 0 與 1 來取代。

由於使用者並不知道多少個數的群集會影響往後分析的結果，因此本系統將提供預警機制，以設定臨界值(critical number)來控制所產生的群集個數，若使用者在開始時不指定上限之群集個數，則會由系統自動運作產生，否則會將原始資料產生群集至使用者所設定之上限，再由臨界值來判斷是否會產生較差的結果。對於分群的架構採階層式的方式，目的在於希望減少過多的群集個數，並且將特徵屬性過濾與標記出來。待達到預定群集個數後，系統除了已保留特徵之屬性外，還會產生系統第一部份之輸出結果，也就是各屬性之成分圖。這不但可以讓使用者先得到初步的結果，也會把有用的資訊往下個部份傳送進行分類的動作。

因此，在視覺化資料分群法與自動標記屬性的技術中，可以滿足本研究的第一部份需求，即自動地將大量資料分成若干個群集，並且能夠把相似性較高的資料聚集在一起，標記出每一群集的主要特徵屬性，然後不斷的重複以上的動作直到滿足最佳的群集個數為止(假如使用者有設定)。

**(二)分類分類決策樹(Classify Decision Tree)**

經過集群分析後得到的結果，可以繼續的運用，許多有效的屬性被保留下來，這些都是足以代表各群集之特徵。接下來的重點是要如何利用這些資訊，藉由分類來找出目標資料群。

在此部份本研究希望能運用分類決策樹的分類方法，將目標分成三種類別：正常群集(normal clusters)、偏差群集(deviation clusters 與可能偏差群集( possible deviation clusters)，由於本研究目的在於找出可能產生偏差之顧客群集，因此在分類決策樹分類法中也將所有的資料分成三類，如此便可以找出哪些顧客群是可能會判離(defect)的。分類完成後必須要做資料分析，本研究希望能夠在找到各群集之後，比較偏差群集與正常群集特徵屬性上之差異，並且明確的找出產生偏差之顧客，當做往後尋找決策建議之參考。

因此，在分類分類決策樹技術中，針對第二部份的需求亦能滿足，除了原始的資訊續用外，將所有的群集再加以分類，企圖找出可能產生偏差之群集，因為目標的分類群相當明顯，故往後可從比較分析後找出原因與對策，亦能從正常群集來繼續開拓未來之潛在顧客群。

**(三)智慧型代理引擎(Intelligent Agent)**

整個系統架構到此部份，產生之結果為能代表各群集之主要特徵屬性，這些屬性可能像是"購買時間"、"居住地區"等等。對於管理決策者來說，要從這些結果知道該如何做決策是不容易的，因此本系統最後一部份則提供助理軟體幫忙於決策知識庫中找尋，利用前半段所分析出後的資訊與決策知識庫進行比對，找出最正確與有效的決策建議。

除了助理軟體的設計外，決策知識庫的建構也是相當重要的，本研究將透過與公司行銷與客服部門主管之經驗以及專家意見，匯整成所有可能發生的問題及相對的因應策略，形成系統決策知識庫。因此當智慧型助理軟體在尋找相關問題之解答時，會進入已建構完成之決策知識庫中，並且根據問題找出對應之建議，提供給管理決策者做參考，此部份將是未來進一步研究的重點。

由於此部份之技術與人工智慧領域中是密不可分的，故將此技術與資料挖掘的結合將使本系統更具有互動性，充份利用助理軟體的特性，滿足系統最後一部份需求，即能夠隨時將發生的問題找出滿意答案，並且期望在未來能夠有學習機制，而不僅只於目前之靜態供應，能夠有動態的互動式學習機制產生，使完整的系統更強大。

**(四)人機介面(Human-Computer Interaction)**

本系統之目標雖然期望達成視覺化分群與自動化分析之結果，但部份仍需要藉由與使用者互動來進行，如設定群集個數、保留比率等，因此在介面的設計上期望亦能做到半自動化分析，使用者可根據系統流程自行操作與設定，並由電腦輔助其分析運作。除此之外，也可完全經由電腦自動化運作產生至最終的結果，但此結果並不保證為最佳化。本系統強調重點為與使用者雙向互動溝通，因為在資料挖掘的過程中，使用者扮演的角色相當的重要，唯有互動溝通才可使系統更人性化，並且由使用者操作與電腦輔助的自動分析，得到具高說服力與可信度的實驗結果。

除此之外，在系統開始運作前，使用者需要設定群集個數與保留比率，若使用者不自行設定，系統將會有預設值，以自動化運作產生至最佳效果為止。

## 三、研究方法

本研究將以三部份為基礎，即視覺化資料分群法(visualization SOM)[3, 5, 6, 7, 10, 14, 15, 16, 19, 18, 19 ]、automatic labeling [11, 12, 13]、分類分類決策樹(decision tree)[8, 21, 24]以及智慧型助理軟體(intelligent agent)。在許多的資料分群方法中，由 Kohonen 所提出的 Self Organization Ma 可以說是較新穎的方法，而分類決策樹則有兩種最常被應用的方法：ID3 與 C4.5。因此，本研究將以上述方法為系統架構之基礎。

### 3.1 視覺化資料分群法(Visualization Self-Organization Maps)

一般來說，許多未經過分類的原始資料在分類之前都需先經過分群步驟，對於現存的分群方法中可以分成階層式(hierarchical)與分割式(partition)兩種，其代表方法分別有最小生成樹(minimum spanning tree) 以及 K 平均資料分群法(k-means clustering 。但是，由 Kohonen 所提出之自我組織對應資料分群法在最近已經被廣為運用，除了剛開始應用於解決工程問題外，也慢慢的被用來做資料的分析。這種方法的優點是可以在分群時減少資料量，並且將這些資料投射(project)到二維的圖形上以減少原始資料的維度(dimension)。

視覺化資料分群法的概念相當近似 k 平均資料分群法。由於每筆原始顧客資料可能會有相當多的屬性，

這些屬性也稱為『維度』，因此觀念為"若在 n 維空間中相近，則投射至二度空間也會相當接近"，此觀念亦可以運用在醫學與天文學上。主要的方法是利用『競爭(competition 』的概念，所有的原始資料都當成輸入向量，而被投射之二度空間的每一個 neuron 均視為輸出，也都是競爭者。而二度空間的 neuron 可以一維 neuron 輸出向量來表示。這每一個 output neuron 都會與所有的輸入向量比較，最接近者則稱為贏家，並且調整至與此輸入向量更接近。從圖 3 中可以明確的看出輸入、輸出向量(競爭者)與 fan-in 向量之關係。每個點都有對應的權重向量(weight vector)，並且每個群集都會將所有的權重向量平均，稱為平均權重向量。

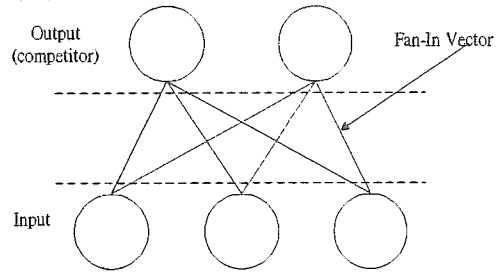


圖 3. Self-Organization Maps 概念圖

除此之外，SOM 也使用了觸發函數(activation function)，其主要的功用是用來表達某一點的輸入向量與權重向量間的距離，一般來說，歐基里德距離(Euclidean distance)是最常被使用的。在經過訓練學習靠近後，最短的距離(最大的觸發函數)會被當成最贏的，稱為最適配對單元(best matching unit)，亦可稱其為最贏向量(winner vector)，因此最後便會被修正的更靠近原始輸入向量。換句話說，整體的 SOM 觀念可用下列公式來表示：

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \quad \text{for each } i \in N_c(t),$$

其中 $t$ 是任一時間，Nc(t)則是最適配對單元的所有鄰近點集合，i 為該集合中的任何一個點，$\alpha(t)$為鄰域核心函數(neighborhood kernel function)，通常界於 0 與 1 之間，也可稱為學習速率(learning rate)，主要目的是要將最適配對單元(BMU)的所有鄰近點做調整，一般有泡沫(bubble)與高斯(gaussian)兩種函數，而以高斯函數最常被使用。

在經過一連串競爭後，除了最贏向量(winner vector)會調整外，其附近之鄰近向量也會跟著調整，這是因為要讓整個集群能夠更接近相近的資料，且因為本身的關聯性，故鄰近的點移動調整仍要繼續維持鄰近關係，以形成更大之群集。然而，由於點與點之間不斷學習與靠近，導致學習到一定程度時便會遇到瓶頸，無法再繼續學習下去，或者是學習成效變化不大時，便可以終止學習。因此，學習機制會影響學習的效率，依照不同的學習機制能讓系統在不一樣的環境下運作，故在系統設計之初就需考慮此問題。

總而言之，視覺化的資料分群法可以說是資料分群視覺化的先驅，藉由此類方法減少資料維度與數量，並且可以產生多個屬性組成圖，讓使用者可以從圖形中更清楚看出群集間的關聯性與差異性，也可以更清楚地找出不同屬性的群集分佈，充份的發揮其『視覺化』的優點。

### 3.2 自動標記資料分群法(Automatic-Labeling SOM)

從 SOM 的觀念於 1989 年被 Kohonen 提出之後，許多學術研究便紛紛加以利用並改良，皆以視覺化概念為基礎來發展新的觀念。其中 Andreas 與 Dieter 於 1999 年提出的 Automatic-Labeling SO 的觀念，是將傳統 SOM 的缺點改進後所形成的。以傳統的視覺化 SOM 資料分群法來說，屬於非監督式(unsupervised)的類神經網路(neural network) 模型，主要以分析高維度的資料而形成低維度的群集。因此藉由 LabelSOM 的方法可以讓使用者從特徵來了解群集的結構，並且自動的標記出特徵屬性。

自動標記(automatic labeling)的目的是在將原始資料中多且複雜的顧客屬性，自動地過濾(filter)出具代表的特徵屬性。所謂標記(label)也就是將屬性貼上標籤，表示該屬性已經被選取，並且在標記後利用計算的數值排出所有的順序重要性。例如該群集之資料原有十五個屬性，但經過自動標記的動作後，只剩下五個可以代表此群集的特徵屬性，系統並且可以根據所計算出的數值來排各屬性的優先順序，如此便可利用這些特徵屬性快速的分析出該群集的特性。

SOM 雖然有視覺化的功能，但是卻仍然無法自動地偵測出各群集之間的界限，而且對於自動標記出原始資料的特徵屬性(features)仍然存在著瓶頸。而 LabelSOM 能夠使被分群後的點集合都會被標記出主要的特徵屬性。而且所期望的是能夠在形成許多群集後根據不同的特徵屬性來描述該群集特徵，而不是在形成群集前先用所有的屬性來描述其特徵，如此才能顯現自動標記的優點。因此，便利用簡單公式來計算權重向量(weight vector)與輸入向量(input vector)中各屬性間的距離，內容是根據權重向量與輸入向量的每個屬性值來計算距離值，值越小顯示該屬性與群集相當接近，表示越能夠表現出此群集之特徵，因此便會自動被標記出來。當然，我們也可以設定一個值來限定標記的屬性個數，例如設定五個 label 則表示只需標記出最重要的五個特徵屬性。LabelSOM 的概念可以表示成下列公式：

$$q_{ik} = \sum_{x_j \in C_i} \sqrt{(m_{ik} - x_{jk})^2}, \qquad k = 1..n$$

qik 表示每個屬性的量化誤差向量(quantization error vector)值。Ci 是所有輸入樣本 Xj 對應到的點 i 之集合，而根號裡面運算權重向量與輸入向量每個相對屬性的距離(mi 表示權重向量的第 k 個屬性值，Xjk 則為輸入向量的第 k 個屬性值)。圖 4 將描述量化誤差向量所含之觀念。

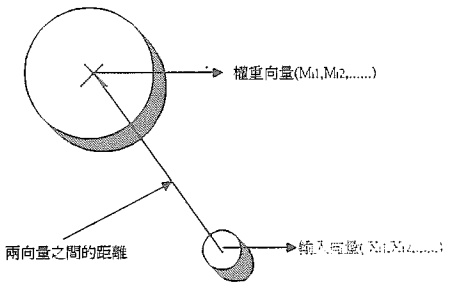


圖 4. 量化誤差向量運算表示圖

因此，利用自動標記屬性的 SOM 資料分群法可以驗証決定性的特徵屬性。除此之外，該系統內部亦提供 AC (Adaptive Coordinates)與 DC (Distance Connections)之附屬功能，目的即是讓靠近群集以圖形顯現，並可讓使用者更清楚的區分各群集間的界限。但最重要的是，它將所有資料進行分析處理，過濾出有用的屬性資訊，也就是所謂的特徵屬性(feature)，以及讓使用者可以清楚

找出群集間明顯的差異以及不同群集之屬性相異性。

### 3.3 2-Level Self Organization Maps

在應用 Automatic Labeling SO　視覺化的分群方法後，本研究嘗試將階層式處理之觀念。系統前端將以階層式之 LabelSOM 資料分群法，全自動化地來減少屬性個數與群集個數，期望能夠將總群集個數空制在合理的範圍之內。並且不斷的過濾找出主要特徵屬性。另外，為能自動地予資料類別標記以利後端分類，我們利用兩段式的 SOM (2-level SOM) 分群方法來解決資料類別自動標記的問題[4]。本研究將嘗試將兩段式 SOM 資料類別自動地標記方法結合多層式 automatic labeling SOM 資料分群法。

兩段式 SOM 來解決資料類別自動標記的方法，乃是先從部份原始資料來做訓練，然後將目標分類的標記定出。例如該實驗將雞蛋分成為受損的、已破壞的與部份損毀的三類。接著根據第一層的分群結果，與所指定之標記形成輸入向量進入第二層 SOM 來做資料類別自動標記，此部份為分類前的訓練步驟，目的是在於讓系統學習何種資料屬於何種標記，並在未來有資料輸入時知道分至何種標記。待系統將訓練資料進行標記學習後便完成了第一個步驟。

對於第二個測試的步驟，也就是完全以電腦自動做資料類別自動標記，所有的資料也將經過兩段式的 SOM 資料分群法，然後根據先前測試後的學習標記結果來進行資料資料類別自動標記。此方法(圖 5)根據該研究結果顯示有高達百分之九十七的成功率，可以成功的將所有測試雞蛋分成三類。然而，此方法仍有一些瓶頸，也就是須為同一物件之分類，無法以多樣物件加以分類。無論如何，此方法除了能解決分類的問題之外，也是一種觀念相當新穎的方法，它只須運用到兩段式的 SOM 便可做到分類法。

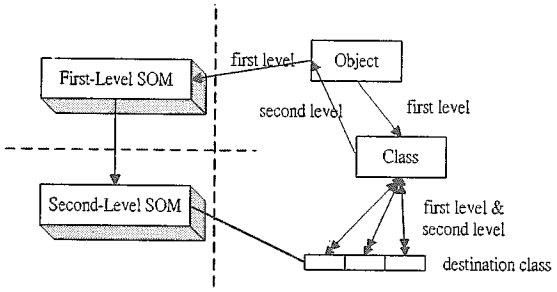


圖 5. 兩段式 SOM 表示圖

### 3.4 SET-Gen 分類分類決策樹(SET-Gen Decision Tree)

分類決策樹(decision tree)在眾多分類方法中可以說是最典型的，它是利用二元樹的觀念來區分相異屬性之同一物件，而且是屬於符號化(symbolic)的分類法，簡單的說就是將非數值屬性轉換成代表符號(如 0 或 1)。整個分類決策樹當中，可以分成終端點(terminal node)與非終端點(non-terminal node)兩部分，終端點為經過分類後的點，其餘則為非終端點。在所有的終端點上有一層等級標記(class label)，目的是區隔終端點與非終端點之界線。分類決策樹是屬於監督式學習，會將存在的問題藉由分類變成有限的。分類決策樹的觀念如圖 6 所示。

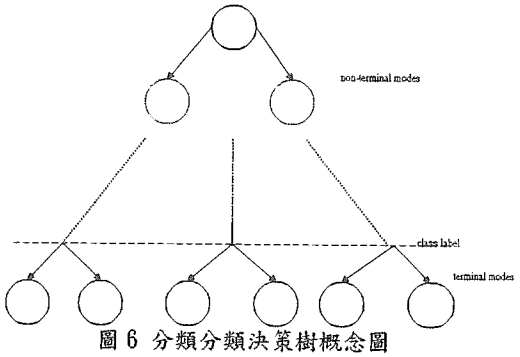


圖 6 分類分類決策樹概念圖

由於目前存在許多不同分類決策樹的方法，其中 Cherkauer 與 Shavli 於 1996 年提出的 SET-Gen 分類分類決策樹，它是依照傳統 C4.5 改良製作，並將輸入特徵屬性(input unique features)的個數減到最少，且所建之分類決策樹較 C4.5 所產建之小，樹越小越能對於資料一般化，且其預測精確程度沒有與 C4.5 相差太多，甚至更為精確。主要是在內部運用了基因搜尋(genetic search)來選擇特徵屬性，假若新的子集合(subset)更適合於原來最差的屬性(worst member)，則就會被新的子集合取代，否則將其丟棄。根據該研究所進行之實驗結果顯示，各方面的統計數據都優於傳統的 C4.5 分類法，顯示其為快速又準確的分類決策樹分類法，未來將逐漸地被廣為運用於各領域中。

對於分類分類決策樹來說，仍然先需要原始資料來進行分類訓練，經由不斷的學習分類後，可以得到預期的分類結果。但有一個問題是要考慮的，若將不相關的屬性當成訓練資料來學習，則有可能造成結果的偏差，即學習到錯誤的資訊，準確性就有可能會不穩定。目前最常使用的兩種分類分類決策樹演算法分別是 ID3 與 C4.5，但由於可能會出現未知(unknown)的屬性，且 ID3 並不能允許有未知的屬性，相反的 C4.5 不但允許未知屬性並有因應的建立分類決策樹的方法，因此功能相對的也較強大。本研究因為仍於研究初期，針對所有的原始顧客資料屬性仍處於篩選階段，可以預期的是大筆的資料存在未知的屬性是無法避免的，因此未來也將運結合功能較強之 SET-Gen 之分類決策樹分類法，找出可能產生偏差之顧客群集。

### 3.5 綜合方法

以上所描述之方法包括自我組織對應視覺化資料分群法、自動標記屬性資料分群法、兩段式資料分群分類法以及分類決策樹，皆依據本研究之需求加以深入探討，由於本系統亦分成三個部份來研究設計，故將上述所有的方法結合起來將形成系統前端與中端的重點。系統後端的決策知識庫建置也是系統中重要的一環，配合簡易的智慧型助理軟體，整體的系統將趨近於完善。本研究也將所四種技術加以結合，期望每個步驟都能順利的運作，且有效的分析資料，並達到預期的結果，最後加上專家知識庫後，系統將變的更有效用。因此，本研究將過去提出之應用技術加以結合，期望能有效的解決未來所遭遇的問題。

### 3.6 範例說明

本範例使用約五百筆的資料來進行研究方法流程說明，其中本研究使用的屬性包括：支持球隊、居住地、年齡、職業、學校、收入、性別等七項。首先須將所有的資料轉成數值來分析，如將居住地北部設為 1、中部

為 2、南部為 3 等，其餘各屬性值則依此類推。接著經由 SOM 視覺化資料分群法得到分群結果與各屬性成份圖(圖 7)。

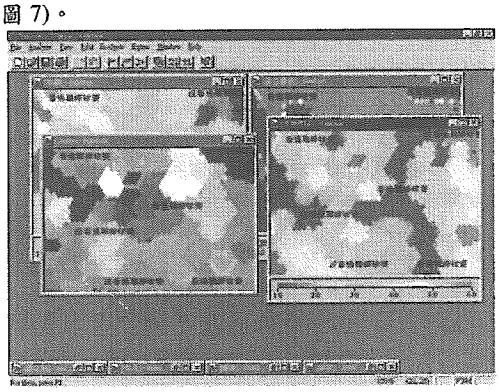

圖 7. Self-Organization Maps 視覺化分群結果

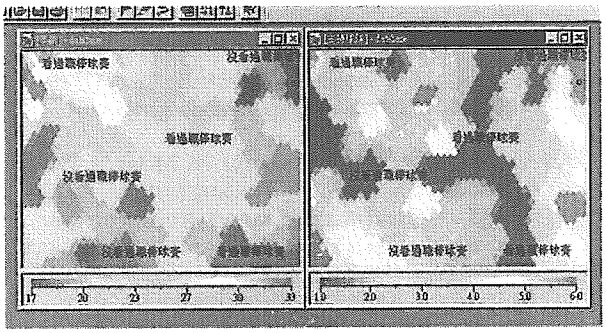

圖 8. 屬性組成圖

根據圖 8 之屬性組成圖中可以比較出年齡與支持球隊的分佈圖，經由 SOM 提供之顏色差異看出不同群集之分佈，本研究測試後發現 17-19 歲年齡層的球迷都多支持味全龍隊，而和信鯨隊的球迷則大多分佈於 22-27 歲。因此，將所有的屬性成份圖加以分析比較後，可以得到許多類似上述的隱藏關聯性。此為系統的第一部份輸出。

由於資料會不斷的經過多層的 SOM 分析，於是自動標記的功能便會將屬性過濾，逐漸的保留有用的特徵屬性。根據自動標記屬性資料分群法的實驗結果顯示，所有的特徵屬性會全部輸出，圖 9 為模擬產生之輸出結果。

| | | | | | |
|---|---|---|---|---|---|
| 年齡、支持球隊、職業 | | | | 居住地、年齡、支持球隊 | 居住地、收入、支持球隊 |
| 年齡、支持球隊、性別 | 支持球隊、性別、居住地 | | | 居住地、年齡、性別 | 收入、年齡、支持球隊 |
| | | | 居住地、年齡、收入 | | |
| | | | 居住地、年齡、收入 | | |
| 收入、年齡、職業 | | 性別、年齡、職業 | | | |
| 收入、年齡、居住地 | | | | 學校、年齡、支持球隊 | 居住地、年齡、學校 |

圖 9. 自動標記出之特徵屬性圖

由於本研究尚處於研究初期，故在此模擬兩段式 SOM 資料類別自動標記法應用至多階層式 SOM 資料分群法。本研究將設定未來分成三種類別，分別是正常群集、偏差群集與可能偏差群集，因此圖 10 將顯示兩段式 SOM 分群分類法應用至多階層式 SOM 資料分群法之架構圖。

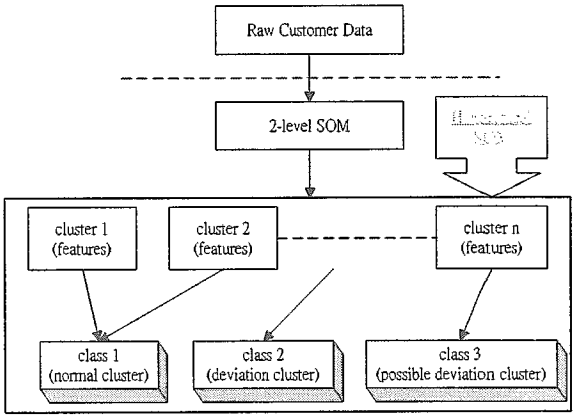


圖 10 兩段式 SOM 分群分類法應用至多階層式 SOM 分群法之架構圖

接著系統並將所有群集之特徵屬性輸入分類分類決策樹，藉由設定三種目標分類群集來進行分類，由於做完自動標記後所有屬性皆已數值化，故進行分類決策樹分類即可直接分類。圖 11 顯示本系統之分類決策樹架構。

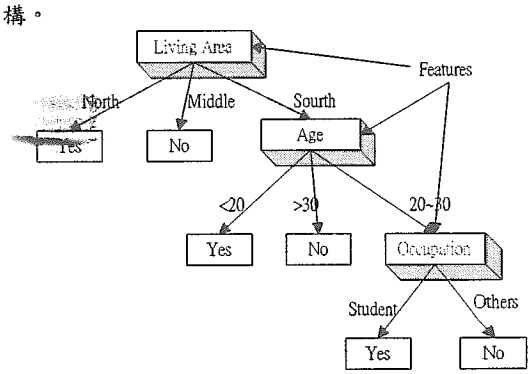


圖 11. 系統分類決策樹架構圖

為模擬的第二部分輸出，說明了住在中部的球迷與住在南部的 30 歲以上球迷是較易流失的群集，而住在南部的年輕上班族是未來最需要開發的潛在球迷群，因為也是最有可能流失的群集。因此，由分類的結果可以很清楚的找出目標可能偏差群集，藉著系統後端助理軟體知識庫的應用，找出產生偏差的原因，並且提供最有效的決策建議。由於系統後端仍處於研究階段，故本研究將以圖 12 來模擬未來第三部份的輸出。

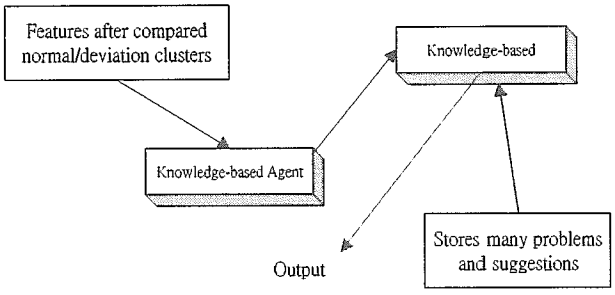


圖 12. 專家知識庫與代理搜尋功能架構圖

因此，根據範例模擬說明後，所有的輸出結果將如上述所示，而未來將推廣至使用各公司的顧客資料來分析，顯示本系統所能應用範圍相當廣泛。且所進行之分析將較一般軟體更為深入與人性化，並且不斷修正使用者需求，期望未來能為所有的公司企業在顧客方面創造更多的利潤。

### 3.7 系統效能評估與討論

本研究在設計完成後將設定下列四項評估指標(metrics)：

1. **群集個數(Cluster Number)**：根據使用者所訂之個數上限來運作，一直到結果出來為止(使用者可以選擇由電腦自動產生)。此為程式運作的評估指標，系統將利用階層式 SOM 來不斷減少群集個數，減少因大量群集所產生之後續分析所遭遇的困難。
2. **屬性個數(Attribute Number)**：最後過濾後的屬性個數若低於設訂值(如預設值為 10)，則表示無法有效從所存在之屬性來分析，即屬性太少則無法過濾出有效的特徵屬性。因此屬性個數也是程式運作的評估指標。
3. **屬性比率(Attribute Ratio)**：將正常集群與偏差集群之相異屬性個數除以總屬性個數，若值太小則表示相異程度太低，因此無法很明顯分析出偏差原因。此為本系統對顧客保留貢獻之評估指標之一，若可以成功地找出相異的屬性來分析，則可以提供公司未來的行銷方向，也可找出過去顧客產生叛離之主要原因。
4. **保留比率(Retention Rate)**：將由最後之偏差集群個數除以總集群個數，與預設值(如設定為 3%)比較，若低於預設安全值則最後放棄搜尋，用意在於能夠真正的做到符合實際的情況化，評估未來是否需要進行進一步之 CR 相關策略，或者放棄目前少量的顧客群以節省成本。觀念可由下列公式來表達：

   ***Retention_Rate =number of possibly defected customersr / total number of customers***

   **If** (Rentention_Rate < default_value) **then** *give u searching suggestions*

本研究結合了四種方法來進行資料的分析，因此最後將經由四項指標來評估系統，預期將會達到較佳的結果。將改善目前無法由電腦自動篩選特徵屬性與分類分析的缺失，也可透過與使用者互動來運作。此四項指標將為本系統未來發展之依據，若所有的顧客保留比率過低，則系統將不給予建議，相同的若屬性個數太少，則系統也無法做進一步的分析，另外兩項指標也是相同的道理，故若以此指標為限制條件的話，系統將會更強大與更互動化。

## 四、後續研究

在未來的後續研究方面，將針對後端的專家知識庫建置以及智慧型助理軟體，預期將朝專家系統來實作，而不單只是簡易知識庫，智慧型代理也將變的更靈活與強大，功能也不只受限於搜尋比對。因此，若將後續研究完成後，本研究之系統架構可以說是一完整的主管資訊系統，提供自動化分析、視覺化處理，並且能有強大的專家系統來支援，讓所提供之建議更有用，更有參考的價值。

## 五、結論

本研究將建構一自動化系統，主要的目的是針對顧客保留與顧客利潤創造，試圖達成視覺化圖形顯現與自動化的分類分析，並且嘗試主動產生有效之決策建議。因此運用了資料挖掘中的視覺化自我組織對應資料分群法，外加上改良的自動標記特徵屬性功能，以及新穎快速之分類決策樹分類法，最後系統後端運用智慧型代理與建立簡易知識庫。將所有的方法串聯起來形成本系統，經由與使用者的雙向互動，學習並分析原始的大量資料，最後轉換成許多有用的資訊呈現給使用者，以期達到自動化學習與提供建議的機制，並在未來後續研究加入專家系統於系統後端，使系統功能更為強大。

## 六、參考文獻


[1] A. Arning, R. Agrawal, P. Raghavan, "A Linear Method for Deviation Detection in Large Databases", Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996

[2] Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant: "The Quest Data Mining System", Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.

[3] Barbro Back, Kaisa Sere, HannuVanharanta, "Analyzing Financial Performance with Self-Organization Maps", In Proc. of Workshop on the Self-Organizing Map (WSOM'97), Espoo, Finland, June 1997.

[4] Bart De Ketelaere, Dimitrios Moshou, Peter Coucke, Josse De Baerdemaeker, "A hierarchical Self-Organization Map for classification problems", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.

[5] Juha Vesanto, "Data Mining Techniques Based on the Self-Organization Map", Thesis for the degree of Master of Science in Engineering, 1997

[6] Kimmo Kiviluoto, Pentti Bergius, "Analyzing Financial Statements with the Self-Organizing Map", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.

[7] K. F. Goser, "Self organising maps for intelligent process control", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.

[8] Kevin J. Cherkauer, Jude W. Shavlik, "Growing Simpler Decision Tree to Facilitate Knowledge Discovery", Appears in Proceeding, Seond International Conference on Knowledge Discovery and Data Mining, Portland, OR: AAAI, 1996

[9] R. Agrawal, R. Srikant: ``Mining Sequential Patterns", Proceeding of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995

[10] Rauber Andreas, "Cluster Visualization in Unsupervised Neural Networks" Diplomarbeit (Master Thesis, in English), Technische Universit Wien, Austria, 1996

[11] Rauber Andreas, "Alternative Ways for Cluster Visualization in Self-Organizing Maps", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.

[12] Rauber Andreas , "LabelSOM: On the Labeling of Self-Organizing Maps" Proceedings of the International

Joint Conference on Neural Networks (IJCNN'99), Washington, DC, July 10 - 16, 1999.
[13] Rauber Andreas , "Automatic Labeling of Self-Organizing Maps: Making a Treasure -Map Reveal its Secrets" Proceedings of the 3. Pacific -Asia Conference on Knowledge Discovery and Data Mining} (PAKDD'99), Bejing, China, April 26--28, 1999. LNCS / Lecture Notes in Artificial Intelligence, LNAI 1574, pp. 228 - 237, SpringerVerlag
[14] Robert Leivian, William Peterson, Mike Gardner, "CorDex : a knowledge DiscoveryTool", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki,.Finland, 1997.
[15] Teuvo Kohonen, Panu Somervuo ,"Self-Organization Maps of Symbol Strings with Application to S peech Recognition", Proceeding of the Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.
[16] Timo Honkela, Samuel Kaski, Krista Lagus, Teuvo Kohonen, "WEBSOM-Self Organization Maps of Document Collections", Proceeding ofthe Workshop on Self-Organizing Maps (WSOM97), Helsinki, Finland, 1997.
[17] Timo Honkela, "Comparisons of Self-Organization Word Category Maps", Proceeding of the Workshop on Self Organizing Maps (WSOM97), Helsinki, Finland, 1997.
[18] Timo Honkela, "Comparisons of Self-Organized Word Category Maps", Proceeding of the Workshop on Self Organizing Maps (WSOM97), Helsinki, Finland, 1997.
[19] Xiaowei Xu, Martin Ester, Hans -Peter Kriegel, Lorg Sander, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Datab ase"
[20] Kaski, Samuel, Method For Exploratory Data Analysis, Samuel Kaski, Teuvo Kohonen, "Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world" In Apostolos-Paul N. Refenes, Yaser Abu-Mostafa, John Moody, and Andreas Weigend, editors, *Neural Networks in Financial Engineering*, pages 498--507. World Scientific, Singapore, 1996
[21] "Building Classification Models: ID3 and C4.5", UGAI97Workshop, http://yoda.cis.temple.edu:8080/UGAIWWW/lectures/C45/
[22] Customer Retention Practices : Solutions http://retention.harrisblackintl.com/solutions/
[23] Customer Retention Associates http://www.customerloyalty.org/
[24] Stuart Russell, Peter Norvig, "Artificial Intelligence Modern Approach", Prentice Hall, p531-p544, 1997