行政院國家科學委員會專題研究計畫 成果報告

基於證據累加的叢集整合技術之強韌化與功能延伸 II
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2221-E-009-179-
執 行 期 間 ：99 年 08 月 01 日至 100 年 07 月 31 日
執 行 單 位 ：國立交通大學資訊工程學系（所）

計 畫 主 持 人：王才沛

計畫參與人員：碩士班研究生-兼任助理人員：蘇裕傑
碩士班研究生-兼任助理人員：魏良佑
碩士班研究生-兼任助理人員：林俞丞
碩士班研究生-兼任助理人員：邱俊予

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 100 年 10 月 31 日

中文摘要：　叢集化是一個可以在沒有分類資訊的資料當中，將相關的資料點區分成叢集的方法。叢集化演算法的種類很多，但並沒有一個方法可以對所有的資料與叢集性質都產生好的結果。叢集整合 (cluster ensemble) 技術是近年的一個新趨勢，其做法是對同一組資料產生多個不同的叢集化結果，再結合這些個別結果來產生一個具有共識的、更穩定也更能代表實際資料分佈的分群。

本計畫是我們前一年度國科會專題計畫（編號：98-2221-E-009-146-；題目：基於證據累加的叢集整合技術之強韌化與功能延伸；期間：98 年 8 月 1 日至 99 年 7 月 31 日）的延續性計畫。我們在本年度的主要成果包含以下兩大部分：（一）以基於 co-association 矩陣的叢集整合方法為基礎，設計一新的資料結構 CA-tree，藉由去除原始個別分群資訊中的多餘性，可以大幅降低此類演算法的運算複雜度而提升效率，同時仍可維持與舊有方法接近的分群準確度。（二）我們將叢集整合應用到有特定叢集形狀的問題，實作了利用叢集整合演算法於偵測線段叢集以及主曲線的演算法。我們相信這些研究成果將對發展叢集整合的應用有明顯的貢獻。

英文摘要：　Clustering is a process that groups unlabeled data points into clusters. There are a large variety of clustering methods, but none can generate good clustering results for all types of data and cluster characteristics. Cluster ensemble is a new trend in recent years. Its approach is to generate multiple clustering results out of the same data set, and then combine the individual clustering results to form a consensus partition of the data that is more stable and more representative of the actual data distribution.

This project is a continuation of a past NSC project (No. 98-2221-E-009-146-； Title: Robustification and Functionality Extension of Evidence-Accumulation-Based Cluster Ensembles； Duration: 2009/8/1 to 2010/7/31). The main contribution of this project includes the following two parts: (1) Starting with cluster ensembles methods based on co-association matrices, we design a new data structure called CA-tree which, by reducing the redundancy in the individual partitions, can significantly reduce the high computational complexity with little loss of clustering accuracy compared with previous methods. 2) We apply cluster ensemble to clustering problems that identify clusters of particular shapes, and implement methods that detect line-segment clusters and principal curves. We believe that the outcome of this project will contribute substantially to developing applications of cluster ensembles.

行政院國家科學委員會專題研究計畫成果報告

基於證據累加的叢集整合技術之強韌化與功能延伸 II
Robustification and Functionality Extension of Evidence-Accumulation-Based Cluster Ensembles II

計畫編號：NSC－99-2221-E-009-179-
執行期間：2010年8月1日 至 2011年7月31日

主持人：王才沛 國立交通大學資訊工程學系(所)

中文摘要

叢集化是一個可以在沒有分類資訊的資料當中，將相關的資料點區分成叢集的方法。叢集化演算法的種類很多，但並沒有一個方法可以對所有的資料與叢集性質都產生好的結果。叢集整合 (cluster ensemble) 技術是近年的一個新趨勢，其做法是對同一組資料產生多個不同的叢集化結果，再結合這些個別結果來產生一個具有共識的、更穩定也更能代表實際資料分佈的分群。

本計畫是我們前一年度國科會專題計畫（編號：98-2221-E-009-146-；題目：基於證據累加的叢集整合技術之強韌化與功能延伸；期間：98 年 8 月 1 日至 99 年 7 月 31 日）的延續性計畫。我們在本年度的主要成果包含以下兩大部分：(一）以基於 co-association 矩陣的叢集整合方法為基礎，設計一新的資料結構 CA-tree，藉由去除原始個別分群資訊中的多餘性，可以大幅降低此類演算法的運算複雜度而提升效率，同時仍可維持與舊有方法接近的分群準確度。(二）我們將叢集整合應用到有特定叢集形狀的問題，實作了利用叢集整合演算法於偵測線段叢集以及主曲線的演算法。我們相信這些研究成果將對發展叢集整合的應用有明顯的貢獻。

關鍵詞：叢集整合、證據累加、共識叢集

Abstract

Clustering is a process that groups unlabeled data points into clusters. There are a large variety of clustering methods, but none can generate good clustering results for all types of data and cluster characteristics. Cluster ensemble is a new trend in recent years. Its approach is to generate multiple clustering results out of the same data set, and then combine the individual clustering results to form a consensus partition of the data that is more stable and more representative of the actual data distribution.

This project is a continuation of a past NSC project (No. 98-2221-E-009-146-; Title: Robustification and Functionality Extension of Evidence-Accumulation-Based Cluster Ensembles; Duration: 2009/8/1 to 2010/7/31). The main contribution of this project includes the following two parts: (1) Starting with cluster ensembles methods based on co-association matrices, we design a new data structure called CA-tree which, by reducing the redundancy in the individual partitions, can significantly reduce the high computational complexity with little loss of clustering accuracy compared with previous methods. 2) We apply cluster ensemble to clustering problems that identify clusters of particular shapes, and implement methods that detect line-segment clusters and principal curves. We believe that the outcome of this project will contribute substantially to developing applications of cluster ensembles.

Keywords: cluster ensemble, evidence accumulation, consensus clustering, co-association matrix

一、簡介與文獻探討

叢集化(clustering)代表的是在一組沒有已知分類的資料當中找出其中的叢集(cluster)的過程。一個叢集是原資料組的一個子集合，其中的資料點具有高度的相關性或是具有相近的性質。我們將由一組資料區分出的所有叢集稱之爲這個資料組的一個分群(partition)。叢集化是一個非監督學習 (unsupervised learning) 的過程，必須靠著資料點之間的關係來找出其中的叢集。即使對於同樣一組資料，由於計算資料點之間關係方法的不同，或者是由這些關係來找出叢集的演算法的不同，或是所使用的參數或初始化條件不同，都會產生不同的叢集化結果。用於叢集化的演算法很多，但是各方法所適用的資料組特性以及所能找出的叢集特性都有其限制。

近幾年的一個趨勢是使用叢集整合 (cluster ensemble) 技術，對同一組資料產生多個不同的叢集化結果，再結合這些個別結果來產生一個具有共識的、更穩定也更能代表實際資料分佈的分群。叢集整合的另一個優點是它非常適合於平行的與分散式的資料分析，甚至儲存在多個不同地點的資料可以先分別作叢集化，再將其結果整合。

早期關於叢集整合的一篇重要論文是 Strehl 與 Ghosh 在 2002 所發表[1]。這篇論文說明了"知識再利用" (knowledge reuse) 用於叢集的概念，在此"知識"所代表的即是個別分群（有可能是在不同的時間點所得到的）。對於叢集整合的優點，實驗結果在許多個別研究中都有，而[2]由理論的角度證明叢集整合在穩定度及正確性方面的優點。

叢集整合的先決條件是要能對同一組資料產生多個不同的分群結果。在這方面一個具代表性的方法可算是使用同一個叢集演算法但是不同的隨機初始化[3,4]。另一個常見的方法是將高維度的特徵向量(feature vectors)投影到多個隨機的低維度次空間(subspace) [5]或隨機的選擇部份的特徵值來進行叢集化 [6]。這一類方法對高維度資料的叢集化特別有用。其他的方法還包括將資料隨機分爲幾個較小的資料組，並且對這些資料組作叢集化[7]。這一類方法的好處是當資料量很大時，每次進行叢集化都只使用部分資料可以較有效率，甚至分散在不同地點的資料可以被分別叢集化，再透過叢集整合得到整體的分群結果[8]。除此之外，當大量資料必須以線上(on-line)的方式叢集化時，也可以透過改變各資料點被處理的順序而得到不同的分群結果再做整合，如[9]。此外也可以使用數個不同的叢集化演算法來產生個別分群再做整合，如[10]中的方法。

Co-association 矩陣可算最常被使用的叢集整合的資料結構，而原因應是由於其觀念非常單純易懂，所需的演算法也單純。代表性的方法包括[1]所提出的 Cluster-based Similarity Partitioning Algorithm (CSPA)以及[3]的 EAC。其中 EAC 使用階層聚合(hierarchical agglomeration)來從 co-association 矩陣算出最終分群。與其他的叢集整合演算法比較，EAC 具有以下幾個優點：（一）利用如單一連結 (single-link; SL) 或平均連結 (average-link; AL) 的階層聚合，可以找出具有延伸性、任意形狀的叢集；（二）不需事先指定最終分群的叢集個數，而可利用階層聚合的最大生命期條件來選擇最終分群。

EAC 所用的 co-association 矩陣最大的問題即是其運算複雜度（至少 $O(N^2)$，N 爲資料量）。由於此矩陣所含的資訊有許多重複性，若能利用這些重複性來降低 co-association 矩陣大小而減少運算量，則可大大提升 EAC 與其他基於 co-association 矩陣的演算法的實用性。這方面的成果是本計畫最主要的貢獻。

一般使用叢集雛形的叢集演算法有許多用於偵測特定形狀（如直線、線段、曲線、長方形、球面/橢球面、甚至基於樣本的任意形狀）叢集的應用。由於有特定形狀叢集的問題一般而言較偵測緊密叢集(compact cluster)更困難，叢集整合是一個極有可能在這些問題上帶來更穩定及可靠的結果的方向。本報告中，亦包含了我們利用叢集整合偵測線段與主曲線的演算法與結果。

整體而言，這個研究計畫的重要性在於，針對叢集整合中最容易實作也最常被使用的基於 co-association 矩陣的方法，一方面透過加速技術提升其在大量資料問題的應用性，一方面將其與用在偵測特定形狀的叢集演算法結合來增加其應用的範圍。我們將對基於 co-association 矩陣的叢集整合的特性有更多了解，也能夠讓這個新方法的優點能夠更有效的應用在許多必須以叢集演算法分析的問題上，而得到比過去更好的結果。

二、研究內容與結果

- 關於 co-association 矩陣的定義：

假設集合 $X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 代表了 N 個資料點。一個 X 的清晰分群 (crisp partition) 可以由共 k 個 X 的子集合 $C_1, C_2, \ldots, C_k$ （即叢集）來表示，每個資料點只屬於一個叢集，而每個叢集至少包含一個資料點。我們用 $P=\{C_1, C_2, \ldots, C_k\}$ 代表這樣一個分群。每一次執行叢集演算法可以得到一個不同的 P。一個叢集整合包含了 H 個個別分群：$P_1, P_2, \ldots, P_H$。每個個別分群當中的叢集個數都可以是不同的，因此我們以 k_h 來代表分群 P_h $(1\le h\le H)$ 當中的叢集個數。

整個叢集整合的 co-association 矩陣 $\boldsymbol{S}=[s_{ij}]$ 的計算方式如下：

$$s_{ij}=\frac{1}{H}\sum_{1\le h\le H} s_{ij}^{(h)} \tag{5}$$

其中

$$s_{ij}^{(h)}=\begin{cases}1, & \lambda_i^{(h)}=\lambda_j^{(h)}\\ 0, & otherwise\end{cases} \tag{6}$$

這裡λ_{ih}代表在 P_h 當中，資料點 $\boldsymbol{x}_i$ 所屬的叢集的編號或標籤(label)。在此我們進一步定義資料點 $\boldsymbol{x}_i$ 的叢集標籤向量(cluster label vector)：

$$\boldsymbol{\lambda}_i=\begin{bmatrix}\lambda_{i1} & \lambda_{i2} & \cdots & \lambda_{iH}\end{bmatrix} \tag{7}$$

- 叢集結果的正確度分析：

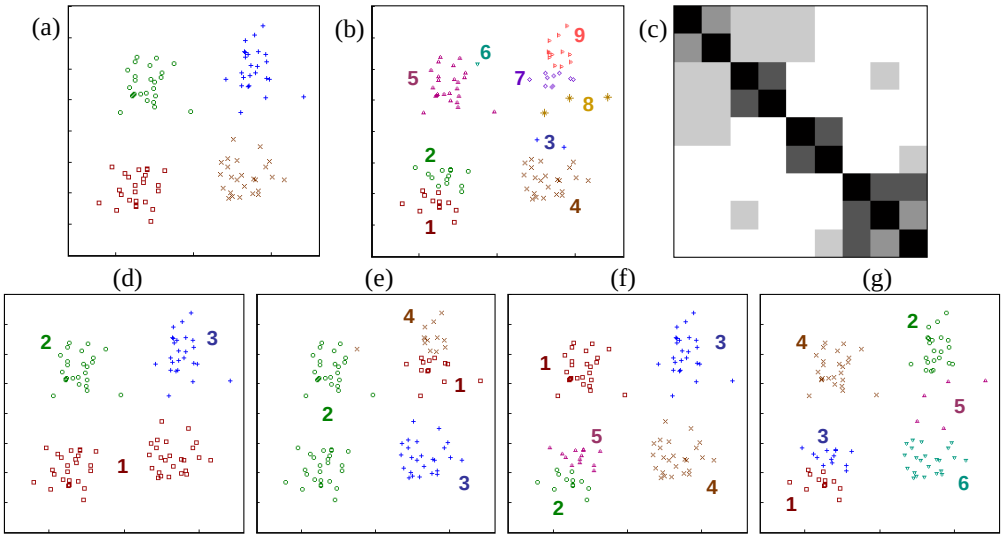
爲了能計算叢集結果正確程度，在此我們必須使用已知每個資料點所屬叢集的模擬資料，以及使用於分類研究的、有分類標準答案的資料組。我們採用的是通稱的叢集準確度 (clustering accuracy)。這是把叢集化結果當成分類結果來與已知的標準答案比較所算出的分類正確程度。這做法的問題在於叢集標籤與分類標籤之間的對應關係並非已知。解決方法是使用 Hungarian 演算法來求得一組最佳的對應關係與其準確度。

- 對基於 co-association 矩陣的叢集整合演算法的加速技術：CA-tree

我們進行加速的基本假設是在於組成叢集整合的各個個別分群之間的相關性或資訊重

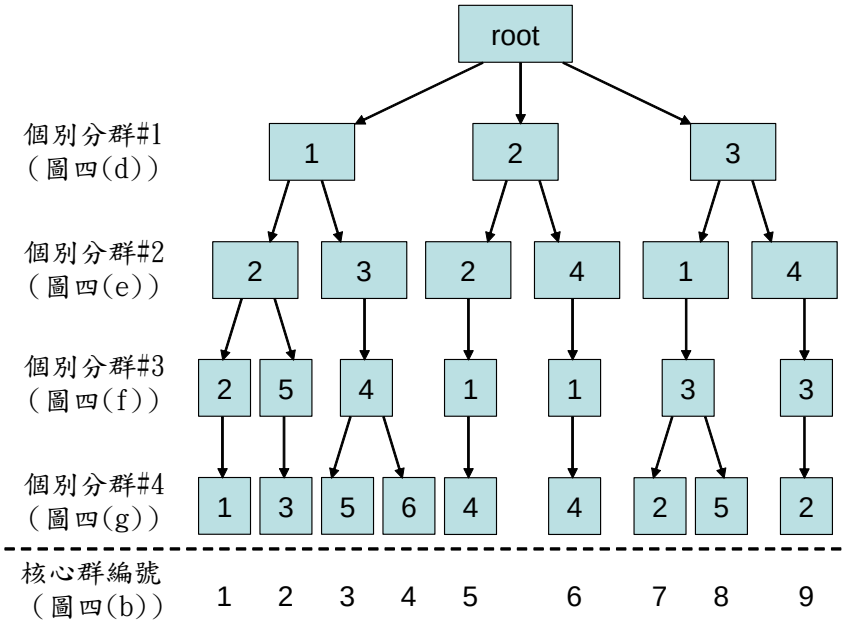
複性。我們針對這一點舉例說明如下：假如有兩個資料點 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 很接近到一個程度，在每個個別分群當中，這兩個資料點都被分配到同一個叢集當中。在這種情形下，就叢集整合而言，$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是無法分辨的。因此，我們可以把 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 合併成一個新的資料點看待，因而減少了得到最終分群的計算量。

以下我們以一個例子來進一步說明。圖一(a)是一個包含 100 個點的資料組。這 100 個點大約平均分配在 4 個叢集當中。圖一(d)-(g)是 4 個使用 k 均值法所得到的分群，其中叢集的總數是從 3 到 6。各圖中皆以不同的顏色與符號代表各點所屬的叢集。在圖一(b)當中我們將資料點分成 9 個核心群 (core group)。一個核心群的定義是其中每個資料點的叢集標籤向量都是相同的，也就是說，這些資料點在所有個別分群中都被分到同一個叢集。因此，在將所有 co-association 矩陣當中重複的行與列合併後，我們得到的是一個 9x9 的矩陣（圖一(c)），而非原來 100x100 的矩陣，而且仍保有所有的資訊。這對降低計算量的效果是顯而易見的。



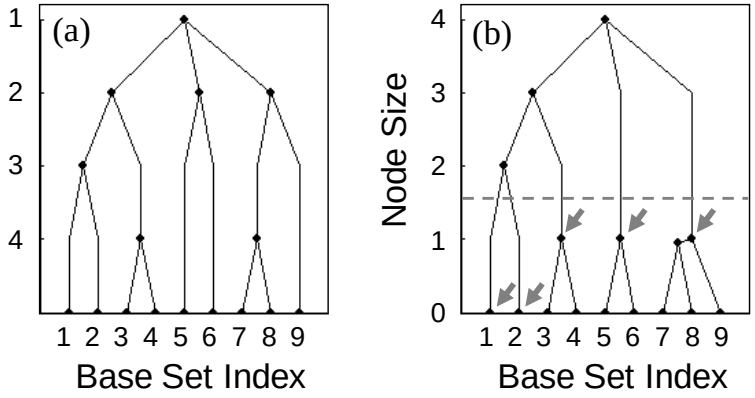
圖一：(a)資料組；(b) 9 個核心群（數字爲核心群的編號）；(c) 由核心群算出的 co-association 矩陣；(d)-(g)四個個別分群（數字爲每個分群中的叢集標籤）。

在以上的範例中，一個直覺的做法是先建立完整 100x100 的 co-association 矩陣，再將重複的行與列去除。然而，這樣做的缺點是我們仍然必須先處理一個有 N^2 個元素的矩陣，因此運算複雜度仍然是 $O(N^2)$，並不能真正達到我們降低複雜度的目的。在這方面我們已設計了一個階層式的方法，藉由建立一個樹狀結構，直接將所有資料點分成核心群。這樹狀結構在根節點(root node)之下的每一層對應到一個個別分群，而每一個葉節點(leaf node)則對應到一個核心群。雖然會因爲各個別分群加入的順序不同而得到不同的樹狀結構，然而所得到的核心群則不受影響。以下圖二是根據圖一的資料與個別分群所產生的樹狀結構（個別分群的順序是圖一(d)-(g)），各節點當中的數字是在該個別分群的中的叢集標籤，而最下方的數字是核心群的編號（對應到圖一(b)）。對於每個核心群，將其從根節點到葉節點的路徑上所有的叢集標籤組合即是其叢集標籤向量。此樹狀結構即是我們的 CA-tree。



圖二：根據圖四的分群所產生的樹狀結構。各節點當中的數字是在該個別分群的中的叢集標籤。

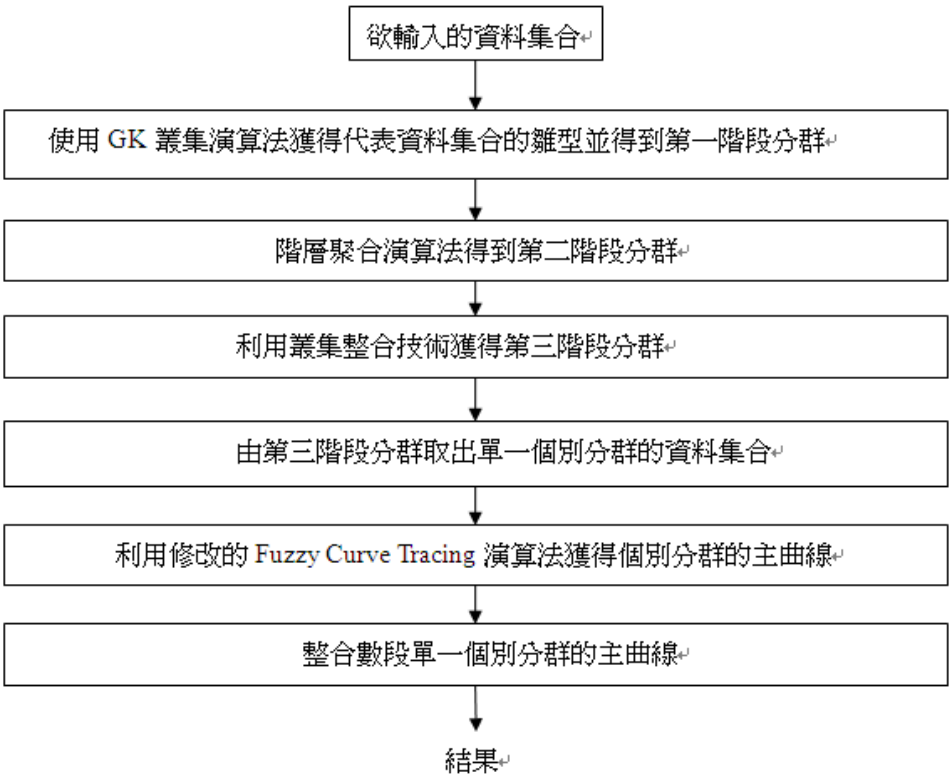
當我們單純考慮運算複雜度與資料量的關係，由個別分群產生此樹狀結構的部分爲 $O(N)$，而由核心群產生簡化的 co-association 矩陣並計算最終分群的部分爲 $O(N_{cg}^2)$，其中 N_{cg} 是核心群的個數。然而對於 N_{cg} 我們只能確定 $N_{cg} \le N$。此外對大量資料與有較多的個別分群的情況而言，N_{cg} 仍然可能太大而降低實用性。因此接下來我們要研究的是進一步將相似的核心群組合得到近似的核心群（即每個近似核心群中各資料點的叢集標籤向量並非完全一樣，但我們可以限制差異的程度），如此可以進一步降低運算量。針對這個目的，我們定義了上圖二中 CA-tree 每個節點在叢集標籤空間中的大小（根據 Hamming 距離計算）。下圖三中的左圖是一個由圖一測試資料得到的 CA-tree，而右圖則對應到每個節點的大小。虛線是一個 threshold，可以藉以取出一組近似的核心群（見箭頭）。



圖三：(a) CA-tree 範例；(b) 對應之節點大小，與取近似核心群的範例。

- 使用叢集整合偵測主曲線

我們主要的做法是使用個別小線段來組合成主曲線，在組合個別線段的階段利用了叢集整合來增加其準確率。我們的流程圖如圖四所示：



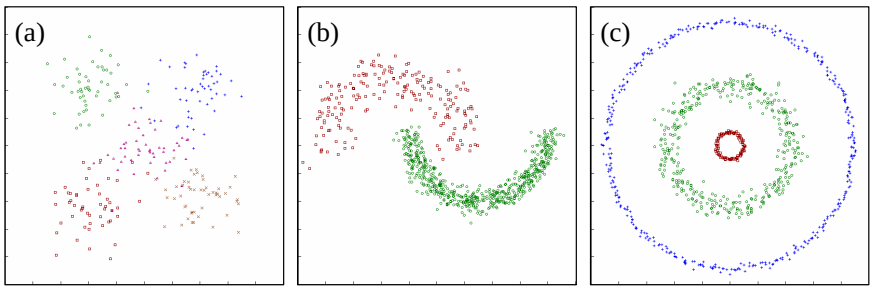
圖四: 使用叢集整合偵測主曲線的流程圖

三、結果與討論

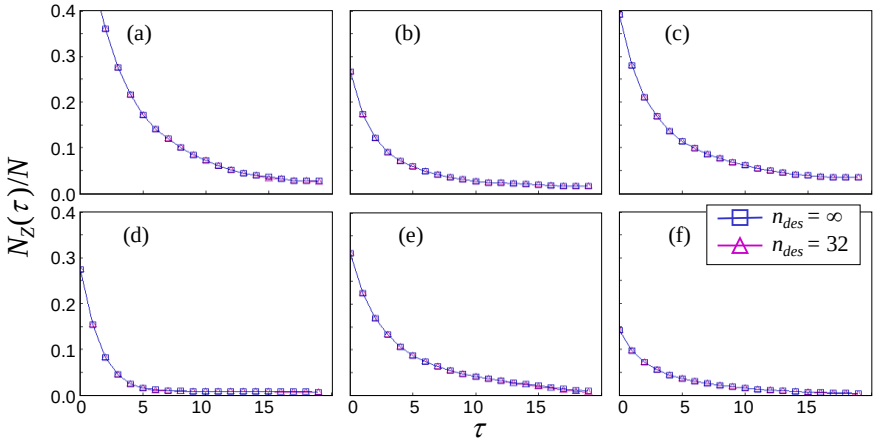
針對 CA-tree 的部分，我們使用六個資料組（見表一）進行結果分析。其中三個是自行產生的資料組（見下圖五）。在圖六及圖七中，我們分別顯示了在不同 threshold（即圖中之 τ）之下，近似核心群個數相對於原資料點個數的比例，以及叢集化準確率的結果。我們可以看出即使只有極少的近似核心群，仍能維持極好的叢集化準確率。

表一：資料組

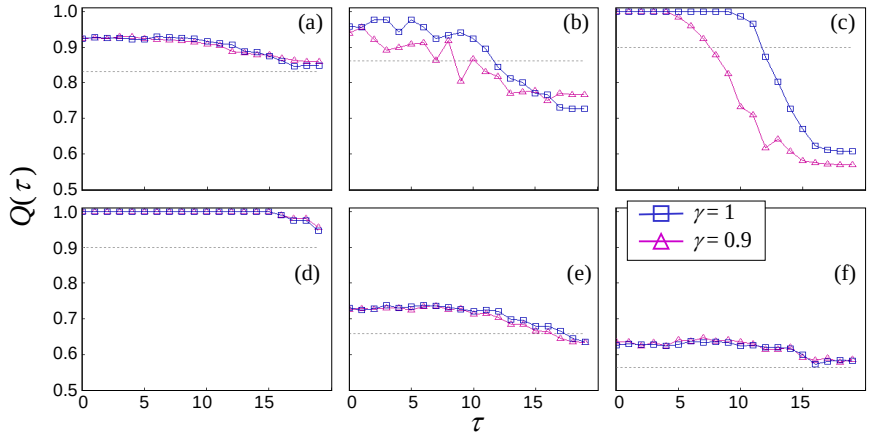
資料組名稱	資料點個數	正確叢集個數
Spherical5	250	16
Half-rings	800	2
3Rings	900	2
8d5k [10]	1000	8
Opt-digits	3823	64
Pen-digits	10992	16



圖五：自行產生之資料組。

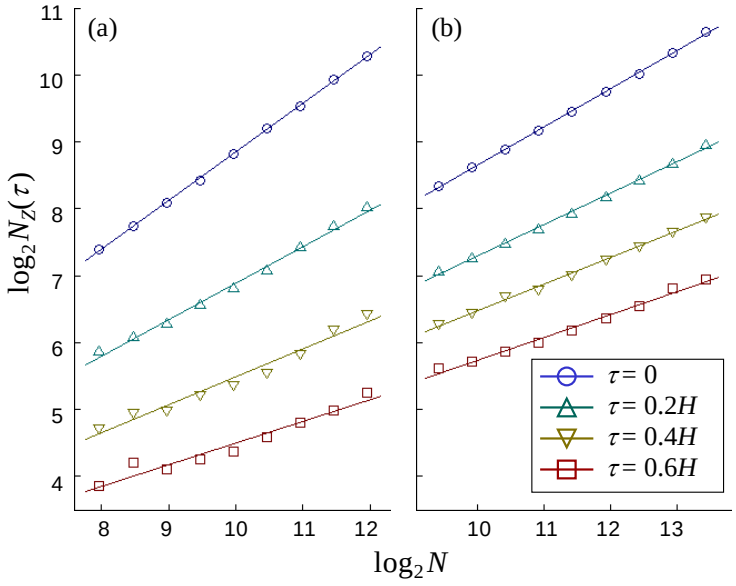


圖六：對於六個資料組，在不同 threshold（τ）之下，近似核心群個數相對於原資料點個數的比例。



圖七：對於六個資料組，在不同 threshold（τ）之下的叢集化準確率。

我們進一步的實驗分析發現 CA-tree 運算量與原始資料點數量之間存在近似於對數的關係（見圖八）。因此，在絕大多數情形下，計算複雜度將可降到 $O(N)$（主要用在產生 CA-tree 本身）。這對效率與實用性乃是極大的改進。爲了進一步驗證，我們將 CA-tree 用在影像的分割上，得到很好的結果（見圖九）。

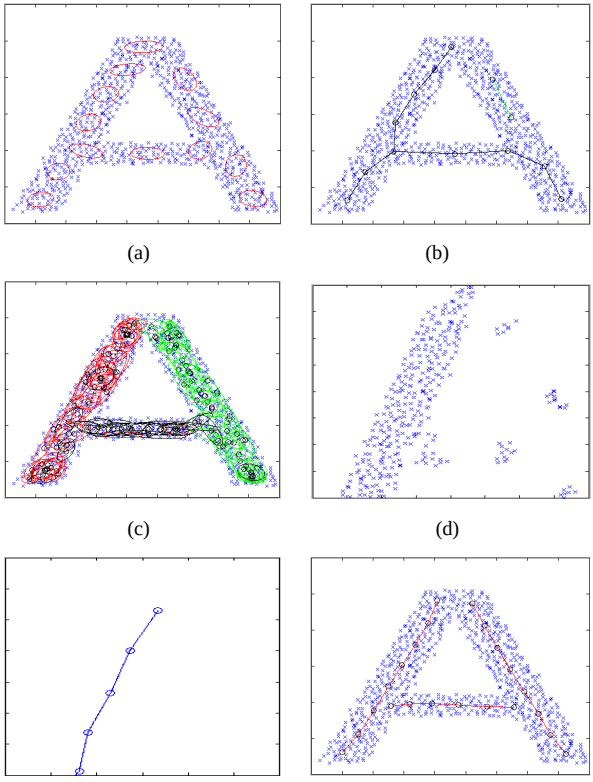


圖八：近似核心群個數與資料點個數之關係。

(a) (b)

圖九：EAC 結合 CA-tree 用在影像像素分割的範例。(a)原圖；(b)分割結果（三群）。

對於以叢集整合偵測主曲線的部分，以下圖十是圖四流程圖的結果範例。



(a) (b)

(c) (d)

(e) (f)

圖十：流程圖（圖四）各個步驟所產生的結果。(a)第一階段分群結果；(b)第二階段分群結果；(c)第三階段分群結果；(d)紅色個別分群的資料集合；(e)從(d)所得到的主曲線；(f)整合數段主曲線所得到的結果。

參考文獻

[1] A. Strehl and J. Ghosh "Cluster ensembles -- a knowledge reuse framework for combining multiple partitions", *J. Machine Learning Research*, vol. 3, pp. 583-617, 2002.

[2] A.P. Topchy, M.H.C. Law, A.K. Jain, and A.L. Fred, "Analysis of consensus partition in cluster ensemble", *Proc. 4th IEEE Int'l Conf. Data Mining* (ICDM), pp. 225-232, 2004.

[3] A.L.N. Fred and A.K. Jain, "Combining multiple clusterings using evidence accumulation", *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 27, pp. 835-850, 2005.

[4] H. Luo, F. Kong, and Y. Li, "Combining multiple clusterings via k-modes algorithm", *LNAI*, vol. 4093, pp. 308 – 315, 2006.

[5] L.I. Kuncheva and D.P. Vetrov, "Evaluation of stability of k-means cluster ensembles with

respect to random initialization", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1798-1808, 2006.

[6] X.Z. Fern and C.E. Brodley, "Random projection for high dimensional clustering: A cluster ensemble approach", *Proc. 20th Int'l Conf. Machine Learning* (ICML), 2003.

[7] B. Minaei-Bidgoli, A. Topchy and W. F. Punch, “Ensembles of partitions via data resampling”, *Proc. 2004 Int'l. Conf. Information Technology*, pp. 188-192, 2004.

[8] P. Hore, L. Hall, .and D. Goldgof, "A Cluster Ensemble Framework for Large Data sets", *Proc. 2006 IEEE Int'l. Conf. System, Man, and Cybernetics*, pp. 3342-3347, 2006.

[9] P. Viswanath and K. Jayasurya, "A fast and efficient ensemble clustering method", *Proc. 2006 Int'l Conf. Pattern Recognition* (ICPR), pp. 720-723, 2006.

[10] P. Kellam, X. Liu, N.J. Martin, C. Orengo, S. Swift, and A. Tucker, "Comparing, contrasting and combining clusters in viral gene expression data," *Proc. 6th Workshop on Intelligent Data Analysis in Medicine and Pharmocology*, pp. 56-62, 2001.

五、計畫成果自評
V. Self-Evaluation

本研究計畫在叢集整合的研究上為一新的方向，有別於先前的相關研究，故具有學術論文發表的價值。CA-tree的結果目前已發表於國際期刊（IEEE Transactions on Systems, Man, and Cybernetics, Part B）。此一計畫的執行，共有四位碩士班的學生參與，對於人才的培養，多有助益。

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：王才沛	計畫編號：99-2221-E-009-179-
計畫名稱：基於證據累加的叢集整合技術之強韌化與功能延伸 II	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	4	4	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	1	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

作為一個較新的領域，叢集整合的研究主要在於其特性與演算法的分析，但在各領域中也逐漸被應用於實用問題。對於叢集整合而言，最常見的演算法是依據 co-association 矩陣，然而這類方法計算量大而不實用。我們研究的主要成果 CA-tree 可以減少運算量並維持準確率，因此對於將叢集整合用在更多實用問題上，有相當大的幫助。