第一章　緒論

1.1 研究背景與動機

隨著資訊科技的進步以及硬體設備的快速升級，在科學、工程設計、經濟管理及控制系統等領域上，對於資料的收集和儲存已經是很容易的事。在龐大資料下，並不是所有的資訊都是有用的，如何在這龐大的資料中找尋隱藏在其中的有用資訊，並做有效率且正確的處理，是一門很重要的課題。

群集分群為資料分析的前置動作。其目的是將大量未整理的資料分群，使得相同群集內的資料具有較高的相似性，而不同的群集的資料相似性會較低。資料分群可將其視為最佳化問題，其求解難度隨著問題規模之增大，而迅速提增。主要目的在於區隔不同及未知類別的資料樣本，將相似度較高的樣本分群在相同的群組裡。目前群集分群的較常用方法為以資料的距離來做為分群的依據，也就是資料點之間距離越近，則越容易分在同一群集中。

在傳統的分群法中較常使用到的便是k-means演算法，但卻有不足之處，須由使用者決定群集數目以及容易受到初始點的影響結果造成落入區域解分群導較得到較差的結果。因此分群的方法，就變成是一個最佳化的問題，換句話說，要順利地找到群聚中心點，使得距離總合的值為最小。在文獻中有許多改良k-means演算法不足之處，利用階層式方法的概念對資料做分裂或聚合來處理初始中心點的位置與群聚數的給定[24-26]，而針對如何跳脫區域解的問題大多以啟發式演算法求解相關議題居多，因此有許多學者利用啟發式算法，如遺傳演算法（Genetic Algorithms，GA）[22]、模擬退火 (Simulated Annealing，SA)[23]以及粒子群演算法(Particle Swarm Optimization，PSO)[20]等，來改進k-means演算法容易落入區域解的問題。

模擬退火演算法具備有全域的搜尋能力，但傳統的模擬退火演算法在擾動方面是採用隨機產生測試方式，若新解較佳則取代舊解，使得傳統的模擬退火演算法缺乏在局部空間中搜尋的能力。因此，衍生出將模擬退火演算法與具有局部搜尋能力較高的技術結合，加強模擬退火演算法的搜尋能力，但隨著問題求解的複雜度增加，一個具有精確的找到全域最佳解且快速收斂到最佳解的演算法勢必需的。

1.2 研究目的

基於以上所敘，k-means演算法具有快速收斂以及概念簡單等優點 ，但k-means演算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。有鑑於此，綜合以上的動機。本文的目的為將使用結合模擬退火演算與正交實驗設計進行資料分群，簡稱為HSAKM，希望利用k-means演算法快速收斂的優點，再以結合模擬退火演算與正交實驗設計來補足k-means演算法易陷入區域最佳解的缺失。期望HSAKM應用於資料分群能有下列優勢:

1. 初始中心點之影響:由於k-means演算法的第一個步驟是在範圍空間中隨機取得初始中心點位置，因此常會導致落入區域解位置導致整體效能穩健性不佳，收斂至較差結果的可能性。因此希望透過模擬退火演算法的跳躍機制能順利跳脫。
2. 距離總值將為最佳:由於k-means演算法常最後之距離總偏移值會不理想，會有誤判資料的機率因而導致分群效果不理想。經由HSAKM分群過後能夠比k-means演算法得到較低的距離總值，藉以提升群集結果之品質。
3. 評估函數：傳統模擬退火演算法需要反覆降溫求解，故所需之評估函數會較多。模擬退火演算法透過與實驗設計法的結合能減少退火的搜尋時間，減少評估次數，並希望能以較少的評估次是達到最佳的分群

效果。

4. 處理多維度的問題: k-means演算法在處理維度特徵高的資料時，往往會結果會有可能分群效果較差。
5. 資料庫處理:在實驗過程中能夠同時解決實際資料庫以及人工資料庫的分群問題，並且能夠優於k-means演算法以及其他方法。

1.3 研究步驟

1. 收集文獻

   收集國內外之相關文獻進行適當的問題分類，並歸納分析各類問題的究現況與其特性。本研究主要之文獻分別為：

   (1) 群集分析，廣泛的回顧目前分群之研究上所使用之研究方法之優缺點及各研究者使用的狀況。

   (2) 回顧各學者應用模擬退火於分群之精神與特色。

2. 系統建構、程式撰寫

   根據本研究之演算模式，以程式語言撰寫演算搜尋程式，將演算流程程式化。

   (1) 收集所需的資料庫，並整理至系統資料庫中。

   (2) 撰寫k-means演算法。

   (3) 撰寫HSAKM程式

3. 系統實驗與比較

   經由模擬以及實際案例進行測試，驗證本研究所應用之理論與其建構之演算模式成效，並和先前相關研究做一比較。

4. 研究論文撰寫

   統合整理本研究之各項資料與結果，彙整本研究之結論並對未來研究方向提出建議，最後撰寫完成書面論文。

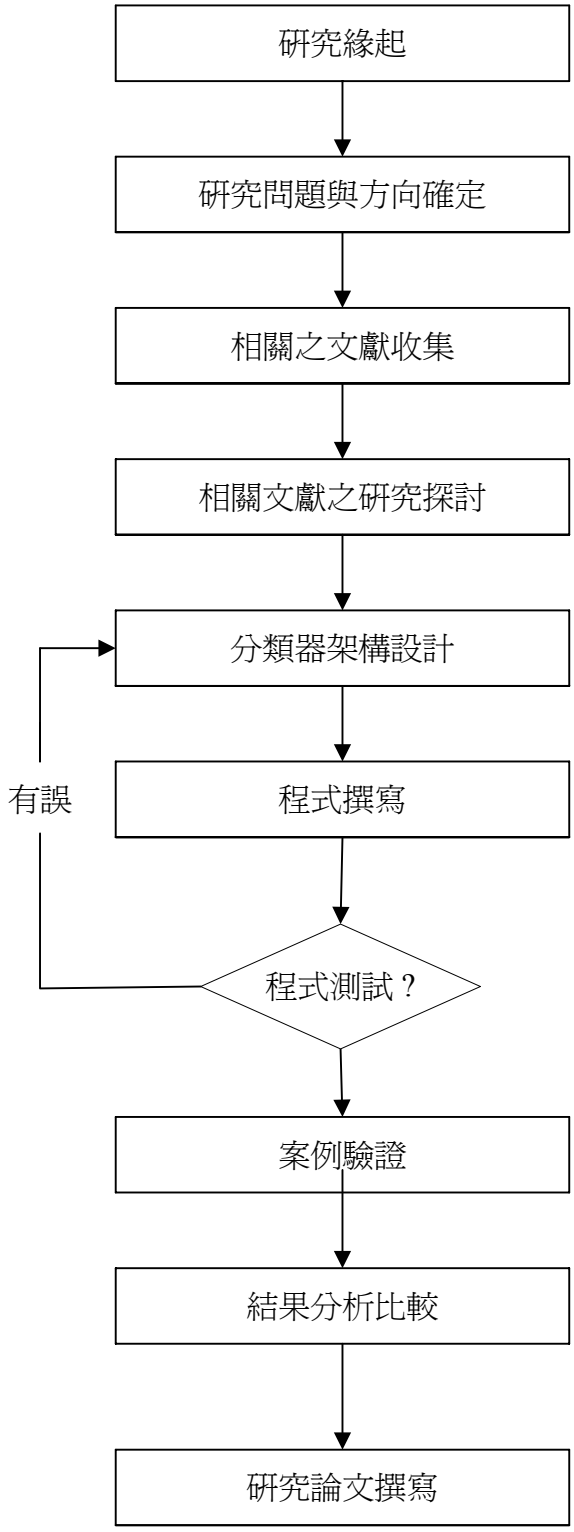

研究緣起
研究問題與方向確定
相關之文獻收集
相關文獻之研究探討
分類器架構設計
程式撰寫
有誤
程式測試？
案例驗證
結果分析比較
研究論文撰寫

圖1-1 研究流程圖

第二章 相關研究

本章節分為兩部份。第一部分是關於資料分群技術之簡介，第二部份為模擬退火演算法之說明

2.1 資料分群技術

群聚技術(Clustering Technology)可以將資料依據分群的目標分成許多群聚；而被凝聚在同一群的資料會有某些特性是相近的；也就是說，被分成不同群聚的資料就會有某些特性會明顯不同。工程與科學界常使用這個技術來處理生物資訊(Bioinformatics)[14]、資料探勘 (Data Mining)[15]、人工智慧(Artificial Intelligence)[16]等等問題。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。群聚技術大致上可以分四類:

(1) 以階層式方式：較常使用的模式有兩種，分別是分裂法以及聚合法，分裂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集分開，直到分成單一群集。而聚合法則與上述方法相反，首先將每一個資料歸為一群，並逐漸將彼此相似度高的群集合併，直到所有的物件合併為單一群集。在分割前必須先指定群聚的數目，然後藉著反覆迭代運算，逐次降低目標函數的誤差值，直到目標函數不再變化，即為最後的分群結果。階層式可處理任意形狀之群集、群集大小差異大等方面之問題，但當所要處理的資料數較多時其時間需求會較高。

(2) 以分割為基礎的方式：一般而言，分割式分群法的目的將資料點切割到數個互不交集的群集中，讓每一群集中的資料點，每一點與群中心的距離偏移值小於其他群集中心。假設現在擁有資料總數為S的資料集

合，將其分配到k個互不交集群聚中，而每一個群集包含N_k筆資料群聚中心為c_k，則該群聚的距離偏移值可以定義為

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - c_k \| \qquad (2\text{-}1)$$

分割式通常一開始會先選擇k個資料點當做初始中心點，然後經由反覆迭代運算找出較佳的群集中心來降低距離總偏移值，直到目標函數不再變化，就達到分群的最後結果。分割式並不適合於任意形狀之群集做分群，因其容易受初始中心點所挑選的位置好壞，將會對結果具有決定性的影響。

(3) 密度導向的方式：大部分以分割式或階層式為基礎的方法都是基於資料點間的距離關係來分群，密度導向的方式則是利用資料點間密度的關係來分群。其主要在固定的範圍內測量所涵蓋資料點的數目是否有達到所設定的目標。如果為達到將視資料集合中較密集的資料為一個群集，而密度低的資料則被視為雜訊，之後再利用反覆地將較小的群集以形成較大的群集。運用密度的方法可以用來過濾雜訊，而且可以對任意形狀之群集做分群，但最大難題是如何去設定適當的涵蓋範圍和密度評估值。

(4) 格子結構為基礎的方式：格子結構為基礎的方式則是將資料所在的空間量化切成多個單位，若某個單位是含有符合一定比例的資料數目便可以考慮串聯其他也含有超過一定比例資料數目的單位，當然這些單位必須是鄰近的，這些彼此鄰近且含有一定比例以上資料的單位便被集合起來形成一個群聚。

以上四種類型的群聚演算法各有其優劣，並不能通用於各種情況，如分割式只能找出類圓形和群集大小相似的群集，而對於影像辨識肉眼即可判斷群集大小就並不適合；這時採用能產生任意形狀和任意大小的階層式

或密度導向的方式較佳。因此有許多學者結合多種群聚技術，取其演算法的優點，成為更好的群聚演算法。圖2-1為各種群集分群的發展概況[27]。

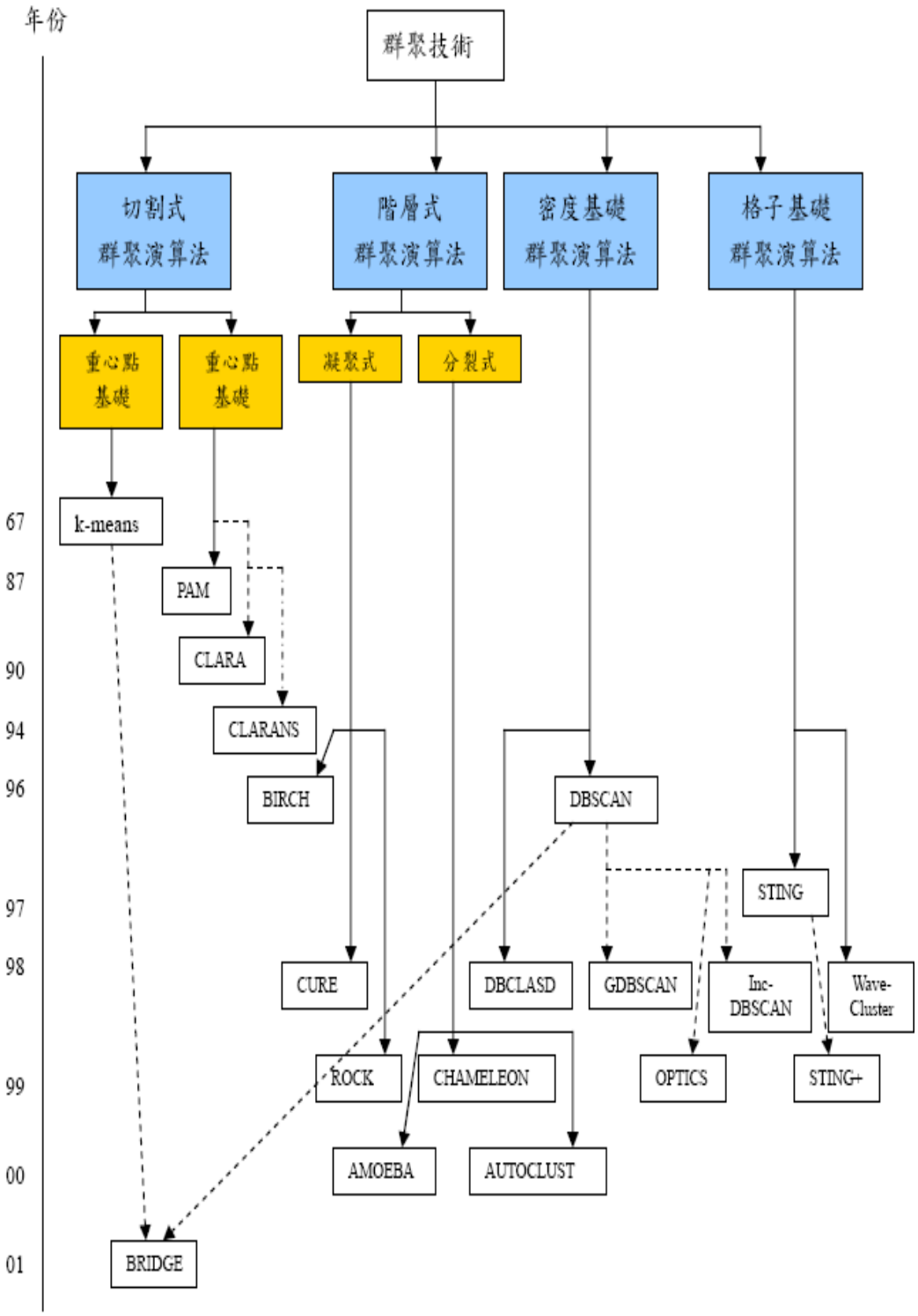


圖2-1 群聚分群的發展概況

2.2 K-Means 演算法

在所有的分割式分群法（partitional clustering）之中，最基本的方法，就是所謂的 K-means演算法。其概念很簡單，就是使用一個反覆迭代的方式讓目標函數越來越小。首先，我們知道目標函數是由群聚的分法和群聚的中心點所決定;如果要同時找到這兩組參數，使目標函數為最小，是一件不容易的事情。因此可以觀察到下列現象：

1. 當群聚的分法固定時，可以很快地找到群聚的中心點，使得目標函數為最小。此時的群中心，就是每一群的平均值。
2. 當群聚的中心點固定時，我們可以很快地找到群聚的分法，使得目標函數為最小。此時第k個群集，就是距離第k個群中心最近的資料點所成的集合。

在上述方法中，我們是先找群集中心點位置，再開始反覆迭代的過程。事實上，也可以先進行任意分群，然後再進行反覆迭代的過程，得到的結果應該很類似。特別要注意的是，在迭代過程中，目標函數應該是隨迭代次數而遞減，直到小於某一個特定值，就不會再變化。其過程如圖2-2所示。以下簡單說明k-means演算法步驟:

1. 由訓練樣本中隨機挑選k個初始中心點m_j， $j=1,2,...,k$

2. 計算各個訓練樣本p_i與m_j之距離，$i=1,2,...,s$

$$D_i^j = \sum_{i=1}^{s}\sum_{j=1}^{k}\left\| p_i - m_j \right\| \tag{2-2}$$

3. 將p_i分配到各自最接近的群聚

$$p_i^w \ ,if \ \left\| p_i - m_w \right\| < \left\| p_i - m_j \right\| \ ,w=1,2,...k,\ w \neq j \tag{2-3}$$

4. 依照各群聚點數量n_w，計算出各群聚新的中心點m_w

$$m_w = \frac{1}{n_w}\sum_{i=1}^{n_w} p_i^w \qquad (2\text{-}4)$$

5. 反覆執行2-4步驟直到每一個群聚中心都不再改變為止。

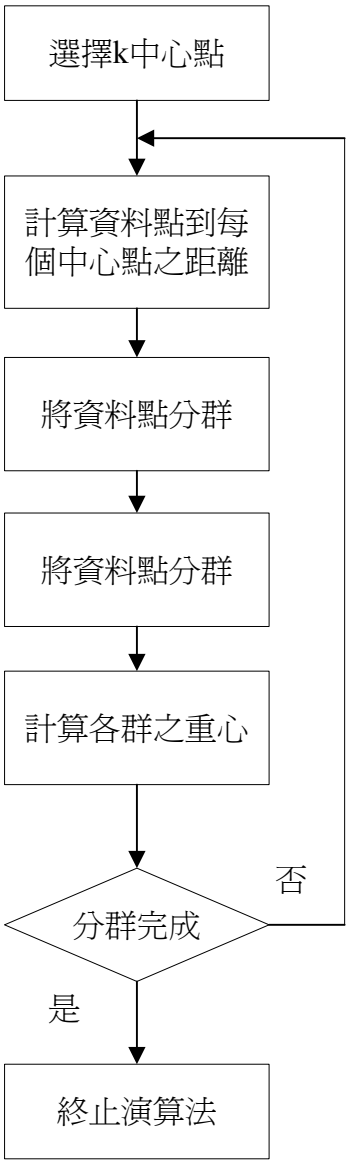


圖2-2 k-means 演算過程

2.3 模擬退火演算法

目前以自然為基礎的演算法來求解全域最佳化問題可說是最有效率並且使用容易，雖然此種演算法不能保證一定能找到全域最佳解，但通常都可以找到一個接近全域最佳解的值，且利用隨機、非直接的方式搜尋全域

最佳解。這類演算法中又以基因演算法和模擬退火法因發展的早，已被廣泛應用在許多科學和工程領域中。本研究運用模擬最火演算法作為擾動跳躍機制，以下將簡單說明模擬退火演算法。

2.3.1 模擬退火演算法介紹

模擬退火演算法是求解全域最佳解的一種方法，並且應用於求解最佳化問題上大多有良好的結果。例如影像處理(Image Processing)[2]、醫療診斷[4]領域等。最早於1953年由美國物理學家Mettopolis發表[5]，主要研究複雜系統且計算其中能量分佈(蒙特卡羅法)。雖然，模擬退火法相較於基因演算法與禁忌搜尋法早提出，但當時並沒有受到研究者的重視。直到1983年Kirkpatrick借用了Mettopolis的方法探討一種旋轉玻璃系統(spin glass system)時，發覺物理系統的能量和一些最佳組合問題的成本函數相當類似。於是，發展出以Mettopolis方法為基礎的一套演算法。這時才將模擬退火法完整的提出，並詳述模擬退火法的基本原理與使用機制

模擬退火法的原理，為一種模擬物理結晶的退火過程，當物體經由加熱到一定溫度後，分子可以隨意組合，是不穩定的；當物體溫度逐漸下降，物體開始冷凝與結晶，在結晶狀態時，系統的能量狀態會最低。一旦達到最低溫度時，分子則會重新以一定的結構進行排列，可以找到最低能量狀態。但是，如果操之過急，快速降溫時會導致達不到最低狀態的非結晶形。因此緩慢下降溫度，使得物體分子在每一溫度時，能夠有足夠時間找到安頓位置，分子活動力也逐漸減弱。最後溫度達到低點，物質凝結為固體，分子組合將不再做改變。

模擬退火法的最大特徵在於它除了可以接受較佳解外，當目前狀況是落入區域解時，模擬退火法會藉由重新加熱的動作，透過隨機的過程，以機率性質的方式判斷是否接受所找到的鄰近擾動點使其能跳脫目前的區域最佳解，而有機會能達到另一個最佳解。故此特徵使得模擬退火法不同

於其他鄰點搜尋法，因而具有跳出局部最佳解進而達到全域搜尋之能力。如圖2-3所示。目前研究趨勢已逐漸傾向運用混合型演算法求解，換句話說，結合不同演算法的特點，來改善單獨採用一種演算法的搜尋效率，更進一步解決單一演算法所無法克服的瓶頸。使其能求得更精確之全域最佳解和較高的搜尋效能。

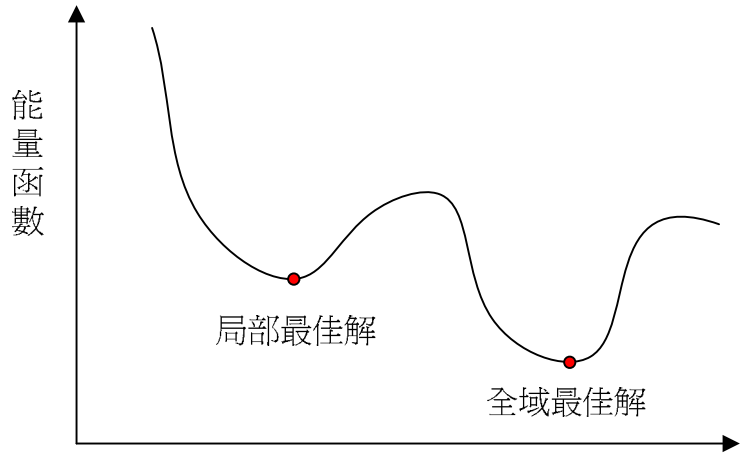


圖 2-3 模擬退火法示意圖

2.3.2 模擬退火演算法的擾動機制

模擬退火演算法的主要精髓在於擾動機制之特性,擾動之作法即是以目前解為中心，在整個可行解之空間中隨機取得單點解作為候選解，此即稱為擾動解。接下來利用波茲曼機率決定是否接受較差解為一新解。波茲曼機率計算公式如下：

$$P = \exp(-\Delta E / kT_j) \qquad (2\text{-}5)$$

(2-5)式中，k為波茲曼常數，通常設定為1。ΔE 為目標函數值差，為第 T_j 次之溫度設定。波茲曼機率將隨溫度之調整而改變，舉例來說：當溫度較大時或者ΔE較小時，有較大的機率接受擾動解。反之，當溫度較小時，容許

擾動解之機率則愈小。一旦溫度趨近於0時，則代表此新解一定為最佳解。理論上，只有在波茲曼機率為1時，才能保證皆能接受所有擾動解為新解。

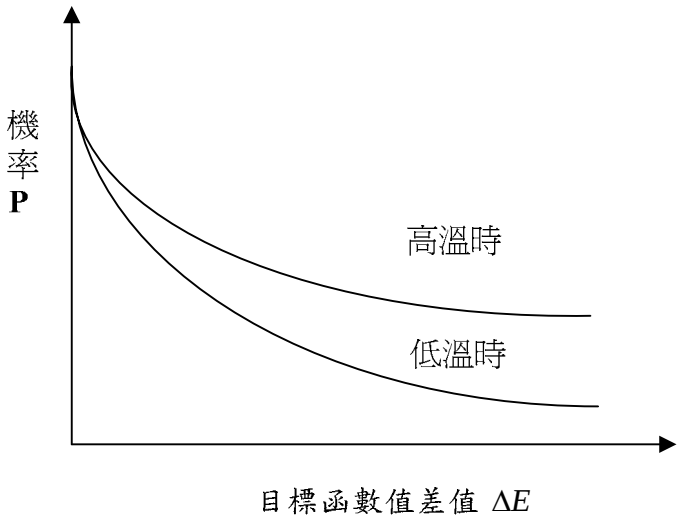


圖 2-4 波茲曼機率的分佈示意圖

2.3.3 模擬退火法之流程

圖2-4 是模擬退火演算法的流程圖。模擬退火法是運用溫度來調整使否要接受較差解之機率，故需引入溫度T來做為運算來模擬熱平衡過程，而此溫度T並不一定要具有實際的物理意義。在演算法開始的初期，需先設定一些參數，如初始溫度、冷卻率（Cooling Rate）和同一溫度執行的次數等。在冷卻機制部分比較有代表性的冷卻方式如下

$$T_i = \frac{T_0}{\ln(1+i)} \tag{2-6}$$

T_0為初始所設定的溫度，其特點為溫度下降緩慢，因此收斂速度也較慢。

$$T_i = \frac{T_0}{\ln(1+\alpha \times i)} \tag{2-7}$$

α為可調之參數，可以改善退火曲線的型態。其特點為在高溫時下降較快，低溫時則下降較為緩慢，所以此方式主要在低溫時做搜尋。

本研究則採用簡單的幾何冷卻規則,假設初始溫度T_0，在第$i+1$次降溫步驟後的溫度為

$$T_{i+1} = \alpha \times T_i \text{，} i = 0,1,... \tag{2-8}$$

其中α冷卻率介於0 到1 之間之常數,通常都取接近1以使得溫度下降不至於太快。使其經過反覆的求解過程，一直到達熱平衡狀態為止。最初先隨機產生初始解X，並且計算其目標函數值$f(x)$。以目前解為中心在其附近中做隨機擾動，產生一個新的狀態若此狀態的能量較原先的低，則接受此新的狀態；反之，則按照波茲曼機率決定是否接受此新的狀態。意即當能量改變為$\Delta E \le 0$，則接受機率為$P(T)=1$；若能量改變為$\Delta E > 0$，則接受機率為$P = \exp(-\Delta E / kT_j)$。接著判斷是否滿足降溫條件，若是，則透過冷卻機制降溫。反之，維持目前溫度。之後判斷是否達到終止條件，例如達到設定的迭代次數或是目前解都不再改變時。

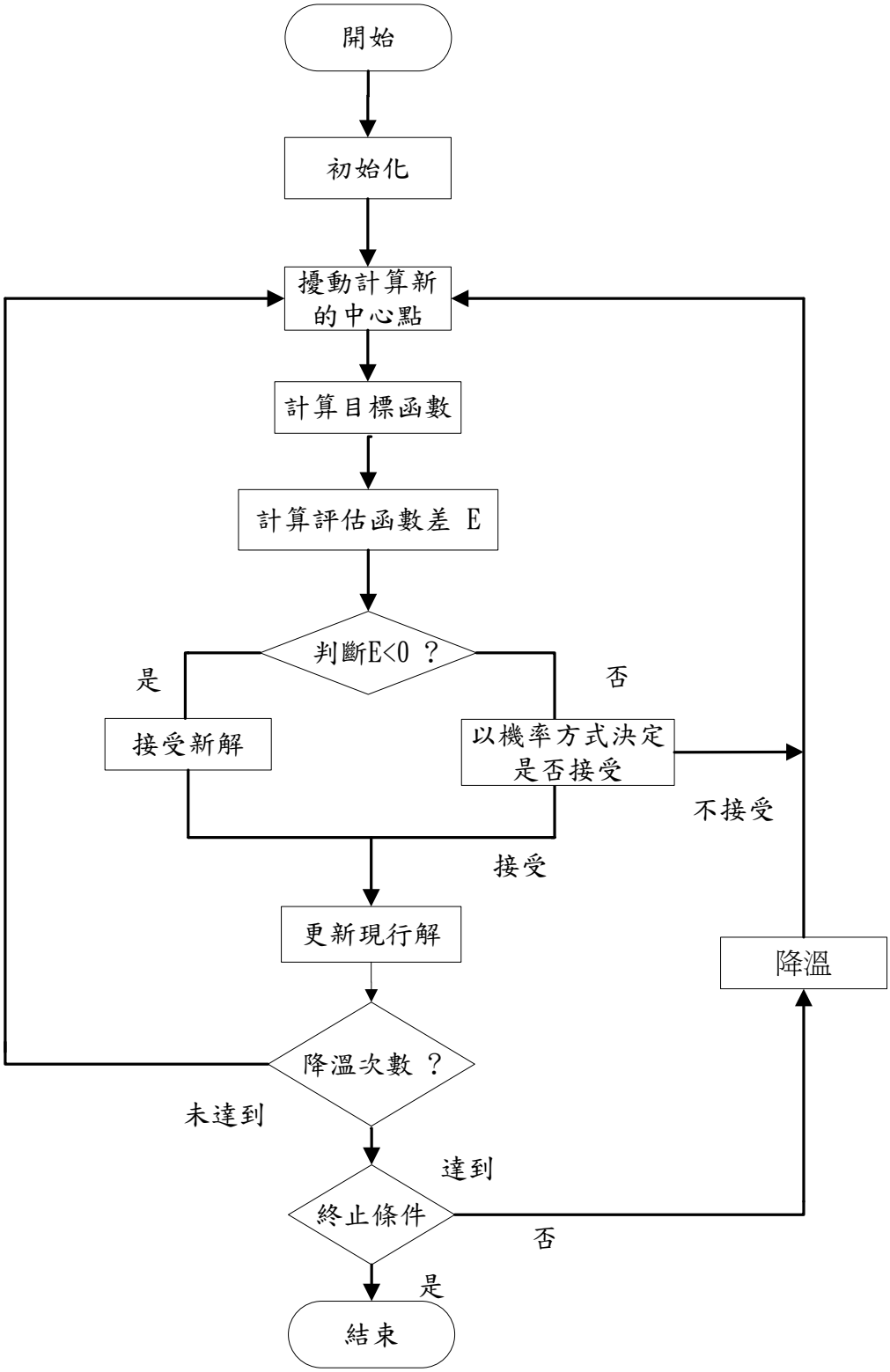


圖 2-5 模擬退火之流程圖

第三章　研究方法

本章將探討以k-means演算法作為基礎，針對上述所之k-means演算法的問題，思考因應改善之問題。因此提出結合模擬退火演算法於k-means演算法中之想法，進一步提升k-means演算法應用於資料分群搜尋成效及準確度。模擬退火演算法是一種啟發式隨機搜尋法，並具有收斂於全域最佳解的能力。本文以搜尋鄰近範圍較佳解的概念，將模擬退火演算法的跳躍擾動機制引入k-means演算法在範圍內找尋最佳的分群結果，因此期望距離總偏移值與收斂狀況能優於傳統k-means演算法。在此將k-means與模擬退火之混合方式稱為HSAKM混合搜尋法。其各部份分別說明於以下幾節。

3.1 正交實驗設計

雖然以自然為基礎的推測演算法一般是可找到接近全域最佳解，但隨著所求解的問題愈來愈困難，我們更需要一個更能準確找到全域最佳解和能更快速收斂到最佳解的高效率演算法。Zhang 和Leung 最早提出將正交實驗設計(Orthogonal Experiment Design，OED)中的正交設計（Orthogonal Design）的技術與基因演算法做結合來處理離散型變數的最佳化問題[7]，之後又利用正交設計及量化（Quantization）的技術處理連續型變數的最佳化問題，實驗的結果發現此方法不僅較強健且所找到的最佳解也較佳，但因只有從正交設計的正交表（Orthogonal Array, OA）所篩選出的實驗組合中選擇出其中最佳的因子組合，但並沒有對其實驗組合進行分析，因此所找到的組合可能並不一定是最佳的因素組合[8]。Tsai所提出的混合的田口基因演算法（Hybrid Taguchi-Genetic Algorithm, HTGA）[9]和Ho *et al.* 所提出的智慧型新解產生機制（Intelligent Generation mechanism, IGM [10]），分別使用訊號雜訊比(Signal-to-Noise Ratio，SNR）和因子分析（Factor Analysis, FA）對所產生的直交表做分析，使可以分析出真正的最佳組合。

Ho et al. 利用直交設計取代了傳統模擬退火法隨機擾動產生新解的方式，並進而應用求解大型積體電路（Very Large System Integration, VLSI）、平面規劃的問題[11]。

3.1.1 正交實驗計法之理論

許多科學實驗上，往往需要藉由許多假設來降低影響實驗結果的因素。但面對現實問題時，便無法使用相同的方式。在完全因子設計法（Full Factorial Design）中，當因子數目增加時，實驗次數會隨之增加。當考量許多會影響實驗結果程度的因素時，較為有效的其中一種方式便是使用正交實驗設計法[12-13]。OED包含兩個重要的部份：正交表與因素分析。利用正交表來取得資料，能讓我們以較少的實驗而獲得更可靠的因素效果估計量可有效降低實驗次數。假設因子數為N ，而每一個因子有三個水準值（Level），則需做3^N次實驗才能將所有的狀況考慮到。但當N 越來越大實驗的次數與複雜度也將大幅提升，如$3^4=81$次 、$3^8=6561$次 等。如此一來將導致所需耗費的成本和時間不符合成本效應的。

以下舉例一個有三個因子，而每一個因子有三個水準值的來說明。若以完全因子設計法來說，則需執行$3^3=27$次 實驗組合，圖3-1 為其實驗表示的空間分布圖。但若以部分因子設計法則僅需執行$3^2=9$次 實驗(如表3-1)，表3-1中的水準值1表示起始點，若水準值為2和3表示將作起始點附近區域的擾動點。Exp.No.欄位呈現為對應到完全因素實驗的實驗編號，包括三個輸入因素A、B 和C。觀察圖3-1中可發現到只需執行相當於完全因子設計法的1/3次數的實驗，取樣的●點為立方體每一面取3點使其保持對稱，且每一個因子的水準值個數均相等使保持平衡，相當於對立方體均勻取樣，此即為正交表具有正交性(Orthogonal)、對稱性(Symmetry)和平衡性（Balance）的特性。而且最佳解在其所包圍的立方體之中，因此可以藉由均勻取樣的實驗，來推測全試驗的最佳組合。經由這9次實驗去推論出和

完全因子設計法27次實驗一樣的結果，而這也就是部分因子設計法能以相當於完全因子設計法較少的實驗次數仍可推論出整體最佳解的原因。

表3-1 三因素三水準正交表

Exp.No	Factor		
	A	B	C
1	1	1	1
2	1	2	2
3	1	3	3
4	2	1	2
5	2	2	3
6	2	3	1
7	3	1	3
8	3	2	1
9	3	3	2

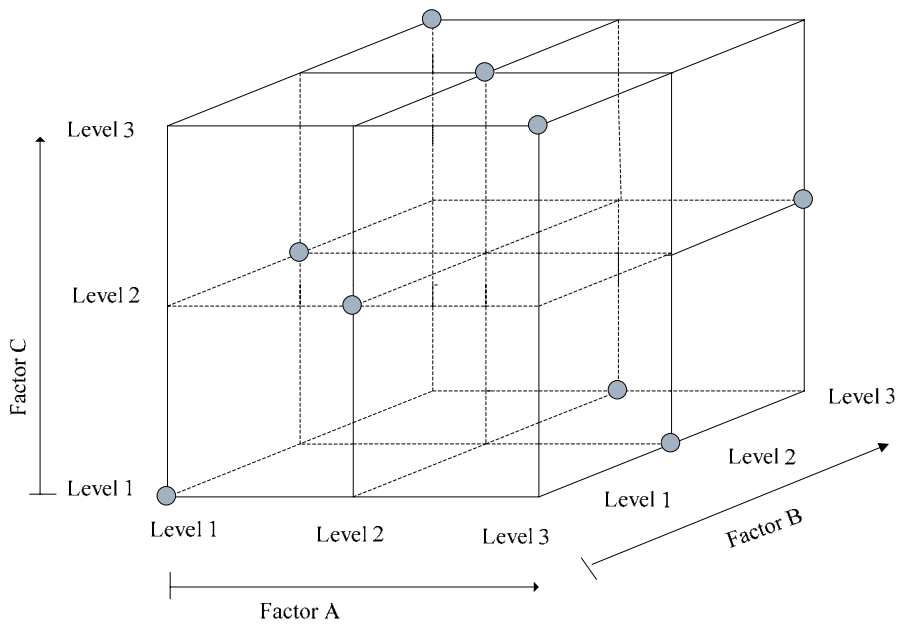


圖3-1 三因素三水準正交表實驗對應至三維搜尋空間的分佈圖

3.1.2 正交表的特性與產生方法

接下來要說明如何產生部分因子設計法的正交表。因為本文只使用到三個水準值，所以接下來所指的正交表皆為指多因子三個水準值的正交表，令因子個數為N，每一個參數有三個水準值，完全因素實驗次數為3^N。若正交表的欄數（或行數）為N，欄數表示所要處理的因子個數，則可以用$(3^{\lceil \log_3(2N+1) \rceil}-1)/2$推論處理$N$個因子所需的實驗次數也就是列數$M$，一個$N$因素三水準的正交表大小為$M$列乘以$N$行，標識為$L_M(Q^N)$，其中$Q$為水準值。舉例來說，當$N$值為5到13的數值，使用$L_{27}(3^{13})$的正交表，但忽略$(M-1)/2-N$行的使用，也就是說只使用到正交表中的Q行。表3.1為一個三因素三水準的正交表，表3-1中的1、2與3代表每一個參數的水準值。三水準正交表的性質可由表中觀察到正交表的特性:

1. 在每一欄中，每一個水準值出現的次數皆為M/Q,如表3-1在每一欄中各個水準值總共出現次數皆為9/3=3次。
2. 在每一欄中，每一個水準值連續出現的次數皆為M/Q^2次，也就是說在表3-1中出現這九組數值(1, 1)、(1, 2)、(1, 3)、(2, 1)、(2, 2)、(2, 3)、(3, 1)、(3, 2)與(3, 3)次數皆為$9/3^2=1$次。而且不管是大小為何其正交表，其每一欄的第一列皆是先填入"1"，如表3-1 的結果。
3. 假如將正交表中的任意兩欄做交換，例如將第第二欄與第三欄互相對調仍然具有以上的性質。
4. 假如將直正交表中任意刪除任意一欄仍然具有正交表的性質。

所以其他大小的正交表也是依此規則類推建立。正交表用來分析個別因素所需實驗次數為M。當使用兩水準正交表時，$N+1 \leqq M \leqq 2N$；而使用三水準正交表時，$2N+1 \leqq M \leqq 6N-3$。兩水準的正交表產生與三水準的正交表產生方式雷同。以下演算法為描述如何產生任何水準的正交表 [8]。

The following algorithm generates the Q-level OA used by OGA with N factors where Q=2, 3. If Q=2, $J = \lceil \log_2(N+1) \rceil$. $[a_{ij}]_{M\times N}$ If Q=3, $J = \lceil \log_3(2N+1) \rceil$. Let the jth column of the OA by a_{ij}. The columns where j=1, 2, $(Q^2-1)/(Q-1)+1$, $(Q^3-1)/(Q-1)+1$, …, $(Q^{J-1}-1)/(Q-1)+1$ are called basic columns, and the others are called nonbasic columns.

Step 1: Construct the basic columns.

for $=1$ To J Do

{ $j=\frac{Q^{k-1}-1}{Q-1}+1$;

for $i=1$ To M Do

$a_{ij}=\left\lfloor \frac{i-1}{Q^{J-k}} \right\rfloor \bmod Q$;} }

Step 2: Construct the nonbasic columns.

for $k=2$ to J do

{ $j=\frac{Q^{k-1}-1}{Q-1}+1$;

for $s=1$ to j-1 do

for $t=1$ to Q-1 do

$a_{j+(s-1)(Q-1)+t}=(a_s\times t+a_j) \bmod Q$; } } }

Step 3: Increase by one for all

$1\le i\le M$ *and* $1\le j\le N$

3.1.3 因素分析

在正交表產生後利用正交表來分析因素對事件影響的效果，稱為因素分析或主效應評估。每一個因素的每一個水準都有主效果，主效果是經由統計該因素的某一水準在不同實驗中，對目標函數的貢獻程度在表3-2 中先產生每一列的輸出 f_i 後，再計算每一欄的主效應(Main Effect)，找出每一欄中，具有較大貢獻度的水準值:

$$S_{jk}=\sum_{i=1}^{M} f_i\times L_k \ ,k=1,.....N \qquad (3\text{-}1)$$

當第k 欄中的第i 列的水準值為j 時，L_k=1；否則，L_k=0。M是實驗的次數。f_i 是正交表的輸出。比較各欄中計算出的三水準值之主效果，當所求問題的是適應函數最大化時，則具較越大主效應的水準值為佳；反之，所求問題為目標函數最小化時，則具較越小主效應的水準值為佳。而每一欄中所找出具最好的主效應所對應到的因素之水準值，即為最佳因素組合。

表3-2 三因素三水準的正交表與因素分析

Exp.No	Factor			Function Cost
	A	B	C	
1	1	1	1	f_1
2	1	2	2	f_2
3	1	3	3	f_3
4	2	1	2	f_3
5	2	2	3	f_5
6	2	3	1	f_6
7	3	1	3	f_7
8	3	2	1	f_8
9	3	3	2	f_9
S_{j1}	S_{11}	S_{21}	S_{31}	
S_{j2}	S_{12}	S_{22}	S_{32}	
S_{j3}	S_{13}	S_{23}	S_{33}	

3.2 增強退火演算法的搜尋能力

本文在增進傳統式模擬退火演算法的搜尋效率，採用結合最正交實驗設計法來分析出在目前維度空間中較佳目標函數值的解，運用其特性對於目前解系統化的產生出一組擾動解，並且從中迅速地推論出可能的最佳解。用此方式來取代傳統是模擬退火演算法每一迭代只對現有解產生一個擾動點的運算，採用亂數隨機產生與測試的方法產生新解。使能夠大幅增進傳統式模擬退火法的求解的效能。並且依據退火的過程，選擇機率較大的

分群中心位置移動，最後會收斂到可能到最佳距離總偏移值。

3.2.1 影響模擬退火演算法效能的因素

在介紹如何增強模擬退火演算法的搜尋能力前，先分析影響模擬退火演算法搜尋的因素，其有四項影響效能的性質如下：

1. 解的表示方法: 在解決問題時，好的表示方式可以使得模擬退火演算法得到更好的解，而對於正交退火演算法，好的表示法更為重要，這是因為正交實驗設計如果變異因素間交互性降低，正交表能夠更準確推測鄰近最佳解，所以如果在設計表示法時特別注意降低交互性，可提升整體求解的效率。
2. 目標函式: 不同問題都有其目標，而如何快速算出目標值，是很重要的，如此可以增進尋找的速度。
3. 在鄰近區域新解的產生機制:產生新解機制通常採用機率分布函數來產生新解，使新解與原來解間的距離不要差太大。如果解與解之間差太大，會使得解在空間中大幅的跳躍，因此會較不容易細部搜尋找到最佳值。但如果解與新解距離太近，在一次迭代中探尋解的空間較小，需要花較多的時間找到全域最佳解，且容易陷入區域最佳解。因此在擾動過程中加入正交實驗法來做產生新解的機制。
4. 冷卻法則: 通常是將目前的溫度乘上一常數，而此常數是介於0與1間接近1的實數。模擬退火演算法的溫度的設定直接影響了接受較差解的機率以及收斂速度。當溫度越高，接受較差解的機率就越高。相反的，則接受機率越低。溫度下降速度往往控制了是否能找到最佳解的原因之一。根據不同的問題需設計不同的冷卻法，好的冷卻法則是有助於搜尋最佳解的效率。

3.2.2 柯西勞倫茲機率分布(Cauchy-Lorentz probability distribution)

以當時點為中心之一為柯西密度機率密度函數 $p(x)$ 定義如下[17]

$$p(x)=\frac{T_t}{\pi(T_t^2+x^2)} \tag{3-2}$$

其中在 T_t 為縮放參數(scale parameter)，對應上式的柯西機率分布函數 $P(x)$ 為

$$P(x)=\frac{1}{\pi}\tan^{-1}(\frac{x}{T_t})+\frac{1}{2} \tag{3-3}$$

柯西機率分佈與高斯分佈之圖形極為相似，但其特點為慢慢地接近x軸卻不收斂於0並且其變異數不存在。

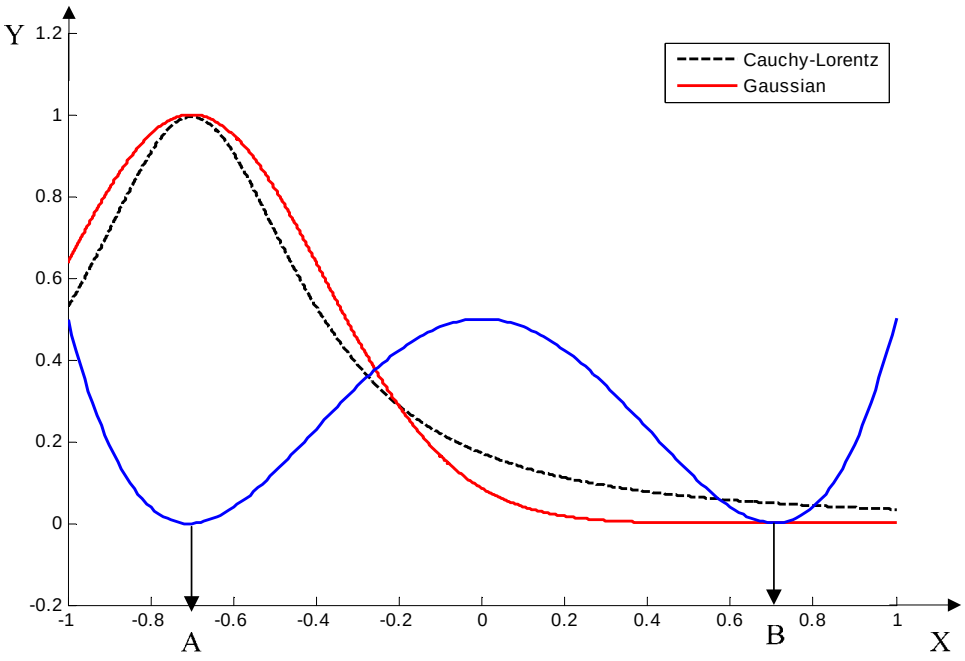


圖3-2 柯西機率分布與高斯分佈之取樣情形

由圖3-2可以觀察到，由於柯西機率分佈會比高斯分佈較慢地接近x軸，因此以柯西機率分佈能夠從A點跳脫到B點附近，比使用高斯機率分佈函數

更有機會跳至較遠的區域點。但相對的，柯西機率分佈函數就比高斯機率分佈函數缺乏微調(fine-tuning)的能力，在本文中則採用動態調整縮放參數的方式解決此一問題。

3.2.3 採用實驗設計法的區域搜尋產生新解

傳統式模擬退火法的擾動的方式是隨機在可行解之空間中隨機產生單個擾動解，為改善演算法的求解效率以及收斂的穩定性。以 i 個因素的目前位置 x_i^0 為出發點利用科西勞倫茲機率分布產生參數之的附近產生兩點 $x_i^1 = x_i^0 - w$ 、 $x_i^2 = x_i^0 - w$ 為擾動點，而其擾動搜尋範圍則以空間維度之上下限所界定。如圖3-4所示。其中 w 為當時溫度之擾動量，U 為是介於0與1間的隨機變數，T_t 為當時之溫度。並且此擾動量會隨著溫度下降而動態產生。

$$w = T_t \tan(\pi(U - \frac{1}{2})) \tag{3-4}$$

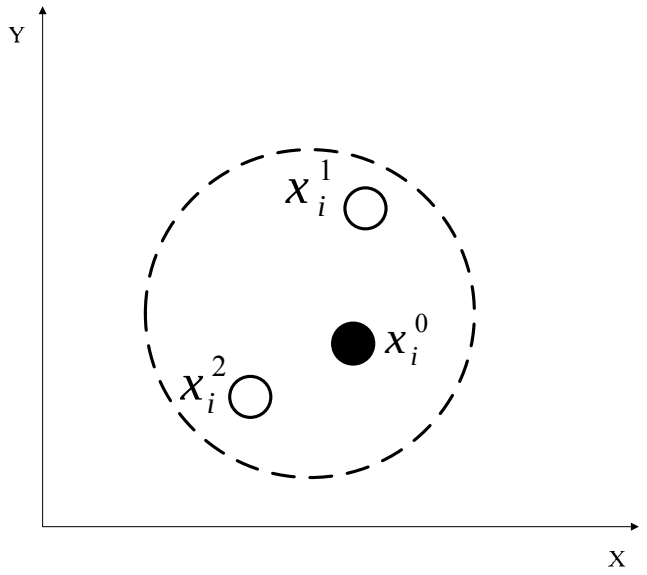


圖3-3 在目前解的區域做擾動示意圖

將所產生之擾動點引入正交表的組合中，產生出M組候選解。經過因素

分析後可以推論出一組最佳的組合，推論步驟如圖3.4所示。因此可以預期系統化取樣鄰近點居並推理最佳移動方向會比隨機測試來得有效率，在較少的評估次數與代數即可搜尋到某局部區域的局部最佳解。利用正交實驗設計法產生新解其步驟詳述如下：

1. 同時考慮k個擾亂步驟，k即是正交表的因素水準值。
2. 建構出適合之正交表，檢查正交表作正交實驗，當因素k 為1表示第t次組合對應為目前點，而k為2時表示第t 次組合對應為擾亂過後的其中一點，以此類推。並計算出各組候選解的評估值。
3. 再由各個因素的主效應評估中找出最佳的組合為推理解，並計算其評估值。
4. 驗證推理解是否與原本解相等。假如是，則在正交表所做的M次實驗中扣掉第一組(因為第一組與原解同)，選剩下最好的一個來當新解候選解。反之，在推理解與正交表所做的M次實驗扣掉第一組解選最好的一組最為最佳的推理解。

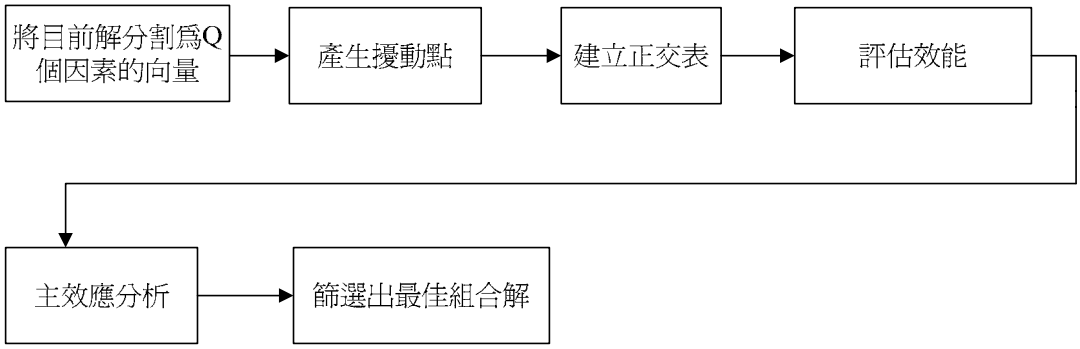


圖3-4 產生新解的擾動機制流程圖

3.3 HSAKM 流程步驟

綜合以上所述，模擬退火演算法藉由正交實驗設計法來加強局部搜尋的能力，可以由部份因素實驗便推論完全因素實驗的結果。並與k-means

結合應用在資料分群上，將距離總偏移值視為目標函數值，已初始k-means演算法的分類結果作為初始解，初始溫度為T_0，針對當時解重複產生新解並計算目標函數值差，決定是否成為下一次迭次的初始解，詳細步驟描述如下。圖3-5為HSAKM流程圖。

步驟1：起始狀態及參數設定：設定初始溫度T ，冷卻率α，群集數K。

步驟1： 產生初始族群:在空間範圍內隨機產生初始中心點,並經由k-means演算法加以分群為初步群集。

步驟3：智慧型擾動挑選出新中心點：

步驟3.1：針對每群集當時之中心點進行擾動，依據中心點之位置產生鄰近的兩點為候選解，計算方式如公式3-4。

步驟3.2：計算正交表中每組候選解之評估值。

步驟3.3：經由因素分析，計算方式如公式3-1。計算出每一個因素的主效應值，並選出每一個因素中較佳的水準值為最後的組合

步驟4：計算目標函數：計算擾動更新中心點之後的J值，與之前尚未進行擾動的J

值做比較，如果為較小則做為此迭代的J'值以及中心點位置。反之，如比較之結果較差則以波茲曼機率來選擇是否接受，最後結果仍然不接受保持原解並回到步驟3。

步驟5：判斷是否滿足終止條件：終止條件為二，一為達到指定之迭代數；二為溫度已達到。若滿足終止條件則進行步驟8，否則進行步驟5。

步驟6：降溫：依據當時的溫度進行降溫，計算方式如公式2-8。

步驟7：產生新群集：透過以上步驟所產生新中心點，再次分群。

步驟8：停止並輸出：判斷是否執到停止條件，若是則停止運算，並輸出該迭代中最佳分群結果及J值。否則回到步驟3。

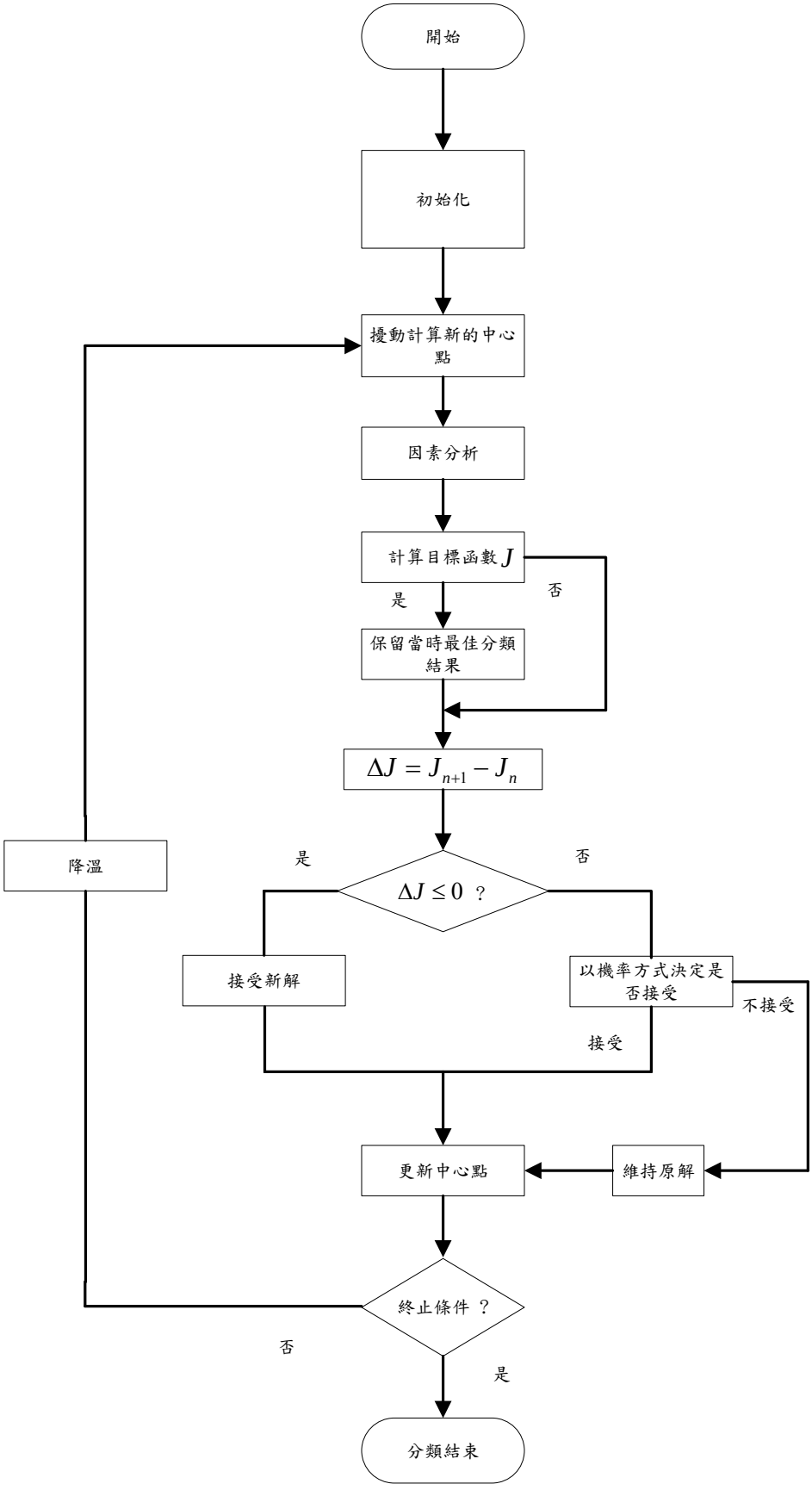


圖3-5　HSAKM流程圖

第四章 實驗結果與分析

為了驗證所提出的方法在效能以及正確率等方面改善，將以人工資料庫與實際資料庫兩種進行模擬實驗分析共分為兩個部份。第一部分測試資料庫說明;第二部份實際模擬比較產生實驗數據，並針對實驗數據進行分析比較。

4.1 實驗資料與參數設計

對於實驗結果之成效評估：採用分群結果之誤判機率、與計算各群集中新點與資料的距離總偏移值來表示，而各成效評估計算方式如式(4-1)、式(4-2)。其中ER為誤判率；A_i與B_i分別為分群前後之資料數；N為總資料數；

J為各群集中新點與資料的距離總偏移值，p_i與c_j為第k個中心點內的資料點與中心點

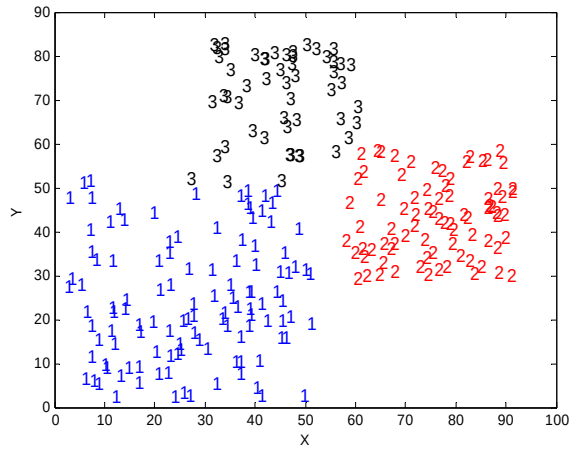
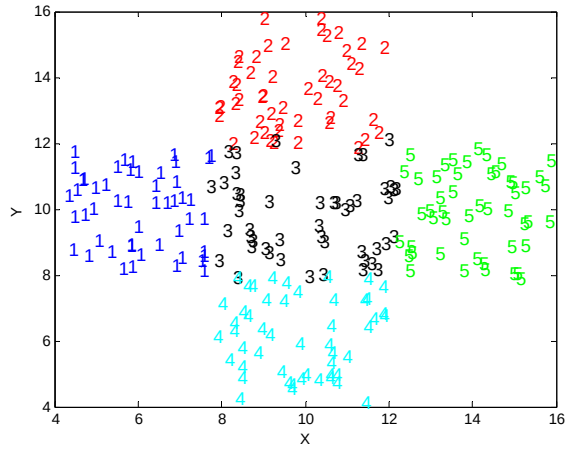
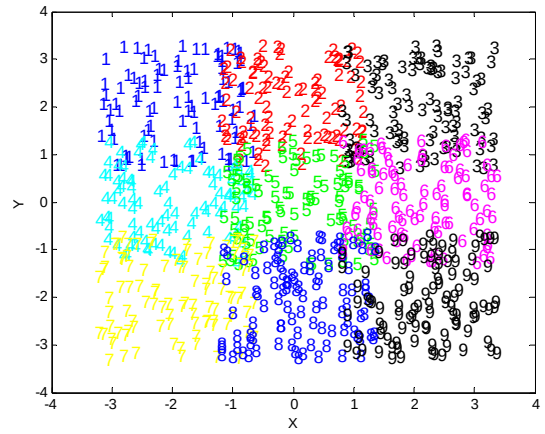
$$ER = \frac{\sum_{i=1}^{K} |A_i - B_i|}{N} \tag{4-1}$$

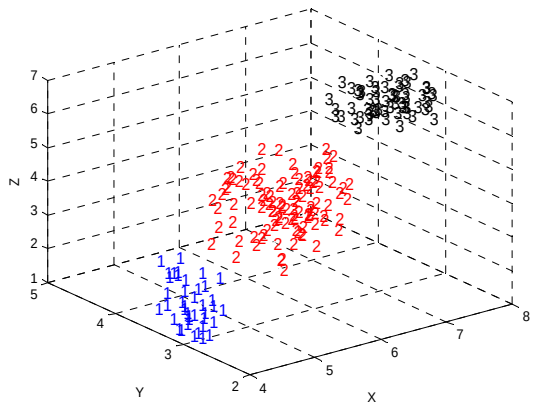
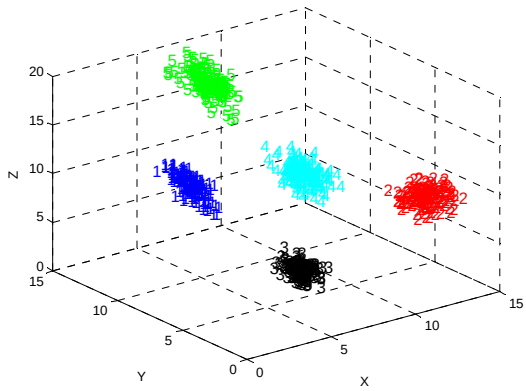
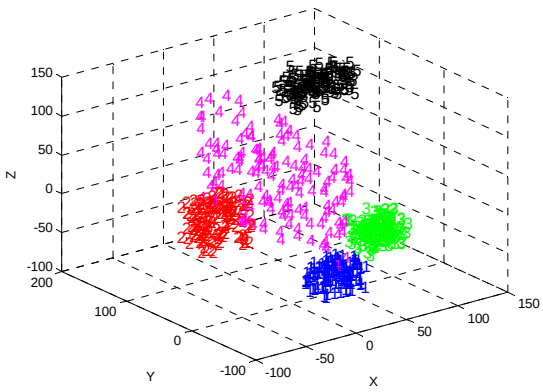
$$J = \sum_{i=1}^{K} \sum_{pi \in C_j} \left\| p_i - c_j \right\|^2 \tag{4-2}$$

4.1.1 人工資料

本節為人工資料庫的介紹，以及相關參數設定實驗所採用之資料，皆由自行撰寫之人工資料產生器所形成。每筆資料皆隨機在範圍空間內產生，按照所需產生不同大小維度的資料庫，以測試分群效果在不同的條件下距離總偏移值函數值與分群正確性之關係，並觀察是否有較佳的分群成果。

表4-1 人工資料庫資料說明

	人工資料庫 1 空間維度:2 群集數目:3 資料數目:120+80+50 群聚大小:不相同 群集重疊:互不重疊
	人工資料庫 2 空間維度:2 群集數目:5 資料數目:50*5 群聚大小:相同 群集重疊:部分重疊
	人工資料庫 3 空間維度:2 群集數目:9 資料數目:100*9 群聚大小:相同 群集重疊:嚴重重疊

	人工資料庫 4 空間維度:3 群集數目:3 資料數目:40+107+63 群聚大小:不相同 群集重疊: 互不重疊
	人工資料庫 5 空間維度:3 群集數目:5 資料數目:150*5 群聚大小:相同 群集重疊:互不重疊
	人工資料庫 6 空間維度:3 群集數目:5 資料數目:150*5 群聚大小:相同 群集重疊:部份重疊

4.1.2 實際資料

實際資料庫則是驗證對於實際且大量的資料是否能成功分群。本文採用四組由UCI Repository of Machine Learning Database [1]所取得的測試資

料。資料庫分別是蝴蝶花(Iris plants)、葡萄酒(Wine)、乳癌(Wisconsin Breast Cancer，簡稱WBC)、避孕器(Contraceptive Method Choice，簡稱CMC)和糖尿病(Pima Indians)，以下將介紹這五個資料庫的相關資訊。表4-2是有關這五個實際資料庫的相關資訊。

表4-2 五組實際資料庫相關資料

	Iris	Wine	WBC	CMC	Pima
資料數目	150	178	683	1473	768
維度	4	13	9	9	8
群體數	3	3	2	3	2

1. 蝴蝶花(Iris Plants)

鳶尾植物資料庫共有150筆四特徵三類的資料，分別為萼片(Sepal) 與花瓣(Petal)的長度(Length)、寬度(Width)4種特徵， Iris Setosa、Versicolour與Virginica這3類鳶尾花的種類所組成的。詳細資料如表4-3與表4-4所示。

表4-3 蝴蝶花資料庫中4種特徵之分佈

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4-4 蝴蝶花資料庫中3類資料之大小與比例

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

2. 葡萄酒(Wine)

葡萄酒資料庫是對義大利同一區域所製作的三種不同品種酒類化學成份分析共有178筆資料，其中包含13種特徵分析，總共分為3種。詳細資料如表4-5與表4-6所示。

表4-5 葡萄酒資料庫中13種特徵之分佈

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	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4-6 葡萄酒資料庫中3類資料之大小與比例

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

3. 乳癌(Wisconsin Breast Cancer)

乳癌資料庫共有699筆資料，其中由於16筆含有遺漏之資訊故而剔除，共整理出683筆資料，其中包含9種特性，共分為良性細胞以(Benign)及惡性細胞(Malignant)兩類所組而成的。詳細資料如表4-7與表4-8所示。

表4-7 乳癌資料庫中9種特性之分佈

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

表4-8 乳癌資料庫中2類資料之大小與比例

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

4. 避孕器(Cmc)

避孕器資料庫共有1473筆資料，其中包含9種特性，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4-9與表4-10所示。

表4-9 避孕器資料庫中9種特性之分佈

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4-10 避孕器資料庫中3類資料之大小與比例

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%

5. 糖尿病(Pima Indians)

糖尿病資料庫來自於美國印地安納的Pima部落，資料來源為女性共有768筆資料，其中包含8種特徵，分類結果分兩類，分別為類別0為沒患有糖尿病以及類別1為患有糖尿病。詳細資料如表4-9與表4-10所示。

表4-11 糖尿病資料庫中8種特性之分佈

屬性	平均值	標準差
Number of times pregnant	3.8	3.4
Plasma glucose concentration	120.9	32.0
Diastolic blood pressure (mm Hg)	69.1	19.4
Triceps skin fold thickness (mm)	20.5	16.0
Hour serum insulin (mu U/ml	79.8	115.2
Body mass index	32.0	7.9
Diabetes pedigree function	0.5	0.3
Age	33.2	11.8

表4-12 糖尿病資料庫中2類資料之大小與比例

群體	所佔資料比數	所佔資料比例
0	500	65.1%
1	268	34.9%

4.2 資料庫操作與結果

以上資料庫之實驗相關條件，總實驗次數為20次，每次實驗迭代次數皆為200次，最後將20次實驗之結果平均來評估分群之績效。

人工資料庫之群集數目為根據人工資料庫之原始設定加以決定，實驗操作k-means與HSAKM所得知結果將以圖表所示以利判斷結果之正確性，其中包含群及數目、指標函數值、已集各個群及中心點所在位置及與資料量等資訊，以下為6種人工資料庫實驗之結果。

‘

人工資料庫 1

表4-13和表4-14為針對工資料庫1分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-13和表4-14分別為k-means演算法以及HSAKM演算法採用相同的初始中心點模擬20次之其中一次之執行結果。

表 4-13 人工資料庫 1 之 k-means 結果

群集數	2		群集分類結果表示圖
指標函數	4569.789394		
各個群集中心點位置與資料量	數量	群集中心點	
	60	15.1217	20.8657
	62	39.1440	30.4192
	128	63.6221	54.6449

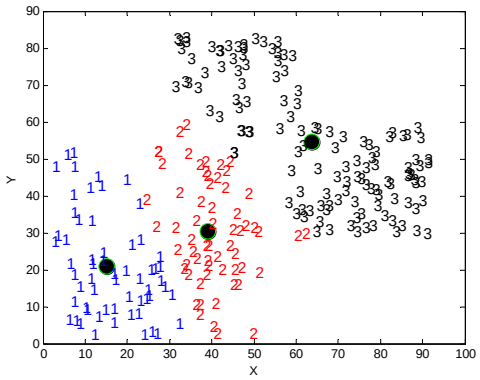
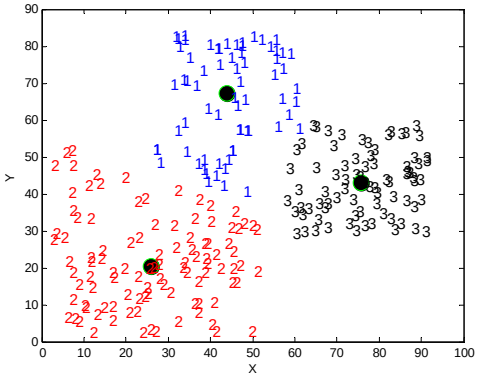


表 4-14 人工資料庫 1 之 HSAKM 結果

群集數	2		群集分類結果表示圖
指標函數	3687.727927		
各個群集中心點位置與資料量	數量	群集中心點	
	118	43.9433	67.1088
	78	25.8769	20.3447
	54	75.6820	43.0034



人工資料庫 2

表4-15和表4-16為針對工資料庫2分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-15和表4-16分別為k-means演算法以及HSAKM演算法採用相同的初始中心點模擬20次之其中一次之執行結果。

表4-15人工資料庫 2 之k-means結果

群集數	5		群集分類結果表示圖
指標函數	382.372019		
各個群集中心點位置與資料量	數量	群集中心點	
	64	6.5258	9.8451
	71	10.1588	6.4980
	35	13.9019	10.4560
	17	14.9283	8.7420
	63	9.9001	13.0477

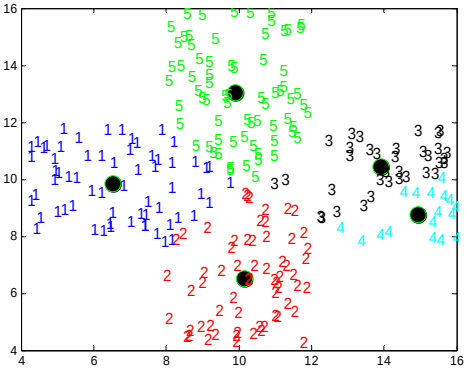
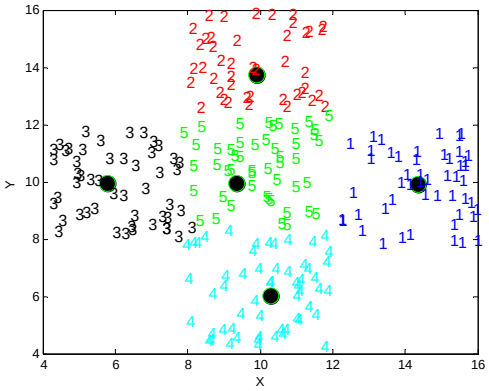


表 4-16 人工資料庫 2 之 HSAKM 結果

群集數	5		群集分類結果表示圖
指標函數	376.911542		
各個群集中心點位置與資料量	數量	群集中心點	
	50	14.4560	10.0380
	44	9.6387	13.9731
	50	5.9108	9.8541
	57	10.1907	6.2610
	49	9.8647	10.6096



人工資料庫 3

表4-17和表4-18為針對工資料庫3分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-17和表4-18分別為k-means演算法以及HSAKM演算法採用相同的初始中心點模擬20次之其中一次之執行結果。

表 4-17 人工資料庫 3 之 k-means 結果

群集數	9		群集分類結果表示圖
指標函數	761.576691		
各個群集中心點位置與資料量	數量	群集中心點	
	102	-1.7629	-0.6571
	117	-1.2865	2.4366
	86	0.7741	1.6426
	71	-0.7997	-0.4227
	117	1.8475	-2.4710
	129	2.4800	1.7353
	87	-2.5137	-1.5679
	102	-0.5817	-2.1552
	89	-2.16184	0.9561

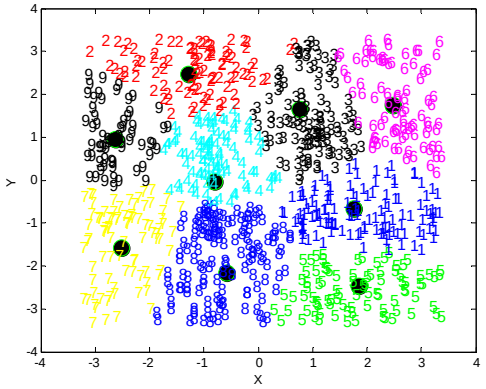
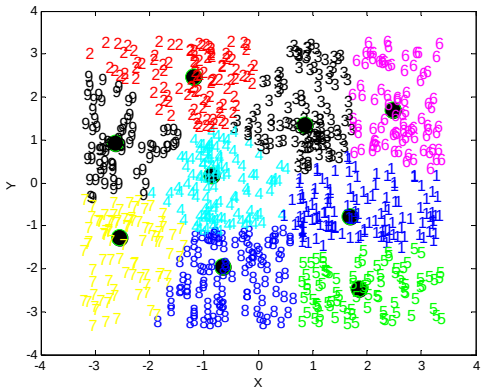


表 4-18 人工資料庫 3 之 HSAKM 結果

群集數	9		群集分類結果表示圖
指標函數	748.976344		
各個群集中心點位置與資料量	數量	群集中心點	
	94	-1.7322	-0.7682-
	116	-1.1692	2.5925
	86	0.9165	1.2805
	102	0.9125	0.5689
	102	1.6985	-2.5373
	126	2.5125	1.6816
	87	-2.5675	-1.2617
	137	-0.7013	-1.6677
	90	-2.7329	0.9558



人工資料庫 4

表4-19和表4-20為針對工資料庫4分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-19為k-means演算法之執行結果;表4-20為HSAKM演算法之執行結果。

表 4-19 人工資料庫 4 之 k-means 結果

群集數	3			群集分類結果表示圖
指標函數	146.141697			
各個群集中心點位置與資料量	數量	群集中心點		
	62	7.2184	3.1693	6.3562
	106	5.5855	3.0673	4.3168
	32	4.3031	3.2223	1.9353

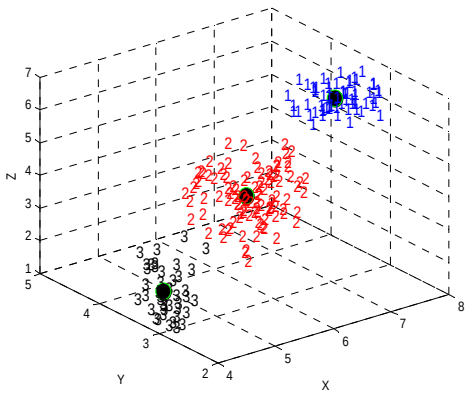
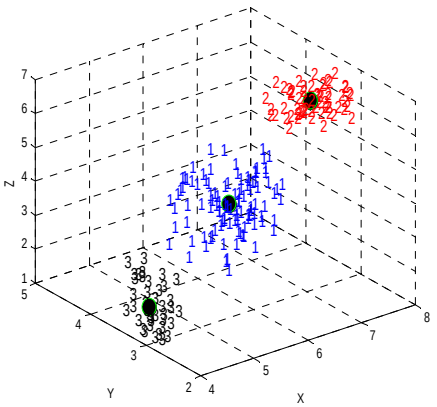


表 4-20 人工資料庫 4 之 HSAKM 結果

群集數	3			群集分類結果表示圖
指標函數	145.499240			
各個群集中心點位置與資料量	數量	群集中心點		
	63	5.6293	3.0753	4.2474
	107	7.2334	3.1642	6.3572
	40	4.2985	3.2342	1.7559



人工資料庫 5

表4-21和表4-22為針對工資料庫5分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-21為K-means演算法之執行結果;表4-22為HSAKM演算法之執行結果。

表 4-21 人工資料庫 5 之 k-means 結果

群集數	5				群集分類結果表示圖
指標函數	1329.683175				
各個群集中心點位置與資料量	數量	群集中心點			
	68	8.8589	15.858	15.858	
	150	9.9893	8.9683	8.9683	
	82	9.0281	14.591	14.591	
	300	4.9942	4.9942	8.9517	
	150	15.001	6.0425	6.0425	

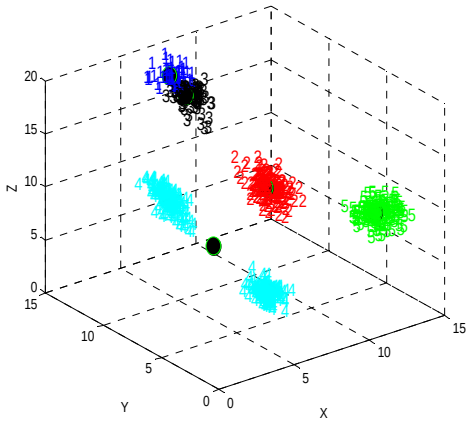
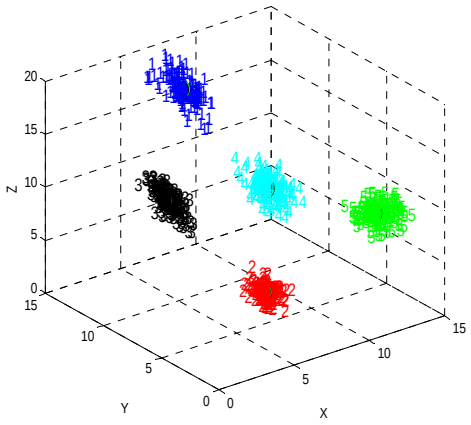


表 4-22 人工資料庫 5 之 HSAKM 結果

群集數	5				群集分類結果表示圖
指標函數	675.171363				
各個群集中心點位置與資料量	數量	群集中心點			
	150	9.0670	14.998	15.014	
	150	6.0588	4.0052	3.9503	
	150	4.0121	9.9664	9.9435	
	150	10.078	8.9604	8.9517	
	150	14.995	6.0525	6.0602	



人工資料庫 6

表4-23和表4-24為針對工資料庫6分別以k-means與HSAKM模擬所呈現出之分群結果，其中表4-23為K-means演算法之執行結果;表4-24為HSAKM演算法之執行結果。

表 4-23 人工資料庫 6 之 k-means 結果

群集數	5			群集分類結果表示圖
指標函數	26849.049374			
各個群集中心點位置與資料量	數量	群集中心點		
	187	76.578	109.81	109.67
	74	79.598	-16.755	-25.830
	76	59.508	-22.353	-16.926
	176	1.1559	-58.110	-27.351
	237	-39.053	50.619	17.203

表 4-24 人工資料庫 6 之 HSAKM 結果

群集數	5			群集分類結果表示圖
指標函數	22432.297007			
各個群集中心點位置與資料量	數量	群集中心點		
	164	86.379	111.05	109.73
	81	66.689	-20.597	-21.577
	181	69.538	-18.937	-21.620
	174	1.8709	-58.527	-35.009
	216	-55.953	41.216	6.6195

兩種方法之群誤差率以及距離總偏移值，如表 4-25 所示。觀察以上之群集分類結果表示，當群集大小相異時，兩種方法皆會發生誤判的情況，距離總偏移值也會提升。雖然兩種方法皆無法找出正確之群集，然而 HSAKM 之誤判率與距離總偏移值依然低於 k-mean。k-mean 在群集之間重疊程度之影響較低時能夠較為順利正確地分群；反之出現誤判之機率越高。k-means 雖然對這 6 種資料庫能快速的分群，但卻時常因為初始中心位置的關係導致無法保證每次皆能分成正確的群集，因而造成相當大的誤差。HSAKM 不受初始點的影響可收斂至最佳中心點位置的區域，並且距離總偏移值皆低於 k-means。因此驗證出 HSAKM 可順利地在人工資料庫進行分群，不論是誤差率、距離總偏移值的結果，皆比 k-means 佳。

表 4-25 對人工資料庫進行平均20次之比較列表

資料庫編號	誤判率		距離總偏移值	
	k-means	HSAKM	k-means	HSAKM
1	12.94%	8.42%	3834.41	3687.73
2	9.62%	4.58%	386.38	377.05
3	45.44%	42.48%	767.01	749.14
4	0.95%	0%	146.14	145.49
5	31.64%	0%	1036.57	675.39
6	21.94%	12.47%	23595.07	22436.86

蝴蝶花資料庫

表4-26為四種方法對Iris Plants資料庫進行在20次迭代後之距離總偏移值值結果。單從最佳距離總偏移值觀察出除了k-means以外皆能收斂於最佳中心點的附近；就平均距離總偏移值而言，HSAKM的平均結果則優於其他三種方法。

圖4-1為其中一次之結果可觀察出以相同的初始中心點，k-means在數次迭代中就能快速收斂，KPSO和KGA與k-means收斂迭代數差異不大，HSAKM雖花較多次數才收斂但可搜尋到最佳中心點附近。綜合以上結果，除了k-means最後落入區域解而導致到達最差距離總偏移值外，其他三種方法皆能順利跳脫。

表4-26 蝴蝶花資料庫平均20次之分群結果

	K-means	KGA	KPSO	HSAKM
最佳距離總偏移值	97.3259	97.2221	96.9521	96.6554
平均距離總偏移值	103.0313	97.2221	98.5096	96.6662
最差距離總偏移值	128.4042	97.2221	122.2788	96.6686

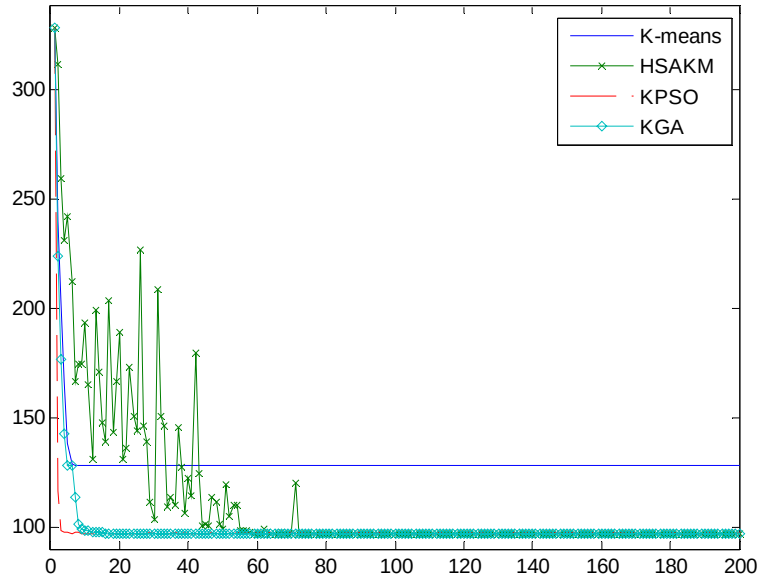


圖4-1 蝴蝶花資料庫其中一次之收斂結果

葡萄酒資料庫

表4-27為四種方法對Wine資料庫進行在20次迭代後之距離總偏移值結果。Wine此資料庫有維度空間較大且第13種維度範圍廣，因此對於此較為複雜的資料庫k-means無法很有效的分群。距離總偏移值觀察出除了k-means以外皆能收斂於最佳中心點的附近；以平均距離總偏移值而言，HSAKM的平均結果優於其他三種方法。

圖4-為其中一次之結果可觀察出以相同的初始中心點，除k-means落入區域解之外其它三種方法皆能快速收斂。KPSO、KGA與HSAKM彼此之收斂迭次數差異不大，而HSAKM卻能搜尋到最佳中心點附近。

表4-27 葡萄酒資料庫平均20次之分群結果

	K-means	KGA	KPSO	HSAKM
最佳距離總偏移值	16555.68	16496.53	16530.53	16292.18
平均距離總偏移值	17983.27	16532.61	16549.45	16292.67
最差距離總偏移值	18436.95	16538.96	16550.45	16300.53

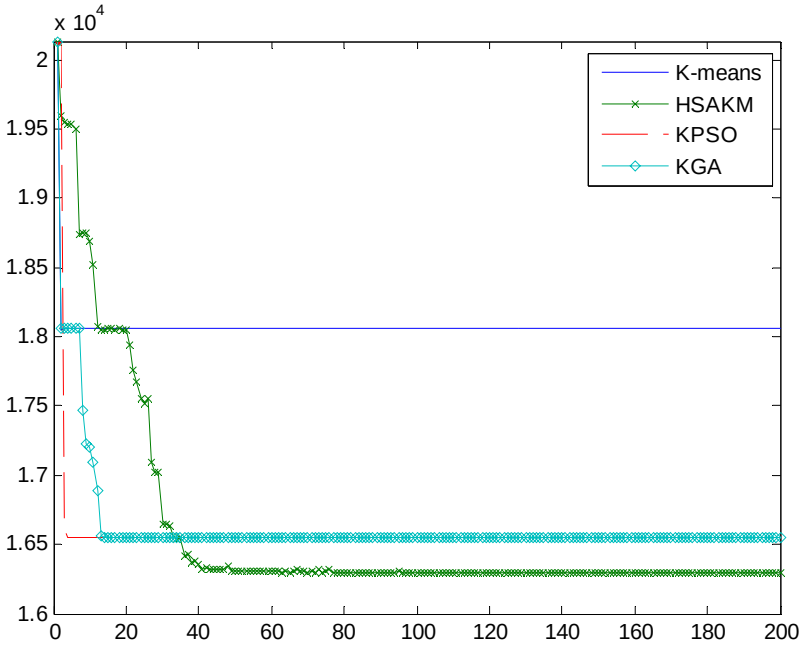


圖4-2 葡萄酒資料庫其中一次之收斂結果

乳癌資料庫

表4-28為四種方法對WBC資料庫進行在20次迭代後之距離總偏移值結果。可觀察出四種方法皆能收斂於最佳中心點的附近，最佳距離總偏移值皆差不多；以平均距離總偏移值而言，HSAKM的平均結果仍然優於其他三種方法。

圖4-3為其中一次之結果可觀察出以相同的初始中心點，除HSAKM其他方法皆能在10次內快速收斂到最佳中心點位置區域範圍附近，雖然HSAKM花較多次數才收斂但比其他方法更能搜尋到最佳中心點附近，並且誤判率也較為低。

表4-28 乳癌資料庫平均20次之分群結果

	K-means	KGA	KPSO	HSAKM
最佳距離總偏移值	2986.96	2984.07	2984.89	2964.38
平均距離總偏移值	2988.13	2985.23	2984.98	2964.38
最差距離總偏移值	2988.43	2988.37	2985.93	2964.38

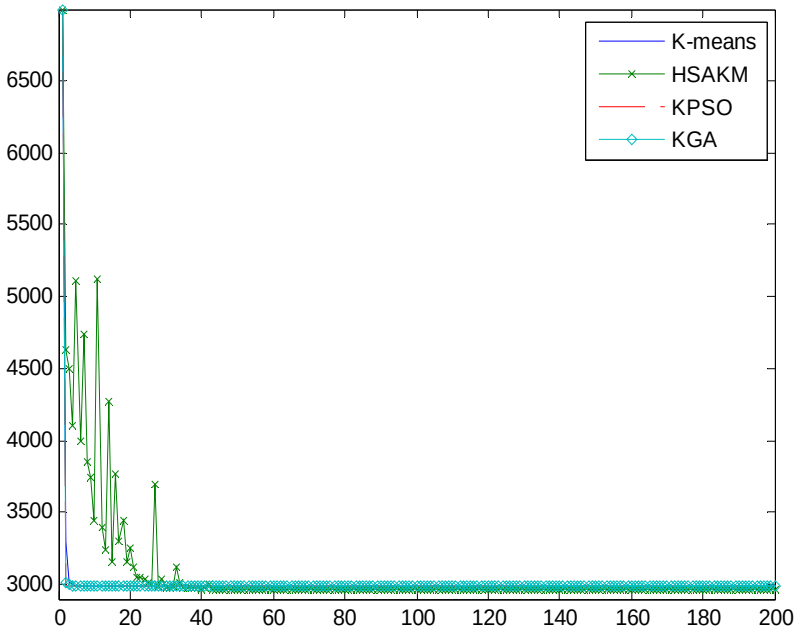


圖4-3 乳癌資料庫其中一次之收斂結果

避孕器資料庫

表4-29為四種方法對CMC資料庫進行在20次迭代後之距離總偏移值結果。可觀察出四種方法最佳距離總偏移值皆差不多；以平均距離總偏移值而言，HSAKM的平均結果仍然優於其他三種方法。

圖4-4為其中一次之結果可觀察出以相同的初始中心點，除HSAKM其他方法皆能在10次內快速收斂到最佳中心點位置區域範圍附近，KPSO和KGA與k-means收斂迭代數差異不大，HSAKM雖花較多次數才收斂但比其他方法更能搜尋到最佳中心點附近。綜合以上結果，除了k-means最後落入區域解而導致到達最差距離總偏移值外，其他三種方法皆能順利跳脫。

表4-29 避孕器資料庫平均20次之分群結果

	K-means	KGA	KPSO	HSAKM
最佳距離總偏移值	5542.33	5542	5538.25	5532.18
平均距離總偏移值	5689.25	5542	5538.86	5532.18
最差距離總偏移值	7040.13	5542	5541.64	5532.18

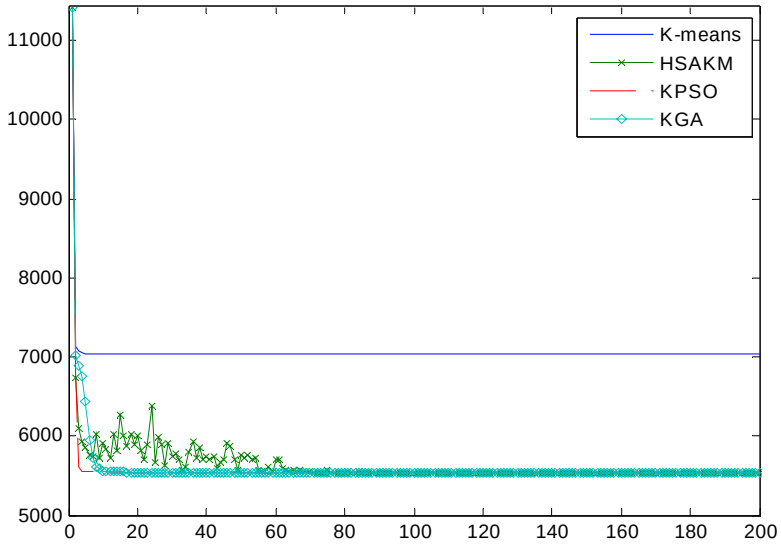


圖4-4 避孕器資料庫其中一次之收斂結果

糖尿病資料庫

由表4-11可看出，由於Pima資料庫的重要性差異很大，而資料中又有大量且不合理的數值，造成資料分佈會有莫大影響，因此測試此資料庫時採取先將資料作正規化後再進行測試。表4-30為四種方法對Pima資料庫進行在20次迭代後之距離總偏移值結果。

圖4-5為其中一次之結果可觀察出以相同的初始中心點，除HSAKM其他方法皆能在10次內快速收斂到最佳中心點位置區域範圍附近，KPSO和KGA與k-means收斂迭代數差異不大，HSAKM雖花較多次數才初步收斂但能夠搜尋到最佳中心點附近。綜合以上結果，k-means仍然有快速收斂的優點以及容易落入區域解的缺點。

表4-30 糖尿病資料庫平均20次之分群結果

	K-means	KGA	KPSO	HSAKM
最佳距離總偏移值	282.7102	282.6346	282.6364	281.9137
平均距離總偏移值	287.1988	282.6483	282.6455	281.9242
最差距離總偏移值	314.6554	282.6619	282.6856	281.9242

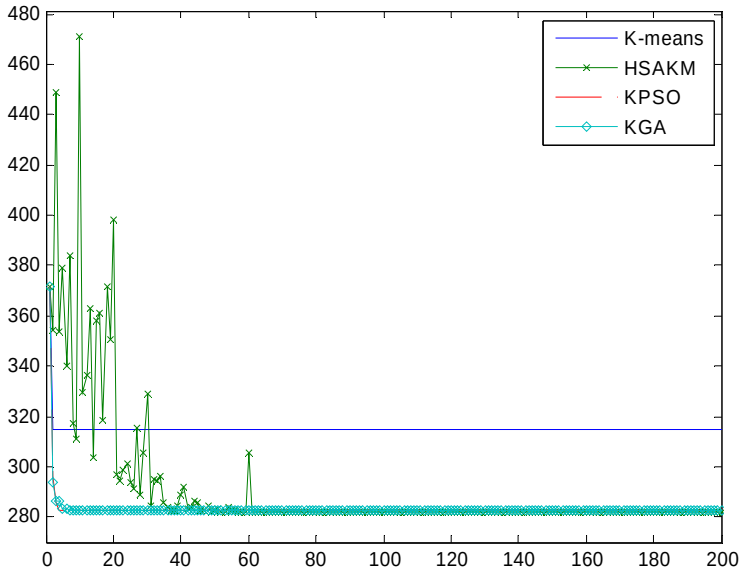


圖4-5 糖尿病資料庫其中一次之收斂結果

誤差率：

表4.31為以四種方法，對五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HSAKM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，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。

表4-31 四種方法對實際資料庫進行平均20次之比較列表

資料庫名稱	誤判率	K-means	KGA	KPSO	HSAKM
Iris	最佳值	10.67%	10.00%	10.00%	10.00%
	平均值	16.83%	10.00%	10.17%	10.28%
Wine	最佳值	29.21%	28.65%	29.21%	28.09%
	平均值	36.89%	29.21%	29.21%	28.81%
WBC	最佳值	3.81%	3.22%	3.37%	3.51%
	平均值	3.92%	3.23%	3.45%	3.51%
CMC	最佳值	54.45%	54.36%	54.33%	54.35%
	平均值	54.51%	54.36%	54.39%	54.37%
Pima	最佳值	33.2%	32.07%	32.07%	32.07%
	平均值	34.47%	32.15%	32.36%	32.12%

第五章 結論與未來研究方向

k-means 演算法是資料探勘、資訊處理應用在進行群集計算時的最常用到的演算法之一，與其他分群方法相較，它具有簡易快速的優點，但也存在一些不足之處。k-means 演算法在為隨機選取初始中心點位置，因而使其效能下降，因此本文採用結合k-Means與模擬退火演算法的方法來處理此問題，產生擾動點的部份則運用正交實驗設計法與擬退火演算法結合，以增進傳統退火演算法的搜尋能力並可以推論出較理想的中心點位置組合。

在第四章中分析比較HSAKM演算法與k-means演算法於多筆資料下之操作結果，就執行效率而言，從實驗結果當中我們可以發現，由於HSAKM運用了退火演算法的擾動跳躍機制，因此降低k-means演算法容易受到初始中心位置的影響。然而，在執行過程中發現，HSAKM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HSAKM雖以群集中心點和資料點之距離來做為評估，雖然有良好的結果，但是誤判率卻偏高。

本論文是以結合k-means與退火演算法的方法來處理資料分群的問題，期望透過HSAKM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. 雖然k-means收斂速度快，但容易陷入區域最佳解並且分類錯誤。
2. 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。
3. HSAKM對於資料分群之距離值，不論最佳或平均值，與本研究其他三種分群法相比，皆為最佳。卻仍存在著不小的誤差。

未來期望能HSAKM與階層式資料分群作結合，希望能有效改善以分割式分群之誤差率缺失。

參考文獻

[1] UCI(University of California，Irvine) Repository of Machine Learning Databases.

Available：ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

[2] W.T.A. Lopes, Madeiro，M.S. Alencar, and B.G. Aguiar Neto, "Simulated Annealing for Robust VQ: Improving Image Transmission through a Fading Channel，" *IEEE Computer Society Press，VI Brazilian Symposium on Neural Networks*，pp.243-248，Rio de Janeiro，Brazil，2000.

[3] J. Ning, S. McClean, and K. Cranley, "3D Reconstruction from Two Orthogonal Views Using Simulated Annealing Approach," *IEEE Computer Society Press，Third International Conference on 3-D Digital Imaging and Modeling*, pp.309-319, 2001.

[4] X.Y. Wang, J.M. Garibaldi, "Simulated Annealing Fuzzy Clustering in Cancer Diagnosis," *Information*，Vol.29,pp.61-70,Number 2005.

[5] N. Metropolis, A.W. Rosenbluth, A. Teller, and E. Teller, "Wquation of State Calculation by Fast computing Machines," *Journal of Chemical Physics*, Vol.21, No. 6, pp. 1087-1092，1953.

[6] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol.220, N0.4589, pp.671-680, 1983.

[7] Q. Zhang and Y.W. Leung, "An orthogonal genetic algorithm for multimedia multicast routine," *IEEE Trans*, *Evolutionary Computation*, col.3, no.1, pp.41-53, Apr. 1998.

[8] Y.W. Leung, Y. Wang, "An Orthogonal Genetic algorithm with Quantization for Global Numerical Optimization," *IEEE Trans.*

Evolutionary Computers. , Vol.5, pp.41-53，Feb. 2001.

[9] J.T. Tsai, J.H Chou, and T.K. Lin, "Tuning the Structure and Parameters of a Neural Network by using Hybrid Taguchi-genetic Algorithm，" *IEEE Trans. Neural Networks*，Vol.17, pp.69-80, Jan. 2006.

[10] S.Y. HO, Y.K. Lin, "OSA: Orthogonal Simulated Annealing Algorithm and Its Application to Designing Mixed H_2/H_∞ Optimal Controllers," *IEEE Trans. Syst.*, Man, Cybern. *A*，Vol. 34，pp.588-600, Sept. 2004.

[11] S.Y Ho , Y.K. Lin, "Orthogonal Simulated Annealing for Floorplanning," *Artificial Intelligence and Application Conference*，pp.469-474, 2002.

[12] D.C. Montgomery, Design and Analysis of Experiments，3rd ed. New York: Wiley, 1991.

[13] C.R. Hicks, Fundamental Concepts in the Design of Experiments，4th ed .TX:Saunders College Publishing, 1993.

[14] T.S. Chen, Y.T. Chen, C.C. Lin，and R.C. Chen, "A combined K-Means and Hierarchical Clustering Method for Improving the Clustering Efficiency of Microarray," *Intelligent Signal Processing and Communications Systems*, pp.405-408, 2005.

[15] M.H. Dunham, Data Ming: Introductory and Advanced Topics, Prentice Hall, 2003.

[16] R.J. Roiger, M.W. Geatz, Data Mining：A Tutorial-Based Primer，Addision Wesley, 2003.

[17] H. Szu, R. Hartley， "Fast simulated annealing," *Phys.Lett.* , Vol.

122，pp.157-162, 1987.

[18] V. Petridis, S. Kazarlis, and A. Bakirtzis, "Varying fitness function in genetic algorithm constrained optimization：The cutting stock and unit commitment problems," *IEEE Trans. Syst.* , Man，Cybern.B, Vol. 28，pp.629-540, Oct.1998.

[19] P.S. Shelokar, V. K. Jayaraman, B. D. Kulkarni, "An ant colony approach for cluster," *Analytica Chimica Acta*, Vol.509, Issue 2, pp. 187-195，May. 2004.

[20] Lizhong Xiao, Zhiqing Shao, Gang Liu, "K-means Algorithm Based on Particle Swarm Optimization Algorithm for Anomaly Intrusion Detection," *Intelligent Control and Automation* Vol.2, pp.5854-5858, June. 2006

[21] Fun Ye, C.Y. Chen, "Alternative KPSO-Clustering Algorithm," *Journal of Science and Engineering*, Vol. 8, No 2, pp.165- 174, 2005.

[22] Ujjwal Maulik, Sanghamitra Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition*, Vol. 33, pp.1455-1465, 2000.

[23] Ujjwal Maulik, Sanghamitra Bandyopadhyay, " Cluster Using Simulated Annealing with Probabilistic Redistribution," *Journal of Pattern Recognition and Artificial Intelligence*, Vol.15, pp.269- 285, 2001.

[24] Zülal Güngör, Alper Ünler, "K-harmonic means data clustering with simulated annealing heuristic," *Applied Mathematics and Computation*, Vol. 184, Issue 2, 15, pp.199-209, Jane. 2007.

[25] Krista Rizman Žalik, "An efficient *k*′-means clustering algorithm，" *Pattern Recognition Letters*, Volume 29, Issue 9, 1, pp. 1385-1391,July

2008.

[26] Adil M. Bagirov, "Modified global *k*-means algorithm for minimum sum-of-squares clustering problems," *Pattern Recognition*, Vol.41, Issue 10, pp. 3192-3199, Oct. 2008.

[27] 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。

個人簡介

姓名：林益民

就讀學校：國立台灣師範大學

系所：工業教育研究所電機電子組

學歷：私立龍華科技大學電機工程學系

專長：智慧型控制

軟體能力：PHP、HTML、C++、VB、MATLAB

研究方向：智慧型控制與資料探勘

著作：

1. “An Intelligent Fuzzy-Neural Diagnostic System for Osteoporosis Risk Assessment,” *International Conference on Computer Science,* Vol. 32, August 2008.
2. “Implementation of Fuzzy B-Spline Membership Function Control Through Mobile Robots,” 2007系統科學與工程會議