挑選具有非線性關聯性之基因選取機制: 以演化式計算模式搭配多目標決策作法

唐瑋君　　蔡毓舜　　鍾翊方
國立陽明大學生物醫學資訊所
g30123003@ym.edu.tw　g39528010@ym.edu.tw　ifchung@ym.edu.tw

摘要

近年來微陣列基因晶片因為能夠檢測基因的表現資料，使其可用於觀察不同組織之間的基因表現狀況或因基因表現改變造成的疾病等而被廣泛使用[1][2]，但因為人類有三萬多個基因，基因之間組合的情況會非常多，因此，從大量的資料中找尋有用的基因用於診斷系統的建立目前是一個大家感興趣的研究議題。由於基因數目很多，如何找出可用的基因組有多種作法，而一般的基因挑選機制通常選出的基因是基因與類別之間關係強烈且希望挑出來的基因數量是較少的，雖然選出的基因有具有線性關係與非線性關係，不過大多數都針對基具有線性關係的基因作探究，因為非線性交互作用的基因間的關係並無固定的規則，在探究基因的生物意義時，往往有較大的挑戰；我們希望能夠將演化式計算模式中結合多目標問題的方法運用在基因挑選機制上面，透過原有的 mRMR 基因挑選機制的多目標定義作延伸，將基因之間不同的問題演變成不同的目標去找出具有非線性關聯性的基因，因此，我們提出以演化式計算模型搭配多目標決策作法(Feature Selection: Evolutionary Computing with Multiobjective Strategy，簡稱 FS-ECMOS)來針對微陣列基因晶片做非線性關聯性之基因的挑選，除了考慮基因對於類別的區分能力之外，我們還考慮基因之間的非線性關聯性以及所選用的基因數目，希望藉由演化式計算的概念，搜尋所有的解空間中符合多目標條件的基因，使我們能夠保留較多具有非線性關聯性的基因，幫助我們更加了解這些具有非線性關係的基因之間的生物意義。

關鍵詞：相互資訊值、特徵選取、微陣列基因晶片、多目標決策作法、基因選取。.

1. 前言

微陣列基因晶片(Microarray)近幾年被廣泛用來檢測基因的表現狀況，常用與觀察生物體內不同組織的基因表現狀況或者相同組織之間因基因表現的改變造成的疾病等。由於微陣列基因晶片的資料提供豐富的基因表現資料，使得這些資料量都非常龐大，因此，在大量的基因表現資料當中，挑選適當數量的基因用於建立模型或者當作診斷的生物標記是個很重要的一個議題。

微陣列基因晶片資料的基因挑選方式，可以概分為三種[3]：(1) Filter：此種方法主要先獨立執行基因的篩選，基於挑選機制(統計檢定、皮爾森相關係數或其他做法等為常用於挑選基因的方法)提供的評估數據，利用排序方式，選出排名前面的基因當作較佳且具有類別區分能力的，最後將這些基因丟入分類器，因為此種方法只有使用到評分機制，與分類器的選用並無關聯，優點是不需要考慮太多的時間複雜度，可以使得計算簡單快速，不過因為這種方法只有對基因跟類別之間的關係作考慮，並無考慮到基因與基因之間的關聯性，所以這是缺點的一部分。(2) Wrapper：這種方法會使用機器學習的技巧，根據學習演算法的效能指標評估學習結果，將能夠提高效能指標數值的基因納入考慮，藉此達到提高效能選取基因的方法，此種方法的優點有考慮到基因與基因之間的關聯性與分類器造成的影響，缺點是因為利用機器學習的方法考量後再使用分類器作評估，花較多的時間。(3) Embedded：此作法會透過用分類器中嵌入的參數來考慮挑選出來的基因是否能夠提高選取基因的效能，Embedded 作法與 Wrapper 作法類似，差別在於兩種作法之間的挑選基準不一樣。

一般基因挑選作法會希望找尋基因與類別之間關係強烈並且希望將挑選的基因數量減少，我們希望將這些不同問題變成不同的目標，與多目標問題作法進行結合。演化式計算法為 wrapper 方法其中之一，此作法在做基因特徵選取方法時，只有解決單目標的問題，並無考慮到多目標問題的結合，我們透過文件探勘，發現在演化式計算模式中結合多目標問題較有名的方法有 NSGA-II [4] 與 MOEA/D [5][6]，而在基因挑選機制上，因 mRMR 作法[7]有定義多目標問題的機制，我們希望透過此作法的多目標定義進行延伸，因此我們主要針對 NSGA-II、MOEA/D 兩種方法與基因挑選機制(mRMR)作探究。

我們發現這些方法一般都是以基因之間具有高線性關係的做保留，而且並沒有強調挑選出來的基因是否全都是具有線性關係的基因，因此我們設定不同的機制探討基因與基因之間是否具有非線性關係，除了希望能夠使用我們的多目標決策作法，透過演化式計算模式針對基因做演化，從微陣列基因晶片資料中找出多組基因解，使這些基因對於類

別有很高的相關性、基因與基因之間也具有非線性(Non-linear interaction relationship)的關係之外，也希望挑出來的基因皆是非線性關係基因的組合，最後讓我們可以針對這些非線性關聯性的基因去做分析，了解非線性關聯性的基因對於分類的能力。

2. 基因挑選機制

在運用基因演算法在做基因挑選機制時，用來探究基因與類別之間的關係與基因之間的關係的方法有兩種：(1) 皮爾森相關係數(Pearson's correlation coefficient) (2) 相互資訊值(Mutual information)；一般皮爾森相關係數方法主要是測量兩連續變數間關係的強弱(計算線性關係情況)，針對有特殊圖形的非線性關係基因是算不出來的，因此我採用相互資訊值(Mutual information)方法來做基因之間的關聯性運算。相互資訊值是利用熵(Entropy)與機率理論(Probability)來測量兩變數之間關聯性的一個方法。

相互資訊值作法：

$$E(X) = -\sum_{i=1}^{n} P(X_i) log_2 P(X_i) \quad (1)$$

首先會先分別計算兩個變數的各自的熵，計算方法如方程式(1)，運用機率概念做運算。

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(X_i,Y_j) log_2 [P(X_i,Y_j)] \quad (2)$$

再計算兩變數聯合在一起後的關聯情況，將兩變數合併在一起計算聯合熵(Joint entropy) H(X,Y)。

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (3)$$

最後透過將兩個變數各自的熵相加後相減聯合熵來得到我們想知道的兩變數之間的相互資訊值 I(X；Y)。

在基因選取的方法中，使用相互資訊值作為搭配比較有名的有“minimum redundancy maximum relevance”(簡稱 mRMR) [7]，此方法訂定兩個目標，目的希望能夠找出基因與類別之間關聯性高的(稱為 Relevance，簡稱 Rel)及基因與基因之間關聯性低的情況(稱為 Redundancy，簡稱 Red)。

目標一：基因與類別之間的關聯性

$$Rel_I = \frac{1}{|S|}\sum_{g\in S} I(c,g) \quad (4)$$

為了了解一組基因與類別之間的關聯性(c 為類別，S 為一組基因)，目標一計算基因與類別之間的相互資訊值，並採用平均方法來看一組基因與類別之間的關聯性，希望關聯性越大越好。

目標二：基因與基因之間關聯性

$$Red_I = \frac{1}{|S|^2}\sum_{g_i,g_j\in S} I(g_i,g_j) \quad (5)$$

由於有些基因與基因之間關係呈現高度關聯性，而我們其實只要使用其中一個基因來做代表用於分類也會得到相近的預測效果，不需用整個基因組合的基因作考量，因此目標二的目的是希望基因與基因之間關聯性越小越好，除了避免只使用部分基因當作代表外，也能避免彼此造成影響，所以在目標二中對於基因與基因之間關聯性的測量，同樣是使用相互資訊值作法來測量基因與基因之間的關係(g_i, g_j 皆為一組基因，兩組基因不相同)，不過會希望基因與基因間關聯性越小越好。

由於 mRMR 基因選取作法想找出高的 relevance 且低的 redundancy 基因，所以 mRMR 基因選取的策略為 Mutual information difference(簡稱 MID)，此方法採用前進選擇法(Forward selection)的作法做基因選取：

$$max\,(Rel_I - Red_I) \quad (6)$$

首先會先挑基因與類別之間關聯大的基因進來，接著第二個要挑的基因，會去看是什麼樣的基因進來，會使得整個組合的 Rel_I 很大，Red_I 很小，運用 MID 方法找出能使值最大的基因，當作第二個被選入的基因，第三個被選進來的基因會先考慮第三個加進來後對於原本 Rel_I 與 Red_I 的影響。

雖然 mRMR 是解決兩目標的問題，但此種作法若將解最大化或最小化，最終其實只會得到一個解，在解多重目標的作法中，是不可能找到一個解可以同時最佳化這些目標，反而是找尋那些解可以同時考慮各種目標，並在目標之間求得合理的情況(這些合理情況的解稱為 Pareto set，如圖 1 所示)。

由於我設定的基因挑選作法需同時考慮多重目標，而 mRMR 作法雖有使用到相互資訊值作為基因挑選機制，不過並沒有實際應用在解多目標問題，且此作法也無提到挑選出的基因有多少數量是與類別具有高度相關性，因此 mRMR 方法並不適用在解多目標的問題。後來，我們發現有演化式演算法(Evolutionary algorithm)可以用來解決多目標問題與求取 Pareto set。

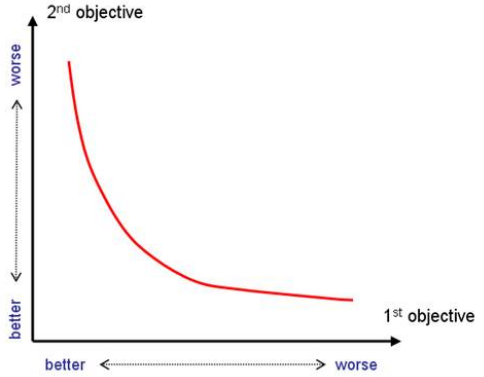

圖 1：Pareto Set 示意圖

3. 多目標演化式演算法

演化式演算法是以達爾文物競天擇的概念發展出來的全域搜尋演算法且被用於解決不同的最佳化問題，此作法的運作機制首先會透過隨機的方式產生初始解，透過適應函數(Fitness function)給予各個解適應各種不同目標的適應值(Fitness score)用以評估解的好壞，並透過突變(Mutation)及交配(Cross-over)來產生新的解，最後用適者生存不適者淘汰的機制來演化，最終目的是希望漸漸找到一個全域最佳解來解決最佳化問題。

演化式演算法用來解多目標問題的方法中較有名兩種方法為 Non-dominated Sorting Genetic Algorithm (簡稱 NSGA-II) [4]與 Multiobjective Evolutionary Algorithm Based on Decomposition (簡稱 MOEA/D) [5][6]。

NSGA-II 方法主要利用隨機產生出來的群體去計算出各自的適應值，根據這些適應值，把這些結果點在空間中，針對非支配解情況(Non-dominated)去作排序，由圖 2 來做說明，圖中同一顏色的解中，我們並無法說明哪個解可以支配誰，根據不同的目標定義，各自有各自好的部分，因此我們同一顏色都當作同等重要，而根據這樣畫出來的結果，我們最終可以得到多組的 pareto set。

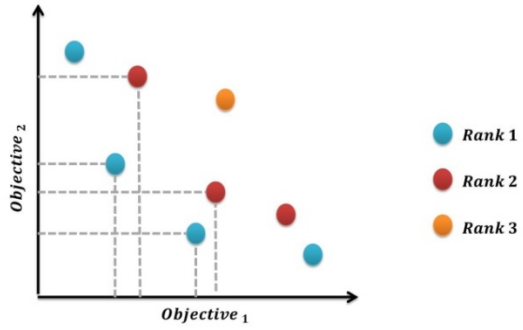

圖 2：NSGA-II 示意圖

MOEA/D 方法也是利用隨機產生出來的群體去進行演化，將隨機群體的資料分成 n 等分，並給予每一等分一個權重值(Weight)，透過權重與鄰居關係的概念把 n 等分的資料化成 n 個單一子目標去進行求解；此作法計算每個解之間的距離，找出每個解的鄰居，如圖 3(a)所示，我們會找尋每一個解附近的鄰居。接著我們會透過適應函數給予這些解適應各種不同目標的適應值，讓每個解與周圍鄰近的鄰居能利用權重和適應值相乘的結果作比較，將好的解取代不好的解並變成下一代繼續演化下去，圖 3(b)為給予隨機群體中每個子目標權重值的示意圖，此範例設定兩個目標讓解進行演化並給予 n 個子目標權重值(不同目標之權重值相加總合為 1)，因此每個子目標有各自目標的權重值與適應值。而每次在進行演化時，還會給予一個參考點(Reference point)用來與目前空間中的解做比較，希望最後解能夠像 pareto graph 一樣往最佳化目標的方向前進，圖 3(c)為參考點的設定方式，圖中的解有各自符合不同目標的適應值，從圖中的 A 點來看，此解在目標一有最佳狀況，B 點在目標二有最佳狀況，因此我們會取 A 點與 B 點各自的適應值做組合，組成此代演化的參考點，供所有解與此參考點做比較並演化。當做完所有演化世代(Iteration)後，最後一代的解就是每個子目標的最佳化解，如圖 3(d)所示，由於此範例最小化兩個目標，因此解的走向會往左下角移動；藍色的點為一開始隨機選取的初始解，紅色點則為經過設定目標演化過後產生的符合每個子目標的最佳化解。所以 MOEA/D 的作法，我們可以想像成是 NSGA-II 方法中最終解集合裡的那一組解(如圖 2 中藍色點)。

由於微陣列晶片資料上面擁有數萬個基因的資訊，當同時將這些資料使用演化式演算法進行演化時，會因為搜尋偌大的解空間且有多種的解的情況組合，使得執行時間非常可怕，而 MOEA/D 是把多維度的資料化簡成單一子目標問題針對解空間作搜尋求解，所以此方法需要較少的執行時間且能夠找到較佳的解，因此我們改寫 MOEA/D 的流程，使其能夠適用於基因選取作法。

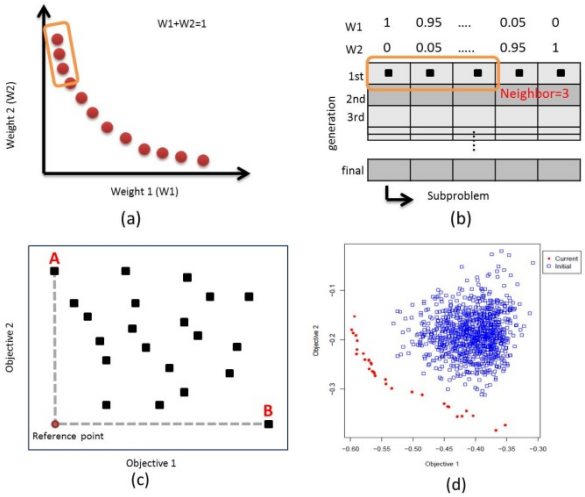

圖 3：MOEA/D 作法示意圖

4. 基因選取機制(FS-ECMOS)

為了將 MOEA/D 作法應用在基因選取機制上面，我們提出以演化式計算模型搭配多目標決策作法(Feature Selection: Evolutionary Computing with Multiobjective Strategy，簡稱 FS-ECMOS)，將每個解定義為與資料中基因數相同長度的二元染色體，以二元染色體編碼的方式來進行演化，當染色體的值為 1 時，代表我們所要挑選的基因，反之，若染色體的值為 0 時，代表我們不挑選的基因。

由於我們目標是希望從微陣列基因晶片資料中找出一組基因，這些基因對於類別有很高的相關性之外，基因與基因之間也具有非線性的關聯性，所以我們訂定以下三個目標：

目標一：一個基因與類別之間的關聯性

$$Rel(g,c) = \left\{\frac{I(g,c)}{H(g)+H(c)}\right\} \times 2 \; ; \; 0 \le Relevance \le 1 \quad (7)$$

對於第一個目標：一個基因與類別之間的關聯性(簡稱 Relevance)，我們採用基因與類別之間的相互資訊值來表示，由於需將資料化成離散的方式做運算，所以當資料型態為連續型資料，我們會先將資料數值轉換至 0 至 1 區間後，再區分為 10 等分的區間，計算每個區間中包含有多少個樣本後再轉換為機率值，估算基因與類別各自的熵(H(g)與 H(c))，並藉此計算基因與類別之間的相互資訊值 I(g, c)，最後用此相互資訊值用以代表基因與類別之間的關聯性(Rel(g, c))。

目標二：一對基因之間非線性的關聯性

$$\mathrm{Dep}(g_i, g_j) = \left\{1 - abs\left(Cor(g_i, g_j)\right)\right\} \times \left\{\frac{I(g_i, g_j)}{H(g_i)+H(g_j)}\right\} \times 2 ; 0 \le Dependency \le 1 \quad (8)$$

針對基因之間的關聯性(簡稱 Dependency)，根據相互資訊值的定義，若兩個基因之間有很大的相互資訊值，那麼就表示兩基因之間具有高度關聯性，同理，若兩個基因之間具有強烈的線性關係，那也表示兩基因之間具有很大的相互資訊值與存在著一定的關係，這樣的一組基因對我們來說是比較不具有參考價值的，因為性質相同的基因，其實只需要拿其中一方的當作代表就好，若同時接拿來參考，有時候造成的結果並不一定比較好，反而會覺得這些基因多餘了，此目標與 mRMR 目標二極大的不同點的地方在於我們是希望找出具有非線性關聯性的基因，因此 dependency 越大越好，而 mRMR 目標二是希望找出基因與基因之間關聯性低的情況，因此 redundancy 越小越好。

因為我們的目標特別著重在於那些具有特殊圖形且擁有非線性關係的基因，且希望找出全都是擁有非線性關係的基因，因此在我們目標二的公式 8 中，我們先針對這 g_i 與 g_j 這兩個不同的基因去算彼此之間的關聯性，而為了保留非線性關聯性的基因，我們設定{1- abs (Cor (g_i, g_j))}用以刪除具有高線性關係的基因，利用此公式用以判別基因之間關係是否具有非線性關聯性。

由於 MOEA/D 挑選的每個解都是一組基因，所以在目標一與目標二的部分，我們參考 mRMR 作法的目標一，使用相互資訊值的平均值來代表一組基因的 relevance 與 dependency，為了使 relevance 與 redundancy 的值能夠比較每組解，我們使用正規化將值控制在 0~1 的區間。

目標三：設定所選用基因的數目

$$no.of.gene = \left\{\frac{1}{[1+(k-n)^2]}\right\}^{exp(-q)} \quad (9)$$

k: targeted number of feature
n: number of feature
q: tunable parameter

針對第三個目標：設定所選用特徵的數目，允許特徵選擇一定數量範圍內的特徵，以控制每個解挑選出的特徵數目，在公式 9，利用我們想取得的基因數目 k 與實際挑選出的基因數目 n 彼此之間差異數量的平方根數與權重參數(q)來控制分數，用以當作評估的條件，讓我們挑出適當數量的特徵數目。

關於 q 值的設定，我們採用 exponential 策略(圖 4)，此策略可以有效控制選用基因的數目在一定範圍之內，假設我們想將選到的基因數目大約控制在比較小的範圍內，那麼我們可以將 q 值設定的比較小，這樣透過目標三算出來的分數，會使不同基因數目的控制分數彼此之間差異很大，使得不在我們想選的範圍之內的基因數目就比較不會被我們挑選到，反之，若我們將 q 值設定的很大，那根據基因數目算出來的控制分數彼此之間差異不大，會使得比較難將選用的基因數量控制在一定範圍之內。

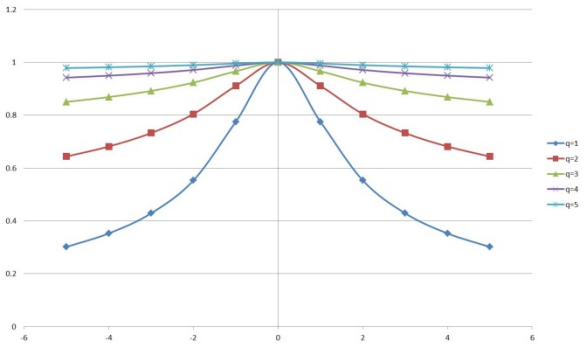

圖 4：權重(q)設定策略

圖 5 為整個搭配多目標作法的基因選取流程圖，由於此方法是只是改變 MOEA/D 的流程使其適用於基因選取，但還是如同一般演化式計算的方法，還需指定族群數目、演化世代、突變機率、交配機率值等。

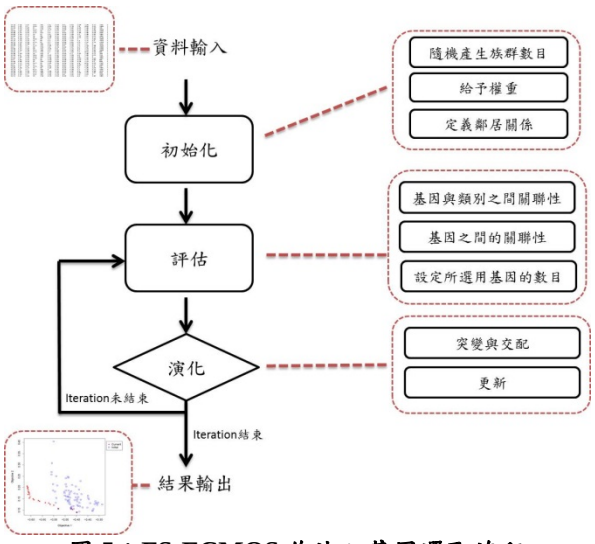

圖 5：FS-ECMOS 作法之基因選取流程

步驟一：初始化

首先會根據設定的族群數目，在解空間中隨機產生族群，接著會將這些隨機產生的族群轉成可以適用在解最佳化的問題，並給予各個族群權重，在 MOEA/D 作法中，此作法能夠比 NSGA-II 執行更快速的原因是因為 MOEA/D 透過鄰居關係的方式，將多目標問題轉化成單一子目標問題去做，所以我們利用歐幾里得距離(Euclidean distance)的方法，去找出與每一個子問題相近的鄰居，進而彼此比較。

步驟二：評估

在此步驟，我們會知道哪些基因是已經被選取到的，因此我們會將這些被選取到的基因去根據我們設定的目標，利用相互資訊值作法去計算 relevance 與 dependency，由於我們希望找到基因與類別之間具有高度相關性與基因之間具有非線性關係的基因，為了最後在圖形呈現上符合我們的視角，我們皆最小化這兩個目標的值，使其畫出來的圖類似 pareto graph，此外，透過設定的第三個目標讓基因只能選擇一定範圍內數量的基因。

步驟三：演化

我們會先從每個子問題與其鄰居中隨機挑選兩個解當作親代(parent)，根據交配機率來設定與解長度相同的二元染色體後隨機選取，若選取到的位置等於 1，那麼親帶會依據這個被選取到的位置，讓彼此同一邊的染色體作交配；而在突變的步驟，首先我們會根據步驟二產生的 relevance 值去針對基因排序，利用與做交配方式一樣的作法，隨機選取一個位置，若此二元染色體的位置的值為 1，那麼便對此位置做突變，同時也在另外沒被選到的位置上作突變，會這樣做的目的是希望能讓解往結果好的方向演化。

當我們做完突變與交配的步驟之後，我們會得到兩條新的子代染色體，這兩條子代染色體會與親代的染色體與鄰居做比較(利用權重和適應值相乘的結果)，將結果比較好的解留下來，繼續做演化。

步驟二與步驟三會重複做直到設定的演化世代結束，演化結束我們得到的族群中有很多個解，每個解都是一個子問題最好的解。

5. 結果

我們使用小圓藍細胞腫瘤(Small round blue cell tumors，簡稱 SRBCT)資料來套用我們的演化式計算模型搭配多目標決策作法 (Feature Selection: Evolutionary Computing with Multiobjective Strategy，簡稱 FS-ECMOS)，SRBCT 資料擁有 2308 個基因，63 個樣本與 4 種不同亞型的腫瘤(EWS、BL、NB 與 RMS)。

在 FS-ECMOS 做法中，我們指定族群數目為 1000、演化世代為 50、突變機率為 0.7、交配機率為 0.9，q 值為 1，並將基因數目控制在 2~5 的範圍內以利觀察探究，為了使挑出來的基因組合性更多，此實驗重複進行 30 次，此外我們透過最近鄰居分類法(Nearest neighbor classification，簡稱 NNC)結合交叉比對(Leave-one-out cross validation，簡稱 LOOCV)策略，針對挑選出來的基因組合去計算對於 SRBCT 資料分類的準確性(Accuracy)。

我們特別針對基因之間、基因與類別之間的資料去計算相互資訊值與皮爾森相關係數。計算之後，發現兩種相互資訊值的分布情況皆大部分落於 0.1~0.3 區間之間，而基因間的皮爾森相關係數分布情況大部分落在-0.3~0.3 區間之間。因此我們認為，在計算基因之間、基因與類別之間的關聯性時，只要相互資訊值大於 0.3，我們就認為彼此之間關聯性很高。基因之間的線性關係值為-0.3~0.3 之間時，則表示基因間線性關係不強烈。

表 1 為透過 FS-ECMOS 作法做基因挑選的部分結果，表中除了列舉該基因組合為哪些基因組合而成的之外，我們還將此組基因拆成成對的方式，針對成對的基因計算基因與類別之間的相互資訊值、基因與基因之間的相互資訊值、基因之間的線性關係以及對於 SRBCT 資料分類的準確性。

Solutions	Gene pair	F1vs. Classes	F2 vs. Classes	b/w Fs	Cor	Group Acc.
1389 1434	1389:1434	0.4682	0.3812	0.3029	0.046	0.8254
1389	1389:1434	0.4682	0.3812	0.3029	0.046	
1434	1389:2159	0.4682	0.2931	0.355	-0.0598	0.8889
2159	1434:2159	0.3812	0.2931	0.2841	-0.1393	
	552:867	0.299	0.3818	0.2984	-0.0034	
552	552:1389	0.299	0.4682	0.2919	-0.2864	
867	552:2159	0.299	0.2931	0.3241	-0.416	0.9524
1389	867:1389	0.3818	0.4682	0.2964	-0.332	
2159	867:2159	0.3818	0.2931	0.3062	0.2094	
	1389:2159	0.4682	0.2931	0.355	-0.0598	

表 1：FS-ECMOS 作法之部分結果

從表 1 的結果中，我們可以列舉幾個有趣的觀察現象：(1)從欄位 3 與欄位 4 來看，基因組合中個別基因與類別之間的相互資訊值很高；(2)欄位 5 的基因組合中的基因與基因之間的相互資訊值很高；(3)從欄位 6 中可以觀察到此結果符合我們目標二的設定，將基因之間高線性關係的組合拿掉，保留基因與基因之間線性關係低的(以平均模式做觀察)；(4)各組基因對於 SRBCT 資料做分類的準確率很高；在這幾個有趣的觀察現象中我們可以得知 FS-ECMOS 作法挑出來的基因組合有符合我們基因選取機制所設定的目標，挑出個別基因與類別之間相互資訊值高、基因與基因之間具有非線性關聯性的基因，且將基因數目控制在 2~5 的範圍之間。

為了瞭解各組基因對於 SRBCT 資料的分類效果，我們使用 Multidimensional Scaling (簡稱 MDS plot)將資料降成兩個維度的情況來對基因做檢視，圖 6 為各組基因中的其中一組(表 1 中的第三組，基因組合為基因 552、基因 867、基因 1389 與基因 2159)，從圖中可以觀察到藉由我們的 FS-ECMOS 作法篩選出來的基因組合可以使不同類別之間有好的區

隔力。

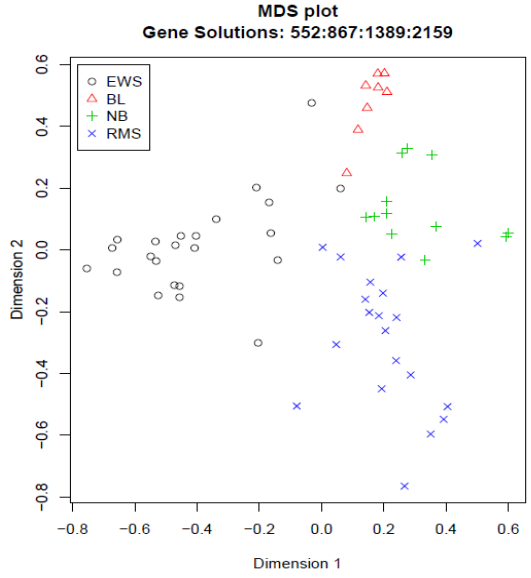

圖 6：FS-ECMOS 作法中 4 個基因組合的 MDS plot

6. 討論與結論

我們除了使用 SRBCT 資料套用我們的 FS-ECMOS 作法之外，還使用了 mRMR 方法中的前進選擇法(Forward selection)策略來做基因選取，表 2 為以 mRMR 作法做出的部分結果，我們選擇透過前進選擇法選出的前四名的基因，將這前四名的基因分別組合，分別以前 2 名基因、前 3 名基因、前 4 名基因的組合情況將各組基因拆成成對的模式，針對這些成對基因計算基因與類別之間的相互資訊值、基因與基因之間的相互資訊值、基因之間的線性關係及使用 NNC 結合 LOOCV 策略計算此組基因對於資料分類的準確性，再運用此結果與我們的 FS-ECMOS 作法做比較。

從表 2 中我們可以觀察到幾個有趣的現象：(1) 欄位 3 與欄位 4 的基因組合中個別基因與類別之間的相互資訊值很高；(2) 欄位 5 的基因組合中的基因與基因之間的相互資訊值很低；在這幾個觀察現象中，我們也可以得知，mRMR 作法挑選的結果符合基因與類別之間關聯性高的及基因與基因之間關聯性低的情況。

Solutions	Gene pair	F1vs. Classes	F2 vs. Classes	b/w Fs	Cor	Group Acc.
1389	1389:846	0.4682	0.363	0.161	-0.2392	0.8413
846						
1389	1389:846	0.4682	0.363	0.161	-0.2392	
846	1389:255	0.4682	0.3346	0.1562	-0.3376	0.8889
255	846:255	0.363	0.3346	0.1172	-0.2606	
	1389:846	0.4682	0.363	0.161	-0.2392	
1389	1389:255	0.4682	0.3346	0.1562	-0.3376	
846	1389:1955	0.4682	0.4036	0.256	-0.3455	0.9048
255	846:255	0.363	0.3346	0.1172	-0.2606	
1955	846:1955	0.363	0.4036	0.1898	-0.0797	
	255:1955	0.3346	0.4036	0.134	-0.1944	

表 2：mRMR 作法之部分結果

根據不同方法觀察到的有趣現象，可以得知 FS-ECMOS 作法除了對於 SRBCT 資料做分類的準確率很高之外，也能讓我們找出基因與基因之間線性關係低的解組合；從我們的 FS-ECMOS 作法與 mRMR 作法來比較，由於我們所設定的目標一(方程式 8)與 mRMR 設定的目標一(方程式 4)類似，，因此我們從表 1 與表 2 中觀察到的現象，均有基因組合中個別基因與類別之間的相互資訊值很高的情況，不過在目標二的地方，由於 mRMR 設定要找尋基因與基因之間關聯性低的情況，所以較沒辦法保留基因與基因之間相互資訊值高的基因組合，因為會被 mRMR 基因選取策略 MID 篩選，而我們的作法則是希望找出具有非線性關係的基因，且保留基因與基因之間相互資訊值高的基因組合，從此目標的結果上我們能夠得知我們方法與 mRMR 作法的差異，此外，從表 1 的欄位 6 也能觀察出我們挑出的解組合的線性關係低(以平均模式作觀察)，符合我們希望找出非線性關聯性基因的目標，而在 mRMR 中作法則無對基因之間的線性關係作探究。

mRMR 的前進選擇法策略最後只能挑出一組基因，而使用我們演化式計算模式的多目標決策作法，因為此作法是搜尋所有的解空間，找出同時符合多個目標的解組合，所以在分類效果上因我們的方法找尋較好的結果，因此在分類效果上會比 mRMR mRMR 作法來得好。

不過，針對我們的方法，我們的解組合目前是使用平均狀況來對線性關係作評估，所以有些基因對的線性關係或許還是會比較高，所以我們希望未來我們能夠再針對我們的目標做一些調整，將那些基因對線性關係高的解組合作篩選，使得挑出的結果能夠更符合我們的目的且能夠再針對基因與基因之間關係為非線性的圖形做更深入的探討。

參考文獻

[1] Isabelle Guyon, Andr'e Elisseeff. (2003) "An Introduction to variable and Feature Selection." Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

[2] Huan Liu, Hiroshi Motoda. (1998) Feature Selection for knowledge Discovery Data Mining. Kluwer Academic Publishers, Norwell, MA.

[3] Yvan Saeys, Inaki Inza and Pedro Larranaga. (2007) "A review of feature selection techniques in bioinformatics." Bioinformatics, vol. 23 pp. 2507-2519.

[4] Kalyanmoy Deb, Associate Member, IEEE, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. (2002) "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II." IEEE Transactions on Evolutionary Computation, vol. 6, pp. 182-197.

[5] Qingfu Zhang, Senior Member, IEEE, and Hui Li. (2007) "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition." IEEE Transactions on Evolutionary Computation, Vol. 11, pp. 712-731.

[6] Hui Li and Qingfu Zhang, Senior Member, IEEE. (2009) "Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II." IEEE Transactions on Evolutionary Computation, Vol. 13, pp. 284-302.

[7] Chris Ding, and Hanchuan Peng. (2003) "Minimum Redundancy Feature Selection for Microarray Gene Expression Data." IEEE, pp. 523-528.