FCLARANS (Faster Clustering Large Applications based on Randomized Search) – 以 Microarray 為例

陳同孝	陳民枝	蔡螢池	陳緯祐	王凱慶	陳師融	施富祥	吳俊岳
國立臺中技術學院資訊科技與應用研究所	國立臺中技術學院資訊科技與應用研究所	國立臺中技術學院資訊科技與應用研究所	國立臺中技術學院資訊工程系	國立臺中技術學院資訊工程系	國立臺中技術學院資訊工程系	國立臺中技術學院資訊工程系	國立臺中技術學院資訊工程系
ttschen@gmail.com	jeanne@ntit.edu.tw	bigear42@gmail.com	lr210032@hotmail.com	p_34520@hotmail.com	fallon90258@hotmail.com	fnank0928@hotmail.com	gol1223@hotmail.com

摘要(Chinese Abstract)

資料探勘的技術可以幫助企業找出作為決策依據或具有價值的知識，而分群演算法是重要的資料探勘技術，其分群結果有利於後續研究發展。而分群演算法中 CLARANS 可得到精準的分群結果，但是隨著資料量與分群數的增多，所需的執行時間也相當程度的增加，所以本研究提出單點比對與距離矩陣兩種方法一起應用於 Microarray 上，可節省下 CLARANS 分群演算法的 95%以上的時間，而且不受分群數影響又不失準確度。

關鍵詞：資料探勘、分群演算法、CLARANS、K-medoids、微生物晶片

Abstract

Data mining technology can help enterprises with decision-making based on the mined knowledge. The clustering algorithm is an important data mining technique where results are grouped in favor of the follow-up research and development. CLARANS is a clustering algorithm with accurate results in grouping. However, the amount of data and clustering increased the time required for implementation. This study proposed using the comparison of single-point with distance matrix in microarray to save more than 95% of the required time in the clustering algorithm CLARANS which will not affect the amount of clustering and accuracy.

Keywords: **Data Mining, Clustering Algorithms, CLARANS, K-medoids, Microarray**

1. 前言 (Introduction)

資訊化時代的來臨，帶來了大量的資料透明化，而這些資料可以藉由資料探勘(Data Mining)的技術，找出具有價值或關聯的知識及規則，將可作為企業決策者改善策略規劃時的依據。資料探勘包含有許多數理統計分析的方法，例如：分群演算法[5]、分類演算法[5]、類神經網路[5]、支持向量機器[9]、統計[10]與人工智慧[1]等。

分群演算法是重要的資料探勘技術，作法是根據欲分群的資料集，依照各種分群演算法，利用資料特性將資料集區分為數個資料群集，而分群結果亦可依使用者的需求來找出關聯規則等知識，做為後續研究發展的來源。並於近年來資料探勘被大量的應用於生物晶片(Micro-array)的資料分析[9]，利用將舊有的生物晶片資料分群或分類，幫助生物學家來分析了解各基因之間的特性及關聯，進而對後續研究有所助益。

目前分群演算法又可分為兩大類：階層式分群演算法(Hierarchical Clustering Algorithms)與分割式分群法（Partition Clustering Algorithms），前者可再細分為聚合法(Agglomerative Algorithm)與分裂法(Divisive algorithm)二種[8]，而後者則是將原始資料以分割的方式，分為數個資料群集，並讓每個資料群集之間沒有關聯，代表每個資料群集為獨立的群聚，而分群結果可依使用者的需求來找出關聯規則等知識，做為後續研究發展的來源。

分割式分群法主要為 K-means 及 K-medoids 兩種，K-medoids 的各種分群演算法

較 K-means 更不易受到雜訊(noise) 及 離群值(outlier)的影響，可得到較好的分群結果，而 K-medoids 中 CLARANS(Clustering Large Applications Based upon Randomized Search)[6]採取大量隨機樣本的方式進行分群，用以獲得較佳的分群結果，對大量資料分群較為有效，亦可應用於更多的範圍，但是在整體分群速度上卻是比 K-means 慢許多。所以本研究以 CLARANS 分群演算法為研究對象，期望在較好的分群結果之外，還能有兼具較佳分群速度。

2. CLARANS分群演算法

PAM(Partitioning Around Medoids)[3]及 CLARA (Clust-ering Large Application)[3]兩分群演算法為 CLARANS 分群演算法的前身，而 CLARANS 分群演算法比 PAM、CLARA 分群演算法更能應用於大量資料，並可得到最佳的分群結果。CLARANS 分群演算法步驟如下：

步驟 1. 輸入分群數 *k* 及 *maxNeightbor* 和 *numLocal* 變數，初始化 *Count* 及 *Iteration* 為的值 0，*minCost* 為∞，*bestNode* 為空集合。

步驟 2. 隨機選 *k* 個資料為 medoid，代表分群中心，而 k-medoids 為集合 *current*。

步驟 3. 在 *current* 集合中隨機挑選 1 個 k-medoid 替換為非 medoid 成為新集合 *S*。

步驟 4. 計算所有 *current* 集合中 k-medoid 與各資料計算距離並加總為 $TC_{current}$，並將 S 集合中各資料與其他資料計算距離並加總為 TC_s。判斷 $TC_{current}<TC_s$ 是否成立，是的話，將 *S* 集合取代 *current* 集合及 TC_s 取代 $TC_{current}$，及 *Count* = 0，並回到步驟 3；否，則 *Count* 累加至下一步驟。

步驟 5. 判斷 *Count* 是否等於 *maxNeightbor*，如果等於，則 *Iteration* 累加後進入下一步驟；不等於的時候，則回到步驟 3。

步驟 6. 判斷 *minCost* 是否大於 $TC_{current}$，如果大於，則 $TC_{current}$ 取代 *minCost*，並以 *current* 集合取代為 *bestNode* 集合；小於則不做變更。進入下一步驟。

步驟 7. 判斷 *Iteration* 是否等於 *numLocal*，等於則輸出最佳解 *bestNode* 及所有分群中心與群內各資料距離加總 *minCost*；不等於，則回到步驟 2。

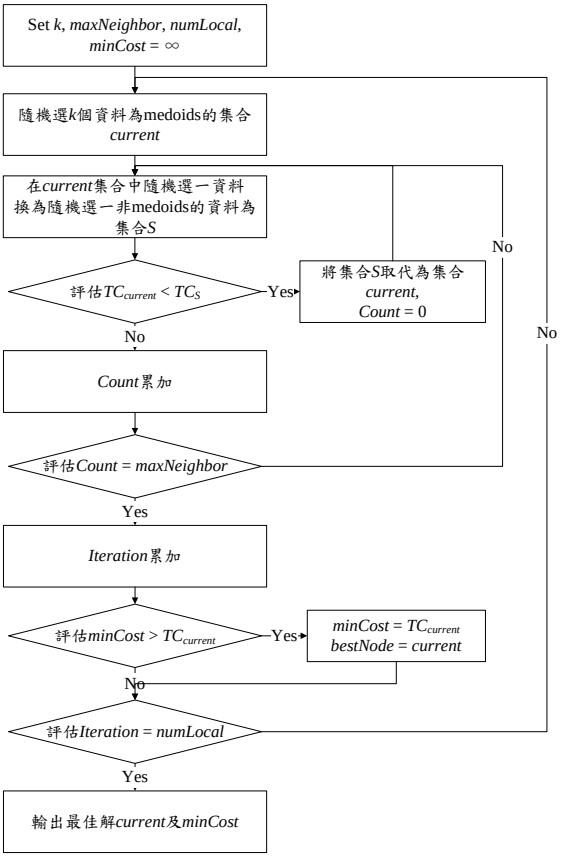

圖 1 CLARANS 分群演算法之流程圖

由圖 1 可得知，CLARANS 分群方式是藉由大量隨機採取樣本的方式，然後新舊集合與各資料的距離加總，留下新舊集合中較佳的分群結果，最後獲得較佳的分群結果，但是相對於大量樣本的運算，也需要相當大量的時間，而時間是資料探勘技術的重要指標，若是有個股票預測分類演算法可以精準的預測隔天的股市漲跌，但是需要一星期的時間完成預測，等到得知預測結果的時候，事實早已發生，所以好的分群演算法不僅需要有好的分群結果，亦需兼顧時間性，固本研究將針對新舊集合與各資料的距離加總部份進行探討。

3. FCLARANS分群演算法

由於 CLARANS 分群演算法的主要時間著重於在比較次數部份，所以本研究將針對此部份進行改進。本研究提出了兩種加速 CLARANS 分群演算法的方法，在本實驗中，將會一起使用來加速 CLARANS 分群演算法。

第一個方法是改進步驟 4 的部份，將原始 k-modeid 的集合與各點距離加總，修改為只計算原 k-mdeoid 與隨機挑選的非 k-medoid 與各點距離加總運算，如表 1 所示，原本運算次數

為 $2 \cdot k \cdot m_k$，k 為欲分群數量，m_k 為非 medoid 屬於第 群的資料量，而本方法將可在加速於比較距離時，只運算的 k 距離加總，如此即可將大量的減少運算，有效的加快分群速度並提升整體效能。

表 1 次數比較表

	運算次數
CLARANS	$2 \cdot k \cdot m_k$
單點比對	$2 \cdot m_k$

第二個方法則是在步驟 1 之後加入一步驟，使用距離矩陣將運需要計算的距離存放在陣列裡，此陣列為 n · n 矩陣，如圖 2 所示，依照此陣列可於演算法開始之前，先經過互相對應的資料兩兩運算，求得資料之間的距離後存入陣列中，此方法可以避免重覆運算，亦能有效的減少運算並加快整體演算法速度。

$$D = \begin{array}{c} \\ 1 \\ 2 \\ . \\ . \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & ... & .... & n \end{array} \\ \begin{bmatrix} 0 & d_{12} & d_{13} & . & d_{1n} \\ d_{21} & 0 & d_{23} & . & d_{2n} \\ . & . & 0 & . & . \\ . & . & . & 0 & . \\ d_{n1} & d_{n2} & . & . & 0 \end{bmatrix} \end{array}$$

圖 2 $n \cdot n$ 距離矩陣示意圖

修改分群演算法為 FCLARANS(Faster Clustering Large Applications based on Randomized Search)，改善時間性的問題，又不失準確度的有效分群演算法。詳細分群演算法步驟如下：

步驟 1. 輸入分群數 k 及 *maxNeightbor* 和 *numLocal* 變數，初始化 *Count* 及 *Iteration* 為的值 0，*minCost* 為 ∞，*bestNode* 為空集合。

步驟 2. 將所有資料 n 與其他資料計算距陣放入 $n \times n$ 矩陣中。

步驟 3. 隨機選 k 個資料為 medoid，代表分群中心，而 k-medoids 為集合 *current*。

步驟 4. 在 *current* 集合中隨機挑選 1 個 k-medoid 替換為非 medoid 成為新集合 *S*。

步驟 5. 計算所有 *current* 集合中隨機選一點 k-medoid 與各資料計算距離並加總為 PTC_{test}，並將 S 集合中各資料與其他資料 *new* 計算距離並加總為 PTC_{new}。判斷 $PTC_{test} < PTC_{new}$ 是否成立，是的話，將 *new* 取代 *current* 集合中的 *test* 及重新計算 $TC_{current}$，及 *Count* = 0，並回到步驟 3；否，則 *Count* 累加至下一步驟。

步驟 6. 判斷 *Count* 是否等於 *maxNeightbor*，如果等於，則 *Iteration* 累加後進入下一步驟；不等於的時候，則回到步驟 3。

步驟 7. 判斷 *minCost* 是否大於 $TC_{current}$，如果大於，則 $TC_{current}$ 取代 *minCost*，並以 *current* 集合取代為 *bestNode* 集合；小於則不做變更。進入下一步驟。

步驟 8. 判斷 *Iteration* 是否等於 *numLocal*，等於則輸出最佳解 *bestNode* 及所有分群中心與群內各資料距離加總 *minCost*；不等於，則回到步驟 2。

4. 微生物晶片(Microarray)

資料探勘需要有足夠的資料量，探勘出來的知識才具有意義，所以本研究將採用史丹佛大學所提供的微生物晶片資料庫(Microarray)[7]中的乳線癌資料庫[7]，此資料庫主要以平滑肌細胞(SMCs)來判別罹患乳線癌的可能性，其平滑股細胞主要分布於人體中的動脈和靜脈壁、膀胱、子宮、男性和女生生殖道、消化道、呼吸道、眼睛的睫狀肌和虹膜等等，其細胞的收縮功能扮演著各器官重要的角色，如排尿、呼吸作用及血管的維護等功用，所以平滑肌細胞的變異常常會暗藏許多器官重大異變的關鍵。

資料庫總共有 4780 基因，如圖 4 所示，共有 60 個維度，如圖 3 所示，[2]建議將資料庫分為兩群，分別為血管(Vascular)和內臟(Visceral)兩群，而血管部份有 2338 筆，內臟部份有 2462 筆。

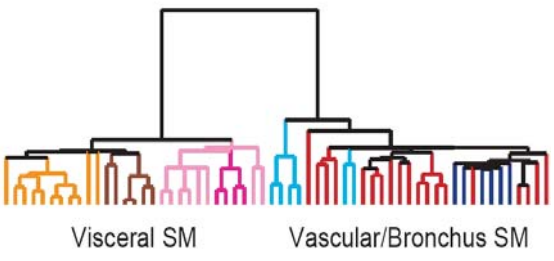

圖 3 乳線癌之樹狀結構圖

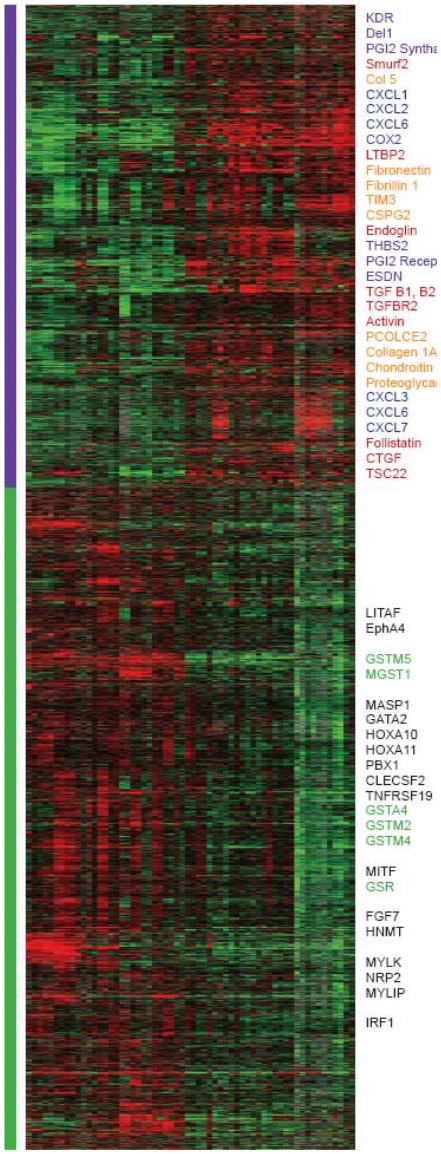

圖 4 微生物晶片資料庫

5. 實驗結果

將 Microarray 資料庫進行 CLARANS 分群演算法及本研究所提出兩種改良方法進行分群，首先是依照[2]的建議，將微生物晶片資料庫分為兩群，但是本研究期望能有更多的實驗數據證明，所以將微生物晶片資料庫再分為三群、四群，並與其樹狀圖分群結果做比對，計算出準確度，準確度公式如公式 1，A 為準確度，r 為與樹狀圖分群結果相同筆數，n 為所有資料筆數。

$$A = \frac{r}{n} \tag{1}$$

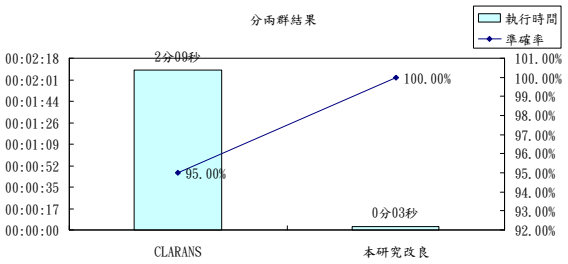

圖 5 分兩群結果

圖 5 為將 Microarray 資料庫分為兩群的結果，CLARANS 分群演算法於分兩群所耗費的時間明顯的多於本研究所提出的改良方法，而且本研究的改良方法並不會讓準確度下降，亦可得到較佳的分群結果。

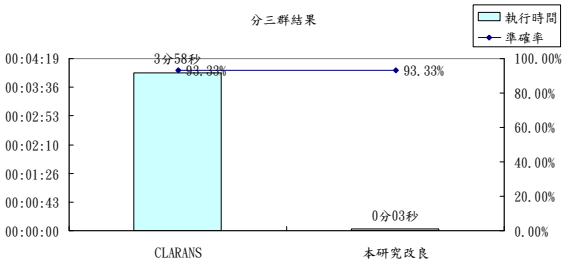

圖 6 分三群結果

當資料分愈多群則原始 CLARANS 分群演算法速度則會愈慢，可從圖 6 看出本研究提出改良方法可以有效的加速分群演算法，而且原始 CLARANS 分群演算法及本研究改良方法皆是 93.33%的準確度，其中錯誤部份為四個維度與原始樹狀結構圖的分群不同，所以本研究提出的改良方法在分三群的時候，速度一樣可以有效改善，而且其準確度與原始 CLARANS 的準確度都很高。

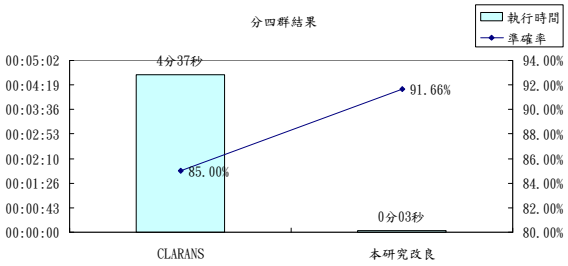

圖 7 分四群結果

由圖 7 可見，當資料分群數量為 4 的時

候，CLARANS 分群演算法的執行時間已經超過 4 分鐘，而本研究所提出的改良方法從分兩群開始，皆只需要三秒鐘，而且分群結果亦與樹狀結構非常吻合。

6. 結論

由實驗結果可以看出本研究所提出兩種方法一起使用可有效的加速 CLARANS 演算法，而分群結果也可與原始的 CLARANS 分群演算法媲美，並可以提升整體分群演算法的速度及效能，不過在分群數愈多的時候，準確率稍微下降，期望在未來能夠對這部份繼續探討，期待能有更好的準確度。

參考文獻 (References)

[1] D. Delen , G. Walker , A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artificial Intelligence in Medicine , Volume 34 , Issue 2 , pg 113 – 127.

[2] Jen-Tsan Chi , Edwin H. Rodriguez , Zhen Wang , Dimitry S. A. Nuyten , Sayan Mukherjee , Matt van de Rijn , Marc J. van de Vijver , Trevor Hastie , Patrick O. Brown, “Gene Expression Programs of Human Smooth Muscle Cells : Tissue-Specific Differentiation and Prognostic Significance in Breast Cancers,” PLoS Genetics, September 2007 , Volume 3 , Issue 9 , e164.

[3] Kaufman, L., and P. J. Rousseeuw. Finding Groups in Data: at Introduction to Cluster Analysis. John Wiley & Son Inc. 1990

[4] Martin Buess, Dimitry SA Nuyten, Trevor Hastie, Torsten Nielsen, Robert Pesich and Patrick O Brown, “Characterization of heterotypic interaction effects in vitro to deconvolute global gene expression profiles in cancer ,” Genome Biology 2007.

[5] M.H. Dunham, Data Ming: Introductory and Advanced Topics, Prentice Hall, New Jersey, 2003.

[6] Ng, Raymond T. and Jiawei Han. 1994. Effcient and Effective Clustering Methods for Spatial Data Mining. Proceedings of the 20th Very Large Databases Conference(VLDB 98) Santiago, Chile.

[7] Stanford Microarray Database, http://smd.stanford.edu/.

[8] T.S. Chen, C.C. Lin, Y.H. Chiu and R.C. Chen “Combined Density- and Constraint-based Algorithm for Clustering,” In Proceedings of 2006 International Conference on Intelligent Systems and Knowledge Engineering, 2006.

[9] T.S. Chen, R.C Chen., C.C. Lin, T.H. Tsai, S.Y. Li, X. Liang, 2005, Classification of Microarray Gene Expression Data Using a New Binary Support Vector System. “IEEE International Conference on Neural Networks and Brain” (ICNN&B) ,pg 485-489.

[10] Vapnik, V.N., 1995, The Nature of Statistical Learning Theory, Springer.