運用資料探勘輔助商品分類之需求預測方法

范有寧
國立臺灣大學資訊管理學系

黃聖祐
國立臺灣大學資訊管理學系

陳靜枝
國立臺灣大學資訊管理學系

摘要

在商業應用中，商品分類幾乎是所有使用與管理商品相關資訊活動的核心。企業普遍會為商品分類，以期透過此種分析模式與歸納方法可以有效提高商品的銷售量並增加企業營收。需求預測在供應鏈管理中扮演重要角色，良好的預測模式將幫助企業有效的存貨管理。然而，以往管理者以質性觀點所建立的商品分類架構無法完全適用於需求預測，貿然將銷售發展趨勢差異太大的商品歸為同一類，會導致類別商品的發展趨勢扭曲或模糊。因此，本研究將提出一以量化觀點為依據的自動化資料探勘模型，建立一套適合需求預測使用之商品分類架構，以達成提高商品銷售量與預測準確度的目標。

本研究以台灣地區知名的茶飲料商與知名連鎖藥妝店的銷售資料為實際案例，進行研究方法之驗證。經實驗結果發現，商品若擁有長期的銷售歷史，且具有明顯的長期趨勢與季節性波動，經由本研究所提出之分析方法可以有效判別商品之間的異同，並加以群集，以提升預測準確度。另外，本研究所提出之研究方法不受產業與商品限制，具有因應不同使用者之背景與未來應用產業的彈性。

關鍵字：資料探勘、需求預測、商品分類、基因演算法、供應鏈管理

Demand Forecasting Using Data Mining Aided Product Classification

Yu-Neng Fan
Department of Information Management, National Taiwan University

Sheng-Yu Huang
Department of Information Management, National Taiwan University

Ching-Chin Chern
Department of Information Management, National Taiwan University

Abstract

Product classification is the core of every information activity related to product management. Companies classify their products according to some attributes for different management functions. Within these functions, demand forecasting is the most critical oneof businessmanagement. Forecasts are essential to the business's decision making and planning processes. Better forecasting can contribute to better price structuring and better inventory management. However, it is a challenging problem owing to the volatility of demand which depends on many factors. Therefore, the study aims to design a classification scheme based on the quantitative characteristics of products and make it more suitable for demand forecasting.

This study proposes a heuristic algorithm, called Data-Mining Aided Product Classification (DMAPC), to deal with aforementioned issues. DMAPC analyzes the sales records by using time-series analysis and searches the optimal product grouping result using GA-based heuristic algorithm. Accordingly, a new classification scheme is constructed by aforementioned processes. The proposed approach is applied to solve two real-world demand forecasting problems from a well-known cosmetic chain retailer and a prestigious tea retailer in Taiwan. The experimental results demonstrate that the proposed approach is proved to enhance the prediction accuracy effectively by applying DMAPC.

Key words: Data Mining, Demand Forecasting, Product Classification, Genetic Algorithm, Supply Chain Management

壹、緒論

分類即是將一群觀察對象依照設定的架構進行指派，使得它們各自屬於某一個群集。這些群集的標籤代表研究者所關心的類別屬性，而分類任務的用途在於新出現的觀察對象可以使用一樣的架構將之歸類，藉此協助決策；或是做為其他分析活動中的步驟，為後續研究提供增益，改善研究的結果。在商業應用裡，商品分類不僅與資料庫設計息息相關，也幾乎是所有使用與管理商品資訊活動的核心，使用不一樣的屬性描述商品會導致相異的商品分類架構，而好的商品分類架構能夠使分類結果更具有意義，進而讓消費者更容易找到心目中最適合的商品並進行購買。

需求管理是供應鏈管理最前端的模組，連結外部市場環境與組織內部的規劃決策系統。需求預測又是需求管理中最重要的功能，預測準確度對企業的獲利影響甚鉅。許多學者針對傳統時間序列趨勢預測方法提出提高預測準確度的方法，其中一個值得注意的是，合併個別商品的銷售紀錄以降低資料的變異性，使用合併的資料進行預測再依適當的比率分配。商品分類架構可以用作合併商品銷售紀錄的依據，在架構中屬於同類的商品即合併成一個類別商品進行預測。然而，任意的商品分類架構並非都適合銷售預測所用。在許多狀況下，廠商的專家或行銷業務人員會依照自己的專業設定分類，但這類型的分類方式往往僅適合財務分析或生產管理所需，貿然採用不適當的分類架構進行銷售紀錄整合，可能會誤將銷售發展趨勢相反的商品置於同一商品分類，以至商品分類整合的銷售發展趨勢產生扭曲或抵銷的效應，進而產生錯誤的預測，導致整體銷售預測模式的正確性與有效性低落。因此，要發展一套適合需求預測的商品分類架構，最重要的就是將各商品連續性的銷售表現納入考慮，以求被分類於同一類別中的商品間具有相似的銷售趨勢，進而達到提高需求預測準確度的目的。

近年來由於電腦硬體運算與儲存科技已發展穩定且廣泛為企業組織所採用，資訊網路與資料庫技術的發展與佈署，更進一步擴大企業所能收集的資料廣度與深度。此外，隨著網路的興起，電子化的出版品數量開始快速增加，例如線上新聞、學術研究論文、電子書、電子郵件訊息、企業內部文件、網頁內容、部落格（Blog）文章，與線上討論區的意見。這些數不盡的電子化資料構成龐大的文字資料庫，不同於商業資料庫中的紀錄，它們多半為半結構化的資料（Semi-structured data），例如每份文件都有標題、作者等等必須記載的資訊，但是摘要與內容這類文字區塊則屬於非結構化，不同文件所使用的描述方式沒有固定的格式。

資料探勘（Data Mining）意指從大量且未經處理的資料中發掘珍貴的知識，近年來被視為一種有效的分析工具與自動化流程，廣泛應用於解決資料充足卻資訊貧乏（Data rich but Information poor）的處境（Han &Kamber 2006）。

商品除了有基本資訊可用於資料分類之外，來自零售商的銷售量紀錄對銷售量預測而言更為重要。這些隨時間不斷累積且具有時間序列（Time series）特性的資料是以往任何知識發掘（Knowledge Discovery）方法所沒有考慮過的。不論是資料探勘或文字探勘（Text Mining）的方法，都是收集大量屬於不同個體的樣本，例如不同的客戶、不同

的生物、不同消費者的購物紀錄、不同文件，而不是同一個觀察對象於長時間的發展趨勢。

因此，從時間序列資料中辨認各項商品銷售模式的相似相異程度並加以區分是本研究的重點。過去少有研究針對調整商品分類架構使其適合供應鏈成員進行多樣商品的整體需求管理與銷售預測，或是研究銷售預測卻只針對少量或單一商品。

因此，本研究將發展一套適合需求管理的商品分類架構，適用於前述的複雜商品銷售發展趨勢與資料型態，以供零售商進行準確的銷售預測，進而達成存貨成本下降，甚至將預測資料分享給上游製造商、配銷商，提升整體供應鏈的效率。同時亦考慮商品之間的關係，例如是否高度相關、相依，還是互補商品，甚至是新商品取代生命週期已經到退出市場的舊商品。相信此套方法模型，能有效協助管理者進行銷售量預測與產品需求管理，進而提升企業整體的營收與提高供應鏈管理的效率。

本研究的最終應用目標屬於需求管理中的銷售預測，商品分類架構的建立屬於銷售預測模式中最先開始的步驟。本研究針對供應鏈末端的零售商，包含一家台灣地區知名的茶飲料商與一知名連鎖藥妝店的銷售資料為實際案例，進行研究方法之驗證，並假設目標商品均具有充分的歷史銷售資料，亦不探討流行性商品或易受促銷活動影響的商品。

貳、文獻探討

一、資料分類之定義

根據Han等學者在（Han &Kamber 2006）中的定義，分類是一種資料分析的任務，透過分類演算法建立分類模型或分類子（classifier），以預測研究者對於觀察目標所關心的類別標籤（categorical labels），例如一件貸款申請是「安全」還是「有風險」，一個消費者會「買」還是「不買」一件特定的商品，一個前來求診的患者應該接受「處方A」、「處方B」還是「處方C」；這些類別可以用離散的數值來表示，其數值順序與大小並無比較上的意義。「預測（prediction）」有時會與分類在字詞上混用，但是在此將之定義為專用於預測連續型數值函式（continuous-valued function），例如預測商品銷售量。兩者藉由預測目標的資料類型加以區分。分類包含兩個主要的階段（Han &Kamber 2006 ; Kotsiantis 2007）:

（一）學習（learning or training）

這個步驟顧名思義是指分類模型將透過分類演算法從一群事先準備好的資料中學習而得，分類模型通常會以分類規則來表示，而分類規則是一群「若則（IF-THEN）規則」的集合。這群資料被稱為訓練集（training set）。每一筆樣本（tuple）以一個n維的屬性向量（attribute vector）表示，例如$X = (x1, x2, \ldots, xn)$。同時有另一個資料庫屬性被指定為類別標籤屬性（class label attribute），擁有離散且無序的數值代表研究者所關心的類別。因為每一個訓練集樣本的類別標籤屬性值都是已知，因此這個步驟也被稱為

監督式的學習（supervised learning）。

（二）分類：

已經建構好的分類模型在這個步驟會經由另一群稱為測試集（test set）的樣本進行準確度檢驗。測試集的樣本通常是隨機從一般的觀察群體中抽取，並且不同於訓練集中的樣本，以避免得到過分樂觀的準確度結果。經過測驗通過研究者所設定的準確度門檻的分類模型，終於能用於預測類別標籤屬性值未知的新樣本；反之，若未達到要求的準確度，則必須返回前一個步驟重新進行學習，採用新的分類演算法或不同的訓練集樣本。

有別於上述的定義，叢集分析（clustering analysis）也是一種分類的方法（Han &Kamber 2006），但是在學習階段所使用的訓練集樣本並不具有已知的類別標籤，憑著分類演算法中的相似度計算將樣本聚集成不同的群集，被稱為非監督式的學習（unsupervised learning）。經過叢集分析所得到的分群結果將不是以準確度作為最終的評量標準，而是以可解釋性與作為其他延伸分析的基礎是否帶來良好的效果來判斷其方法的優劣。

二、 資料分類之準備流程

在正式將所收集到的樣本引入資料分類分析流程之前，樣本資料可以經過下列前處理（preprocessing）的步驟，以提升分析流程與結果的品質（Han &Kamber 2006）。

（一）資料清理（data cleaning）：

這個步驟的目的在於去除或減輕雜訊（noise）對資料分析的影響。樣本資料在收集的過程中，可能因為使用者或資訊系統流程的疏失，導致某些屬性欄位數值有缺漏或是不合理。這些品質不佳的樣本稱之為雜訊，在資料前處理階段可以將之去除，但是如果樣本數量稀少或取得成本較高，較不適用；或將缺漏的屬性欄位填入最有可能的數值，但是此做法必須謹慎控制避免對分析結果造成扭曲的影響。多數的分類演算法都有掌控雜訊的機制，但是經過清理的資料仍然可以減低學習階段的困擾。

（二）相關性分析（relevance analysis）：

儲存於資料庫中的資料在平時交易作業的時候經常會記錄相當多樣的屬性，但是真正在分析時會發現許多屬性欄位彼此之間存在高度的相關性。這些與分類模型相關卻無法提供新資訊的多餘欄位（redundant column），或是根本與分類模型無關者（irrelevant column）（Guyon&Elisseeff 2003 ; Liu & Yu 2005）必須加以處理。特徵選擇（feature selection）不同於特徵抽取（feature extraction），前者僅是選出最佳的屬性子集合，後者則牽涉到將屬性轉換並重新組合成新的屬性。有許多學者針對樣本的特徵選擇提出許多研究，希望能增進預測效能、提出更快更有成本效益的分類模型，並且對產生資料的流程提出更好的解釋。特徵選擇方法有兩大方向（Guyon&Elisseeff 2003 ; Yuan et al. 1999）:一是獨立於主要的歸納分析演算法之外，用過濾器（filter）的方式去除不相關的

特徵；另一個則是在進行歸納演算法的同時，以包裝（wrapper）的方式評估所選擇的特徵子集合，根據適度（fitness）的比較進行調整。Yuan（1999）提出同時採用兩種方法並結合基因演算法的兩階段特徵選擇方法。Hanczar （2003）針對基因序列分析中，樣本數很少，但是基因數龐大的特殊情況，提出從相似樣本的群集中選出代表該群集的原型（prototype）作為資料分類學習階段的訓練集。Swiniarski與Skowron （2003）在粗略集合（rough set）與主成分分析（principle components analysis, PCA）引入特徵選擇的方法協助分類模型的建構。Liu 與 Yu （2005）則在中將現存的特徵選擇演算法依照搜尋策略、評量條件、資料探勘任務分做一個三維的分類架構，並建構了一個統一的平台協助對特徵選擇演算法細節無知的研究者選擇適當的特徵選擇方法。

（三）資料轉換（data transformation）：

同樣屬於數值資料的欄位，為了避免數量級的差距過大造成屬性之間相對的重要性受到扭曲，透過標準化（normalization）將資料轉換至固定間距內（Han &Kamber 2006），例如將原始數值範圍數萬至數百萬的年收入屬性數值轉換為0.0至1.0之間。這種資料轉換方法通常適合應用於衡量樣本點之間的距離。另一個方向是普遍化（generalization），將原本連續的數值資料轉換為離散的範圍類別，例如將年收入依照設定的界線轉換為低中高三種類別。如此一來因為原始的資料受到壓縮，也就降低了分類模型在學習階段的輸入輸出動作，提升學習速度。

本研究假設所收集到的資料不存在屬性值缺漏或可能因為記錄失誤產生品質不佳的狀況，因此可略過資料清理的步驟，將所有資料引入分類模型中。銷售歷史記錄資料的特性為，描述單一資料列的屬性很少，資料筆數相對龐大，所以屬性篩選能提供的助益不顯著。本研究中最重要的資料前處理步驟為，將原始資料轉換為適用於本研究所提出的分類模型。

三、資料探勘分類方法

在資料探勘領域中，分類方法（Classification）被視為一種監督式學習（Supervised Learning）的知識發掘，這類型的發掘方式是具有目標導向性的，在執行分類前有特定的變數是希望被推測的，且希望將某組特定的分類套用至原始資料，並找出特定的關係。常用的分類方式包含許多種，下列將簡述幾種常用的分類方式，並比較各方式間的優劣及適用性。最後，將說明本研究採用基因演算法（Genetic Algorithm）的原因。

貝氏分類模型（Bayesian Classifiers）屬於傳統以統計為基礎的研究方法。其採用統計中的貝氏理論，計算一個樣本在已知的屬性表現下屬於各類別標籤值的條件機率，當一個樣本屬於某一個類別標籤值的機率大於屬於所有其他類別標籤值時，貝氏分類模型就將該樣本歸類為那個類別（Han &Kamber 2006）。當樣本的屬性符合貝氏分類模型的假設，而且相關的機率資料可以取得的情況下，貝氏分類模型與其他分類模型相比擁有最小的分類錯誤率以及最短的分類模型建構時間與分類時間（Han &Kamber 2006）。但是因為樣本屬性之間經常具有相關性，在為這些假設鬆綁的同時，貝氏分類模型的整體

準確度會下降，若為了確保假設而進行屬性轉換則會使得分類方法複雜化失去快速分類的優點。然而，本研究所針對的商品銷售記錄資料有高度的自我相似性，同時也是貝氏分類模型所無法處理的連續型數值，所以統計基礎的貝氏分類方法並不適用於本研究。

決策樹（Decision Tree）是一個適合產生分類規則的方法（Kotsiantis 2007）。決策樹是一個類似流程圖的樹狀結構，包含根節點（root node）、內部結點（internal node），與葉節點（leaf node），節點與節點之間以分支連結。每個內部的節點代表對指定屬性的測試，每一條分支代表測試的結果，葉節點則代表一個類別標籤的值。決策樹的建構關鍵在於每個內部節點應該擺放的屬性測試，選擇的標準是根據每個屬性對分類所能提供的資訊增量（information gain）（Han &Kamber 2006）。決策樹的建構不需要專業領域知識與參數設定，建構與分類過程相當簡單快速，產生的規則也容易被研究者所解讀吸收，而且大致上擁有不錯的準確度，因此獲得廣泛的使用，或當作新分類方法的評比標準。但是決策樹的建構需要品質良好的資料，否則會因為雜訊的影響產生許多小分支，必須經過修剪（tree pruning）才能增進準確度，因此本研究並未採用決策樹做為分類方式。

關聯規則探勘（Association Rules）是一種以規則為基礎的分類模型（Han &Kamber 2006），將常見的關聯規則分析應用於資料分類任務上，將每一個樣本屬性與值的配對（attribute-value pair）視為一個項目（item），在訓練集中尋找頻繁項目集合（frequent itemset）並檢驗其信心（confidence）與支援（support），藉此建構分類模型。此種分類方法必須輸入離散的屬性值，如果該屬性為連續性的數值，也必須先間隔離散化。本研究所使用的樣本資料包含大量的連續型數值，並且擁有時間序列的特性，同一件商品會因為時間不同而有不一樣的屬性表現；另外，本研究所需要的結果並非完整的分類規則，只需要經過分類的商品群集能有效提升後續的銷售預測模式的正確性，因此規則基礎的分類方法不適合單獨使用於本研究。

類神經網路（Neural Network, NN）源自於心理學家與神經生物學家為了在計算模型上模擬大腦神經元所開發出來，類神經網路可以視為一組互相連結的單元（unit），它們之間的連結有輸入輸出的區別還有不同的權重（Li & Wang 2004）。長久以來，類神經網路被批評所建構出來的分類模型難以解讀，而且學習時間較長，選擇此方法進行分類模型建構時必須考慮結果的時效性是否重要；但是在另一方面，類神經網路對於雜訊資料的容忍度很高，即使不清楚樣本屬性之間的相關性也可以使用，再加上分類演算法有平行運算的特性，可以加以發揮縮短運算所需的時間。類神經網路分類模型與規則基礎方法相反，輸入與輸出都需要數值類型的屬性值，如果要處理非連續數值的屬性，也必須加以編碼才能引入分類模型。本研究所要分析的資料型態雖然適用，但是仍然需要加以調整，加速分類模型的建立，因此未被採用。

支援向量機（Support Vector Machine, SVM）是一種新興的分類方法，適用於區分在樣本空間內呈線性或非線性分佈的資料，並且以優異的分類準確度吸引研究者的注意與應用。主要的做法是將原始樣本資料經過非線性轉換到一個高層次（維度數目較少）的空間，再尋找一個能區隔訓練集樣本的最佳線性超平面（Hyper plane）。

另外，尚有以集合為基礎的分類方法，主要包含:約略集合分類法（rough set）與模

糊集合分類法（fuzzy set）兩種。此種以集合為基礎的分類概念能增進分類模型處理不精確的規則或是不完整的樣本資料的能力（Li & Wang 2004 ; Mohanty&Bhasker 2005）。某種程度上可以提升分類模型的強健性，但是同時也增加了衝突發生的情況。如使用模糊集合將原本間隔沒有重疊的屬性數值判斷條件鬆綁，符合間隔重疊區域的樣本將同時符合多個分類規則，可能導出衝突的分類結果。由於本研究最後期望的分類結果是每個商品只屬於一個分類，因此不允許集合所產生的衝突結果。

最後，基因演算法（Genetic Algorithm）（Carvalho&Freitas 2004 ; Yuan et al. 1999）的引入，將研究者所期望的可行解編排成如同生物基因的形式，同時考慮多組可行解，然後模擬生物界交配（crossover）與突變（mutation）的方式造成基因序列變換產生子代（新的可行解），並以適者生存的法則，根據研究者所定義的適度（fitness）作為演算法終止的條件。基因演算法的分類規則搜尋法可能顯得難以控制，不像其他分法是由一個起始解開始漸漸往最佳解的方向搜尋，但是也正因為這種跳動的特性，將創造跳脫區域最佳解（local optimal）而得到全域最佳解（global optimum）的機會，或是避開雜訊樣本的干擾，並且多點搜尋的策略將有效縮短複雜問題的解題時間。適度函式的設計與子代數目的設定是使用基因演算法的重要參數，影響最後所得到的解的品質與求解速度。本研究所必須分析的樣本無論在數量上與屬性維度上都可稱為複雜問題，因此適合引入基因演算法提升分類模型建構時的速度與強健性。

四、預測與分類方法

銷售資料運用在行銷上最常見的功能是預測，亦即使用歷史的銷售資料預測市場未來的發展，進而預測企業未來銷售的消長，做為企業規劃未來發展的預算基礎或是各門市各商品的補貨建議量（Black 2004 ; Sheikh 2002 ; Taylor 2004）。雖然對未來的直覺判斷也可以提供管理者粗略的指標，在大方向上做決策，但是直覺很少可以直接轉變成精確的銷售數字，而各種預測的方法可彌補這種不足。例如超級市場通常在前一年就必須先預測下一年度每種商品的銷售量，由此計劃下一年度配合的賣場架位、倉庫儲位、人員調度、進貨與存貨及促銷活動等等。正確有效的預測，一方面可以使管理者節省時間及成本，另一方面更可以增加銷售擴展利潤。對於零售業而言，正確的預測可說是行銷策略制定的一大利器。

預測的方法首先必須將歷史資料加以分析，從資料中找出過去銷售上之特殊性質，再由這些性質中推算未來的情況。通常歷史資料都具有時間的特性，因此稱為時間序列，亦即是經過長期觀察某些變數的結果。數量化的預測方法即是以上述時間序列的資料為主要分析對象而演化出來的。主要的方法有以下幾種（Black 2004 ; Sheikh 2002 ; Taylor 2004），簡介如下：

（一）前一年同期平移法（Seasonal naiive）：

本方法是將前一年同期之銷售資料做為本期之銷售量。這個方法可能適用於沒有變化的季節性商品，如冰品及火鍋料等，由於方法較為簡單，很多企業或公司都採用這個

方法來計算未來銷售狀況。

（二）前幾年同期平均法（Same Period Averaging）：

本方法是將前面年份同期所有之銷售資料做平均數處理，例如若共有三年銷售歷史資料，則二月份的銷售預測就可以將此三年二月份資料共三筆做平均。這個方法可能適用於季節性非常明顯之商品，如在時裝中春、夏、秋、冬非常明顯之服裝，又如只有夏季出產的水果荔枝與只有冬季銷售之火鍋料等，由於方法特別，在特殊商品的銷售預測計算上比其他方法都有成效。

（三）移動平均法（Moving Average）：

本法必須先決定期數（p），如三期移動平均（p=3）、四期移動平均（p=4）或五期移動平均（p=5）等，期可泛指日、週、月或季。本研究在選擇銷售預測方法時不指定期數，由系統自動搜尋p = 1, 2, 3, …., 6，找出最佳期數，並以此作銷售預測。

（四）指數平滑法（Exponential Smoothing）：

本法需先決定平滑係數及序位（order）。與移動平均法相同，指數平滑法主要目的也是看出移動趨勢。這個方法較適用於沒有變化的長效型商品，如衛生紙等，由於方法較為簡單，很多企業或公司都採用這個方法來計算未來銷售狀況。

（五）趨勢指數平滑法（Exponential Smoothing with Trend）：

此方法需先決定平滑係數，及趨勢係數。與指數平滑法相同，趨勢指數平滑法主要目的也是看出移動與成長或衰退趨勢。

然而，在企業實際執行預測時，銷售量會因為市場的機制與消費者的反應所產生的購買行為而在記錄上產生上升或下降的趨勢表現；長期發展趨勢則可能表現出季節性的成長與下跌或是更長期的循環表現；也可能受特殊事件影響，例如節慶、促銷活動、社會新聞事件，而產生短期異常的發展趨勢導致短期缺貨的問題。商品與商品之間的關係是互補、互斥、還是替代的關係都會影響該商品在分類架構中的歸屬。由相關的研究文獻可知（Geurts&Whitlark 1999；Lapide 1998），在商品銷售預測改進模式中，將單項商品的銷售記錄整合成分類商品銷售記錄可以降低資料的變異性，取得顯著的發展趨勢，而不適當的商品分類架構會造成發展趨勢迥異的單項商品被歸類為同一分類，使得分類商品的銷售發展趨勢受到扭曲或抵銷的效應。由上述可知，商品本身的分類架構，對於判斷未來的銷售趨勢有顯著的影響，若只採用傳統的預測方法，而未考慮商品分類的議題，在實際執行銷售需求預測將產生許多限制之處。因此，從時間序列資料中辨認各項商品銷售模式的相似相異程度並加以區分將是本研究的重點。本研究將透過調整商品分類架構使其適合供應鏈成員進行多樣商品的整體需求管理與銷售預測，以增進企業整體的預測分析結果。

然而，過去許多學者運用前述的資料探勘的分類方法協助各領域的研究，包括醫學研究中區別不同類型的DNA表現（Hanczar et al. 2003；Lee & Lee 2003；Yuan et al.

1999）、網際網路中網頁內容自動分類（Wakaki et al. 2004）、資訊安全中偵測入侵的網路行為（Li 2005），或是商業分析中的資料庫行銷（Lo 2002）。但是大部份的研究所做的分類僅僅是將觀察對象指派給各個預先定義的類別，然後進行解讀，並沒有延伸應用到數值預測上。近年來，有學者Cardoso等人曾經針對報紙的銷售量預測與各銷售點的供補貨建議（Cardoso &Gomide 2007），但也是僅針對單一商品的研究。而本研究的目標，則是尋求足以面對流通業中多達萬種商品的解決方案，而非以單一商品的觀點進行銷售量預測。因此先將商品進行分類將是不可或缺的前處理過程，否則將導致無法負擔的計算與時間資源成本。

五、需求管理相關研究

需求管理在供應鏈中扮演著相當重要的角色，其主要目的是平衡顧客的需求與整體，是供應鏈中重要的管理流程 （Croxton et al. 2002）。將對的管理流程運用在合適的地方，可以事前採取行動，以滿足供給需求及降低執行計畫時所產生危機與對組織的傷害。將有助於提升整體供應鏈營運的彈性（Flexibility）及降低不確定狀況所造成的變異性（Variability）。

好的需求管理機制能夠協助企業或組織事先掌握需求的改變，進而採取適當的行動。就需求管理而言，重要的觀念是找出一些方法可以降低需求的變異性及改善營運環境，使組織更有彈性。降低需求變異性將有助於企業制定計劃時的連貫性與一致性，同時可以降低成本。提高營運彈性是指當組織內部或外部環境改變的時候，能夠幫助企業快速的回應，並提供適配的行動方案 （Lambert 2006）。

於今日快速變遷的社會環境中，以顧客導向而衍生的變異性是無法避免，因此需求管理的目的就是以更有效率及快速的方式，滿足顧客的需求。這無疑能直接影響企業、供應商與顧客的利潤。適當的流程改善，能幫助將對的商品提供給需要的顧客，有助於銷售業績及顧客忠誠度的提升 （Zinn& Liu 2001）。由此可知，好的需求管理機制將有助於需求預測，準確的預測模式將降低原物料及成品的存貨，可以降低物流成本、提高資產利用率。這些流程的改善不僅僅是依賴企業內部的活動，更需要依靠企業外部供應鏈上夥伴的配合。因此需求管理不能只是完全依靠企業內部的流程，而是必須跨越組織、跨職能的整合。

Chopra與Meindl （2003）指出，預測需求是所有供應鏈中策略與規劃決策的基礎。Sheikh （2002）提出的生產資源規劃（Manufacturing Resource Planning, MRP II）架構中亦指出，需求管理（Demand Management）是所有規劃模組的起點。不論是主生產排程（Master Production Scheduling, MPS）、物料需求規劃（Material Requirements Planning, MRP），或是採購運輸排程的研究，都假設擁有良好的需求管理，也就是準確的銷售預測，然後才宣稱可以透過他們的研究與提出的模式方法達成整體供應鏈管理的目標。

然而，在過往的研究中，研究需求管理、改進銷售預測準確度者並未考慮商品分類架構是否適合預測所需，僅依靠廠商或專家所提供的解釋性分類架構進行銷售記錄整

合，可能造成個別商品之間的銷售發展趨勢差異模糊扭曲類別商品整體的銷售發展趨勢；研究分類演算法者僅專注於處理研究對象的靜態屬性，尚未將研究對象延伸至隨時間累積變動的時間序列資料。因此，本研究試圖改進前述兩者不足之處，找出符合最小距離模型的最佳分類架構。為了更有效率地解決最適產品分類架構問題，本研究將採用啟發式演算法，以期在可接受的時間範圍內求得最佳解或近似最佳解。

六、需求預測成果評估方法

應用於需求預測的研究都希望提升預測準確度。因應不同的情境與資料特性，預測準確度有數種量化的評估方式，以下列四種最為常見（陳靜枝&蔣明晃 2005；Keller 2005）：

（一）平均方差（mean squared error, MSE）：

預測值與已知歷史記錄差之平方平均值。平方的做法雖然避免了高估與低估效應互相抵消的問題，但是也使得大誤差更為顯著。

（二）平均絕對差（mean absolute deviation, MAD）：

預測值與已知歷史記錄差之絕對平均值。計算方法簡單，但是會受觀察對象數量級不同的影響，例如預測數千到數萬之間的誤差與預測數十到數百的誤差，可能前者所得出的MAD較大但是並不代表預測準確度較低。

（三）平均絕對百分比差（mean absolute percent error, MAPE）：

已知歷史記錄與預測值差之比值絕對平均值。採用比值的作法避免了數量級不同的影響，將誤差絕對值轉換為與歷史記錄的百分比。但是所有使用比值的計算公式都必須考慮分母不得為零的限制，因此採用MAPE時也必須注意。

（四）最大絕對差（largest absolute deviation, LAD）：

預測值與已知歷史資料記錄之差最大絕對值。與前述三種方法相比，計算方式最為簡單。但是只留下最大絕對值的作法在多數的情境下捨棄了太多資訊，容易導致偏差的分析結論，或受原始資料誤差影響。

依照學者Kahn （1998） 的研究顯示，四種比較標準中以MAPE最廣受一般企業採用，且皆是使用數量進行計算。本研究將採用平均絕對百分比差（MAPE）來驗證商品分類架構對於銷售量預測準確度的改善效果，其目的在於去除不同商品銷售量數量級之間的差異，並用實際銷售值為分母的公式，以避免過度高估銷售的狀況。

本研究將運用資料探勘的分類方法找出最符合目標的商品分類架構，並且應用於商品銷售量預測分析。資料的前處理步驟中，資料轉換是本研究的重點，使得原始銷售歷史記錄可以適用於所提出的模型。另外，透過上面的整理比較，本研究將結合距離基礎分類方法，並搭配基因演算法提升最佳解的搜尋效率，最後在各個需要評比的階段使用MAPE做為量化的標準。

參、問題描述與最小距離群集模型

一、問題描述

本研究將分析商品基本資訊與銷售歷史紀錄，進而建構一商品分類架構。依此分類架構，每個商品會被歸屬於一個類別，然後同類別的商品銷售紀錄會在後續的預測分析中整合形成單一的類別銷售量進行預測，最後再將類別預測銷售量分配給單項商品做為最終的銷售量預測。此分類架構所預期產生的商品分類是：銷售表現相近的商品會屬於同一個類別。如此將避免進行銷售預測時，在整合分類商品銷售歷史紀錄時，誤將銷售表現迥異的商品置於同一分類，造成該分類商品銷售表現的扭曲或產生互相抵銷的效果，然後因此產生錯誤的預測，導致整體銷售預測模式的正確性與有效性低落。

在商品資訊方面，可分為預先定義的資訊，即上市時就已經確定，由供應商所提供的資訊，包括建議售價、銷售地區、廠商定義的商品管理分類架構；與隨著該商品持續出現在門市通路上不斷累積的銷售量歷史紀錄，形成一串連續性的時間序列數值資料。

為了去除上市時間先後不同所造成的歷史紀錄資料量不一的影響，本研究採取時間序列趨勢分析方法，將每項商品的銷售量歷史紀錄轉換為下列四種指標：長期趨勢指標（T）、循環性指標（C）、季節性指標（S）、不規則性指標（I）。其中的循環性指標與不規則性指標因為難以有效量化其效應，故在本研究中不列入考慮。因此，可以將將銷售量（Y）寫作一以時間（t）為自變數的線性乘法模式：

（1）$Yt = (\beta 0 + \beta 1t)\, St + \varepsilon$

其中S_t代表在一年之中，固定週期長度的漲跌波動；β_0與β_1代表去除季節性波動影響之後所呈現的長期趨勢。

二、研究範圍與限制

本研究的最終應用目標屬於需求管理中的銷售預測，商品分類架構的建立屬於銷售預測模式中最先開始的步驟。因此本研究針對供應鏈末端的零售商，以一般商品流通業為主要研究對象，目標商品必須具有充分的歷史銷售資料。

本研究不探討流行性商品或易受促銷活動影響的商品，亦即不分析流行期間的起訖與量化效果幅度；促銷活動與預期之外新聞事件爆發對商品銷售量的影響視為不規則變動，在分析的過程中以其他方法去除。

本研究假設所收集之歷史銷售資料具有良好的品質，即不存在錯誤或不完整的資料。零售商擁有足夠的資訊科技建設與能力，能記錄每日的商品銷售記錄，並且透過網路連線將各分店記錄匯集至一整合資料庫，呈現單一商品在最小預測時距中全區銷售量總和的觀點。

本研究針對長銷型商品，希望是擁有長達一年以上銷售記錄的商品，才具有足夠的資訊進行有效的季節性與長期趨勢分析，但是當面對銷售不滿一年的商品，或銷售量起伏變動劇烈，無法有效以長期趨勢與季節性來描述時，將降低本研究所提出的效果。

三、最小距離群集模型

根據上述之問題描述，本研究以最小距離群集模型來呈現問題。

（一）參數

P：　表示待分類商品所構成之集合

$N(P)$：　表示待分類商品數目

M：　表示欲區分的群集數目

M_{min}：　表示群集數目之下限。$M_{min} \geq 1$

M_{max}：　表示群集數目之上限。$M_{max} \leq N(P)$

G：　表示所有可行的分類結果構成之集合

G^m：　表示指定分為m群所有可行分類結果構成之集合

G^mC：　表示前述分類結果中，屬於群集編號C的商品所構成之集合，$C = 1 \sim m$

$N(G^m{}_C)$：表示前述分類結果中，屬於群集編號C的商品數目

C_i：　表示商品i所屬於的類別群集編號，$i \in P$，$1 \leq C_i \leq m$

B_i：　表示長期趨勢分析中，商品i所採用描述長期趨勢的係數個數。根據每項商品所選擇的最適關係式，各自有不一樣的係數個數：若選擇線性，則$B_i = 2$（β_0、β_1）；若選擇二次式，則$B_i = 3$（β_0、β_1、β_2）

K_i：　表示季節性分析中，商品i所採用一年的季數。若單季為三個月，則$K_i = 4$；若單季為一個月，則$K_i = 12$

S_{iq}：　表示商品i第q期的季節性指標，$i = 1 \sim N$，$q = 1 \sim K_i$

T_{ib}：　表示商品i第b個長期性指標，$i = 1 \sim N$，$b = 1 \sim B_i$

（二）決策變數

（1）TS_i：表示描述商品i銷售發展趨勢的特徵向量。

若使用某個參數設定時，有不需要的屬性，則該屬性值指定為0。例如使用簡單線性迴歸與四期季節性描述時，T_{i3}，也就是時間序列模型中係數β_2，與季節性指標$S_{i5} \sim S_{i12}$皆指定為0。同時，因為模型中用來描述趨勢線截距的參數β_0並不影響後續預測的準確度，因此從特徵向量中移除。又為了避免描述趨勢線斜率的係數數量級過大的影響，導致季節性指標的差異被掩蓋，因此趨勢線斜率的係數必須經過標準化，將係數除以所有待分類商品的係數平均值$(\bar{\beta}_1 、 \bar{\beta}_2)$，形成一平均值為1的長期趨勢指標。

（2）$TS_i = （T_{i2}, T_{i3}, S_{i1}, S_{i2}, S_{i3}, S_{i4}, \ldots, S_{i12}）$

d_{ij}：表示商品i與商品j在特徵向量空間內的距離。

根據各商品所選擇的最適長期趨勢關係式與季節性分析的季節數不同，使用不同參數描述發展趨勢的商品即屬於不同類別。因此，將原始銷售量歷史記錄轉換為時間序列分析指標之後，這些商品至少可依長期趨勢分析分作遞增與遞減；依季節性分析分作無季節性、四期季節性、十二期季節性三個類別。不同參數設定下所計算得出的指標之間並不存在合理的轉換關係，例如在長期趨勢分析中，採用一次線性關係與二次函式所得

出的β_1所代表的意義完全不同（在一次線性中代表遞增或遞減趨勢，在二次函式中則代表二次曲線最高最低點水平位移的幅度），或是在季節性分析中透過資料縮減或添補的方法使得四期季節性與十二期季節性得以互相比較。所以，唯有歸為同一類別的商品可以計算之間的距離，而商品i與商品j兩者之間的距離。

（3）$dij = |TS_i - TS_j|^2$

藉由採取平方和的方法去除商品之間指標值差異正負號加總時的抵銷效應，同時突顯較大的dij對整體模型的影響應該愈大，應該優先考慮將彼此之間距離較遠的商品分為不同類別。

（三）限制式

（4）$M_{min} \le m \le M_{max}$

若群集數目太少，會不足以將發展趨勢相異的商品區分為不同類別，來避免類別商品發展趨勢扭曲的問題，因此需要選擇M_{min}做為下限；若群集數目太多，會使得每個商品自成一類，無法達成合併商品記錄來提升預測準確度的效果。M_{min}與M_{max}的選擇方法留待演算法中討論，經過不同的情境實驗分析之後提出建議值。

（四）兩階段目標函式

（4）兩階段目標函式

第一階段：最小化同群集內的樣本間距離總平均

（5）$Min\ aTDG = \frac{2}{\sum_{c=1}^{m} N(G_c^m)(N(G_c^m)-1)} \sum_{i=1}^{N(P)-1} \sum_{j=i+1}^{N(P)} d_{ij},\ \forall g \in G^m$

$aTDG$代表在分類結果g下，所有商品與屬於同一群集的其他商品彼此之間的距離平均，屬於不同群集的商品之間的距離則不加以計算。在解讀意義上，希望被歸為同一群的商品彼此愈相似愈好。

第二階段：最大化不同群集之間的距離總平均

（6）$Max\ AGD = \frac{2}{m(m-1)} \sum_{k=1}^{m-1} \sum_{j=k+1}^{m} d_{\bar{k}\bar{j}},\ \text{for } M_{min} \le m \le M_{max}$

AGD代表所有群集之間的平均距離，而每個群集用於計算距離的虛擬代表點$(\bar{k}、\bar{j})$的座標為同一群集的所有商品屬性指標的平均值。在解讀意義上，希望不同群集之間的差異愈大愈好。

本研究若只考慮第一階段的最小化目標，其最佳化方向會建議每個商品自成一個類別，$aTDG$等於零。但是這並非最適合的分類結果，一來將失去合併銷售記錄提升預測準確度的效果；二來若有銷售記錄較短的商品加入分類，無法將它放入適當的類別，提供資訊給後續分析使用。因此加入第二階段最大化目標，只有當這兩階段目標達成平衡時，才是最適當的分類結果。

肆、啟發式分類演算法

研究提出一啟發式演算法，以時間序列分析進行分類所需之資料轉換，並根據其分析結果建構分類階層。另外更以基因演算法為基礎搜尋最適之分類結果，以建立商品分類架構。最後，希望將此分類架構運用於需求預測，進而提升預測之準確度。

演算法流程主要分為三大部分，如圖1所示：第一部分為前置作業，將單一商品每日銷售紀錄轉換為時間序列指標；第二部分為分類演算法，根據前一部分所轉換的指標將銷售發展趨勢相似的商品加以群集分類，建構一分類架構；第三部份為預測效果評估，依照丁恬文（2007）所提出的銷售預測流程，使用第二部分所建構的分類架構整合銷售紀錄，再次計算分類商品最佳參數並產生預測值，最後以MAPE做為預測準確度的量化標準。

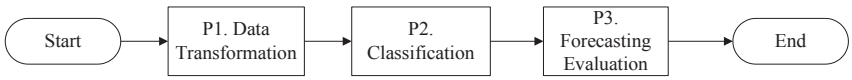

圖1：啟發式分類演算法流程

一、前置作業

各商品因為上市日期前後不一，所擁有的銷售歷史紀錄量也因此不同，僅以銷售歷史紀錄無法形成品質良好的資料引入資料探勘方法模型進行分類。為了去除這個問題，本研究採用如圖2所示的資料轉換流程，將各商品的銷售歷史紀錄轉換為時間序列指標。為了找出最適合描述商品銷售發展趨勢的時間序列模型，本研究假設銷售可能具有12期季節性、4期季節性或無季節性，去除季節性之後以簡單線性與二次式分析長期趨勢。最後，以平均絕對百分比誤差（MAPE）做為評量標準，為每個商品選擇一組預測誤差最小的，也就是最適合描述該商品銷售發展趨勢的指標組合。

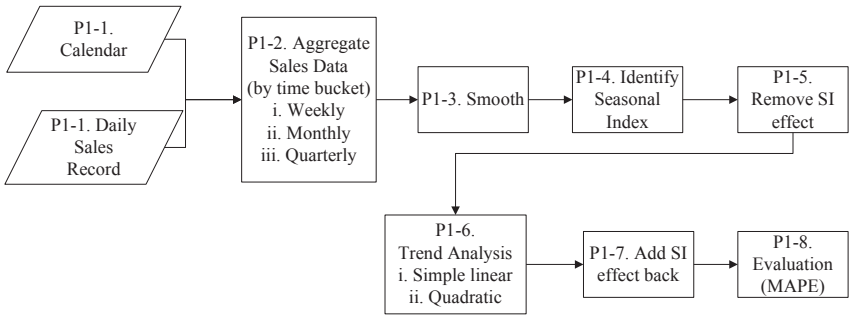

圖2：資料轉換流程

二、分類演算法(DMAPC)

在此階段，將設計一階層式分類架構。前兩層為規則基礎的分類，最後一層則以上一節所定義的最小距離群集模型為標準，將相近（相似）的商品群集成為一類，如圖3所示。

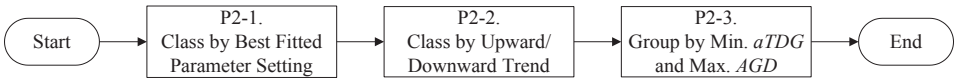

圖3：階層式分類架構建立流程

同的時間序列指標設計模式下所產生的參數值彼此之間並不存在合理的轉換方式，因此分類架構的第一層（P2-1.）即為依照各商品最適合的分析模式組合予以區分。本研究所希望避免的不適當的分類結果，其中之一即是銷售發展長期趨勢向上成長與向下衰減的商品被歸為同一類，將造成分類商品整合的銷售發展趨勢互相抵銷，長期趨勢因此顯得趨緩，無法產生準確的銷售預測。基於前述理由，在此分類階層（P2-2.）將前一階層所區分的各個子分類中的所有商品依照長期趨勢指標表示成長或衰減再予以區分。分類階層第三層（P2-3.）是最主要的分類步驟，使用資料探勘的方法，藉由定義的「距離」遠近來判斷兩商品的銷售發展趨勢是否相似，而適合聚集為一個分類，為後續的銷售預測流程創造良好的基礎。

本研究採用基因演算法為基礎的最適分類搜尋方法。大致來說，針對同一群商品，在數個分類結果中，經過評估選出較好的分類結果進行調整來產生新的分類結果。如此一來改善了純粹使用規則基礎（rule-based）的分類方法缺乏彈性的缺點——一旦決定了，就無法再調整分類結果，即使根據目前的規則只能找到區域最佳解（local optimal）而不是全域最佳解（global optimum）。

此方法的流程如圖4所示，共包含兩個迴圈：一開始指定分群數目（m），將分類結果進行編碼，使之成為「染色體」的形式，然後創造出兩條起始的染色體。每次產生新的染色體都要進行評估，捨棄最差的染色體並找出最優質的染色體進行基因演算法中的交配（crossover）或以突變（mutation）來產生子代，也就是新的分類結果。評估然後繁衍的循環不斷重複直到終止條件被滿足為止。內層迴圈的最佳分類結果進行第二階段評估，然後重新設定m，再次進行內層基因演算法搜尋，直到外層迴圈的限制被滿足。以下詳細說明各細部流程：

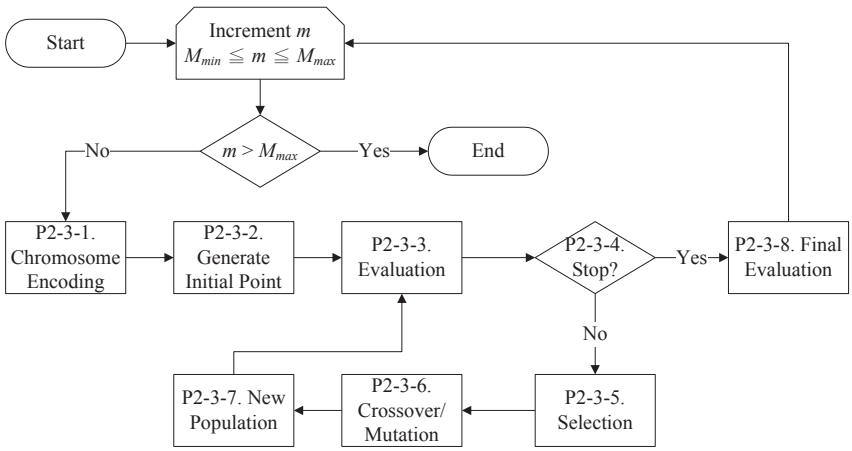

圖4：最小群集分類演算法流程

P2-3-1. 染色體編碼

由於基因演算法皆以染色體（chromosome）為執行之個體，在演算法開始之前，必須先定義染色體的編碼方式，用以表示不同的分類結果，然後依此建立資料結構，以便進行後續繁衍的動作。本研究的染色體編碼方式如表1所示，假共有N個商品，則染色體會有N個基因；位置i的基因代表任意排列的商品清單中列於第i個商品所屬的群集C_i，染色體即C_i所組成的序列。群集的編號僅具標示區別的意義，並不表示順序，C_i相同表示所對應的商品屬於同一群。

表1：染色體編碼範例（10個商品共分成3群）

i	1	2	3	4	5	6	7	8	9	10
Item No.	A	B	C	D	E	F	G	H	I	J
C_i	1	1	2	1	1	3	2	3	1	2

P2-3-2. 產生初始染色體

當決定了染色體的表現方式之後，要產生初始的染色體以啟動演算法，做為「祖先」繁衍第一代。因為效率是本研究很重要的評量指標，因此初始染色體的品質對最後產生的結果有很大的影響。本研究採用兩種方式各產生一組初始染色體，隨機產生與規則基礎：（假設要將N個商品分為m群）

1. 隨機產生：染色體的每個基因透過擲一個公平的m面骰子來決定C_i，表示每個商品屬於任一群集的機率相等。
2. 規則基礎：藉由觀察商品的屬性值，令距離較遠的商品屬於不同的群集，然後各自招攬較近的商品與之形成同一群集，以避免傳統距離基礎分類方法會綁定彼此距離較近的樣本不再分開的缺點。

P2-3-3. 評量染色體

因演算法的精神來自於模擬自然界「物競天擇，適者生存」的規則，所有產生的染色體都必須經過設計的適度函式（fitness function）評量，以決定誰應該被淘汰，誰應該留下來繁衍下一代。本研究採用的適度函式即最小距離群集模型中的第一階段目標函式，群集內總平均距離（*aTDG*）。適度函式值愈小表示該染色體品質愈好，因為他最符合本研究的目標。

P2-3-4. 檢查終止條件是否滿足

演算法是否停止繁衍與評量新的染色體將會在這個步驟決定。本研究選擇兩種終止條件：其一為跟效率息息相關的染色體繁衍代數，若可用的代數消耗殆盡，表示分類演算法已耗盡可接受的時間長度，因此將目前最好的染色體分類結果傳出；其二為最佳的分類結果連續維持某特定代數，若該分類結果在多少次產生後代的的過程中，仍然沒有比他更好的分類結果，演算則終止。

P2-3-5. 篩選染色體

經過評量之後，且演算法尚未達到終止條件時，為了保持族群的品質，並保留新成員加入的機會，會將適度最差（aTDG最大）的前兩名染色體從族群中移除。

P2-3-4. 檢查終止條件是否滿足

為了尋求更好的分類結果，在這個步驟透過染色體交配或突變來產生新的染色體。

1.交配：交配的目的在於希望子代（offspring）保留優良親代（parent）的部分特徵，也就是分類結果，然後因此得到更好的適度表現。在進行交配時，以經過評量之後適度最好的兩條染色體做為親代，在N個基因裡隨機選擇一個點，做為交配的分界點進行單點交配（one-point crossover），如表2所示。

表2：染色體交配範例(*i*=7為交換點)

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Parent1	1	1	2	1	1	3	2	3	1	2
Parent2	1	1	1	1	2	1	1	1	3	1
Offspring	1	1	2	1	1	3	1	1	3	1

2.突變：突變則不保留任何跟族群裡現有染色體相關的訊息，當演算法決定在此代產生突變，將如同隨機產生初始染色體的做法產生一個突變種（mutant），希望能產生跳躍的效果，一舉將解的搜尋區域帶往另一個更好的地帶。

P2-3-7. 更新族群成員

完成染色體篩選與子代的繁衍之後，移除適度最差的染色體，將經由交配或突變產生的子代加入族群，以形成新一代的族群，然後再次進行評量的循環。

P2-3-8. 終止後評量

當分類階層第三層的演算法達到終止條件，將用前述所提之第二階段目標函式評量

最後產出的染色體，群集間距離總平均（*AGD*）。

在此分類階層搜尋最適分類結果有一個限制，就是必須在固定分群數（*m*）之下進行搜尋，然而*m*也限制了分類結果的樣貌，因此也必須調整*m*——從M_{min}到M_{max}——進行搜尋，最後以*AGD*做為比較基準，能夠產生最大*AGD*值的*m*與其對應的染色體就是那N個商品在這分類階層中最適當的分類結果。

三、複雜度分析

本研究之演算法主要包過前置作業、分類架構建立與預測效果評估。其中前置作業在每次執行本演算法時只會做一次，預測效果評估也只會做一次，對整體演算法效率影響極小；最複雜的部分即是以基因演算法為基礎的分類架構建立演算法，本節將針對此部分做複雜度分析，重要參數如下：

N(P)	表示待分類商品數目
T_i	商品*i*所擁有的銷售歷史紀錄總期數
m	表示欲區分的群集數目
M_{min}	表示群集數目之下限
M_{max}	表示群集數目之上限
GC	產生有效演化的代數
p	產生突變的機率

前置作業的部分，必須將每個待分類的商品的銷售歷史紀錄轉換為時間序列指標，每個商品有T_i期紀錄，共*N(P)*個商品，所需的矩陣運算的時間複雜度為*O(T*N(P))*。在分類演算法中，以亂數產生初始染色體所需的時間極短，而以規則基礎產生的初始染色體需要建立商品兩兩之間的距離表，以及多次搜尋距離表中最大值，其時間複雜度為*O(N(P)2)*；在染色體演化的過程中，每一代的交配與突變動作皆相當簡單，其時間複雜度可寫作*O(N(P))*；最不理想的狀況即是必須完全耗盡系統所設定的有效演化代數*(GC)*才能達成此迴圈停止條件，產生品質足夠好的解。前述 分類動作所預設的區分群集數目*(m)*必須從M_{min}試到M_{max}，才能決定P集合中的商品的最佳分類結果，因此分類架構建立的時間複雜度為$O(N(P)^2 + (M_{max}–M_{max})*GC*N(P))$。

經過上述前置作業與分類架構建立的時間複雜度分析，所有變數皆不在指數項，因此本研究所提出的演算法不為NP演算法。

伍、實例分析

本研究的最終應用目的是為了提高商品銷售量預測的準確度，針對擁有足夠銷售紀錄的一般商品，使用陳靜枝與蔣明晃(2005)所建置的需求預測學習系統進行驗證。首先

以銷售歷史紀錄中移除最後一個月的資料做為訓練資料集進行最佳預測模型學習，最後模擬產生最後一個月的預測銷售量，與資料庫中的實際值比較，以MAPE做為預測誤差的量化標準。

在此實例分析中，我們選擇知名茶飲料商與知名連鎖藥妝店之目的是考量兩大不同的商品性質。以茶飲料商為例，其商品性質多屬於具有長時間銷售歷史之長銷型商品；而藥妝產業之商品，則是屬於銷售歷史較短之流行性商品。因此這樣的案例設計方式，主要是用以證明，此研究方法足以套用至不同銷售類型之商品，藉以概化此研究方法之價值。在此實例分析中，我們將比較不同的分類架構，包括廠商所提供的分類架構與本研究所提出之DMAPC建立之分類架構。由實驗結果得知，根據不同的分類架構，將導致統整得出的類別銷售量擁有不同的趨勢發展，並可能適合於不同的預測模式。因此最後將以單項商品的預測誤差做為標準，比較使用不同分類架構所產生的差異。

一、實際案例-知名茶飲料商

此案例所提供商品與銷售歷史資訊共包括319個商品，原本以三層的分類架構區分各項商品——產品型態、產品口味、產品包裝規格，總共有58個小分類。茶飲料商品具有生命週期長，銷售量大且季節性波動趨勢明顯，同類商品在長期趨勢與季節性表現上相似，因此不調整分類架構即可得出可接受的預測結果。

經過DMAPC分析之後，產生30個小分類，少於原本的58個。其中以選擇了12期季節性的商品佔了半數，選擇4期季節性佔了三分之一，其餘才是無明顯季節性。

仔細比較廠商所提供的分類架構與DMAPC所分析的分類結果，可以發現原本屬於同類別的商品，其實在長期趨勢的正負成長與季節性波動描述上皆有所不同，如表3所示，應該將其分開。

表3：分類結果差異分析範例

ItemNo	Trend (P: Positive, N: ative)	Seasonality
A	P	12
B	N	12
C	P	12
D	P	12
E	N	4
F	N	0
G	N	12
H	N	12
I	P	12
J	P	12

學習階段的評量結果，以累積銷售金額排名前百大商品為比較對象，原本的分類架構的平均預測誤差為113.99%，而使用DMAPC所建立的分類架構的平均預測誤差為63.54%，表現最差的是隨機任意群集的結果，平均預測誤差高達195.46%。

使用學習後建議的最佳預測模式與參數實際模擬預測往後30天的銷售量，同樣以累積銷售金額排名前百大商品為比較對象，原本的分類架構所產生的平均預測誤差為67.90%，而使用DMAPC分析所得的平均預測誤差為81.61%，最差的仍然是使用隨機指派的結果，為119.52%。若個別比較則可以發現，前百大暢銷商品中，使用DMAPC有73項商品的預測誤差低於使用原本的分類架構，或是不大於10%。

藉由此案例可以發現，DMAPC面對長期趨勢與季節性明顯且紀錄時間長、銷售量穩定的商品集合時，可以有效分辨出應該群集在一起的商品，進而有效提升預測準確度。

二、實際案例-知名連鎖藥妝店

此案例在其擁有的通路店面販售不同品牌的彩妝商品及保養品，為了管理方便，廠商以品牌及產品用途予以分類，共有39個小分類。此案例的商品銷售量起伏變動大，且經常有短期的異常銷售高峰出現；另外，商品生命週期短，款式種類繁多且雜亂，下市之後通常由新商品取代舊商品的地位，因此較難以預測。

經DMAPC分析後，此案例共約450項商品被分作30個小分類，略少於原本所分的39個小分類。個別商品在分析過程中所選擇的長期趨勢與季節性組合模式以無季節性與12期季節性佔了絕大多數。因為本案例的銷售紀錄特性，比較結果發現，DMAPC的效果不大，多數商品因為紀錄過短的關係無法分析出有效的季節性波動，所以原本同類的商品在DMAPC的分類結果中仍然屬於同類。

本研究在此案例中，再加入單項商品直接進行學習預測的情境與其他分類架構比較。預測百大暢銷商品未來30天銷售量，以DMAPC所產生的預測值最為準確，平均預測誤差為32.95%，優於原本以品牌分類的39.63%；若不做銷售紀錄合併直接進行學習，所產生的平均預測誤差最大，達57.59%。

三、效率分析

本研究所提出的啟發式演算法以第二階段的最適分類結果搜尋最為耗時，其所要花費時間受系統參數設定影響，包括搜尋範圍（M_{min}～M_{max}）與搜尋停止條件（演化總代數與最佳解停留代數）。

假設有N個待分類的商品，且N大於30，完整的搜尋範圍應該在兩個不分類的極端狀況之內，也就是將（M_{min}, M_{max}）設為（2, N-1）。但是經過實驗發現，最適分類結果的分群數都不曾出現於超過$N/2$，甚至不曾超過10，如圖5所示。因此建議將搜尋範圍設定為（2, $N/2$-1）；若N的數目不多，即可將所有非極端分類群集數搜尋過。

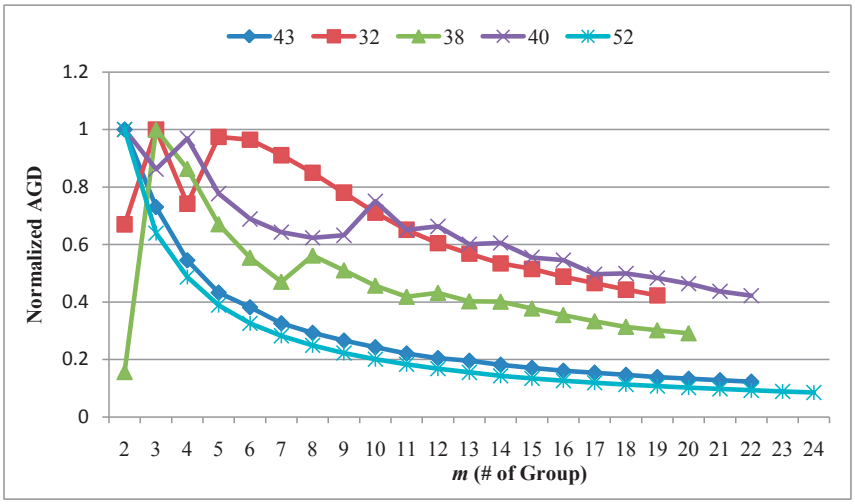

圖5：建議搜尋範圍（數列代表不同的待分類商品數目）

以基因演算法為基礎的最適解搜尋方法以競爭為原則進行，從這一代進行到下一代必須以一個新的可行競爭者出現才算數。經實驗發現，產生的初始染色體其品質優良可以保留在族群裡數十代，或是在前期就經由交配產生更好的染色體。因此建議將演化總代數設為200即可。若要更加縮短搜尋時間則可設定最佳解停留代數，惟此代數必須小於總演化代數的一半才有意義。此外，使用前述建議系統參數進行分類，兩個實際案例都可以在十分鐘內完成分類架構建立，也證明本研究所提出之啟發式分類演算法其執行效率是足以被接受並可使用於真實世界的案例中。

陸、結論

本研究透過分析商品的銷售發展趨勢，以時間序列指標——長期趨勢及季節性波動——做為該商品的特徵，藉以群集相似商品並建立分類架構，進而達成改善商品預測準確度的目標。由文獻探討得知，過去對於時間序列資料預測方式之研究，皆未探討分類架構對整合商品銷售紀錄的影響，因此常發生將長期趨勢發展相反的商品歸為一類，使得整個類別商品的長期發展趨勢遭受扭曲或模糊，進而致使預測準確度降低；另外，在過往的資料探勘研究中，也少見將研究範圍延伸至時間序列型態的資料，由於不同的觀察對象並非在所有的觀察時間內都有觀察值，導致此種資料型態難以引入資料探勘的分析模型中。因此，本研究所提出的模型架構，試圖解決過往研究不足之處，將商品銷售歷史紀錄轉換為數個時間序列指標，以形成描述該商品銷售發展趨勢的特徵向量，使得原本擁有不同數量屬性的商品可以適用於資料探勘分析模型，並加以計算彼此之間的相似度。

本研究針對最適分類結果提出一兩階段最佳化目標模型：在內層固定分群數目時，尋求群集內樣本距離總平均最小化；然後以不同群集之間距離總平均最大者為外層目標。在這個模型中，必須搜尋的可行解集合大小將以待分類商品數目為指數的速度成長，以全域搜尋法在商品數目達到實際應用規模前，就無法使用有限的運算資源於可接受的時間範圍內找到最佳解。本研究針對此限制提出一啟發式演算法，以基因演算法為基礎，大幅縮短近似最佳解搜尋時間，並且足以處理具有實際應用規模的商品數量。

透過兩個實際案例的測試結果得知，本研究可以在合理的時間內找出一組可行的分類結果，而新的分類架構在匯入需求預測學習系統後，經驗證得知將有效提升學習系統的品質，並為類別商品選擇較好的預測模式，以提升整體之預測準確度。

在實務層面上，適合引入本研究的商品類型並未限制，商品本身只要擁有足夠的銷售歷史紀錄進行資料轉換即可，因此本研究所提出的方法可以適用於各種產業類別的商品並具有因應不同使用者之背景與不同產業之彈性。本研究建議商品項目龐雜的使用者採用此分類分析方法，來區分不知道如何分類的商品集合，以得到較好的需求預測結果，並且同時藉此觀察商品在銷售發展趨勢上的異同，可協助改進企業供應鏈管理中的重要功能，如:存貨管理、訂補貨策略制定與調整等，將有助於提升整體供應鏈效率。

本研究主要針對長銷型商品，希望產品是擁有長達一年以上的銷售紀錄，具有足夠的資訊進行有效的季節性與長期趨勢分析。因此當面對銷售歷史不滿一年的商品，或銷售量起伏變動劇烈，無法有效以長期趨勢與季節性來描述時，將降低本研究所提出的效果。因此期許在未來的研究中，可以考慮加入其他商品銷售趨勢分析因素，例如去除流行性影響，使整體銷售趨勢更加明顯，或另外研究另一套方法處理銷售紀錄較短的商品與其他商品之間的相似度定義與分析。

參考文獻

1. 丁恬文，民96，流通業協同規劃預測補貨解決方案，國立台灣大學資訊管理研究所碩士論文。
2. 陳靜枝、蔣明晃，民94，需求預測模式之研究期末報告，財團法人工業技術研究院。
3. Black, K.*Business Statistics for Contemporary Decision Making* (Fourth Edition), John Wiley & Sons Inc., Hoboken, New Jersey, USA, 2004, pp. 598-645.
4. Cardoso, G., and Gomide, F. “Newspaper demand prediction and replacement model based on fuzzy clustering and rules,” *Information Sciences* (177:21), 2007, pp. 4799-4809.
5. Carvalho, D. R., and Freitas, A. A. “A hybrid decision tree/genetic algorithm method for data mining,” *Information Sciences* (163:1-3), 2004, pp. 13-35.
6. Chopra, S., and Meindl, P. *Supply Chain Management: Strategy, Planning, and Operation* (Second Edition), Pearson Education International, USA, 2003.

7. Croxton, K. L., Lambert, D. M., Garcia-Dastugue, S. J., and Rogers, D. S. "The Demand Management Process," *The International Journal of Logistics Management* (13:2), 2002, pp. 51-66.
8. Geurts, M. D., and Whitlark, D. B. Six Ways to Make Sales Forecasts More Accurate," *The Journal of Business Forecasting Methods & Systems* (18:4), 1999, pp. 21-23.
9. Guyon, I., and Elisseeff, A. "Six Ways to Make Sales Forecasts More Accurate," *Journal of Machine Learning Research* (3), 2003, pp. 1157-1182.
10. Han, J., and Kamber, M. *Data Mining: Concepts and Techniques* (Second Edition), Morgan Kaufmann Publishers, USA, 2006.
11. Hanczar, B., Courtine, M., Benis, A., Hennegar, C., Clément, K., and Zucker, J. D. "Improving Classification of Microarray Data using Prototype-based Feature Selection," *ACM SIGKDD Explorations Newsletter* (5:2), 2003, pp. 23-30.
12. Kahn, K. B. "Benchmarking Sales Forecasting Performance Measures" *The Journal of Business Forecasting Methods & Systems* (17:4), Winter 1998/1999, pp. 19-23.
13. Keller, G.*Statistics for Management and Economics* (Seventh Edition), Thomson Brooks/ Cole, USA, 2005.
14. Kotsiantis, S.B. "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*(31:1), 2007, pp. 249-268.
15. Lambert, D.M. "Supply Chain Management–Processes, Partnership," *Performance* (Second edition), The Supply Chain Management Institute, pp. 59-76, 2006.
16. Lapide, L., "New Developments in Business Forecasting: Forecasting Is about Understanding Variations," *The Journal of Business Forecasting Methods & Systems* (17:4), 1998, pp. 29-30.
17. Lee, Y., and Lee, C. K. "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics* (19:9), 2003, pp. 1132-1139.
18. Li, R., and Wang, Z. "Mining classification rules using rough sets and neural networks," *European Journal of Operational Research* (157:2), 2004, pp. 439-448.
19. Li, X.B. "A scalable decision tree system and its application in pattern recognition and intrusion detection," *Decision Support Systems* (41:1), 2005, pp. 112-130.
20. Liu, H., and Yu, L. "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, (17:4), 2005, pp. 491-502.
21. Lo, V. S. Y. "The True Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing," *ACM SIGKDD Explorations Newsletter* (4:2), 2002, pp. 78-86.
22. Mohanty, B.K., and Bhasker, B. "Product classification in the Internet business—a fuzzy approach," *Decision Support Systems* (38:4), 2005, pp. 611-619.

23. Sheikh, K. *Manufacturing Resource Planning (MRP II) with introduction to ERP, SCM,* and CRM (International Edition), McGraw-Hill, Singapore, 2002.
24. Swiniarski, R. W., and Skowron, A. "Rough set methods in feature selection and recognition," *Pattern Recognition Letters* (24:6), 2003, pp. 833-849.
25. Taylor, B. W. III*Introduction to Management Science* (Eighth Edition), Pearson Education, Inc., Upper Saddle River, New Jersey, USA, 2004, pp. 691-722.
26. Wakaki, T., Itakura, H., and Tamura, M. "Rough Set-Aided Feature Selection for Automatic Web-Page Classification," *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004)*, 2004, pp. 70-76.
27. Yuan, H., Tseng, S.S., Gangshan, W., and Fuyan, Z. "A Two-phase Feature Selection Method using both Filter and Wrapper," *Proceeding of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC '99 Conference),* Tokyo, Japan, 1999, pp. 132-136.
28. Zinn, W., and Liu, P. C. "Consumer response to retail stockouts," *Journal of Business Logistics* (22:1), 2001, pp. 49-71.