利用主成份分析改進 K-medoid 分群演算法不穩定性之研究-以 Microarray 為例

陳同孝	陳民枝	陳秉豐	陳泊郡	陳則翔	陳昕謨
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
tschen@gmail.com	jeanne@ntit.edu.tw	rxwayen@gmail.com	jc7003@hotmail.com	q204227@hotmail.com	loveliw520@yahoo.com.tw

摘要(Chinese Abstract)

分群演算法中的 K-medoid 分群演算法雖然具有分群方法簡單、執行快速等優點，然而不穩定分群結果卻是其最為人垢病的缺點。本論文所提出的 KPMCA 是以 K-medoid 分群演算法為基礎，結合主成份分析的特性，達到改善分群結果不穩定的缺點。在我們的實驗中，以生物晶片為分群資料來源，在經由多次分群實驗後，KPMCA 擁有高達 95.2%的平均分群正確率，大幅改善 K-medoid 分群演算法不穩定的缺點。KPMCA 具有穩定可靠的分群結果、分群計算處理快速，無需設定額外參數等優點，因此將 KPMCA 其應用於生物資訊領域中，相信對於基因體的研究能有很大的幫助。

關鍵詞：資料探勘、分群演算法、K-medoid、生物晶片

Abstract

K-medoid clustering algorithm is a simple and quick implementation, but the unstable clustering result is a serious drawback. In this paper, K-PCA-medoid clustering algorithm (KPMCA) combines the conception of the K-medoid clustering algorithm and principal component analysis to slave the problem. In our experimental result, KPMCA has as much as 95.2 percent average rate in the correct grouping results, it can effectively improve the drawback of K-medoid clustering algorithm, and our method doesn't need to add any new parameter. So, we believe that using KPMCA in bioinformatics can be very helpful for research of gene.

Keywords: **Data Mining, Clustering Algorithms, K-medoid, Microarray**

1. 前言 (Introduction)

自從人類基因體計劃(Human Genome Project, HGP) [13]的完成，人類對於基因體的研究有了更深一層的了解。然而目前所完成的工作僅限於對人類四十六條染色體定序(sequencing)完畢，但定序只是提供讓人一窺基因全貌的機會。人類估計約有十萬個可表現的基因，基因又可分為結構基因、調節基因和操縱基因三種，因此想要瞭解基因、應用基因與了解其功能及作用，變成了當成最熱門的研究領域。

因應人類基因組計劃完成後帶來的研究熱潮，吸引愈來愈多的專家學者投入此研究領域，使著相關的基因功能分析技術便應運而生，也帶動了全球生物晶片的蓬勃發展。幾項新發展的生物晶片技術，如生物晶片(microarray)[12]、基因晶片(DNA chip)及實驗室晶片(Lab-on-a-Chip)等，使我們能更快地瞭解這些基因的功能及其彼此間的交互作用 。

其中，生物晶片(Microarray) [12]是做為大量篩檢及平行分析基因表現的工具。生物晶片具備快速、方便、經濟、省時等特性，適用於大量基因表達、篩檢及比對等研究。因此在病原體基因檢測、基因表現比較、基因突變分析、基因序列分析、及新藥物開發等研究領域中，都是常用的生物晶片技術之一。

在這股研究的潮流中，獲得了大量的原始資料，而這些原始資料是需要儲存、處理和分

析等處理後才會具有意義的。以人力去處理和分析這些大量、無一定規則的原始資料，是一件費時費力的工作，因此，如何運用資訊處理技術將這些資料轉換成有價值的資訊，造就了另一門的生物資訊學(Bioinformatics)的興起。生物資訊學是一門結合生物學、計算機科學及資訊科技所形成的新研究領域，生物資訊學強調整合、分析資料庫中的生物資訊，尋找致病基因，預測基因的功能。其中利用資料探勘[8]中的分群演算法[1][10]來協助大量基因資料進行分析與統計，已逐漸形成重要的基因分析應用工具。

分群演算法的用途在於不需要事先知道資料該分成幾個已知的類型，而可以依照資料間彼此的相關程度，將資料分成幾個相異性最大的群組(Cluster)，而各群組內的資料彼此間相似性高，透過這樣的特性。完成資料分群的目的。因此可以利用分群演算法來協助生物晶片進行分群工作，找出基因間的相關性。

現有的分群演算法眾多，諸如 Hierarchical 分群演算法[2][7]，K-means 分群演算法[3][5][10]，K-medoid 分群演算法[6]等，但這些方法皆不盡完美，會有需指定分群數目、分群結果的正確率與分群執行速度緩慢等問題，因此在分群演算法的研究中，如何有效改善上述問題仍然是個值得討論與研究的問題。

K-medoid 分群演算法屬於分群演算法中的分割演算法，它以集群內最接近中心位置的物件為集群的中心點，每一回合都只針對扣除作為集群中心物件外的所有剩餘物件，重新尋找最近似的集群中心。K-medoid 分群演算法與其它分群演算法相比，K-medoid 分群演算法的優點是方法簡單、運算時間短、不易受離群或雜訊資料的影響，但其不穩定的分群結果卻是它最大的缺點。

因此，本研究將以生物晶片資料做為分群資料，採用 K-medoid 分群演算法為分群演算法基礎，希望能提出一個有效改善分群正確率的方法，藉此提升分群結果的穩定性，減少每次分群結果的差異度，提供一個可信賴的分群結果。

後續的章節將會先進行文獻探討，接著介紹本論文所提出的 K-PCA-Medoid Clustering Algorithm (KPMCA)與將其應用於生物晶片資料分群的實驗結果，最後是本論文的結論。

2. 文獻探討

2.1 K-medoid 分群演算法

K-medoid 分群演算法[6]基本上和 K-means 分群演算法[3][5][10]是類似的，不同在於 K-means 分群演算法是以集群內各物件的平均值為集群的中心點，而 K-medoid 分群演算法是以集群內最接近中心位置的物件為集群的中心點，每一回合都只針對扣除作為集群中心物件外的所有剩餘物件，重新尋找最近似的集群中心。

接下來詳細說明 K-medoid 分群演算法的步驟：

一、輸入要分 k 群(n>k)；n=資料物件個數。

二、決定分群數後，隨機取出 k 個 medoid 資料 。

三、計算出剩餘非 medoid 資料與各 medoid 資料的距離，將其與距離最近的 medoid 劃分於同一群內 。

四、計算 medoid 資料與群內其他非 medoid 資料的距離總和 D_1。再從群內，隨機找出一個非 medoid 資料，並算出該非 medoid 與群內其他資料的距離總和 D_2 。

五、如果 $D_2<D_1$，則取代原先的 medoid 資料成為新的 medoid 資料，再將其與距離最近的非 medoid 資料劃分為同一群。

六、重複步驟四~六，直到滿足終止條件 。

K-medoid 分群演算法與常用的 K-means 分群演算法比較後可以發現，K-medoid 分群演算法具有下列優點：運算時間短、分群所需時間短、分群方法簡單、不易受離群或雜訊資料的影響、用實際的資料來表示群集的中心點、仍可適用於無法以數值表示的資料。

正因為 K-medoid 分群演算法簡單快速及不易受離群或雜訊資料的影響的特性，仍吸引不少學者以它為改進的基礎，提出如 PAM [3]、CLARA [4]、 CLARANS[5]等方法，藉此改進 K-medoid 分群演算法需先指定 k 的數目與因隨機選擇的方式，造成分群結果不穩定等缺點。

2.2 主成份分析

主成份分析[11]是一種維度簡化的技術，假

設資料中有 p 個變數，主成份分析可以只用 $m(m<p)$個主成份來描述原始資料，達到減少資料維度的目的。主成份分析是一個線性轉換。這個轉換可以將資料變換到一個新的坐標系統中，得到轉換後的主成份，其中第一主成分佔總變異量最大的比例，是所有主成份中最能表現出原始資料的特性。而第二主成份解釋第一主成分未能解釋的總變異量，之後依次類推。但這 p 個主成份間彼此並不相關。

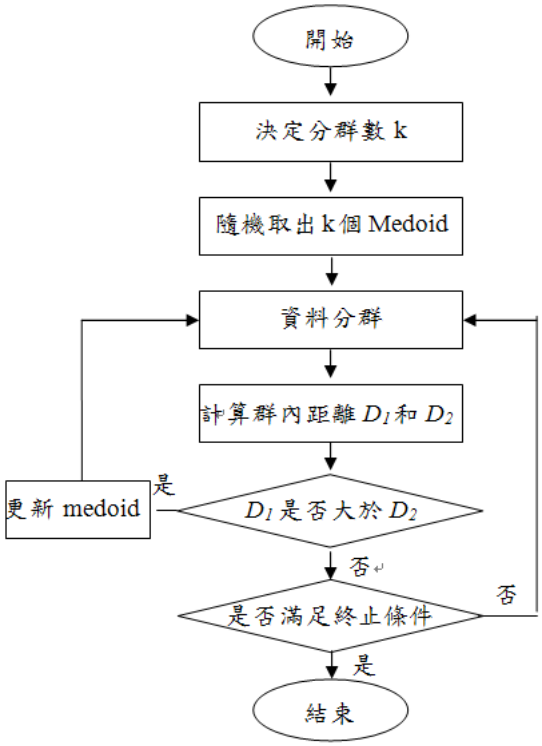

圖 1 K-medoid 分群演算法流程圖

舉例來說，假設有 2 個特徵值 x_1 與 x_2 的 10 筆資料，如圖 2 所示。將 x_1 與 x_2 代入有一線性方程式 x_1*，會使原始資料投影至一個新的座標軸上，由圖中可以發現原來散佈於四處的資料點經由投影轉換後，將會座落於圖 2 中直線 x_1*上，而各筆資料代入線性方程式 x_1*所得到的值，我們稱之為投影長。在我們的方法中，將利用投影長做為我們選擇初始 medoid 的依據，以達到選出較佳的初始 medoid 的目的。

3. K-PCA-Medoid Clustering Algorithm (KPMCA)

從前面的文獻探討中，我們可以發現 K-medoid 分群演算法在進行分群的過程中，在初始 medoid 的選擇與各回合中選出比較的資料的步驟中，皆是使用亂數來決定。初始 medoid 的選擇是透過亂數的方式來決定的結果，造成容易受到所選擇的初始 medoid 位置好壞，造成後續分群的結果不一定會朝好的方向前進，每次分群的結果不穩定或是需要執行更多的回合數才能達到終止條件等缺點。因此，如何找出一個好的初始 medoid，對於 K-medoid 分群演算法的分群結果來說，是一項重要的關鍵影響因子。

本論文所提出的 K-PCA-Medoid Clustering Algorithm (KPMCA)，主要是針對 K-medoid 分群演算法的分群步驟中的決定初始 medoid 的部分進行改進，透過結合主成份分析的概念，來找出適當的初始 medoid，改善原始 K-medoid 分群易受到亂數選擇初始 medoid 的位置，而影響結果的因素，進而達到提升分群結果的穩定度，加速達到終止條件所需耗費的時間的目的。

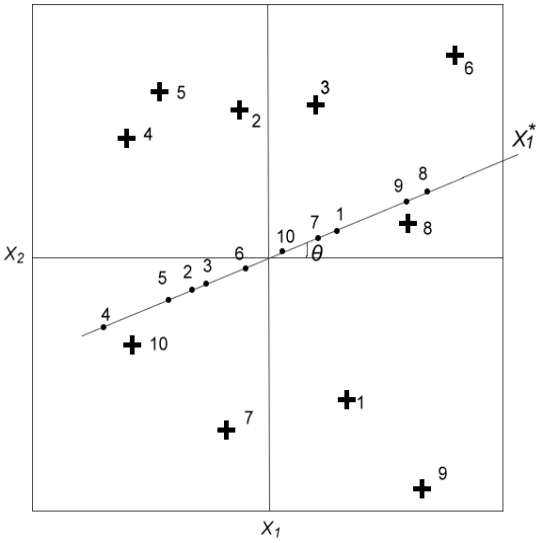

圖 2 PCA 示意圖

KPMCA 與原始 K-medoid 分群演算法分群流程大致相同，不同的部份在於以 PCA 決定初始 medoid，取代原始 K-medoid 分群演算法以隨機的方式決定，接下來將說明本論文如何透過結合主成份分析的概念，來找出適當的初始 medoid。詳細步驟說明如下：

一、利用主成份分析資料，取得各資料第一主成分的投影長。

在此步驟中，我們將所得到的微陣列資料用主成份分析得到該資料的第一主成份，因為第一主成份是變異量較大的主要成份，較其它成份更能表示原始資料的特性，因此我們只採用第一主成份做本論文

尋找初始 medoid 的依據。因些假設有 n 筆資料，經由主成份分析後，我們可以得到原始資料的所有主成份，其中 pca_i 表示第 i 筆資料的第一主成份投影長。

二、 將 pca_i 依大小排序得到一維陣列 A。
在得到經由主成份分析後所得到的各筆資料第一主成份投影長後，我們將依投影長大小依序排序原始資料，得到一個一維陣列 A。

三、 依分群數 k，將陣列 A 分成 k 個區塊，找出分隔各區塊的點 p_S。
根據 K-medoid 分群演算法中所設定要的分群數 k，將一維陣列 A 分成 k 個區塊，找出分隔各區塊的點 p_S 與其對應的原始資料，其中 k 個區塊就需要 k+1 個點 p_s (p_s, s=0,1,2,..., k)來區隔。

四、 各區塊的中間點 $middle_s$ 所對應的原始資料成為初始回合的 medoid。
在找出各區塊的分隔點後，我們要找出各區塊的中間點做為 K-medoid 分群演算法初始回合的 medoid。假設現在要找出第 s 個區塊的中間點 $middle_s$ 所對應的原始資料，其計算方式如(1)。

$$middle_s = p_{s-1} + (p_s - p_{s-1})/2 \qquad (1)$$

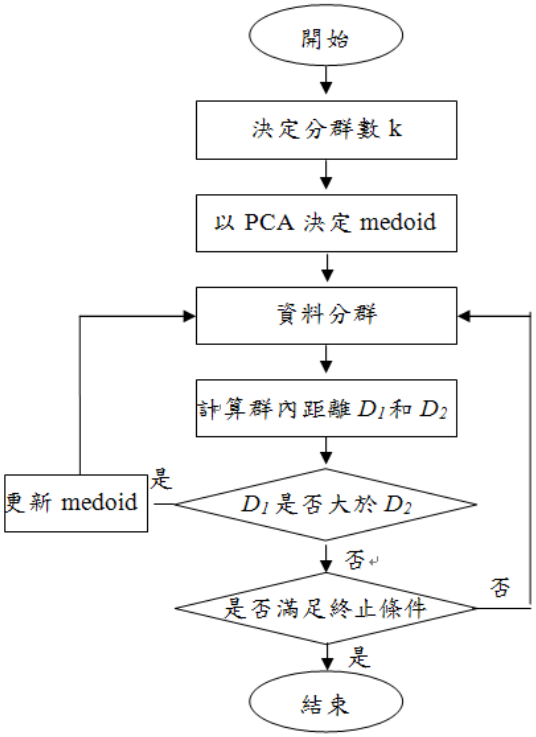

圖 3 KPMCA 流程圖

4. 實驗結果

在本論文中，我們使用的實驗資料是來自 Research Genetics (Huntsville AB, USA) (http://www.resgen.com/) 所提供的微陣列資料，內容是 5 個成人腎臟的基因表現。腎臟是一個複雜的器官，在腎臟中，各種不同的功能以固定的方式被分開，如腎元的細胞或相關的結構。資料庫總共有 1548 基因，如圖 4 所示，共有 33 個維度將其分為 5 個群集。經由 Hierarchical 分群演算法的分群處理後，分群結果如圖四所示，其中 A 群集為腎小球(Sieved glomeruli)，B 群集為皮層(cortex)，C 群集為髓質(medulla) ，D 群集為鐘乳石狀(papillary tips) ，E 群集為腎盂(renal pelvis)。

本論文將使用此成人腎臟微陣列資料為資料，來實驗比較原始 K-medoid 分群演算法與 KPMCA 兩者的分群結果正確率與分群結果穩定性。

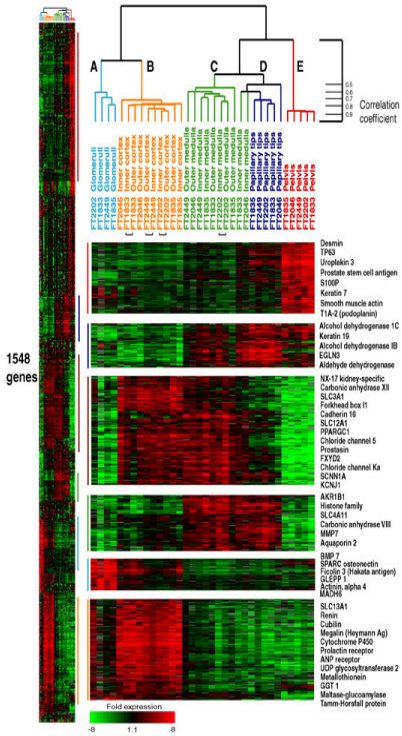

圖 4 微生物晶片資料庫

為了公平比較原始 K-medoid 分群演算法與 KPMCA 的分群結果，我們將各做 50 次分群實驗來測試分群結果正確率。分群結果正確率計算式如(2)所示，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。

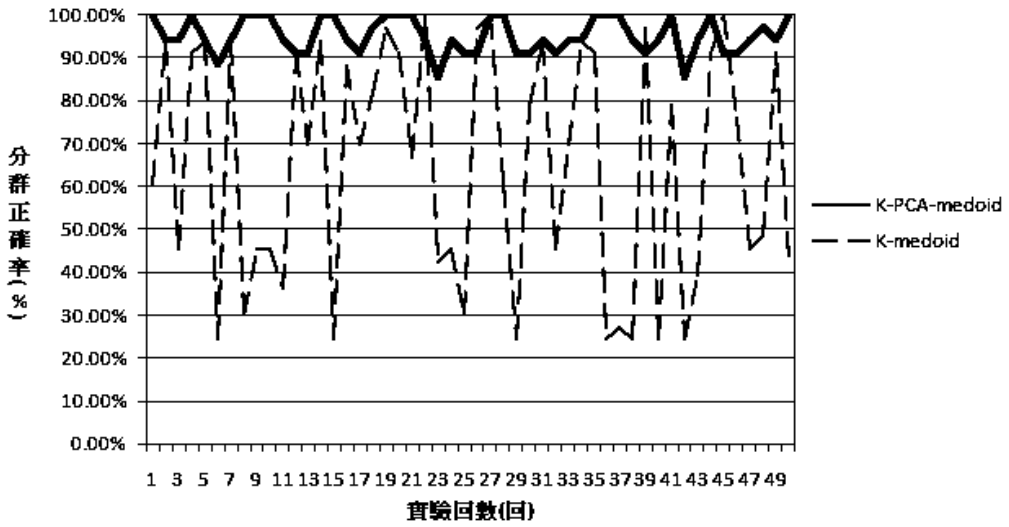

圖 5 分群結果正確率

$$A = \frac{r}{n} \tag{2}$$

從圖 5 中，我們可以明顯發現到原始 K-medoid 分群演算法在這 50 次的分群結果正確率相當的不穩定，最好的情況下，可以將資料全部分至正確群集中，分群正確率可達到百分之百；但也有發生多次分群正確率不到百分之三十的情況，無法將資料正確的分群。會發生這種情況的原因主要是受到原始 K-medoid 分群演算法的初始 medoid 的選擇與各回合中選出比較的資料的步驟中，皆是使用亂數來決定的影響，所以在執行多次分群實驗時，會有不同分群結果的情況產生，因而造成分群結果的不穩定。

相對來說，在從圖 5 中，我們可以明顯發現到本論文所提出的 KPMCA 在這 50 次的分群結果正確率相當高，即使是最差的情況下，仍可維持百分之八十八以上的正確率。從表 1 中，也可輕易比較出在這 50 次的分群實驗中，比起原始 K-medoid 分群演算法的分群結果正確率大幅度的上下跳動，平均分群正確率只有 64.7%，KPMCA 的分群結果正確率是相當穩定的，50 次的分群實驗中的平均分群正確率高達 95.2%，由此可見 KPMCA 每回合的分群結果是穩定且可信賴的。

表 1 平均分群正確率

	KPMCA	K-Medoid
平均分群正確率	95.2%	64.7%

5. 結論

從本論文的實驗結果中，可以明顯發現到本論文所提出的 KPMCA，在加入利用主成份分析決定初始 medoid 後，其分群結果較原始 K-medoid 分群演算法來的穩定，在 50 次的實驗中，KPMCA 的平均分群正確率仍可維持在 95.2%，遠勝於原始 K-medoid 分群演算法的 64.7%，可以有效改善原始 K-medoid 分群演算法，因為透過亂數的方式來決定初始 medoid，導致容易受到初始 medoid 位置好壞，造成對後續分群結果有很大的影響。而本論文的 KPMCA 相較於其它以原始 K-medoid 分群演算法為基礎的方法，並不需要額外設置與調整參數，仍維持與原始 K-medoid 分群演算法一樣，只需輸入所要的分群數目即可，可以大幅減少測試與調整參數的時間，盡速完成分群的工作。KPMCA 具有穩定可靠的分群結果、分群計算處理快速，無需設定額外參數等優點，相信將其應用於生物資訊域領中，可以有效快速地完成基因表現比較、基因突變分析、基因序列分析等工作，加快人類對於基因體的研究能有了更深一層的了解。

參考文獻 (References)

[1] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: A survey," ***ACM comput.Surv.***, Vol.31, pp. 264-323,1999.

[2] J. Grabmeier, A. Rudolph, "Techniques of Cluster Algorithms in Data Mining," ***Data***

Mining and Knowledge Discovery journal**,** Vol.6(4), pp. 303-360, 2002.

[3] M. Ester, H. -P. Kriegel, X. Xu, "Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification," ***In Proc. 4th Int. Symp. Large Spatial Databases (SSD'95)***, pp. 67-82, August 1995.

[4] P.W. Huang, P.L. Lin, H.Y. Lin, "Optimizing storage utilization in R-tree dynamic index structure for spatial databases," ***The Journal of Systems and Software*****,** Vol.55(3), pp. 291-299, 2001.

[5] R. Ng, J. Han, "Efficient and effective clustering method for spatial data mining," ***In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94)***, pp. 144-155, September 1994.

[6] S. Basumallick, J. S. K. Wong, "Design and implementation of a distributed database system," ***Journal of System Software,*** Vol.34(4), pp. 21-29 ,1996.

[7] S. Guha, R. Rastogi, K. Shim, "Cure: An efficient clustering algorithm for large databases," ***In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)***, pp. 73-84, June 1998.

[8] A. Berson, S. J. Smith, *Data Warehousing, **Data Mining, and OLAP***, McGraw-Hill, 1997.

[9] J. Han, M. Kamber, ***Data Mining: Concepts and Techniques, Academic Press***, 2001.

[10] L. Kaufman, P. J. Rousseeuw, ***Finding Groups in Data: An Introduction to Cluster Analysis***, John Wiley & Sons, 1990.

[11] Nunnally, J. C., and Bernstein, I. H., ***Psychometric theory***, McGraw-Hill, 1994.

[12] DNA microarray, http://en.wikipedia.org/wiki/DNA_microarray.

[13] Human Genome Project, http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.