

朝陽科技大學
資訊管理系

碩士論文

群聚參數與群聚適切性的分析與應用

The Analysis and Applications of Cluster Parameters
and Cluster Validation

指導教授：陳榮昌 博士

研究 生：王浩永

中華民國九十三年七月十四日

朝陽科技大學資訊管理系
Department of Information Management
Chaoyang University of Technology

碩士論文
Thesis for the Degree of Master

群聚參數與群聚適切性的分析與應用
The Analysis and Applications of Cluster Parameters
and Cluster Validation

指導教授：陳榮昌博士(Rong-Chung Chen)
研究 生：王 浩 永(Hao-Yun Wang)

中華民國九十三年七月十四日

14, July 2004



博、碩士論文授權書
(國科會科學技術資料中心版本, 93. 2. 6)

本授權書所授權之論文為本人在 朝陽科技 大學(學院) 資訊管理 系所 組 九十二 學年度第 二 學期取得 碩 士學位之論文。

論文名稱：群聚參數與群聚適切性的分析與應用

同意 不同意

本人具有著作財產權之論文全文資料，授予行政院國家科學委員會科學技術資料中心（或其改制後之機構）、國家圖書館及本人畢業學校圖書館，得不限地域、時間與次數以微縮、光碟或數位化等各種方式重製後散布發行或上載網路。

本論文為本人向經濟部智慧財產局申請專利（未申請者本條款請不予以理會）的附件之一，申請文號為： ，註明文號者請將全文資料延後半年後再公開。

同意 不同意

本人具有著作財產權之論文全文資料，授予教育部指定送繳之圖書館及本人畢業學校圖書館，為學術研究之目的以各種方法重製，或為上述目的再授權他人以各種方法重製，不限地域與時間，惟每人以一份為限。。

上述授權內容均無須訂立讓與及授權契約書。依本授權之發行權為非專屬性發行權利。依本授權所為之收錄、重製、發行及學術研發利用均為無償。上述同意與不同意之欄位若未鈞選，本人同意視同授權。

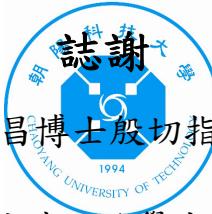
指導教授姓名：陳榮昌 副教授

研究生簽名：王浩永
(親筆正楷)

學號：9114635
(務必填寫)

日期：民國 93 年 07 月 14 日

1. 本授權書(得自
<http://sticnet.stic.gov.tw/sticweb/html/theses/authorize.html> 下載或至
<http://www.stic.gov.tw> 首頁右下方下載) 請以黑筆撰寫並影印裝訂於書名頁之次頁。
2. 授權第一項者，請確認學校是否代收，若無者，請個別再寄論文一本至台北市(106-36)和平東路二段 106 號 1702 室 國科會科學技術資料中心 黃善平小姐。(電話:02-27377606 傳真:02-27377689)。



本文承蒙指導教授陳榮昌博士殷切指導，在學業與論文研究上諸多啟迪，使本論文終能順利完成。於學生遇到挫折時耐心的給予關懷與支援、細心輔導學生解決困難，謹誌於此，以此表示由衷之敬意與謝忱。此外，更感謝中正大學游寶達教授、本校資訊管理系王淑卿教授及網路通訊所呂瑞麟教授，在計畫口試及論文口試期間對本論文不吝指正並提供寶貴的建議。

在研究的路上還要感謝育臣學長、升助學長、思齊學長、健三學長、溢桐學長、小麥學長的引領與鼓勵，與布魯斯、塔克、小查、大郭、文慶、傑克達、小白、伸豐的並肩奮戰，蓓蓓、果子、恰吉、家惠、惠菁、玉昀、淑蓉的體貼關心，淨雅、慧敏、小語、KIKI的包容鼓勵與阿玄的多媒體倉庫、俊德的愛車、yy的勸敗、啟琳的團購、孝的JAZZ、靜婷的小道消息與大里霧峰地區餐飲業的無限加飯，在朝陽漫長的兩年裡所有關於喜怒哀樂的平凡小事，能有你們的陪伴我想我一直是幸運地。

最後感謝撫育我的父母親及家人，由於你們的鼓勵讓我無後顧之憂的實現自己的夢想，謝謝你們支持我走過求學生涯中最有意義的時期；僅以本文獻給所有關心我的人。



本研究目的在於針對現今群聚技術的不足提出一個資料導向為基礎之選擇群聚演算法類型方法與一個新群聚適切性評估式，以降低盲目選擇群聚演算法及其輸入參數所造成的錯誤。隨著資訊化的普及，使得企業收集了大量的資料，透過對此資料的分析可以挖掘出許多有用的決策資訊，成為企業獨特的商業智慧。其中常使用群聚技術來挖掘出隱藏在資料內部的資訊，藉由此技術鑑別出資料中相似的群體，並幫助使用者發現資料中分佈特徵與感興趣的關係；然而隨著資料庫維度及資料量不斷的上升，使得群聚技術及其輸入參數的選擇已成為一種試誤的流程，有鑑於此本研究提出兩種方法解決之，其一為藉由輸入資料的密度、混亂度及分離度三指標，輔助使用者有效地選擇群聚演算法之方法；其二為提出一利用群聚間的密度率、凝聚率及鑑別率進行群聚結果的評估，以挑選出最合適的輸入參數，最後透過本研究提出的兩個方法使得群聚技術的應用上更符合實際的需求並提升群聚品質的目的。

關鍵字：群聚技術、群聚參數、群聚適切性評估



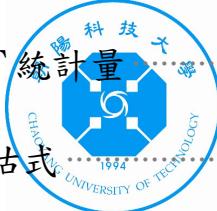
The research proposes a data oriented approach for choosing the type of clustering algorithms and a new cluster validity index for choosing their input parameters.

The clustering technology is often used to discover the patterns and interested relations (or the hidden information) in a data set. However, with the increasing of data complexity, the traditional cluster technology has been changed to a try-and-error process. For this reason, we propose two approaches to improving these problems. Firstly, we use three indexes which include density rate, entropy, and coefficient of variation to measure the input data so as to help the users to choose an appropriate type of clustering algorithm. On the other hand, we propose a new criterion which is based on both the factors of between-clusters and within-cluster to evaluate the results of clustering algorithms. The experimental results show that our new approaches do improve performance of clustering both on the consideration of practicability and quality.

Keywords: Clustering Technology, Cluster Parameters, Cluster Validation



第壹章 緒論	01
一、研究背景與動機	01
二、研究目的	03
三、研究架構	04
四、論文架構	04
第貳章 文獻探討	06
一、群聚技術回顧	06
(一) 切割式群聚類型	07
(二) 階層式群聚類型	08
(三) 密度式群聚類型	10
(四) 格子式群聚類型	11
(五) 混合式群聚類型	12
(六) 群聚演算法類型與群聚品質之比較	14
二、群聚適切性評估	17
(一) 群聚演算法參數問題描述	17
(二) 群聚適切性量測準則	18
(三) 群聚適切性評估式	19
(1) CH 評估式	19



(2) Hubert's Γ 統計量評估式	19
(3) Dunn's 評估式	20
(4) DB 評估式	21
(5) 混合型評估式	22
 三、模糊控制理論	23
(一) 模糊控制器基本架構	23
(1) 模糊化單元	24
(2) 模糊規則庫	24
(3) 模糊推論引擎	24
(4) 解模糊化單元	25
 第參章 研究方法	26
 一、資料導向為基礎之選擇群聚演算法類型方法	26
(一) 各種群聚演算法類型適合之資料分佈	27
(二) 資料密度指標	30
(三) 資料混亂度指標	31
(四) 資料分散度指標	33
(五) 資料指標與群聚演算法類型	34
(六) 資料導向為基礎之群聚演算法選擇模型	35
 二、群聚適切性評估式	38



(一) 密度率.....	41
(二) 鑑別率.....	42
(三) 凝聚率.....	43
(四) 適應型群聚適切性評估式.....	45
第肆章 實驗與結果.....	47
一、資料導向為基礎之選擇群聚演算法方法.....	47
(b) 資料密度指標.....	47
(c) 資料混亂度指標.....	48
(d) 資料分離度指標.....	51
(e) 實驗範例.....	55
二、群聚適切性評估式.....	58
第五章 結論	68



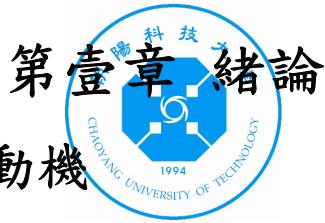
圖 1-1 群聚處理流程步驟	02
圖 1-2 研究方法流程圖	04
圖 2-1 鏈結效應	10
圖 2-2 密度式群聚演算法示意圖	11
圖 2-3 格子式演算法示意圖	11
圖 2-4 STING 示意圖	12
圖 2-5 SOM 網路圖	13
圖 2-6 不同參數的群聚結果	18
圖 2-7 模糊控制器基本架構	23
圖 3-1 合適於切割式群聚演算法的資料分佈	27
圖 3-2 合適於階層式群聚演算法的資料分佈	28
圖 3-3 合適於密度式群聚演算法的資料分佈	28
圖 3-4 Isothetic rectangle 空間示意圖	31
圖 3-5 資料導向之群聚演算法選擇模型	36
圖 3-6 原始資料各輸入變數之隸屬函數	36
圖 3-7 兩群聚重心點距離相等，但群聚半徑不同	39
圖 3-8 兩群聚最遠點距離相等，但群聚半徑不同	39
圖 3-9 兩群聚最近點距離相等，但群聚半徑不同	39

圖 3-10 群聚資料空間示意圖	41
圖 3-11 Single link 距離相等, Centroid link 距離不同	43
圖 3-12 群聚凝聚率示意圖	43
圖 3-13 群聚內資料點分佈圖	44
圖 3-14 適應型群聚適切性評估式示意圖	46
圖 4-1 混亂度資料一之資料分佈	48
圖 4-2 混亂度資料二之資料分佈	48
圖 4-3 混亂度資料三之資料分佈	49
圖 4-4 混亂度資料四之資料分佈	49
圖 4-5 混亂度資料五之資料分佈	50
圖 4-6 混亂度資料六之資料分佈	50
圖 4-7 分離度資料一之資料分佈	51
圖 4-8 分離度資料二之資料分佈	51
圖 4-9 分離度資料三之資料分佈	52
圖 4-10 分離度資料四之資料分佈	52
圖 4-11 分離度資料五之資料分佈	53
圖 4-12 分離度資料六之資料分佈	53
圖 4-13 分離度資料七之資料分佈	54
圖 4-14 分離度資料八之資料分佈	54

圖 4-15 範例一資料分佈	55
圖 4-16 範例一分群結果示意圖	56
圖 4-17 範例二資料分佈	56
圖 4-18 範例二分群結果示意圖	57
圖 4-19 適切性評估式資料一之資料分佈	59
圖 4-20 適切性評估式資料一 (a) 之分群結果	59
圖 4-21 適切性評估式資料一 (b) 之分群結果	60
圖 4-22 適切性評估式資料二之資料分佈	61
圖 4-23 適切性評估式資料二 (a) 之分群結果	62
圖 4-24 適切性評估式資料二 (b) 之分群結果	63
圖 4-25 適切性評估式資料三之資料分佈	64
圖 4-26 適切性評估式資料三 (a) 之分群結果	65
圖 4-27 適切性評估式資料三 (b) 之分群結果	66



表 2-1 群聚演算法與參數需求比較表	16
表 3-1 資料導向之選擇群聚演算法規則庫	37
表 4-1 資料密度分析表	45
表 4-2 適切性評估資料一 (a) 之評估值	59
表 4-3 適切性評估資料一 (b) 之評估值	61
表 4-4 適切性評估資料二 (a) 之評估值	63
表 4-5 適切性評估資料二 (b) 之評估值	64
表 4-6 適切性評估資料三 (a) 之評估值	65
表 4-7 適切性評估資料一 (b) 之評估值	67



一、研究背景與動機

資訊科技的日漸普及使得企業能夠收集到大量且複雜的交易記錄，這些記錄中隱含著豐富的資訊，透過完善的分析（如統計方法、資料挖掘或機器學習等方法），我們便可以發現許多有用的決策知識，並且隨著這些知識的累積而逐漸形成為企業獨特的商業智慧。其中群聚技術（Cluster technology）便是一項經常被利用來挖掘隱藏資訊的重要方法，它藉由鑑別出資料集合中相似的群體幫助使用者了解群聚中共同特徵與令人感興趣的關係。而近年來隨著群聚技術的成熟，其應用領域也逐漸廣泛，但是在各種群聚演算法的選擇及其參數的適切性等，仍因使用環境的不同而受限，所以本研究將針對以上的限制提出一些解決的概念與方法。

隨著群聚相關研究與實務應用上的落實，群聚技術不僅是獨立的一環，更形成了所謂的群聚處理流程[10]（圖 1-1），它包含特徵選擇、選擇群聚演算法、群聚結果的適切性與解釋群聚結果，詳述如下：

- (1) 特徵選擇：其目的在於選擇出對於後續的分析能夠提供較高貢獻度的特徵屬性，由於過多無用的特徵屬性不但造成計算資源上的負擔甚至會誤導分析的結果，所以通常要先經過前置處理的作業來將資料複雜度降低。

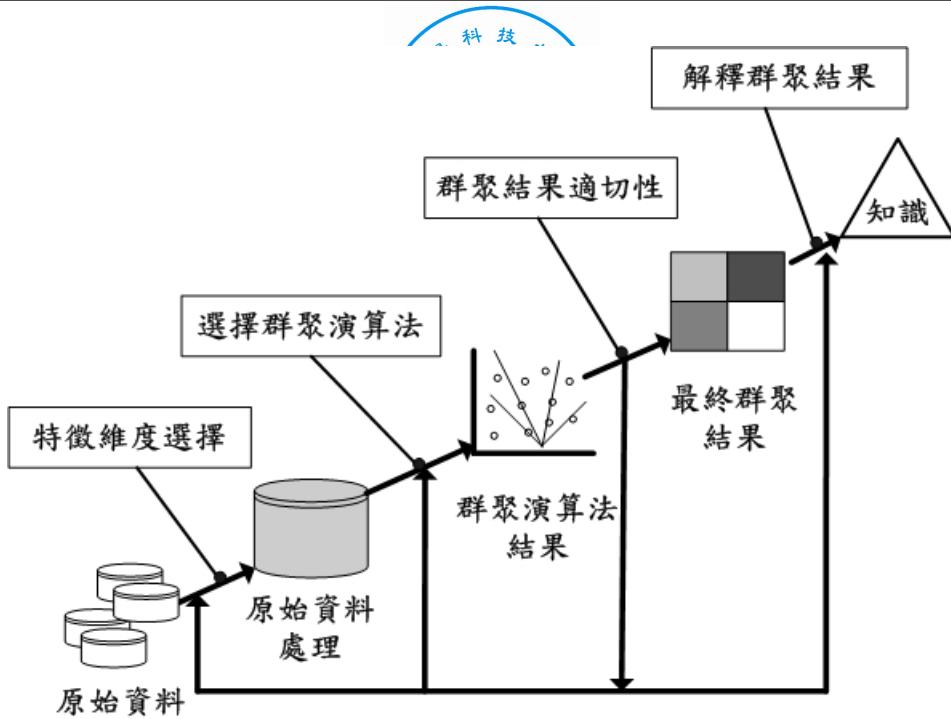


圖 1-1 群聚處理流程步驟 [10]

(2) 選擇群聚演算法：群聚演算法的目的在於分析資料的特徵，將性質相似的資料分配在相同群聚中，並且將那些性質相異的資料分配在不同的群聚中，其瓶頸在於如何從眾多的群聚演算法中選擇一個較合適的演算法；另外，大部分的群聚演算法皆需要輸入參數，不同的參數組合會造成群聚品質的優劣；如此使得群聚技術變成一種試誤（Try-and-error）的過程，這也導致其群聚結果信賴度降低。

(3) 群聚結果適切性：此步驟是為了修正群聚演算法中參數選擇而造成的錯誤群聚結果，其方法在於使用合宜的評估準則及方法以鑑別出適當的群聚結果。然而，其中最大的問題在於人為介入所產生的決策錯誤，當使用者不了解自身需求時，便無法選

擇出適用的評估準則，這也將導致錯誤的群聚結果。

- (4) 解釋群聚結果：在許多應用群聚技術的情況中，使用者為了解釋正確的群聚結果，便有賴於此應用領域中的專家與其他系統結果整合後加以解釋和闡明，但是過多人為的介入反而提高解釋的誤差與風險。

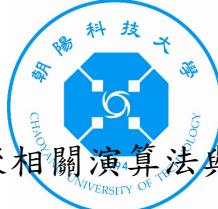
由以上四步驟中可知每個步驟皆有其困難與不足之處[10]；本研究將針對選擇聚演算法與群聚結果的適切性，此二步驟進行討論與研究。

二、研究目的

在上一節中提到許多的群聚實際應用應用的問題（例如群聚參數的選擇、群聚演算法結果的不確定性以及群聚技術在應用時過多的人為介入等），因此本研究將針對現今常用的群聚演算法與群聚適切性評估式研究和探討，並提出一個資料導向選擇群聚演算法的方法以及提出較具適應性的群聚適切性評估式，以解決盲目選擇群聚演算法所造成的錯誤與選擇群聚參數時遭遇的問題。根據上述動機本研究欲達到的目的如下所示：

- (1) 研究與探討各類型群聚演算法的特性與優劣。
- (2) 提出一個資料導向之選擇群聚演算法類型的方法。
- (3) 提出較具適應性的群聚適切性評估式。

三、研究架構



本研究架首先針對群聚相關演算法與群聚適切性評估式進行詳細討論，接著針對其特性作一番分析並試著提出方法解決其不足之處；最後提出一個藉由對輸入資料中特性分析進而找尋出選擇群聚演算法的方法，並根據群聚間與群聚內的量測方法，提出一個較具適應性的群聚適切性評估式，本研究之方法流程圖如圖 1-2。

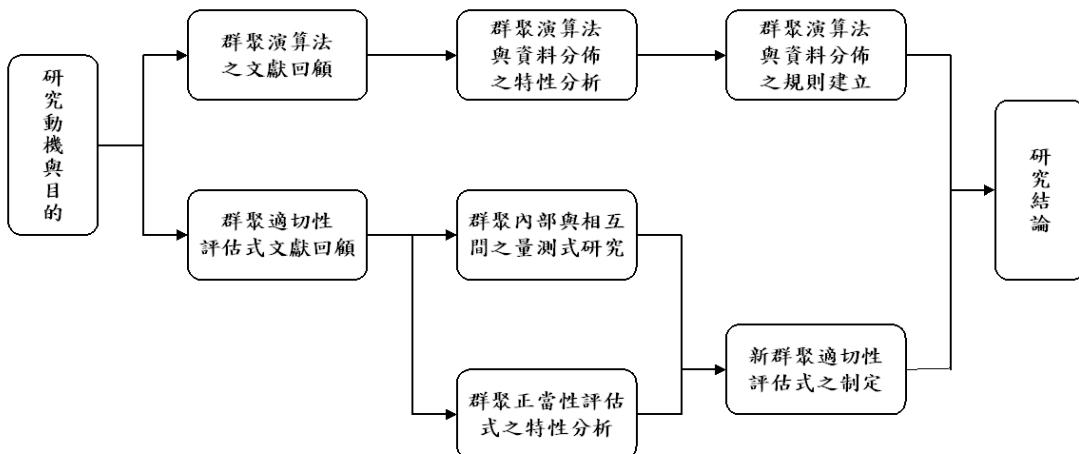


圖 1-2 研究方法流程圖

四、論文架構

本論文架構依照上節中的研究架構將其區分為五個章節，第壹章為緒論，將說明本研究之動機、目的與步驟及整篇論文架構；第貳章為文獻探討，將針對以往群聚演算法與群聚適切性評估式進行描述與分析；第參章為研究方法，此章中本研究提出一個以資料導向之選擇群聚演算法類型的方法與一個根據群聚內部及相互間的量測方法為

基礎的群聚適切性評估式；第肆章為實驗結果，對於提出的方法進行實驗模擬和驗證；第伍章為結論與未來研究方向，其內容敘述研究成果與貢獻，並提出未來可研究與改進之相關工作。



第二章 文獻探討

此章中將文獻回顧區分為三部份；第一部份為常見群聚技術回顧，第二部份為常見群聚適切性評估式描述與分析，第三部分為模糊控制理論，詳述如下。

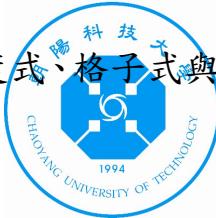
一、群聚技術回顧

在資料探勘的領域中群聚技術是重要的技術之一，因為藉由群聚技術可以將性質相似的資料分配在相同群聚中，而性質相異的資料則分配在不同的群聚中；尚可針對大型資料庫進行群聚，例如地理資訊、空間資料庫、多媒體資料庫等中探索出資料的特徵分佈；但隨著大型資料庫的資料筆數與維度的增加，使得大型資料庫之群聚演算法效率越來越受到重視。

另外，群聚技術與分類（Classification）是差異在於使用者必須使用群聚演算法來找尋資料內部特徵，而其群聚結果能讓使用者能夠解釋其意義，反之分類是事先定義其類別的標籤，根據類別分類從結果而論，兩種方法皆是讓資料分類，從過程的角度觀察，兩種方法運作的模式是相反的[11]。

在群聚技術中有許多種不同的類型的群聚演算法，當使用者在選用群聚演算法時，必須了解其演算法特性、限制，並應用於適合的領域上，才能獲得到正確的群聚結果；本研究使用[11]的分類方式將其

區分為切割式、階層式、密度式、格子式與混合式的群聚演算法類型，詳述如下。



(一) 切割式群聚類型 (Partition based)

切割式群聚演算法的是將相似的資料切割到同一個群組裡，以建立各個群聚的集合；此類型演算法企圖把原始資料切割成 k 個子集合，因此在相同的集合內和其他相異集合的成員必定明顯地不相同；而相同集合內的成員具有一定程度的相似之處，這樣的集合我們稱之為一個群聚，其優點在於演譯過程中直觀而且具說服力。

常見的切割式演算法有 K-means[15]、PAM [15] 、CLARA[15] 及 CLARANS[22]各別詳述如下。K-means 它是以群聚重心作為群聚的代表點（Representative object），然而它的群聚結果易受到雜訊（Noises）或是離群值（Outliers）所影響。另外，以選擇最靠近重心之資料點作為代表點的 PAM，則對於小型資料集合有不錯的處理能力，但是隨著資料筆數增加，處理的效率也隨之降低，所以便有針對大型資料庫採用取樣方式的 CLARA 與 CLARANS，然而取樣方法將會受到樣本數量多寡以及挑選樣本方式所影響。尚有針對資料中具有不同屬性類型的群聚演算法，如針對非連續型屬性的 K-mode 及具備 K-means 與 k-medoids 優點及非連續型屬性的 K-prototypes。

(二) 階層式群聚類型 (Hierarchical based)

階層式演算法類型是以樹狀架構呈現資料點間相似或相異的程度，其表示方法分別為由下至上的凝聚法（Agglomerative）及由上至下的分裂法（Divisive）。以凝聚法為基礎的階層式群聚演算法在初始狀態先將單一資料點皆視為一個群聚，接著依資料點間相似程度做為合併條件，每一回合將兩個相似度最高的群聚進行合併，直到所設定的終止群聚數目為止。而分裂法與凝聚法相反，初始時將所有資料點視為同一個群聚，開始依資料點間相異的程度往下做分裂的動作，直到群聚數目為設定終止條件為止。

常見的階層式演算法有CHAMELEON[14]、CURE[3]及ROCK[3]；此類型演算法中有以凝聚為基礎的CURE演算法，它固定選擇 C 個資料點作為代表點，如此可取得單一代表點與全部代表點間群聚品質與效能間的平衡位置。而採用資料互相鏈結(Interconnectivity)關係的ROCK可以解決以相似函數為基礎中，資料間相似卻分配在不同群聚中的矛盾現象；另外CHAMELEON針對CURE與ROCK中僅採用群聚內相似程度或群聚間相異程度的現象提出以相對互相鏈結(Relative interconnectivity)與相對相似(Relative closeness)為基礎的階層式群聚演算法，並大幅度的提升群聚外型的辨識能力。

一般而言，大部分階層式群聚演算法皆為凝聚法[29]，但是不論凝聚法或分裂法的階層式群聚演算法都花費大量的計算成本，因為需要計算每一個代表點與其他代表點間成對的相似程度，分別常用以下四種的方法評估兩群聚間的相似程度，詳述如下。

- (1) Centroid link：以兩群聚間的重心點最為代表點，再以兩代表點間的距離來表示彼此間的相似程度。
- (2) Complete link：以兩群聚間最遠的兩個資料點最為代表點，再以兩代表點間的距離來表示彼此間的相似程度。
- (3) Single link：以兩群聚間最近的兩個資料點最為代表點，再以兩代表點間的距離來表示彼此間的相似程度。
- (4) Average link：以兩群聚間所有資料點彼此的平均距離表示彼此間的相似程度。

當資料集合中群聚分佈明顯時，個別使用以上四種評估式的群聚結果是相似地；反之當群聚分佈受到雜訊影響時，利用上述評估方法的群聚結果會造成極大的差異，如圖2-1中所示，群聚間有細長的雜訊相互連結，此現象稱為鏈結效應（Chaining effect）。當使用Single link評估此資料分佈時，會造成錯誤的群聚結果，如圖2-1 (a)；但若使用Complete link來評估時便能獲得正確的群聚結果圖2-2 (b)，由此可知單一群聚相似度評估式並非適用所有資料分佈中。

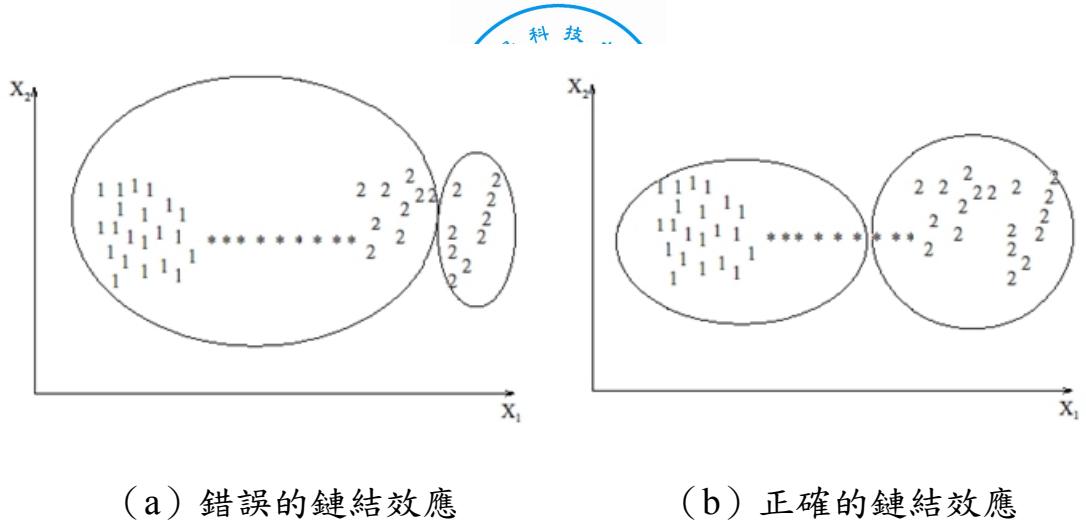


圖2-1 鏈結效應 [26]

(三) 密度式群聚類型 (Density based)

密度式群聚演算法利用相鄰區域的觀念來發現群聚之方法，當相鄰區域內資料點的密度到達預先設定的門檻值時，便自動形成一個群聚（如圖2-2所示）。此外，密度基礎的群聚演算法較切割式群聚演算法具有去除或控制偏差值與雜訊的能力，而且在參數正確的前提下，能分辨出任意外型的群聚分佈[2][18]。

目前較常見的密度式群聚演算法有DBSCAN[8]，OPTICS[1]；其中DBSCAN為第一個密度為基礎的群聚演算法，它是利用當群聚發生時，其群聚密度必定優於整體密度的觀念建構而成的；另外OPTICS是將DBSCAN的觀念延伸，利用密度為基礎的分群順序進而有效的發現群聚，其二者皆具有相似的密度概念。

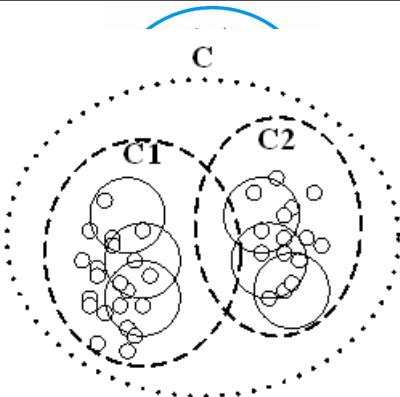


圖2-2 密度式群聚演算法示意圖 [8]

(四) 格子式群聚類型 (Grid based)

格子型式的概念是將包含資料點的資料空間，量化為格子狀 (Cell) 的概念，藉由計算格子內的統計資訊，進行群聚的方法，如 圖2-3所示。每一個格子內皆包含其資料的統計資訊（如平均值、標準差、最大值與最小值），因為在在演算法演譯的過程無需面對所有 資料點，所以能夠大量的減少演算的時間，提高其此類型群聚演算法 效能，尤其擅長於大量且高維度資料庫。

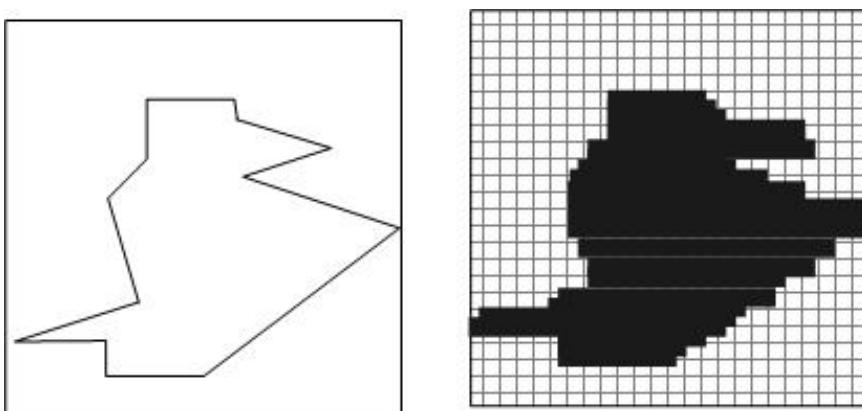


圖2-3 格子式演算法示意圖[27]

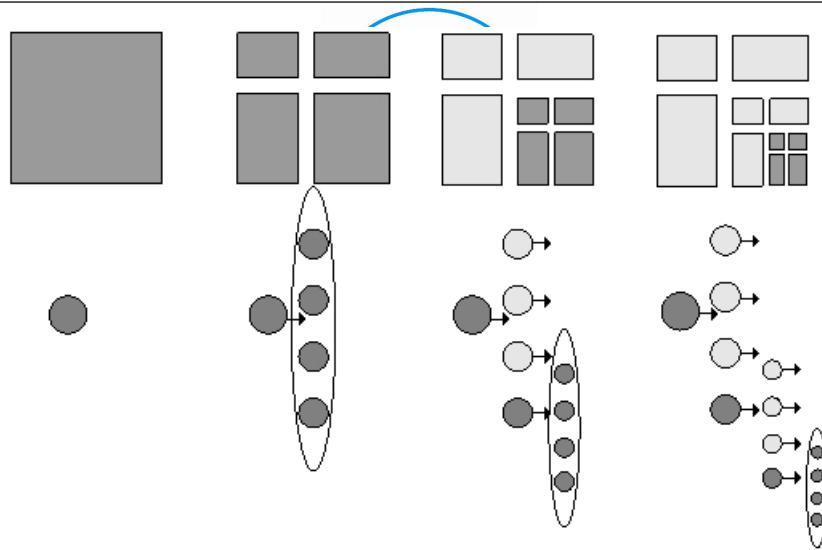


圖2-4 STING示意圖 [27]

目前常見的格子式群聚演算法有STING[27]、WaveCluster[24]；其中STING是由上而下的將資料空間切割成格子狀，再以樹狀結構呈現出來，接著利用廣度搜尋將格子內的群聚作合併，而將群聚結果呈現出來，如圖2-4所示，另外WaveCluster則是將密度與小波轉換的概念進行群聚程序。

(五) 混合型群聚類型

在混合型群聚演算法中，包括類神經網路（Artificial neural network）、演化式計算(Evolutionary computation)及進階搜尋(Advance search)等方法；由於種類繁多，本研究僅針對類神經網路中的自我組織映射網路（Self-organizing map）介紹。

SOM是一種非監督式學習的網路模式，由Kohonen提出後，被廣泛的運用在資料挖掘與時間序列分析等領域上[13]。其運作原理如

如圖2-5所示；當訓練資料被輸入至此網路時，每筆資料間均透過網路連結，並將彼此間的歐式距離結果傳遞到輸出層。而在輸出層內的每一個單元皆比較兩兩間的歐式距離（Euclidean distance），直到發現最大值時，稱此值為優勝單元（Winner）。接著調整輸入層與輸出層之間的連接加權值，直到優勝單元的影響力可以傳遞到每一個輸入屬性。當鄰近距離越大者，其鄰近係數（Neighborhood parameter）則越小，其修正的連接加權值也越小。

初始時，鄰近區域（Neighborhood）越大其修正的幅度也就越大，當訓練過程結束時，鄰近區域將會越縮越小，直到鄰近區域為零或者到達設定收斂條件時，整個分群過程就結束了。

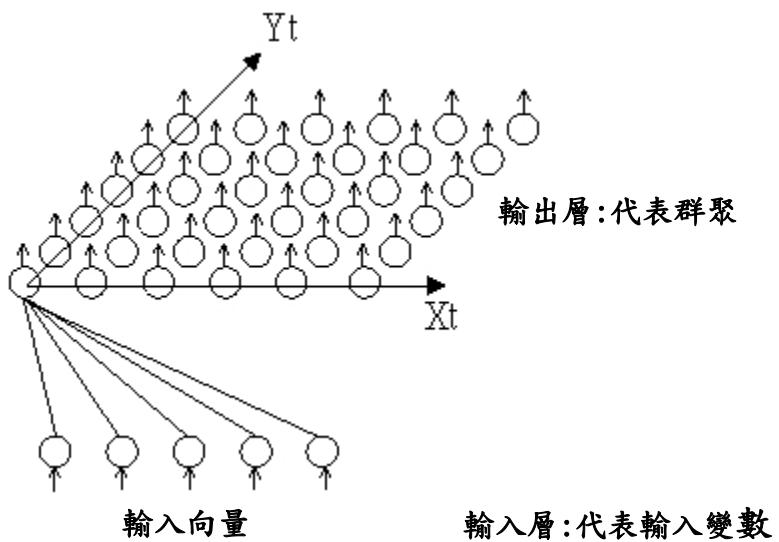


圖 2-5 SOM 網路圖

(六) 群聚演算法類型與群聚品質之比較

以上介紹的各種群聚演算法皆有其先天的特性與限制，但本研究將焦點放在各群聚演算擅長處理的群聚外型與處理效能上，接下來針對各種群聚演算法類型的群聚品質與效能一一介紹如下[2][18]。

- (1) 切割式群聚演算法：善於處理群聚大小相似而且低雜訊的凸狀圖形，但是拙於處理非凸圖形且高維度的資料分佈。
- (2) 階層式群聚演算法：對於簡單且大小相異的凸狀圖形有著不錯的效能，但不適合應用在高維度與大型資料分佈。
- (3) 密度式群聚演算法：可以發現任意外型的群聚分佈包括非凸圖形、巢狀圖形及雜訊的辨別，但是當群聚內部不均勻密度分佈及大型資料庫時，此類群聚演算法效能及品質低落。
- (4) 格子式群聚演算法：適用大型高維度資料庫的使用，但是其產生群聚的依據在於”發現”群聚分佈，但是並不保證其群聚外型的正確性，如此將降低其群聚外型的鑑別能力。
- (5) 混合型群聚演算法中的 SOM 網路：群聚品質與切割式群聚演算法相近，但是當處理大型高維度資料庫時 SOM 却可以在較短的時間內滿足收斂條件；換句話說便是 SOM 與切割式群聚演算法雖然群聚品質相似，但是 SOM 尤其適合大型高維度資料分佈。

以上分析整理如表 2-1，其中包含各類型群聚演算法的特性，分別為時間複雜度、處理屬性、群聚外型、處理雜訊能力與輸入資料順序等特性。

由此表 2-1 中可得知密度類群聚演算法相較於其他類的演算法對於各種群聚外型分佈的發覺表現最為優異，其原因在於其他演算法類型均以歐式距離為基礎，如此將導致其群聚演算法僅能發現以圓型為主的群聚分佈，但是在實際資料中除了凸狀（Convex）分佈外，尚有非凸狀（Non-convex）分佈等；但唯有密度式群聚演算法能發現非凸外型的群聚分佈，原因於衡量群聚發生之準則並非單純的以歐式距離公式為主，尚加入了密度或其他概念，才能克服發現非凸外型群聚分佈。

表 2-1 群聚演算法與參數需求比較表

類型	演算法	時間複雜度	群聚外形				處理 雜訊 能力	資料 輸入 順序
			處理 屬性	圓形	非凸 圖形	巢狀 圖形		
切割式	K-means	$O(n*k*t)$ <small>n:資料筆數 k:群聚個數 t:迴圈數</small>	數值型	可	否	否	否	無影響
	CLARA	$O(k(40+k)^2 + k(n-k))$	數值型	可	否	否	否	差
階層式	CURE	低維度: $O(n^2)$ 高維度: $O(n^2\log n)$	數值型	可	可	否	收縮程度 (α)	尚可
	BIRCH	$O(n)$	數值型	可	否	否	群聚半徑 (EPS) 分支因素 ($branching factor$)	尚可
密度式	DBSCAN	$O(n\log n)$	數值型	可	可	可	EPS 最小相鄰數量 ($Min-Pts$)	佳
	SOM	$O(n)$ <small>n:特徵向量數</small>	數值型	可	否	否	k	無影響

二、群聚適切性評估



群聚適切性 (Cluster validity) 是群聚流程中最重要的議題之一，原因在於不同輸入參數組合之下特定群聚演算法，如何能夠擁有最合宜群聚結果，而群聚適切性評估式能幫助使用者選擇出合宜的群聚結果。接下來本研究將敘述群聚適切性的基本概念與各種群聚適切性所採用的評估方法。

(一) 群聚演算法參數問題描述

如同其他演算法，群聚演算法之輸入參數直接的影響其輸出群聚結果，所以唯有輸入合適的參數組合才能獲得正確的群聚結果，但有些群聚演算法受限於先天的限制，即使輸入合適的參數組合，仍有可能輸出不合適的群聚結果。

以圖 2-6 為例；原始資料分佈為圖 2-6 (a)，其中我們可以直觀的發現最佳群聚數量應為三群，但是若群聚演算法依群聚幾何關係或密度仍可能將群聚結果區分為圖 2-6 (b)，其群聚結果並非最合適的分群分佈；或是使用者輸入不合適的參數，仍可能將群聚結果區分為圖 2-6 (c) 形成四個群聚，依然並非最佳的群聚結果。

有鑑於此；便有學者提出能夠幫助使用者在不同參數間，選擇出最合宜的群聚結果的評估準則，即為群聚適切性評估式。

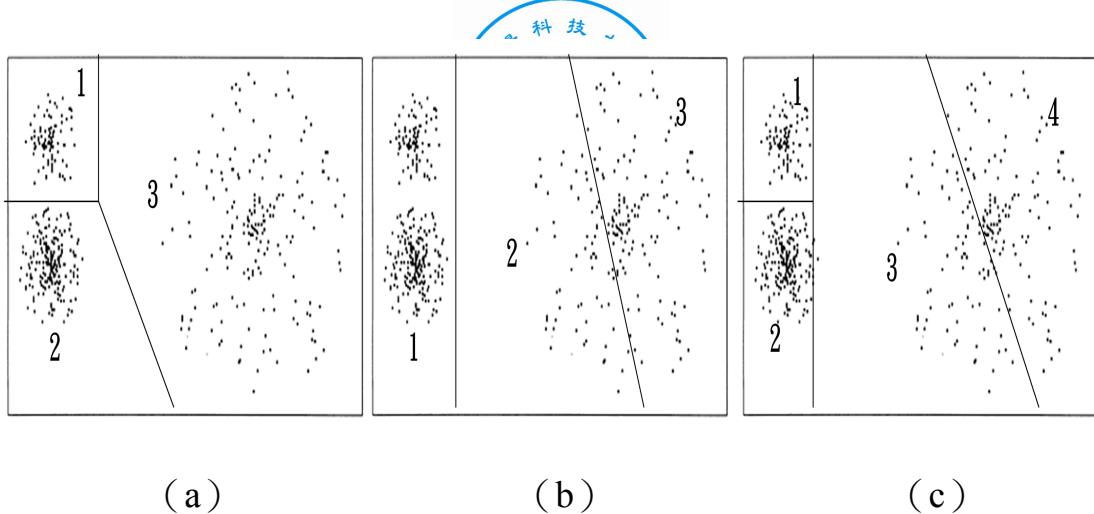


圖 2-6 不同參數的群聚結果

(二) 群聚適切性量測準則

一般直觀認定中，一個優良群聚結果必備的條件為在相同群聚內的資料點其相似度要高，而分屬於不同群聚的資料點必須能夠明確的區分開(Well separation and compact)；便有學者[21]提出兩個評定群聚結果及選擇最佳化群聚參數的準則，分別為群聚內的緊密度與群聚間的分離度，分述如下。

(1) 緊密度 (Compactness)：同一個群聚內的資料點應該盡可能地靠近(相似)，換句話說即是同一個群聚內的資料點其變異應該降至最小。

(2) 分離度 (Separation)：分屬於不同群聚的資料點應儘可能地遠離(相異)，使其相似度降至最低；有四種常用方法用來衡量群聚間的分離度，便是第一節中介紹的Centroid link、Average link、Complete link與Single link。

(三) 群聚適切性評估式



目前常被使用到的群聚適切性評估式包含 CH 評估式、 Γ 統計量、

Dunn's 評估式、DB 評估式與混合型評估式，並詳述如下。

(1) CH (Calinski Harabasz) 評估式

CH 評估式[4]如公式 2-2：

$$CH \text{ index} = \frac{\text{trace } B / k - 1}{\text{trace } W / n - k} \quad (2-2)$$

$$\text{trace } B = \sum_{K=1}^K n_k \times \|z_k - z\|^2 \quad (2-3)$$

$$\text{trace } W = \sum_{K=1}^K \sum_{i=1}^{n_k} \|x_i - z_k\|^2 \quad (2-4)$$

其中 n_k 為群聚 k 中資料點數； z 為所有資料點的中心點； z_k 為群聚 k 的中心點。

$\text{trace } B$ (公式 2-3) 表示資料點所屬群聚之群聚中心點至全體資料中心點之歐式距離，即表示所有群聚間的分離程度。 $\text{trace } W$ (公式 2-4) 表示資料點至所屬群聚之群聚中心點的歐式距離，即群聚內部的緊密程度；所以為了符合上節中的群聚適切性量策準則，當群聚結果越合適時，則 CH 評估值越小。

(2) Hubert's Γ 統計量

Γ 統計量[19]的公式如公式 2-5：

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N P(i, j) \times Q(i, j) \quad (2-5)$$

$$-1 \leq \Gamma \leq 1$$

其中 $M=N(N-1)/2$ ；同時令矩陣 $P=[p(i,j)]$ 與矩陣 $Q=[q(i,j)]$ 各為 $N \times N$ 大小矩陣，而 $p(i,j)$ 則以歐式距離表示元素 i 與 j 之間的相似度；另外 $q(i,j)$ 則是當元素 i 與 j 被分配至同一個群聚時其值為 1，當元素 i 與元素 j 被分配至不同群聚時其值為 0。

由以上的描述中可以發現當資料相似度高而且被分配至相同群聚時其 Γ 統計量會上升，所以可知為了符合群聚適切性量策準則當群聚結果越合適時， Γ 統計量值越高；但由於 Γ 統計量值落於 -1 至 1 之間，所以繼續發展出正規化 Γ 統計量[19]，其式如公式 2-6。

$$\bar{\Gamma} = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N (P(i,j) - \mu_p) \times (Q(i,j) - \mu_Q) / \delta_p \times \delta_Q \quad (2-6)$$

$$0 \leq \bar{\Gamma} \leq 1$$

μ_P 與 μ_Q 分別為矩陣 P 與 Q 中所有元素的平均數， δ_p 與 δ_Q 分別為矩陣 P 與 Q 中所有元素的標準差；藉由此正規化方法可將 $\bar{\Gamma}$ 統計量值控制在 0 至 1 之間，當群聚結果越合適時，其 $\bar{\Gamma}$ 統計量值將越接近 1。

(3) Dunn's 評估式

Dunn's 評估式[7]其目標在於鑑別出群聚內部緊密與群聚間明確分離的群聚分佈，其公式定義如公式 2-7：

$$v_D = \min_{i=1,\dots,n} (\min_{j=i+1,\dots,n} (\frac{\delta(C_i, C_j)}{\max_{k=1,\dots,n}(\Delta(C_k))})) \quad (2-7)$$

$$\Delta(C_k) = \max_{x_i, y_i \in C_k} \|x_i - y_i\|^2$$

$$\delta(C_i, C_j) = \min_{x_i \in C_i, y_j \in C_j} \|x_i - y_j\|^2$$

其中 n 為資料點數， k 為群聚數量， C_i 與 C_j 表示第 i 個群聚與第 j 個群聚， x_i 與 y_j 表示隸屬於 C_i 的資料點和隸屬於 C_j 的資料點；由其公式 2-7 可知，當群聚結果越合適時，Dunn's 評估值會隨之上升。由於其 Dunn's 評估式在計算群聚的緊密度與分離度時，僅取最大值與最小值，這將降低其計算複雜度，使之較適合使用在大量資料的大型資料庫中，同時也將降低對偏差值得敏感度。

(4) DB (Davies-Bouldin) 評估式

DB 評估式[5]的概念與 Dunn's 評估式的概念相似，兩者皆利用群聚間離散程度與群聚內聚程度作為衡量依據，其式如公式 2-8：

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j, i \neq j} \left(\frac{s_{i,j} + s_{j,q}}{d_{ij,t}} \right) \quad (2-8)$$

$$s_{i,j} = \left(\frac{1}{|C_i|} \right) \sum_{x \in C_i} \|x - z_i\|^2 \quad (2-9)$$

$$d_{ij,t} = \|z_i - z_j\|^2 \quad (2-10)$$

其中 k 為群聚個數； S_{ij} 為隸屬於群聚 i 的資料點至到群聚 i 重心的平均歐式距離，如公式 2-9； $d_{ij,t}$ 為群聚 i 與群聚 j 之重心間的歐式距離，如公式 2-10。由式 2-8 可知，當群聚結果越合適時，其 DB 評估值將越小。

(5) 混合型評估式

以上介紹的群聚適切性評估式中，皆使用資料點間的幾何關係為

基礎，但是在大部分的情況下，僅使用幾何關係作為基礎的評估式其群聚結果往往表現不佳，所以便有學者提出不同類型的評估式，藉由非幾何關係量測方法的選擇，組合出新的群聚適切性評估式；例如[3]中以 Dunn's 評估式為基礎，使用 Centroid link、Average link、Complete link 與 Single link 來衡量群聚間的分離程度及使用最長距離、最短距離、重心距離與平均距離衡量群聚內部緊密程度，其概念在於藉由選擇不同的量測方法，進而建構出群聚結果適切性較佳的群聚適切性評估式。

尚有學者[23]以 Dunn's 評估式與 DB 評估式為主配合上圖形理論中的最小擴張樹、RNG (Relative neighborhood graph) 與 GG (Gabriel graph) 來量測群聚內部緊密程度，也是藉由不同量測指標與既有的群聚適切性評估式重新組合，試圖提出群聚結果適切性較佳的群聚適切性評估式。

三、模糊控制理論



模糊控制 (Fuzzy control) [16] 是模糊理論中，一項極為重要的應用，由 Mamdani 根據 Zadeh 提出的語意分析法 (Linguistic approach) [28] 及模糊推論 (Fuzzy inference)，使用 IF~THEN 規則形式敘述操作員的操控經驗，並且與感測元件輸入資訊做近似對照，成功的應用在蒸汽機自動運轉控制上。

(一) 模糊控制器基本架構

模糊控制器其基本架構共包括四個部分；模糊化單元 (Fuzzifier)、模糊規則庫 (Fuzzy rule base)、模糊推論引擎 (Fuzzy inference engine) 與解模糊化單元 (Defuzzifier)。當系統藉由感測器將外界輸入資料的明確值藉由模糊化單元轉化為適當的模糊資訊，模糊推論引擎則是整個模糊控制器的核心，它根據所得到的模糊資訊以及模糊規則庫中專家所建立的模糊規則，模擬人類思考決策的方式，解決問題，最後解模糊化單元則將模糊推論引擎所推論出的模糊資訊，轉化為外界所能接受的明確資訊。

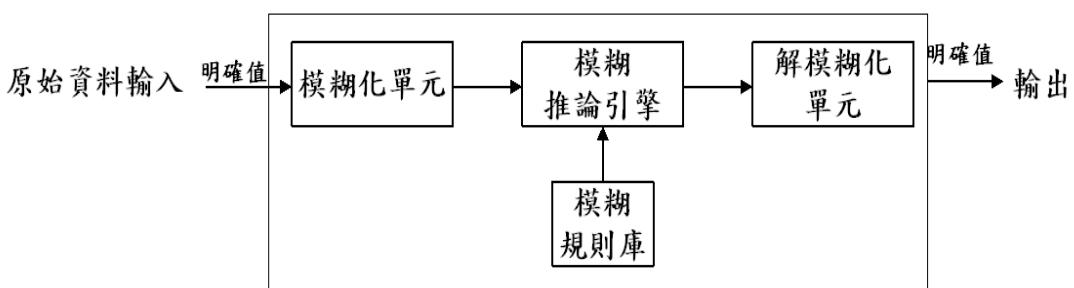


圖 2-7 模糊控制器基本架構



(1) 模糊化單元

模糊化單元將外界輸入的明確資料轉換成適當的語意式模糊資訊，透過模糊化的步驟，模糊系統能將輸入系統中的實際變數從技術層次轉換成語言層次的語言變數 (Linguistic variables)。

(2) 模糊規則庫

模糊規則庫由一組以IF~THEN形式的模糊規則所建立，每一條規則包括前提項 (Antecedent) 及結論項 (Consequent)；當前提項成立時，可推得結論項成立，藉此描述系統的輸入與輸出關係。

建立規則庫有兩種方法，其一是經由專家根據其專業知識所建立，但專家無法完整提供所有的規則，所以無法處理所有可能出現的情況。其二為經由學習法則從過往資料中學習，獲得模糊規則。但在資料變數過多時，將會產生大量的規則，而造成無效率的推論結論。

(3) 模糊推論引擎

模糊推論引擎是模糊控制器的核心，藉由近似推論或模糊推論的過程，以模擬專家的決策模式。由規則庫產生的推論過程，可分為兩個部分，第一部份是規則條件 (IF) 的部份，定義目前狀況下，其規則是否有效；第二部份是規則動作 (THEN) 的部份，定義所採取的結論動作。由於推論是一項主觀的判斷，所以在選擇方法上，也會因為應用領域的差異而有所不同，目前常用的推論方法有Max-Min推論

法、Max-Prod推論法、Mamdani推論法等。

(4) 解模糊化單元



解模糊化單元是將模糊推論後產生的結論，轉化為明確數值的過程。根據各特徵值所對應變數的語言與模糊規則庫內的規則作比對，找出被觸發的模糊規則，進而得到模糊推論的結果，此結果再透過解模糊化的動作，轉換成各類別的隸屬程度，再藉由模糊規則，找出最可能的結果作為輸出。常見的解模糊化方法有重心法（Center of Gravity Method）、最大值平均法（Mean of Maximum）、面積中心法（Center of Area Method）等。

第參章 研究方法

在此章中我們針對群聚處理流程中的選擇群聚演算法與群聚結果適切性此二步驟中的不足，本研究所提出以資料導向為基礎之選擇群聚演算法類型方法與一個較具適應性的群聚適切性評估式，詳述如下。

一、資料導向基礎之選擇群聚演算法方法

此節將敘述本研究如何分析原始資料分佈，並由中獲取選擇合適群聚演算法的資訊，藉此資訊幫助使用者選擇合適的群聚演算法類型。

由第貳章中表 2-1，可知每種群聚演算法類型都有其優缺點及其擅長處理的資料型態，但是一般使用者在不熟悉群聚演算法適用的環境及限制下，極易做出錯誤的決定，而選擇不合適群聚演算法類型進行群聚分析，如此將使群聚結果不可信賴甚至錯誤。為此本研究提出三個有效的資料指標 能夠輔助使用者選擇群聚演算法，其指標分別為資料密度、資料分散度與資料混亂度，藉由以上三種指標輔助使用者選擇出適合此資料分佈的群聚演算法類型，以下針對各群聚演算法適合之資料分佈與資料密度、資料分散度、資料混亂度三指標作詳細說明如下。

(一) 各群聚演算法類型合適之資料分佈

由表 2-1 與過去研究[2][18]調查中發現，相同種類的群聚演算法雖然有著不同的演譯方法，但是由於概念與原理是相似地，如此將使得屬於相同種類的群聚演算法具有相似特性，包含群聚外型與演算效能。本研究將針對切割式、階層式、密度式、格子式與混合式中的 SOM 群聚技術其擅長處理的群聚外型，以具體的方式加以敘述。

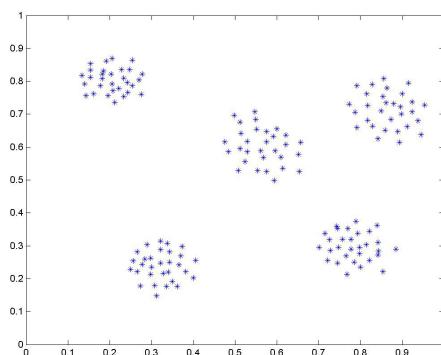


圖 3-1 合適於切割式群聚演算法的資料分佈

本研究由[2][18]中得知切割式群聚演算法善於處理，群聚大小相似而且低雜訊的凸狀圖形，其原因在於此類型群聚演算法的目標式大部份皆以歐式距離公式為主，這將使得此類型群聚演算法善於分辨出群聚大小相似的凸狀圖形的主因，但也因此必須計算兩兩資料點間歐式距離造成的高時間複雜度，使得此類型群聚演算法不善於處理大型資料庫；本研究將以上的資料特性轉化為圖 3-1，並將其此分佈視為合適於切割式群聚演算法的資料分佈之一。

科 業

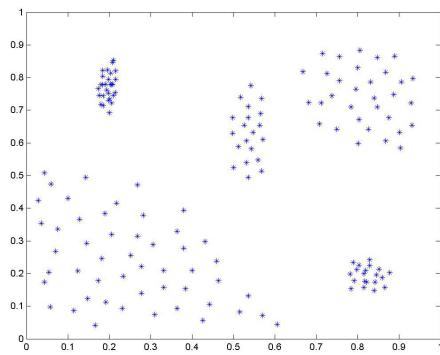


圖 3-2 合適於階層式群聚演算法的資料分佈

接著是階層式群聚演算法，此類型的群聚演算法對於簡單且大小相異的凸狀圖形有著不錯的效能[2][18]，但與切割式群聚演算法相同的是階層式群聚演算也大多使用歐式距離做為其量測其相似度的依據，如此也使得其時間複雜度升高；另外與切割式群聚演算法不同的是階層式群聚演算法具有由下至上的凝聚概念或由上至下的分裂概念或兩者兼具，如此將造成此類群聚演算法較切割式群聚演算法更擅長處理大小相異的凸狀群聚外型；本研究將此敘述轉化為圖 3-2 並將其視為合適於階層式群聚演算法的資料分佈之一。

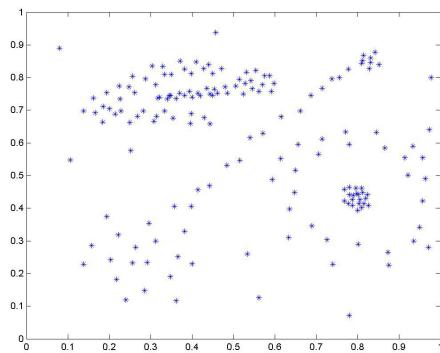


圖 3-3 合適於密度式群聚演算法的資料分佈



另外，密度式群聚演算法善於發現任意外型的群聚分佈，包括非凸圖形、巢狀圖形及雜訊的辨別[2][18]。此類群聚演算法類型發展的動機在於切割式與階層式群聚演算法，兩者皆受限於歐式距離公式，僅能發現凸狀的群聚分佈，但這無法克服所有的資料分佈外型。為此密度式群聚演算引入密度的概念，在參數正確的前提下，能夠發現任意群聚外型分佈包括非凸圖形甚至於雜訊的辨別與處理，但是此類群聚演算類型計算複雜度尚不足以應付大型資料庫；而本研究將此敘述轉化為圖 3-3，並將其視為合適於密度式群聚演算法的資料分佈之一。

尚有針對高維度與高資料筆數的大型資料庫所設計的格子式群聚演算法，此類群聚演算類型產生群聚的依據在於”發現”群聚，但是並不保證其群聚外型的正確性[2][18]；善於處理大型資料庫原因在於格子式群聚演算法之群聚概念是由上而下的利用廣度搜尋將格子內的群聚作合併，並統計紀錄於格子中的資訊，作為群聚合併的依據；此類群聚演算類型的缺點在於分辨群聚外型的邊緣不是水平就是垂直，儘管大幅度的降低其時間複雜度，但卻使其群聚結果的品質及正確性下降；本研究將格子式群聚演算法視為專門處理高資料筆數與高資料維度的大型資料庫的群聚演算法類型。

最後是混合類群聚演算法中的 SOM 網路模型，此群聚演算法的

其群聚結果與切割式群聚演算法相近，但是 SOM 却更適合應用在大型高維度資料庫[2][18]，其原因在於 SOM 處理高資料筆數的資料庫時，其收斂速度較切割式群聚演算法快速；所以本研究將 SOM 演算法視為能夠適用在高資料筆數資料庫中的切割式群聚演算法。

(二) 資料密度指標

本研究由過去密度式群聚演算法相關研究中，發現當資料中有群聚發生時，其群聚本身的密度必定高過於整體資料的密度[29]；有鑑於此，本研究將資料密度視為能夠幫助使用者選擇群聚演算法類型的重要指標之一，其理由一為當資料密度上升時，其群聚也較容易被發現，其因為整體資料密度上升時，表示其單位面積中所包含的資料點數也增加，同時導致群聚內部的緊密程度提升，如此群聚會較容易被群聚演算法發現。其二為群聚演算法的時間複雜度，其因可由表2-1中得知大部分群聚演算法的時間複雜度皆與資料點數有關，這表示當資料密度上升時，其演譯時所花費的時間成本也隨之增加。

接下來本研究將資料密度指標定義為單位面積下 Isothetic rectangle 空間所包含的資料點數，其式如下：

$$\text{Density of data} = \frac{\| \text{data} \|}{\text{Isothetic rectangle of data}} \quad (3-1)$$

其中 Isothetic rectangle of $data$ 為所有資料的 Isothetic rectangle 空間大小； $\|data\|$ 為總資料點數

Isothetic rectangle 空間指的是一個將所有資料完全含括的空間，以圖 3-4 為例，欲求圖 3-4 資料之 Isothetic rectangle 空間時，先將其資料投影至 x 軸，求其極大值及極小值，分別為 0.3 與 0.8；再將投影至 y 軸，求其極大值及極小值，分別為 0.12 與 0.7，而其 Isothetic rectangle 空間大小為 $(0.8 - 0.3) \times (0.7 - 0.12)$ 其值為 0.29。

本研究所定義的資料密度指標（公式 3-1），可以成功的發現整體資料在單位面積中的密度大小。

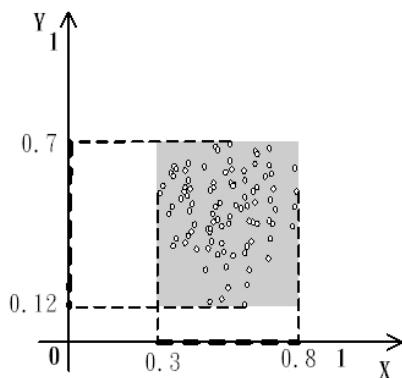


圖 3-4 Isothetic rectangle 空間示意圖

（三）資料混亂度指標

本研究由近期具有高品質的群聚演算法中 [9][12][20] 發現一共同的概念為皆以熵函數（Entropy）[9] 做為群聚演算法的目標函數，藉由演譯其群聚演算法以獲得到最佳的群聚結果；而熵函數最初被使用在熱力學中，其目的在於評估整個熱力環境中的混亂活動程度。

本研究繼承資訊理論 (Information theory) 中資訊含量的概念，
應用熵函數評估資料的混亂度，其式如下：

$$E(X) = \sum_{x=1..n} p(x) \times \log \frac{1}{p(x)} \quad (3-2)$$

其中 $p(x)$ 為 x 事件出現的機率；從公式 (3-2) 可知，假設 x 為資料中的某一事件，而 $p(x)$ 為 x 事件發生的機率，則 $E(x)$ 可視為 x 在資料庫中的混亂程度。

在熵函數為基礎的群聚演算法中皆以熵函數做為群聚演算法的目標函數，藉由演譯其群聚演算法以獲得到最佳的群聚結果；換句話說，在群聚演譯過程中，各階段將獲得到不同的群聚組合，但是唯有最佳的群聚結果可以將整體的混亂程度降至最低，也就是符合收斂條件的分群結果。在熵函數為基礎的群聚演算法中，[9]利用熵函數為目標函數，獲得到品質較佳的分群結果，甚至分辨出非凸群聚外型及非監督式學習的目的，另外在[12]中則是使用連續型熵函數（Renyi's entropy）也獲的不錯的實驗結果；尚有[20]將 Renyi's entropy 使用至模糊群聚演算法中，也提升模糊群聚的群聚品質。

本研究由[9][12][20]討論中發現當熵函數中 x 發生的機率越低與整體資料的維度增加時，將使得熵函數值上升；另外[9][12][20]發現當熵函數上升時，除了表示分群結果尚有改進空間外，也表示演譯的

複雜度也將隨之增加；有鑑於此本研究也採用熵函數作為資料混亂度指標以評估原始資料混亂程度，並將此混亂程度視為幫助使用者選擇群聚演算法類型的重要指標之一。

(四) 資料分散度指標

資料分散度指標的目標是為了量測原始資料分離的程度；而資料分散度與群聚的關係在於當資料分散程度高時表示其資料彼此間相似的程度低，則適合將其區分為較多的群聚；資料分散程度低時，表示其資料彼此間相似的程度較高，則適合將其分群數量降低。

一般常見的分散程度統計量測指標如全距、四分位數、標準差與變異數等，僅能衡量資料的絕對分散程度，因為都受到平均數大小及不同衡量單位的影響，而不能進行資料間相對分散程度的比較，所以本研究採用一種能客觀評估資料間相對分散程度的方法為統計量測指標中的變異係數(Coefficient of variation)進行資料分散程度的量測，變異係數定義如下：

$$CV \text{ of data} = \frac{\sigma \text{ of data}}{\mu \text{ of data}} \quad (3-3)$$

其中 $\sigma \text{ of data}$ 表示資料的平均數； $\mu \text{ of data}$ 為資料的標準差，利用資料間的平均數與變異數間比例來衡量資料相對的變異程度，所以本研究採用變異係數作為量測資料分離度的指標，並將此指標視為幫

助使用者選擇群聚演算法類型的重要指標之一。



（五）資料指標與群聚演算法類型

本研究中的所選擇群聚演算法的概念是以效能與品質為前提之下，依據第一節所整理的各群聚演算法類型所擅長的資料分佈型態再計算欲分群的資料的密度、分散度與混亂度三指標，便可鑑別出此資料分佈並作為選擇合適群聚演算法類型之依據。

以下便針對各群聚演算法類型，以資料密度、分散度與混亂度三指標與各群聚演算法類型加以分析。

(1) 切割式群聚演算法：以一合適此類型之群聚演算法資料分佈性進行分析，如圖 3-1 中其資料分佈型態為群聚大小相似而且低雜訊的凸狀圖形；而由本研究提出之三指標的角度進行分析時可知因為資料群聚大小相似而且低雜訊，所以資料混亂度不高，另外資料間分離明顯而且低雜訊所以其分散度為高的。

(2) 階層式群聚演算法：以一合適此類型之群聚演算法資料分佈性進行分析，如圖 3-2 中其資料分佈型態為簡單且大小相異的低雜訊凸狀圖形分佈；而由本研究提出之三指標的角度分析因為資料群聚大小相異而且低雜訊，所以資料混亂度是低的，另外群聚間大小相異所以資料分散度較切割式群聚演算法略低。



(3) 密度式群聚演算法：以一個合適此類型之群聚演算法資料分佈性進行分析，如圖 3-3 中其資料分佈型態為為複雜且任意外型群聚分佈包括大小密度相異的高雜訊資料分佈；而由本研究提出之三指標的角度分析，因為複雜且任意外型群聚分佈，所以資料混亂度高，另外群聚大小密度相異的高雜訊資料分佈所以其分散度較低。

(4) 格子式群聚演算法：合適此類型之群聚演算法資料分佈為高資料筆數與高資料維度的複雜型資料分佈，而由本研究提出之三指標的角度分析，因為高資料筆數與高資料維度所以資料混亂度較高，另外由於資料筆數較高，相對的造成資料密度也較高。

(5) 混合式群聚演算法中的 SOM 群聚演算法：由於其資料特性繼承切割式群聚演算法的資料分佈，但卻能處理高資料筆數的資料；而由本研究提出之三指標的角度分析，因為與切割式群聚演算法的特性相似，所以其分散度高；另外分佈高資料筆數所以資料混亂度與資料密度較高。

(六) 資料導向為基礎之群聚演算法選擇模型

根據前節中可知，由原始資料中密度、混亂度與分離度指標中，可以找出適合的群聚演算法類型；因此本研究將此概念與模糊控制結合建構一資料導向之群聚演算法選擇模型；如圖 3-5。

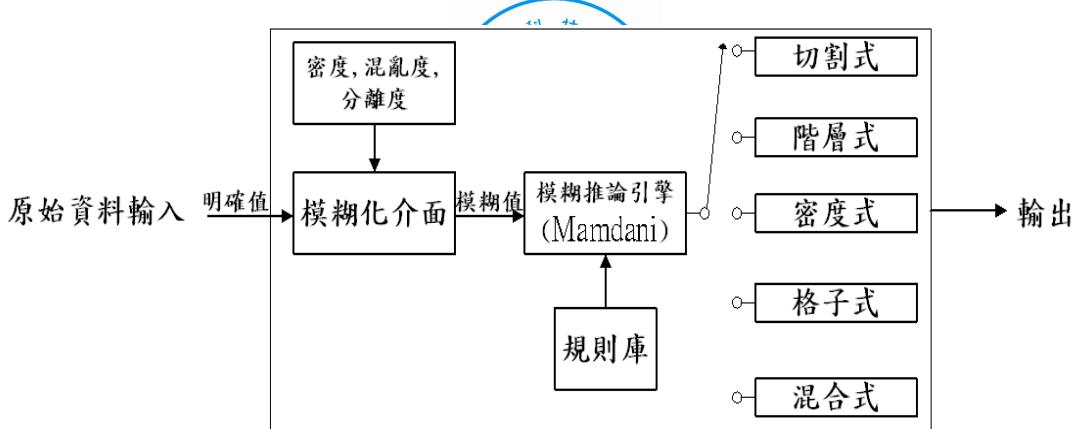


圖 3-5 資料導向之群聚演算法選擇模型

(1) 定義輸入與輸出變數

本研究定義欲分群的原始資料的密度 (Density)、混亂度 (Entropy) 與分離度 (CV) 三指標值作為資料導向之群聚演算法選擇模型的輸入變數，而輸出變數為各類型的群聚演算法，包括切割式、階層式、密度式、格子式與混合式群聚演算法。

(2) 定義模糊變數與隸屬函數

模糊化介面中為了將明確變數轉換為模糊變數，必須採用適當的隸屬函數 (Membership function) [28]。本研究中採用梯形隸屬函數並定義出兩個模糊變數，分別為高 (High) 與低 (Low)，對每一個輸入變數選擇適當的隸屬函數，如圖 3-6 所示。

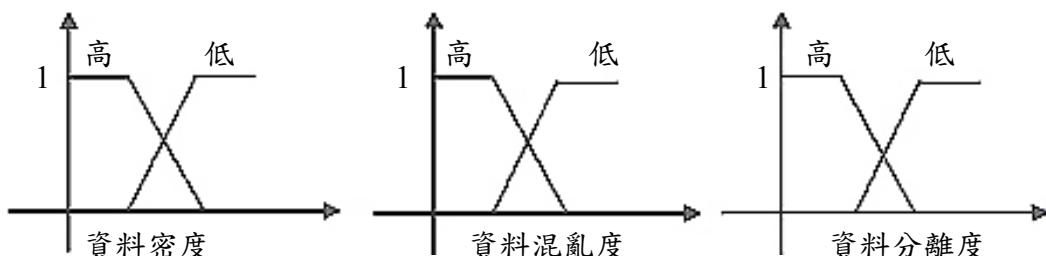


圖 3-6 原始資料各輸入變數之隸屬函數

(3) 設計模糊規則庫



建立模糊控制系統，規則庫為模糊控制之核心，具有模擬專家決策之能力。此步驟利用模糊推論發法去推論一組描述操控法則的語言控制規則，以求得控制所需的模糊量。本研究中所設計的模糊規則之輸入變數密度、混亂度與分離度各有兩個隸屬函數，其中我們採用的 Mamdani 方法的 IF-THEN 推論系統應用於模糊推論規則中，總計有七條控制規則，詳見表 3-1。

表 3-1 資料導向之選擇群聚演算法規則庫

資料 密度	資料 混亂度	資料 分散度	群聚演算法 類型
低	低	低	密度式
低	低	高	階層式
低	高	低	切割式
低	高	高	密度式
高	低	低	N/A
高	低	高	格子式
高	高	低	混合式
高	高	高	格子式

二、群聚適切性評估式



第貳章中介紹之群聚相關技術及群聚適切性評估式中，不論群聚演算法或適切性評估式皆具有共同目標，但是卻使用許多不同方法來完成此目標，因此藉由第貳章詳細討論每種方法的優缺點及適用的環境，本研究歸納出以下三點，敘述如下。

(1) 在比較群聚品質較佳與較差的群聚演算法後發現，其因在於品質較差的群聚演算法均僅使用資料點中的幾何關係或統計為基礎。而且在表 2-1 中，以幾何關係或統計為基礎的群聚演算法僅能發現以圓型為基礎的資料分佈，但是唯有以密度為基礎的群聚演算法能同時發現凸狀分佈與非凸狀分佈；原因於其衡量群聚發生的準則並非單一歐式距離公式或統計觀點為主，尚加入了密度或其他概念。由以上兩點可以推知，一個適應性較佳的群聚適切性評估式，除了考慮緊密度與分離度之外；尚可加入度量密度的量測準則。

(2) 在第貳章中介紹傳統量測群聚間分離度的方法分別為 Centroid link、Average link、Complete link 與 Single link，而此類方法均僅選擇單一代表點來代表整個群聚，並且僅單純的考慮到群聚與群聚間的距離；基於以上兩理由提出其不足之處，並詳述如下[29]。在此考慮三種狀況分別如圖 3-7、圖 3-8 與圖 3-9，

當圖 a 與圖 b 中兩群聚之半徑大小均有明顯差異時，不論使用 Centroid link、Complete link 或 Single link 量測其群聚間離散程度時其結果皆相同，但這卻是不客觀的。因為當群聚的半徑不同時，其半徑的長度相對地會影響到群聚間的分散的程度，

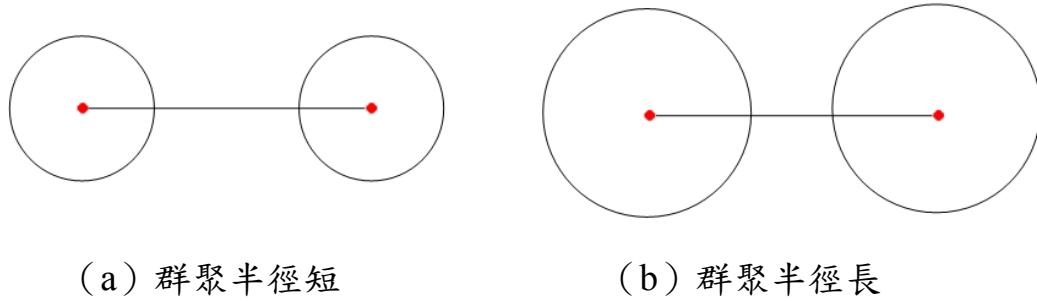


圖 3-7 兩群聚重心點距離相等，但群聚半徑不同

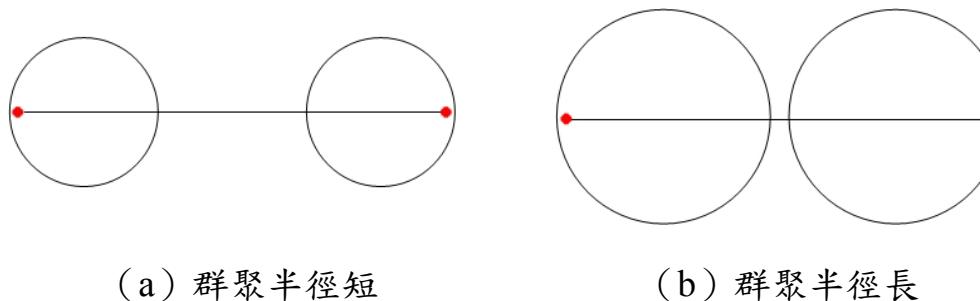


圖 3-8 兩群聚最遠點距離相等，但群聚半徑不同

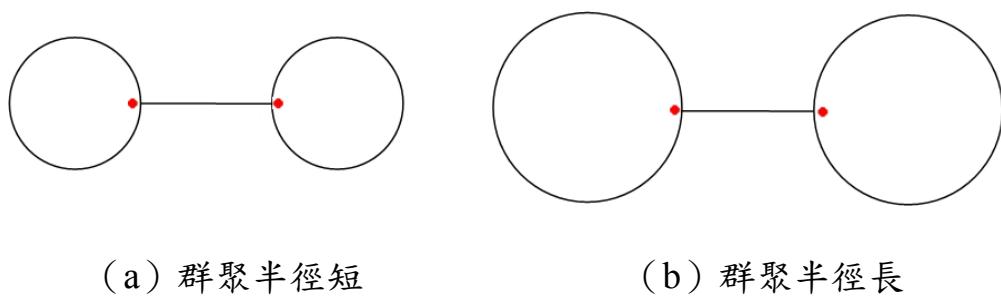
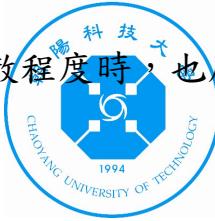


圖 3-9 兩群聚最近點距離相等，但群聚半徑不同

所以在衡量群聚間分散程度時，也應該包含群聚半徑所涵蓋的範圍。



(3) 以應用領域之觀點切入群聚技術時，如地理資訊、空間資料庫、影像資料庫與醫療資料庫等..皆經常應用到群聚技術，但是當地理資訊或醫療影像評估群聚結果時，僅利用群聚分散程度和緊密程度是不實際地，因為當分析此群聚結果時，尚有一個重要因素，即是群聚的凝聚程度。

以醫療影像與地理資訊應用為例，分析患者的斷層掃描影像時，若發現群聚的產生即有可能在此處發生病變，但是要判定其為良性或惡性腫瘤時，則需進一步進行切片。但是尚有另一個方法為分析群聚結果是否向內凝聚，即越接近群聚中心的密度越高時，則為惡性腫瘤的機率也較為升高。另外在地理資訊應用中分析山坡植被影像時，也可利用群聚的凝聚力找尋山坡植被產生病變或病蟲害的發源地，以利進行實地勘查。由以上二例中可以得知當引用群聚適切性評估式進行評估群聚結果時，必須將群聚凝聚力納入考量。

綜觀以上三點，本研究發現一個較佳的群聚評估式除了考量緊密度與離散度以外，尚須考慮到群聚凝聚力與密度率。

(一) 密度率 (Density rate)



本研究引用[29]中相對的密度來計算密度率的概念，本研究計算單一群聚的密度與整體資料空間的密度之關係，如此才能顯現出相較於整體密度而言，其某一群聚的資料密度程度為何。

單一群聚密度計算方法是先將資料集合內所有的資料點數除以資料點所分佈的空間，而在我們的研究中則是以先計算出所有資料空間，並將標準化成為 1，接著計算出經過標準化的各單一群聚資料空間，如圖 3-10 所示，假設整理空間經過標準化後的值為 1，則群聚 c1 的資料空間為 0.075，另外群聚 c2 的資料空間為 0.015。

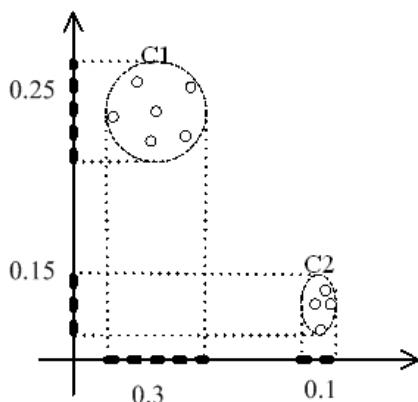


圖 3-10 群聚資料空間示意圖

由於本研究引用個別群聚密度與整體資料密度間的關係顯現出相較於整體資料而言，其某一群聚的資料密度程度為何，其定義如公式 3-1 所示。

$$\text{Density rate of } c_i = 1 - \left(\frac{\text{density of data}}{\|c_i\|} / \frac{\|c_i\|}{S_{ci}} \right) \quad (3-1)$$

$$0 \leq \text{Density rate of } c_i \leq 1$$

其中 $\|c_i\|$ 為第 i 個群聚中所包含的資料點個數； S_{ci} 為第 i 個群聚所佔的資料空間大小；當 $Density rate of c_i$ 越大則表示該群聚相對於整體群聚評估式流程中整體密度而言的密度是較高。

(二) 鑑別率 (Discrimination rate)

有鑑於上節中群聚間的分離度因群聚半徑不同而造成判斷的瑕疵，所以本研究不採用單一代表點方法計算群聚間的分離度，改為同時採用重心點與最近點之雙代表點方法，用以鑑別群聚間的分離度。

本研究引用[29]中評估分離度的方法，採用雙代表點的方法計算兩群聚分散程度，本研究定義鑑別率計算方式如公式 3-2 所示。

Discrimination rate between $c_i, c_j =$

$$\frac{\text{Min}\{\text{DIS}(c_i, c_j)\}}{u_i - u_j} \quad (3-2)$$

$$0 \leq \text{Discrimination rate between } c_i, c_j \leq 1$$

其中 n 為所有資料點數； k 為群聚數量； i 與 j 各為群聚 i 與群聚 j 中資料點； u_i, u_j 為群聚 i 的重心點與群聚 j 重心點；當 *Discrimination rate* 越接近 1 則代表鑑別效果越佳。以圖 3-11 為例，當以 single link 為衡量群聚間的分離程度時，3-11 (a) 與 3-11 (b) 兩者分離程度皆為 2，則將兩者的離散程度視為相同，但這樣是不合理的，因為圖 3-11 (a) 中群聚半徑較小，雖然兩者分散程度都為 2，但相對而言，圖

3-11 (a) 的分離程度應該是較圖 3-11 (b) 高；而採用本研究採用的鑑別率評估式，則可以分別得到 0.5 與 0.33，便可合理的表現出群聚分離程度。

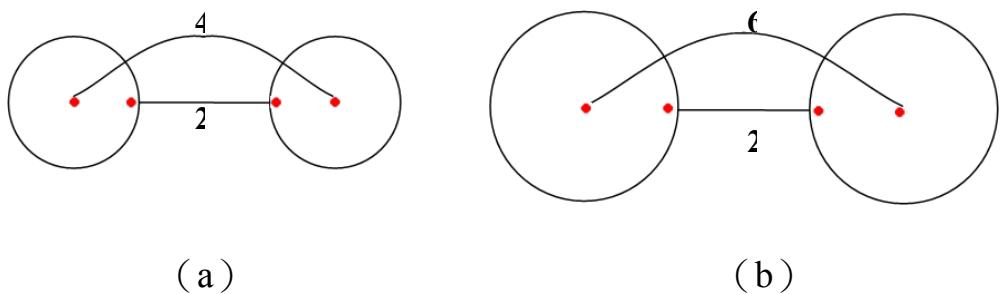


圖 3-11 兩群聚最近點距離相等，但重心點距離不同

(三) 凝聚率 (Agglomerate rate)

凝聚率之目的在於評估群聚內資料點向內集中的程度；以圖 3-12 為例，有兩個資料空間與密度皆相等的群聚，可以觀察到群聚 c2 較群聚 c1 凝聚，其原因在於當兩群聚在相同的條件下，群聚 c2 靠近中心點的資料點數較多，此時我們稱群聚 c2 的凝聚率高於群聚 c1。

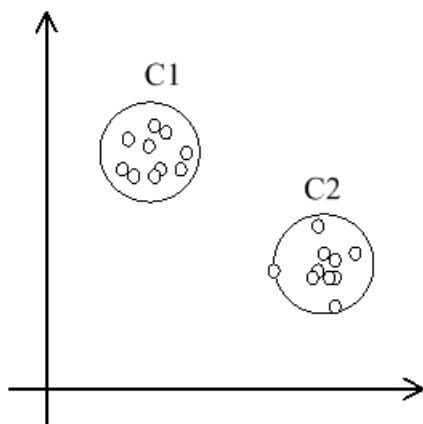


圖 3-12 群聚凝聚率示意圖

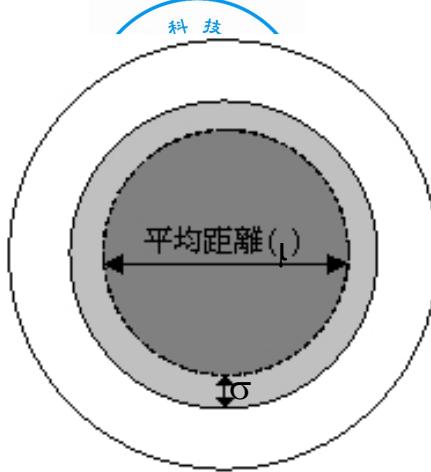


圖 3-13 群聚內資料點分佈圖

為了量測群聚內凝聚率的概念，本研究引用[29]中凝聚力與統計中標準差與平均數觀念以評估群聚中的凝聚程度；首先定義一群聚內資料點個數為 n ，接著計算資料點間彼此的平均距離為 μ 與標準差 σ ，最後便可計算出落於平均距離加上一倍標準差內之資料點個數，如圖 3-13 所示藉由此資料點數的多寡，便可量測出群聚的凝聚率。本研究將上述轉變為公式 3-3 詳述如下

$$\text{Agglomerate rate of } c_i = \frac{\|C_i\|_{\text{caliber}=\mu+\sigma}}{\|C_i\|} \quad (3-3)$$

$$0 \leq \text{Agglomerate rate of } c_i \leq 1$$

其中 $\|c_i\|$ 為第 i 個群聚內的資料點個數； $\|c_i\|_{\text{caliber}=\mu+\sigma}$ 為落於第 i 個群聚內資料點間的平均距離加上一倍標準差內之資料點個數。

由上述公式中，可知當越多資料點落於群聚內平均距離加一倍標準差的範圍內，表示其凝聚力越高。

(四) 適應型群聚適切性評估式

本研究以群聚內部緊密度與群聚間分離度為基礎，加入群聚密度率、鑑別率與凝聚率三種群聚特性指標，並提出一個適應性群聚適切性評估式，其式如下公式 3-4，詳述如下。

$$I_{adapt} = \frac{1}{C_2^k} \sum_{\substack{1 \leq i \leq k \\ 1 \leq j \leq k \\ i \neq j}} e^{\text{Disc}(c_i, c_j) + \{\text{Agg}(c_i) \times \text{Den}(c_i)\} \times \{\text{Agg}(c_j) \times \text{Den}(c_j)\}} \quad (3-4)$$

$$1 \leq I_{adapt} \leq 7.389$$

其中 k 為群聚數量， $\text{Disc}(c_i, c_j)$ 為群聚 i 與群聚 j 之間的鑑別率， $\text{Agg}(c_i)$ 為群聚 i 的凝聚率， $\text{Den}(c_i)$ 為群聚 i 的密度率；群聚結果越合適時，即群聚間分離度越高且群聚內緊密度越高的準則之下，其 I_{adapt} 值越大表示群聚結果越合適。

I_{adapt} 群聚適切性評估式係以群聚間分離程度與群聚內部緊密程度兩者間相互關係作為基礎點；其指數中的 $\text{Disc}(c_i, c_j)$ 是以群聚間的鑑別率來辨識出群聚間的分離程度，而 $\text{Agg}(c_i)$ 與 $\text{Den}(c_i)$ 則考慮一對群聚間，各別凝聚力與密度之特性的總合，以此來評斷群聚內部緊密的程度；再將分散程度與群聚緊密程度二者加總，最後將 C_2^k 組的群聚結果評估值加總，以此評估值評斷此次群聚結果的合適與否。

以圖 3-14 為例，當資料被區分為三個群聚時，先各別計算出每

個群聚之密度與凝聚率，再計算成對群聚間鑑別率，分別為群聚 c_1 與群聚 c_2 間、群聚 c_2 與群聚 c_3 間與群聚 c_3 與群聚 c_1 共 $C_2^3 = 3$ 組，便可得到其此次分群結果評估值，並與其他分群結果進行評估，最終選擇出一個最合適的分群結果。

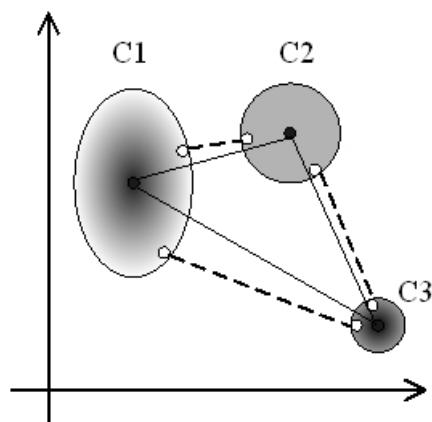


圖 3-14 適應型群聚適切性評估式示意圖

第肆章 實驗與結果

一 資料導向為基礎之選擇群聚演算法方法

此實驗目的為利用本研究所提出輔助使用者選擇群聚演算法類型的資料指標，對原始資料進行分析，讓使用者較明確的了解欲分群資料之特性，並幫助使用者選擇合適的群聚演算法類型。

實驗模擬實驗平台為 IBM 相容之 PC，其 CPU 為 Pentium 4 2.4GHZ，RAM 為 512MB，作業系統採用 Windows 2000，開發工具為 Borland JBuilder X 與 Math Works Matlab 6.5。

（一）資料密度指標

本實驗共測試了五個資料庫，各為 50、100、150、200 與 250 筆的二維正規化資料，五個資料庫內容皆為亂數產生，其個別資料密度為表 4-1。

表 4-1 資料密度分析表

	資料庫 1 (n=50)	資料庫 2 (n=100)	資料庫 3 (n=150)	資料庫 4 (n=200)	資料庫 5 (n=250)
資料密度	54.44	105.2	157.39	223.3	271.01

由以上的實驗結果中可得知，在相同的單位下，其隨著資料點增加，其密度也相對上升；而由第貳章中表 2-1 可知，大部分的群聚演算法時間複雜度皆與資料點數成正比，所以我們可得到當資料密度越

高時其群聚演算法計算的時間成本將隨之升高；換句話說一個優良的群聚演算法必須有能力處理高筆數的資料分佈即密度高的資料分佈。

(二) 資料混亂度指標

本實驗針對六個 150 筆資料的二維正規化資料資料，其實驗結果與分析，詳述如下。

資料一之資料分佈如圖 4-1，其分佈特性為完全隨機產生，所以並無明顯的群聚產生，其混亂度為 3.23。

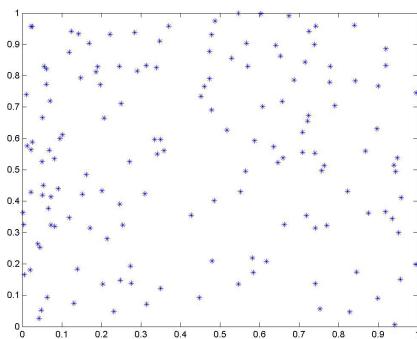


圖 4-1 混亂度資料一之資料分佈

資料二之資料分佈如圖 4-2，其分佈特徵為為四個分割不明顯且高雜訊的資料分佈，其混亂度為 3.06。

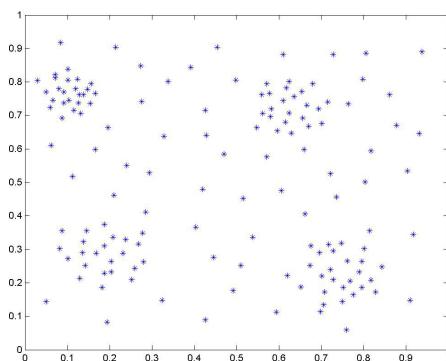


圖 4-2 混亂度資料二之資料分佈

資料三之資料分佈如圖 4-3，其分佈特徵為一個分割不明顯且高雜訊的資料分佈，其混亂度為 2.84。

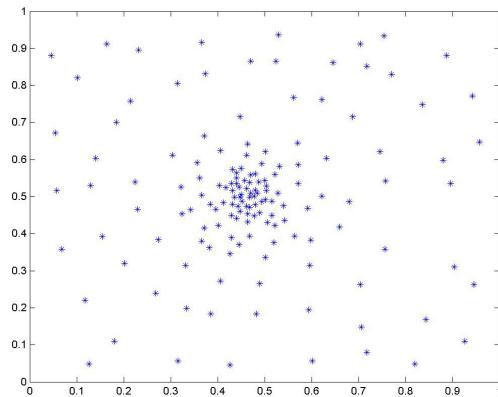


圖 4-3 混亂度資料三之資料分佈

資料四之資料分佈如圖 4-4，其分佈特徵為一個分割不明確且高雜訊但向內集中的資料分佈；如圖 4-4，其混亂度為 2.45。

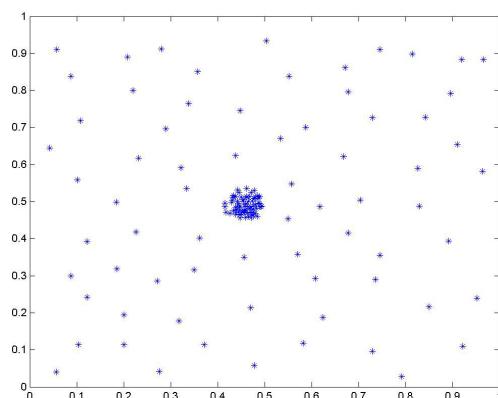


圖 4-4 混亂度資料四之資料分佈

資料五之資料分佈如圖 4-5，其分佈特徵為三個分割明確且低雜訊的資料分佈，其混亂度為 2.28。

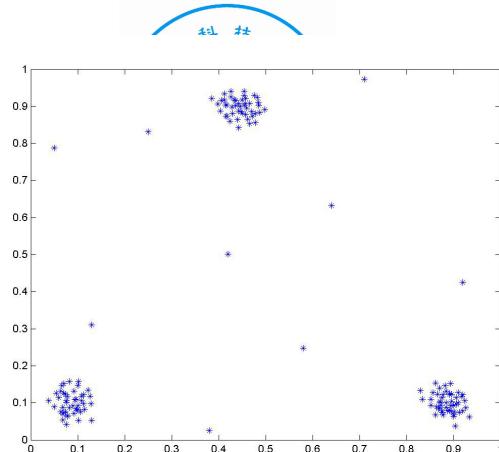


圖 4-5 混亂度資料五之資料分佈

資料庫六之資料分佈如圖 4-6，其分佈特徵為三個分割不明確但密度與外型大小相異的資料分佈，其混亂度為 2.84。

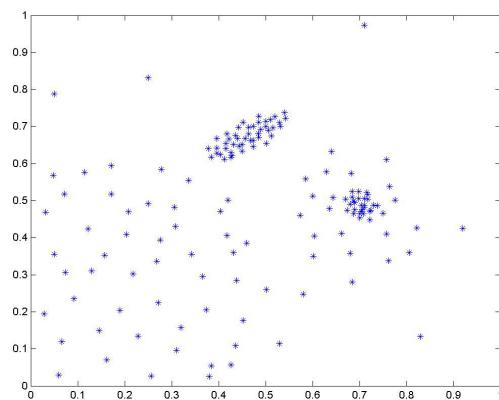
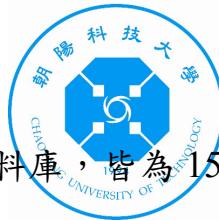


圖 4-6 混亂度資料六之資料分佈

由實驗中可得知，當資料中其群聚切割越明顯及雜訊越低時，其混亂程度降低；換句話說一個優良的群聚演算法必須有能力處理群聚切割不明顯及高雜訊的資料分佈即混亂程度高的資料分佈。

(三) 資料分離度



本實驗共實驗了八個資料庫，皆為 150 筆二維正規化資料，其實驗內容與分析，詳述如下。

資料一之資料分佈如圖 4-7，其分佈特性為完全隨機產生，並無明顯的群聚產生，其分離度為 0.47。

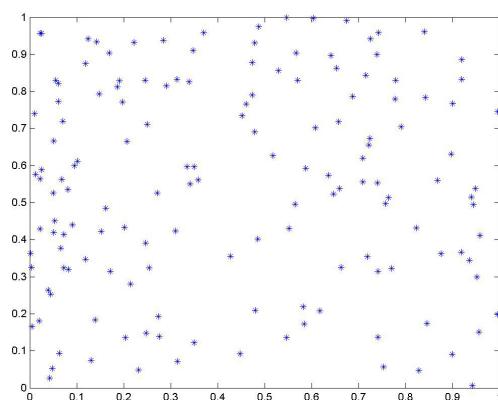


圖 4-7 分離度資料一之資料分佈

資料庫二的資料分佈如圖 4-8，其資料分佈為八個分割不明確且高雜訊的，其分離度為 0.48。

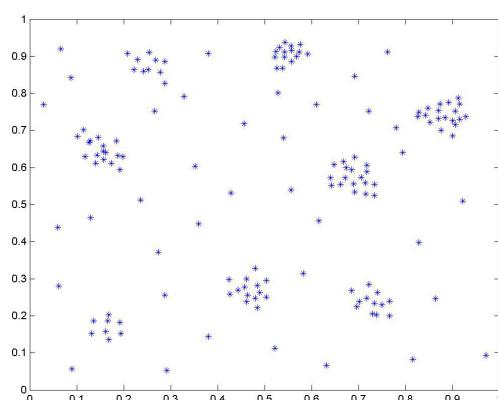


圖 4-8 分離度資料二之資料分佈

資料庫三的資料分佈如圖 4-9，其分佈為四個分割不明確且高雜訊的資料分佈，其分離度為 0.48。

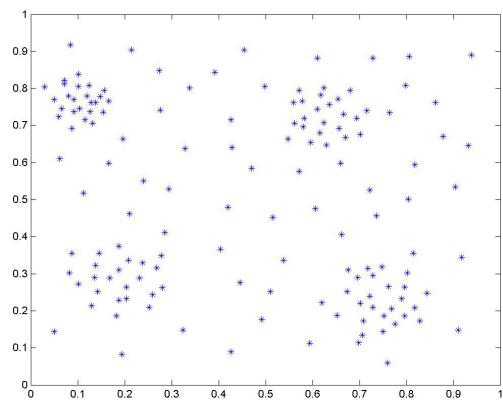
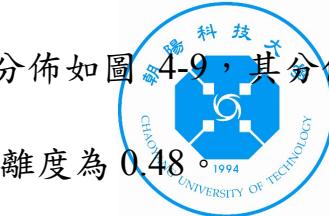


圖 4-9 分離度資料三之資料分佈

資料庫四的資料分佈如圖 4-10，其分佈為兩個分割不明確且高雜訊的資料分佈，其分離度為 0.57。

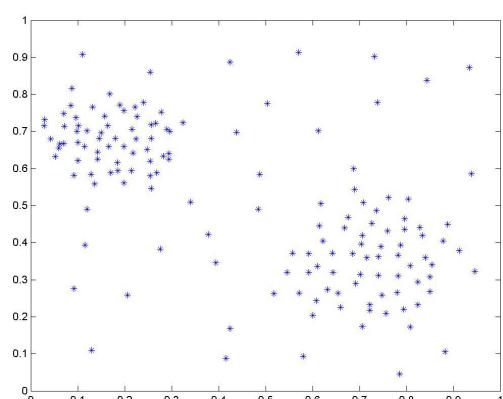


圖 4-10 分離度資料四之資料分佈

資料庫五為 150 筆二維資料，其分佈為一個分割不明確且高雜訊的資料分佈；如圖 4-11，其分離度為 0.62。

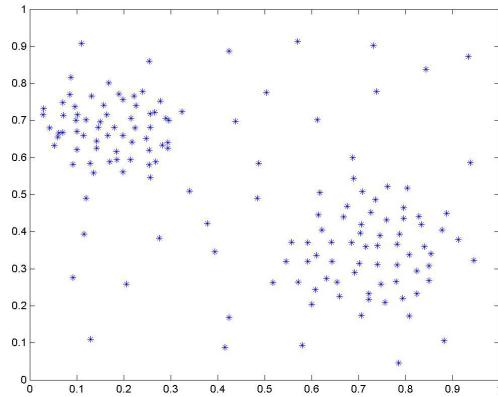


圖 4-11 分離度資料五之資料分佈

資料庫六為 150 筆二維資料，其分佈為七個分割明確且低雜訊的資料分佈；如圖 4-12，其分離度為 0.51。

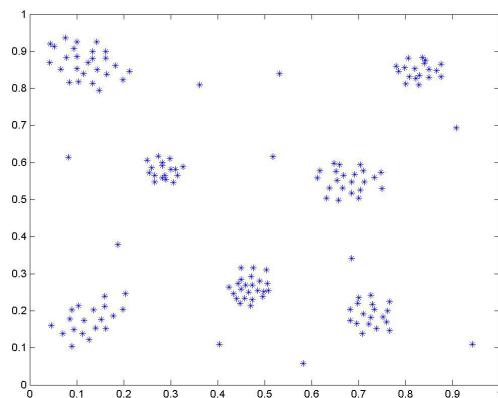


圖 4-12 分離度資料六之資料分佈

資料庫七為 150 筆二維資料，其分佈為三個分割明確且低雜訊的資料分佈；如圖 4-13，其分離度為 0.62。

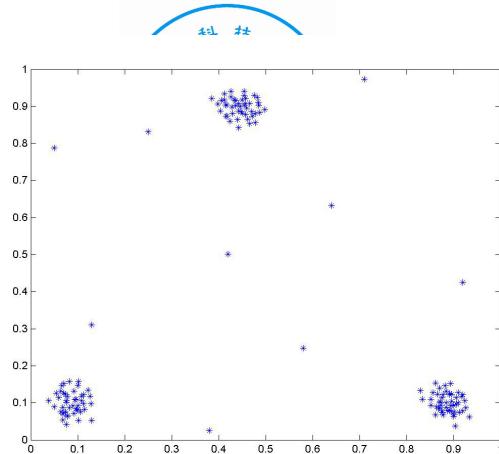


圖 4-13 分離度資料七之資料分佈

資料庫八為 150 筆二維資料，其分佈為三個分割不明確且低雜訊的資料分佈；如圖 4-14，其分離度為 0.7。

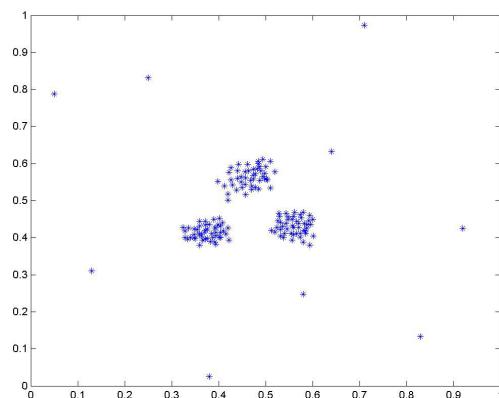


圖 4-14 分離度資料八之資料分佈

由以上實驗中可知當原始資料分佈中的群聚間，分割越明顯即群聚彼此間分離的越遠時，其分離度越大。

(四) 實驗範例



本實驗以群聚演算法效能與品質為前提下，以本研究提出的資料密度、分散度與混亂度指標，針對欲分群之原始資料進行分析，再以[2][18]所提出的各群聚演算法類型的特性為基礎所建立的規則，如表4-2，以輔助使用者選擇合宜的群聚演算法；其範例架構圖如圖3-5。本實驗模擬兩組資料分佈；其一為二維資料分佈，其資料筆數為150，合適分群數為三群而每個群聚的大小與密度皆相似，同時並有雜訊發生；其資料分佈如圖4-15。

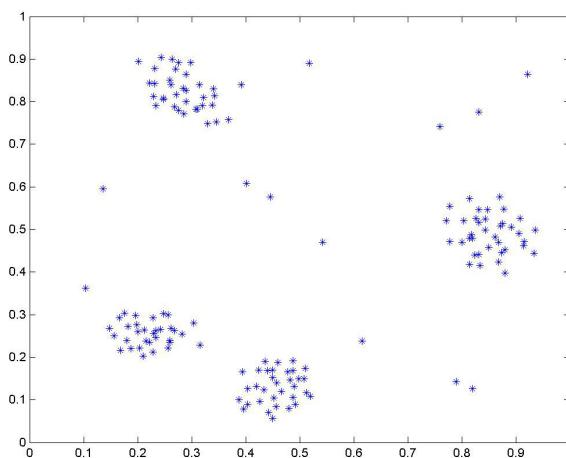
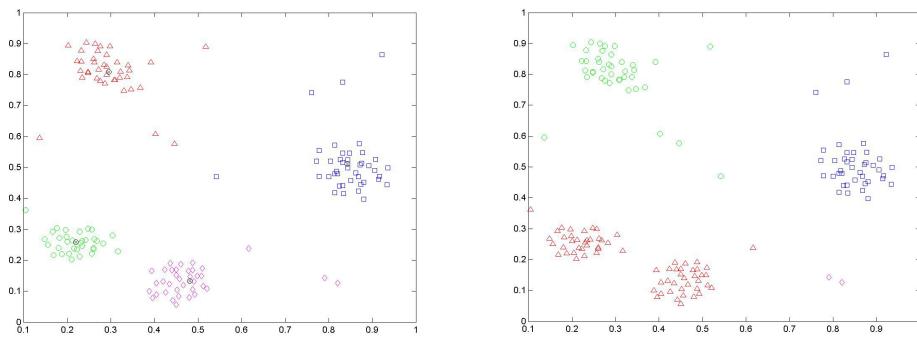


圖 4-15 範例一資料分佈

經由資料分析後，其密度為160.26、混亂度為2.52、分離度為0.539，經由本研究所提出之群聚演算法選擇模型發現其適用的演算法類型為切割式群聚演算法，接著本研究採用切割式群聚演算法與階層式群聚演算法進行分群，獲得分群結果如圖4-16。



(a) 切割式群聚演算法結果 (b) 階層式群聚演算法結果

圖 4-16 範例一分群結果示意圖

由圖 4-16 (a) 中可以發現切割式群聚演算法結果在輸入正確的群聚參數（分群數為 3）下，其分群結果皆為正確；但是階層式群聚演算法結果圖 4-16 (b) 在輸入正確的群聚參數（分群數為 3）下，其分群結果出現大量不適合的群聚結果。

另一組為二維資料分佈，其資料筆數為 150，合適分群數為三群每個群聚的大小與密度皆不相同，同時並有雜訊發生；其資料分佈如圖 4-17。

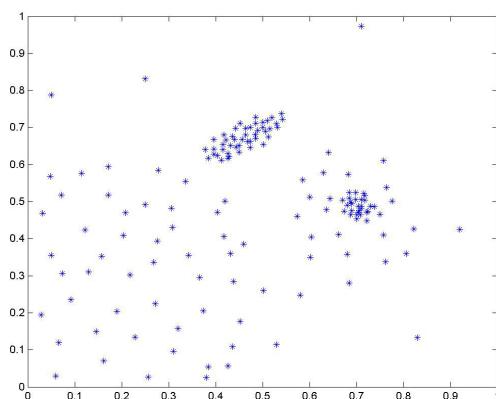
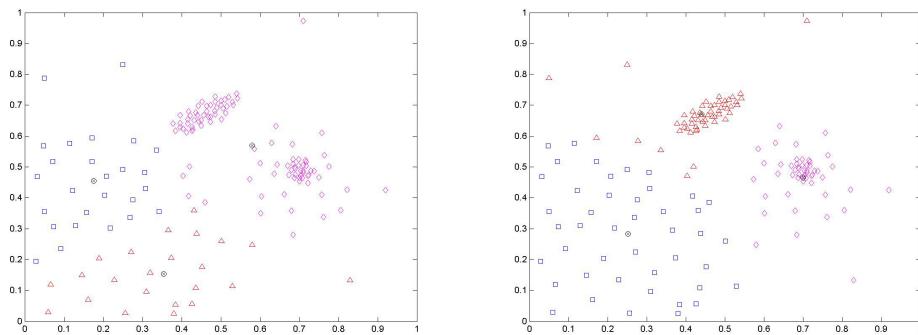


圖 4-17 範例二資料分佈

經由資料分析後，其密度為 155.65、混亂度為 2.8、分離度為 0.56，
 經由本研究所提出之群聚演算法選擇模型發現其適用的演算法類型
 為階層式群聚演算法，接著本研究採用階層式群聚演算法與切割式群
 聚演算法進行分群，獲得分群結果如圖 4-18。



(a) 切割式群聚演算法結果 (b) 階層式群聚演算法結果

圖 4-18 範例二分群結果示意圖

由圖 4-18 (a) 中可以發現切割式群聚演算法結果在輸入正確的
 群聚參數（分群數為 3）下，出現大量不適合的群聚結果；但是階層
 式群聚演算法結果圖 4-18 (b) 在輸入正確的群聚參數（分群數為 3）
 下，其分群結果皆為正確。

二 適應型群聚適切性評估式

本實驗包含 6 筆資料分佈，區分為三組各別為群聚間的外型、密度及群聚內密度，而每組中各分為無雜訊影響及有雜訊影響的資料分佈；其中資料筆數為 200 筆的二維資料，並利用 K-means 群聚演算法將其區分為二至五群。透過此次實驗將本研究提出的群聚適切評估式與 Dunn 評估式及 DB 評估式做比較，經由不同的群聚適切性評估式挑選出合適的群聚結果，實驗詳述如下。

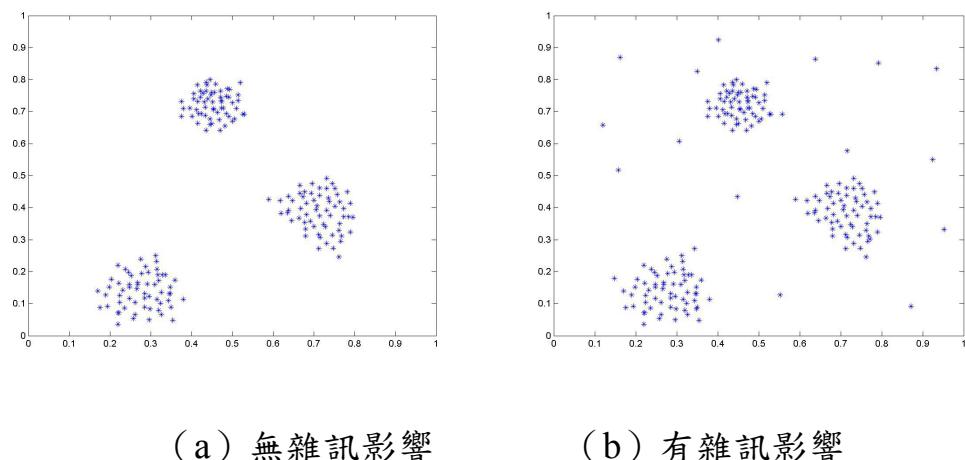


圖 4-19 適切性評估式資料一之資料分佈

圖 4-19 (a) 中是三個外型大小相似的群聚所構成，其中群聚間分隔的非常明顯，並無明顯雜訊干擾；其合適的分群數量為三群。本研究使用 K-means 群聚演算法將圖 4-19 (a) 區分為二群、三群、四群及五群（如圖 4-20 所示）。

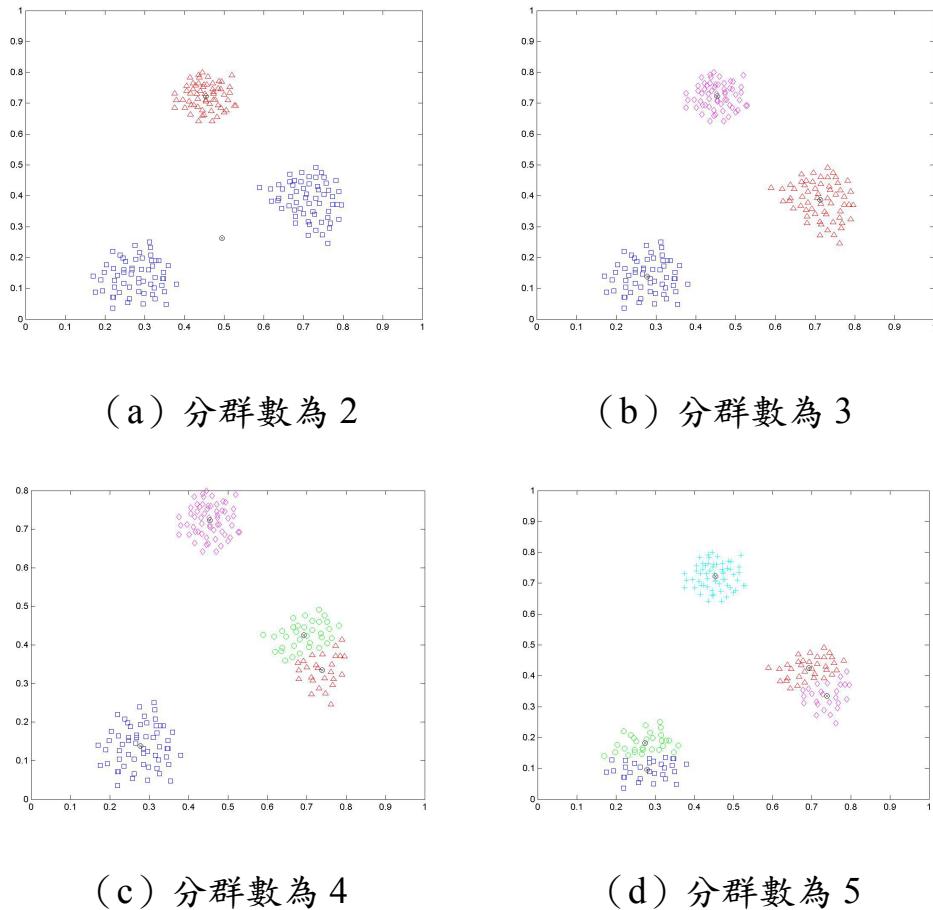


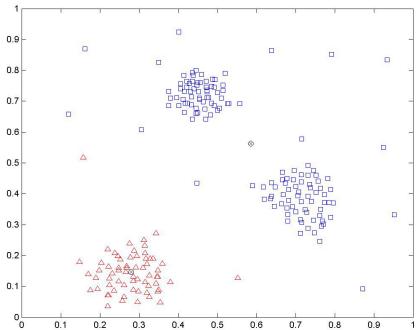
圖 4-20 適切性評估式資料一 (a) 之分群結果

由表 4-2 可知本研究提出之群聚適切性評估式、Dunn 評估式及 DB 評估式，皆選擇符合實際的情況的分群數三群。

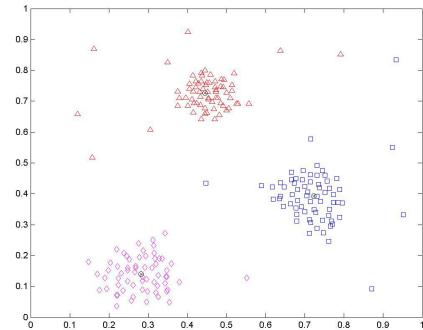
表 4-2 適切性評估式資料一 (a) 之評估值

分群數	I_{adapt}	Dunn's	DB
2	4.2511	0.35	1.009
3	5.157	0.9925	0.2812
4	5.0068	0.0768	0.6265
5	5.1224	0.082	0.9028

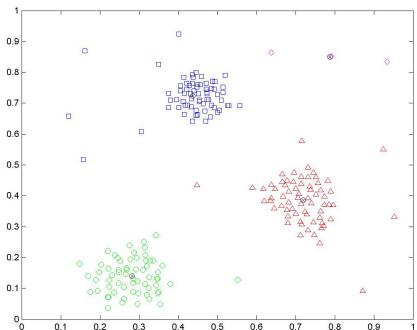
圖 4-19 (b) 由三個數量相似的群聚所構成，其中群聚間分隔的非常明顯，並有明顯雜訊干擾；其合適的分群數量為三群。本研究使用 K-means 群聚演算法區分為二群、三群、四群、五群（圖 4-21 所示）。



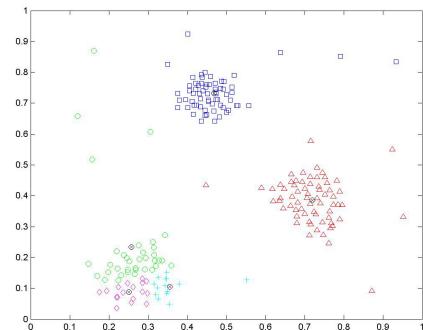
(a) 分群數為 2



(b) 分群數為 3



(c) 分群數為 4



(d) 分群數為 5

圖 4-21 適切性評估式資料一 (b) 之分群結果

由表 4-3 可知本研究之適切性評估式與 DB 評估式將會選擇分群數量為三群，符合實際的情況；而 Dunn 評估式將會選擇分群數量為四群，不符合實際的情況。

表 4-3 適切性評估式資料一 (b) 之評估值

分群數	I_{adapt}	Dunn's	DB
2	3.5245	0.1377	0.899
3	4.37	0.1922	0.3667
4	4.2388	0.2557	2.4441
5	2.3075	0.014	0.815

圖 4-22 (a) 由三個數量相異的群聚所構成，其中群聚間分隔非常明顯，並無明顯雜訊干擾；其合適的分群數量為三群。

本研究使用 K-means 群聚演算法將圖 4-22 (a) 其區分為二群、三群、四群、五群（如圖 4-23 所示）。

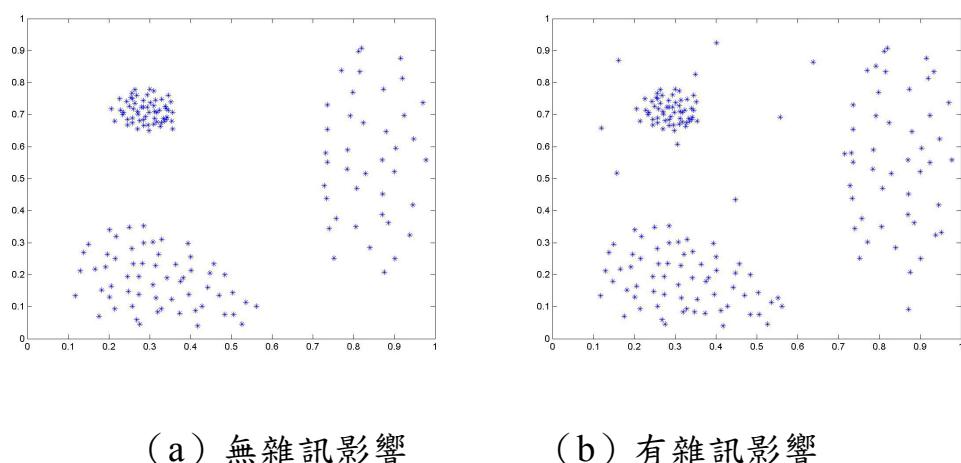


圖 4-22 適切性評估式資料二之資料分佈

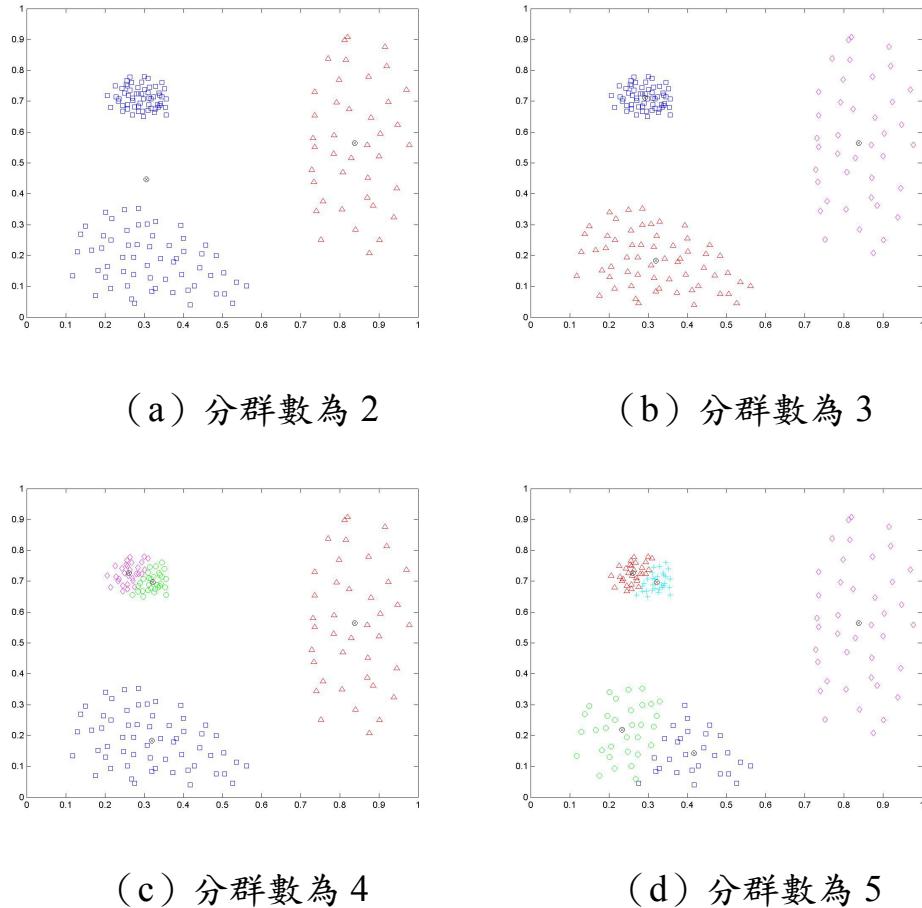


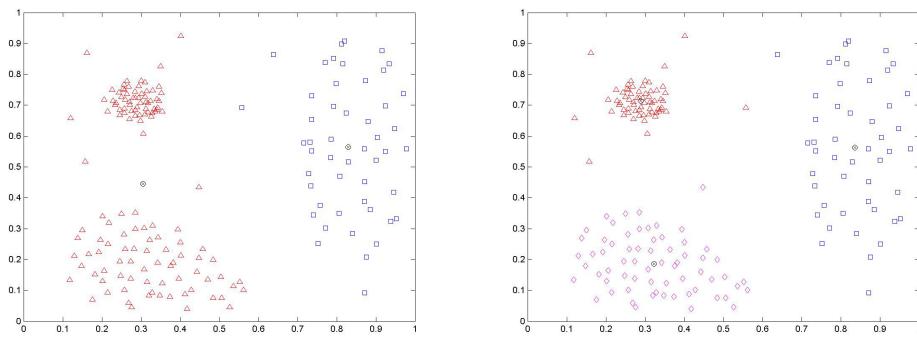
圖 4-23 適切性評估式資料二（a）之分群結果

由表 4-4 可知利用本研究提出之適應型適切性評估式與 DB 評估式將會選擇符合實際的情況的分群數量為三群；而 Dunn 評估式將會選擇不符合實際的分群數量為四群。

表 4-4 適切性評估式資料二（a）之評估值

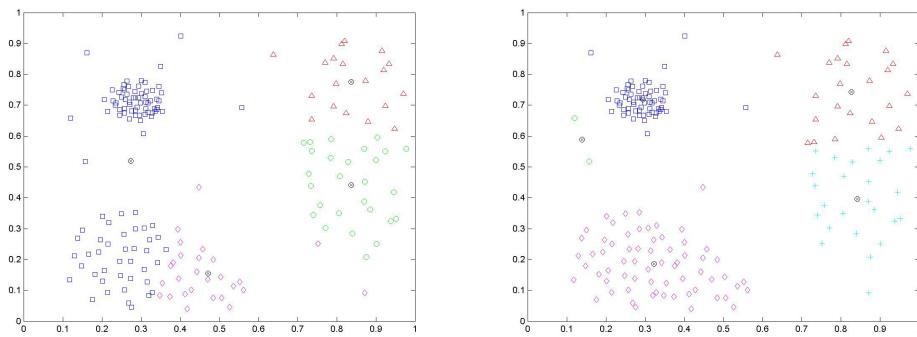
分群數	I_{adapt}	Dunn's	DB
2	3.735	0.3114	1.2729
3	4.5339	0.3447	0.4638
4	4.3344	1.0903	0.5693
5	4.3882	0.0145	0.8833

圖 4-22 (b) 由三個數量相異的群聚所構成，其中群聚間分隔的非常明顯，並有明顯雜訊干擾，其合適的分群數量為三群。本研究使用 K-means 群聚演算法將其區分為二群、三群、四群、五群，其結果如圖 4-24 所示。



(a) 分群數為 2

(b) 分群數為 3



(c) 分群數為 4

(d) 分群數為 5

圖 4-24 適切性評估式資料二 (b) 之分群結果

由表 4-5 可知利用本研究提出適切性評估式與 DB 評估式將會選擇符合實際情況的分群數量為三群；而 Dunn 評估式將會選擇不符合實際情況的分群數量為二群。

表 4-5 適切性評估式資料二 (b) 之評估值



分群數	I_{adapt}	Dunn's	DB
2	2.79	0.2269	1.3393
3	2.85	0.2226	0.5029
4	2.0102	0.001	0.6087
5	2.4849	0.0059	0.6551

圖 4-25 (a) 由三個密度相異的群聚所構成，其中群聚間分隔的非常明顯，並無明顯雜訊干擾；其合適的分群數量為三群。

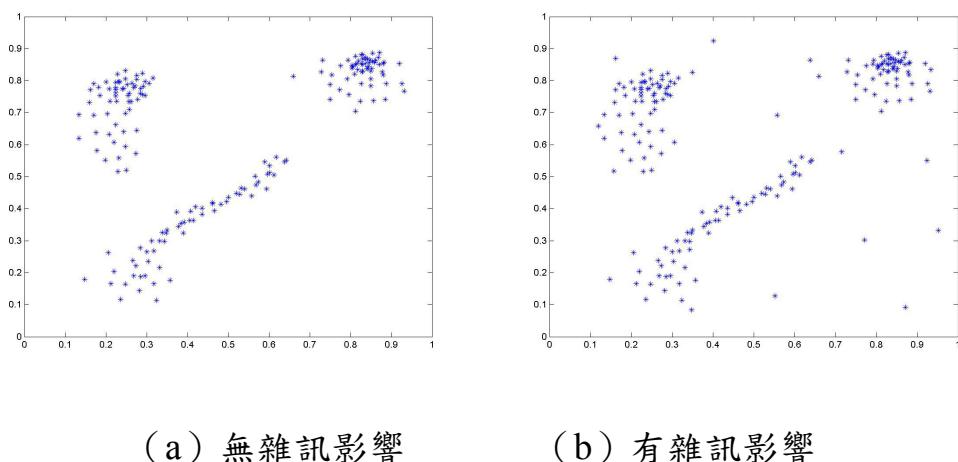


圖 4-25 適切性評估式資料三之資料分佈

本研究使用 K-means 群聚演算法將圖 4-25 (a) 區分為二群、三群、四群、五群；其結果如圖 4-26 所示。

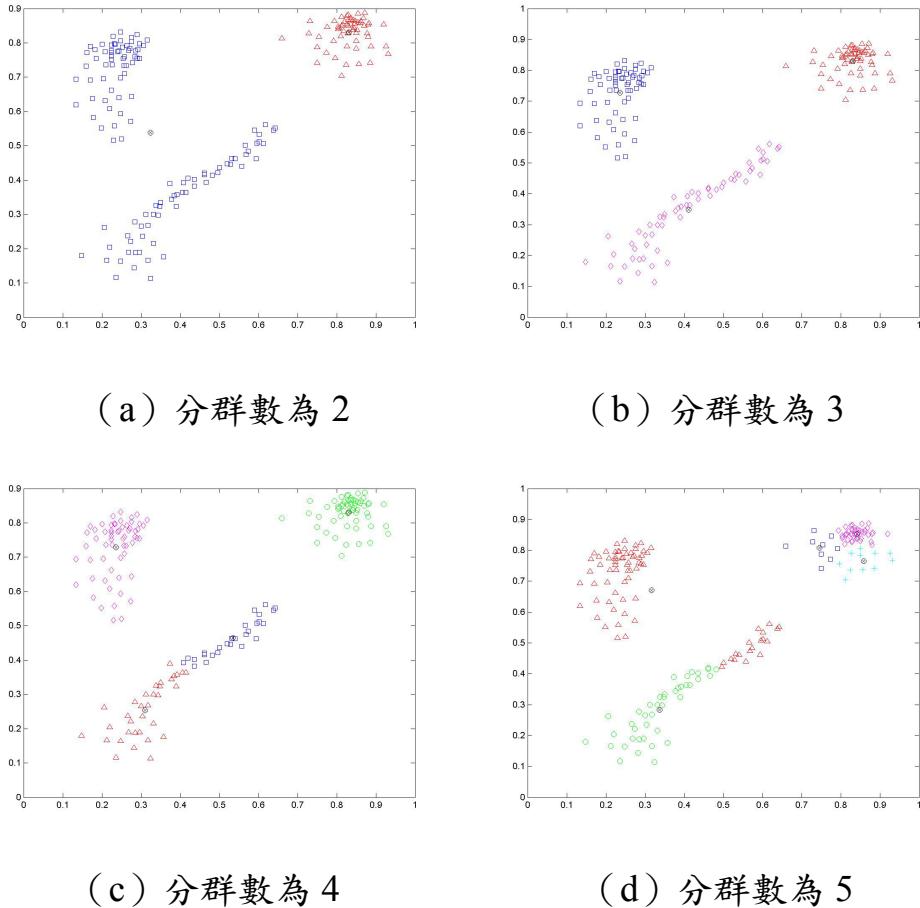


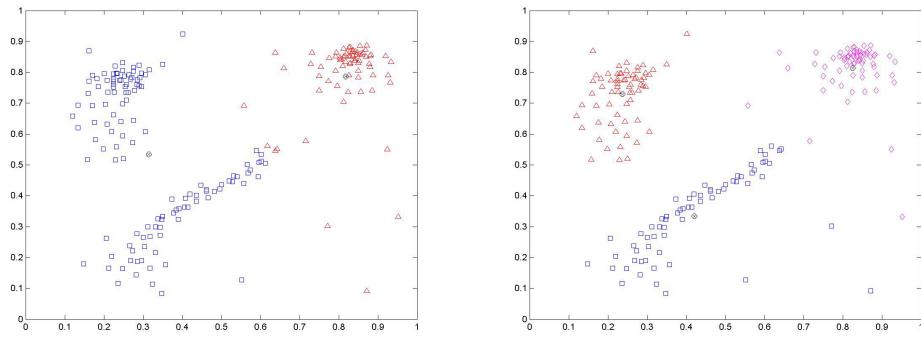
圖 4-26 適切性評估式資料三（a）之分群結果

由表 4-6 可知利用本研究提出之適切性評估式，將會選擇符合實際的情況的分群數量為三群；而 DB 評估式與 Dunn 評估式皆選擇不符合實際的情況的分群數量為四群與二群。

表 4-6 適切性評估式資料三（a）之評估值

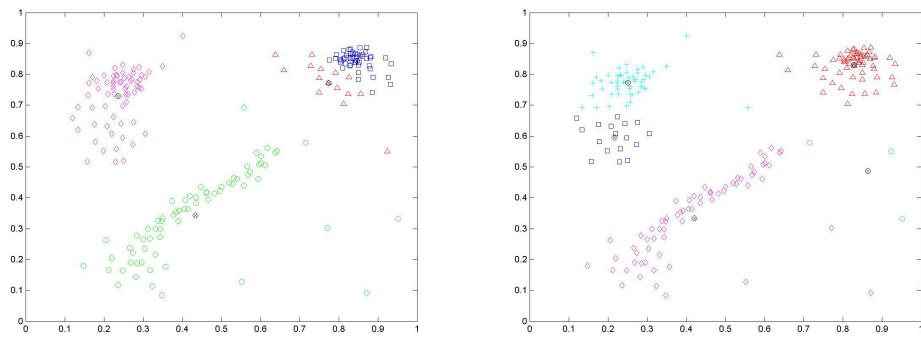
分群數	I_{adapt}	Dunn's	DB
2	2.91	0.302	0.763
3	3.055	0.2905	0.4917
4	3.001	0.0868	0.4397
5	2.0752	0.0001	4.6524

圖 4-25 (b) 由三個密度相異的群聚所構成，其中群聚間分隔的非常明顯，並有明顯雜訊干擾；其合適的分群數量為三群。本研究使用 K-means 群聚演算法將其區分為二群、三群、四群、五群；其結果如圖 4-27 所示。



(a) 分群數為 2

(b) 分群數為 3



(c) 分群數為 4

(d) 分群數為 5

圖 4-27 適切性評估式資料三 (b) 之分群結果

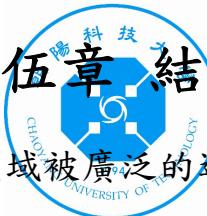
由表 4-7 可知利用本研究提出之適切性評估式、DB 評估式及 Dunn 評估式皆同時選擇符合實際的情況的分群數量為三群。

表 4-7 適切性評估式資料三 (b) 之評估值

分群數	I_{adapt}	Dunn's	DB
2	2.15	0.0383	0.9612
3	2.79	0.1075	0.5348
4	2.2889	0.0005	1.4871
5	2.1431	0.0002	0.9456

以上實驗是將本研究所提出的群聚適切性評估式與傳統 DB 評估式及 Dunn 評估式進行交互比較各種資料分佈，其中包含不同群聚外型 不同群聚大小、不同群聚密度及群聚內不同密度及雜訊的各種資料分佈，唯有本研究所提出的群聚適切性評估式適應任何資料分佈均能選擇合適的群聚數量及群聚演算法輸入參數。

第五章 結論



隨著群聚技術在各個領域被廣泛的運用，許多不足之處逐一浮現，如部分維度與群聚演算法的選擇、群聚演算法參數的設定及合理的應用領域等。本研究首先針對群聚相關演算法與群聚適切性評估式進行詳細的討論，接著將對其特性作一番分析；接著本研究藉由群聚相關技術的分析，提出提高群聚技術品質及實用性的兩個方法，其一為利用輸入資料的密度、混亂度及分離度三個指標的配合，找尋出輔助使用者選擇群聚演算法類型的方法；其二為根據群聚相互間與內部資訊的量測方法，提出一較佳適應性的群聚適切性評估式。

本研究所提出的方法屬於較創新的方法，因此其中尚有許多待改進及可繼續探討的空間，以下就對未來可改進及研究的方向做說明：

(1) 納入目前新興的群聚演算法類型進行分析：由於近期被提出群聚演算法類型；如啟發式群聚演算法、競爭式群聚演算法類型或熵函數為基礎的群聚演算法類型，皆具有不錯的群聚品質，所以仍可加入新概念的指標使其架構適應其目前未納入的群聚演算法類型。

(2) 引入真實資料庫並由中找出更具代表性的權重：藉由導入真實資料庫的訓練，以得到適切性評估式中評估群聚間與群聚內兩者間的權重比例，使得此適切性評估式獲得更佳的適應性。



- [1] Ankerst M., Breunig M., Kriegel H.P. and Sander J., “optics: Approach to Spatial Data Mining”, *VLDB'97*, 1997, Vol 18, pp. 144-155.
- [2] Berkhin P., Survey of Clustering Data Mining Techniques, Technical Report, 2002.
- [3] Bezdek J.C. , Pal. N.R., “Some new indexes of cluster validity”, *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 1998, Vol. 28(3) ,pp 301-315.
- [4] Calinski T., Harabasz. J., “A dendrite method for cluster analysis”, *Communications in Statistics*, 1974, Vol. 3, pp.1-27.
- [5] Davies, DL, Bouldin, D.W., “A cluster separation measure”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, Vol. 1(2).
- [6] Dunn .J.C., “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters”, *J. Cybernet.*, 1974, Vol. 3(3), pp. 32-57.
- [7] Dunn J. C., “Well separated clusters and optimal fuzzy partitions”, *J.Cybern*, 1974, Vol. 4, pp. 95-104.
- [8] Ester M., kriegel H.P., Sander J. and Xu X., “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *Knowledge Discovery and Data Mining*, pp.226-231.
- [9] Gokcay E., Principe J.C., “Information theoretic clustering”, *PAMI*, 2002, Vol. 24, pp. 158-171,

- [10]Halkidi M., Batistakis Y. and Vazirgiannis M., “On Clustering Validation Techniques”, *Journal of Intelligent Information Systems*, 2001, Vol. 17(2), pp.107-145.
- [11]Han J., Kamber M., Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, 2000.
- [12]Jenssen R., Hild K. E., Erdogmus D., Principe J. C. and Eltoft T., “Clustering using Renyi's entropy”, *IJCNN2003*, 2003, pp. 523-528
- [13]Juha V., Esa A., “Clustering of the Self-Organizing Map”, *IEEE Transactions on Neural Networks*, 2000, vol.11 (3), pp586-600
- [14]Karypis G., Han E.H., “Chameleon: Hierarchical Clustering Using Dynamic Modeling”, *IEEE Computer*, 1999, Vol. 32(8).
- [15]Kaufman L., Rousseeuw P.J., Finding Groups in data: an Introduction to clustering Analysis, *John Wiley & Sons*, 1990.
- [16]Klir G. J., Ute S. C. and Bo Y., ”Fuzzy Set Theory: Foundations and Applications”, *Prentice Hall*, 1997
- [17]Kohonen T., “Self-Organizing Formation of Topologically Correct Feature Maps”, *Biological Cybernetics*, 1982, Vol.43, pp. 59-69.
- [18]Kolatch E., Clustering Algorithms for Spatial Databases: A Survey, *Dept. of Computer Science, University of Maryland, College Park*, 2000
- [19]Lawrence H., Arabie P., “Comparing partitions”, *Journal of Classification*, 1985, vol. 2, pp.193-218.
- [20]Li R.P., Mukaisono M., “A maximum-entropy approach to fuzzy clustering”, *Fuzzy IEEE*, 1995, pp.2227-2232
- [21]Michael J. A., Gordon L., “Data Mining Techniques For marketing

- Sales and Customer Support”, *John Wiley & Sons*, 1996.
- [22] Ng R.T., Han J., “CLARANS: A Method for Clustering Objects for Spatial Data Mining”, *Transactions on Knowledge and Data Engineering*, 2002, pp.1003-1016.
- [23] Pal N.R., Biswas J., “Cluster Validation using Graph Theoretic Concepts”, *Pattern Recognition*, 1997, Vol. 30(6), pp. 847-857.
- [24] Sheikholeslami G., Chatterjee S., and Zhang A., “WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases”, *VLDB98*, 1998, pp. 428-439.
- [25] Theodoridis S., Koutroubas K., *Pattern recognition*, Academic Press , 1999
- [26] Tian Z., Raghu R. and Miron L., “BIRCH: A Efficient Data Clustering Method for Very Large Databases”, *SIGMOD*, 1996, pp.103-114.
- [27] Wang W., Muntz Yang R., “STING: A Statistical Information grid Approach to Spatial Data Mining”, *VLDB'97*, pp.353-358.
- [28] Zadeh L. A., “Fuzzy sets”, *Information and Control*, 1965, Vol. 8(3), pp. 338-353.
- [29] 林育臣，「群聚技術之研究」，碩士論文，朝陽科技大學資訊管理系，2002年。
- [30] 譚嘉慧，「模糊分類適切性分析」，碩士論文，中原大學數學系，1999年。