利用資訊增益與瀰集演算法於基因微陣列之特徵選取與分類問題

楊正三
成大醫學工程系
p8896117
@mail.ncku.edu.tw

莊麗月
義守大學
化學工程系
Chuang
@isu.edu.tw

陳禹融
國立高雄應用科技大學
電子工程系
b3356629
@yahoo.com.tw

楊正宏
國立高雄應用科技大學
電子工程系
Chyang
@cc.kuas.edu.tw

摘要

特徵選取在解決分類問題上，舉凡機器學習、圖形辨識、調適性控制等研究都必須先利用特徵選取選出重要的特徵以進行前處理動作。本研究是以常見人體疾病之基因微陣列(Microarray)的資料集為主，基因微陣列其資料擁有低資料量、高特徵維度的特性，導致在計算分類時會造成冗長的運算時間。在本文中，我們利用資訊增益對每筆資料集計算出其所有特徵的資訊增益值，再用瀰集演算法做特徵選取與K-NN 分類，經由本研究所提出之兩個架構的特徵選取後，實驗結果顯示所提之方法有助於降低運算時間及提高分類正確率。

關鍵詞：微陣列、 Filter、 Wrapper、 特徵選取、 瀰集演算法。

Abstract

Feature selection to solve classification problem, now ubiquitous in machine learning, pattern recognition and data mining. In this article, these microarray datasets are general cancer-related of human, the characteristics of microarray datasets with information on low volume and the features of the high-dimension, therefore lead to a long time in calculated classification. This correspondence presents a novel memetic algorithm (MA) and information gain (IG) for a classification problem. We aimed that used to IG calculated the IG values to all feature at each dataset and then used memetic algorithm and K-NN for these feature. After our proposed two stage feature selection, we can find out this experimentation the proposed method outperforms existing methods in terms of classification accuracy and computational efficiency.

Keywords: Microarray, Filter, Wrapper, Memetic algorithm.

1. 前言

基因表現資料的分類與預測對於現今的生物醫學及生物學家在基因研究領域中是不可或缺的，主要原因是因為正確的分類結果能幫助生物學家進一步處理複雜的生物問題，及提高基因表現資料的使用效率。目前也廣泛的運用在基因鑑定、細胞分化、藥品研發、癌症分類等等。由於基因表現資料擁有低資料量、高特徵維度的特性，因此，若進行基因微陣列的分類必然是一件艱難的任務，所以特徵選取將會是一個有效降低運算時間且不影響分類正確率的前處理方法。

近幾年特徵選取已成為不少研究的焦點，主要是由一個大量的特徵集合中，能有效挑選出助於辨識的特徵子集合，此項技術也充分運用在機器學習、圖形辨識、調適性控制[1][2]等用途上。而特徵選取主要是針對過濾每個問題中含有不恰當、多餘的、或是被認為雜訊的資料，以減少分類時所花費的運算時間，因此更可以改善此問題的正確率[4] ，本研究前處理的部份是將所有基因微陣列做一次資訊增益(Information gain)，主要是將龐大的資料集透過資訊增益的運算後，可得到維度較小的基因微陣列，這是為了改善分類的效果以及提高重要特徵子集合的辨識。

特徵選取分為Filter 與Wrapper 兩種架構[5]；Filter 主要目的是將每個資料利用本身獨有的特徵，分別去計算其權重值，經由特徵集合D 中挑選出權重值較高的特徵子集合d 作為研究資料，常見Filter 方法有Information gain[4] 、 Relief[6] 、 gain ratio[4] 、 chi-square[7] 。Wrapper 特徵選取的方法是藉由持續的增加或刪除特徵，且利用學習演算法中的目標函數評估特徵集合的優劣，再利用分

類演算法去計算被挑選出的特徵子集合之正確率，常見Wrapper 的方法有基因演算法(Genetic Algorithm, GA)[8]、粒子族群最佳化演算法(Particle swarm optimization, PSO)[9]等。

在Zhu *et al*. [3]的研究中有提到，通常在預測分類正確率時，Wrapper 所得之結果會較 Filter 來的優秀，所以將這兩個架構結合在一起必可以互相補強彼此。本實驗之動機是將這兩個架構結合以提高分類效果，並利用7 筆基因微陣列進行實驗，過程中利用資訊增益搭配瀰集演算法做基因微陣列之特徵選取分類問題。本研究的架構分為兩個部分：第一部分先利用資訊增益計算出每個基因的資訊增益值，透過門檻值挑選，將被選取之特徵再進行第二部分；第二部份我們利用瀰集演算法將第一部份選出之特徵，再進行一次特徵選取，並搭配 K-NN 分類法以及LOOCV(leave-one-out cross validation)[4]作為正確率之評估，藉由這兩個架構之分類計算後，可得到較少的基因數量，以及更好的分類正確率。

2. 相關研究

2.1 資訊增益 (Information Gain, IG)

Information gain 是由 Quinlan 於 1979 年首先提出[4]，以資訊理論 (Information theorem) 為基礎，在計算資訊增益時會挑選出很多種值，主要是做為 ID3 (Interactive dichotometer 3, ID3）決策樹歸納演算法用於判斷分割點的屬性，建構決策樹的依據。

由於資訊增益可用於判斷此資料集的特徵屬性以及與其他類別之間差異的特點，也可藉由此特點將龐大的特徵集合縮減為容易處理的特徵子集合，這對處理分類問題時，可有效縮短分類的時間，因此在處理特徵選取時，資訊增益便是一個相當好的前處理。然而在處理特徵選取時，資訊增益會計算出所有特徵之資訊增益值，這些值主要是判斷所有特徵中，哪些特徵是對我們分類有幫助，而哪些特徵是必須刪除，所以必須設立一個門檻值，若得之資訊增益值超過設定之門檻值，此特徵就會被挑選出來，相反的，未達到門檻值標準的特徵必須被淘汰。舉例說明:假設分類的問題中包含了資料 N={1, 2, 3 …, n}、特徵維度 D={1, 2, 3, …, d}以及類別 K={1, 2, 3, …, k}，若欲求得單一特徵之資訊增益值，必須先得到兩個相關聯的數值，這兩個相關性的數值稱為熵值(Entropy)，如下方程式(1)與(2)，由這兩個 Entropy 的差值即為資訊增益，如方程式(3)

$$Entropy\ (N) = \sum_{i=1}^{K} P_i \log_k (\frac{1}{P_i}) = -\sum_{i=1}^{k} p_i \log_k p_i \quad (1)$$

$$Entropy\ (D_j) = \sum_{i=1}^{|D_j|} \frac{D_{ji}}{N} \times Entropy\ (D_{ji}) \quad (2)$$

$$IG(D_j) = Entropy(N) - Entropy(D_j) \quad (3)$$

方程式(1)中，Entropy(N)主要計算出整個分類問題中具有的總資訊含量，以此總資訊含量作為單一特徵的資訊增益根據。其中，N 表示為資料數量，K 表示為類別數量，P_i 表示為第 i 個類別在 N 比中出現的機率。

方程式(2)中，Entropy(D_{ji})是計算第 j 個特徵維度中，第 i 種數值、類別與資料數量之間的資訊含量，將特徵中不同數值、類別與資料數量之間的資訊含量加總即能表示單一特徵之總資訊含量。其中，D_{ji} 表示為第 j 個特徵維度中包含有 i 種不同的數值，N 代表為資料數量。

方程式(3)中，IG(D_j)表示在分類問題中第 j 個特徵所獲得的資訊增益，是由整個資料集中總資訊的含量，與第 j 個特徵的總資訊含量兩者之間的差值。

經由以上三個方程式運算後所得到之結果，便可獲得每個特徵的資訊增益，而每個特徵經由門檻值的篩選後，便可得到一組具有相當豐富分類辨識資訊的特徵子集合(Feature subset)[4]。

2.2 基因演算法 (Genetic Algorithm, GA)

基因演算法是在1975 年由John Holland所提出[8]，是仿照達爾文“適者生存”的自然法則，而基因演算法的進化步驟依序為：染色體編碼 (Initial encoding)、初始化族群(Initial population) 、計算適應函數值(Fitness)、選擇(Selection) 、交配(Crossover) 、突變(Mutation) 、替換(Replacement) 、終止條件，由這幾個演化程序來解決最佳化問題；首先，利用隨機產生族群數，再經由交配的過程中，隨機選出兩條染色體，互相交配，產生下一代，或是利用突變的特性，對單一的染色體進行突變，完成以上步驟稱之為一次迭代，當滿足迭代次數或是符合終止條件時便會獲得基因演算法之最佳解。

2.3 粒子族群最佳化 (Particle Swarm Optimization, PSO)

粒子群最佳化演算法是Kennedy 和 Eberhart 於1995 年時提出[9]，PSO 的概念源

自群體行為理論，啟發自觀察鳥類群體行動時，試著搜尋出未知的最佳解，且透過個體間特別的訊息傳達方式，使整個團體朝同一方向、最佳解而去[10]，是模仿此類生物行為來尋求群體最大利益的方法。PSO 的基本演算模式如下：首先以均勻分佈(Uniform distribution)隨機產生初始粒子群，每一個粒子都是一個求解問題的侯選解，粒子群會參考個體的最佳經驗，以及群體的最佳經驗，選擇修正的方式，完成一次迭代，經過不斷的迭代之後，粒子群會漸漸接近最佳解，直到滿足迭代次數或符合終止條件，便可得到最佳的粒子族群。

3.研究方法

本文是採用人體常見的疾病資料集進行實驗。針對此龐大的基因微陣列資料集，本研究先利用資訊增益將重要的特徵基因挑選出來，被篩選後的特徵，再由第二階段利用瀰集演算法做特徵選取，最後配合 K-NN 以及搭配 LOOCV 來計算出分類正確率。

首先我們利用 WeKa 軟體[15]計算出所有基因微陣列裡的特徵所能夠代表本身的的資訊增益值，經過計算後，若得到的資訊增益值越高，代表該基因對於與其他類別的區別性越高，也就表示對之後所作分類計算具有較重要的影響。在計算過程中，計算出每個特徵的資訊增益值後，須設定一個數值當作門檻值，用來區別基因的重要性。從研究中發現，利用資訊增益計算後大部分基因的資訊增益值皆為0，這表示基因微陣列資料中具有許多對於類別結果的區別沒有重要的影響性，因此在本研究中，我們將資訊增益值的門檻值設定為 0，當基因的資訊增益值有大於 0 的即被挑選出來進行下一階段的計算，而未超過門檻值的基因特徵將不被取用於下一個階段的計算；做這個步驟的目的，是為了將具有龐大特徵的資料集先做一次的挑選，但被挑選的特徵只會有部份可通過門檻值，此步驟不僅可縮短運算時間，並可提高分類正確率。

舉例說明，一個具有 9 個特徵數量的微陣列資料集中 $F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9$，而經由資訊增益之後，此 9 個資訊增益值分別 $F_1=0$, $F_2=0.4$, $F_3=0$, $F_4=0.9$, $F_5=0$, $F_6=1.2$, $F_7=0.6$, $F_8=0.5$, $F_9=0$，得知多數資訊增益結果為 0，因此本實驗門檻值設定為 0，發現超過門檻值條件的有 5 個，分別為 F_2, F_4, F_6, F_7, F_8，而下個階段只會將這 5 個基因繼續執行選擇的動作。

3.1 瀰集演算法 (Memetic algorithm, MA)

瀰集演算法是由 Dawkins' "在自私的基因" 一書中所定義的[11]，基因為生物遺傳的單位，而「瀰」可定義為一個能傳達「文化傳遞單位」的概念名詞，或是能代表「模仿」的單位，因為文化演化的速度以及造成的改變比生物演化更為驚人。瀰集演算法是根據文化演化的概念結合了多點搜尋特色以求解組合最佳化問題，並且有著不錯的表現。其與基因演算法有些許相似，但瀰集演算法在個體將資訊傳遞予下一代之前，會融合配適其擁有的資訊，也就是瀰集演算法結合了區域搜尋(Local Search, LS)的處理[12]，以區域最佳解的求解空間取代可行解的求解空間。

瀰集演算法，在作區域搜尋之前會先經過一段複雜的演化[13]，這是為了幫助染色體以及後代得到一些經驗。這對瀰集演算法來解決特徵選取的分類問題是一大益處並能提高分類正確率。

本實驗是利用隨機產生 30 組族群數，接著對產生出來的族群進行區域搜尋，因此可以得到族群的區域最佳解，然後仿照 GA 的交配與突變來產生後代，而這些後代,也會從區域搜尋中找出區域最佳解，有了那麼多的區域最佳解[14]，因此可以藉由這些解獲得區域中最好的全域最佳解。

3.2 K 個最近鄰居法（K-nearest neighbor, K-NN）

K個最近鄰居法是由Fix 和Hodges在1951年所提出，K-NN演算法是一種機器學習的演算法；利用特定的距離測量方式如：歐幾里得距離(Euclidean distance)，來找出最接近預測點的K個已知點，利用此K個已知點的結果來當作預測點的結果。在分類問題中，可由已知類別對未知類別進行分類預測，經由計算未知類別資料與已知類別資料的最小距離，依序可得知有K個已知類別資料點與未知類別資料點的最小距離，此K個已知類別資料點的類別結果，其出現頻率最高的類別即是此未知類別資料的分類結果。

目前在分類計算的研究中最常被使用的正確率評估方式有K-fold評估法與留一交叉評估法 (LOOCV)；在不同的資料集下，必須採用不同的評估方法，以避免產生過度訓練(Overfitting)，導致不客觀的正確率，通常在處理基因微陣列時都會使用K-fold評估法，而針對只有小數量的資料集，則是採用留一交叉評估法(LOOCV)，在本研究中，因為先使用資訊增益篩選過後，所以得到的資料會比原本的微陣列小很多，所以會採用留一交叉評估法來評估

分類正確率。

3.3資訊增益-瀰集演算法

本研究我們採用11個人體常見的疾病資料集進行實驗，這些龐大的基因微陣列資料集，首先使用資訊增益將重要的特徵先篩選出來再利用瀰集演算法來處理特徵選取問題，最後配合K-NN搭配LOOCV來計算出分類正確率。此階段將細分為好幾個部份來處理特徵選取及分類之問題，此階段的流程如圖1所示：

3.3.1 隨機產生初始族群數

在本文中，我們初始的族群數設為30組，並且對這30 組作染色體編碼，本研究的編碼方式是以二進位的字串$S=F_1F_2F_3\ldots F_i$，i=1, 2, …, n 來表示，一個位元代表一個特徵，位元值為1的，表示該特徵被選取，位元值為0 的，表示該特徵不被選取，舉例說明：經由資訊增益篩選後一個二進位的字串為S=10011 代表被選擇的特徵集合$X=\{F_1, F_4, F_5\}$而不被選的特徵集合為$Y=\{F_2, F_3\}$。

3.3.2 區域搜尋

本實驗的區域搜尋與傳統式的區域搜尋不同在於傳統區域搜尋是利用實數來解決問題，以及隨機選擇執行區域搜尋的次數，因此本研究為了避免前述之問題發生，而採用隨機產生該資料集特徵數量的 1/10 作為區域搜尋的次數。

區域搜尋其步驟為，先計算出最初始的分類正確率，再隨機選取一個維度判斷該特徵之位元是否為“0”(不被選取)，若為“0”就將該特徵替換為“1”(被選取)並計算其正確率。此時將未替換之正確率與替換後之正確率相比較，若替換後之正確率較未替換之優異，此時正確率被更變為替換後之正確率，倘若沒有正確率沒替換前好，該特徵就還原不動，抑或是本身就為“1”則換下一個被選機選取之維度繼續去執行區域搜尋，以此類推。

舉例說明如圖2，假設原始染色體之分類正確率假設為60%，之後隨機選到第3個維度，而判斷出維度3之特徵為“0”，此時將此特徵替換為“1”，再計算替換過後之新染色體之正確率假設為70%，則發現替換後的正確率大於原始正確率，那新染色體及正確率均會被替換，如圖2狀況一所示；但若替換後之正確率假設為50%時，發現替換後的正確率反而效果不佳，則該特徵以及正確率均不被替換，如狀況二所示；若隨機選取到的維度為4，發現該特徵為“1”，

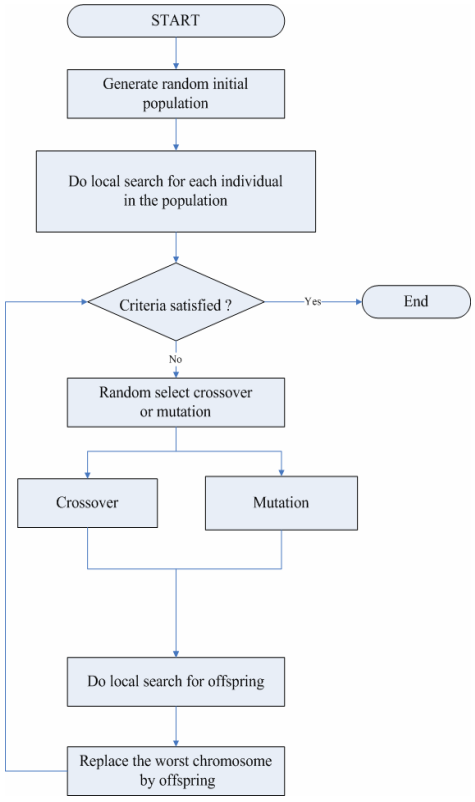
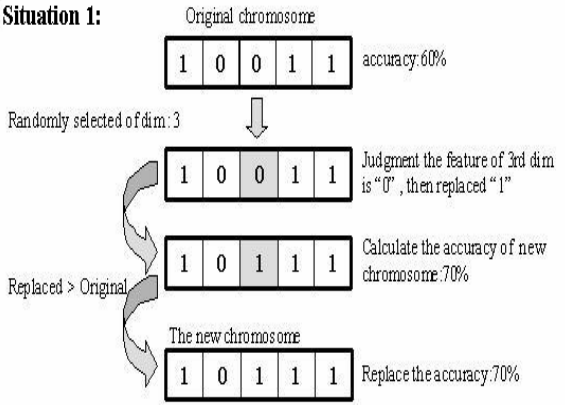


圖1. 瀰集演算法的流程圖

此時將此特徵替換為“1”，再計算替換過後之新染色體之正確率假設為 70%，則發現替換後的正確率大於原始正確率，那新染色體及正確率均會被替換，如圖 2 狀況一所示；但若替換後之正確率假設為 50%時，發現替換後的正確率反而效果不佳，則該特徵以及正確率均不被替換，如狀況二所示；若隨機選取到的維度為 4，發現該特徵為“1”，此時就隨機選取下一維度繼續進行區域搜尋，如狀況三所示，以此類推。

本文中針對二進制的特徵在計算正確率，發現若該特徵為“1”將它替換為“0”對此正確率改善不大，是因為任何運算式對“0”做運算其效果較不彰顯，但若該特徵為“0”，而將它替換為“1”時發現對正確率的結果會大為提高。



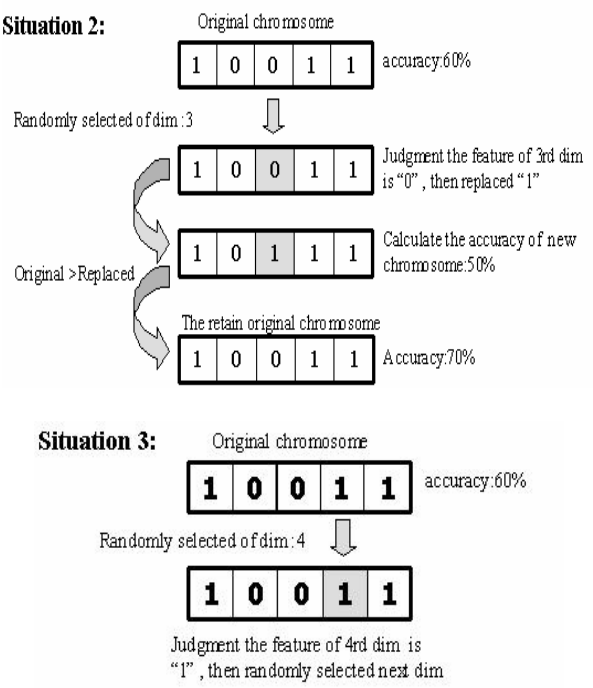


圖 2. 特徵替換示意圖

以下為區域搜尋之 Pseudo code。

```
Procedure Local Search
1. Begin;
2.  Random select a value j =1/10 feature
3.  numbers;
4.  For a given chromosome i calculate
5.   fitness_KNN(i);
6.   Using Loocv calculate classification
7.   accuracy of pre_chromosome;
8.   For j=1 to number of variables in
9.       chromosome i;
10.    If chromosome's feature j =0 then
11.      change the feature j=1;
12.      calculate fitness_KNN of the
13.      new_chromosome;
14.      Using Loocv calculate
15.      classification accuracy of
16.      new_chromosome;
17.      If pre_ fitness> new_fitness then
18.        retain the pre_ fitness;
19.      Else if replace the worst
20.            chromosome;
21.      End If;
22.    End If;
23.  Next j;
24. End;
```

3.3.3 交配

交配的方式是採用雙點交配，隨機產生兩個介於[1,D-1]的亂數，交換兩母代之間的位元數，因而產生不同於兩母代的新子代。

3.3.4 突變

突變的方法是對每一個染色體中的位元皆隨機產生一個亂數值，若位元數為 1 且亂數值小於突變率，則將位元數由 1 變 0，若位元數為 0 且亂數值小於突變率，則將位元數由 0 變 1。

3.3.5 替換

經由交配或突變過後的子代，必須與原本選擇的兩個母代做比較，若子代優於兩個母代，則子代將取代較差的母代；若子代介於兩母代之間，則此子代必取代較差的母代，若子代劣於兩母代，則與群族中最差的母代比較，若子代優於母代則取代母代，若母代優於子代，則保留母代捨棄子代。

4. 實驗結果與討論

本研究之資料集是來自 http://www.gems-system.org，我們針對 7 筆疾病基因微陣列做分類。由相關文獻得之 Wrapper 的方法會比 Filter 的方法所得到的結果還優異[3]，所以本研究，結合了 Filter 以及 Wrapper 的觀念各自補強彼此不足，發現不僅縮短運算的時間，更提高每筆資料集的分類正確率，在 Huang *et al.*[16]的研究中說明，針對基因微陣列的資料集，選擇越小數量的基因將可有效的提升分類正確率。

首先，利用資訊增益對資料集進行篩選，目的在於從全部基因中挑選出對於區別不同分類結果有重要影響性的基因，然後再進行分類計算，相較於不進行選擇而直接進行分類計算的結果更為優異。表1中記載著11筆基因微陣列的原始基因數量，以及表2中利用資訊增益篩選出的基本資料，其中符合門檻值且佔比例最高的為Brain_tumor2資料集，它佔比例中的43.07%，而最少的是Leukemia1資料集，它只佔了原始特徵數量的15.92%。

第二階段是針對這些被篩選出的特徵再進行第二次的特徵選取，本研究中我們設定的族群數為30，迭代數為100次。利用Brain_tumor1來說明，各個演算法之分類結果，如圖3所示，經由第一階段資訊增益篩選過後的特徵數量有1612個佔比例的27.23%，首先在初始化時分類正確率的評估是採用K-NN搭配LOOCV做分類正確率的計算(K=1)，其值在84.44%到94.44%之間，而我們加入Sort排序，其目的是將分類正確率做排序，這麼一來可確保每次都是由最好的正確率開始進行交配或突變(本實驗的交配率為：1.0，突變率為：0.1)，所以，使它的正確率不會被不好的取代，只會被更好的替換；第一次的迭代會先由94.44%進行交配或突變的工作，而在第57次迭代時其正確率以提高至95.56%，最後到了100次迭代後，發現結果為

95.56%，由下圖3得知利用瀰集演算法所得到的結果為95.56%，而基因演算法及粒子族群最佳化均之正確率均為93.33%，由結果得知瀰集演算法之正確率比基因演算法之正確率還優異許多。

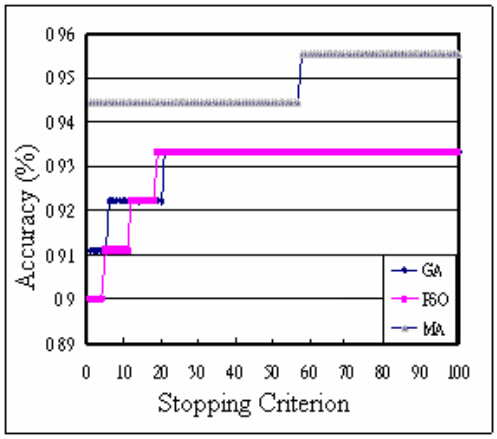


圖3.各演算法之分類結果

由表2可得知在Leukemia1此資料集效果沒有基因演算法好，原因是此筆資料集在未經過資訊增益時原始特徵數量為5327，而經過資訊增益後特徵數量為848，此特徵數量只佔了原始特徵數量的15.92%，且我們只取原資料集的1/10維度進行區域搜尋，發現此時分類正確率為98.61%比基因演算法還少了1.39%，但經由實驗證明，提高此筆資料維度至1/5發現其分類正確率效果也隨著提高至100%。

本實驗發現資訊增益搭配瀰集演算法的分類正確率效果比資訊增益加上基因演算的分類正確率還多出平均的2.32%，以及資訊增益加上粒子族群最佳化的效果更是多出平均的3.33%，和文獻中的分類方法的效果還優異；從表2我們比較了其他文獻的所提出的方法，在Statnikov *et al.*的研究中[17]，在分類器部分區分為Non-SVM與MC-SVM兩種，其中Non-SVM的部分結果最好的是K-NN其正確率為82.81%，而MC-SVM中最好的是OVR其正確率為93.67%，也由表1中我們得知只透過資訊增益選擇出的基因進行分類計算，得到的分類正確率為86.35%，雖然也得到良好的分類正確率，但仍然無法與MC-SVM部分的分類結果要來的更好從結果得知，因此本研究所提的方法確實可以有效的選擇出對於分類計算具有影響性的基因，因此之後在計算分類結果時，確實可以有效的提升分類正確率。

瀰集演算法較其他兩種演算法優異是因為基因演算法與粒子族群最佳化演算法沒有加入區域搜尋的步驟，由於基因演算法與瀰集演算法非常的相似，但瀰集演算法加入區域搜尋，而做區域搜尋是為了執行每一個染色體以改善其經驗和得到族群的區域最佳解，而交配與突變的運算也應用在產生後代，這些後代也會從區域中得到最好的區域最佳解[14]。

本研究更在瀰集演算法裡的區域搜尋中改善了傳統式的區域搜尋，傳統的區域搜尋是利用實數來解決問題，以Brain_tumor2為例，傳統的方法會隨機選擇執行區域搜尋次數，因此得知正確率為84%，而我們所提供的區域搜尋次數有所限制，會隨機針對此筆資料的1/10個特徵數量，進行區域搜尋，而所得之正確率為98%，結果較傳統式的優異。

而本結果所得的整體平均之正確率比基因演算法和粒子族群最佳化以及利用MC-SVM分類器中最好的正確率來的優秀。由於基因演算法以及粒子族群最佳化演算法必須對所有經過資訊增益後的特徵數量進行特徵選取，這會使運算的時間更為冗長，而本研究的演算法只需對所有特徵數量的1/10進行特徵選取，因此更是大幅提升運算效率，降低運算時間。

5. 結論

在本文中，我們提出的演算法先使用資訊增益進行資料篩選，再利用瀰集演算法搭配K-NN進行特徵選取及計算分類正確率。實驗結果顯示，我們所提出的方法可以有效的降低運算時間以及提高運算效率，並且可以有效的提升分類正確率。與文獻上所提供之方法比較之結果資料皆優於Non-SVM與MC-SVM所提正確率更提升至98.39%。希望在未來的研究，我們可以藉由測試不同區域搜尋的維度，以便有效的改善分類正確率。

參考文獻

[1] Oh, I.-S., J.-S. Lee, B.-R. Moon, "HybridGenetic Algorithms for Feature Selection," ***IEEE Trans. Pattern Analysis and Machine Intelligence***, Vol.26, No.11, Nov, 2004.

[2] Vafaie, H., and K. De Jong, "Robust Feature Selection Algorithms," ***In Proceedings of the 5th IEEE International Conference on Tools for Artificial Intelligence***, Boston, MA:IEEE Press, pp.356-363, 1993.

[3] Zhu, Z., Y. S. Ong, and Dash M. Dash, "Wrapper-Filter Feature Selection Algorithm Using a Memetic Framework," ***IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS -PART B: CYBERNETICS,*** Vol. 37, No.1, pp.70-76, 2007.

[4]Quinlan. J. R., "Induction of decision trees," ***Machine Learning***, No. 1, pp. 81-106, 1986.

[5] Dash, M. and H. Liu, "Feature selection for classification," ***Int. J. Intell.Data Anal.***, vol. 1, no. 3, pp. 131–156, 1997.

[6] Robnic- ˇ Sikonja M. and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," ***Mach. Learn***, vol.53, no. 1/2, pp. 23–69, 2003.

[7] Liu, H. and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," ***in Proc. 7th IEEE Int. Conf. Tools With Artif. Intell***, pp.388-391, 1995.

[8] Holland, J.H., "Adaptation in Nature and Artificial Systems," ***MIT Press***, 1992.

[9] Kennedy, J. and R. Eberhart, "Particle swarm optimization," ***Procee dings of the IEEE international conference on neural networks*** (Perth,Australia), pp.1942–1948, Piscataway, NJ: IEEE Service Center, 1995.
Shi, Y., R. Eberhart, "A modified particle swarm optimizer," ***Proceedings of the IEEE international conference on evolutionary computation.Piscataway,*** pp. 69–73, 1998.

[11] Dawkins, R., "The selfish gene," ***Oxford: Oxford University*** Press, 1976.

[12] Moscato, P., M.G. Norman, "A memetic approach for the traveling, salesman problem—implementation of a computational ecology for, combinatorial optimization on message-passing systems," ***Valero M, Onate E, Jane M, Larriba JL, Suarez B, editors.International conference on parallel computing and transputer, application.*** Amsterdam, Holland: IOS Press, pp. 177–86, 1992.

[13] Merz, P., B. Freisleben, "A genetic local search approach to the quadratic assignment problem," *In:* ***Ba¨ck CT, editor. Proceedings of the 7th international conference on genetic algorithms. San Diego, CA: Morgan Kaufmann;*** pp.465–72, 1997.

[14] Elbeltagi, E., T. Hegazy, and D. Grierson, "Comparison among five evolutionary-based optimization algorithms," ***Advanced Engineering Informatics***, vol. 19, pp. 43-53, 2005.

[15] Frank, E., M. Hall, L. Trigg, G. Holmes, I.H. Witten, "Data mining in bioinformatics using Weka,"***Bioinformatics***, Vol. 20, pp. 2479-81, 2004.

[16] Huang, H.L., C.C. Lee, and S.Y. Ho, "Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers,"accepted by ***BioSystems,*** 2006.

[17] Alexander, R. Statnikov, Constantin F. Aliferis, Ioannis tsamardinos, Douglas Hardin, Shawn Levy, "A comprehensiveevaluation of multicategory classification methods for microarray gene expression cancer diagnosis," Bioinformatics(21): pp.631-643, 2005.

表**1. 7**筆基因微陣列經過資訊增益後所篩選出的特徵數量與分類正確率

檔案名稱	原始特徵數量	經過IG 計算後特徵數量	比例	IG-KNN
11_Tumors	2533	3181	25.38 %	83.33
Brain_Tumor1	5920	1612	27.23 %	88.89
Brain_Tumor2	10367	4465	43.07 %	78.00
Leukemia1	5327	848	15.92 %	93.06
Leukemia2	11225	4596	40.94 %	91.67
SRBCT	2308	669	28.99 %	98.80
DLBCL	5469	882	16.13 %	93.51
Average				86.35

表**2** 利用**IG_GA, IG_PSO, IG_MA** 分別對**7**筆基因維陣列資料集做比較

方法	Non-SVM			MC-SVM					IG_GA	IG_PSO	IG_MA
資料集	KNN	NN	PNN	OVR	OVO	DAG SVM	WW	CS	KNN	KNN	KNN
11_Tumors	78.51	54.14	77.21	94.68	90.36	90.36	94.68	95.30	92.53	93.68	**96.55**
Brain_Tumor1	87.94	84.72	79.61	91.67	90.56	90.56	90.56	90.56	93.33	93.33	**95.56**
Brain_Tumor2	68.67	60.33	62.83	77.00	77.83	77.83	73.33	72.83	88.00	84.00	**98.00**
Leukemia1	83.57	76.61	85.00	97.50	97.32	96.07	97.50	97.50	**100.00**	98.61	98.61
Leukemia2	87.14	91.03	83.21	97.32	95.89	95.89	95.89	95.89	98.61	95.83	**100.00**
SRBCT	86.90	91.03	79.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
DLBCL	86.96	89.64	80.89	97.50	97.50	97.50	97.50	97.50	**100.00**	**100.00**	**100.00**
Average	82.81	78.21	78.32	93.67	92.78	92.60	92.78	92.80	96.07	95.06	**98.39**