行政院國家科學委員會專題研究計畫　成果報告

演化式半監督式群聚分析演算法之研究

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-032-023-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學電機工程學系

計畫主持人：余繁

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 31 日

演化式半監督式群聚分析演算法之研究

Evolutionary-based Clustering with Supervision

計畫編號：NSC 94-2213-E-032-023
執行期限：94 年 08 月 01 日至 95 年 07 月 31 日

主持人：余繁教授　　淡江大學電機工程系
e-mail：fyee@mail.tku.edu.tw

一、摘要

工程應用上關於監督式學習與非監督學習這兩者在其出發點或執行策略上是有所差異的；以群聚分析為主要應用的監督式學習法則是在已有預設的資訊或知識的情況下，利用演算法去發掘問題的核心所在，而非監督式學習法則卻是在不預設立場的情況下，讓系統從大量的資料中去發現其隱含的有用信息。以影像分割為例，它常被定位在典型非監督式法則的應用範疇，影像中的像素通常視為色彩空間中的獨立物件；當影像資料所處的時域或頻域空間中之資料點具有良好的分隔特性時，此時非監督式的群聚分析演算法在物件分類上便可以有很好的表現；不過若是其資料群聚間發生重疊現象時，非監督式的切割法則很難得到正確的邊界；在這種情形下，若是可以採用監督式的分類方法(supervised classification)將會有效得多，但是監督式的分類方法所面臨最大的困難卻是我們需要大量經過人工方式加標(labeled)的資料才能進行訓練。本計畫中我們發展以演化式群聚分析演算法為基礎的半監督式學習法則(Semi-supervised learning)，它的作法是在傳統非監督式法則的目標函數中併入了了少量的加標訓練資料之成本(cost)評估，以期能夠同時補強二種演算法的一些特定缺點。

關鍵詞：半監督式學習、演化式計算

二、簡介

半監督式學習法則(Semi-supervised learning)試圖整合監督式與非監督式學習法則的利處，它採用一些初始知識結合非監督式群聚分析進行訓練，這些初始的知識包含了切割類別數目以及每一類別中少量的加標資料，這種方法已經成功的應用在群聚分析[1]、SVM分類器[2]以及影像切割上[3]。舉K-Means演算法為例，若我們欲將資料群集 $X=\{x_1,x_2,...,x_N\}$ 作 C 個合適的區隔(其中 $2\le C\le N-1$)，其目標函數可表示為

$$J_{kmeans}=\sum_{i=1}^{C}\sum_{x_j\in C_i}\left\|x_j-z_i\right\|^2 \qquad (1)$$

但當我們考慮訓練樣本配對的限制因素時，該目標函數法則便無法直接適用。此處我們再假設 M 為"必須連結(must-link)"集合，$(x_i,x_j)\in M$ 代表 x_i 和 x_j 應屬於同一個群聚；而 K 為一個"不能連結(cannot-link)"集合，$(x_i,x_j)\in K$ 代表 x_i 和 x_j 應該分屬於不同的群聚。此時半監督式的K-means演算法之目標函數可修正為[4]

$$J_{pckmeans}=\sum_{i=1}^{C}\sum_{x_j\in C_i}\left\|x_j-z_i\right\|^2+ \\ +\sum_{(x_i,x_j)\in M}w_{ij}\mathrm{I}[l_i\neq l_j]+\sum_{(x_i,x_j)\in K}\overline{w}_{ij}\mathrm{I}[l_i=l_j] \qquad (2)$$

式子中 I 設定為一指示函數，$I[true]=1$ 以及 $I[false]=0$。以模糊 C-means 方式實現的半監督式學習法則是屬於另類的群聚分析型式[5]，它是模糊 C-means 演算法的擴展。在半監督式群聚分析(Semi-supervised clustering)的處理工作中，我們使用"必須連結(must-link)"以及"不能連結(cannot-link)"這兩者樣本與類別間的限制作為成本(cost)的最佳化評估條件之一。有名的 FCM 目標函數如下：

$$J_{FCM}=\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m\left\|x_j-z_i\right\|^2 \qquad (3)$$

此處 $m>1$，使目標函數 J_{FCM} 最小化的必要條件之(μ,z)如下：

$$\mu_{ij} = \frac{\left[1/\left\|x_j - z_i\right\|^2\right]^{1/(m-1)}}{\sum_{k=1}^{c}\left[1/\left\|x_j - z_k\right\|^2\right]^{1/(m-1)}} \quad (4)$$

以及

$$z_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m} \quad (5)$$

W. Pedrycz 等人[6]根據上述 Fuzzy c-means 的目標函數更新如下：

$$J_{SSFCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m\left\|x_j - z_i\right\|^2 + \alpha\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij} - f_{ij}b_j)^m\left\|x_j - z_i\right\|^2 \quad (6)$$

在這裡，α 是保持監督以及非監督式資料之間平衡的尺度因子，而 f_{ij} 則是加標資料 j 和群聚 i 的歸屬度， b_j 則是用來區分加標與未加標資料之間的布林變數，若是布林變數 b_j 為0，則此目標函數將轉成標準型式的Fuzzy c-means架構。而切割矩陣變成

$$\mu_{ij} = \frac{1}{1+\alpha}\left\{\frac{1+(1-b_j\sum_{l=1}^{c} f_{lk})}{\sum_{l=1}^{c}\left[\left\|x_j - z_i\right\|^2 / \left\|x_j - z_l\right\|^2\right]} + \alpha f_{ij}b_j\right\} \quad (7)$$

三、演化式半監督式群聚分析演算法則

半監督式群聚分析(Semi-Supervised clustering)使用少量的加標樣本以強制指定其歸屬群聚外，它也應在目標函數中加上未能分類的成本項，以便同時進行群聚間緊密程度與群聚資料同質性的估測；也就是說我們期望得到最小化的目標函數應為群聚分佈 (Cluster_Dispersion) 與群聚不純度(Cluster_Impurity)的線性組合。當目標函數中去除 Cluster_Impurity 這一項指標時，其結果即為單純的非監督式架構；反之，當我們僅考慮 Cluster_impurity 這一項指標時，它便轉變成試圖最小化錯誤分類結果的監督式架構了。

在 PSO 演算法中，族群中的每一個個體在解空間(solution space)的移動位置是根據自我過去最佳經驗(pbest_p)與群體最佳行為($gbest$)進行機率式的修正調整，其式子可描述如下[7]:

$$V_{p,d}^{t+1} = \tau\cdot V_{p,d}^{t} + c_1 * rand() * (pbest_{p,d}^{t} - X_{p,d}^{t}) + c_2 * rand() * (gbest_{d}^{t} - X_{p,d}^{t}) \quad (8)$$

此處 X_p 是個體的位置向量（解集合），p 是指向族群中某一個個體的指標，d 是向量維度,τ 用來決定迭代過程中移動速度遞減的幅度，t 意指現在狀態而 $t+1$ 代表下一次迭代的狀態，c_1 和 c_2 則是控制個體受到自我過去經驗(pbest_p)與群體最佳解($gbest$)比率的參數。當個體的移動速度 V_p 確定之後，其位置向量 X_p 則可據以更新為

$$X_{p,d}^{t+1} = X_{p,d}^{t} + V_{p,d}^{t+1}. \quad (9)$$

接下來我們將敘述所建立之演化式半監督式架構之步驟如下：

Step1) 隨機產生初始群中所有個體的位置向量 X_p 以及移動速度 V_p，此處位置向量 $X_p = \{c_{p,j}, j = 1,2,...,K\}$ 包含 K 個群聚中心 $c_{p,j}$,而且 $c_{p,j} = (c_{p,j1}, c_{p,j2},..., c_{p,jn})$ 均為 n 維的向量。

Step2) 計算每一個個體之適應函數值，此處我們設計適應函數 F 如下：

$$F = k_o\cdot(\gamma + [\sum_{i=1}^{K}\sum_{x_j\in C_i}\left\|x_j - z_i\right\|^2 - s_1\cdot\sum I[corr(x_i,c_j)] + s_2\cdot\sum I[error(x_i,c_j)]])^{-1} \quad (10)$$

其中 k_o、γ 是正實數，s_1、s_2 是可調係數值；其中 $corr(x_i,c_j)$ 代表加標資料中的 x_i 分類正確($x_i\in C_j$)，而 $error(x_i,c_j)$ 則表示加標資料 x_i 最後的分類結果錯誤。I 為指示函數，當 $I[true]=1$ 而且 $I[false]=0$。

Step3) 將每一個個體目前求得之適應函數值與其所記憶之最佳適應函數值進行比較，若目前之適應函數值較之前最佳結果為佳，則以目前位置取代個體記憶之最佳所在位置，以目前適應函數值取代個體記憶之最佳值：

$$\text{pbest}_p^{t+1} = \begin{cases} X_p^{t+1} & \text{if } F(X_p^{t+1}) \ge F(\text{pbest}_p^{t}) \\ \text{pbest}_p^{t} & \text{if } F(X_p^{t+1}) < F(\text{pbest}_p^{t}) \end{cases} \quad (11)$$

Step4) 比較個體最佳解所求得的適應函數是否優於群體所記憶的最佳值，若判斷條件成立則將群體所記憶之最佳位置與最佳值重設為目前的結果，反之群體最佳解維持原先狀態：

$$gbest^{t+1} = \begin{cases} \text{pbest}_p^{t+1} & \text{if } F(\text{pbest}_p^{t+1}) \ge F(\text{gbest}^{t}) \\ \text{gbest}^{t} & \text{if } F(\text{pbest}_p^{t+1}) < F(\text{gbest}^{t}) \end{cases} \quad (12)$$

Step5) 根據 Equation 8 和 Equation 9 調整所有個體移動的速度與位置。

Step6) 重覆 Step2)-Step5) 直至達到預設的

迭代次數為止。

三、實驗結果

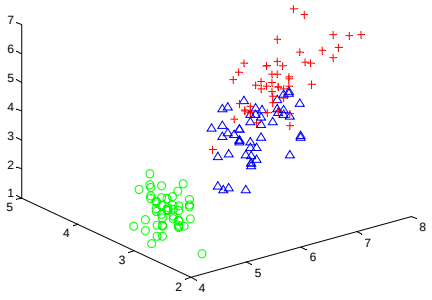
為驗證上述所提架構之有效性，我們以群聚分析領域中最具代表性的 Iris data 進行測試。Iris data 共包含 150 筆資料，區分為 Iris setosa, Iris versicolor 以及 Iris virginica 三個類別，每一個類別中均有 50 筆資料，其中有兩類的資料分佈區域相互重疊，因此以單純的非監督式法則如 K-means 或 FCM 進行分類時，所得到的錯誤分類結果均在 16 筆以上。本實驗中，我們預設 PSO 之參數 $c_1 = c_2 = 1.5$，而且為了達到較佳的收斂效果，我們在每次迭代過程中令 τ=0.75，以逐漸調整個體移動的速度值。適應函數中 $k_o = 50$，γ=20，$s_1 = 0.25$ 以及 $s_2 = 1.8$。

Table 1 為所提方法針對 Iris data 得到的分類結果；我們從 Iris versicolor 和 Iris virginica 二個類別中分別抽取 100 筆, 34 筆和 25 筆進行加標，結果發現其分類錯誤率確實與加標資料的學習有關，而且從表中我們還可看到所有 150 筆資料與 centroids 之 mean-square-error (*MSE*)隨著分類錯誤數而升，這說明了 Iris data 的不同類別群聚之間確有混淆情況；也因此，未必 MSE 誤差較小者代表有良好的分類效果。

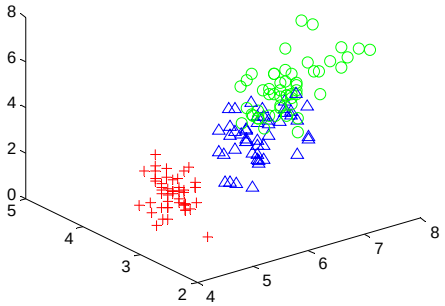
Fig.1 和 Fig. 2 分別以 3D plot 以及 one-dimensional 的方式來呈現資料的分類結果，在 Fig.1(a)中，真實的 Iris data 分佈情況也說明了 Iris versicolor 和 Iris virginica 二個類別之間混淆的現象，圖示的分類結果可以協助我們比較各種不同數目加標資料對於分類結果的影響。

Table 1 所提方法之分類結果

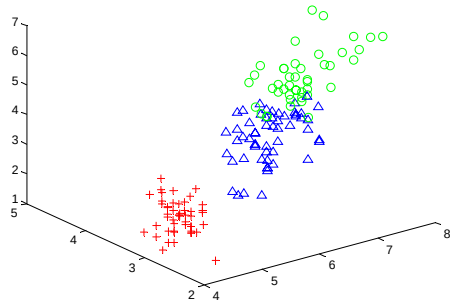
	加標資料 = 100		加標資料 = 34		加標資料 = 25	
分類結果	正確	錯誤	正確	錯誤	正確	錯誤
	146	4	140	10	136	14
MSE	102.6592		97.9849		97.6598	
centroids	(4.9608, 3.4893, 1.4179, 0.3063; 6.4106, 2.8230, 5.5318, 2.1189; 5.9869, 2.7984, 4.0956, 1.2498)		(5.0231, 3.4036, 1.4483, 0.2672; 6.5922, 2.9876, 5.6266, 2.0537; 5.9400, 2.7542, 4.2904, 1.3214)		(5.9287, 2.7601, 4.3227, 1.3620; 4.9930, 3.4527, 1.4531, 0.2657; 6.7542, 2.9939, 5.6943, 2.0385)	



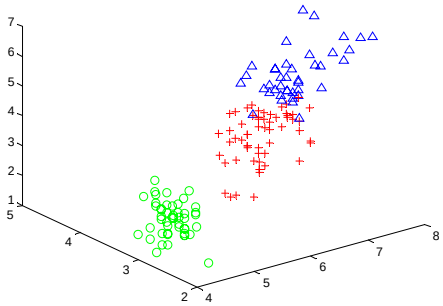
(a)真實的分佈 (3D plot)



(b) labeled data = 100 (3D plot)

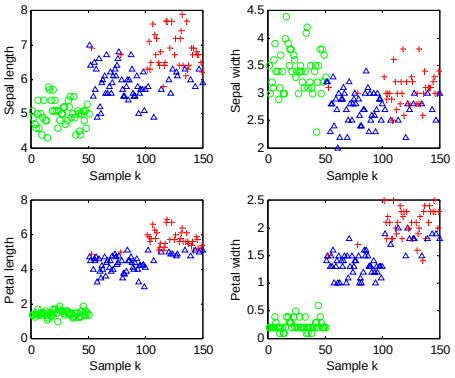


(c) labeled data = 34 (3D plot)

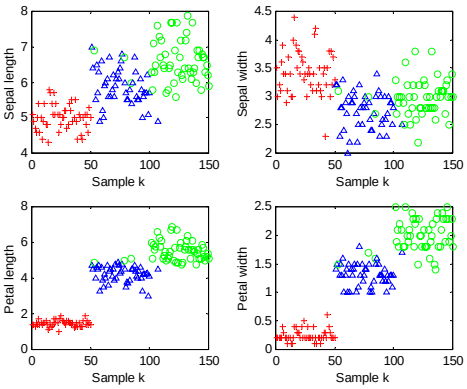


(d) labeled data = 25 (3D plot)

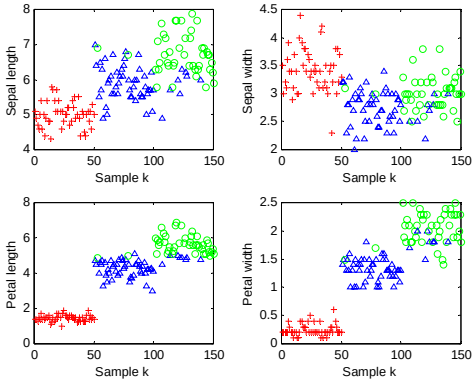
Fig. 1 IRIS data 的 3D plot



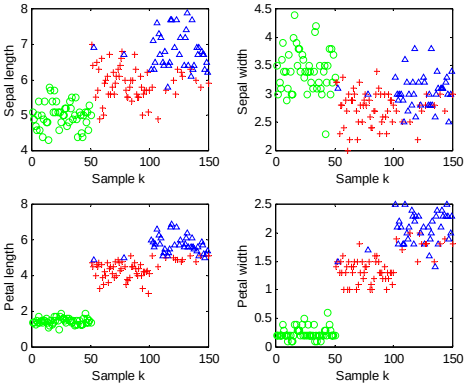
(a)真實的分佈



(b) labeled data = 100



(c) labeled data = 34



(d) labeled data = 25

Fig. 2 Iris data 在各個維度的資料分佈情形

四、結論

本計畫中，我們建立一個以 PSO 為基礎的演化式半監督式群聚分析架構，它可透過自我演化的過程自動找尋超高維度最小化目標函數的近似最佳解，並應用在資料的分類等工作上。PSO 是以模擬鳥類或魚類覓食的群體行為來達到系統自我演化的目的，可以自動找尋向量空間中的最佳解；然而其搜尋過程並非是漫無目的的，而是根據所面臨的問題轉化之適應函數來決定其搜尋的方向。適應函數數值代表著解集合對於外在環境的適應度，適應度愈高表示在此環

境下此組解集合愈佳。在我們所建立的半監督式學習架構中，合併加標與未加標訓練資料的成本評估指標即將轉化成生物群體（不同的目標函數最小化策略）行動中每一個個體（其中一個目標函數最小化策略）移動方向的適應函數評估值。本計畫之執行除了有助於建立半監督式學習法則之新架構外，也可進一步將之推廣至其他相關應用的研究工作。

五.參考文獻:

[1] A. Amar, N. T. Labzour, and A. M. Bensaid, "Semi-supervised hierarchical clustering algorithms." In *6th Scandinavian conference on Artificial Intelligence*(*SCAI'97*), pp. 232-239, Helsinki, Finland, 1997.

[2] K. P. Bennett and A. Demiriz. "Semi-supervised support vector machines." *Advances in Neural Information processing Systems*, pp.368-374, MIT Press, Cambridge, MA, 1998.

[3] A. M. Bensaid, L. O. Hall, J. C. Bezdek, and L. P. Clarke, "Partially supervised clustering for image segmentation," *Pattern Recognition*, 29(5), pp. 859-871, 1996.

[4] Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney, "Integrating Constraints and Metric Learning in Semi-Supervised Clustering", *Proceeding of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[5] A. M. Bensaid, L. O. Hall, J. C. Bezdek, and L. P. Clarke, "Partially supervised clustering for image segmentation," *Pattern Recognition*, 29(5), pp. 859-871, 1996.

[6] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE trans. On systems, man and cybernetics*, pp787-795, 1997.

[7] 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.