第一章 緒論

1.1 研究動機

自1990年代中期起,為因應電子商務之興起及政府積極鼓勵產業電子化的影響,國內企業相繼導入電子化,由中華民國對外貿易發展協會公佈,我國資訊服務業近年來的複合成長率約為10%,2005年資訊服務業產值已經超過新台幣1846億元,預計2008年成長到新台幣3,000億元。代表社會上對於電子資訊的需求與日劇增,電子商務化對於企業更是勢不可擋。企業經由電子化過程後,往往倚重電子資料庫,並將顧客與商品以及相關資料,以電子檔案方式儲存,當資料龐大時,將資料分類、分群、關聯性建立,變得十分重要。資料分群的結果,可以幫助簡化繁雜的資料、建立資料的分類規則,使得資料接收者,能以此訊息加以判斷及決策。經營者往往可從龐大的資料中,將分群的資料轉換成商業訊息,由此可知其重要性。

目前資料分群已被廣泛的使用在各領域,經由國內博碩士論文查詢可知,近十年間,有關分群方面的論文,就已超過150篇以上,著名的外文資料庫IEEE超過17000筆,而SDOS則超過7700篇文獻,是從事分群相關的研究,其應用層面相當廣泛,例如:資料庫管理、醫藥、社會研究、影像處理、生物資訊...等等。

1.2 研究目的

資料分群,是資料分析的前置動作。資料分群是一種將資料點分類成群的方法,其主要的目的在於找出資料中較相似的幾個群體,並找出各個群聚的中心點。資料透過分群可以將群體資料中性質一樣的資料歸類為群體,使得同一群體內的資料相似性高,不同群體內的資料的相似性低。而驗證分群結果方式通常是以迭代次數以及誤差率計算。其中迭代次數是以效率為績效指標,誤差率則是計算將資料分錯的比率。

目前較常使用的分群方法為階層式與分割式,階層式是將一群尚未分群的資

料，透過分裂或合併找出資料中關聯性；分割式分群法，則是將一群尚未分群的資料，分割成數個群體，以便找出資料的關聯性。

在傳統的分割式分群法較常使用到的便是K-Means分群法，但我們發現K-Means分群法不能處理大量的資料分群數目，與多維度資料分群，並且也無法解決資料點重疊的狀況，因此稱不上是一個完善的分群技術。而另一啟發式演算法則是粒子群體最佳化演算法(Particle Swarm Optimization, 簡稱PSO)，此一方法為較新的演算法，優點是較不易陷入區域最佳解，缺點則是所需運算時間較長，且需要較大的群體數目。上述兩種方法都存在著極需改善的問題，有鑑於此，本論文將使用整合Nelder-Mead單體法(Nelder-Mead Simplex Method，簡稱NM)與粒子群體最佳化演算法(Hybrid Nelder-Mead Simplex Method and Particle Swarm Optimization，簡稱NM-PSO)進行資料分群，目前尚未有文獻以此演算法進行資料分群，而以NM-PSO處理多區域函數(Fan *et al*.,2004)和影像處理而言(Zahara *et al*., 2005)，其效果都優於其他演算法，不僅所需群體較小，對於降低誤差率和提升運算速度而言，也有大幅度的改善。並且將整合K-Means、PSO (簡稱KPSO)與K-Means、NM-PSO(簡稱KNM-PSO)，希望利用K-Means快速收斂的優點，再以NM-PSO來補足K-Means易陷入區域最佳解的缺失，因此本研究將同時使用KNM-PSO演算法、KPSO演算法、NM-PSO演算法、K-Means分群法和PSO演算法，應用於資料分群，並比較其結果，期望KNM-PSO演算法應用於資料分群，有下列優勢:

1. 所需評估函數運算次數較少。
2. 處理多維度的問題。
3. 同時解決實際資料庫與人工資料庫的分群問題。
4. 需要群體數較少。
5. 距離總值將為最佳。

1.3 研究流程

本論文所探討的是整合Nelder-Mead單體演算法與粒子群體最佳化演算法為基礎，解決多維度資料分群的問題，研究流程如圖1.1 所示，以下則為本研究流程說明。

1. 確立研究架構

   經過長時間討論與思考，本論文以NM-PSO演算法為主架構，將應用其演算法於特定領域。

2. 確立應用領域

   熟讀相關論文及文獻，將使用NM-PSO演算法，應用領域鎖定於資料分群，並且深入瞭解資料分群的精神以及應用層面。

3. 收集相關文獻

   收集NM-PSO、K-Means、PSO與Nelder-Mead單體法以及任何與資料分群演算法有關之文獻，同時整理各種利用演算法為基礎解決分群上的問題，並且歸納出相關的理論及彙整出K-Means、Nelder-Mead單體法與PSO 演算法的缺失。

4. 撰寫程式

   將K-Means、PSO與NM，以Matlab程式語言撰寫，經過實際的接觸，發現並了解相關問題。將Nelder-Mead單體法與PSO演算法，整合為NM-PSO演算法，再將這三種分群技術，整合KNM-PSO演算法與KPSO演算法，並且尋找適用之實際資料庫以及設定人工資料範圍，以便實驗測試。

5. 實驗與數據分析

   將K-Means、PSO、NM-PSO、KNM-PSO、KPSO，應用於上述資料庫分群，分析到達全域最佳解之速度和比例，將數據表格化比較，並且以實際圖示。

6. 結論與未來研究方向

以實驗數據證明NM-PSO，是否能有效的應用於資料分群，並且預期結果將優於K-Means、PSO。而KNM-PSO之結果將明顯優於上述四種分群技術。最後提出自我看法，以及未來研究方向。

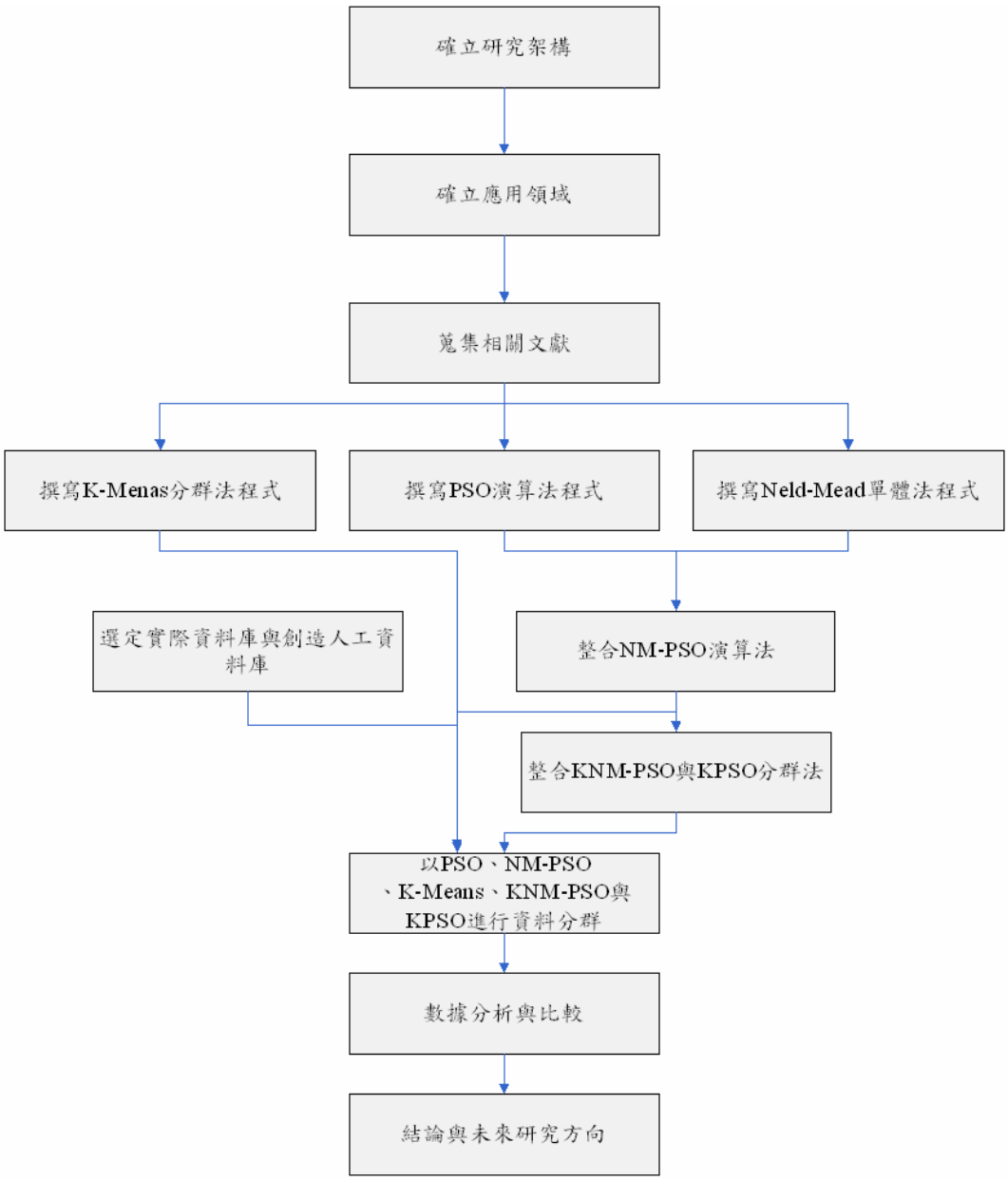


圖 1.1 研究流程圖

第二章 文獻探討

本章將分為兩個部分。第一部分是關於資料分群技術介紹與探討，其中包含了過去五年之中部份使用演算法於分群技術之文獻、階層式分群、分割式分群以及K-Means分群法。另外一部分則介紹粒子群體最佳化、Nelder-Mead單體法。

2.1 資料分群技術與啟發式演算法應用

近年來，分群技術已經廣泛被應用於各式各樣領域上的分析及探討，它是屬於數理統計中主成分分析(Principle Component Analysis)的一環(Agrawal *et al*., 1998)。 分群技術是用來探索資料聚集關係的一個工具，在真實世界中資料往往都具有群聚的現象，也就是資料會呈現群體的分佈情形，而資料的分群探索就是希望能挖掘出這個現象(Duda and Hart, 1973)，(Jain and Dubes, 1988)。資料分群，便是將多維度空間的資料點，以其特性轉換為數據化，並以資料群體中心點歸類。舉例來說，天文學家試圖瞭解恆星的亮度與溫度之間的關聯性，如圖2.1所示。可知眾多的恆星，落在以亮度和溫度所構成的二維空間中。而經由分群的結果，落在三個群集中。每一個群集內，溫度與亮度之間的關係是接近的，但群集之間的溫度與亮度關係卻不同。經由此分群動作，天文學家便可從中了解恆星的亮度與溫度之間的關聯性 (彭文正，2001)。

目前有許多學者使用啟發式演算法來處理資料分群的問題，表2.1則整理出部份啟發式演算法之列表，啟發式演算法一般的理論基礎是以觀察自然及社會現象而得到啟發。而表2.2則為近五年之間，應用基因演算法與粒子群體最佳化演算法及其相關改良型啟發式演算法，處理資料分群技術之部分文獻，在此提出基因演算法，是因為其理論基礎，與粒子群體演算法較相近。由表2.2可知，已有學者使用此兩種演算法來處理資料分群問題得文獻。因此本篇論文將使用何怡偉博士於2004年提出，整合Nelder-Mead單體法與粒子群體最佳化演算法，來處理資料分群之技術，並預期結果將優於其他啟發式演算法。

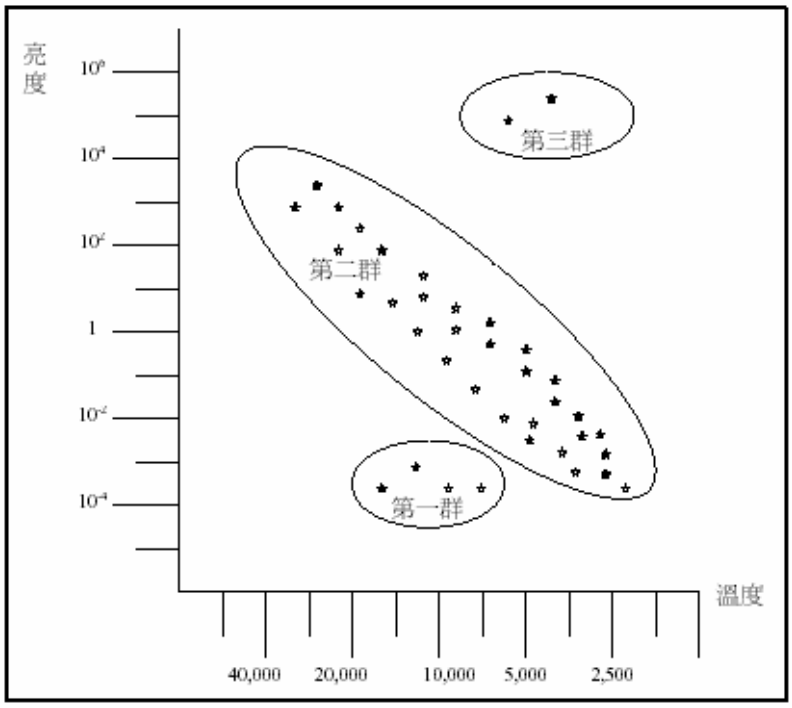


圖2.1 恆星溫度與亮度分群圖

資料來源：彭文正譯，2001

表 2.1 部份啟發式演算法列表

提出學者	年份	演算法	概念源起
Holland, J.	1975	基因演算法	達爾文演化論
Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P.	1983	模擬退火法	高溫狀態退火結晶
Glover, F.	1989	禁忌演算法	仿造人類智慧
Dorigo, M. and Colorni, V.	1992	蟻群演算法	螞蟻間訊息之傳遞
Kennedy, J. and Eberhart, R. C.	1995	粒子群體最佳化演算法	鳥群及魚群遷徙

表 2.2 部份分群技術之文獻表

學者	年份	應用方法	評估方法	資料庫型態
Kennedy, J.	2000	改良式粒子群體最佳化演算法	誤判率、總距離值	實際
Sanghamitra B., Ujjwal, M.	1999	基因演算法	總距離值	實際、人工
葉承銓	2002	適應性基因演算法	總距離值、誤判率、迭代次數	實際、人工
Merwe, D. V. and Engelbrecht, A.	2003	粒子群體最佳化演算法	總距離值、誤判率	實際、人工
陳慶逸	2003	粒子群體最佳化演算法	總距離值、誤判率	人工
Cheo, C. Y. and Ye, F.	2004	粒子群體最佳化演算法	總距離值	人工
陳孟佐	2004	整合階層式基因與粒子群體最佳化演算法	誤判率	實際、人工

2.1.1 階層式分群

分群技術一般可分為階層式(hierarchical)以及分割式(partitional)兩種方法。早期較著名的聚合型演算法有單一連結法(Single-link Clustering)、完整連結法(Complete-linkage Clustering)、比較質心法(Comparison of Centroids Clustering) (Kaufman and Rousseeuw, 1990)，而它們共同的問題是：只要前面有一個步驟決定錯誤，之後便無法改正，而導致錯誤的分群。

階層式分群法較常使用的模式有兩種，分別是分裂式階層群集法（Divisive）以及凝聚式階層群集法（Agglomerative），分裂式階層群集法進行分群的方向是由上往下，先將所有物件歸類為同一群，逐漸將屬性差異大的群體分開，直到分成單一物件。而凝聚式階層群集法則與上述方法相反，首先將每個物件歸類為一群，並逐漸將屬性相似度高的群體合併，直到所有的物件合併為單一群體。

如圖 2.2 所示，資料點集合$H\{X_1,...,X_9\}$，若以凝聚為例，第一次凝聚，資料點X_1、X_2、X_3合併為群體A，資料點X_4、X_5則合併至群體B，最後資料點X_6、X_7、

X_8、X_9責備合併為群體C。經過了第一次合併，將 9 個資料點以屬性接近程度分為 3 個群體，而第二次合併，則是將群集A、B中的 5 個資料點，包含X_1、X_2、X_3、X_4、X_5，合併至群體D，最後，經由第三次合併，將D和C兩群體進行合併至 H，此為凝聚式階層群集法的基本流程。反之，由上而下進行分裂的方法便是分裂式階層群集法。

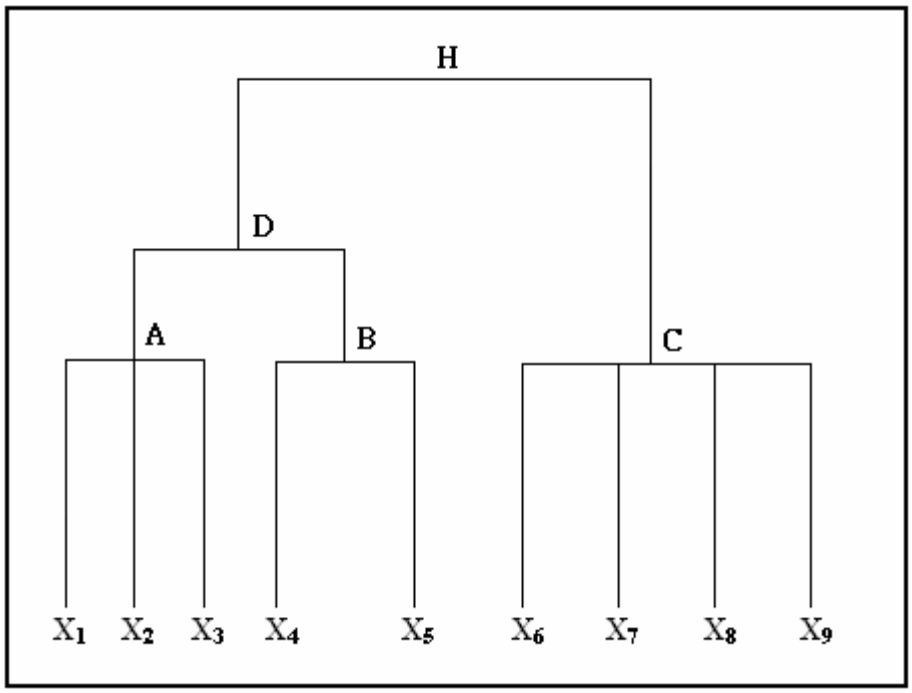


圖2.2 凝聚式與分裂式階層群集法示意圖

2.1.2 分割式分群

分割式群聚演算法是發展最早的分群技術，運用這一類的演算法的使用者必須先決定所要分割的群聚數目，再以重心點基礎（Centroid-based）或中心點基礎（Center-based）的方式進行分群(林育臣，2002)。簡單的來說，切割式分群法，是在一個 N 空間中，將資料切割成 K 群集。我們先定義一N維度空間中的 n個資料點，資料集合為$S=\{X_1, X_2, ..., X_n\}$，而$C_1, C_2, ..., C_k$ 則表示 K個群集的中心點。因此，資料分群必須滿足下列條件：

$$C_i \neq \Phi \; for \; i = 1, ..., K \tag{2.1}$$

$$C_i \cap Cj = \Phi,\ for\, i = 1,...,K,\ j = 1,...,K\, and\, i \neq j\, and \bigcup_{i=1}^{K} C_i = S \qquad (2.2)$$

2.1.3 K-Means分群技術

分割式分群法中最常被使用的便是 K-Means 分群法，K-Means是一個經常被使用且較簡單的分群方法。基本的概念便是以每個資料點到群體中心距離，並將資料點分派到距離最短的集群，找出每一個群聚的重心，也就是新的群體中心點，再加以分群，重覆此動作，當中心點不再改變，便可結束實驗。其完整的演算法，為以下所示：

1. 隨機產生起始群體中心。
2. 待分群的資料點為 n 個,分別是 Xi， i =1,2,...,n ,計算資料點與群體中心的歐基里德距離，群體C_j 類中的群體中心為Vj，當下列公式(2.3)成立時，將點Xi 歸屬於群C_j ：

$$\left\| X_i - V_j \right\| < \left\| X_i - V_p \right\|,\ p = 1,2,.....k\, and\, p \neq j \qquad (2.3)$$

3. 依據公式(2.4)計算每一個群體的重心並當作新的群體的中心：

$$V_j = \frac{1}{N_j} \sum_{X_i \in C_i} X_i \qquad (2.4)$$

4. 重新執行步驟2，直到群體中心點不再改變，才可結束演算步驟。

K-Means分群法的演算流程雖然相當簡單，但不能處理大量的資料分群數目，並且也無法解決資料點重疊的狀況，是十分為人詬病的，因此不能稱的上是一個完善的分群技術。

2.2 粒子群體最佳化演算法

粒子群體最佳化演算法是於1995年由Kennedy和Eberhart所提出(Kennedy and Eberhart,1995)。PSO的概念起源群體行為理論，其概念如圖2.3所示(Reynolds, 1987) ，兩位學者自觀察鳥群或魚群行動時，透過個體間特別的訊息傳遞方式，使整個團體朝同一方向、最終目標而去，是社會化的仿效同伴行為以及自我意識來尋求最佳化途徑的方法。圖2.4則是PSO演算法概念表示圖，首先隨機產生初始粒子群，每一個粒子都是一個求得最佳解的侯選者，粒子群會參考個體的最佳經驗，以及群體的最佳經驗，即是圖中白色個體，經過不斷的修正之後，粒子群會漸漸接近最佳解。

圖 2.3 模擬鳥群飛行示意圖

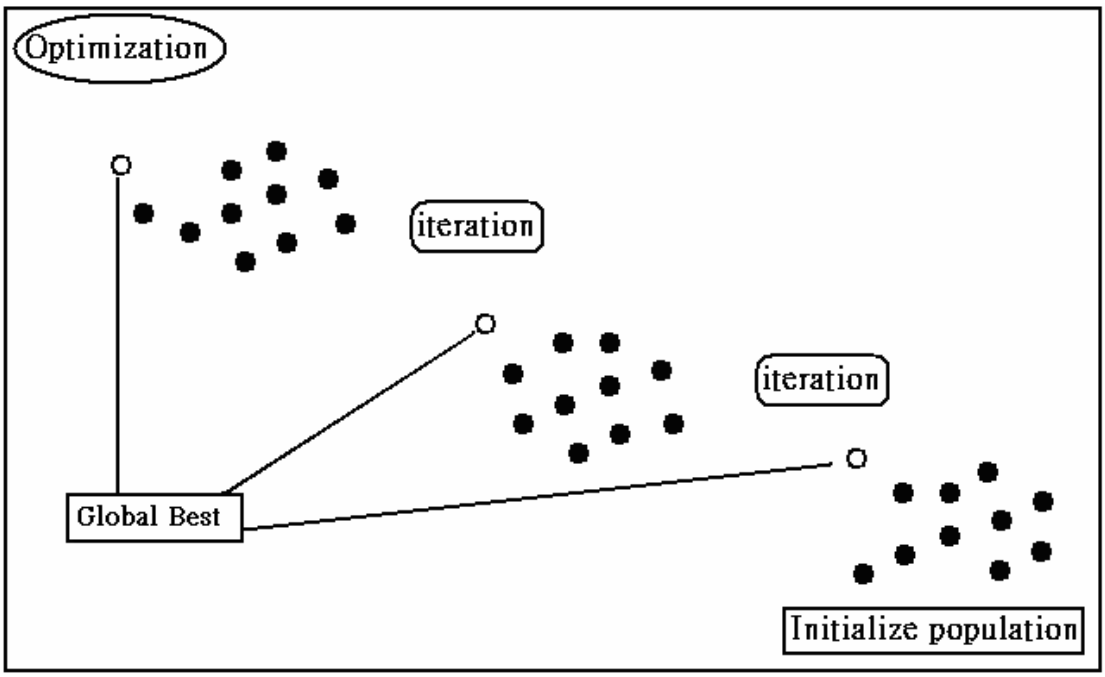


圖 2.4 PSO 概念表示圖

PSO 演算法演算流程如以下所示:

1. 設計出所需的評估函數，使得每個粒子所搜尋到的位置組都有一對應的數據值。
2. 決定粒子數目，通常設定為 5N，初始位置及速度以亂數產生，並設定 PSO 演算法中的所

   有參數，以及決定迭代次數。
3. 將粒子的所在位置代入評估函數，每個粒子依傳回評估值，並記錄全體粒子中最佳的位置，以及粒子個別最佳位置。
4. 執行 PSO 的計算式(2.5)來得到每個粒子每個參數維度新的移動速度，並以計算示(2.6)改變每個粒子的位置。

$$V_{id}^{New} = w \times V_{id}^{old} + C_1 \times rand() \times (p_{id} - X_{id}^{old}) + C_2 \times rand() \times (p_{gd} - X_{id}^{old}) \quad (2.5)$$

$$X_{id}^{New} = X_{id}^{old} + V_{id}^{New} \quad (2.6)$$

$rnad()$ 是介於 0 到 1 之間的亂數，w 為加速度，產生方式為 $[0.5 + (rand()/2)]$，V_{id}^{old} 為目前的速度，C_1 與 C_2 則是社會與自我認知權

重，皆設定為2，p_{id} 為個別粒子紀錄中曾經到達的最佳解，p_{gd} 是全部群體粒子紀錄中曾經到達的最佳解，X_{id}^{old} 表示目前粒子所在的位置，從公式(2.5)中可得更新後的速度V_{id}^{New}，再利用公式(2.6)便可求得新的群體位置 X_{id}^{New} 。

5. 重複步驟 3 及步驟 4 評估函數值以及更新速度及位置，比較是否優於之前所紀錄之全域以及區域最佳解，若較佳則取代之，反之，則以之前紀錄最佳值進行運算。
6. 重複步驟 3.至 6.直到演算回合數符合演算停止條件，便可求得群體最佳參數。

PSO 的主要精神在於讓每個粒子獨立搜尋最佳解的可能，並記錄個體之最佳解，因此每個粒子都會擁有的個別最佳解，粒子依照這些個別最佳解而去修正下一次搜尋的粒子速度與位置，這正是所謂的「認知模式」。若每個粒子只依照群體最佳解來修正下一次的粒子速度與位置，就是所謂的「社會模式」。而在綜合上述兩種模式中所用的數學模型，便是合併了社會模式及認知模式的綜合模式，意即粒子會同時參考群體以及個體的最佳解來修正下次的搜尋方向及速度。

所以粒子的下一次移動速度除了被一定程度的自我認知影響外，也將受群體社會化認知而做一部分的修正，並且可以經由參數的設定，改變兩種模式的重視程度。所以一開始雖然粒子群平均分佈在參數空間中，並且各自尋找個體最佳解，然而經由數次代迭後，大多數粒子將會逐漸移動至全域最佳解附近，而形成一個粒子群，並往全域最佳解的逼近，如此的社會模式則能避免粒子群陷入區域最佳解，而有些離粒子群較遠的粒子，則是認知模式的原因，假使有一粒子群外的粒子在另一位置中，找到更佳的全域最佳解，則社會模式將會使所有的粒子向其逼近。若干次迭代後，粒子群將會再度的逼近於新的全域最佳解附近，因此 PSO 最後將可經由社會及認知兩種模式，找到全域最佳解，如圖 2.5 所示。

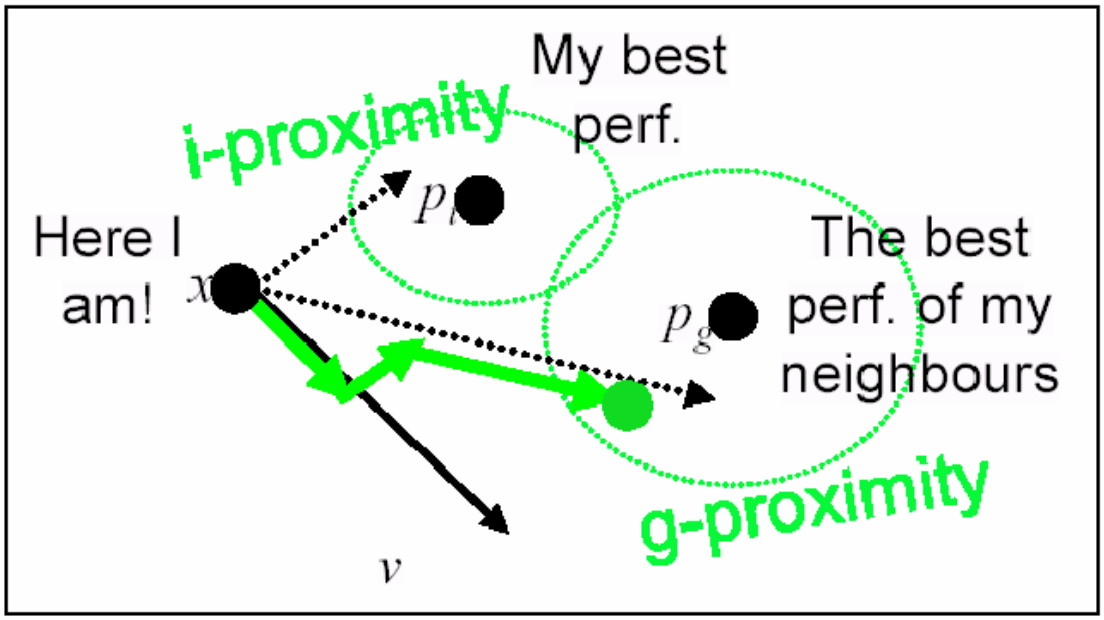


圖 2.5 粒子受自我與社會認知之影響圖

粒子群體最佳化演算法的優勢有以下兩點:

1. 理論簡單，無複雜的數學計算式。因此在實作方面，只需評估函數和參數與決定群體大小，使其自行搜尋最佳解即可。這樣一個簡單的理論基礎，對於程式設計者而言十分輕鬆，就可完成此演算法的撰寫。
2. 全體搜尋，並且同時考慮全域及區域最佳解，修正其速度和位置，使得個別粒子不容易陷入區域最佳解。

2.3 Nelder-Mead 單體法

Nelder-Mead 單體法，由 Spendley (Spendley, 1962) 以及 Box (Box, 1957)所提出，之後由 Nelder 和 Mead 於 1965 所改良。Nelder-Mead 單體法在搜尋時能藉由群體內的資訊往最佳解的方向移動，因此如果它選擇的路徑是正確的，便能將最佳解包圍起來，一旦包圍住便能達到快速收斂的效果。不過，如果遇到的問題區域最佳解太多的話，它很難找到真正的最佳解，且很容易陷入區域最佳解，這也是它最大的缺點(李漢祥，2002) 。

其原理是比較在各點經由運算逐漸向最佳點逼近的一種演算法。Nelder-Mead單體法在一開始時必須將各點代入評估函數，找出最佳、次佳、最差的目標函數後，經過反射、擴張、收縮、逼近，會產生新的點，在重複上述的步驟後，當各點十分逼近函數值時則到達收斂條件，演化過程便終止。

Nelder-Mead單體演算法實際演化步驟如下:

1. 定義Nelder-Mead單體法相關參數

   選定評估函數*(f)*，以及各參數設定，其中包括了:反射參數α、擴張參數γ、收縮參數β、逼近參數δ。

2. 初始產生與函數評估、排序

   若問題所需解決的維度為N維，則在N維的空間中，產生N+1個點。將其點帶入評估函數中，將得到N+1組解並依小到大排序。

3. 計算中心點

   依排序結果由第1至第N個資料點之總和並除以N，以公式(2.7)得到中心點P_{cent}位置。P_i即代表所有的資料點。

$$P_{cent} = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{2.7}$$

4. 反射 (Reflection)

   已排序後資料點，第1、第2以及第N+1，即P_{high}、$P_{sec\ hi}$和P_{low}，以公式(2.8)進行反射動作，如圖2.6(a)所示，反射參數α須大於0，一般設定為1。得到所反射的點P_{refl}之後，代入評估函數，若$f_{low} < f_{refl} < f_{sec\ hi}$，則$P_{refl}$取代$P_{high}$，並且再一次進入步驟4，若$f_{refl} < f_{low}$，則進行至步驟5，若$f_{refl} > f_{sec\ hi}$或$f_{refl} < f_{high}$則進入步驟6。

$$P_{refl} = (1+\alpha) - \alpha P_{high} \tag{2.8}$$

5. 擴張 (Expansion)

若經由反射動作之後，$f_{refl} < f_{low}$，則以 P_{refl} 和 P_{cent} 進行擴張，如圖2.6(b)所示，依照下列公式(2.9)所示，擴張參數 γ 必須大於1，一般設定為2，經過擴張動作後，便得到新的擴張點 P_{exp}，代入評估函數後，若 $f_{\exp} < f_{low}$ 則以 P_{exp} 取代 P_{high}，並判斷是否達到收斂條條件，若否，則 P_{refl} 取代 P_{high} 並回到步驟4。

$$P_{exp} = \gamma P_{refl} + (1-\gamma) P_{cent} \tag{2.9}$$

6. 收縮 (Contraction)

當 $f_{refl} > f_{\sec hi}$ 和 $f_{refl} \le f_{high}$ 時，則 P_{refl} 取代 P_{high} 並依公式(2.10)向外收縮，如圖2.6(C)所示，若 $f_{refl} > f_{high}$,依公式向內收縮時 P_{high} 不需要被 P_{refl} 取代，如圖2.6(d)所示。收縮參數 β 設定需介於0至1之間，依般設定為0.5。產生新的點 P_{cont} 之後，若 $f_{cont} \le f_{high}$ 則 P_{cont} 取代 P_{high} ，則回到步驟4重新運算，若 $f_{cont} > f_{high}$ 則進入步驟7。

$$P_{cont} = \beta P_{high} + (1-\beta) P_{cent} \tag{2.10}$$

7. 逼近 (Shrink)

若在步驟6時 $f_{cont} > f_{high}$ ，則進入此步驟，全部的點使用公式(2.11)，向 P_{low} 逼近，如圖2.6(e)(f)所示。逼近參數 δ 設定範圍為0至1之間，一般設定為0.5。逼近之後再代入評估函數，並且重複步驟4。

$$P_i \leftarrow \delta P_i + (1-\delta) P_{low} \;\; i = 1,2,...,N+1, \;\; i \neq low \tag{2.11}$$

8. 停止條件 (Stopping Criterion)

不斷重複步驟4至步驟7，直到變異縮小至下列公式(2.12)所示，即可達到實驗停止條件。

$$\left[\sum_{i=1}^{N+1}\frac{(f_i-\overline{f})^2}{N+1}\right]^{1/2}\le\theta \qquad (2.12)$$

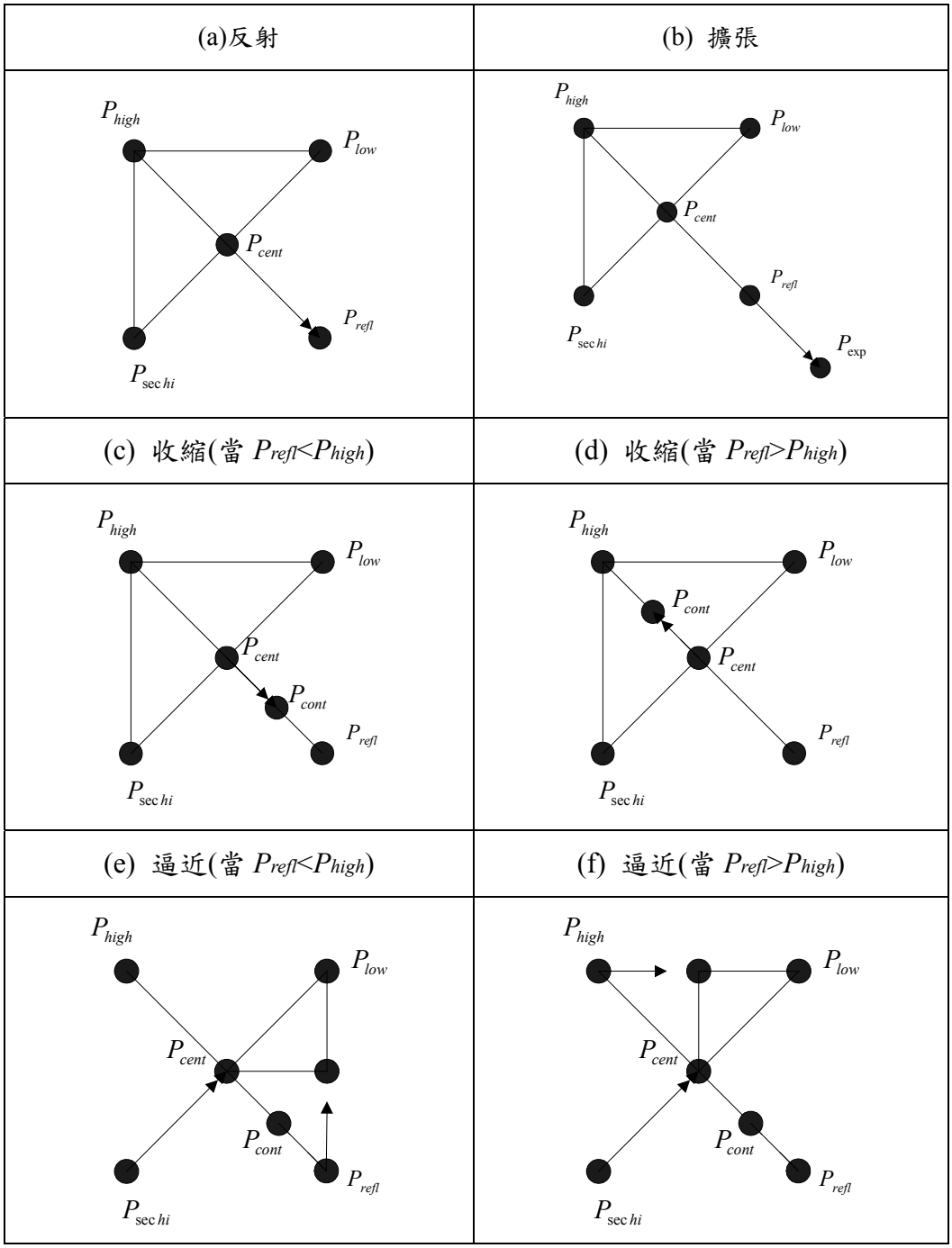


圖2.6 Nelder-Mead單體法運算流程圖

資料來源:何怡偉，2004

第三章 研究方法

本章節將探討NM-PSO演算法與整合KPSO分群法和KNM-PSO分群法，並將之應用於資料分群。

圖3.1則是比較當NM、PSO以及NM-PSO以方程式(3.1)進行測試，所得到收斂的結果。從圖中可知，NM-PSO收斂的速度明顯優於其它兩種方法。另外圖3.2是三種方法尋找最佳解之路徑表示圖，可發現NM-PSO從起始點至最佳解之路徑，幾乎為一條直線，代表以NM-PSO尋找最佳解時，陷入區域最佳解的機會較低(Fan *et al.*,2004)。因此由上述可知，NM-PSO不管是收斂速度、尋找最佳解，都是明顯優於其它演算法。也因此，本篇論文所使用之分群方法，鎖定為NM-PSO與KNM-PSO，也因為目前尚無任何文獻是使用NM-PSO與KNM-PSO進行資料分群，希望透過此篇論文之研究，能對分群技術有所貢獻。

$$GP(x_1,x_2)=[1+(x_1+x_2+1)^2\times(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2)]$$
$$\times[30+(2x_1-3x_2)^2\times(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2)] \quad (3.1)$$

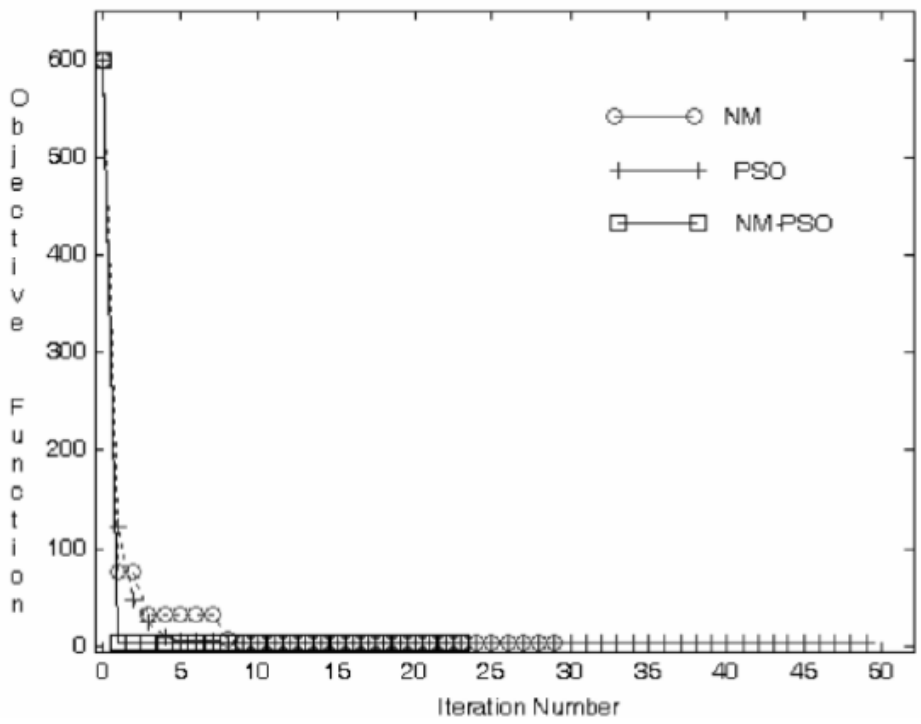


圖3.1 以方程式GP進行測試，三種方法之收斂速度表示圖

資料來源：Fan *et al*, 2004

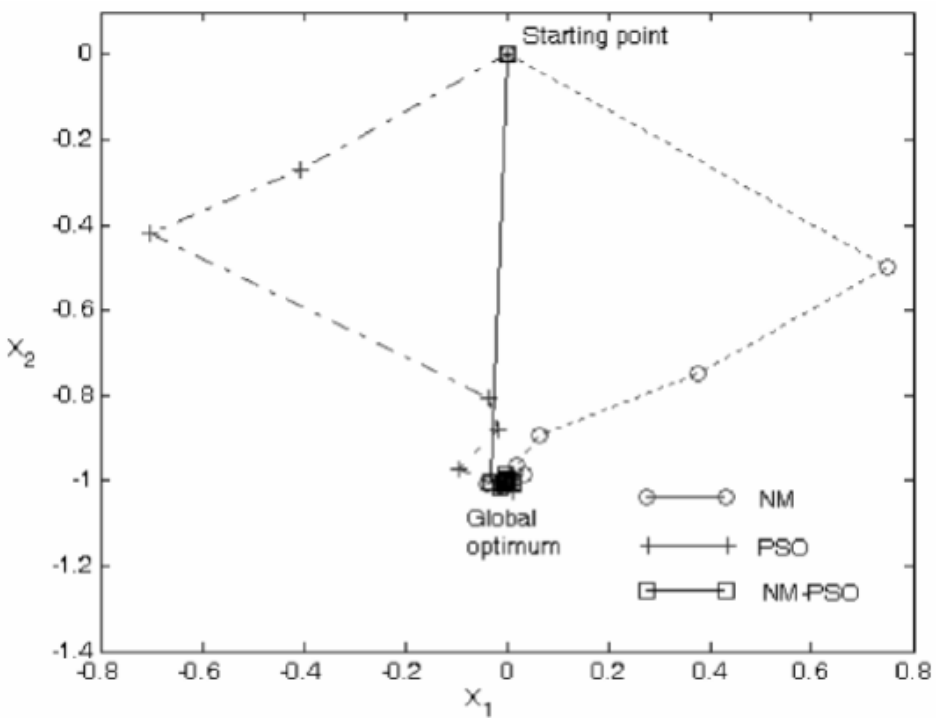


圖3.2 以方程式GP進行測試，三種方法尋找最佳解之路徑表示圖

資料來源：Fan *et al*, 2004

3.1 **NM-PSO演算法**

NM-PSO演算法於2004年，由何怡偉博士所提出(何怡偉，2004)，其概念為整合啟發式與傳統式演算法，因為Nelder-Mead單體法在搜尋時，速度雖然快，但易陷入區域最佳解。而PSO演算法雖然較不易陷入區域最佳解，但需要較大的群體數目，因此降低了速度。結合了兩種演算法而成的NM-PSO演算法，也同時結合了其優點，不僅計算速度快，也能正確的找到最佳解。

以下將清楚的介紹NM-PSO的演算流程。假定目前所需處理的問題為N維度，先產生3N+1組群體，並以優劣排序後，依序可分為N個、第N+1個以及2N個，首先將最佳的N個予以保留，再以NM演算法，以N個和第N+1個加以計算後，可得到更新過後的第N+1個，並將此結果保留。接著將保留的N個、第N+1個以及初始排序較差的2N個，以PSO演算法加以運算，但是運算結果並不更新已保留的N個以及第N+1，只更新較差的2N個。重覆此演算流程，直到滿足演算停止條件。

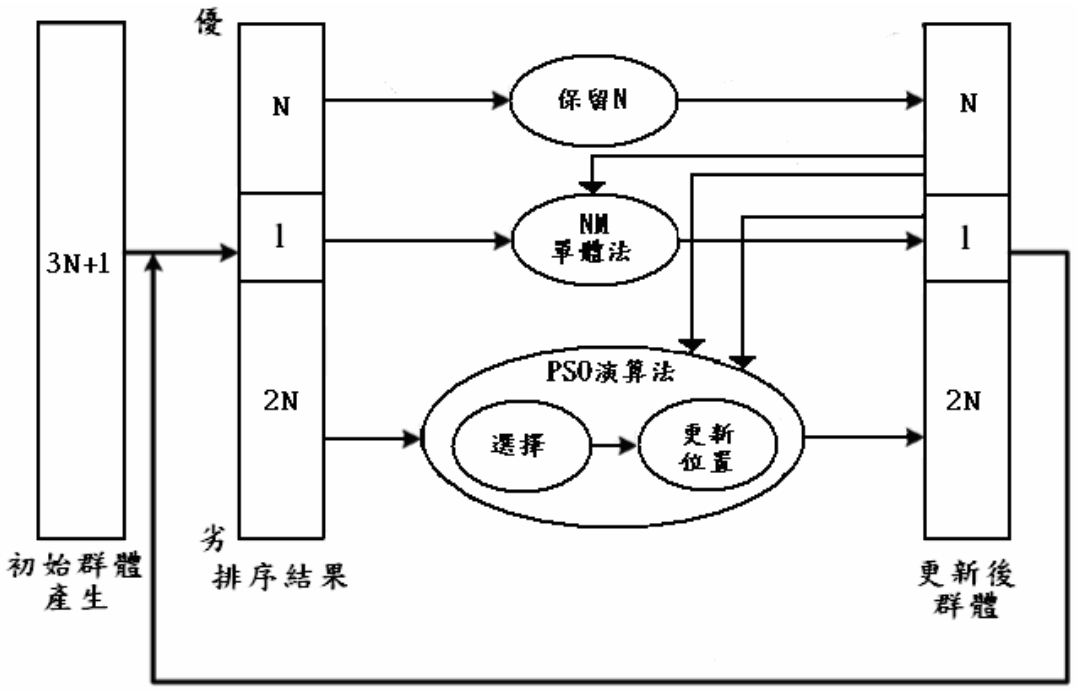


圖3.3 NM-PSO演算流程圖

資料來源： Fan et al, 2004

NM-PSO演算法實際實驗流程如以下步驟所示:

1. 定義實驗

   定義實驗參數，選定評估函數*(f)*，設定實驗參數以及實驗停止條件。

2. 初始產生

   若實驗目標函數維度為N，依實驗條件，則產生3N+1維群體候選解。

3. 排序

   將各個群體代入評估函數中，並以優劣加以排序，分別是N個、第N+1個以及2N個。

4. 保留較佳解

   評估函數後，將結果較佳的N個予以保留至已更新的群體，等待與其他較差的解進行更新。

5. 使用NM演算法進行更新

   使用NM演算法，將已保留的N個與第N+1個進行更新，其結果將取代第N+1，並予以儲存保留至更新群體。

6. 使用PSO演算法進行更新

將已放置更新群體的與尚未更新且排名較差的2N個以PSO演算法進行更新，但已更新過的N個和第N+1個群體，不會被更新，只有尚未更新的2N個會受更新改變，並將更新後的2N亦儲存至更新群體。

7. 停止或繼續運算

如此便完成了一次NM-PSO之演算程序，並檢視是否達到實驗停止條件，若滿足條件，則停止實驗，否則回到步驟3繼續演算，一般而言，實驗停止條件是以迭代次數或評估函數值不再改變或收斂為依據。

由上述可知，NM-PSO演算法便是結合兩種演算法的模式加以運算，並只需3N+1的群體大小便能運算，NM-PSO演算法提出時間為2004年，目前仍處於發展期，相信未來會有更多學者引用此方法於各種領域。

3.2 NM-PSO於資料分群的架構

本論文以NM-PSO運用於資料分群，下列將說明，如何以NM-PSO所提出方法進行實驗。

1. 參數設定

由於NM-PSO是結合了兩種演算法而成，因此在參數設定方面必須同時考慮，通常是以NM及PSO一般參數設定為主。

2. 設定評估函數

對一般分群技術而言，評估分群的績效指標，通常是以各群體中心與同一群的資料點距離加總以及分群誤判率。分群誤判率，即是錯判資料點於錯誤群體的比率。而當距離總和越小時代表分群越成功，而此距離一般而言對於誤判率也有相當大的影響。公式(3.2)為分群評估函數F，z代表分群中心點，共有K個，x則是待分群資料點，共有n個。

$$F = \sum \left\| x_j - z_i \right\|,\ i = 1,...,K,\ j = 1,...,n \tag{3.2}$$

3. 族群初始化

首先，必須判斷欲分群之資料庫之維度，並決定分群數目，維度為d，分群數目為K，N則為K與d相乘之結果，如下列(3.3)所示，並且產生出始群體大小則為3N+1。

$$N = K \times d \tag{3.3}$$

4. 進行實驗與停止條件

實驗所有相關參數以及資訊獲得後，便依照NM-PSO演算法之流程進行實驗。一般來說，分群是以迭代次數停止來做為實驗停止條件，當然也可以以到達評估函數值來決定實驗是否停止。

3.3 KPSO與KNM-PSO分群法

整合K-Means分群法、PSO演算法與NM-PSO演算法成而的KPSO分群法和KNM-PSO分群法。目的是將K-Means分群法，快速收斂的優點充分利用，而易陷入區域最佳解的缺點，則使用NM-PSO演算法與PSO演算法加以補強。其整合理論相當簡單，如圖3.4所示首先使用K-Means分群法加以分群，當所得結果收斂時，再使用PSO演算法或NM-PSO演算法，尋找出最佳解，若K-Means所得結果陷入區域最佳解，則PSO與NM-PSO演算法，將跳出區域最佳解，另尋全域最佳解，若在K-Means所得結果，介於全域最佳解之區域內，則PSO與NM-PSO演算法，將在範圍內，找尋全域最佳解。也就是因為有此優點，因此期望距離結果與收斂狀況能優於NM-PSO與PSO。

表3.1將比較本論文使用之五種分群技術之優缺點。從表得知，NM-PSO與，雖然其理論基礎較為複雜，但是NM-PSO之結果與速度優於PSO與K-Means，而KNM-PSO則都優於其他四種方法。

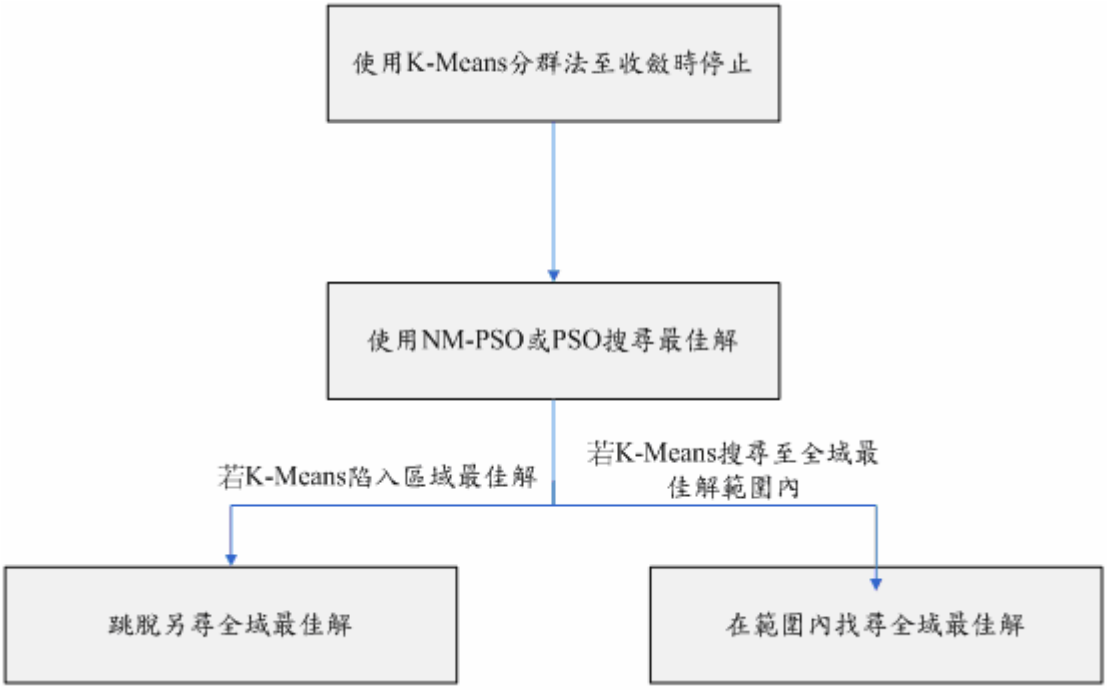


圖3.4 KNM-PSO與KPSO演算流程圖

表3.1 比較實驗使用之五種分群技術之優缺點

	優點	缺點
K-Means	1.理論簡單，撰寫容易 2.執行速度快	1.無法解決多維度問題 2.無法處理資料點重複問題
PSO	1.較不易陷入區域最佳解 2.理論簡單，撰寫容易	1.運算時間較長 2.所需群體數量較大
NM-PSO	1.需要群體較小 2.求得全域最佳解機率較大 3.執行速度快	1.理論基礎較複雜
KPSO	1.執行速度比 PSO 演算法快 2.結果優於 PSO 演算法	1.仍無法解決區域最佳解問題 2.執行速度仍比 NM-PSO 與 KNM-PSO 慢
KNM-PSO	1.執行速度為最快 2.求得全域最佳解機率最大 3.只需與 NM-PSO 相同群體大小	1.理論基礎較複雜

第四章 實驗結果與分析

4.1 實驗資料與參數設計

本節為實驗資料的介紹，以及相關參數設定。本實驗資料庫分別是:人工資料庫和實際資料庫兩種。以人工資料庫作為分群對象的原因是可以按照所需產生不同大小維度的資料庫，以測試分群效果。而對於實際資料庫進行分群，則是證明對於實際且大量的資料可否成功分群。總之，以兩種資料庫來進行，最主要就是要驗證KNM-PSO與NM-PSO分群法，是否在不同的條件下，分群的效果都能優於其他三種分群方法。

透過以表4.1，可以清楚的了解有關此實驗之相關條件。迭代次數以10N次為基準，而實驗總次數為20次，上述所指的是針對一個資料庫，進行總和20次的分群，所得的結果再加以平均，是最客觀的數據。而評估分群之績效指標，是採用計算各群體中心與資料的距離總和，以及分群誤判機率為評估方法。

表4.1 實驗設定

	各項設定
迭代次數	10N
實驗總次數	20
群體大小	3N+1
評估指標	各群總距離、誤判率

4.1.1 人工資料

人工資料方面，分別產生屬性為2維吉3維的測試資料，同時資料的隨機產生的方式，是使用均勻分配。共產生總數400筆與600筆的資料，將其整理如表4.2所示。

表4.2 二組人工資料庫相關資料

	維度	資料點數	群體數
人工資料庫 1	2	600	4
人工資料庫 2	3	250	5

人工資料庫1:

以下列方程式產生2維、600個資料點，4群資料，以下列方程式(4.1)以產生。資料分佈狀況如圖4.1所示。

$$\mu = \begin{pmatrix} m_i \\ 0 \end{pmatrix}, \sum = \begin{bmatrix} 0.50 & 0.05 \\ 0.05 & 0.50 \end{bmatrix} \tag{4.1}$$

$$i = 1,...,4 \qquad m_1 = -3,\ \ m_2 = 0,\ \ m_3 = 3 \ and \ m_4 = 6$$

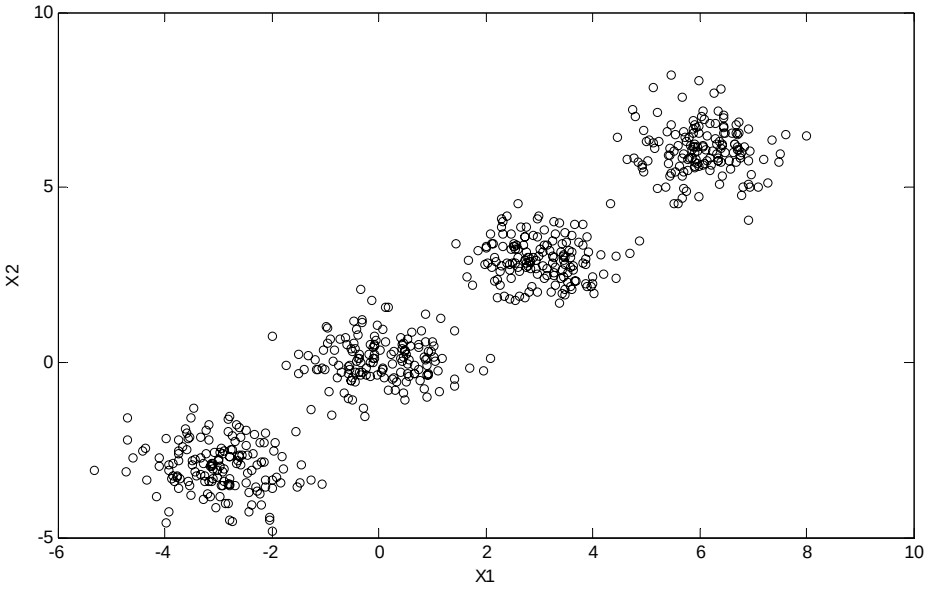


圖4.1 人工資料庫1資料分佈圖

人工資料庫2:

產生3維、250個資料點，5群資料，以下列方程式(4.1)以產生。資料分佈狀況如圖4.2所示。

$$z_i \begin{cases} 1 & Uniform(85, 100) \\ 2 & Uniform(70, 85) \\ 3 & Uniform\ (55, 70) \\ 4 & Uniform(40, 55) \\ 5 & Unifiorm(25, 40) \end{cases} \tag{4.2}$$

$$z_1, z_2, z_3, z_4, z_5 \in (25, 100)$$

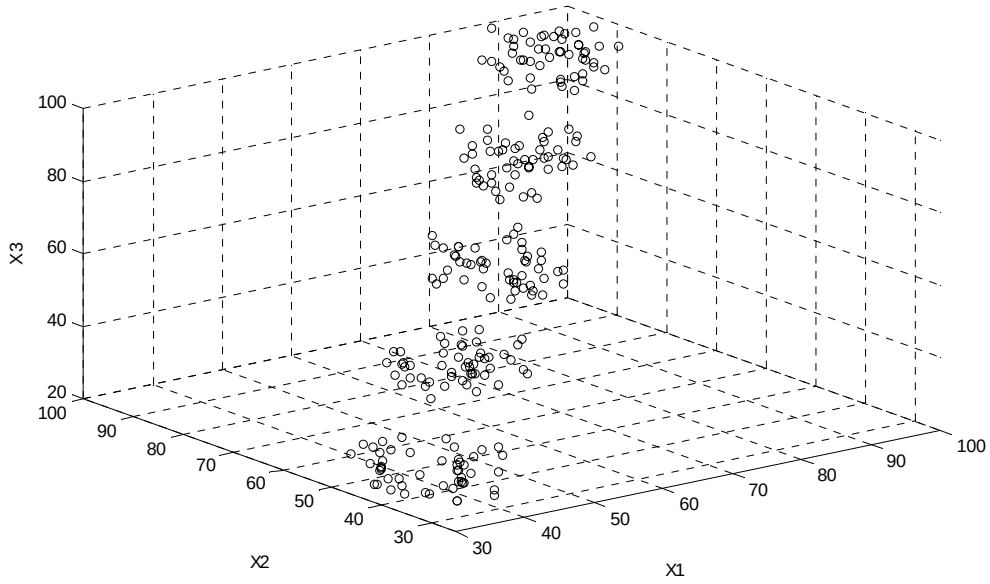


圖4.2 人工資料庫2資料分佈圖

4.1.2 實際資料

實際資料庫方面本研究採用七組由美國加州Irvine 大學資訊與電腦科學系中所提供的測試資料。資料庫名稱分別是蝴蝶花(Iris plants)、葡萄(Wine)、乳癌(Breast Cancer)、玻璃(Glass) 、避孕器(Contraceptive Method Choice，簡稱Cmc)、母音(Vowel)與天然石油(Crude Oil)，以下將介紹這七個資料庫的相關資訊。表4.3是有關這七個實際資料庫的相關資訊。

表4.3 七組實際資料庫相關資料

	Iris	Wine	Cancer	Glass	Cmc	Vowel	Oil
資料數目	150	178	683	214	1473	871	56
維度	4	13	9	9	9	3	5
群體數	3	3	2	7	3	6	3

1. 鳶尾植物(Iris Plants)

鳶尾植物資料庫共有150筆資料，記錄中包含了萼片(Sepal) 與花瓣(Petal)的長度(Length)、寬度(Width)4種屬性， Iris Setosa、Versicolour與Virginica這3種鳶尾花的種類所組成的。詳細資料如表表4.4與表4.5所示。

表4.4 Iris Plants資料庫中4種屬性之分佈

屬性	最小	最大	平均值
Sepal Length	4.3	7.9	5.84
Sepal Width	2.0	4.4	3.05
Setal Length	1.0	6.9	3.76
Setal Width	0.1	2.5	1.20

表4.5 Iris Plants資料庫中3群資料之大小與比例

群體	所佔資料比數	所佔資料比例
Iris Setosa	50	33.33%
Versicolour	50	33.33%
Virginica	50	33.33%

2. 葡萄酒(Wine)

葡萄酒資料庫共有178筆資料，其中包含13種屬性，總共的葡萄酒種類有3種。詳細資料如表4.6與表4.7所示。

表4.6 Wine資料庫中13種屬性之分佈

屬性	最小	最大	平均值
Alcohol	11.03	14.83	1.9382
Malic acid	0.74	5.8	13.001
Ash	1.36	3.23	2.3363
Alcalinity of ash	10.6	30	19.495
Magnesium	70	162	99.742
Total phenols	0.98	3.88	2.2951
Flavanoids	0.34	5.08	2.0293
Nonflavanoid phenols	0.13	0.66	0.36185
Proanthocyanins	0.41	3.58	1.5909
Color intensity	1.28	13	5.0581
Hue	0.48	1.71	0.95745
OD280/OD315 of diluted wines	1.27	4	2.6117
Proline	278	1680	746.89

表4.7 Wine資料庫中3群資料之大小與比例

群體	所佔資料比數	所佔資料比例
群體1	59	33.15%
群體2	71	39.89%
群體3	48	26.96%

3. 乳癌(Breast Cancer)

乳癌資料庫共有699筆資料，扣除掉遺失的資料，共整理出683筆資料，其中包含9種屬性，是由良性細胞以(Benign)及惡性細胞(Malignant)兩種所組成的。詳細資料如表4.8與表4.9所示。

表4.8 Breast Cancer資料庫中9種屬性之分佈

屬性	最小	最大	平均值
Clump Thickness	1	10	4.42
Uniformity of Cell Size	1	10	3.15
Uniformity of Cell Shape	1	10	3.21
Marginal Adhesion	1	10	2.83
Single Epithelial Cell Size	1	10	3.23
Bare Nuclei	1	10	3.54
Bland Chromatin	1	10	3.44
Normal Nucleoli	1	10	2.86
Mitoses	1	10	1.60

表4.9 Breast Cancer資料庫中2群資料之大小與比例

群體	所佔資料比數	所佔資料比例
Benign	444	65.01%
Malignant	239	34.99%

4. 玻璃(Glass)

Glass資料庫共有214筆資料，記錄玻璃所包含的9種化學元素，資料庫由7種不同的玻璃組成。沒有任何的遺失資料詳，細資料如表4.10與表4.11所示。

表4.10 Glass資料庫中9種屬性之分佈

屬性	最小	最大	平均值
RI	1.51	1.5339	1.51
Na	10.73	17.38	13.40
Mg	0	4.49	2.68
Al	0.29	3.5	1.44
Si	69.81	75.41	72.65
K	0	6.21	0.49
Ca	5.43	16.19	8.95
Ba	0	3.51	0.175
Fe	0	0.51	0.57

表4.11 Glass資料庫中7群資料之大小與比例

群體	所佔資料比數	所佔資料比例
Float Building Windows	70	32.71%
Float Vehicle Windows	17	7.94%
Non-float Building Bindows	76	35.5%
Non-float Vehicle Windows	0	0%
Containers	13	6.07%
Tableware	9	4.20%
Headlamps	29	13.55%

5. 避孕器(Cmc)

乳癌資料庫共有1473筆資料，其中包含9種屬性，經由分群可將資料歸納為不使用(No-use)、長期使用(Long-term)以及短期使用(Short-term)。詳細資料如表4.12與表4.13所示。

表4.12 Cmc資料庫中9種屬性之分佈

屬性	最小	最大	平均值
Wife's age	16	49	32.5384
Wife's education	1	4	2.9586
Husband's education	1	4	3.4297
Number of children ever born	0	16	3.2614
Wife's religion	0	1	0.8506
Wife's now working	0	1	0.7495
Husband's occupation	1	4	2.1378

Standard-of-living index	1	4	3.1337
Media exposure	0	1	0.0740

表4.13 Cmc資料庫中3群資料之大小與比例

群體	所佔資料比數	所佔資料比例
No-use	629	42.7%
Long-term	333	22.61%
Short-term	511	34.69%

6. 母音(Vowel)

母音資料庫是由871筆印地安語言之母音資料所組成的，其中包含3種音頻的屬性，將分類為{δ, a, i, u, e, o}六種母音。詳細資料如表4.14與表4.15所示。

表4.14 Vowel資料庫中3種屬性之分佈

屬性	最小	最大	平均值
F1	250	900	470.4822
F2	700	2550	1514.6842
F3	1800	3200	2561.0218

表4.15 Vowel資料庫中6群資料之大小與比例

群體	所佔資料比數	所佔資料比例
δ	72	8.27%
a	89	10.22%
i	172	19.57%
u	151	17.34%
e	207	23.77%
o	180	20.67%

7. 天然石油(Crude Oil)

天然石油資料庫是共有56筆資料，每筆資料中皆有5種屬性，分別是釩(Vanadium)、鐵(Iron)、鈹(Beryllium)、飽和碳氫化合物(Saturated Hydrocarbons)與芳香族碳氫化合物(Aromatic Hydrocarbons)。將可分成三群。詳細資料如表4.16與表4.17所示。

表4.16 Crude Oil資料庫中5種屬性之分佈

屬性	最小	最大	平均值
Vanadium	1.2	11	6.1804
Iron	5.6	52	27.0464
Beryllium	0	1.5	0.3414
Saturated Hydrocarbons	3.06	9.25	5.2911
Aromatic Hydrocarbons	2.22	13.01	6.4336

表4.17 Crude Oil資料庫中3群資料之大小與比例

群體	所佔資料比數	所佔資料比例
Wilhelm	7	12.5%
Sub-Mulinia	10	17.86%
Upper	39	69.64%

4.2 實驗結果

實驗結果方面，將呈現K-Means、PSO、NM-PSO、KPSO與KNM-PSO五種分群之分群成果。本論文將以距離、收斂速度、誤差率以及評估函數運算次數，作為結果之比較。

4.2.1 人工資料庫實驗結果

人工資料庫1:

表4.18呈現出人工資料庫1之分群結果，KNM-PSO與KPSO，不論是在最佳距離值或是平均距離值，與其他三種方法相比，都達到最佳的結果，另外也發現了，以KPSO分群亦有不錯的結果。

圖4.3則是收斂狀況，由此可知，最快速收斂的是KNM-PSO與NM-PSO，其次是K-Means，但K-Means雖然在迭代次數不到20就收斂了，其結果卻是較差的。而比較NM-PSO與PSO，可以輕易的發現，NM-PSO約在迭代次數50次就收斂了，而PSO則是到最後還沒有完全收斂，並且結果也比NM-PSO差。

表4.18 人工資料庫1之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	515.883	515.883	515.883	515.928	516.039
平均距離值	515.883	515.883	515.883	627.743	721.567
距離值標準差	7.14E-08	5.6E-05	7.14E-08	180.235	295.843
最佳距離值(排名)	1	3	1	4	5
平均距離值(排名)	1	3	1	4	5

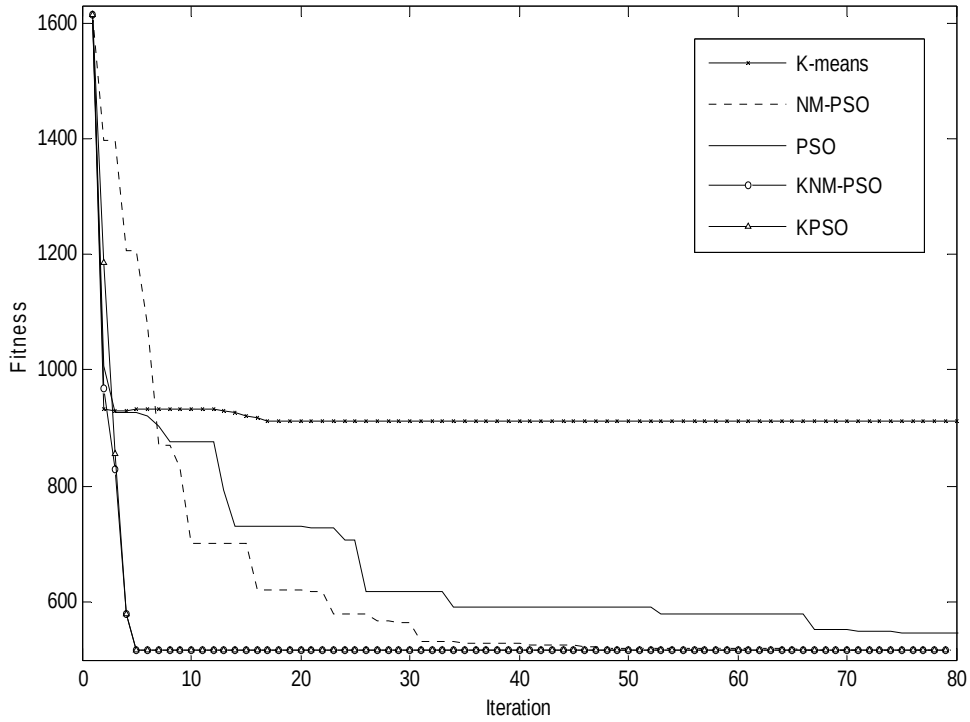


圖4.3 人工資料庫1之分群收斂圖

人工資料庫2:

表4.19呈現出人工資料庫2之分群結果，由此可知，以最佳距離值而言，只有K-Means較差，而其他四種方法結果皆相同。但若以平均值而論，KNM-PSO結果為最佳，依序是NM-PSO、KPSO、PSO與K-Means。

圖4.4為五種方法的收斂狀況，最快收斂為K-Means，依序為KNM-PSO與KPSO、NM-PSO、PSO。從中可知KNM-PSO與NM-PSO收斂速度明顯優於PSO，並且平均距離值亦優於其他三種分群技術。

表4.19 人工資料庫2之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	1743.2	1743.2	1743.2	1743.2	1746.9
平均距離值	1746.9	2067.3	1910.4	2517.2	2762
距離值標準差	3.598	343.643	296.221	415.022	720.659
最佳距離值(排名)	1	1	1	1	1
平均距離值(排名)	1	3	2	4	5

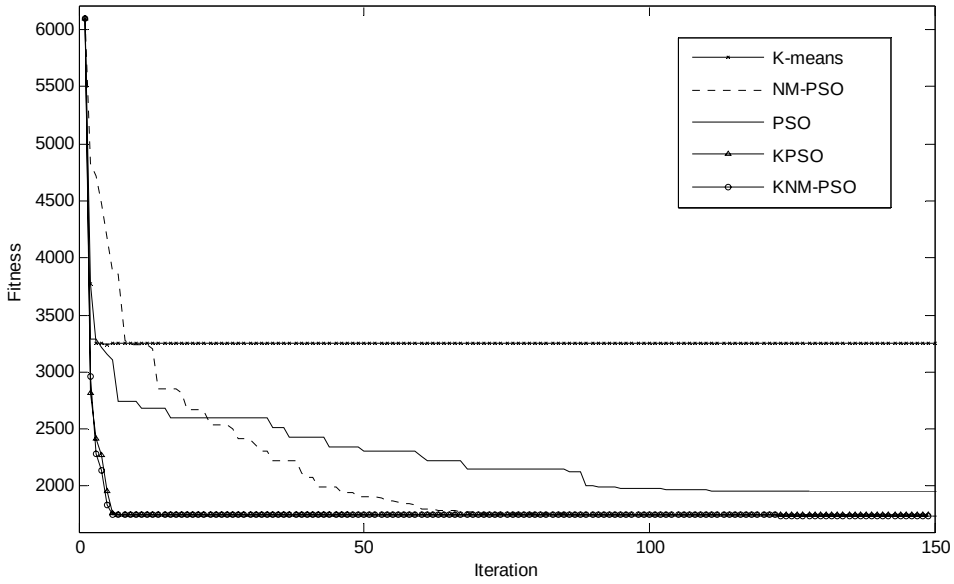


圖4.4 人工資料庫2之分群收斂圖

計算評估函數總數:

計算評估函數總數之比較，由表4.20可知，K-Means雖然十分快速，但分群結果卻不佳，也驗證了文獻探討中所提及，K-Means雖然速度較快，但結果卻往往不穩定。若以其他四種方法而論，KNM-PSO所需計算次數最少，而NM-PSO也與KNM-PSO相差不多，而KPSO與PSO則是需要較多次數的評估函數計算。由這些相關數據可知，KNM-PSO與NM-PSO在運算上的確可降低運算的時間，而以距離為評估函數的結果，KNM-PSO優於KPSO，而NM-PSO亦優於PSO與K-Means。

表4.20 對人工資料庫進行分群之計算評估函數總數表

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
計算評估函數總數(人工資料庫1)	1996	2976	2265	3240	80
計算評估函數總數(人工資料庫2)	7051	10881	7392	11325	150

誤差率與分群圖示:

五種方法之分群誤差率，如表4.21所示，對於資料庫1之分群，除了PSO與K-Means之外，其他三種方法如圖4.5所示，分群結果完全正確，圖4.6則是PSO演算法之分群結果，從中可知，分群結果並不完全正確，而圖4.7則是K-Means分群成果，只將資料點分為三群，因此結果誤差甚多。

在資料庫2中，KNM-PSO分群結果也是完全正確(圖4.8)，其次是NM-PSO(圖4.9) 、KNM-PSO(圖4.10)、PSO(圖4.11)，最差則是K-Means(圖4.12)。從資料庫1與資料庫2的測試中可得知，KNM-PSO分群結果，不論是距離或是誤差率的表現都是最佳的，NM-PSO所測試的誤差率則優於KPSO與PSO，而K-Means對於這兩個人工資料庫，雖然能快速的分群，但卻時常無法分成正確群體數，而造成巨大的誤差。因此可證明，KNM-PSO在以人工資料庫進行分群，不論是誤差率、計算評估函數總數、收斂速度與距離總值的表現，都是最佳的。

表4.21 對人工資料庫進行分群之誤差率列表

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳誤差率 (人工資料庫1)	0.00%	0.00%	0.00%	0.00%	0%
平均誤差率 (人工資料庫1)	0.00%	0.00%	0.00%	7.57%	13%
誤差率標準差 (人工資料庫1)	0	0	0	12.18	17.78
最佳誤差率 (人工資料庫2)	0%	0%	0%	0%	20%
平均誤差率 (人工資料庫2)	0%	10%	4.04%	22%	34%
誤差率標準差 (人工資料庫2)	0	10.32	8.52	11.35	13.45

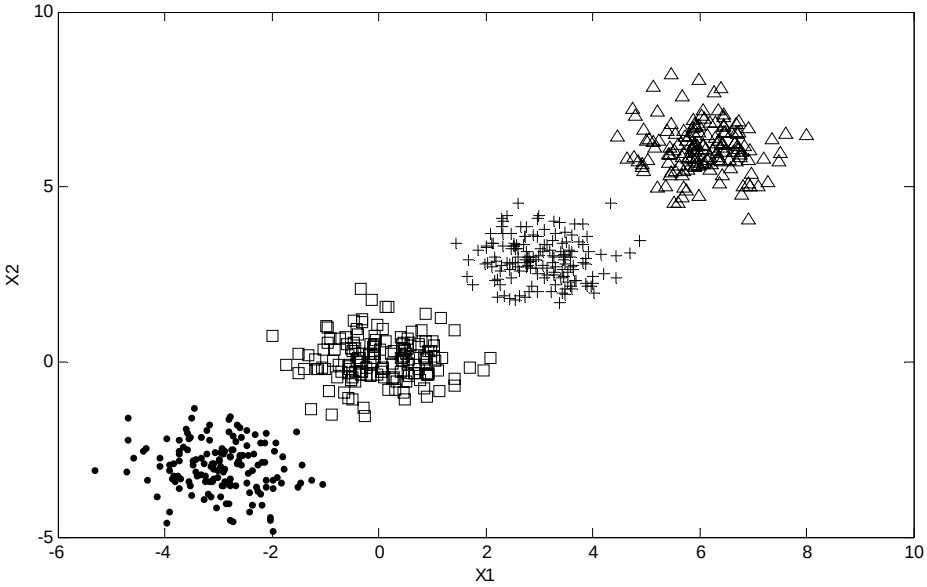


圖4.5 KNM-PSO、KPSO、NM-PSO對於人工資料1之分群結果圖

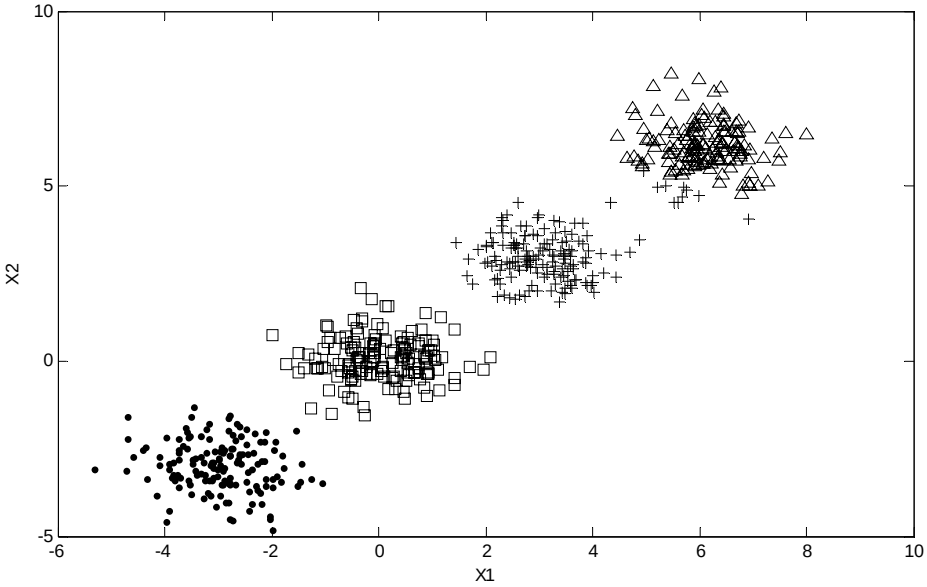


圖4.6 PSO對於人工資料1之分群結果圖

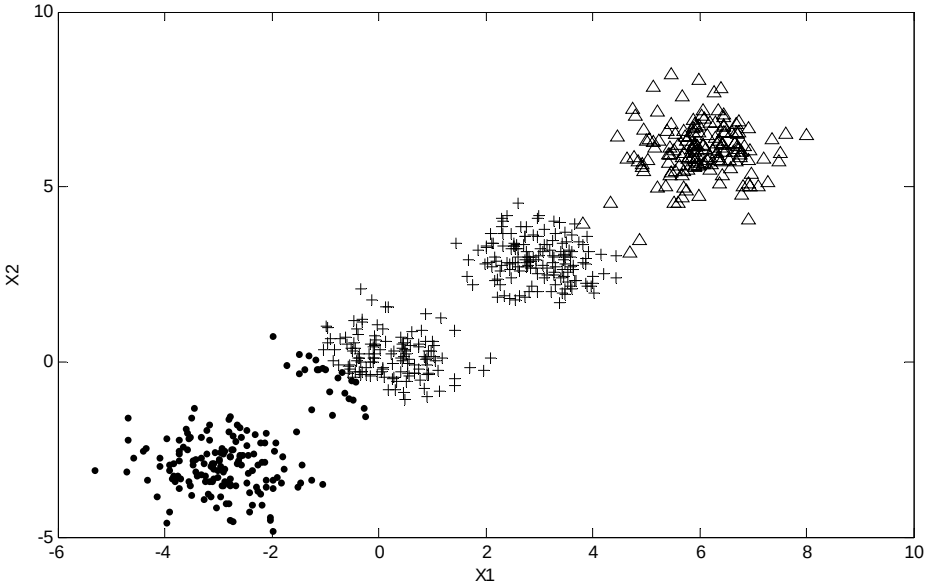


圖4.7 K-Means對於人工資料1之分群結果圖

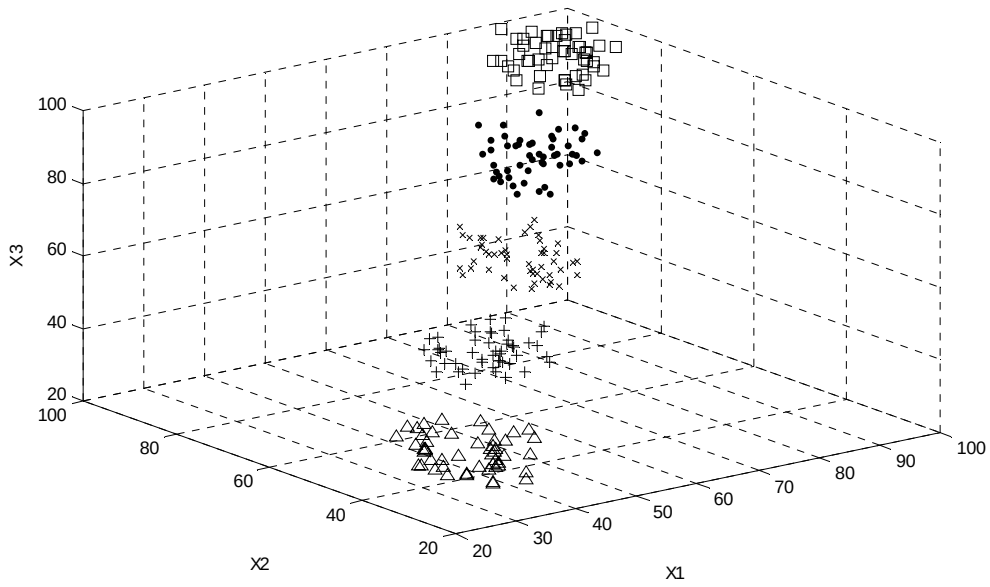


圖4.8 KNM-PSO對於人工資料2之分群結果圖

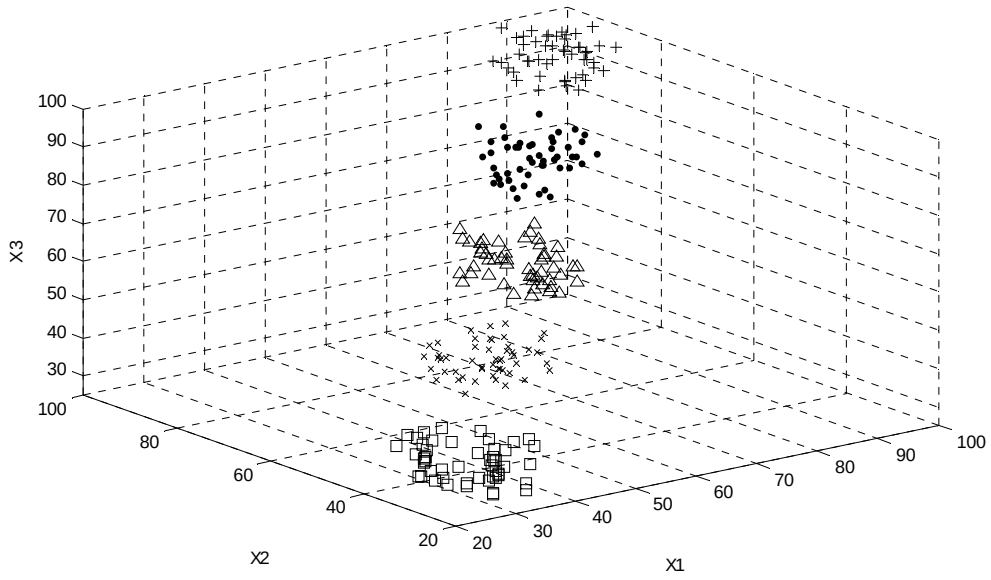


圖4.9 NM-PSO對於人工資料2之分群結果圖

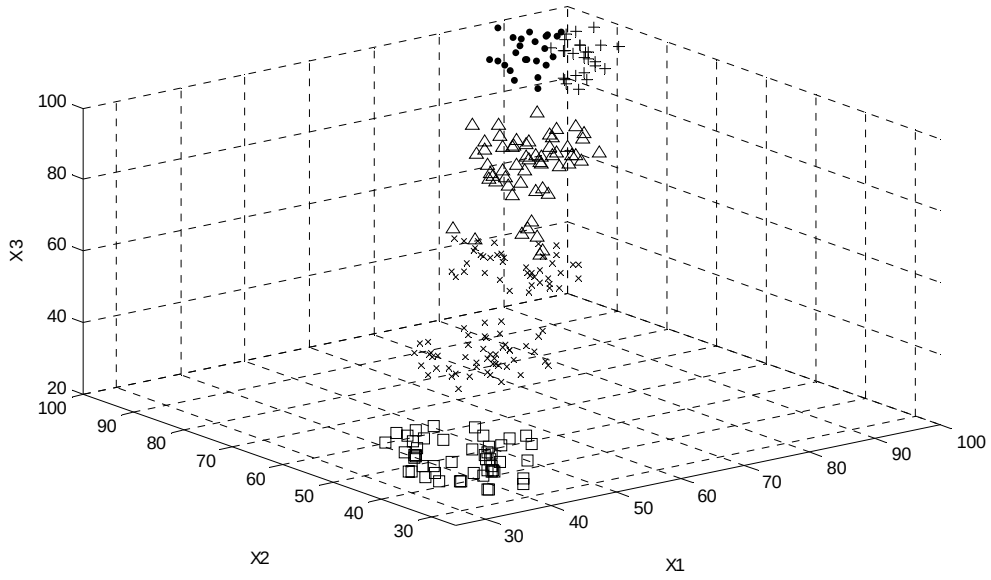


圖4.10 KPSO對於人工資料2之分群結果圖

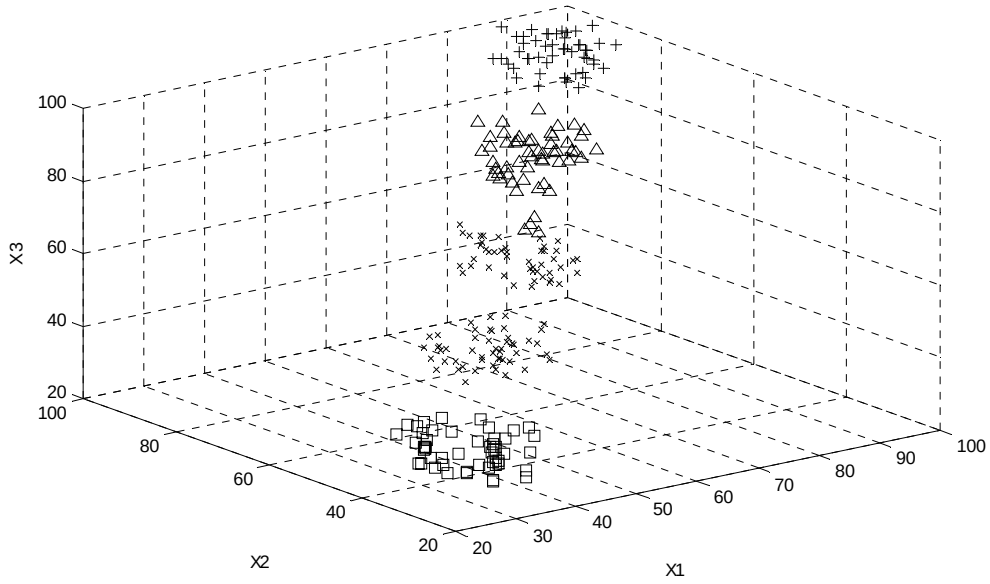


圖4.11 PSO對於人工資料2之分群結果圖

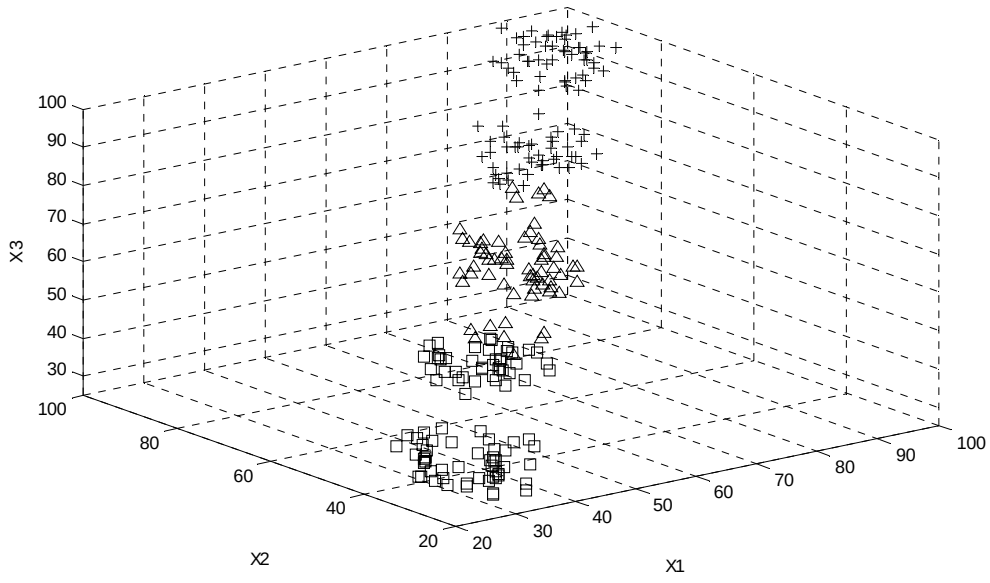


圖4.12 K-Means對於人工資料2之分群結果圖

4.2.2 實際資料庫實驗結果

Iris Plants資料庫:

表4.22為五種方法對Iris Plants資料庫進行分群之距離結果，發現除了K-Means之外，其他四種方法，得到最佳距離值都相同，但若以平均距離值而論，就有明顯的差異，KNM-PSO表現最佳，其次是KPSO，而NM-PSO又明顯的表現比PSO與K-Means好。

圖4.13為對Iris資料庫進行分群之收斂圖，K-Means在迭代是數10次之內便收斂了KNM-PSO與KPSO收斂狀況差異不大，值得注意的是NM-PSO約為迭代次數50次左右，就已初步收斂。而PSO則到最後都尚未收斂。因此可證明，NM-PSO在對Iris資料庫進行分群時，不論是在收斂數值與迭代次數都明顯優PSO。

表4.22 Iris Plants資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	96.6555	96.6555	96.6555	96.6555	97.3259
平均距離值	96.6668	96.7567	100.7245	103.514	106.047
距離值標準差	0.008	0.0716	5.8154	9.6893	14.1177
最佳距離值(排名)	1	1	1	1	5
平均距離值(排名)	1	2	3	4	5

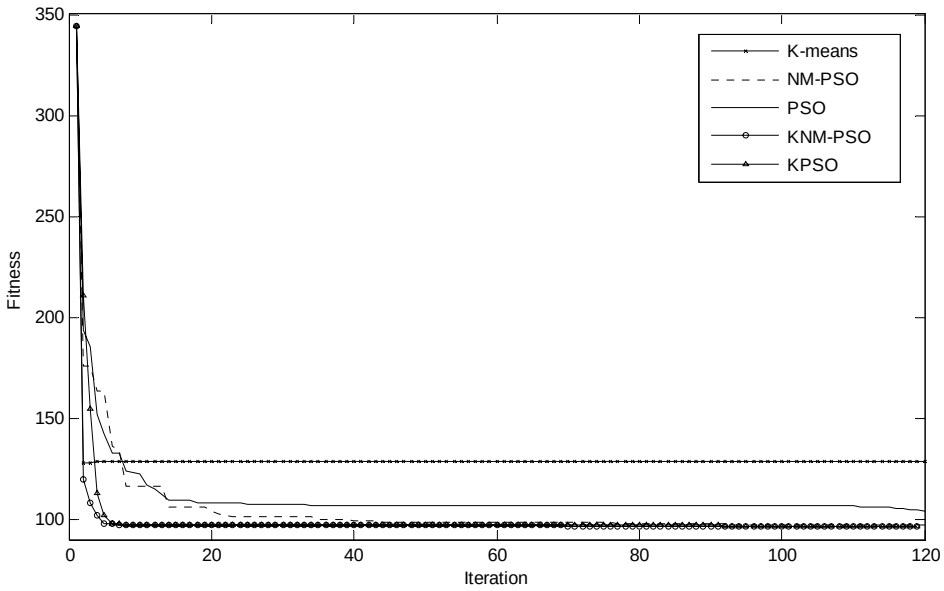


圖4.13 Iris資料庫之分群收斂圖

Wine資料庫:

由表4.23可知KNM-PSO、KPSO與NM-PSO在Wine資料庫中，最佳距離值所得到的結果皆相同，PSO與其落差不大，但K-Means就顯示出相當的缺失。在平均值方面，可知排名結果與Iris資料庫相同，但值得注意的是K-Means在這方面更明顯的劣於其他四種分群法，歸咎原因，可從資料庫型態中了解，因為此資料庫維度較大，型態較複雜，也證明了文獻中所提及的，K-Means無法對維度大的資料庫，進行有效的分群。

由圖4.14可知五種分群方法的收斂狀況，KPSO與K-Means收斂似乎速度相當快速，但是結果卻比都KNM-PSO來的差，KNM-PSO在K-Means初步收斂後，仍然不斷的尋找最佳解。而NM-PSO與KNM-PSO的狀況也相同，PSO雖然收斂的比NM-PSO快，但找到的解卻不是最佳的，因此從中可發現，NM-PSO對於跳出區域最佳解搜尋的能力優於其他分群方法。

表4.23 Wine資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	16292	16292	16292	16294	16555.68
平均距離值	16293	16294	16303	16311	18061
距離值標準差	0.4634	1.6981	4.2755	22.9823	793.2142
最佳距離值(排名)	1	1	1	4	5
平均距離值(排名)	1	2	3	4	5

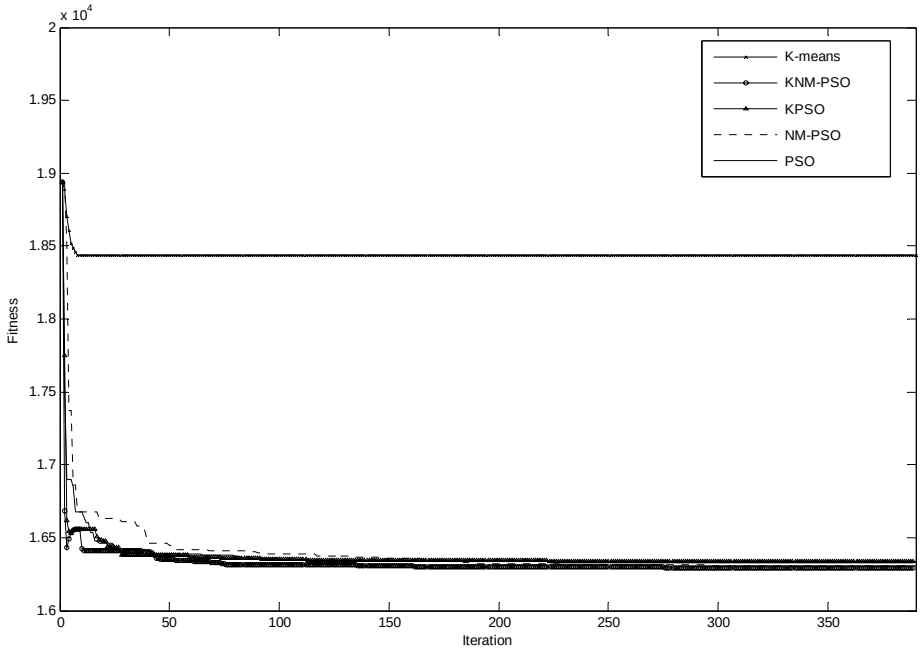


圖4.14 Wine資料庫之分群收斂圖

Breast Cancer資料庫:

在表4.24中顯示，最佳距離值的排名與其他資料庫實驗的結果，並沒有太大的差異，值得注意的是，K-Means在平均距離值方面所呈現的結果，優於PSO。其原因是，此資料庫只需將資料點分為兩群，屬於維度較低的型態，因此K-Means會有較好的表現，但仍比NM-PSO差。

收斂狀況而言，圖4.15顯示出，KNM-PSO、KPSO與K-Means，收斂速度約略相同，並可發現NM-PSO的收斂速度與數值則明顯優於PSO。

表4.24 Breast Cancer資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	2964.5	2964.5	2965.585	2976.3	2987
平均距離值	2964.7	2965.8	2977.7	3334.6	2988.3
距離值標準差	0.1499	1.6254	13.7339	357.659	0.4637
最佳距離值(排名)	1	1	3	4	5
平均距離值(排名)	1	2	3	5	4

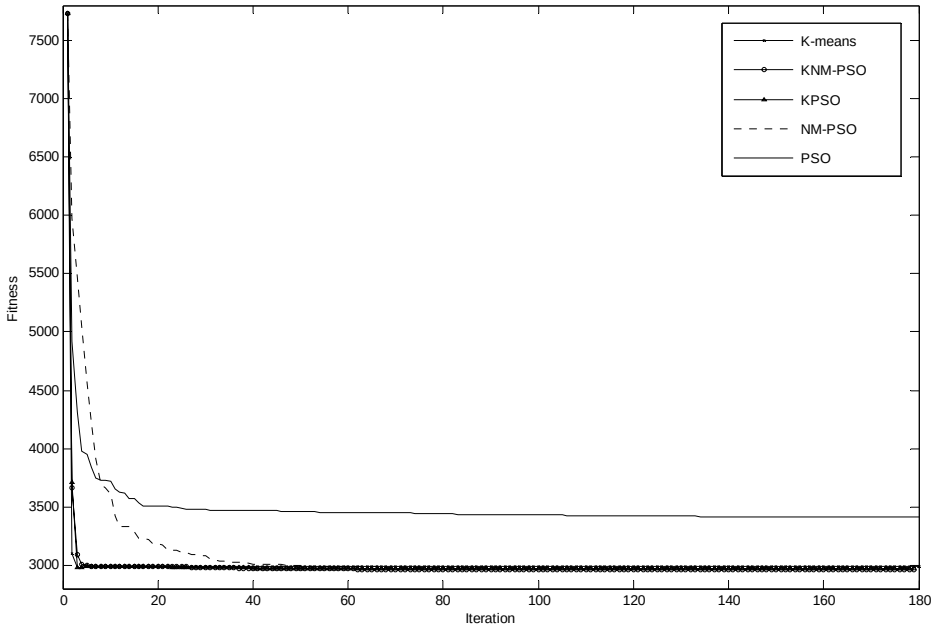


圖4.15 Breast Cancer資料庫之分群收斂圖

Glass資料庫:

Glass是此實驗中，維度最大的資料庫，KNM-PSO分群仍為最佳，而最佳距離值，K-Means則優於NM-PSO，但是平均距離值，NM-PSO仍然比K-Means與PSO來的好。

在收斂狀況方面，圖4.16中可發現，由於此資料庫維度較大，因此NM-PSO與PSO，到了最後都未完全收斂，因此判定，這兩種分群法需要更多的迭代次數。

表4.25 Glass資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	199.6818	203.3707	243.451	271.290	215.6775
平均距離值	200.50	207.3485	248.9633	291.329	260.3996
距離值標準差	2.2601	5.1243	6.8282	12.3323	36.8238
最佳距離值(排名)	1	2	3	4	5
平均距離值(排名)	1	2	4	5	3

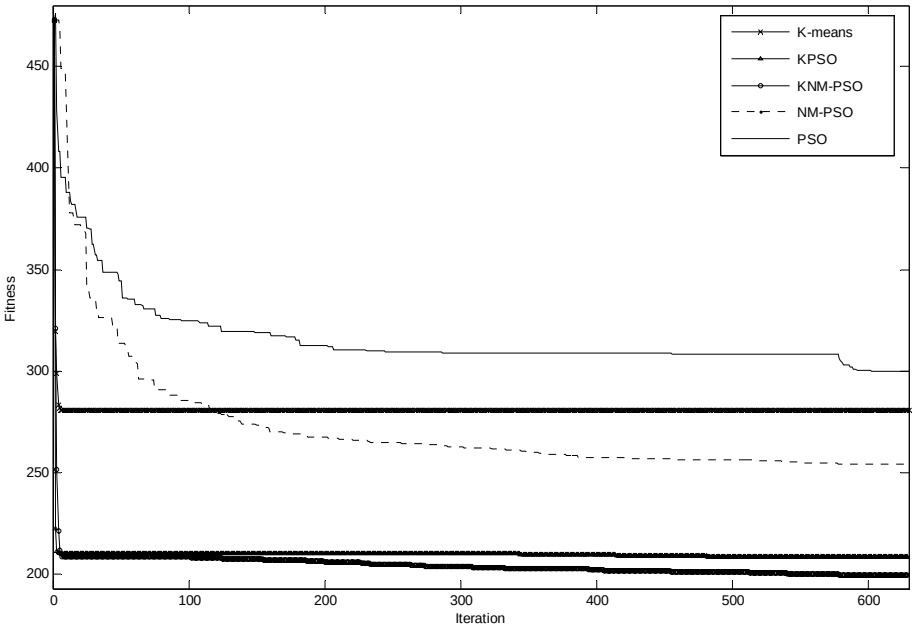


圖4.16 Glass資料庫之分群收斂圖

Cmc資料庫:

在表4.26中顯示，最佳距離值排名依序為KNM-PSO、KPSO、NM-PSO、PSO與K-Means，並且其數值差異性不大，但由平均距離值就可發現，NM-PSO，明顯優於PSO與K-Means。

在收斂狀況方面，K-Means在圖4.17中收斂最快，但卻陷入了區域最佳解。KNM-PSO與KPSO在一開始收斂值接近，但仔細檢視收斂圖至最後，可發現KNM-PSO跳脫出區域最佳解，並尋找較佳解。

表4.26 Cmc資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	5532.4	5532.887	5537.3	5538.5	5542.2
平均距離值	5532.7	5532.9	5563.4	5734.2	5693.6
距離值標準差	0.2277	0.0896	30.2732	289	473.141
最佳距離值(排名)	1	2	3	4	5
平均距離值(排名)	1	2	3	5	4

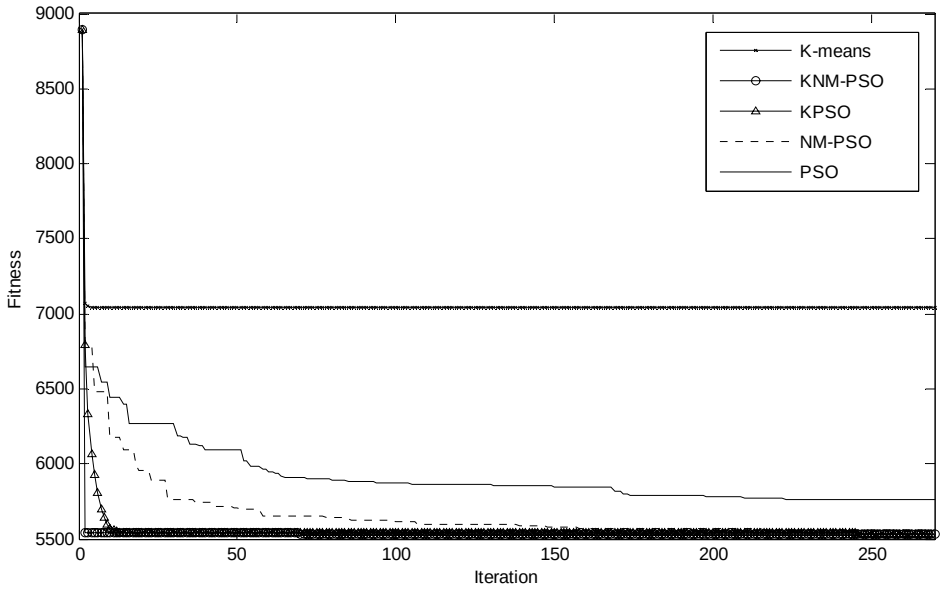


圖4.17 Cmc資料庫之分群收斂圖

Vowel資料庫:

表4.27為五種分群方法之收斂圖，可輕易的發現KNM-PSO，不論是在最佳、平均距離值方面，皆非常明顯的小於其他四種方法，而PSO在此資料庫中明顯表現為最差。

圖4.18表現出五種方法的收斂狀況，KNM-PSO、KPSO與K-Means約在迭代次數10次左右便大致收斂了，NM-PSO則在迭代次數約100次大致收斂，反觀PSO則至最後都尚未收斂，分析其原因有二:其一是PSO有可能陷入區域最佳解，無法跳脫，另外一個可能性，則是PSO需要較多的迭代次數。經過比較，可以輕易的比較出NM-PSO與PSO之特性與優劣，NM-PSO能快速的收斂，並且只需3N+1的群體數量，因此，不論是收斂值或運算速度，都比PSO要好上許多。

表4.27 Vowel資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	149005	149206.1	149240	163882	149422.3
平均距離值	149141.4	149375.7	151983.9	168477	159242.9
距離值標準差	120.3759	155.5549	4386.433	3715.73	91631
最佳距離值(排名)	1	2	3	5	4
平均距離值(排名)	1	2	3	5	4

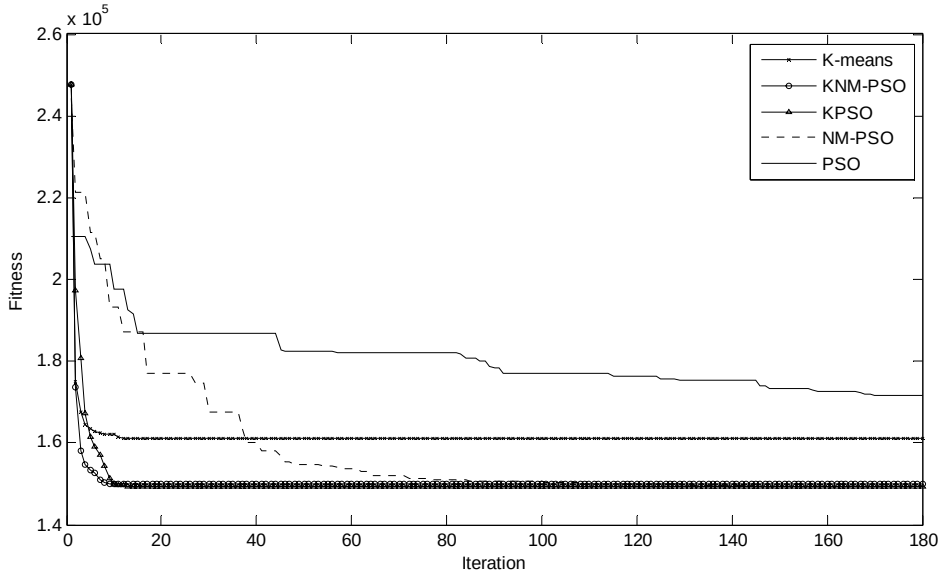


圖4.18 Vowel資料庫之分群收斂圖

Crude Oil資料庫:

五種方法對Oil資料庫進行分群，結果如表4.28所示，KNM-PSO仍得到最佳的結果，值得注意的是，NM-PSO在此資料庫中的表現，優於其他三種方法，並且優於已經過整合分群法的KPSO，因此證明KNM-PSO與NM-PSO的確能有效處理分群問題。

而圖4.19更能發現，KNM-PSO在K-Means初步收斂後，能夠有效尋找最佳解，至於K-Means仍然有著快速收斂的優點與陷入區域最佳解的缺點，而NM-PSO收斂狀況和數值也相當不錯。

4.28表 Crude Oil資料庫之分群距離

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳距離值	277.149	277.4506	277.1995	279.074	279.1956
平均距離值	277.2889	277.7666	277.5867	285.509	287.3552
距離值標準差	0.095	0.3255	0.3659	10.3088	25.4061
最佳距離值(排名)	1	3	2	4	5
平均距離值(排名)	1	3	2	4	5

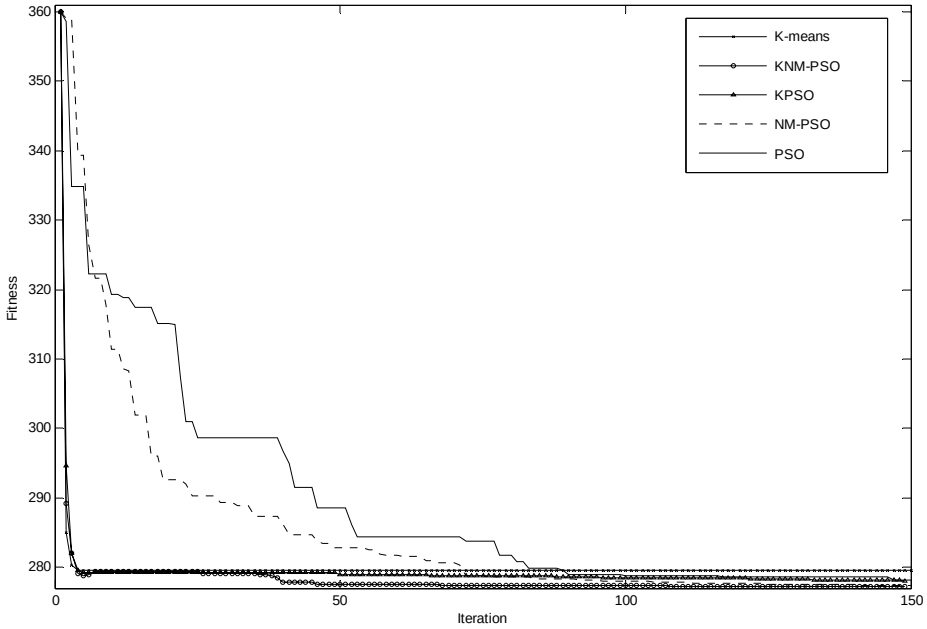


圖4.19 Crude Oil資料庫之分群收斂圖

計算評估函數總數:

計算評估函數總數，其實就代表著運算速度，表4.29是以五種方法，對七個實際資料庫進行分群時，所計算評估函數的總數，其結果與上述表4.20相同，仍然是K-Means所需最低，接著是KNM-PSO、NM-PSO、KPSO與PSO。K-Means雖然是所需的計算評估函數總數最低，但K-Means原本就不算是啟發式演算法的一種，而且結果往往都比其他四種方法差。而KNM-PSO是除了K-Means之外，所需計算評估函數總數最低，並且距離值最好的一種分群方法。

表4.29 對實際資料庫進行分群之計算評估函數總數表

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
計算評估函數總數(Iris資料庫)	4556	6906	4836	7260	120
計算評估函數總數(Wine資料庫)	46459	74305	47309	76245	390
計算評估函數總數(Cancer資料庫)	10149	15756	10485	16290	180
計算評估函數總數(Glass資料庫)	119825	195625	121773	198765	630
計算評估函數總數(Cmc資料庫)	21597	34843	23027	36585	270
計算評估函數總數(Vowel資料庫)	9291	15133	10501	16290	180
計算評估函數總數(Oil資料庫)	7057	10807	7394	11325	150

相關文獻進行比較:

為了比較KNM-PSO是否為一有效之分群法，因此與相關進行分群之文獻，以距離值進行比較，但由於每篇文獻所使用之資料庫不盡相同，因此只針對相同之資料庫進行比較。

表4.30就是以Adaptive GA+K-Means(葉承銓，2001)、GA(Ujjwal and Sanghamitra, 1999)、KGA (Sanghamitra and Ujjwal, 2000)與本論文所提出的KNM-PSO進行比較，共有六個資料庫，以平均距離值比較。並不難發現KNM-PSO在對每個資料庫進行分群的表現都比其他三種好。而在Oil資料庫中，由表4.28得知，NM-PSO之距離值為277.5867，就已優於上述三種分群方法。而KNM-PSO則是更優於其他方法。

表4.30 與相關文獻之距離比較表

	KNM-PSO	Adaptive GA+K-Means	GA	KGA
平均距離值(Iris)	96.6668	97.05	97.10	97.10
平均距離(Cancer)	2964.7	2976.58		
平均距離值(Glass)	200.50	200.65		
平均距離(Vowel)	149141		149363.61	149378.029
平均距離值(Oil)	277.149		278.9651	278.965

誤差率:

表4.31為以五種方法，對七個實際資料庫進行資料分群之誤差率結果，包含了最佳和平均誤差率以及誤差率標準差，可以發現，以五種方法，對每一個資料庫分群後的誤差率都不盡相同，雖然KNM-PSO對每一個資料庫進行分群，以距離為評估函數之結果都是最好的，但在誤差率的表現並非如此，就拿Glass資料庫來說，最佳距離值是KNM-PSO為最佳，但是最佳誤差率，卻是以NM-PSO進行分群的9.35%，而Cmc資料庫所得到的結果也是一樣，最佳距離值仍是KNM-PSO為最佳，但最佳誤差率則是距離值表現較差的PSO，表現的最好，所以從中可以發現，資料中心點與資料點的距離，並不是和誤差率有絕對的關係，對於這方面的問題，經由相關文獻的閱讀以研究，得到其原因，並將在第五章，做詳盡的說明。

表4.31 對人工資料庫進行分群之誤差率列表

	KNM-PSO	KPSO	NM-PSO	PSO	K-Means
最佳誤差率 (Iris資料庫)	10.00%	10.00%	8.00%	10%	10.67%
平均誤差率 (Iris資料庫)	10.20%	10.07%	11.13%	12.53%	17.80%
誤差率標準差 (Iris資料庫)	0.32	0.21	3.02	5.38	10.72
最佳誤差率 (Wine資料庫)	28.09%	28.09%	28.09%	28.09%	29.78%
平均誤差率 (Wine資料庫)	28.48%	28.37%	28.48%	28.71%	31.12%
平均誤差率 (Wine資料庫)	0.27	0.4	0.27	0.41	0.71
最佳誤差率 (Cancer資料庫)	3.66%	3.66%	3.51%	3.66%	3.95%
平均誤差率 (Cancer資料庫)	3.66%	3.66%	4.28%	5.11%	4.08%
誤差率標準差 (Cancer資料庫)	0	0	1.1	1.32	0.463
最佳誤差率 (Glass資料庫)	12.62%	12.62%	9.35%	11.21%	11.00%
平均誤差率 (Glass資料庫)	30.45%	32.17%	40.89%	45.59%	37.71%
誤差率標準差 (Glass資料庫)	14.13	15.52	15.58	15.62	13.75
最佳誤差率 (Cmc資料庫)	54.31%	54.38%	54.38%	54.24%	54.45%
平均誤差率 (Cmc資料庫)	54.38%	54.38%	54.47%	54.41%	54.49%
誤差率標準差 (Cmc資料庫)	0.054	0	0.0644	0.13	0.0351
最佳誤差率 (Vowel資料庫)	40.64%	40.64%	40.07%	41.45%	42.02%
平均誤差率 (Vowel資料庫)	41.94%	42.24%	41.96%	44.65%	44.26%
誤差率標準差 (Vowel資料庫)	0.92	0.95	0.98	2.55	2.15
最佳誤差率 (Oil資料庫)	23.21%	23.21%	23.21%	23.21%	23.21%
平均誤差率 (Oil資料庫)	24.29%	23.93%	24.64%	24.29%	24.46%
誤差率標準差 (Oil資料庫)	0.92	0.72	0.75	1.73	1.21

第五章 結論與未來研究方向

5.1 研究結論

本論文是以結合K-Means與NM-PSO的方法來解決資料分群的問題，期望透過K-Means快速收斂的特性，與NM-PSO跳脫區域最佳解並搜尋全域最佳解之能力，透過有系統的整合，能有效將資料加以分群，並得到以下結論:

1. K-Means收斂速度雖快，但易陷入區域最佳解與分錯群組。
2. NM-PSO之收斂速度，明顯優於PSO。
3. NM-PSO對於資料分群之距離值，明顯優於PSO與K-Means。
4. 以尚未與K-Means整合之分群法而論，NM-PSO分群成效為最佳。
5. KNM-PSO之收斂速度，明顯優於PSO、NM-PSO。
6. 距離總值與誤差率之間並沒有絕對的關係，其原因是實際資料庫之資料分佈狀況並不一定呈規則性分佈，因此距離總值較佳，誤差率不一定低。
7. KNM-PSO對於資料分群之距離值，不論最佳或平均值，與本研究其他四種分群法相比，皆為最佳。

5.2 未來研究方向與應用

本論文使用KNM-PSO分群法以分割式分群進行分群技術，從中發現，對於誤差率的判斷能力較差，其原因是由於實際資料庫中，資料點與群體中心之間的距離，和分群結果沒有絕對的關係，而分割式分群法雖以群體中心點和資料點之距離為評估函數，雖有良好的結果，但是誤差率卻偏高。以圖5.1為例，圖中以兩種標示代表兩群資料之分佈狀況以及正確分群結果，而圖5.2則是以群體內與中心點之距離為指標之分群結果，由此兩圖可知，若以距離，為評估函數，則圖5.2之結果將優於圖5.1，但以誤差率而言，圖5.1之分群方式完全正確，而圖5.2卻存在著不小的誤差。

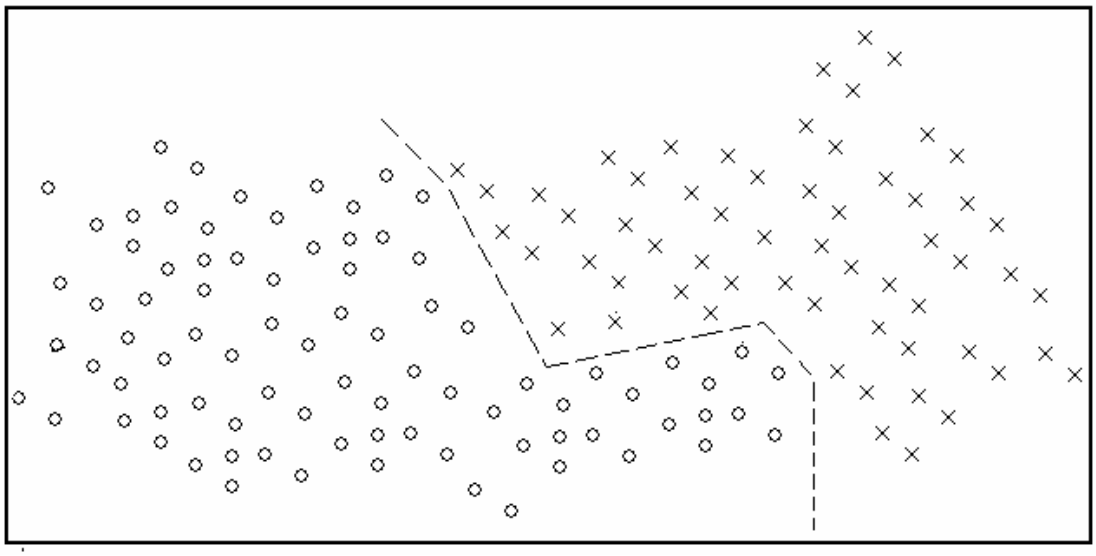


圖5.1 正確分群圖

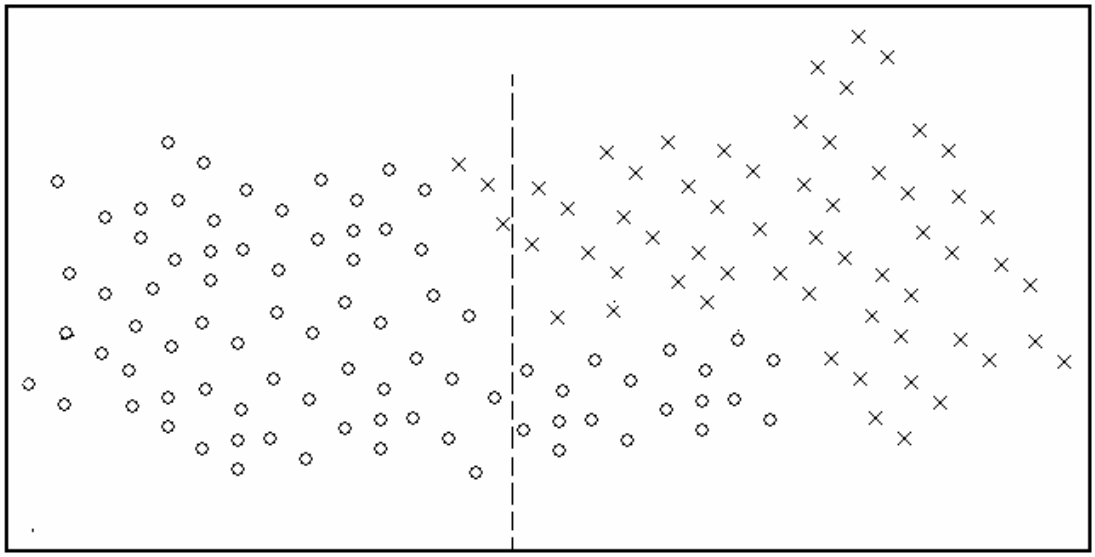


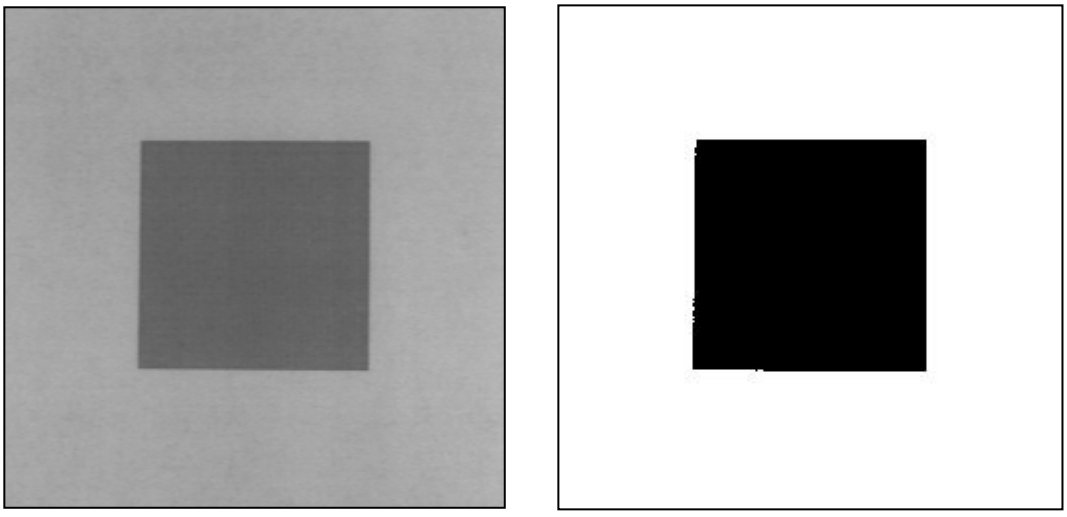
圖5.2 以距離分群結果圖

經由閱讀過相關文獻指出，由於階層式分群是以資料點性質來加以分群，，誤差率大多都優於分割式分群法。因此未來研究方向將建議使用NM-PSO演算法，應用於階層資料分群，希望能有效改善以分割式分群之誤差率缺失。

階層式分群法有下列優點:

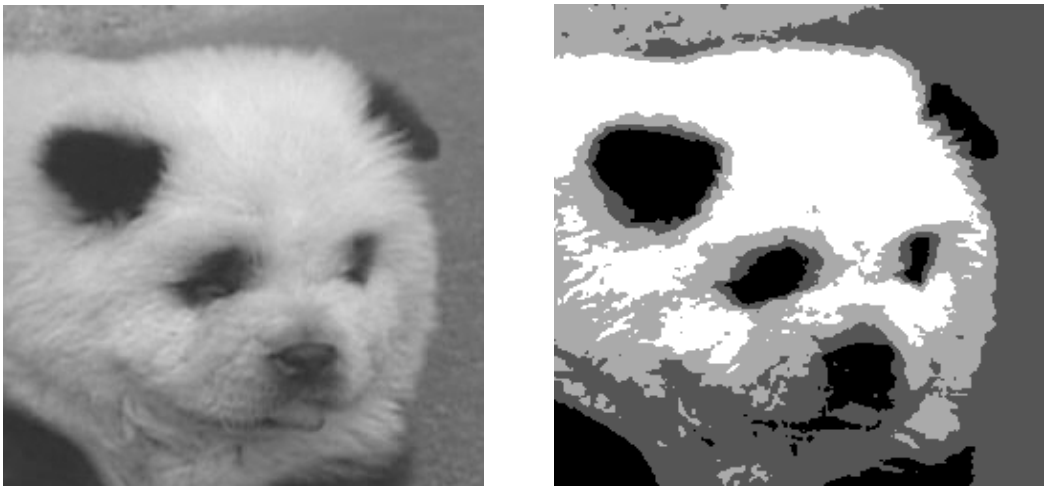
1. 透過分群樹狀圖，可了解資料間相關性的完整架構與極端值之判斷，與分群的過程，將有助於分群資料之分析。
2. 由於階層式分群是以各個資料點之屬性進行分群，因此充分考慮點與點之間的關係，而並非是以中心點進行分群，因此將有助於誤差率之改善。

在相關應用方面，本論文已將此分群技術，應用至影像處理。主要的目的是希望能夠透過尋找色階切割點，來將圖片中之主體與背景以不同顏色來區分，圖5.3(a)為原始圖片1，經過資料分析後，以色階133為切割點進行處理，結果為圖5.3(b)所示，從中可發現，背景與主體物件以黑白兩色，十分明顯的區隔開。圖5.4(a)則為一原始熊貓之圖片2，經過分析決定將其色階分為四群，因此經由實驗可得知切割點為91、140與179。從圖5.4(b)可知，經過影像處理後，可將大致的輪廓呈現，但仍然有些顏色較相近的地方，尚未能完全區分。也因此驗證了，以分割式分群法進行分群，誤差率仍有改善的空間。



(a) (b)

圖5.3 原始與經處理的圖片1 (切割點:133)



(a) (b)

圖5.4原始與經處理的原始圖片2 (切割點:91、140、179)

參考文獻

英文文獻

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). "Automatic subspace clustering of high dimensional data for data mining applications," *ACM SIGMOD Int'l Conference on Management of Data*, 94-105.

Box, G. E. P. (1957). "Evolutionary operation: a method for increasing industrial productivity," *Applied Statistics*, Vol. 6, 81–101.

Cheo, C. Y. and Ye, F.(2004). "Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis," *lnternational Conference on Networking, Sensing Control*, 789-794.

Dorigo, M. and Colorni, V. (1992). "Distributed Optimization by Ant Colonies," *In Proceedings of European Conference on Artificial Life*, 134-142.

Duda R. O., and Hart P. E. (1973). *Pattern classification and scene analysis*, John Wiley & Sons, New York.

Merwe, D. V. and Engelbrecht, A. (2003). "Data cluster using particle swarm optimization," *Evolutionary Computation, 2003 Proceedings., IEEE International Conference* Vol. 1, 215-220.

Glover, F. (1989). "Tabu search─part I," *ORSA Journal on Computing*, Vol. 1, 190-206.

Jain, A. K. and Dubes, R. C. (1988). *Algorithm for clustering data*, Prentice Hall, New Jersey.

Holland, J. (1975). *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, New York.

Kennedy, J. and Eberhart, R. C. (1995). "Particle swarm optimization," *Neural Networks, 1995 Proceedings., IEEE International Conference*, Vol. 4, 1942-1948.

Kennedy, J. (2000). "Stereotyping: improving particle swarm performance with cluster analysis," *Evolutionary Computation, 2000 Proceedings., IEEE International Conference,* Vol. 2, 1507 – 1512.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). "Optimization by simulated annealing," *Science*, Vol. 200, 671-680.

Nelder, J. A. and Mead, R.(1965). "A simplex method for function minimization," *Computer Journal*, Vol. 7, 307-308.

Reynolds, C. W. (1987). "Flocks, herds and schools: a distributed behavioral model," *Computer Graphics*, Vol. 21, 25-34.

Sanghamitra, B. and Ujjwal, M. (2000). "Genetic algorithm-based clustering technique,"*Pattern Recognition*, Vol. 33, 1455-1465.

Fan, S. K., Liang, Y. C. and Zahara, E. (2004). “Hybrid simplex search and particle swarm optimization for the global optimization of multimodal function,” *Engineering Optimization*, Vol. 36, 401–418.

Sanghamitra B., Ujjwal, M. (2002). “An evolutionary technique based on K-Means algorithm for optimal clustering in R^N,”*Information Sciences*, Vol.146, 221-237.

Spendley, W., Hext, G. R. and Himsworth, F. R. (1962). “Sequential application of simplex designs in optimization and evolutionary operation,”*Technometrics*, Vol. 4, 441-461.

Ujjwal, M., Sanghamitra B. (1999). “Genetic algorithm-based clustering technique”*Pattern Recognition*, Vol. 33, 1455-1465.

Zahara, E., Fan, S. K. and Tsai, D. M. (2005).“Optimal multi-thresholding using a hybrid optimization approach,” *Pattern Recognition*, Vol. 26, 1082-1095.

中文文獻

何怡偉(2004)，*Nelder-Mead搜尋法處理無限制式及隨機最佳化問題之研究*，博士論文，元智大學工業工程與管理系，桃園。

李漢祥(2002)，*協同運算研究(1)–多單體基因演算法*，碩士論文，東海大學工業工程學系，台中。

林育臣(2002)，*群聚技術之研究*，碩士論文，朝陽科技大學資訊管理系，台中。

陳孟佐(2004)，*混合階層式遺傳演算法與粒子群優演算法之資料分群技術*，碩士論文，樹德科技大學資訊管理系，高雄。

陳慶逸(2003)，*Particle swarm optimization algorithm and Application to pattern Recognition*，中國海事商業專科學校，改善師資案-各科教師研究計劃成果報告，台北。

彭文正譯(2001)，資料採礦─顧客關係管理暨電子行銷之應用，M. J. A. Berry, G. Linoff 原著，數博網資訊股份有限公司，維科出版社，台北。

葉承銓(2002)，*應用適應性基因演算法於資料分群的問題*，碩士論文，樹德科技大學資訊管理系，高雄。