利用叢集整體性達到符合人類感知的模糊 c-means 分群法

An Integrity-based Fuzzy C-means Clustering Algorithm Conforming to Human Perception

賴彥豪
中興大學資訊科學與工程學系
yanhao.lai@gmail.com

林芬蘭
靜宜大學資訊工程學系
lan@pu.edu.tw

黃博惠
中興大學資訊科學與工程學系
powhei.huang@msa.hinet.net

***摘要*—本篇論文中，我們利用叢集的整體性提出符合人類感知之模糊 c-means 分群法，用來克服非對等叢集大小上的數量差異問題。傳統模糊 c-means 分群法會傾向等量每個叢集大小，以求得數學上的最佳解，但並非是最符合人類感知的分群結果。條件式模糊 c-means 分群法利用叢集大小的比例來平衡叢集間的影響力，其效能會取決於初始群心的位置。為了提升對於非對等叢集大小問題的準確率，我們利用叢集之間與叢集內部的資料點分佈來定義叢集之整體性，並且以叢集的角度與資料的觀點來決定每個資料點的權重值。實驗結果說明，基於叢集整體性的模糊 c-means 分群法，不但能夠有效地抑制較大叢集內資料的影響力，同時也能夠將錯誤的初始群心進一步修正至符合人類感知的正確位置上。**

***關鍵詞*—分群演算法、模糊 c-means 分群法、條件式模糊 c-means 分群法、基於整體性之模糊 c-means 分群法、數量差異問題。**

***Abstract*— Traditional fuzzy c-means (FCM) that uses a least squared errors function as its objective function tends to partition a dataset to clusters of equal population for obtaining a mathematically optimal solution; however, such results are not satisfactory for cases of unequal clusters from human perception. csiFCM introduces a ratio of cluster sizes as the condition value to balance the influence of different clusters, but its performance is predominantly depending on the initial center positions. In order to produce clustering results conforming to human perception even for unequal size cases, we propose an integrity-based FCM method, which determines the contribution of each data by considering the data relationship of both inter- and intra- clusters, to improve the insufficiency of conditional FCM on solving quantity variation problem. Experiment results show that our proposed integrity-based FCM clustering method not only can suppress the influence of dataset in larger clusters effectively, but also can update the wrong initial centers to the positions conforming to human perception.**

***Keywords*—clustering; fuzzy c-means; conditional fuzzy c-means; integrity-based fuzzy c-means; quantity variation problem.**

一、簡介

分群(clustering)是一種將未知的資料點集合依據特定的距離量度來將資料點分類成群的

非監督式學習方法[6]。分群演算法主要是依據資料點於特徵空間中的分佈情形，將相似的資料點群聚在一起，形成數個叢集(clusters)。使得每一個叢集內部所有的資料點都具有非常相似的特性，而在不同叢集之間的資料點所擁有的特徵值是不相似的。由於分群演算法具有將相同性質的資料點聚集成群的特性，因此除了可以被使用於影像切割和影像分類的議題上，同時也已經被廣泛的使用在電腦視覺與圖形識別的領域中。

模糊 c-means(fuzzy c-means, 簡稱 FCM)，是根據明確分群法(hard c-means 或稱 k-means)所衍生而來的一種分群演算法[2]。此方法是透過加入了模糊邏輯的概念，讓每個資料點不再絕對地屬於任何一個叢集，而是以一個介於 0 與 1 之間的數值來表示其隸屬於某個叢集的程度。模糊 c-means 屬於一種分割式的分群法(partitional clustering)，先給定欲分群之叢集數目，再利用疊代的方式，找出最佳的叢集群心來求得最小化的目標函式值(objective function)，以及在數學上的最佳分群結果。由於在真實的世界中，物件的歸屬往往不具有明確的分界線，如果利用模糊 c-means 分群法的特性，對於不明確的部份加以描述，則能夠比傳統明確分群法保留更多的資訊，同時也能夠獲得較好的分群結果。

雖然模糊 c-means 是一種有效的分群法，但是在模糊 c-means 分群法中所採用的目標函式(最小平方差)會傾向於在分群結果中等量最後每個叢集內的資料個數，以便最小化目標函式值。因此模糊 c-means 分群法在處理具有非對等叢集大小(unequal size)的問題時，較小叢集的群心位置往往會被牽引至鄰近較大的叢集區域，而造成最後的分群結果雖然能夠獲得最小的目標函式值(數學上的最佳解)，但是並不符合人類的感知，我們稱這個問題叫做“非對等叢集大小的數量差異(quantity variation)問題”。

為了解決非對等叢集大小問題上的數量差異，條件式模糊 c-means 分群法(conditional fuzzy c-means)使用權重的方式來降低在較大叢集中所有資料點的影響力，進而平衡不同大小叢集的影響力，如此一來就能夠防止較小叢集的群心朝向鄰近較大叢集的方向移動。Bensaid 等作者[1]提出一個半監督式的模糊 c-means 分群法(semi-supervised FCM)，簡稱 ssFCM。此方法是透過一組已知分群結果的訓練資料集合(training set)，預先決定每個叢集的大小，並且讓較小叢集中的資料集合具有較大的權重值。然而，在真實的案例中，資料點的歸屬以及叢集的大小往往是無法預先得知的。因此 Noordam 等作者[7]為了解決此預先資訊的問題，提出了一個不易受到叢集大小所影響的模糊 c-means 分群法(cluster size insensitive version of FCM)，簡稱 csiFCM。此方法是將每次疊代後的分群結果經過解模糊化的程序(defuzzification)，自動計算出每個叢集的大小，並且根據叢集大小的比例來指派不同的條件值(conditional value)給予每個叢集，作為下一次疊代過程中每個資料點的權重值。因此，在分群的過程中，具有較小條件值的資料集合(較大的叢集)將會對整個分群結果提供較小的貢獻程度，使得最後分群的結果在非對等叢集大小的案例中，不會因為目標函式的因素，造成最後分群結果中等量每個叢集的大小。

儘管 csiFCM 分群法能夠自動化地決定每個叢集之間的條件值來解決非對等叢集大小的問題，但是此分群法的效能是取決於資料集合的分佈情形與叢集群心的初始位置。在一個具有非對等叢集大小的案例中，如果叢集之間的大小差異很大以及叢集之間的距離相距不遠，並且初始群心幾乎座落於等分叢集的位置時，csiFCM 分群法所採用的條件值計算方式是不足

以有效的抑制較大叢集內資料集合的影響力，而使得csiFCM分群法所產生的結果會類似於傳統模糊c-means分群法，形成一個叢集大小幾乎相等的分群結果。

在近年來，對於模糊 c-means 分群法的研究，大多著重於在空間域上整合鄰域之間的關係來增加對於雜訊上的處理[3][4][5][8]。然而，這些方法都是基於傳統模糊c-means的方法，利用最小平方差的目標函式來求得分群結果。因此將這些分群演算法使用於非對等叢集大小的問題時，同樣地也會遭受到數量差異(quantity variation)上的困擾，進而產生一個傾向於等量的分群結果。因此，為了解決此問題，在本篇論文中，除了利用叢集大小的比例外，還進一步考慮每一個叢集的整體性以及其內部資料點的分佈情形來決定每個資料點的權重值，以達到更符合人類感知的模糊c-means分群結果。基於整體性之模糊 c-means 分群法主要是透過叢集的整體性來調整資料點的歸屬程度，當叢集整體性高的時候，我們直接採用叢集大小的比例來給予權重值，讓較小的叢集能夠擁有較大的影響力；反之，如果叢集整體性低的時候，則代表叢集內部具有“不純淨(impure)”資料點的存在，因此我們將加強叢集內部“純淨(pure)”資料點的權重值，同時降低不純淨資料點的影響力來提升分群結果的準率性。藉由考量整體性的因素，我們不但可以保留 csiFCM 分群法的優點，同時也能夠改善csiFCM分群法中不足的地方。

本篇論文的內容主要可分成：第二節探討傳統模糊 c-means 分群法會遭遇到所謂的數量差異(quantity variation)問題以及csiFCM分群法對於解決非對等叢集大小問題上不足的地方。第三節說明我們所提出之整體性模糊 c-means 分群法如何有效的達到一個更符合人類感知的分群結果。第四節進行相關實驗並且分析比較各個分群法的結果。第五節為本篇論文之結論與未來工作之展望。

二、相關研究探討

(一) Fuzzy c-means (FCM)原理

模糊c-means分群法是由Bezdek等作者[2][2]所發展出來的一種分群方法，主要是加入模糊邏輯的觀念，進一步的提升明確分群法(hard c-means)分群的結果。假設給予一組具有n個資料點的集合，每個資料點以一個多維度的特徵向量來代表，因此整個資料點集合可以被表示為$X=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)$。我們預期可將資料集合分成$c$個叢集$(\mathbf{c}_1,\mathbf{c}_2,...,\mathbf{c}_c)$，而這$c$個叢集的群心可表示為$(\mathbf{v}_1,\mathbf{v}_2,...,\mathbf{v}_c)$。在模糊c-means分群法中，我們利用一個大小為$c\times n$的隸屬矩陣U(membership matrix)來表示每個資料點隸屬於每個叢集的程度，並且根據此一隸屬矩陣來定義出目標函式J(objective function)：

$$J(U,\mathbf{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|\mathbf{x}_j-\mathbf{v}_i\right\|^2. \qquad (1)$$

其中，m是一個介於$[1,\infty)$之間控制模糊程度的係數，如果m趨近於1，則愈接近明確的分群，如果m趨近於∞，則分群結果愈模糊，一般m都設定為2。||*||代表計算資料點$\mathbf{x}_j$和群心$\mathbf{v}_i$的距離函式，一般皆採用歐幾里得距離。另外針對任一個資料點$\mathbf{x}_j$而言，其對於所有叢集的隸屬程度之總合必定為1，如下所示：

$$\sum_{i=1}^{c}u_{ij}=1,\ \ \forall j=1,2,...,n. \qquad (2)$$

為了要得到最佳化的目標函式J，可以依據Lagrange Multiplier方法，求得新的隸屬矩陣U如下：

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|\mathbf{x}_j - \mathbf{v}_i\right\|}{\left\|\mathbf{x}_j - \mathbf{v}_k\right\|}\right)^{2/(m-1)}}. \quad (3)$$

以及重新計算每個叢集的群心值如下：

$$\mathbf{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}^m}, 1 \le i \le c. \quad (4)$$

根據上述的定義，傳統模糊 c-means 分群法的步驟為：

(A) 任意選取 c 個初始群心。

(B) 根據方程式(3)，計算出隸屬矩陣 U。

(C) 根據方程式(4)，計算出新的群心值。

(D) 根據方程式(1)，計算新的目標函式 J 值。

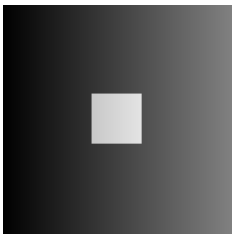
(E) 重覆執行步驟(B)到(D)，直到(J_{pre}-J_{now})值小於預設的門檻值ε，則結束本演算法。

(F) 最後利用隸屬矩陣 U，進行解模糊化程序，求得每一個資料點最後所屬的叢集。

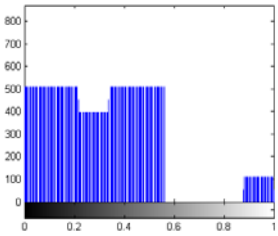
(二) 非對等叢集大小之數量差異問題

以模糊 c-means 為基礎的分群演算法大多是利用最小平方差來求得資料點與群心的距離。因此這些方法都會具有一個共通的缺點，就是儘可能的會將群心的位置調整至能夠等分每個叢集大小的位置，以便獲得最佳的目標函式值。根據計算群心位置的方程式(4)可得知，對於某一特定的叢集而言，具有較小隸屬程度的資料點所能影響此叢集群心最後位置的程度較小。因此，一般而言，每一個群心的位置應該會朝向具有較大隸屬程度的資料區域移動。然而，在非對等叢集大小問題中，對於較小叢集的群心而言，雖然在較大叢集中的資料點對於此叢集具有較小的隸屬程度，但是在較大叢集中的資料點個數遠多於較小叢集中的資料點個數。因此在更新較小叢集群心的位置時，會因為較大叢集中所有資料點集合的隸屬程度之總合大於較小叢集中資料點集合的隸屬程度之總合，所以最後會造成較小叢集的群心位置被錯誤的更新至鄰近較大的叢集區域。這樣的現象我們稱之為“數量差異(quantity variation)問題”。

圖 1 和圖 2 是一個具有非對等叢集大小問題的例子，我們使用傳統的模糊 c-means 分群法來解釋所謂數量差異的問題。圖 1(a)中是一組具有漸層效應的資料集合，此一資料集合可以被分成 2 個叢集：背景與前景。其中背景的部份是由 64200 個具有較低特徵值(範圍從 0 到 128)的資料點所組成；而前景的部份則是一個座落在中間由 3136 個具有較高特徵值(範圍從 200 到 228)的資料點集合形成的正方形區塊。而圖 1(b)則是由圖 1(a)的資料集合經過正規化後所產生出來的特徵分佈直方圖。我們可以從分佈圖得知，此資料集合以人類感知而言，其正確的分群結果應該是被分成左右 2 個叢集。左邊是具有較低特徵值的背景，以及右邊具有較高特徵值的前景。然而，由於模糊 c-means 分群法會傾向於等量每個叢集的大小，因此最後較小的叢集(右邊特徵值較大的部份)群心位置會慢慢的往較大叢集(左邊特徵值較小的部份)方向移動。圖 2 是利用傳統模糊 c-means 分群法在不同疊代次數的時候所產生的分群結果。首先，我們將初始群心手動地設定在最佳的位置上(0.3,0.95)，利用此群心位置可以獲得一個最好且最符合人類感知的分群結果，如圖 2 中初始階段的結果(藍色區域代表背景的叢集，紅色區域則是前景的叢集)。在經過連續不斷的更新群心位置以及隸屬矩陣後，模糊 c-means 分群法為了得到最小的目標函式值(從 1573.6 到 857.2)，叢集的群心會慢慢的從視覺上最佳的位置(0.3,0.95)往具有較多資料點個數的區域移動(0.15,0.49)。最後，如圖 2 中所示，在經過多次疊代之後，背景中特徵值較大的區域會漸漸地與前景的部份分類成為同一個叢集。透過這個簡單非對等的例子，我們可以知道在非對等叢集大小問題中，較小的叢集群心位置在更新的時候，會因為叢集之間的大小差異而造成隸屬程度之總合上的問題，而導致較小叢集會被較



(a)



(b)

圖 1、具有非對等叢集大小問題之範例。(a)具有漸層效應的資料集合。(b)正規化後的特徵分佈直方圖。

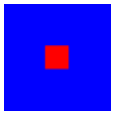
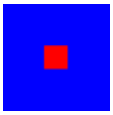
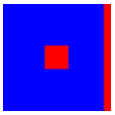
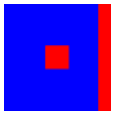
疊代次數	初始階段	第 1 次	第 2 次
群心位置	(0.3,0.95)	(0.27,0.86)	(0.25,0.79)
目標函式	1573.6	1441.7	1363.6
分群結果			
疊代次數	第 4 次	第 6 次	停止階段
群心位置	(0.22,0.63)	(0.18,0.53)	(0.15,0.49)
目標函式	1256.3	894.1	857.2
分群結果			

圖 2、針對圖 1 之資料集合，利用傳統模糊 c-means 分群法之分群結果。

大叢集內資料點所影響，使得較小叢集的群心會朝著具有數量較多的叢集方向移動，進而等量最後的分群，產生一個不符合人類感知的分群結果。

除了模糊 c-means 分群法會受到數量差異的問題之外，在近年來以模糊 c-means 分群法為基礎所提出來的整合區域空間關係之分群法[3][4][8]，同樣地在非對等叢集大小問題上也會產生和傳統模糊 c-means 分群法相同的結果(以圖 1 的例子而言，最後叢集的群心都會座落在大約(0.15,0.49)的位置)。雖然這些方法能夠有效的處理有關雜訊的問題，但是由於所使用的群心位置與隸屬矩陣更新的方式，仍然與傳統模糊 c-means 分群法類似，因此仍然會遭遇到非對等叢集大小的問題。

(三) csiFCM 分群法之不足

Noordam 等作者[7]所提出的不受叢集大小影響之模糊c-means分群法，簡稱csiFCM，主要是用來解決傳統模糊c-means分群法對於非對等叢集大小的問題。其原理是利用每次疊代後的解模糊化分群結果來計算出每個叢集之大小，再利用叢集大小之比例來指定一個較小的條件值f_j給予在較大叢集內的資料點集合，以便降低較大叢集內資料點對於分群結果的影響力。此條件值是一個介於 0 到 1 之間的數值，如果資料點集合是屬於最小的叢集，則此資料點集合是不受限制的(f_j=1)。為了要達到不受叢集大小的影響，csiFCM分群法將傳統模糊c-means分群法求得隸屬矩陣U的方程式修改成如下：

$$u_{ij} = \frac{f_j}{\sum_{k=1}^{c}\left(\frac{\|\mathbf{x}_j - \mathbf{v}_i\|}{\|\mathbf{x}_j - \mathbf{v}_k\|}\right)^{2/(m-1)}}. \qquad (5)$$

其中條件值f_j是利用解模糊化程序後的分群結果來計算出叢集大小之比例所決定的，如下：

$$\mathcal{S}_i = N_i / N. \qquad (6)$$

$$f_j = \frac{1-\mathcal{S}_i}{\max_{l=1,2,\ldots,c}\left(1-\mathcal{S}_l\right)}. \qquad (7)$$

S_i代表經過解模糊化程序後第i個叢集的大小比例。N_i是在最後分群結果中第i個叢集的資料點個數，N則是所有資料點集合的總數。因此，每一個資料點對於所有的叢集而言，會有一個固定的條件值。同時，對於被分類到同一個叢集內的資料點集合也會有相同的條件值。而此條件值，在每經過一次的疊代後，就會根據新的解模糊化分群結果重新加以計算。

根據 csiFCM 演算法，我們可以發現 csiFCM 分群法的效能是取決於叢集大小的比例因素。因此，如果在初始階段中，群心的位置能夠座落在合適的位置上，並且得到一個較好的叢集比例值，那麼將能夠有效的解決非對等叢集大小的問題。相反地，如果群心的初始位置是座落在不適當的位置，而所計算出來的叢集大小比例值與真實叢集大小剛好相反或是大小比例值幾乎相等(1:1)，那麼 csiFCM 分群法將會產生與傳統模糊 c-means 分群法相似的結果。

雖然在多數的非對等叢集大小問題上，csiFCM 都能夠得到正確的分群結果。但是在叢集資料分佈較接近以及初始叢集大小的比例不足以抑制較大叢集中資料點集合的隸屬程度之總合，仍然無法有效的解決非對等叢集大小的問題。圖 3 是針對圖 1 的例子，利用 csiFCM 分群法進行不同疊代次數所產生的分群結果。首先，我們將初始群心值設定在(0.3,0.6)的位置。從圖 1(b)中，我們可以得知，以此群心位置作為初始值時，有部份的背景資料集合會被歸類成前景的叢集，而使得在前景叢集中的資料分佈，會有許多的資料點是非常接近背景叢集(在 0.45 到 0.6 之間的資料點)。因此造成在初始階段所求得的叢集比例大小不足以能夠防止較小叢集的群心(前景)被牽引至較大的叢集(背景)。最後在經過連續疊代後，我們可以發現，csiFCM 的分群結果會跟傳統模糊 c-means 分群法一樣，逐漸的等量每個叢集的大小。

圖 4 是利用 Noordam 等作者[7]在實驗 1 中所使用的資料集合以及加上一些變化來說明 csiFCM 分群法對於非對等叢集大小問題之不足之處。圖 4(a)和(b)的資料集合與分佈是參考 Noordam 等作者所使用的實驗資料集合所產生的，其中圖 4(a)的資料集合具有 43 個資料點，可以分成 2 個叢集。較大的叢集包含 40 個資料點，而較小的叢集包含了 3 個資料點。圖 4(b)的資料集合是由 3 個叢集所組成，最大的叢集含有 40 個資料點，第二大的叢集具有 6 個資料點，而最小的叢集是由 3 個資料點所組成。在圖 4 中，標示點“◆”代表經過傳統模糊 c-means 分群法所得到的最後群心位置，而這些群心位

疊代次數	初始階段	第 2 次
群心位置	(0.3,0.6)	(0.25,0.58)
叢集大小	(49600,15936)	(45504,20033)
分群結果		
疊代次數	第 5 次	停止階段
群心位置	(0.20,0.53)	(0.15,0.49)
叢集大小	(40896,24640)	(34744,30792)
分群結果		

圖 3、利用 csiFCM 分群法對圖 1 之例子進行分群的結果。

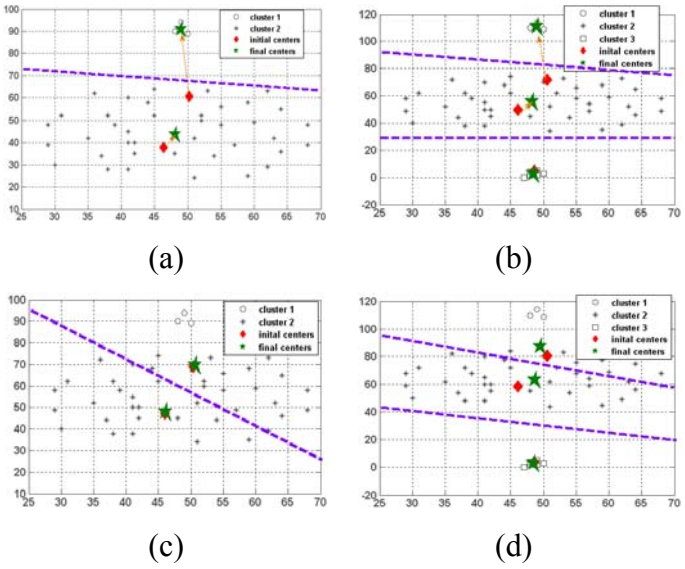


圖 4、利用 Noordam 等作者[7]所使用之資料集合說明 csiFCM 對於非對等叢集大小問題之不足。

置被拿來作為 csiFCM 分群法的初始群心位置(在 csiFCM 演算法中，此作法是為了避免遭遇區域最小值)。標示點"★"則是經過 csiFCM 分群法後各個叢集的群心位置。虛線則是代表最後叢集與叢集之間的分界處。在圖 4(a)和 4(b)的例子中，由於叢集與叢集之間的距離較大，且初始叢集大小有足夠比例，因此利用 csiFCM 分群法能夠有效的防止較小叢集的群心往較大的叢集方向移動。然而，如果我們將叢集之間的距離縮小，如圖 4(c)和 4(d)中，分別將屬於最大叢集的資料點集合往上移動 10 個單位，使其更接近較小的叢集。如此一來，在圖 4(c)中的初始叢集大小比例會幾乎等於 1:1，而使得 csiFCM 分群法所使用的條件值無法產生效果。而在圖 4(d)中，雖然初始的叢集大小比例與圖 4(b)類似，但是卻因為 2 個叢集之間的距離縮小，而使得此一比例仍然無法有效的抑制較大叢集內資料點的影響力。因此在圖 4(c)和 4(d)的二個例子中，雖然是利用 csiFCM 分群法進行分群，但是最後的結果仍然是類似於傳統的模糊 c-measn 分群法。

三、基於整體性的模糊 c-means 分群法

根據人類感知的觀點而言，分群的目的就是將所觀察到的資料點指派到不同的叢集內，而使得在每一個叢集內的資料點彼此之間能夠有很高的相關性，進而產生一個具有高整體性(integrity)的資料點集合。因此，不同於傳統模糊 c-means 分群法與 csiFCM 分群法，我們將以提升叢集的整體性為考量，使得每個叢集內的資料點集合能夠存在最大的相關性，並且在叢集之間的資料集合能夠具有最小的相似性。

(一) 整體性的定義

整體性(integrity)所代表的意義就是在同一個叢集內資料特徵的一致性與相關性。因此，在本篇論文中，每一個叢集的整體性 $\mathcal{I}_i$(integrity)可以由(1)叢集之間以及(2)叢集內部之關係組合而成。根據考量叢集之間與內部兩個部份，我們可以定義出：(1)叢集的純淨度 $\mathcal{I}_{\mathcal{P}i}$(purity)，此一部份是利用計算與最接近之叢集間的不相似程度來求得，如果不相似度愈高，代表叢集的純淨度愈高；(2)叢集的緊密度 $\mathcal{I}_{\mathcal{C}i}$(compactness)，此一部份是計算叢集內部資料點集合的相似程度來求得，如果相似性愈高，代表叢集的緊密度愈高。因此整體性的公式可以表示為：

$$\mathcal{I}_i = \frac{1}{2}\cdot\left(\mathcal{I}_{\mathcal{P}i} + \mathcal{I}_{\mathcal{C}i}\right). \tag{8}$$

其中叢集的純淨度 $\mathcal{I}_{\mathcal{P}i}$ 和叢集的緊密度 $\mathcal{I}_{\mathcal{C}i}$ 分別定義如下：

$$\mathcal{I}_{\mathcal{P}i} = \frac{1}{|\mathbf{A}_i|}\sum_{j\in\mathbf{A}_i}\mathcal{P}_{j,i}. \tag{9}$$

$$\mathcal{P}_{j,i} = \frac{abs\left(\left\|\mathbf{x}_j - \mathbf{v}_i\right\| - \left\|\mathbf{x}_j - \mathbf{v}_k\right\|\right)}{\left\|\mathbf{v}_i - \mathbf{v}_k\right\|}. \tag{10}$$

$$\mathbf{v}_k = \arg\min_{l=1,2,\ldots,c;l\neq i}\left\{\left\|\mathbf{v}_i - \mathbf{v}_l\right\|\right\}. \tag{11}$$

以及

$$\mathcal{I}_{\mathcal{C}i} = 1 - \sqrt{\frac{1}{|\mathbf{A}_i|}\sum_{j\in\mathbf{A}_i}\left(\left\|\mathbf{x}_{j,i} - \mathbf{v}_i\right\| - \mu_i\right)^2}. \tag{12}$$

$$\mu_i = \frac{1}{|\mathbf{A}_i|}\sum_{j\in\mathbf{A}_i}\left\|\mathbf{x}_{j,i} - \mathbf{v}_i\right\|. \tag{13}$$

其中所有的資料點**x**與群心**v**的值皆已正規化至 0 到 1 之間。$\mathbf{x}_j$代表資料點j的特徵向量，$\mathbf{v}_i$代表第i個叢集群心的特徵向量，而$\mathbf{v}_k$則是距離$\mathbf{v}_i$最接近的群心。$\mathcal{I}_{\mathcal{P}i}$ 代表第i個叢集的純淨度，$\mathcal{P}_{j,i}$ 代表資料點j經過解模糊化後被分類到第i個叢集的純淨度，其計算的方式是利用與群心$\mathbf{v}_i$的距離

和與最接近的群心$\mathbf{v}_k$的距離相減所求得，當距離差愈大代表此資料點對於所屬的叢集而言，其純淨度愈高。然而，因為對於每個資料點$\mathbf{x}_j$而言，群心$\mathbf{v}_i$與$\mathbf{v}_k$的距離並不相等，所以我們必須要除以$\|\mathbf{v}_i-\mathbf{v}_k\|$來對純淨度進行正規化。$\mathcal{I}_{Ci}$是利用資料點與群心的距離標準差來呈現叢集$i$的緊密度，當距離的標準差愈小，代表緊密度愈高。$\mathbf{A}_i$是經過解模糊化程序後被分類到第$i$個叢集的所有資料點集合，而$|\mathbf{A}_i|$代表此集合內資料點的個數。求得資料點集合$\mathbf{A}_i$的方式如下：

$$\mathbf{A}_i=\left\{x_j \,\middle|\, u_{ij}>u_{lj};\ \ l=1,2,...,c;\ \ l\neq i\right\}. \quad (14)$$

其中u_{ij}是第二節中所定義的資料點x_j隸屬於第i個叢集的程度值。

(二) 整體性模糊 c-means 分群

根據第二節中針對 csiFCM 分群法所探討的結果，我們可以知道雖然利用叢集大小比例是能夠解決非對等叢集大小的問題，但並不適用於所有的情形。其主要的原因在於，分群的好壞，不僅須考慮叢集大小的比例(以叢集的觀點來看)，還必須考慮到每一叢集內部資料點的分佈情形(以資料點的觀點來看)。因此為了保留 csiFCM 分群法之優點，並且改善其不足之處，我們利用「叢集的觀點」以及「資料點的觀點」同時配合整體性的概念來調整對於每個資料點的權重值，概念如下：

(1) 當叢集具有高整體性的時候，即代表此叢集的分群結果是符合人類感知的。因此可以直接採用叢集大小比例 (叢集的觀點)來平衡每一群的權重值即可。當叢集具有高整體性時，代表叢集內部所存在的“不純淨(impure)”資料點很少，因此我們不需要針對叢集的內部資料點集行權重的調整(資料點的觀點)。如圖 5 中叢集 1 所示，我們可以發現在叢集 1 內的所有資料點之間彼此具有很大的相關性，使得其整體性會比叢集 2 來的高。由於其整體性高，所以我們不需要針對叢集 1 內部資料點的權重進行調整，直接使用叢集大小比例作為權重值即可。

(2) 當叢集具有低整體性的時候，代表叢集的分群結果是較不盡理想的。而不好的分群結果往往會造成叢集大小比例的值無法有足夠的力量抑制較大叢集內資料點的影響(叢集的觀點)。當叢集具有低整體性時，代表在叢集內部會有許多“不純淨(impure)”的資料點存在。因此為了提升分群的結果，我們需要加強“純淨(pure)”資料點的權重，使得叢集內部的資料分佈能夠更加的集中(資料點的觀點)。如圖 5 中叢集 2 所示，我們可以發現在叢集 2 中的資料集合可以被分為二個區域(Ⅱ和Ⅲ)。其中，雖然區域Ⅱ的資料點個數較多，但是其距離區域Ⅰ非常相近，因此對於叢集 2 來說，區域Ⅱ的資料點集合屬於“不純淨(impure)”的資料集合，而區域Ⅲ的資料點集合才算是在此一階段中的“純淨(pure)”資料集合。因此我們必須透過加強區域Ⅲ資料點的權重來使得分群結果更能符合人類的感知。

基於上述的概念，我們可以得到權重調整公式如下：

$$W_{ij}=W_{\mathcal{C}_{ij}}\cdot W_{P_{ij}}. \quad (15)$$

其中$W_{\mathcal{C}_{ij}}$與$W_{P_{ij}}$分別定義如下：

$$W_{\mathcal{C}_{ij}}=\frac{1-\mathcal{S}_i}{\max\limits_{l=1,2,...,c}\left(1-\mathcal{S}_l\right)}. \quad (16)$$

以及

$$W_{P_{ij}}=\frac{\mathcal{D}_{j,i}}{\max\left(\mathcal{D}_{j,i}\right)}. \quad (17)$$

$$\mathcal{D}_{j,i} = \exp^{\left(1-\mathcal{I}_i^{'}\right)\cdot \mathcal{P}_{j,i}}. \qquad (18)$$

$$\mathcal{I}_i^{'} = \frac{\mathcal{I}_i - \min\limits_{l=1,2,\ldots,c}\left(\mathcal{I}_l\right)}{\max\limits_{l=1,2,\ldots,c}\left(\mathcal{I}_l\right) - \min\limits_{l=1,2,\ldots,c}\left(\mathcal{I}_l\right)}. \qquad (19)$$

$W_{C_{ij}}$ 是以叢集的觀點來看，其代表的是經過正規化過後的叢集大小比例值，愈小的叢集會具有愈大的值。在經過解模糊化過程後，對於被分類至叢集 i 的所有資料點 j 都會擁有相同的條件值。同時，對於某一資料點 j 而言，此條件值會作用到所有的叢集。換句話說，也就是對於所有的叢集 i，當資料點 j 固定時，$W_{C_{ij}}$ 的值會是相等的。而 $C_{P_{ij}}$ 是以資料點的觀點來看，其代表經過正規化後的純淨度，愈純淨的資料點會具有愈大的值。每一個被分類到第 i 個叢集的資料點 j，會根據其純淨度與其所屬叢集 i 的整體性給予條件值。將權重調整公式代入整體性模糊 c-means 的概念後，可得知：

(1) 當整體性高的時候，我們只需著重於叢集大小比例作為考量的依據。所以在整體性高的時候 $\left(1-\mathcal{I}_i^{'}\right)$ 會趨近於 0，使得條件值主要是根據 $W_{C_{ij}}$ 在改變。

(2) 當整體性低的時候，我們必需額外考量叢集內部資料的分佈情況。所以在整體性低的時候，$\left(1-\mathcal{I}_i^{'}\right)$ 會遠大於 0，使得 $W_{P_{ij}}$ 能夠提升"純淨(pure)"資料點的權重，讓群體內的資料分佈能夠更加的集中。

最後，根據上述所計算出來的權重值，我們將傳統的隸屬矩陣 U 調整成如下來求得一個較符合人類感知的分群結果：

$$u_{ij} = \frac{W_{ij}}{\sum\limits_{k=1}^{c}\left(\frac{\left\|\mathbf{x}_j - \mathbf{v}_i\right\|}{\left\|\mathbf{x}_j - \mathbf{v}_k\right\|}\right)^{2/(m-1)}}. \qquad (20)$$

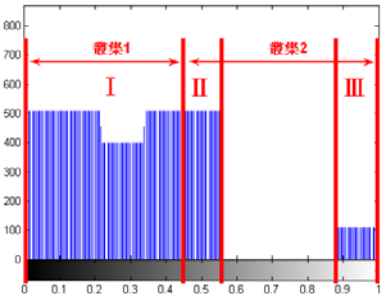


圖 5、基於整體性之模糊 c-means 分群法概念說明圖。

(三) 整體性模糊 c-means 分群演算法

在本篇論文中所提出之基於整體性的模糊 c-means 分群演算法，其步驟如下：

(A) 任意選取 c 個初始群心。

(B) 根據方程式(3)，計算出隸屬矩陣 U。

(C) 利用解模糊化程序，指派每一個資料點到所屬的叢集。

(D) 利用方程式(6)，計算每一個叢集的大小以及其比例。

(E) 利用方程式(8)到(14)，計算每一個叢集的整體性。

(F) 利用方程式(15)到(19)，計算每一個資料點新的權重值。

(G) 利用方程式(20)，求得新的隸屬矩陣 U。

(H) 利用方程式(4)，計算出新的群心值。

(I) 利用方程式(1)，計算出新的目標函式 J 值。

(J) 重覆執行步驟(B)到(I)，直到(J_{pre}-J_{now})值小於預設的門檻值ε或是落於區域最小值(local minima)，則結束本演算法。

(K) 最後利用隸屬矩陣 U，進行解模糊化程序，求得每一個資料點最後所屬的叢集。

四、實驗結果與比較

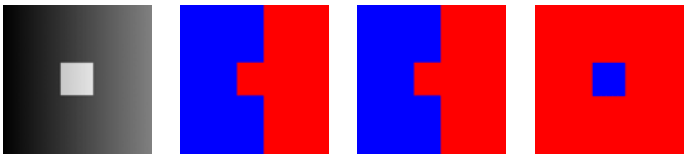
在本篇論文中，我們利用具有不同叢集個數的一維以及二維的非對等叢集大小的資料集合來作為分群的實驗資料。並且利用此實驗資

料來比較傳統模糊 c-means 分群法、csiFCM 分群法以及我們所提出之基於整體性的模糊 c-means 分群法對於非對等叢集大小問題的效能。在實驗中，為了讓 csiFCM 分群法與我們所提出之分群法具有相同的初始群心位置，我們皆以傳統模糊 c-means 分群法所得到的分群結果來作為初始階段的群心位置，而不是以亂數來取得。

(一) 一維資料集合之效能比較

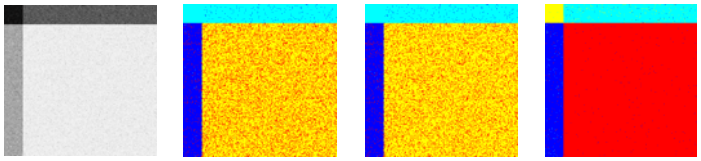
在第一個實驗中，我們利用兩組具有一維特徵值的資料集合來進行分群演算法的效能比較。圖 6 是具有 2 個叢集的資料集合，其資料分佈直方圖如圖 1(b)所示。從分佈圖中，我們可以得知，最符合人類感知的群心位置應該是座落在(0.28,0.95)。由於傳統模糊 c-means 分群法傾向於等量每個叢集的資料點個數，因此利用傳統模糊 c-means 分群法的結果會將此資料集合一分為二，如圖 6(b)所示。由 csiFCM 演算法可知，csiFCM 分群法的結果會取決於始初叢集比例的大小。在此範例中，當利用傳統模糊 c-means 分群法所產生出來的群心作為初始起始群心時，所得到的叢集比例大小幾乎為 1:1(34744:30792)，因此無法有效的抑制較大叢集內資料點的影響力，所以仍然無法將群心位置調整至理想的位置。利用 csiFCM 分群法所產生的結果如圖 6(c)所示。然而，我們所提出的基於整體性之模糊 c-means 分群法，能夠針對叢集內資料點的分佈進行調整，而不是只考慮叢集大小比例因素，因此能對更有效的得到符合人類感知的分群結果。圖 6(d)是利用我們所提出之分群法所得到的分群結果。在圖 6(b)、(c)和(d)中最後的群心位置分別座落於(0.15,0.49)、(0.15,0.49)和(0.22,0.92)的位置。

圖 7 是一個具有 4 個叢集的一維資料集合。此一集合包含了 65536 個資料點，最小的叢集(左上角的區塊)具有 1024 個資料點，每個資料點的特徵值為 0；第二大的叢集有 2 個，分別是由 7168 個具有特徵值 85 和 170 的資料點所構成；而最大的叢集則是由 50176 個特徵值為 255 的資料點所組成。我們在此資料集合中加入平均值為 0 和變異數為 0.01 的高斯雜訊來



(a) (b) (c) (d)

圖 6、具有漸層效應的一維資料集合。(a)具有 2 個叢集的原始資料集合。利用(b)傳統模糊 c-means 分群法，(c)csiFCM 分群法，(d)本論文之分群法，所產生的分群結果。



(a) (b) (c) (d)

圖 7、具有高斯雜訊(平均值為 0，變異數為 0.01)的一維資料集合。(a) 具有 4 個叢集的原始資料集合。利用(b)傳統模糊 c-means 分群法，(c)csiFCM 分群法，(d)本論文之分群法，所產生的分群結果。

迫使每個叢集之間的資料分佈會更加靠近，甚至有些許的重疊。從圖 7(b)和(c)的分群結果可以得知，傳統模糊 c-means 分群法和 csiFCM 分群法會將最小的的叢集(左上的區域)與其鄰近的叢集(上方的區域)錯誤的分類在一起。而我們所提出的分群法則能夠將資料集合正確的分成 4 個叢集，如圖 7(d)所示。在圖 7(b)、(c)和(d)中，最後的群心位置分別座落於(77,166,230,254)、(77,163,227,254)和(6,87,167,248)的位置上。

(二) 二維資料集合之效能比較

在第二個實驗中，使用與圖 4 中相同的資料集合來比較 csiFCM 分群法與我們所提出之分群法的分群結果。圖 7 中所使用的初始群心(如標示點“◆”所示)與圖 4 中的初始群心位置相同。在圖 7(a)和(b)的例子中，我們所提出的分

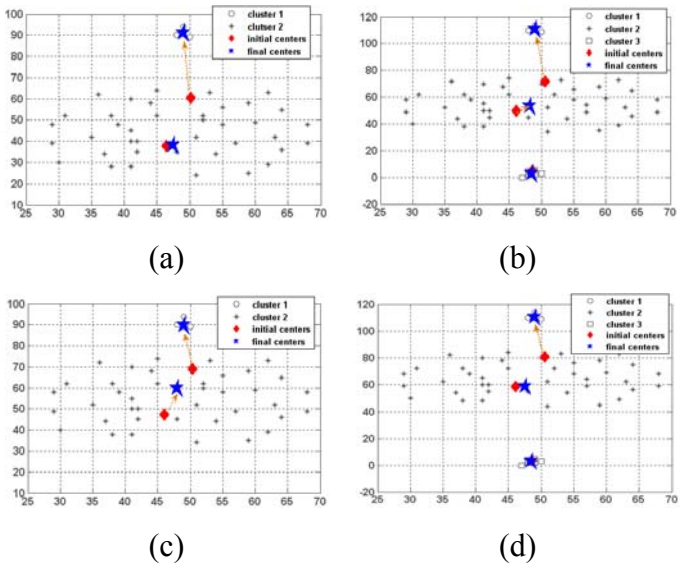


圖 8、利用本論文所提出的分群法對於圖 3 所使用的資料集合進行分群效能之比較。

群法 (如標示點"★"所示)與 csiFCM 分群法(如圖 4(a)和(b)所示)都能夠有效的將錯誤的初始群心修正至符合人類感知的位置上。但是在圖 7(c)和(d)的例子中，當較大叢集的資料點分佈往較小的叢集靠近的時候，csiFCM 分群法無法有效的將誤錯的初始群心修正至正確的位置上(如圖 4(c)和(d))，而我們所提出的方法則能夠正確的獲得正確的分群結果，如圖 7(c)和(d)中藍色星形標示點所示。

雖然在這二組實驗中，我們是利用傳統模糊 c-means 分群法的結果來作為初始階段的群心位置。但是實際上，在我們所提出的分群法是不需要預先執行傳統模糊 c-means 分群法，而是可以利用亂數的方式來選擇初使群心位置。因為我們所提出之分群法不僅僅是考量叢集大小比例，還有考慮叢集內部資料點的分佈。因此不會因為初始的叢集大小比例不好就無法獲得正確的結果。

五、結論與未來展望

在非對等叢集大小(unequal size)案例中，數量差異(quantity variation)的問題是所有基於模糊 c-measn 分群法都會遭遇到的問題。因為使用最小平方差作為目標函式的時候，分群結果會傾向於等量每一個叢集中的資料點個數。因此會使得較小叢集的群心位置被牽引至鄰近較大的叢集區域，藉此來獲得較小的目標式值。為了要解決數量差異的問題，我們提出了利用叢集的整體性來達到符合人類感知的模糊 c-means 分群法。有別於 csiFCM 分群法只考量叢集大小的比例關係，我們利用叢集之間與叢集內部的資料點分佈來定義出叢集之整體性，並且以叢集的觀點與資料點的層面來決定每一個資料點的權重值。使得每一個資料點的權重值不在只是取決於叢集之間的大小，同時還包含了資料集合的分佈情況。因此在相同叢集內的資料點會因為其純淨度的不同，而使得權重值也有所不同。如此一來，既能夠保留 csiFCM 分群法的優點，也可以改善其不足之處。從分群的實驗結果中，我們可以發現，基於叢集整體性的模糊 c-means 分群法，不但能夠在非對等叢集大小的案例中抑制較大叢集內資料點的影響力，同時也能夠將錯誤的初始群心進一步的修正至符合人類感知的正確位置上。儘管我們所提出的分群法能夠有效的解決數量差異上的問題，但是此方法尚未對雜訊的問題加以考量。因此，我們未來將整合空間域的概念，來提升對於雜訊上的處理，形成一個更強健的分群法。

誌謝

本研究感謝行政院國家科學委員會(專題計畫編號：NSC-98-2221-E-126-009)的研究計經費補助。

六、參考文獻

[1] A. M. Bensaid, L. O. Hall, J. Bezdek, and L. Clarke, "Partially supervised clustering for image segmentation," Pattern Recognition, vol. 5, pp. 895-871, 1996.

[2] J. Bezdek, "Pattern recognition with fuzzy objective functions," Plenum Press, New York, 1981.

[3] K. S. Chuang, H. L. Tzeng, S. Chen, J. Wu, and T. J. Chen, “Fuzzy c-means clustering with spatial information for image segmentation,” Computerized Medical Imaging and Graphics, vol. 30, pp. 9-15, 2006.

[4] S. Chen and D. Zhang, “Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure,” IEEE Transactions on Systems, Man, and Cybernetics-Part B-Cybernetics, vol. 34, pp. 1907-1915, 2004.

[5] W. Cai, S. Chen, and D. Zhang, “Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation,” Pattern Recognition, vol. 40, pp. 825-838, 2007.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern Classification, 2nd edition”, Wiley-Interscience, 2000.

[7] J. C. Noordam, W. H. A. M. van denBroek, and L. M. C. Buydens, “Multivariate image segmentation with cluster size insensitive fuzzy c-means,” Chemometrics and Intelligent Laboratory Systems, vol. 64, pp. 65-78, 2002.

[8] D. L. Pham, “Spatial models for fuzzy clustering,” Computer Vision and Image Understanding, vol. 84, pp. 285-297, 2001.