行政院國家科學委員會專題研究計畫　成果報告

PSO 仿生物技術最佳化演算法在群聚分析的應用

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-032-027-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：淡江大學電機工程學系(所)

計畫主持人：余繁

計畫參與人員：陳慶逸,李易聰

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 11 月 2 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告 □期中進度報告

PSO 仿生物技術最佳化演算法在群聚分析的應用

計畫類別：■個別型計畫　　□ 整合型計畫
計畫編號：NSC 92－2213－E－032－027
執行期間：　92 年 08 月 01 日至　93 年 07 月 31 日

計畫主持人：余　繁
共同主持人：
計畫參與人員：陳慶逸　　淡江大學電機工程系
　　李易聰　　淡江大學電機工程系

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：淡江大學電機工程學系

中　華　民　國　93　年　　10　月　　28 日

PSO 仿生物技術最佳化演算法在群聚分析的應用

Particle Swarm Optimization algorithm and Its Application to Clustering Analysis

計畫編號：NSC 92－2213－E－032－027
執行期限：91 年 08 月 01 日至 94 年 07 月 31 日

主持人：余　繁　教授　　淡江大學電機工程系
計畫參與人員：陳慶逸　　淡江大學電機工程系
李易聰　　淡江大學電機工程系
e-mail：fyee@mail.tku.edu.tw

一、摘要

在模式識別領域中，群聚分析一直是用來鑑別資料結構相當重要的一項工具，另外也被廣泛的應用在色彩量化(Color quantization)、影像壓縮等問題上，它可以協助使用者從大量的資料中挖掘資料間的結構、簡化資料的複雜性，進而能夠了解並擷取資料背後所隱含的資訊。但是由於真實情況的資料分佈可能是任意的形狀與大小，因此目前並沒有任何一個群聚演繹法則可以解決所有的群聚問題。本計畫提出一個以演化計算為基礎的非監督式群聚分析方法-Alternative KPSO clustering，它結合粒子群最佳化(Particle swarm optimization, PSO)、K-means 演算法以及強健的距離估測公制(metric)來自動估測群聚中心值；比起傳統的K-means及Fuzzy c-means演算法容易發生受限於雜訊環境或困在區域最佳解的情況，我們所提的架構除了可以有效率地搜尋系統之近似最佳解外，也具有更為強健的群聚分析之解題能力。

關鍵詞：群聚分析，粒子群最佳化

Abstract

Clustering analysis aims at discovering groups and identifying interesting distributions and patterns in data sets. It can help the user to distinguish the structure of data and simplify the complexity of data from mass information. A particle swarm optimization-based clustering technique that utilized the principles of K-means algorithm and a new metric, called Alternative KPSO-clustering, is proposed in this article. We attempt to integrate the effectiveness of the K-means algorithm for update centroids, with the capability of PSO to bring it out of the local minima, and we utilize a new matric to replace the Euclidean norm in PSO-clustering procedures. Finally, the effectiveness of the Alternative KPSO-clustering is demonstrated on some artificial and real life data sets.

Keywords: clustering analysis, *pattern recognition, PSO*

二、計畫緣由與目的

群聚分析是模式識別領域中用來鑑別資料結構相當重要的一項工具；這種技術出現在許多地方，例如模式識別中的非監督式學習法則、生物學中的分類法、圖學理論中的切割方法等，其目標就在於在一堆未知資料中粹取具意義的群體。K-means 演算法是應用相當普遍的群聚分析方法[1]，它以圓形分割的方式將資料明確地分類到所屬群聚去，其分群的基礎就在於歐幾里德空間中樣本與其群聚中心誤差平方和的最小化；Fuzzy c-means 演算法則是由 K-means 演算法所擴展而來的[4]，一般認為它具有比 K-means 演算法更佳的特性，目前已經成為群聚分析演算法中最廣為人知及功能最強大的方法。雖然 K-means 演算法的應用極為普遍，但同樣地它們也遭遇到一些問題；由於 K-means 目標函數(Objective function) 是屬於非凸(Not convex)的函數，因此它也可能包含區域最小值。當演化法在執行最小化目標函數的過程時，所得到的解很有可能被困在區域最小值而無法得到全域的最佳解[5]。易言之，K-means 演算法的性能與群聚中心初始值的決定有很大的關聯。另外，在真實的問題中，資料可能存在著各式各樣可能的分佈，並呈

現出不同的密度或尺寸大小，而這往往會影響到群聚分析之成效，一般群聚分析的方法，常會傾向於忽略密度或尺寸較小的群聚，當資料集具有兩個數目差異很大的群集時，若使用降低 MSE 的方法來進行資料分類，大的群集將會被分割開來，而造成分類結果的錯誤；而且由於這些傳統群方法所使用的歐幾里德距離對於雜訊或局外點(outliers)相當的敏感，因此，它們也會在某些特定問題上發生解題能力不夠強健的問題。

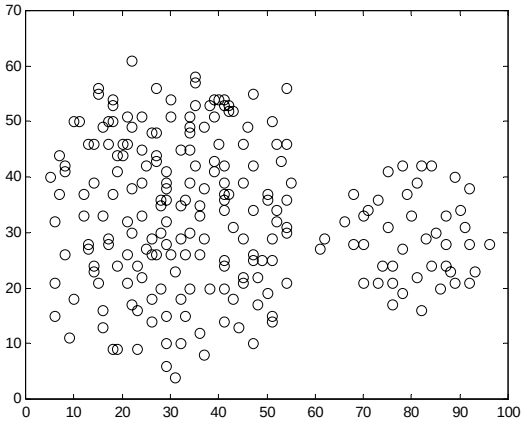

Fig.1 包含一大一小兩個球型群聚之資料集。

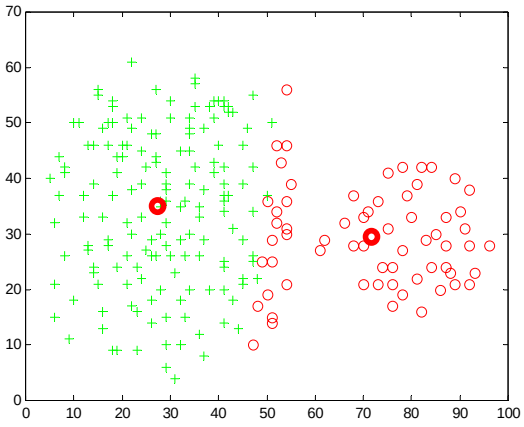

Fig.2 傳統 Fuzzy c-means 對於不同大小群聚問題之錯誤分類結果。

為了有效克服上述的問題，在本計畫中，我們提出了一個新的演化式群聚分法方法，它不受資料於空間中分佈的密度或群聚大小之影響，並具有穩定搜尋向量空間近似最佳解的能力；可解決傳統群聚分析演算法掉入局部最小化解或不夠強健的問題。

三、應用 PSO 演算法之群聚分析技術

1. 粒子群最佳化演算法（PSO）

粒子群最佳化演算法是一種以族群動力學為基礎的最佳化方法[2,3]，它的基本概念來自於社會行為的模擬。在一個社會化的群體中，每一個個體的行為不但會受到其過去經驗和認知的影響，同時也會受到整體社會行為影響。在粒子群最佳化演算法中每一個個體在搜尋空間中各自擁有其方向和速度，並且根據自我過去經驗與群體行為進行機率式的搜尋策略調整。其作法如下：

$$V_{id} = V_{id} + c_1 * rand(\ ) * (P_{id} - X_{id}) + c_2 * rand(\ ) * (P_{gd} - X_{id}) \qquad (1)$$

$$X_{id} = X_{id} + V_{id} \qquad (2)$$

此處 d 是搜尋空間中變數的維度，i 是群體中的個體，V_i 是速度向量，X_i 是位置向量。而 P_i 是個體所經歷過之最佳解位置，P_g 則是個體所處之整個鄰域所記錄的最佳解位置。參數 c_1 以及 c_2 分別是自我認知與社會模式的學習率。

2. Alternative KPSO-clustering

本節中，我們將介紹如何利用 PSO 的搜尋能力來協助我們在 n 維的歐幾里德空間 R^n 中，依資料的相似特性自動地將 N 筆資料區分成 K 類群聚，並分別決定其群聚中心向量。首先，我們令 PSO 演化族群中的每一個個體之編碼值為實數值所構成的字串序列(string sequence)，它代表了 K 個群聚中心。對於 n 維的空間而言，每一個個體的長度是 $K * n$ 個字元。而隨機產生的初始族群也就代表了各組不同的群聚中心向量值[6,7]。

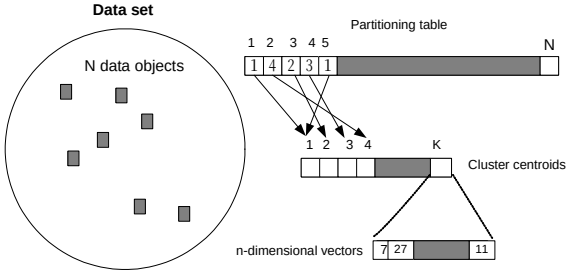

Fig.3 Alternative KPSO-clustering 個體編碼型式示意圖

確定個體的字串編碼之後，再來依下列 PSO 之步驟執行：

Step1) 決定初始族群（population）之個體數目以及相關參數；對第 i 個個體而言，它具有隨機給定的位置 X 以及速度 V。此處，個體的位置 X 值即是我們所欲求得之各群聚的群聚中心值。

Step2) 計算每一個個體之適應函數值。其作法是分別度量資料集中 N 筆資料樣本與 K 個群聚的距離，並依下面條件將樣本歸類至其最接近的群聚，此處我們採用[8]所提出的距離公制來度量相似性：

$$d(x,z) = 1 - \exp(-\beta \|x - z\|^2) \qquad (3)$$

而適應函數則定義如下：

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} \left\{ 1 - \exp(-\beta \|x_i - z_j\|^2) \right\} \qquad (4)$$

$$\text{其中 } \beta = \left(\frac{\sum_{i=1}^{N} \|x_i - \bar{x}\|^2}{N} \right)^{-1}, \quad \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (5)$$

$$fitness = \frac{k}{J + J_o} \qquad (6)$$

此處 k 為預設常數，而 J_o 為一微小的常數值。

Step3) 將每一個個體求得之解與其經驗中記錄的個體最佳解進行比較，若目前之解較之前最佳結果為佳，則以之取代個體最佳解。此外，若目前求得之解優於群體最佳解，則將群體最佳解重設為目前的結果。

Step4)將群體最佳解求得之值以單步(one step)的 K-means 演算法加以取代[9,10]：

$$Z_j^* = \frac{1}{n_j} \sum_{x_i \in C_j} x_i, \quad j = 1,2,...,K \qquad (7)$$

為了節省計算量與維持 PSO 能繼續往搜尋空間之近似最佳解微調的收斂效果，此處我們僅建議在整體迭代過程的前幾次迭代中執行本步驟即可。

Step5) 依 PSO 式子(1)、(2)修改族群中各個體的位置和速度。

Step6) 重覆 step2 至 step5 等步驟，直至滿足所設定的終斷條件後才結束迴圈的執行。

整個運作過程如 Fig.4 所示。

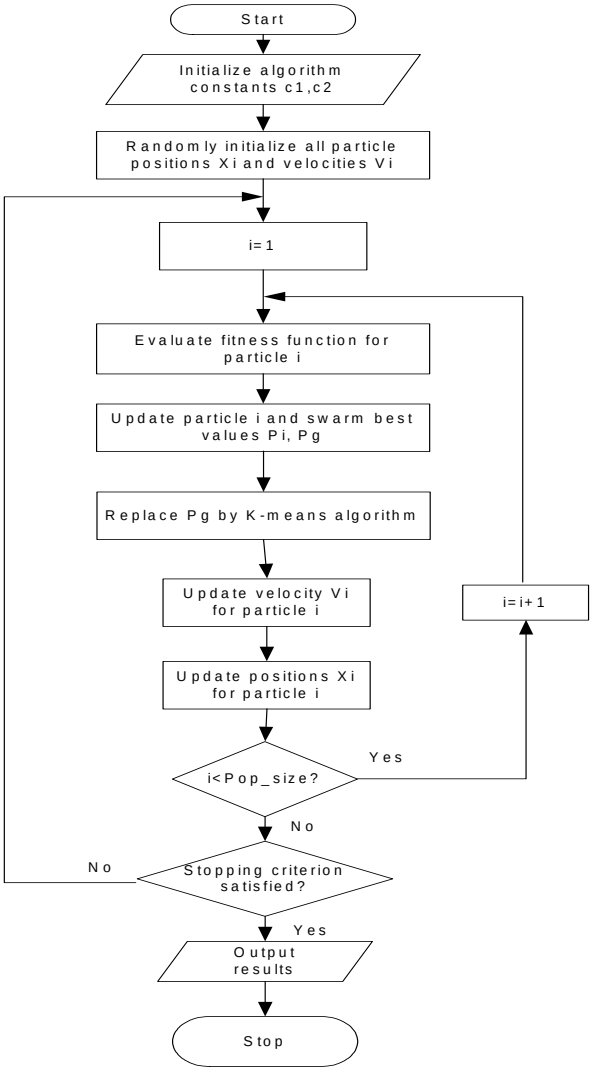

Fig.4 Alternative KPSO-clustering 流程圖

三、實驗結果

為驗證上述所提架構之有效性，我們將分別使用一些不同類別與維度的資料集進行測試：

模擬一:

Example 1. Fig.5 的資料集[10]由 250 筆三維

的資料組成，集合中包含五個球型群聚，而且群聚間彼此相鄰；從分類結果中可以很明顯的看出以 Alternative KPSO-clustering 分類的結果能夠符合我們的預期。

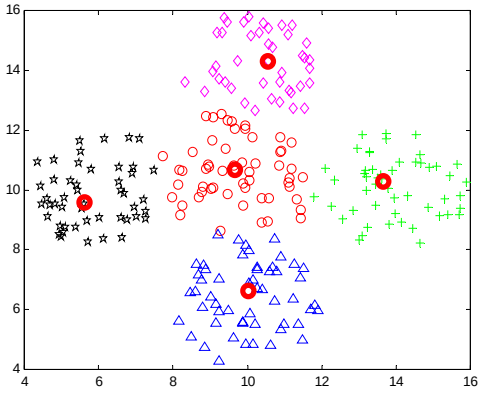

Fig.5 所提架構對於包含五個球型群聚的 250 筆二維資料之分類結果

Example 2. 在 Fig.6 所示的資料集則由 450 筆三維的資料所組成，共分為三類，而且此三類別的資料在立體空間中各個二維平面的映射是彼此交錯的。以 Alternative KPSO-clustering 進行群聚分析可將三個長條狀的群聚分佈資料成功分類出來。

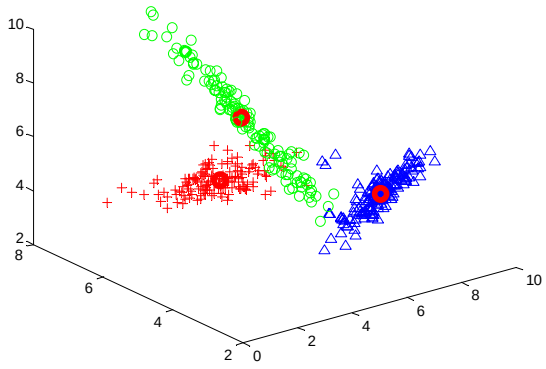

Fig.6 所提架構對於包含三個呈現長條狀群聚分佈的 450 筆三維資料之分類結果

Example 3. 此處我們以 IRIS 資料集當作測試的資料，IRIS 是具有四個維度與三個不同類別的資料，總共有 150 筆資料，每一種類別有 50 筆資料。Fig.7(a)是分別以 IRIS 資料集四個維度來表現所提架構的分類結果；而 Fig.7 (b)則以三維的圖形來表現分類結果。在 Fig.7(c)中顯示了以 PSO 進行群聚分析的過程中，結合 step 4 單步(one step)K-means 運算確實有助於系統的收斂速度。

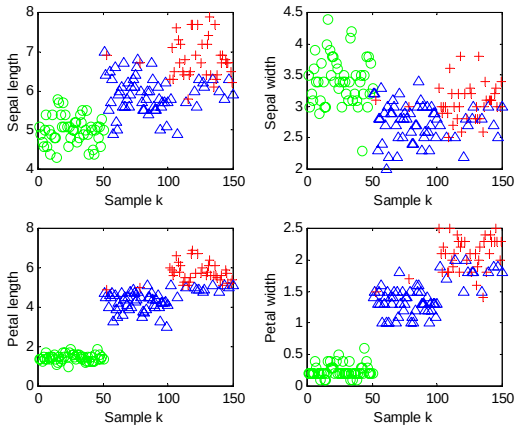

Fig.7(a) 分別以 IRIS 四個維度來表現所提架構的分類結果。

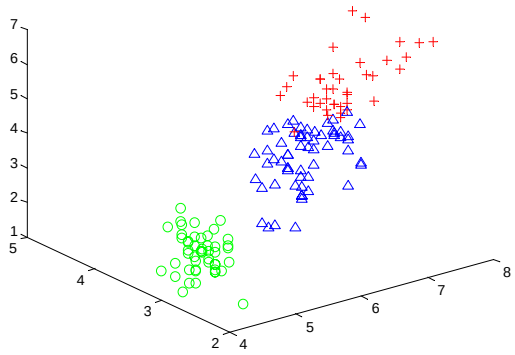

Fig.7(b) 以三維的圖來表現 IRIS 的分類結果

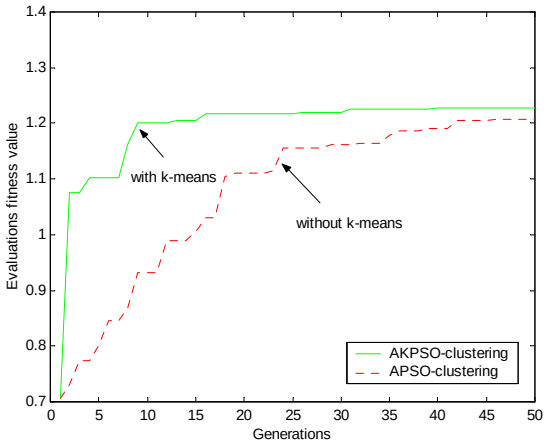

Fig.7(c) IRIS 資料集的收斂情況比較。

模擬二： 在這個模擬中，我們將驗證 Alternative KPSO-clustering 針對不同尺寸群

聚資料的分類能力。

Example 4. 此處我們以二個尺寸和密度有所差異的球型群聚進行群聚分析，Fuzzy c-means 很明顯的會導致不正確的切割結果（Fig.8(b)），而 Alternative KPSO-clustering 卻可正確將兩個群聚分類（Fig.8(a)）。

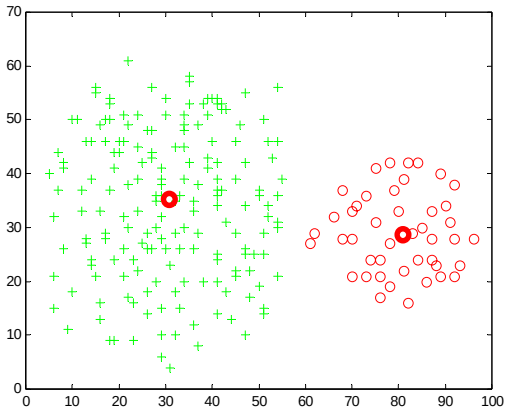

Fig.8(a) Alternative KPSO-clustering 對於不同尺寸大小和密度的群聚分類結果。

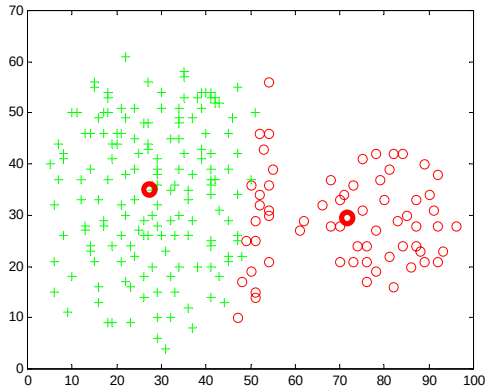

Fig.8(b) Fuzzy c-means 對於不同尺寸大小和密度的群聚分類結果。

Example 5. 此處我們以三個尺寸大小不同的球型群聚進行群聚分析；同樣的，只有 Alternative KPSO-clustering 可以正確的將群聚成功分類。

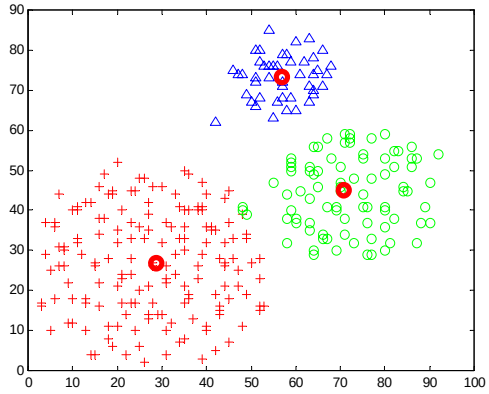

Fig.9(a) Alternative KPSO-clustering 對於三個不同尺寸大小的群聚之分類結果。

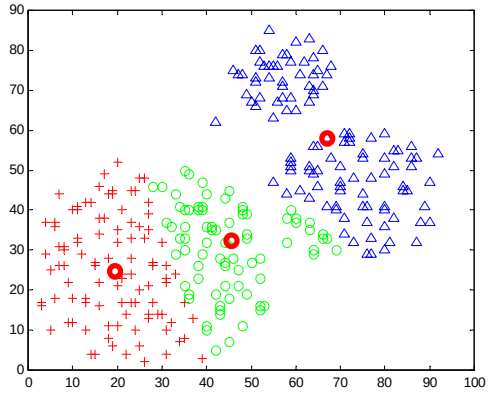

Fig.9(b) Fuzzy c-means 對於三個不同尺寸大小的群聚之分類結果。

模擬三:

Example 6. 我們產生兩筆不同數目的資料在兩個平面上，其中一個在 z=0 平面，而另一個在 z=1 平面。在 Fuzzy c-means 的分類中，其群聚中心並未落在兩個資料平面上(Fig.9(b))，因此發生許多分類錯誤的情況。而所提架構則能得到令人滿意的分類結果(Fig.9(a))。

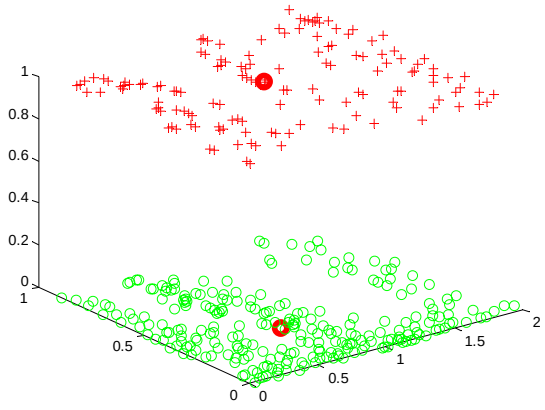

Fig.9(a) Alternative KPSO-clustering 對於二個不同平面資料的分類結果。所求得之群聚中心分別為
(0.9890, 0.5760, **1.0000**),
(0.7073, 0.3129, **0.0000**)。

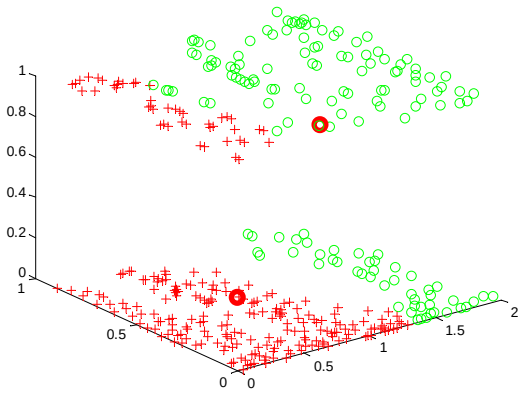

Fig.9(b) Fuzzy c-means 對於二個不同平面資料的分類結果。所求得之群聚中心分別為
(0.5491, 0.3608, **0.1063**),
(1.3690, 0.4874, **0.7590**)

四、結論

本計畫中，我們提出了一個以演化計算為基礎的群聚分析技術，它較不受限於資料在空間中分佈的密度或群聚大小之影響，而且有助於解決傳統 K-means 演算法避開區域最佳解以及受到初始值的影響。從一些資料集的模擬實驗中，我們可驗證所提出的演化式群聚分析技術確實有良好的效能；但相對的，比起傳統群聚分析演算法而言，它也必需付出較大的計算量以及計算時間。

五.參考文獻:

[1] S.Z. Selim, M.A. Ismail, “K-means type algorithms: a generalized convergence theorem and characterization of local optimality,” *IEEE Trans. Pattern Anal. Mach. Intell.6*, pp. 81-87, 1984.

[2] J. Kennedy, R. Eberhart, "Particle Swarm Optimization," *Proc. of IEEE international Conference on Neural Networks (ICNN)*, Vol.IV, pp.1942-1948, Perth, Australia, 1995.

[3] R. Eberhart, J. Kennedy, “A new optimizer using particle swarm theory,” *Proc. 6th Int. Symposium on Micro Machine and Human Science*, pp.39-43, 1995.

[4] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.

[5] S.Z. Selim, M.A. Ismail, “K-means type algorithms: a generalized convergence theorem and characterization of local optimality,” *IEEE Trans. Pattern Anal. Mach. Intell.6*, pp. 81-87, 1984.

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

[7] Ching-Yi Chen and Fun Ye , “Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis,” *IEEE ICNSC 2004*, Taipei, Taiwan, March, 2004, pp.789-794.

[8] Kuo-Lung Wu, Miin-Shen Yang, “Alternative c-means clustering algorithms,” *Pattern Recognition, vol. 35*, pp. 2267-2278, 2002.

[9] Ching-Yi Chen and Fun Ye ,“K-means Algorithm Based on Particle Swarm Optimization,” *2003 International Conference on Informatics, Cybernetics, and Systems*, I-Shou University, Taiwan, ROC. Dec, 2003, pp.1470-1475.

[10] S. Bandyopadhyay and U. Maulik, ``Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification," *Pattern Recognition*, vol.35, pp. 1197-1208, 2002.

附件一

已投稿至 ICNSC2004 研討會並且獲得接受及於會議中發表論文

Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis

Ching-Yi Chen, Fun Ye

2004 INTERNATIONAL CONFERENCE ON NETWORKING, SENSING, AND CONTROL
March 21-23, 2004, Taipei, Taiwan

附件二

已投稿至 ICICS2003 研討會並且獲得接受及於會議中發表論文

K-means Algorithm Based on Particle Swarm Optimization

Ching-Yi Chen, Fun Ye

2003 International Conference on Informatics, Cybernetics and Systems
December 14-16, 2003, Kaohsiung, Taiwan, ROC