國立交通大學

資訊管理研究所

碩士論文

應用Trie結構動態時間扭曲法於台灣股市分鐘資料行為分析

Using Trie-structure Dynamic Time Warping on the

Taiwan Stock Intra-day Index Time Series Analysis

研 究 生：葉 志 彥

指導教授：陳 安 斌 博士

中華民國九十三年六月

摘要

從 2003 年的諾貝爾經濟學獎的論文可知，金融市場的時間序列存在著某些行為與規則。而這些行為與規則，除了使用傳統的統計數學工具被發現外，也可以應用圖形識別的方法學搜尋。

圖形識別的問題，包含了影像處理、語音辨識、時間序列的趨勢分析等領域。影像處理與語音辨識應用在商業、娛樂相當的廣泛。而隨著資訊科技的進步，圖形識別的應用與技術方法，也更加成熟。圖形識別中的距離度量（Distance measures），是影像處理、語音辨識中不可或缺的一個函式。距離度量的定義不同，往往會影響了圖形識別技術效果的好壞。

距離度量是圖形識別的一種基本量度工具，應用在分群（Clustering）與相似度搜尋(Similarity search)上。動態時間扭曲法（Dynamic time warping）便是一種距離度量的方法學，由於其在語音辨識的優異標現，1994 年 Berndt 與 Clifford 將這個方法學帶入了資料挖掘的領域中。雖然這個距離度量相較於傳統的距離度量而言，在資料分析上，有更好的表現，然而其距離計算成本卻因為其演算法的時間雜度而變大，這也限制了它即時處理的表現。

本研究提出一個改良過後的動態時間扭曲法—trie 結構動態時間扭曲演算法，並利用階層分群法，並將其應用於股市每日分鐘資料的行為預測上，並且和其他兩種距離度量比較其預測正確率的表現與處理的時間長短。

經實證的結果發現改良過後的動態時間扭曲演算法在預測準確率上較歐幾里德距離來的好，而和原本的動態時間扭曲法相近，而處理時間也較原來的動態時間扭曲法短。因此，trie 結構的動態時間扭曲法可以應用於及時的金融預測上。

關鍵字：圖形識別、距離度量、時間序列、台灣股價加權指數、分群法、動態時間扭曲法

Abstract

According to the 2003 Nobel Economic Prize, some behaviors and rules exist in time series at financial market. However, these behavior and rules can be found not only by traditional statistic or mathematical tools but by pattern recognition methodologies.

The problem of pattern recognition includes image processing, speech recognition, time series data analysis and so on. Image processing and speech recognition have been widely applied to business and entertainment According to the progress of information technology, the methodologies of pattern recognition are enhanced with high-speed computation and high-volume storage devices.

The distance measures in image and speech processing of pattern recognition are necessary and important tools. Distance measures can be applied to clustering and similarity search. Dynamic time warping is one of distance measures, which is used and well-performed at speech recognition. In 1994, DTW was introduced in data mining domain by Berndt and Clifford. Although it performs well than traditional distance measure, such as Euclidean distance, cost of computation can be large because of the algorithm of DTW. This cost can limit the performance while using DTW on real-time analysis.

An improved DTW, trie-structure DTW is proposed in this research. By using hierarchical clustering, trie-structure DTW will be applied to analysis of time series of minute-data in TAIEX(Taiwan Stock Exchange Corporation Capitalization Weighted Stock Index). The classic DTW and Euclidean distance will be compared with trie-structure DTW in this research.

After experiments, using trie-structure DTW would get better performance than E Euclidean distance measure. Furthermore, the time cost of trie-structure DTW is less than the classic DTW and it's possible to use the improved DTW on real-time financial prediction.

Keyword: Pattern Recognition, Distance Measures, TAIEX, Cluster Analysis, Dynamic Time Warping Method

目錄

表目錄

圖目錄

一、緒論

1.1 研究動機

人類對於周遭的事物，存在著認知的本能[1]。人類和動物都有知識，但是人類的知識較動物而言，更為精密。藉由認知的功能，人類將知識存入記憶中，並且隨時去運用這些知識；也就是說，人類能藉著分析舊的經驗得到新的知識，再將新的知識儲存起來。隨著資訊科技的發展，讓機械學習的想法也隨之而生，於是這些認知的動作開始被模式化，而目前圖形識別的問題也因其所含的知識經驗而被歸類至機械學習之一。

圖形識別的問題，包含了影像處理、語音辨識、時間序列的趨勢預測等領域。影像處理與語音辨識應用在商業、娛樂相當的廣泛。而隨著資訊科技的進步，圖形識別的應用與技術方法，也更加成熟。

圖形識別中的距離度量（distance measures），是影像處理、語音辨識中不可或缺的一個函式。距離度量的定義不同，往往會影響了圖形識別技術效果的好壞。動態時間扭曲法（dynamic time warping）便是距離度量的一種。動態時間扭曲法與隱藏式馬可夫模型（Hidden Markov Model）同為語音辨識中最常被應用的方法學，兩種方法應用於不同的語音類型，均有不錯的效果。隱藏式馬可夫模型的應用，除了在語音辨識外，也應用於股市資料的分析[3]。同樣地，動態時間扭曲法也被應用於股市圖形的比對上[4]。與馬可夫模型相較，動態時間扭曲的運算較為簡單，但是因為計算的時間複雜度為$O(n^2)$，其運算時間成本仍然是相當高的[1]。

因此，本研究提出 Trie 結構動態時間扭曲法，並將此改良過後的動態時間扭曲法應用於股市資料分析，以瞭解改良後方法的效果。

1.2 研究目的

本研究的目的，在於檢驗不同的距離度量下對於分群的影響。利用測試資料檢驗分群的好壞，藉以瞭解距離度量對於分群的影響，並以實驗證明本研究提出改良的距離度量方法學應用在股市資料分析方面優於其他兩種傳統方法學。

此外，對於股市資料分析，本研究運用資料挖掘(Data mining)中資料前處理與資料轉換的步驟，將原始的台灣股市大盤指數分鐘資料，轉為每一個序列代表一天的資料。藉由這處理，並搭配改良方法學的使用，本研究期望使用 trie 結構動態時間扭曲法的距離度量提升預測正確率，並期望能減少動態時間扭曲法在距離度量運算上所耗的時間。

1.3 問題的定義與論文架構

本研究的問題定義如圖 1-1 所示。

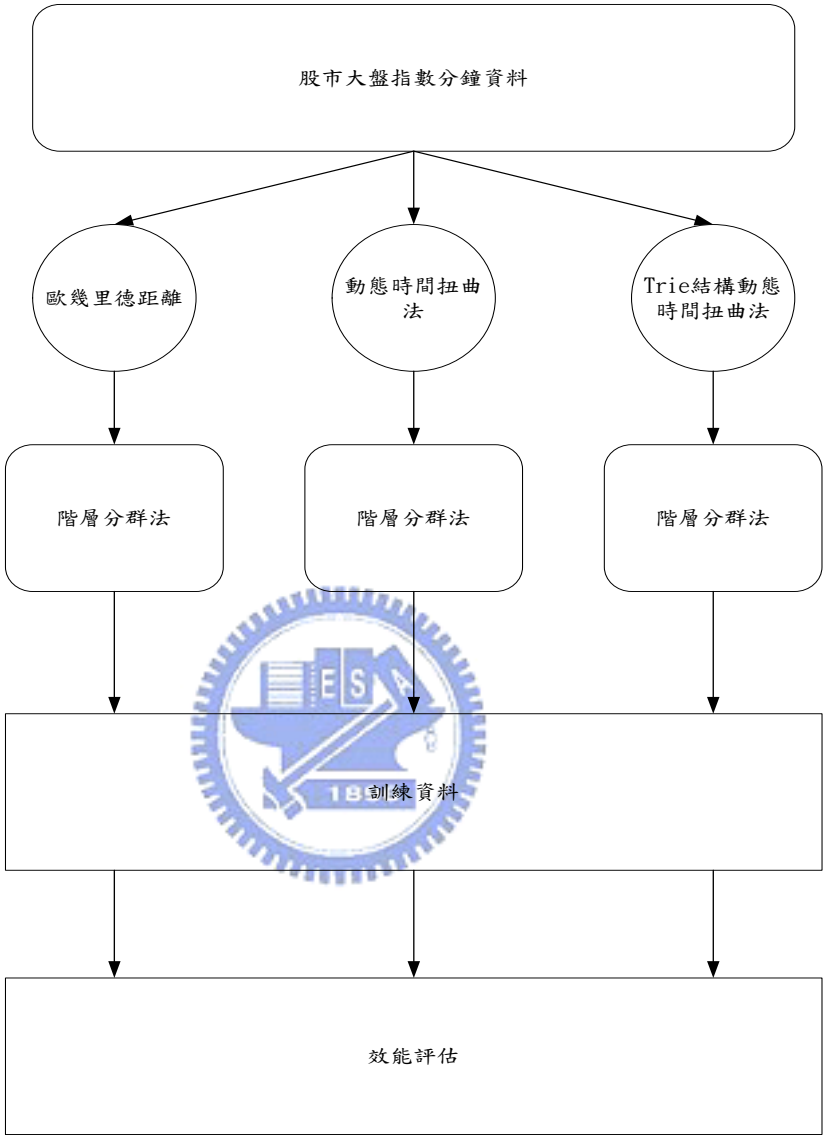

圖 1-1. 問題的定義

本研究的問題定義在評估圖 1-1 中三個距離度量對於大盤指數資料分群與預測效果的好壞。效能評估包括實驗的時間以及訓練資料預測的正確率。

第二章是關於圖形識別、時間序列、距離度量、以及距離度量在分群上應用的相關文獻。第三章介紹本研究所需要用到的方法學與實驗架構，包括 3.1 節的歐幾里德距離，3.2 節的動態時間扭曲法，3.3 節的 trie 結構動態時間扭曲法，3.4 節的階層分群法，以及 3.5 節實驗的設計與流程，資料前處理的過程。第四章為實驗結果的討論與檢驗。第五章為全文的總結與未來可研究方向的討論。

二、文獻回顧

本研究主要提出一個改良的距離度量（Distance measure），trie 結構的動態時間扭曲法（trie-structure Dynamic Time Warping）應用在股市每日分鐘資料的分析，並輔以分群(Cluster)的方式來觀察此距離量度是否比原來的動態時間扭曲法與傳統上所使用的歐幾里德距離所做出的分群更有效果。

距離度量為圖形識別領域的範圍之一，因此，2.1 節先介紹圖形識別的問題與定義。2.2 節介紹時間序列的問題的類型，並且開始回顧距離度量相關的文獻。2.3 節則是介紹距離度量在金融方面的應用

2.1 圖形識別的問題與定義

圖形識別(pattern recognition)是一門綜合性的學門，圖形識別的應用包括:語音識別(speech recognition)、資料挖掘(data mining)、字體識別(character recognition)等。圖形識別的應用廣泛而且不同，但圖形識別的基本問題:測量(measurement)、特徵抽取(feature extraction)、一般化(generalization)以及辨識度的訓練(training for discrimination)卻是這些應用所共有的[5]。圖 2-1 便是一般圖形識別的辨識模型。

圖 2-1 中，原始資料(Raw data)是一群未經處理的資料。特徵萃取的步驟則是將原始資料轉換為特徵向量，如語音辨識裡將聲波資料先轉成頻率資料，時間序列的資料在標準化時也需要做標準化(Normalization)。當把特徵向量萃取出後，將特徵向量當作輸入，經由圖形識別系統去找尋該輸入的特徵向量是否符合一些預先定義圖形的特徵向量。當該特徵向量符合辨識系統中的某個類別，該特徵向量便被系統歸類而擁有該類擁有的特性。

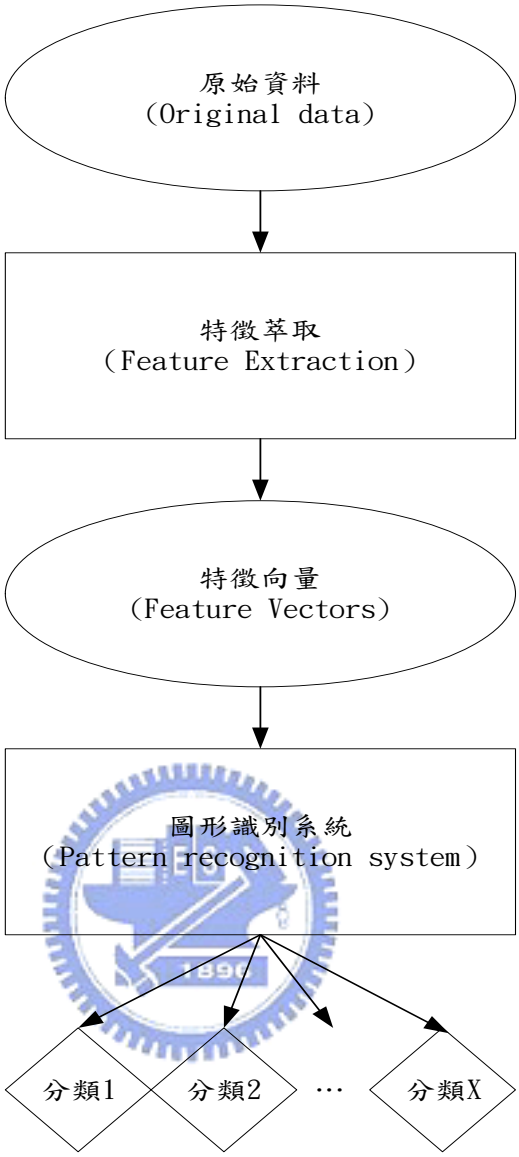

圖 2-1. 對於一般圖形識別的辨識模型

圖形辨識系統需要先經過訓練(training)才能用來辨識圖形，就像人類需要經驗學習才能知道哪些東西可以吃，哪些東西不能吃。訓練分為兩種，監督式學習(Supervised Learning)與非監督式學習(Unsupervised Learning)。監督式學習與非監督式學習的差異點在於對於原始資料的類別種類、類別個數的已知與未知。監督是學習的圖形辨識系統可以藉由每一次圖形的識別，修正系統內的參數；而非監督式學習則是由特徵向量間的相似度特性作為分群的依據，不能修正系統的架構與參數。

人類本身就是一個精密複雜的圖形識別系統。在人類每日的生活中，圖形識別的行為就會不斷的發生，舉凡人類眼睛所看到的物件，耳朵所聽到的聲音，在頭腦裡所做的推理與思考等。這些物件、聲音、推理就是所謂的圖形(patterns)，而這些圖形隨著型態的不同可以分成兩類—具體的圖形(concrete patterns)與抽象的圖形(abstract

patterns)[5]。具體的圖形泛指人類可以看見的圖形，如影像，可以聽的到的圖形，如聲音；抽象的圖形指人類心理所想的、推理的，如問題的解法、思想的辨證、理論的思考。

根據圖形識別的問題，圖形識別的定義在於使用方法學去辨識圖形，並且將辨識的圖形分成多種類別(Classes)[5]。這些被標記類別的圖形，可以被使用在決策的輔助上。舉例來說:若人類對於圖形(patterns)將其分為可以吃的東西一類，不能吃的東西一類，則人類先做辨識，再做分類。往後人類接觸到新的圖形(pattern)，會利用眼睛識別，在利用頭腦中的知識庫做圖形的分類，當圖形分類完了，最後做出決策—這個東西到底能不能吃的決策。

當然，圖形識別被應用在各種研究領域(Research domain)。各領域的知識(domain knowledge)是不相同的，如語音辨識[7]的輸入是聲音的頻率，而影像辨識中的指紋辨識[8]則是抓取指紋圖像中的特徵點，時間序列則是序列圖形(sequential pattern)的辨識。本研究之中，被使用的圖形識別方法學是屬於時間序列問題方面的方法學。

2.2 時間序列的研究

時間序列為一連串的資料，這一連串的資料是伴隨著時間的行進不斷產生的。資料的型態可以為數值或是符號。

時間序列問題的研究已經被研究超過十年以上。從資料庫應用的角度，時間序列的問題主要被分成四個部份[9]:趨勢分析(trend analysis)[10][11]、相似度搜尋(similarity search)[12][13][14][15][16]、序列圖形挖掘(sequential pattern mining)[17][18][19]以及週期性分析(periodicity analysis)[20]。

趨勢分析的基本定義為假設時間序列上時間點t的點估計T，存在某個函式G，使得$T = G(t)$。典型的趨勢分析代表為移動平均法，n為移動平均法的參數，令時間序列$T = \{s_1, s_2, \ldots s_i\}$，若$n = 3$則移動平均的的算式為(2-1)

$$Y_i = \frac{s_i + s_{i+1} + s_{i+2}}{3} \qquad (2\text{-}1)$$

則新的T為$\{Y_1, Y_2, Y_{3,\ldots} Y_{i-2}\}$。圖 2-2 表示為移動平均與原來序列。

$$\{\frac{s_1 + s_2 + s_3}{3}, \frac{s_2 + s_3 + s_4}{3}, \frac{s_3 + s_4 + s_5}{3}, \frac{s_4 + s_5 + s_6}{3}, \ldots, \frac{s_{i-2} + s_{i-1} + s_i}{3}\} \qquad (2\text{-}2)$$

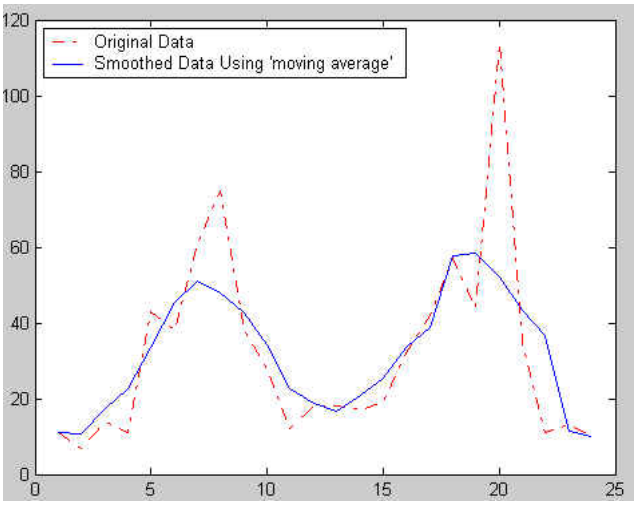

圖 2-2. 原始資料與其移動平均線

相似度搜尋則是利用距離度量的方式，依據給定的圖形，去找出輸入資料中最相近的圖形。

序列圖形挖掘則在資料挖掘中是屬於符號序列的資料挖掘。找出在序列中最常發生的圖形是序列圖形挖掘的主要目的。

週期性分析則是找出時間序列中週期性行為的存在。如經濟的景氣循環、地震的行為、自然界的現象。

另一方面，從研究問題的角度，時間序列的問題被分為四個主題[21]: 索引(indexing)、分群(Clustering)、分類(Classification)、分段(segmentation)。索引的定義跟相似度搜尋的定義相同，均為給定一個時間序列 T，根據距離的度量（distance measure），搜尋與給定時間序列 T 最接近的時間序列。分群為藉由使用距離度量(distance measure)，讓一群序列資料分成一個以上的群聚(groups)。分類定義為給定一個未知分類的時間序列 T，藉由預先被定義好一個以上的類別(Class)，給予此未知類別的序列 T 一種類別。分段（Segamentation）定義為：給定一個含 n 個資料點的時間序列 T，建構一個模型（model）M，使得模型 M 可產生 K 個區段(segament)，K<n，並且使這 K 個區段的圖形趨近於原來的序列 T。

本研究將重點放在相似度搜尋上面，因此，在 2.3 中，相似度搜尋相關的文獻將被列舉與說明。此外，在研究過程中所遇到時間序列的基本問題，乃是時間序列資料選取的頻率[22]。例如，股市的資料要每天選取，每月選取，或是每一分鐘選取作為一個時間序列的資料。而取樣頻率的好壞，對於分析與決策會有重大的影響；取樣過多，相當於沒有過濾資訊，取樣過低相當於濾掉大部分的資訊。

2.3 距離度量

兩個時間序列的相似度(similarity)，是由距離度量（distance measure）來計算。因此，良好的距離度量在對於相似度搜尋的結果正確與否，是一個重要的影響因素。

距離度量實際上是一個數值函數（Numerical function）[23]。它的輸入是兩個向量，輸出是距離。這種函數一般有三種特性。分別為：

$$d(x,y) \geq 0; d(x,y) = 0 \quad if \quad x = y \tag{2-3}$$

$$d(x,y) = d(y,x) \tag{2-4}$$

$$d(\mathrm{x},\mathrm{y}) + \mathrm{d}(\mathrm{y},\mathrm{z}) \geq d(x,z) \tag{2-5}$$

d 為距離度量之函數，x、y、z 為三個向量。而(2-5)式為數學上為人所熟知的三角不等式，也就是兩邊和大於第三邊。

距離度量在聲音與影像處理的方法學上是一個基礎。在語音處理領域裡，它被應用在語音辨識、聲紋認定、編碼方面[24]；在影像處理方面，它被應用在影像圖形的比較[25]。

歐幾里德距離（Euclidean distance）與其衍伸的標準歐幾里德距離(standard Euclidean distance)因為其簡單性與運算的快速的特性，而被廣泛應用。在時間序列的應用方面，歐幾里德距離在時間序列資料庫的快速搜尋[13]、規則發現[26]、比對[18]都具有相當的效果。

標準化歐幾里德距離則是歐幾里德距離的衍生，相關的距離度量還有街道距離(city block distance)、Mahalanobis 距離、Mincowski 距離等。這些距離量度被用在分群與資料挖掘上，均有不錯的成果[27]。

動態時間扭曲法（Dynamic time warping，DTW），常見於訊號處理的研究中[1]，但從 1994 年，Berndt 以及 Cliford 將其使用在資料挖掘領域中，並利用 DTW 去發現時間序列中相似的圖形[28]。

雖然動態時間扭曲法在辨識時間序列的圖形較傳統的距離度量來的好，但是計算其距離量度的演算法卻是時間複雜度$O(n^2)$，這也使得動態時間扭曲法的應用受到限制。動態時間扭曲法與歐幾里德距離應用於資料分群上，Keogh[29]的研究顯示DTW能避免歐

幾里德距離的盲點。但是，DTW的運算過慢，也成了應用上的一大缺點。

因此，對於這個問題，有從硬體平行運算(Parallel implemention)方面解決的方法[30]，也有從演算法本身去找出趨近動態扭曲距離的值[29]，語音辨識上也利用 TSM 的方式由輸入資料的處理去加快 DTW 的運算[1]。

Keogh 的研究，是從距離度量的特性去找出趨近於動態時間扭曲距離的值。由於 DTW 擁有對於時間點校正的優點，對於距離的計算也更佳精確，然而 DTW 卻不符合(2-5)式的三角定理，因此要使用三角不等式估計兩點間的 DTW 距離並不可行。

而 Keogh 利用 PAA(Piecewise Constant Approximation)變形的方式，找出 DTW 扭曲距離的上限與下限，估計真正的 DTW 距離。這個方法的好處是可以利用循序搜尋法(Sequential Search)搜尋出 DTW 的上限與下限值。

本研究所提出之 Trie 結構[31]動態時間扭曲法是從輸入資料方面著手，將兩筆資料經 trie 結構的搜尋，找出兩筆資料前端相同部分，然後將後端的分枝做 DTW 的距離度量，詳細的步驟在 3.3 節被闡述。由於 trie 結構的建立，DTW 省去前端比對的時間，增加了計算 DTW 距離的速度。

而 trie 結構動態時間扭曲，對於相似度搜尋方面亦有良好的表現[32](註 1)。圖 2-3 中，trie 結構動態時間扭曲法被應用於相似度搜尋的搜尋時間效果。

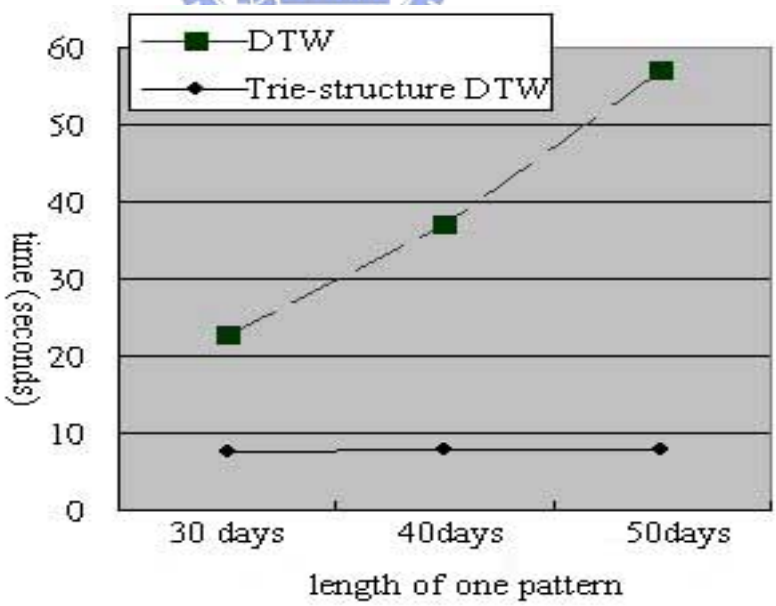

圖 2-3. Trie 結構 DTW 與傳統 DTW 應用於相似度搜尋的效能比較
(資料欄來源：[32])

1

[1] 註：參考文獻[32]為本研究的前身，已投稿AIA 2004 研討會

2.4 距離度量所應用之分群法

分群法是無監督的學習(Unsupervised Learning)方法學。經由分群的方法，將資料點分成一個一個的群聚(groups)。而分群和分類的不同是資料點的類別屬性。分類是將資料點加上類別的標記(labeled)；而分群並不對資料點做特殊的標記，分群只在顯示某些資料點是在一個群聚之內。

分群的基本條件就是資料(data)和距離度量（或相似度）[23]。而根據 Cormack(1971)對於分群法的分類，分群可分為五大類的方法[33]。

第一類為階層法(Hierarchical techniques)，這類方法的目的將每個資料點經由距離度量矩陣的建立，形成樹狀結構(tree，如圖 2-4)。當樹狀結構建立後，再從樹的某一階層去分出各群[34]。階層法在 3.4 節會有詳細的介紹。

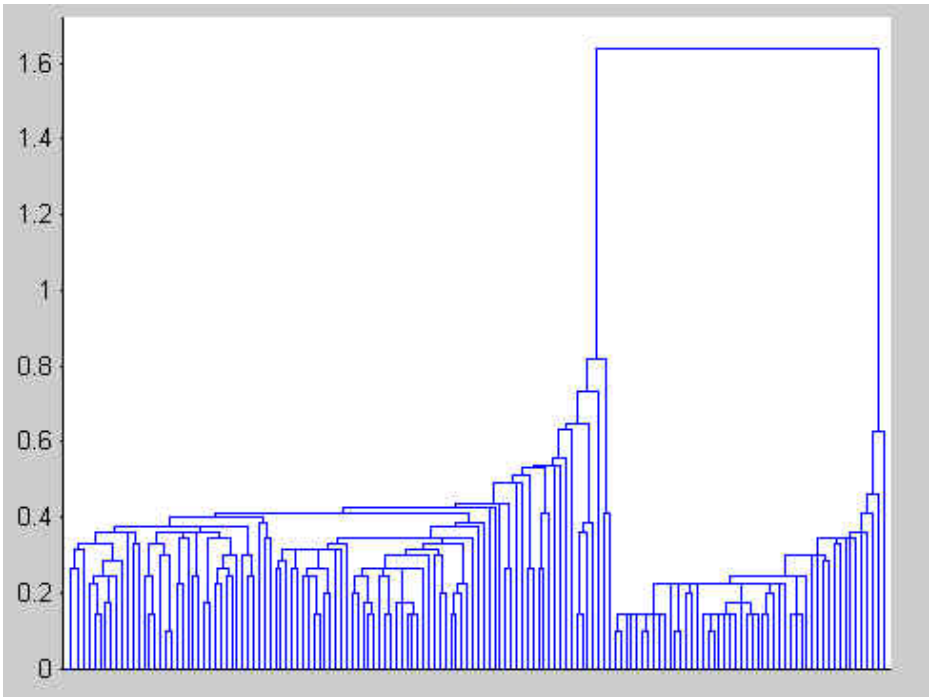

圖 2-4. 階層分群法資料的樹狀階層結構

第二類為最佳化法(Optimization techniques)，此類的分群法避免了初始的分群不佳的問題，但是群的個數需要先被確定才能進行分群，如 *k*-means 即是這一類的分群法。Tapas K.,et al.便針對傳統的 Lloyd 演算法，使用更簡單與有效的方式去執行演算[35]。圖 2-5 與圖 2-6 為 *k*-means 分群法的種子(seeds)與群聚的分佈示意圖。

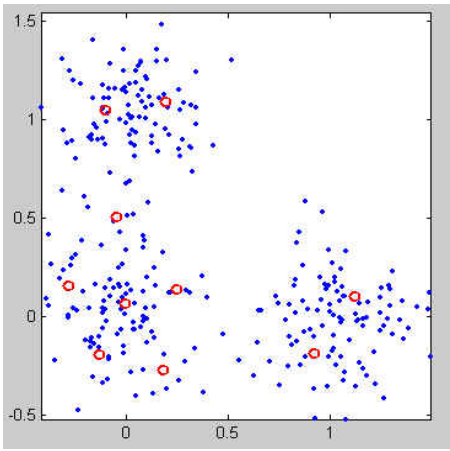

圖 2-5. *k*-means 分群法的種子分佈圖

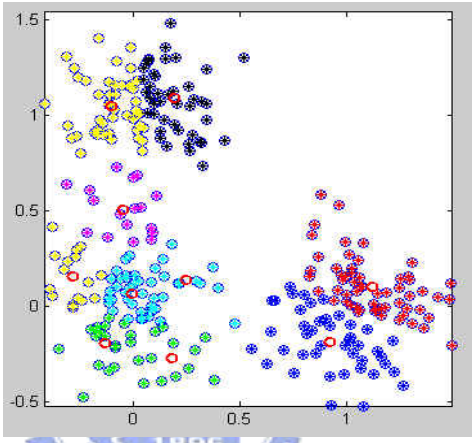

圖 2-6. *k*-means 分群法的群分佈圖

第三類為密度搜尋法，此類方法為搜尋資料點在某個區域的密度，利用區域的密度去分出群聚[36]。

第四類為叢聚法(Clump techniques)，此類方法允許群與群間相互重疊，群與群之間並非互斥的[37]。第五類為無法歸為前四類的分群法[38][39]。

2.5 圖形辨識在金融分析上的應用

圖形識別應用在金融分析上，可分為人工智慧以及其他圖形識別的方法。人工智慧的方法，有類神經網路、基因演算法等。圖形識別的方式則是以訊號處理與資料挖掘的方法為主要的處理方法學。

人工智慧有問題定義的界定問題與缺乏常識性(common sense)的特性[1]。可以被電腦接受的問題定義，是有一個很明白確實的方法來確定答案對不對，相反的，人類日常生活上碰到的問題常常無法界定的很好。如請一個人來規劃速食店店面的設計圖，怎麼判斷這個設計好不好，就是一個無法界定得很好的問題。有些設計一看就是錯的，如沒有設計廁所。因此，問題定義後的解答是要能被評估的。

缺乏常識性是人工智慧方法的另一個特性。當人工智慧的方法學應用在某類的資料分析時，如果想分析其他類資料的關連性，就必須要匯入其他資料，並定義兩類資料的關連性。雖然如此，人工智慧的方法學應用在被定義好的問題上，仍有極佳的表現，典型的例子便是 MYCIN 專家系統。

類神經網路為模仿人類神經元的資訊傳輸而發展出的方法學，雖然它的內部運作是黑箱作業，但它在解決特定問題的表現上，卻有相當優異的表現。

吳勝修（民 92）以類神經網路的模型改善傳統時間不變性資產組合的策略[40]。曾士育(民 92)則是以自組織映射圖神經網路去施行金融投資決策的資料挖掘[41]。王湘蕙(民 91)則是以分群法分析電子類股的風險類型[42]。

而這些研究的目的都在於將雜亂無章的資料整理出一個可以做為決策的法則。而資料挖掘技術的開發，更把金融分析從數學統計的領域拉到了資訊科技分析的領域[21]。而資料挖掘可以發現資料的關連性，而這一點正補足了人工智慧的缺點。

同時，訊號處理領域的方法學亦應用在金融分析上。蔡文智(民 91)以隱藏馬可夫模型對於期貨做當日沖銷交易的決策分析[3]。許育嘉(民 91)，發展出運用小波轉換對於財務序列的預測[44]。這也代表了越來越多用來解決其他非金融問題的方法學被應用到金融分析領域。

三、研究方法

3.1 歐幾里德距離

歐幾里德距離為最常見的的距離度量。其定義為(3-1)所示。

$$D(P,Q) = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2} \tag{3-1}$$

P與Q為兩組長度均為k的向量，其中$P = (p_1, p_2, \ldots, p_k)$，而$Q = (q_1, q_2, \ldots, q_k)$。歐幾里德距離所衍生的標準化歐幾里德距離(standardized Euclidean disatance)其定義為(3-2)。

$$SD(P,Q) = \sqrt{\sum_{i=1}^{k}\frac{(p_i - q_i)^2}{\sigma_i^2}} \tag{3-2}$$

此處σ_i^2的定義為p_i與q_i的變異數。從另一角度來說，若存在一向量集合$V = \{V_s | s = 1,2,\ldots n\}$，且每一向量的長度均為$k$，每一向量中的元素表示為$V_{si}$，$i = 1,2,\ldots k$，則$\sigma_i^2$為$V_{si}$的變異數，$s = 1,2,\ldots,n$。算式如(3-3)所示

$$\sigma_i^2 = \frac{\sum_{s=1}^{n}(V_{si} - \overline{V_{si}})^2}{n-1}, i = 1,2,\ldots,k \tag{3-3}$$

$\overline{V_{si}}$為V_{si}的算數平均數。因此，對於向量集合V中的任兩向量V_r與V_s，其標準化歐幾里德距離為(3-4)

$$SD(V_r, V_s) = \sqrt{\sum_{i=1}^{k}\frac{(V_{ri} - V_{si})^2}{\sigma_i^2}} \tag{3-4}$$

3.2 動態時間扭曲法

動態時間扭曲法的目的是在找尋兩向量間最小累積距離的扭曲路徑，其優點在於能將不同長度的向量做距離量度。其累積距離演算法如表 3-1 所示。

表 3-1. 動態時間扭曲法的累積距離矩陣的演算法

```
DTW distance Algorithm C(i,j)=DTW(R , T)
      //count the Euclidean distance

for i=1 to length(R)
    for j=1 to t
        A(i,j) = distance(R(i),T(j));
    endfor
endfor

for i=1 to length(R)
    for j=1 to r
        m = min(C(i-1,j-1),C(i,j-1),C(i-1,j));
C(i,j) = A(i,j)+ m
    endfor
endfor
```

計算出累積距離矩陣後，便可以依據累積距離矩陣化畫出扭曲路徑。以一個實例來說明動態時間扭曲法。兩個向量為 *RP* = (2,4, 4, 6,2)，*TP* = (1, 2, 2, 4, 1)，則其各點的距離矩陣為表 3-2。

表 3-2. DTW 距離矩陣

2	1	0	0	2	1
6	5	4	4	2	5
4	3	2	2	0	3
4	3	2	2	0	3
2	1	0	0	2	1
RP / *TP*	1	2	2	4	1

經由表 3-1 第二個 For-Loop 的演算，得到累積距離矩陣如表 3-3。

表 3-3. 累積距離矩陣

2	13	9	9	5	4
6	12	9	9	3	6
4	7	5	5	1	4
4	4	3	3	1	4
2	1	1	1	3	4
RP / *TP*	1	2	2	4	1

由累積距離矩陣可畫出扭曲路徑。扭曲路徑主要受到以下的限制。

1.連續性(Continuity):

假設扭曲路徑中某點的座標為(i, j)，則前一點必須為$(i-1, j)$、$(i, j-1)$或$(i-1, j-1)$也就是說扭曲路徑中兩相鄰的點在座標上也必須是相鄰的。

2.無變化性(Monotonicity):

若扭曲路徑中連續兩點的座標先後為(i, j)、(i', j')，則 $i'-i >= 0$，$j-j' >= 0$，這說明了扭曲路徑上的各點是依照時間先後而排序的。

3.界線條件(Boundary Conditions)

若兩相比較的序列長度分別為 m 與 n，則扭曲路徑的起點座標必為(1,1)而終點座標必為(m,n)。

使用這 3 個限制，表 3-3 的累積距離矩陣可被畫出扭曲路徑，如表 3-4。

表 3-4. 扭曲路徑示意圖

2	13	9	9	5	■
6	12	9	9	■	6
4	7	5	5	■	4
4	4	3	3	■	4
2	■	■	■	3	4
RP / *TP*	1	2	2	4	1

扭曲路徑座標為(1,,1), (2, 1), (3, 1), (4, 2), (4, 3), (4, 4), (5, 5)。而扭曲距離 *WD* 的計算方式為式(3-5)。

$$WD = \frac{\sqrt{CD}}{K} \tag{3-5}$$

CD 為累積矩陣中最右上座標所表示的累積距離，而 *K* 為扭曲路徑中座標的個數。上例中的扭曲距離 WD 便是 $\frac{\sqrt{4}}{7} = \frac{2}{7}$ 。而累積距離的運算也說明了動態時間扭曲法對於不同時間長度的兩個序列亦可計算其累積距離與扭曲距離，這個特性使得動態時間扭曲法比歐幾里德距離來的有彈性，因為歐幾里德距離的運算必須要兩序列長度相同。

3.3 Trie 資料結構

Trie的讀音為try(讀法跟pie的方式相同)，是屬於搜尋樹的一種(Binary Search Tree)。對於資料挖掘的資料來源—資料庫的資料處理、儲存、取用，已經成為一個相當重要的議題。而這個問題的另一面，就是儲存空間與存取時間之間的交換關係(Trade-off)。直接定址(Direct addressing)的方式，具有著短暫存取時間，但卻需要大量的儲存空間；相對地，另外一種方式是二元樹搜尋方式，若搜尋的個數為n，則需要$O(\log_2 n)$的搜尋時間。此外，常見的線性搜尋(Linear search)則需$O(n)$的搜尋時間。但這些方法，都需要面對對於存取速度與儲存空間的取捨。

Trie 的結構儲存與搜尋速度處在這個取捨的中間地帶。建構一個 trie 結構，應用在編譯器的字彙分析器獲釋自然語言的詞構學學目錄等均有不錯的效果。而這個結構的原理應用在時間序列上，可建構出一個時間序列的 trie 結構。而利用這個 trie 結構，時間序列的搜尋與比較可以較為簡單的被執行。

Trie 的例子如圖 3-1 所示。

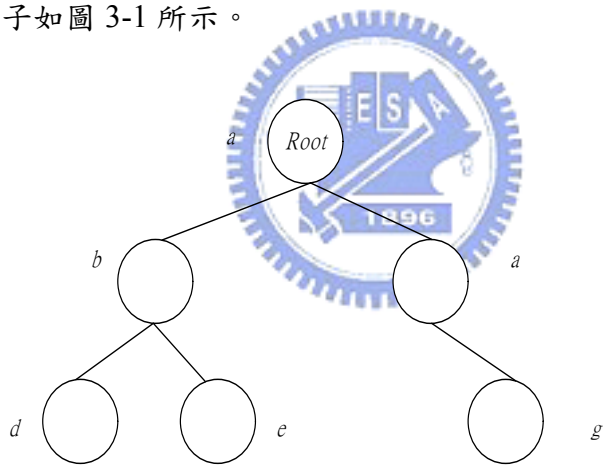

圖 3-1. Trie 結構示意圖

圖 3-1 中，Trie 的結構包含了 3 個序列，分別為(*a, b, d*)、(*a, b, e*)、(*a, a, g*)。

利用這個結構，可以將時間序列的變化模式直接比對，並且可以將之建入資料庫之中，因為 trie 結構已經將時間序列做過初步的分類。而利用 trie 的資料擷取速度與資料儲存空間的特性，可以將前端相似的序列找出，並直接比對後端部分。如圖 3-1 中，由於 trie 結構內含(*a, b, d*)與(*a, b, e*)，因此在搜尋與(*a, b, d*)相似的序列，只需要尋找 b 節點以下的部分。

3.4 Trie 結構動態時間扭曲法

Trie結構動態時間扭曲法的原理在於對資料的前端比對，再對於剩餘不符合的序列做動態時間扭曲比對。舉例來說，若有兩個相同長度的時間序列TS_1與TS_2，在計算它們的距離度量之前，先對TS_1與TS_2兩個序列做正規化(Normalization)，以比較兩序列前端相同的部分。

資料的正規化使用下列的函數決定，$N(t)$為時間序列經過正規化的資料點。給予一個時間序列T ($T = \{t_1,t_2,...t_n\}$)，序列的平均為μ，標準差為σ。式(3-5)中，s是決定極端值範圍的參數，h為大於s的參數，序列T中的點經過標準化後的值為$N(t)$。

$$N(t_i) = \begin{cases} 0 & t_i > \mu + s\sigma \\ \dfrac{t_i - \mu + s\sigma}{h\sigma} & \mu - s\sigma \le t_i \le \mu + s\sigma \\ 1 & t_i < \mu - s\sigma \end{cases} \tag{3-5}$$

因此，時間序列 T 經由轉換式(3-5)後，其值的範圍落在 0 到 1 之間。經過式(3-5)函式轉換的時間序列 T，必須再經由一個參數 *interval*，轉換其資料。

整個轉換流程舉個實例說明。現在一時間T_w = (5,7, 4, 6, 4, 3, 9)，經由式(3-5)轉換後為式(3-6)。若*interval*訂 0.2，則T_w為式(3-7)。

$$T_w = (0.49, 0.62, 0.56, 0.28, 0.42, 0.35, 0.77) \tag{3-6}$$

$$T_w = (0.4, 0.6, 0.4, 0.2, 0.4, 0.2, 0.6) \tag{3-7}$$

而經由正規化的過程後，兩時間序列前端比較方式以例子說明。給定兩個時間序列T_1與T_2經由上述轉換的過程後為P_1與P_2。

$$P_1 = \{0.2, 0.4, 1, 0, 0.6 , 0.8 ,0 ,0 ,0, \ldots\} \tag{3-8}$$

$$P_2 = \{0.2, 0.4, 1, 0, 0.6, 0.8, 1, 1, 0, \ldots \} \tag{3-9}$$

P_1與P_2的前六個資料點都是相同的，因此DTW的距離運算需要由第 7 個資料點後的資料開始運算。而前端相同資料點的個數在表 3-5 的演算法中中定義為*match_points*。當*match_points*的個數越多，後面需要運算的DTW扭曲距離所需時間就越短，所以會比傳統的DTW扭曲距離的運算來得快。以圖來說明，則為圖 3-2。在圖中，動態時間扭曲法的時間複雜度為$O(n^2)$，但是trie結構動態時間扭曲法的時間複雜度卻分為兩部分，第一

部份的時間複雜度為*O(n)*。因此，當第一部份所佔的比率越大，時間也節省的越多。

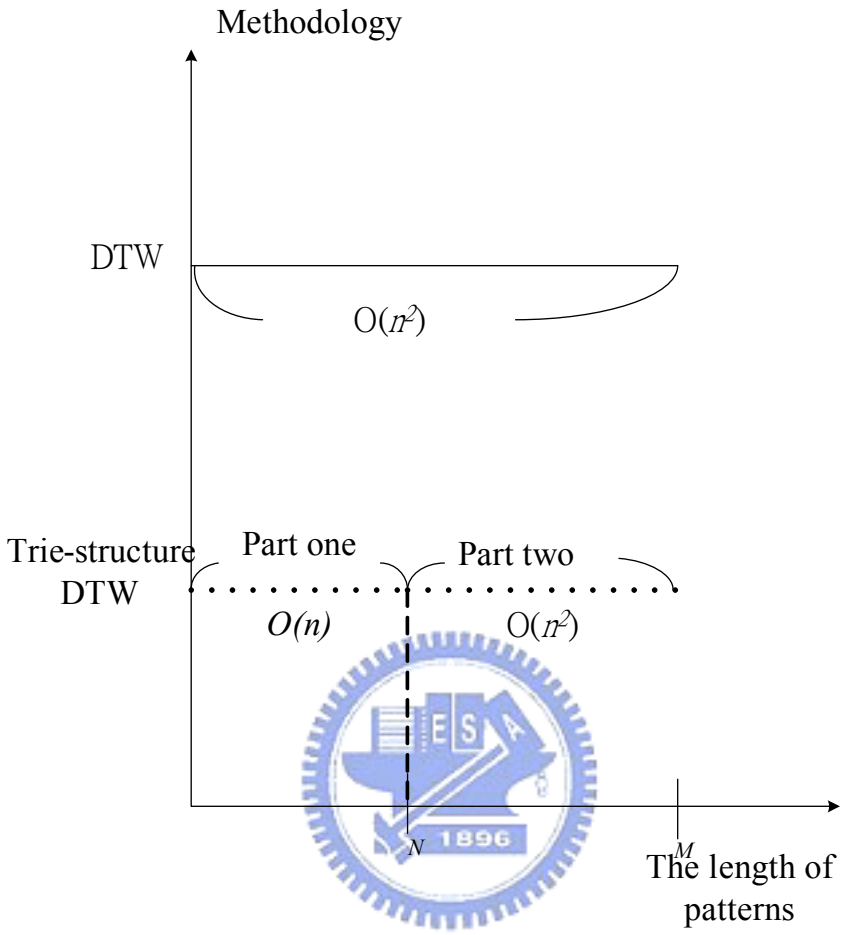

圖 3-2. DTW 與 trie 結構 DTW 的比較

表 3-5. Trie 結構動態時間扭曲法尋找相似圖形的演算法

finding match patterns algorithm

FMP(R, T, n_days, interval, threshold)

*/*trie table is a data structure to store the match patterns. R, T are long-period time-series, interval is used at normalization. threshold is about match_day to modify the trie-strucutre table*

R_i is a n_days subset of R

*T_j is a n_days subset of T */*

***for** each R_i in R*

***for** each T_i in T*

*/*both patterns stored in Data attribute for normalized R_i,T_j and Source attribute for R_i ,T_j*/*

compare normalized R_i and T_j to find the "match_day"

***if** match_day of normalized R_i and $T_i \geqq \lambda$*

*/*the condition is defined as "match_points" between two patterns*/*

do the two DTW measures between R and T excluding their first x days

save result into trie structure

*/*stored in Link attribute*/*

endif

***if** two DTW measures >= thresholds about DTW*

remove the match pattern from the Link attribute

endif

endfor

endfor

3.5 階層分群法

本研究所採用的分群法為階層分群法。而階層分群法的原理可以分為三個部分。第一部分為找尋各樣本彼此間的距離度量；第二部分為利用分群的演算將全部樣本點的關係建構成一個樹狀階層結構；第三部分將樹狀結構從某個階層切割分群。

第一部份：尋找各樣本點之間的距離度量。給定 n 個樣本點，則此 n 個樣本點會有 n*(n-1)/2 個距離度量。

第二部份：階層分群法主要分為兩類，凝聚法(Agglomerative Methods)與分裂法(Divisive Methods)。凝聚法是由下而上，從每個樣本點各為一個群開始，樣本點彼此開始凝聚成群；而分裂法是由上而下，將整個原始樣本點分割成一小群一小群。以圖來看便如圖 3-3，左半部為凝聚法，半部為分裂法。

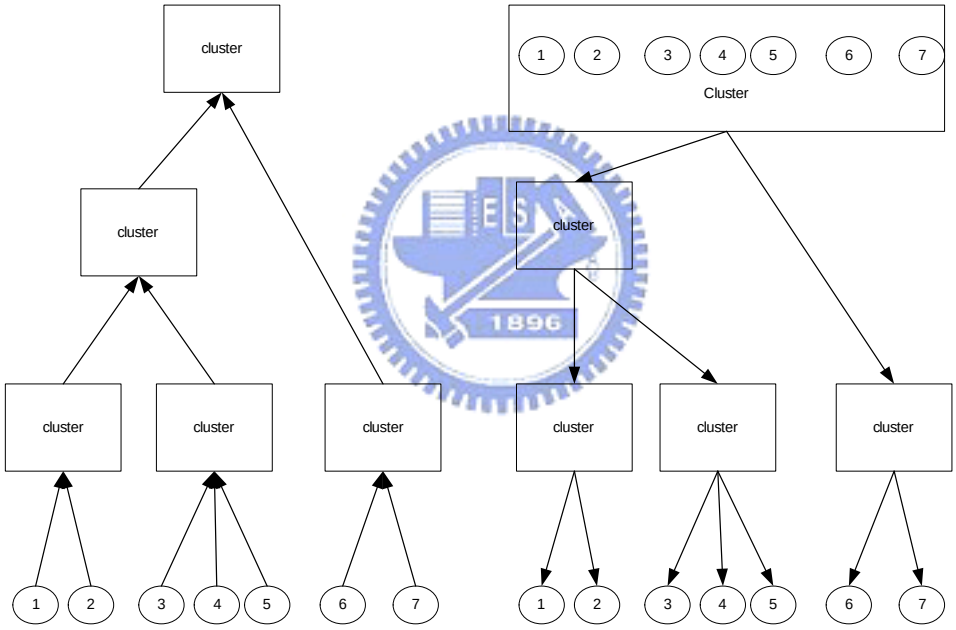

圖 3-3. 凝聚法與分裂法

群聚法常見的有最近相鄰法(the Nearest Neighbor)和群中心分析(Centroid Cluster analysis)。

最近相鄰法凝聚的原則為根據點與點間的距離量度，選取最小的距離量度為凝聚的依據。一距離度量矩陣 *D* 為(3-10)所示。

$$D_1 = \begin{bmatrix} 0 & 2 & 6 & 10 & 9 \\ 2 & 0 & 5 & 9 & 8 \\ 6 & 5 & 0 & 4 & 5 \\ 10 & 9 & 4 & 0 & 3 \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix} \quad (3\text{-}10)$$

D 為一對稱矩陣。矩陣每一列與每一行代表樣本點的索引。(3-10)中除了對角線符合(2-3)的特性外，$D(i, j)$中最小的為 $D(1,2)$距離為 2，所以表示第一個與第二個樣本點將合成一群。則點與群的距離為：

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5 \quad (3\text{-}11)$$

$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 9 \quad (3\text{-}12)$$

$$d_{(12)5} = \min\{d_{15}, d_{25}\} = d_{25} = 8 \quad (3\text{-}13)$$

則距離度量矩陣變為：

$$D_2 = \begin{bmatrix} 0 & 5 & 9 & 8 \\ 5 & 0 & 4 & 5 \\ 9 & 4 & 0 & 3 \\ 8 & 5 & 3 & 0 \end{bmatrix} \quad (3\text{-}14)$$

從(3-14)中可以發現$D_2(4, 5)$的距離為最小，所以把第四個樣本點與第五個樣本點聚成一群。則點與群或群與群的距離為：

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5 \quad (3\text{-}15)$$

$$d_{(12)(4,5)} = \min\{d_{14}, d_{15,} d_{24}, d_{25}\} = d_{25} = 8 \quad (3\text{-}16)$$

$$d_{3(45)} = \min\{d_{34}, d_{35}\} = d_{34} = 4 \quad (3\text{-}17)$$

則距離度量矩陣變為：

$$D_3 = \begin{bmatrix} 0 & 5 & 8 \\ 5 & 0 & 4 \\ 8 & 4 & 0 \end{bmatrix} \quad (3\text{-}18)$$

則階層的結構圖(Dendrogram)為圖 3-4 所示。

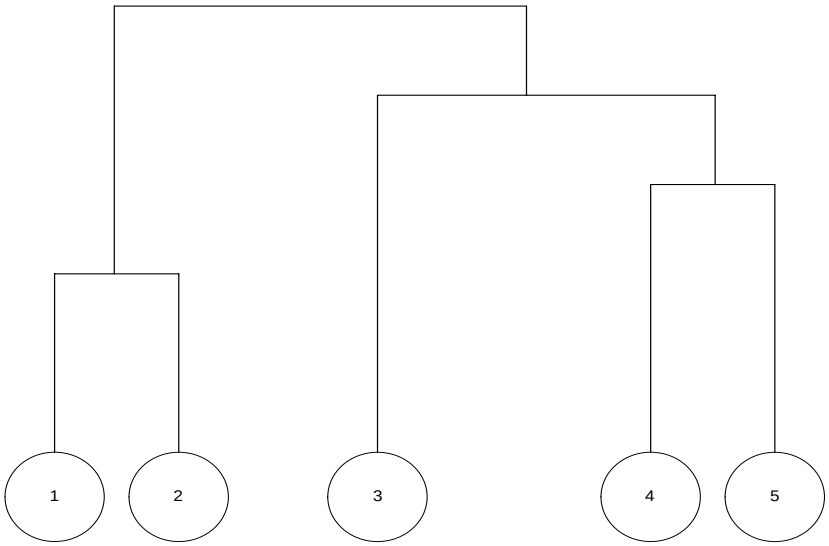

圖 3-4. 最近相鄰法分群樹狀階層結構

(資料來源：[22])

群中心分析法，則利用為一群的點與點或是為一群的群與群，計算其群中心座標。舉一實例，令平面上有 5 點分別為(1, 1), (1, 3), (5, 4), (8, 1), (9, 1)。則距離矩陣為：

$$D = \begin{bmatrix} 0 & 2 & 5 & 7 & 8 \\ 2 & 0 & 4.12 & 7.28 & 8.25 \\ 5 & 4.12 & 0 & 4.24 & 5 \\ 7 & 7.28 & 4.24 & 0 & 1 \\ 8 & 8.25 & 5 & 1 & 0 \end{bmatrix} \quad (3\text{-}19)$$

由(3-19)，第四個點(8, 1)與第五個點(9, 1)間擁有最小的距離度量值，因此第四點第五點構成一群，計算第四點和第五點座標的平均，為(8.5,1)。因此分群的問題為(1, 1)，(1, 3)，(5, 4)，(8.5, 1)四點分群。距離矩陣變為：

$$D = \begin{bmatrix} 0 & 2 & 5 & 7.5 \\ 2 & 0 & 4.12 & 7.76 \\ 5 & 4.12 & 0 & 4.61 \\ 7.5 & 7.76 & 4.61 & 0 \end{bmatrix} \quad (3\text{-}20)$$

由 (3-20)可知，第一點與第二點擁有最小距離值 2，因此第一點與第二點構成一群，計算第一點與第二點的平均為(1, 2).。因此分群的問題變成(1, 2)，(5,4)，(8.5, 1)三點分群。距離矩陣變為：

$$D = \begin{bmatrix} 0 & 4.47 & 7.57 \\ 4.47 & 0 & 4.61 \\ 7.57 & 4.61 & 0 \end{bmatrix} \tag{3-21}$$

依照以上演算的步驟，最後可以畫出階層結構圖如圖 3-5。

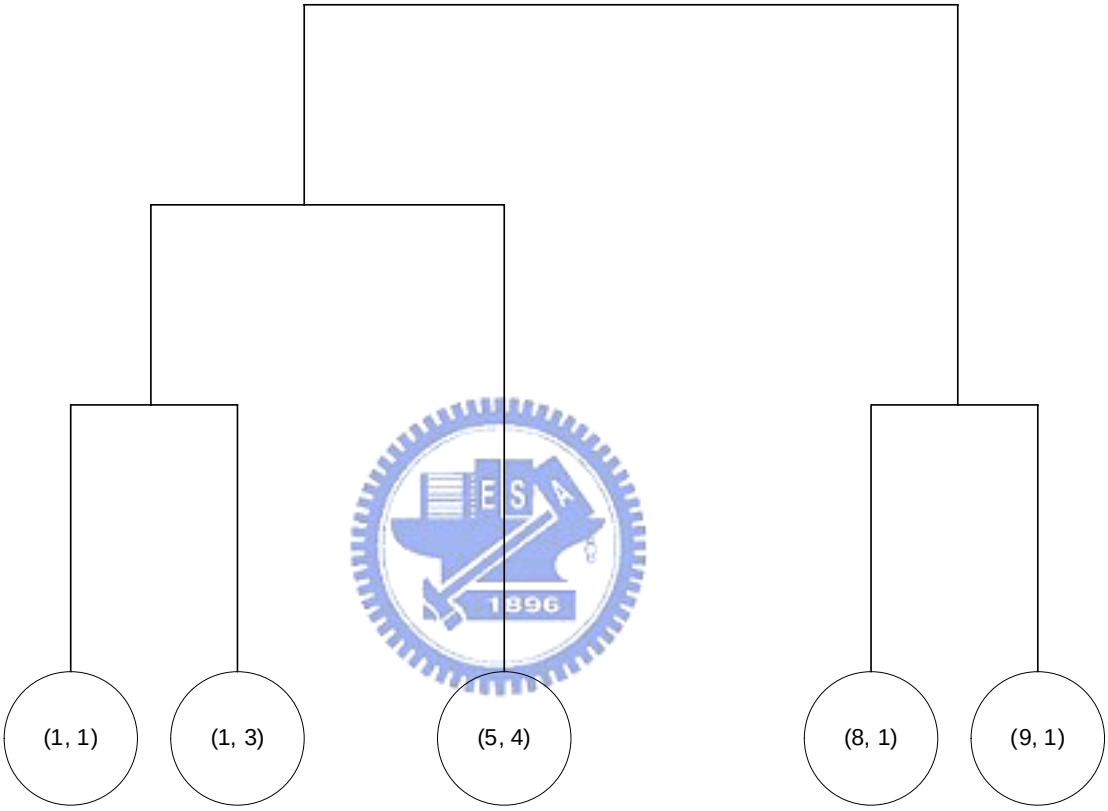

圖 3-5. 群中心分析法樹狀階層架構

在分裂法方面，假設有 7 個資料點，形成距離度量矩陣 *MX*，式(3-22)

$$MX = \begin{bmatrix} 0 & 10 & 7 & 30 & 29 & 38 & 42 \\ 10 & 0 & 7 & 23 & 25 & 34 & 36 \\ 7 & 7 & 0 & 21 & 22 & 31 & 36 \\ 30 & 23 & 21 & 0 & 7 & 10 & 13 \\ 29 & 25 & 22 & 7 & 0 & 11 & 17 \\ 38 & 34 & 31 & 10 & 11 & 0 & 9 \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{bmatrix} \tag{3-22}$$

*MX(i,j)*表示為第*i*個點到第*j*個點的距離。分裂法的第一步是將資料點先分成兩群，因此，必須從 7 個點中挑出一點*m*當成另一群，稱為分裂群(splinter group)，而原來那一群稱為主群(main group)。而點*m*會使得下式(3-23)C_m達到最大值

$$C_m = \sum_{i \neq m} MX(m,i) \tag{3-23}$$

在這個例子中，只有第 1 點符合這個條件，其C_1=156 為最大值。

第二步驟利用主群和分裂群的點建立下列表格。

表 3-6. 主群分裂群距離表之一

主群內的點	主群內的點到分裂群內的點之平均距離(A)	主群內的點到主群內的點之平均距離(B)	(B)-(A)
2	10	25	15
3	7	23.4	16.4
4	30	14.8	-15.2
5	29	16.4	-12.6
6	38	19	-19
7	42	22.2	-19.8

(資料來源：[22])

表格建立後，觀察(B)-(A)的值，將其非負的最大值選出，為 16.4，因此將第 3 點加入分裂群中。主群為第 2 點、第 4 點、第 5 點、第 6 點、第 7 點；分裂群為第 1 點與第 3 點。

重複第二個步驟，則建立表 3-7。

表 3-7. 主群分裂群距離表之二

主群內的點	主群內的點到分裂群內的點之平均距離(A)	主群內的點到主群內的點之平均距離(B)	(B)-(A)
2	8.5	29.5	21
4	25.5	13.2	-12.3
5	25.5	15	-10.5
6	34.5	16	-18.5
7	39	18.7	-20.3

(資料來源：[22])

選取使(B)-(A)為非負的最大值之點，為第 2 點。因此主群為第 4、第 5、第 6、第 7 點，而分裂群為第 1、第 3、第 2 點。重複步驟二，直到(B)-(A)均為負值為止，如表 3-8。

表 3-8. 主群分裂群距離表之三

主群內的點	主群內的點到分裂群內的點之平均距離(A)	主群內的點到主群內的點之平均距離(B)	(B)-(A)
4	24.3	10	-14.3
5	25.3	11.7	-13.6
6	34.3	10	-24.3
7	38	13	-25

(資料來源：[22])

至此，原本的 7 個點被分為兩群（1, 3, 2）與(4, 5, 6, 7)。而這兩群再依照以上兩個步驟重複，直到主群分裂群距離表的(B)-(A)均為負為止。

在本研究中，所採用的凝聚法乃是群中心分析法。第四章便介紹資料的前處理、實驗的結構與流程等。

四、實驗架構、流程、結果與討論

4.1 資料來源與資料前處理

時間序列的資料來源來自證券基金會每日每分鐘的台灣加權指數，資料期間自民國 90 年 1 月 2 號到民國 93 年 4 月 8 號共 804 天。

以原始資料而言，每日有 271 個資料點（早上 9 點到下午 1 點 30 分）。若以每日為一個時間序列的單位，則這個時間序列的長度為 271。這樣的長度會使得距離度量的計算時間過長，另外一點，股價每分鐘的變動是極微小的，因此每隔數分鐘取一個點，可以更佳代表每日股價的變化情況。

在本研究中，資料點的選取為每隔 10 分鐘取一個點，因此每日有 28 個點，也就是說，代表每一天股價指數的時間序列長度為 28。圖 4-1 為使用 271 點代表一天與使用 18 點代表一天的圖形比較。

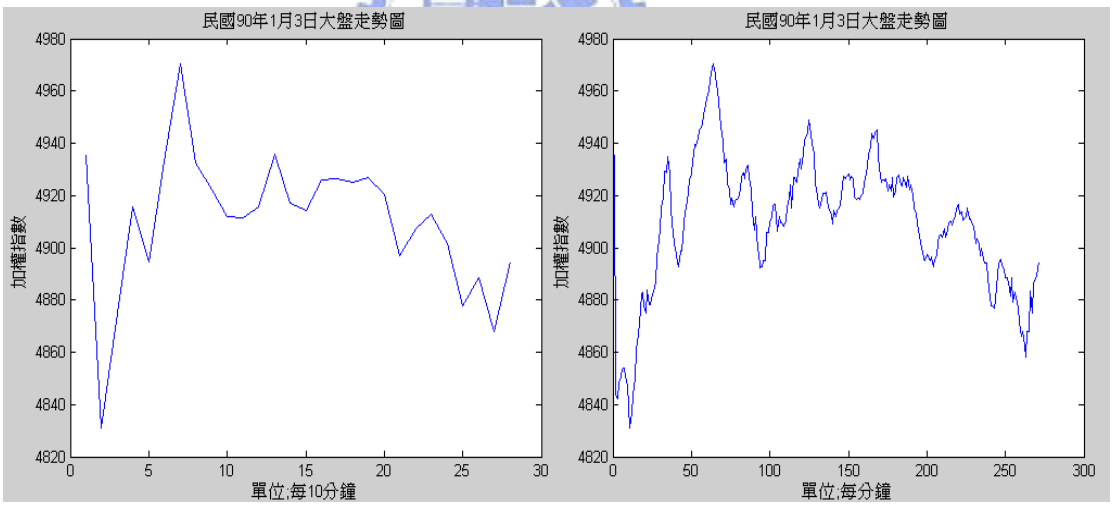

圖 4-1. 每天點數的圖形比較：28 個點對 271 個點

雖然台灣股市有股價漲跌幅百分之七的限制，這些原始資料值的差異仍非常大，也因此，資料前處理必須被執行。為了不忽略開盤時跳空的現象，每一天的資料點值會被轉換為與昨日收盤指數差異的比率，其算法如(4-1)。

$$T_n = (T_p - T_{yc}) / T_{yc} \quad (4\text{-}1)$$

T_n：今日指數轉換後的值
T_p：今日指數
T_{yc}：昨日收盤價

訓練資料的範圍除了以天為主外(例如 800 天裡 500 天為訓練資料)，每日的 9 點到 12 點的 19 個點為一天訓練資料的點。

同時，為了分成三類（漲類、平類、跌類）定義每日的漲跌值，從 12 點的 1 點 30 分的點被成為漲跌的定義區間。以圖 4-2 表示漲跌值的計算方式。

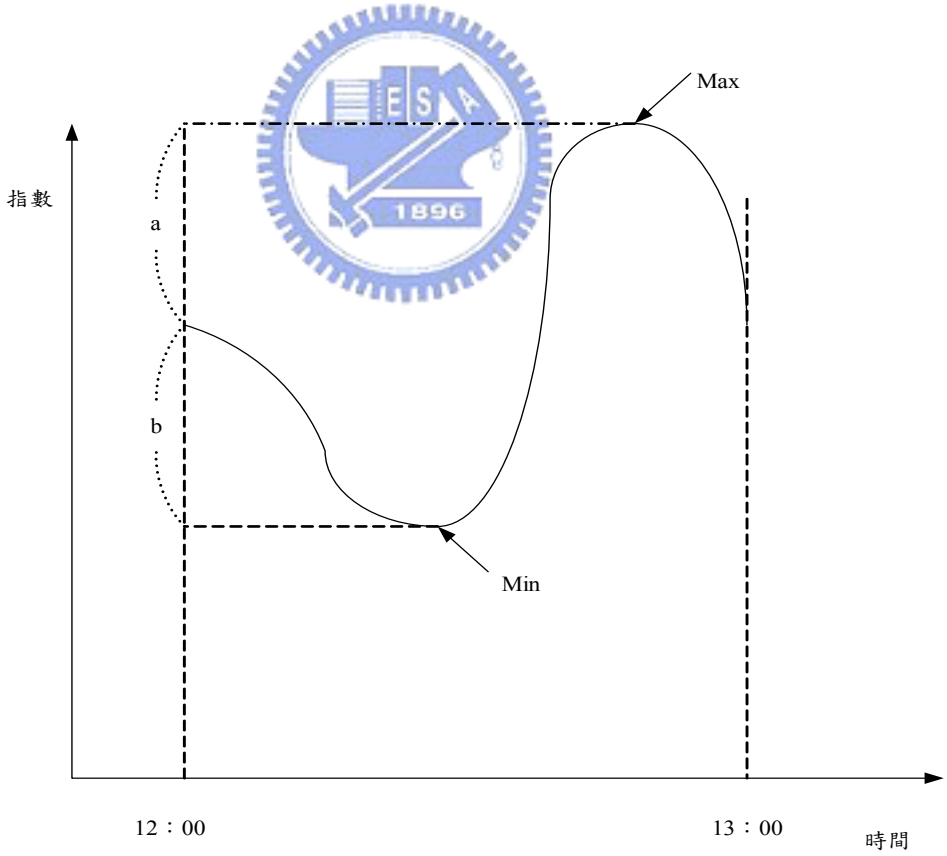

圖 4-2. 漲跌計算示意圖

漲跌值為(a-b)，a 代表 12 點到 1 點 30 分之間的最高指數與 12 點指數相較的漲幅(單位百分比)，同理 b 為 12 點到 1 點 30 分之間最低指數與 12 點指數的跌幅(單位百分比)。

漲跌值大於 0 的意義表示當天漲的趨勢大於跌的趨勢，而漲跌值小於 0 的意義則相反。以此漲跌值，對於訓練其間所佔的天數中，取其第 33 個百分位數，與第 67 個百分位數。用這兩個百分位數來當作漲跌值分為漲類、平類與跌類的區隔值。

用一個例子來說明這個概念。若訓練天數為 500 天，則這 500 天都存在一個(a-b)的值，並將這個值表示為P_j，j = 1,2 ,3, …, 500。P_j在這 500 天中的第 33 個百分位數為P_{33}而第 67 個百分位數為P_{67}。依循下列規則分類：

若P_j小於P_{33}，則將訓練資料歸為跌類。

若P_j大於P_{67}，則將訓練資料歸為漲類。其餘訓練資料則為平類。

4.2 實驗架構與流程

本研究的實驗結構為圖 4-2 與圖 4-3 所示。原始資料經過資料前處理後，經由比例分配分為訓練資料(Training Data)與測試資料(Testing Data)。

訓練資料的流程：

如圖 4-2，訓練資料經過分類，分成 3 類，然後針對每一類做距離度量矩陣。因此實驗前段，會有 3 組距離度量矩陣。資料轉換成距離度量矩陣後，實驗開始進入分群的階段。此時，階層分群法會針對三個距離矩陣去做分群，也就是說，三個樹狀階層結構會藉由階層分群法被產生出來。

這些產生出來的群，必須經過進一步的處理。若群內點的個數為 1，則捨棄這個群。這個被捨棄的群不會延續到測試階段。若訓練群內的個數大於 1，則求其群中心，並求出群中心與群內點的最大距離為群半徑，群中心與群半徑會在測試流程中作為某些規則的判斷。

而本研究所採用的距離度量有三種：歐幾里德距離、動態時間扭曲距離、trie 結構動態時間扭曲距離。而實驗的目的在於測試哪一個距離量度的分群表現最好。

測試資料的流程：

如圖 4-3，測試資料將與各類中的群相比較，並選出各類群的最小距離。因為訓練資料被分為 3 類，因此最小距離值有 3 個：D_{up}、D_{dull}、D_{down}。圖 4-4 為判斷訓練資料屬於何類的決策樹。圖 4-4 中，R_{dull} , R_{down}, R_{up}為測試資料所對到群之群半徑。為了節省空間與更容易被瞭解，決策樹只畫出D_{up}為最小值的情況。

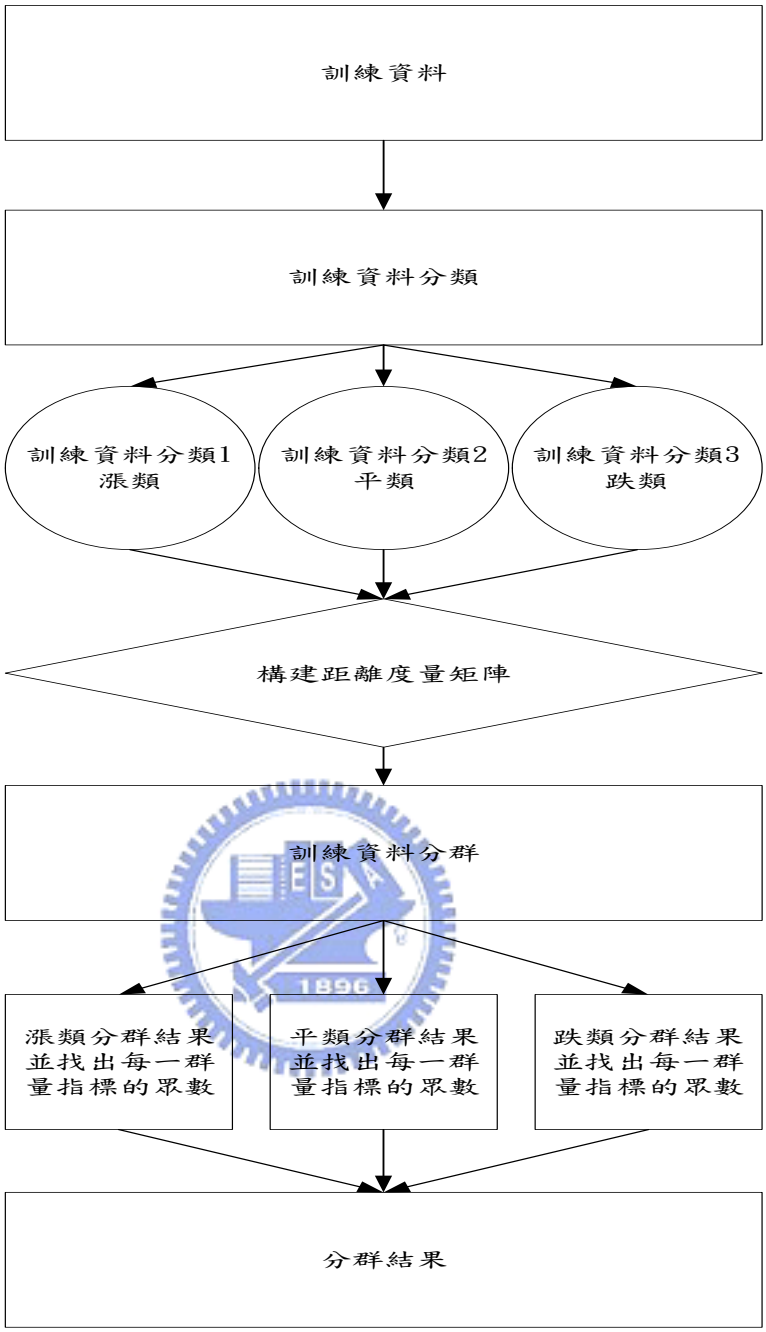

圖 4-3. 資料訓練流程架構

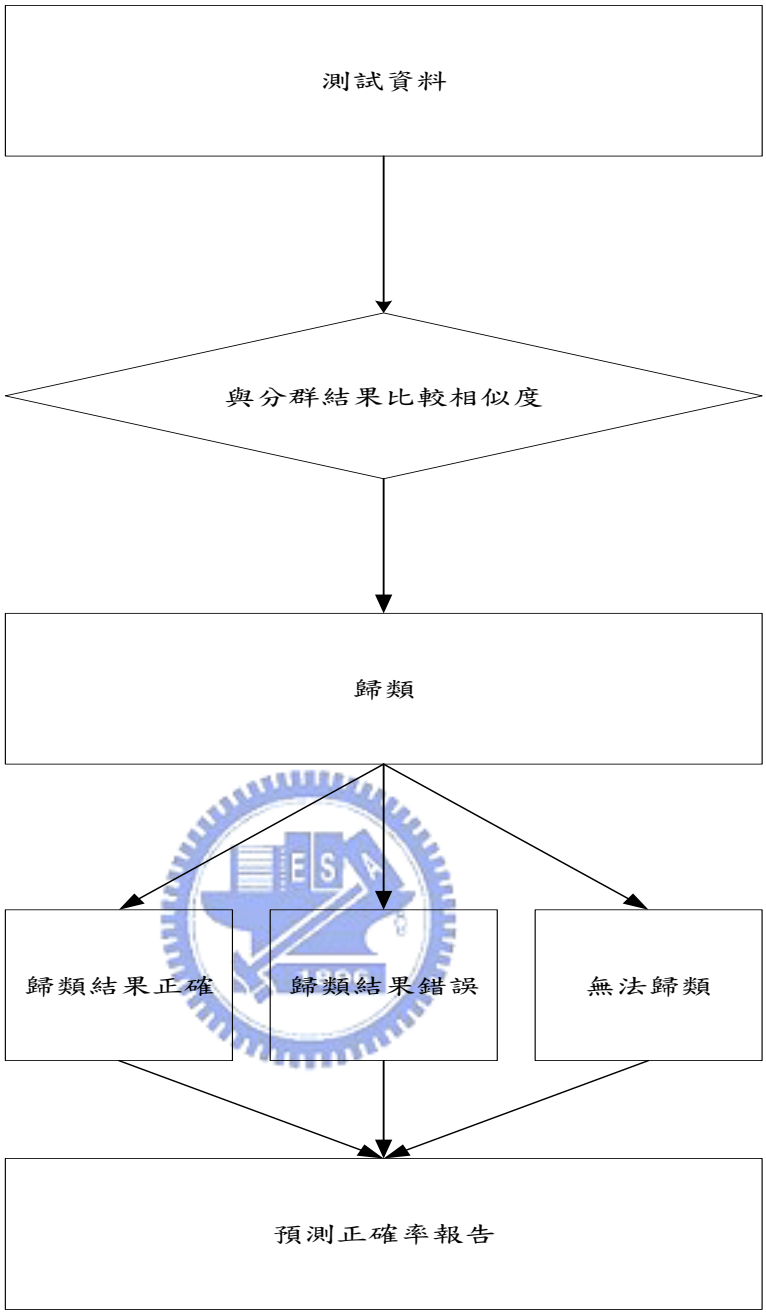

圖 4-4. 資料測試流程架構

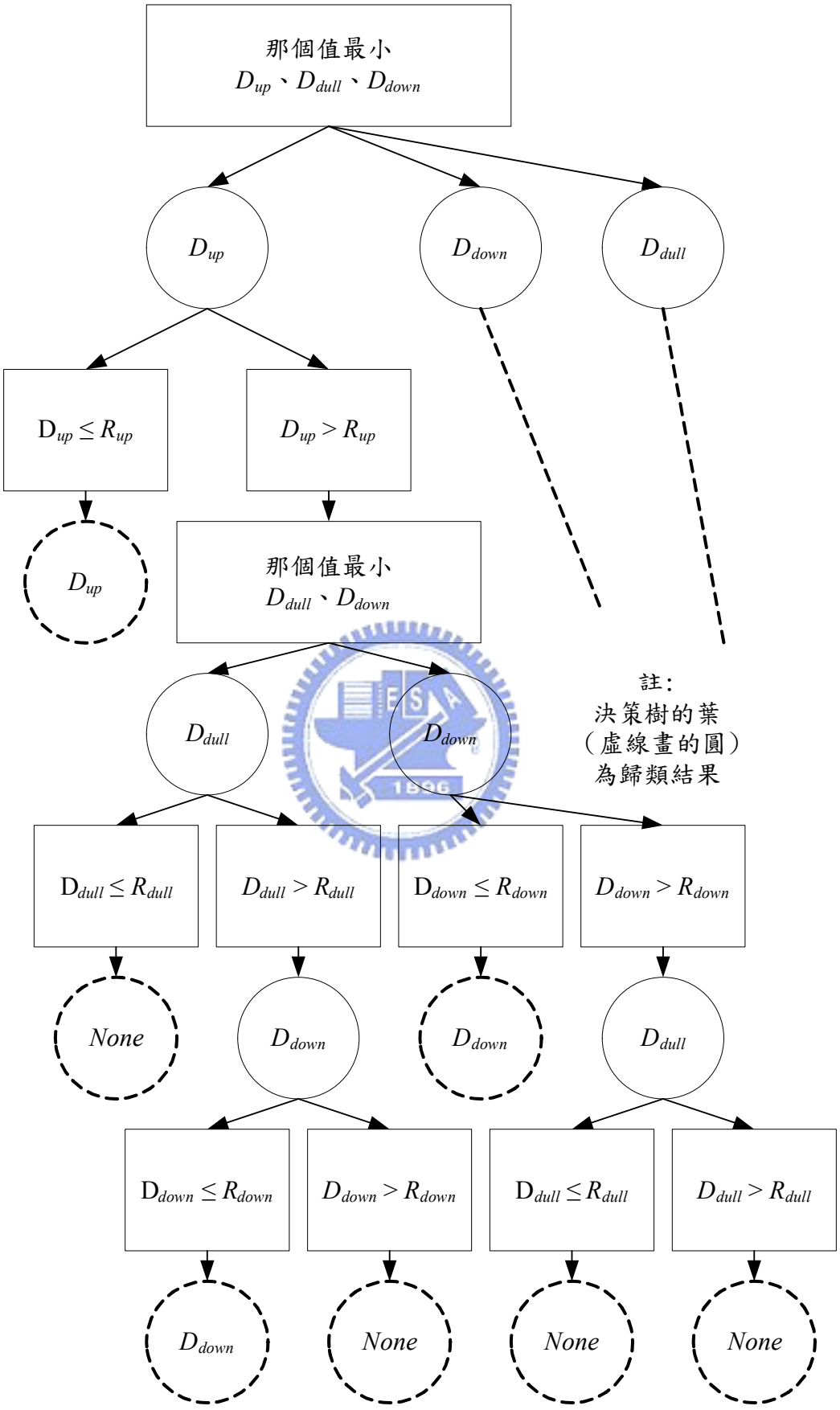

圖 4-5. 測試資料的決策樹

4.3 實驗的結果

實驗分為依據距離量度分為 3 組分別為歐幾里德距離、動態時間扭曲距離以及 trie 結構動態時間扭曲距離，而依據訓練天數與測試天數比分為 5 組；所以全部有 15 組資料。如表 4-1、表 4-2、表 4-3、表 4-4 與表 4-5。

表 4-1. 距離度量與股市分鐘資料分析結果

	Euclidean	DTW	Trie-structure DTW
執行時間(秒)	1.17	195	156.39
成功預測個數	47	186	169
失敗預測個數	167	152	155
可預測個數	214	338	214
成功率(%)	21.9	55.02	52.16
訓練天數	300	300	300
測試天數	503	503	503

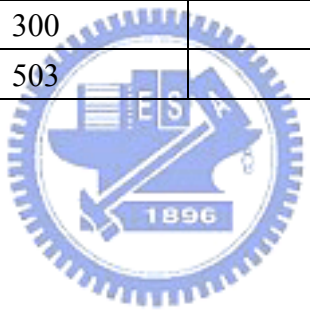

表 4-2. 距離度量與股市分鐘資料分析結果

	Euclidean	DTW	Trie-structure DTW
執行時間(秒)	1.4	240.6	198.9
成功預測個數	34	141	139
失敗預測個數	115	113	123
可預測個數	149	254	262
成功率(%)	22.8	55.5	53.05
訓練天數	400	400	400
測試天數	403	403	403

表 4-3. 距離度量與股市分鐘資料分析結果

	Euclidean	DTW	Trie-structure DTW
執行時間(秒)	2.03	290.34	250.4
成功預測個數	32	111	104
失敗預測個數	100	82	84
可預測個數	132	193	188
成功率(%)	24.2	57.51	55.3
訓練天數	500	500	500
測試天數	303	303	303

表 4-4. 距離度量與股市分鐘資料分析結果

	Euclidean	DTW	Trie-structure DTW
執行時間(秒)	3.03	405.3	332.5
成功預測個數	29	90	65
失敗預測個數	59	68	51
可預測個數	88	158	116
成功率(%)	32.9	56.9	56.03
訓練天數	600	600	600
測試天數	203	203	203

表 4-5. 距離度量與股市分鐘資料分析結果

	Euclidean	DTW	Trie-structure DTW
執行時間(秒)	4.79	507.1	424.4
成功預測個數	20	41	40
失敗預測個數	35	27	26
可預測個數	55	68	66
成功率(%)	36.3	60.2	60.6
訓練天數	700	700	700
測試天數	103	103	103

4.4 實驗的討論

本研究對於距離度量方法學的比較，從兩個方面去探討，一個是時間，一個是預測的正確性。這兩個面向以圖 4-4 與圖 4-5 表示。

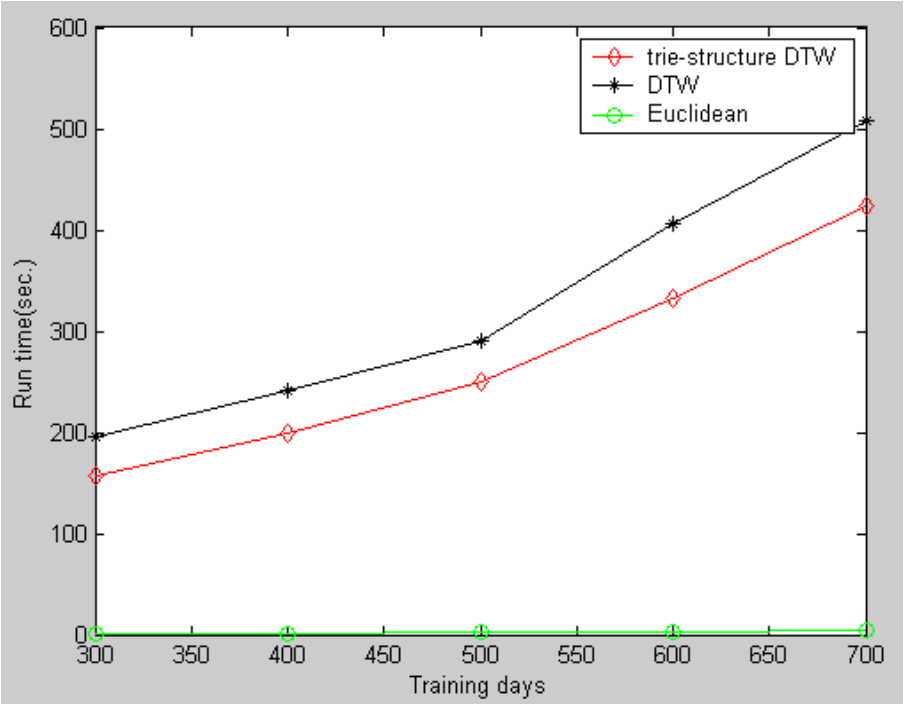

圖 4-4. 距離度量與時間

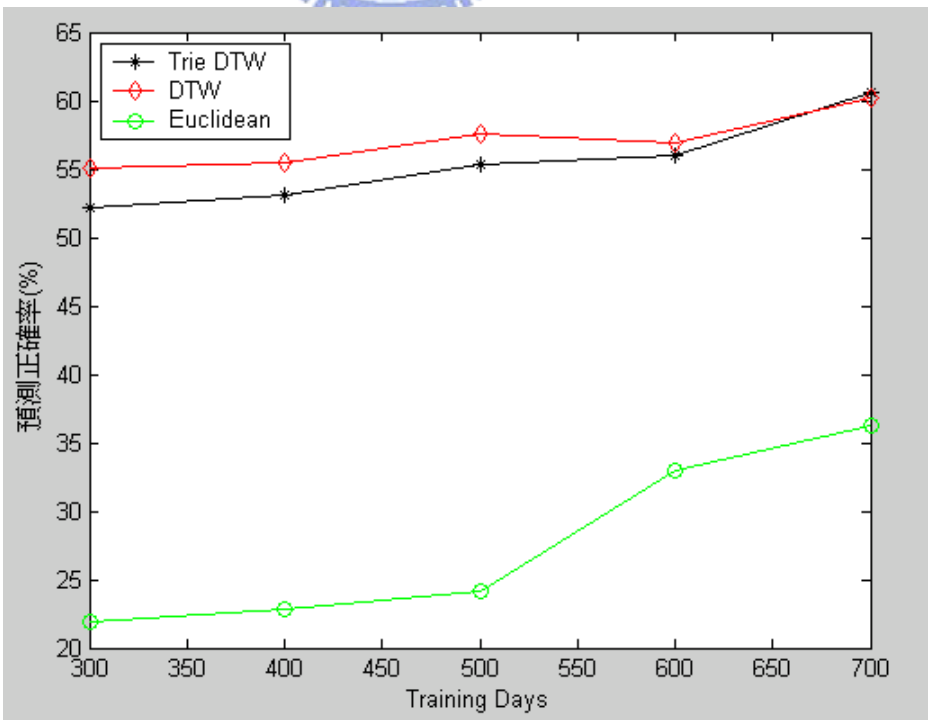

圖 4-5. 距離度量與準確率

使用不同的距離度量，確實會對分群的結果產生影響，進而影響測試的準確率與距離度量的運算時間。同時，掌握越多的訓練資料，也對於準確率的提升有一定的影響。計算時間方面，trie 結構的動態時間扭曲法的確在時間的運算上較傳統的動態時間扭曲法來的少，這個現象在訓練天數越多的情況下，更加明顯。同時，trie 結構 DTW 預測的準確率並沒有因此明顯下降。

五、結論

5.1 結論與本研究貢獻

動態時間扭曲法，應用在訊號處理的領域有極佳的表現。然而，這個方法並非只能應用在訊號處理。資料挖掘的興起，使得大多數圖形識別的技術能夠跨足其他非相關的領域。財經資料的分析，引入圖形識別的概念後，有進一步的成效。

而動態時間扭曲法，被引入分析金融資料上，無論是相似度的搜尋或是距離度量的運算方面，均有不錯的成效。然而其運算時間過長使得它應用在及時的預測或是大量資料的處理卻也是它被詬病之處。在這競爭激烈的時代，時間就是金錢，掌握更多資訊，就有更多的競爭力。

因此，本研究主要是提出一個方法，來改善動態時間扭曲法運算過慢的缺點，並將這個方法應用在時間序列的資料挖掘上面。而這個方法相較傳統的歐幾里德距離在時間序列的資料挖掘上有更佳的表現。

而結果則顯示，trie 結構動態時間扭曲法的使用，可以得到與動態時間扭曲法相近的預測效果，但運算時間卻縮短了，這使得動態時間扭曲法可以應用在及時的預測上。雖然時間在本研究未被大幅度的縮減，但是應用在相似度搜尋方面確有很大的成效。這個原因在於當執行序列間的相似度搜尋時，圖形序列的產生源於 sliding window 的方式

5.2 未來發展

本研究的未來發展，主要可以分成四點。

第一點，資料的前處理，一直是資料挖掘中的重要課題。良好的資料前處理可以使得資料挖掘所分析的結果更有可信度。而本研究的資料前處理，將資料分為漲平跌三類，可以再被改良，已提高分群的有效性與預測的正確性。

第二點，動態時間扭曲法的再改良，使得改良後距離度量更接近傳統的動態時間扭曲法，並且加速運算速度，仍是本研究未來持續發展的方向。Trie 結構概念的再延伸，如把兩個比對相同的序列當成一個點，以更大的尺度來看待 trie 結構，則是本研究未來的另一個重要方向。而同時，從輸入資料的面向、演算法執行的面向、硬體改良的面向，這三個面向去研究如何改良動態時間扭曲法。

第三點，分群法是掌握行為的方法之一，尋找更加有效的分群法，並確實掌握『有意義』的圖形(Patterns)，進一步的提升預測的正確率，是未來發展的另一個方向。

第四點，雖然本研究方法應用於分群預測上，執行時間較於動態時間扭曲法並未大幅度的減低，但是應用於相似度搜尋時，卻可以降低大部分的搜尋時間。因此，應用本研究方法於相似度搜尋，再由相似度搜尋延伸至預測的應用上，是本研究積極延展的一個方向。

參考文獻

[1] Henry Gleitman著，心理學，洪蘭譯，遠流出版社，台北，1991，民國 86 年。

[2] Wong, P.H.W., et al., "Reducing Computational Complexity of Dynamic Time Warping-based Isolated Word Recognition with Time Scale Modification", Signal Processing Proceedings, Pages:722 - 725 vol.1, 1998.

[3] 蔡文智，『以隱藏式馬可夫模型應用於股市單日交易預測上』，國立交通大學，碩士論文，民國 91 年。

[4] 劉科成，『使用動態規劃搜尋股市線形之探討與實作』，暨南國際大學，碩士論文，民國 89 年。

[5] J. T. Tou, "Feature Extraction in Pattern Recognition", Journal of pattern recognition, Vol1, pp3-11, 1968.

[6] C. W. Therrien, Decision estimation and Classification : an Introduction to Pattern Recognition and Related Topics, John Wiley & Sons, New York,1989.

[7] 李上銘，『語音辨認中基於主成份分析之進一步技術』，國立台灣大學，碩士論文，民國 90 年。

[8] 丁鎮權，『指紋辨識系統設計』，淡江大學，碩士論文，民國 92 年。

[9] Jiawei Han, Micheline Kamber, Data Mining: Concept and Techniques, Morgan Kaufmann), p418-427,2002.

[10] C. Chartfield. The Analysis of Time Series: An Introduction, 3rd ed. New York, Chapman and Hall, 1984.

[11] R. H. Shumway, Applied Statistical Time Series Analysis. Eaglewood Cliffs, NJ: Prentice Hall, 1988.

[12] R. Agrawal, et al., "Efficient Similarity Search in Sequence Databases". In Proc. 4th Int. Conf. Foundation of data organization and algorithms, Chicago, II, Oct. 1993.

[13] R. Agrawal, et al., "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time Series Databases", In Proc. 1995 Int. Conf. Very Large Data Bases(VLDB'95), page 490-501, Zurich, Switzerland, Sept. 1995.

[14] R. Agrawal, et al.,."Query Shapes of Histories". In Proc. 1995 Int. Conf. Very Large Data Bases(VLDB'95), page 502-514, Zurich, Switzerland, Sept. 1995.

[15] S. Park et al., "Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases", In Proc. 2000 Int. Conf. Data Engineering(ICDE'00), pages 23-32, San Diego, CA, Feb. 2000.

[16] C.-S. Perng, et al., "Landmarks: A New Model for Similarity-based Pattern Querying in Time Series databases", In Proc. 2000 Int. Conf. Data Engineering(ICDE'00), pages 33-42, San Diego, CA, Feb. 2000.

[17] R. Agrawal, and R. Srikant, “Mining Sequential Patterns”. In Proc. 1995 Int. Conf. Data Engineering (ICDE’95), pages 3-14, Taipei, Taiwan, Mar. 1995.

[18] C. Faloutsos, et al., “Fast Subsequence Matching in Time Series Databases” In Proc. 1994 ACM-SIGMOD Conf. Management of data, pages 419-429, Minneapolis, MN, May 1994.

[19] M. J. Zaki, et al., “PLANMINE: Sequential Mining for Plan Failures”, In Proc. 1998 Int. Conf. Knowledge discovery and data mining(KDD’98), pages 369-373, New York, Aug.1998.

[20] J. Han, et al., “Efficient Mining of Partial Periodic Patterns in Time-series Database”, In Proc. 1999 Int. Conf. Data Engineering(ICDE’99), pages 106-115, Sydney, Australia, Apr. 1999.

[21] Keogh, E. and Kasetty, S.,. “On the Need for Time-series Data Mining Benchmarks: A Survey and Empirical Demonstration” In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 102-111 July 23 - 26, 2002.

[22] Robert Goodell Brown, Smoothing, Forecasting, and Prediction of Discrete Time Series, Prentice Hall, 1962

[23] Brain Everitt, Cluster Analysis 2rd, p17, Heinemann Educational Books, London,1980.

[24] Augustine H. Gray Jr., John D. Markei, “Distance Measures for Speech Recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. Assp-24, No. 5, pages 380-391, Oct. 1976.

[25] Phillip Juffs, et al., “A Multiresolution Distance Measure for Images”, IEEE Signal Processing Letters, Vol. 5, No. 6, June, 1998

[26] Das G.,et al., “Rule Discovery from Time Series”, International Conference of Knowledge Discovery and Data Mining, pages 16-22, 1998.

[27] 劉致和，『臺灣地區燙傷住院治療型態之研究－－應用階層式集群分析於全民健保資料庫』，台北醫學院，碩士論文，民國 91 年。

[28] Berndt, D., Clifford, J., ”Using Dynamic Time Warping to Find patterns in Time Series”, AAAI-94 Workshop on Knowledge Discovery in Database, pages 229-248, 1994.

[29] Keogh, E., “Exact Indexing of Dynamic Time Warping”, In 28th International Conference on Very Large Data Bases, pages 406-417, 2002.

[30] Gregory, N. Stainhaouer, George Carayannis, “New Parrallel implementations for DTW algorithms”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 38 april 4 pages 705-711, 1990.

[31] Kurt Maly, “Compressed Tries”, Communications of the ACM, Vol 19, No. 7, July,pages 409-415, 1976

[32] Chen A. P., et al., “Applying Trie-Structure to Improve Dynamic Time Warping on Time-Series Stock Data Analysis”, International Conference on Artificial Intelligence and Application, 2004

[33] Cormack, R. M., ”A review of Classification”,. Journal of the Royal Statistical Society, Series A, 134, No. 4, pages 321-367, 1971

[34] Johnson, S. C., “Hierarchical Clustering Schemes”, Psychometrika, 32, 241-254, 1967

[35] Tapas Kanungo, et al., “An Efficient *k*-Means Clustering Algorithm: Analysis and Implementation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No.7, July, 2002

[36] Carmichael, J. W., et al., “Finding Nature Clusters”, Syst. Zool., 17, pages144-150, 1968.

[37] Needham, R. M., “Automatic Classification in Linguistics”, The Statistician, 17, pages 45-54, 1967.

[38] Sawrey, W. L., et al., “An Objective Method of Grouping Profiles by Distance Function”, Educ. Psychol. Measur., 20, pages 651-674, 1960.

[39] Gower J. C. and Ross, G. J. S., “Minimum Spanning Trees and Single Linkage Cluster Analysis”, Applied Statistics, 18, pages 54-64, 1969.

[40] 吳勝修，『應用股票趨勢技術分析於動態投資組合保險中之操作策略』，國立交通大學，碩士論文，民國 92 年。

[41] 曾士育，『以自組織映射網路探勘金融投資決策之研究』，國立高雄第一科技大學，碩士論文，民國 92 年

[42] 王湘蕙，『時間數列資料分群方法—在探討台灣上市電子公司股票特性之應用』，國立台北大學，碩士論文，民國 91 年。

[43] S. Benninga and B., Czaczkes, Financial Modeling, Cambridge, MA:MIT Press, 1997

[44] 許育嘉，『結合小波分解和小波神經網路於非定性財務時間序列之預測』，國立交通大學，碩士論文，民國 91 年