使用 BPSO-TS 於基因選擇與分類問題

楊正三
國立成功大學
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

李榮杰
國立高雄應用科技大學
電子工程系
1095320149
@cc.kuas.edu.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

基因微陣列上包含了眾多的基因資訊，經過分析可以獲得許多疾病的訊息資訊，對於人類診斷與治療疾病有很大的幫助。基因微陣列是屬於高特徵維度、多類別屬性與低資料量特性的資料集。降低基因微陣列的特徵數量是最主要的課題。在分基因微陣列上，希望找到一組能夠代表原本基因組合的基因子集合。本文研究中，採用混合了粒子組群最佳化與禁忌搜尋法內，來解決基因微陣列基因選擇跟分類的問題。並搭配以 K 最近鄰居法為分類器，搭配留一交叉驗證法來計算分類錯誤率。根據實驗結果顯示，在 5 個基因微陣列中，提出的方法能夠有效的降低基因數量並且與文獻上提出的方法相比，能獲更低的分類錯誤率。

關鍵詞：基因微陣列，粒子族群最佳化，禁忌搜尋法，K 最近鄰居法，留一交叉驗證法。

Abstract

Microarray data contain the genetic information, many diseases can be analyzed, and the diagnosis and treatment of human diseases will be very useful. The challenge to select relevant genes from microarray data is because of theirs high-dimensional features, multi-class categories and sample size. For classification, the main task is to decrease the microarray data dimensionality first. In order to analyze microarray data, our goal is to find the optimal subset of features (genes) to represent the original set of features. In this study, we use binary particle swarm optimization (BPSO) and tabu search (TS), hybridizing two kinds of algorithms to solve the microarray data selection and classification. The K-nearest neighbor method with leave-one-out cross-validation serves as a classifier. The proposed BPSO-TS approach is compared to five microarray data from the literature. Based on the experimental results, the proposed method could not only effectively reduce the number of genes and obtain lower classification error rates.

Keywords: Microarray, binary particle swarm optimization, tabu search, K-nearest neighbor, leave-one-out cross-validation.

1. 前言

基因微陣列(Gene Microarray)是生物醫學上一個重要技術，近年來已發展的越來越重要了，此技術已形成重要的研究領域。基因微陣列對於醫學發展是非常重要的一環，主要的原因在於它可以同時針對數以萬計的基因進行分析，同時企圖藉由分析所產生的結果尋找出基因彼此之間的關係，並提供生物學家監測基因微陣列資料的能力，進一步做為分類與預測的目的，對於人類疾病預測跟治療都有極大的貢獻。基因微陣列是個非常龐大的資料量，具有高特徵維度、多類別屬性與樣本數較少的特性。樣本數少的特性使得分類演算法在測試與訓練時，無法獲得較高的分類正確率。基因微陣列的分類與預測對於現今的生物醫學與分子生物學的基因研究領域已經是不可或缺的，其主因在於正確的分析結果能夠幫助生物學家進一步的處理複雜的生物問題，以提高基因表現資料的使用效率。由於基因表現資料擁有高特徵維度、樣本數較少的特性，因此，若直接進行基因微陣列的分類勢必會造成運算時間過長的問題，故基因微陣列在進行分類前，採取特徵選取將會是一個有效降低運算時間且盡可能不影響分類正確率的前置處理方法。

特徵選取是一種從原始特徵集合中挑選出一個特徵子集合的技術，此技術可被視為進行分類處理前一個前處理的技術。特徵選取是屬於最佳化問題，屬於 NP-hard 問題內的一

種。我們在解決特徵選取的問題上，不可能對全部特徵集合進行詳細計算，那會造成大量的時間花費跟成本[2]。因此特徵選取最主要的目的是希望能夠移除掉不必要的特徵，只留下對分類辨識有幫助的特徵，希望能得到對分類辨識有幫助特徵的最佳特徵子集合。對分類不相關的特徵對分類辨識是沒有幫助的，藉由移除掉這些特徵後，分類計算的效率則會提高並且不會因為這些對分類無幫助的特徵影響了分類錯誤率。

在基因微陣列分類問題上，可以把它視為特徵選取的問題，如何選擇出一組最佳的基因特徵組合是一個非常複雜的工作。一個好的演算法就顯的重要了，經過演算法的運算我們可以得到一組近於最佳解，這代表了此組合可能是診斷某個疾病最佳的基因組合方式。用於解決特徵選取的方法，越來越多最佳化的演算法應用於此，包括有模仿自然界演化的基因演算法 (Genetic Algorithm) [8] 、具 Swarm Intelligence 的粒子族群最佳化(Particle Swarm Optimization) [11]及具有記憶功能的禁忌搜尋法(Tabu Search) [9]等較常見的演算法。

在本文中針對基因微陣列的特徵選取上，我們提出了混合兩種不同演算法的方式。主要是以粒子族群最佳化為基礎並結合禁忌搜尋法，使得特徵選取的效能與分類的結果皆能優於未進行特徵選取的方法。本研究利用兩種不同演算法各自的優點，利用禁忌搜尋法幫助粒子族群最佳化的粒子得到在解空間較好的搜尋位置，以便能跳脫區域解，進而能得到趨近於全域最佳解的解。並搭配以 K 最近鄰居法做為適應性函數(Fitness function)，並使用留一交叉驗證法(Leave-one-out Cross-Validation, LOOCV)的評估方式對分類錯誤率進行計算。我們使用了 5 個基因微陣列資料來進行特徵選取的實驗，根據實驗結果此方法不但能有效減少特徵(基因)數量並且能夠文獻中其他方法得到更低的分類錯誤率。

2.研究方法

本研究混合了兩種不同的演算法來進行特徵選取，主要是以粒子族群最佳化為基礎並結合了禁忌搜尋法(BPSO-TS)。首先以粒子族群最佳化進行第一次的特徵選取，在粒子族群最佳化的迭代過程中，加入了禁忌搜尋法，進一步的對粒子族群最佳化所產生的特徵子集合給予最佳化，同時，再藉由禁忌搜尋法進行第二次的特徵選取，如此，直到粒子族群最佳化的迭代結束為止。如此可以獲得比單一使用一種演算法還要更少的特徵數量跟更低的分類錯誤率。

粒子族群最佳化與禁忌搜尋法的適應值函數皆以 K 最近鄰居法搭配留一交叉驗證法。以下針對所提出的方法給予說明介紹：

2.1 K 最近鄰居法

K 最近鄰居法(K-Nearest Neighbor, K-NN)是由 Fix 和 Hodges 在 1951 年[1]所提出的，屬於監督式學習(Supervised Learning)的一種。一個資料若在特徵空間中的 K 個最鄰近的資料中的大多數屬於某一個類別，則該資料也屬於這個類別，距離計算方式則是採用歐基里德距離 (Euclidean distance) 或是馬式距離(Mahalanobis distance)。在本研究中，採用了歐基里德距離。該方法只依據最鄰近的一個或者幾個資料的類別來決定待分類的資料所屬的類別。K 是影響分類效果最主要的因素 [10]，K 會依所要測試的資料的不同而進行不同值的選取，一般來說 K 越大需要越久的運算時間和可能較低的分類錯誤率。本文所用的基因微陣列是高特徵維度與多類別的，為了考量計算時間，因此我們採用 1-NN 即一個鄰居的方式來縮短計算時間。留一交叉驗證(Leave-one-out cross-validation, LOOCV)，評估方式進行分類時，若有 n 筆資料，在每一次的進行評估時只有採用 1 筆測試資料，而另有 n － 1 筆當作訓練資料，每一次訓練這 n - 1 筆的訓練資料將可建構一個分類器，再藉由測試資料驗證分類器的正確性並且獲得測試資料的類別。本研究 BPSO-TS 是採用 1-NN 搭配 LOOCV 的方法來計算基因微陣列分類的錯誤率做為適應函數值。

2.2 粒子族群最佳化

粒子族群最佳化 (Particle Swarm Optimization, PSO) [6]是 Kennedy 和 Eberhart 於 1995 年所發展出來的一種進化計算的技術。不像基因演算法所使用的基因運算子，選擇、交配與突變。PSO 是模擬鳥類飛行與魚類游水的族群特性。在 PSO 的解空間中，一個解視為一個粒子，每個粒子都有自己所在的位置與速度。粒子移動時，粒子調整自身的位置與速度依據的是粒子本身的最佳經驗(*pbest*)和群體的最佳經驗(*gbest*)。在搜尋空間中，每一個粒子都有各自負責的搜尋區域，每個區域都將朝向最佳化的目標做搜尋，因此，大多數的情形下，粒子族群最佳化能夠較快的收斂並尋找

到全域最佳解。

PSO 原本設計解決實數最佳化的問題。為了能夠擴展 PSO 解決離散與二進制的問題，Kennedy 和 Eberhart 又提出了二進制版本的 Binary PSO (BPSO)[7]。

2.2.1 產生每一個粒子初始的位置與速度

在粒子族群最佳化的搜尋空間中，假設有 P 個粒子，每一個粒子皆有其位置 X_i 與速度 V_i，其中 i=1,2, …, P。粒子位置 $X_p=\{X_1, X_2,\ldots, X_p\}$。粒子位置表示方式二進制的方式，隨機產生 0 或 1 的值，0 代表此特徵(基因)不選擇，1 代表此基因選擇。粒子速度 $V_p=\{V_1, V_2,\ldots, V_p\}$。是一個均勻的隨機亂數產生介於[0,1]的數值。

2.2.2 計算每一個粒子的適應函數值

在 BPSO 中，其適應函數值是以每一個特徵子集合所能獲得的分類錯誤率做為適應函數值，以此尋找最佳的特徵子集合。而所採用的適應函數為 1-NN 搭配以 LOOCV 的方式。

2.2.3 更新目前 BPSO 的 *pbest* 與 *gbest* 的值

在 BPSO 中，若某一個粒子在目前的迭代所求算的適應值比該粒子在之前的疊代所產生的適應值都更加接近最佳化，則表示目前該粒子所在位置較接近於最佳化粒子的位置，因此，會將目前的位置稱作為該粒子自身的最佳經驗(*pbest*)。另外，至目前的迭代次數為止，搜尋空間中所有粒子的最佳值則稱作群體的最佳值(*gbest*)。

2.2.4 更新每個粒子目前的位置與速度

每一個粒子在每一次的迭代中都必須更新自身目前的位置與速度。BPSO 採用的 3 個更新公式如下：

$$v_{id}^{new}=w\times v_{id}^{old}+c_1\times rand_1\times\left(p_{id}-x_{id}\right)+c_2\times rand_2\times\left(p_{gd}-x_{id}\right)$$

$$S\left(v_{id}^{new}\right)=\frac{1}{1+e^{-v_{id}^{new}}}$$

$$\text{if } \left(rand<S\left(v_{id}^{new}\right)\right) \text{ then } x_{id}^{new}=1;\ \text{else } x_{id}^{new}=0$$

其中，$i=1, 2, \ldots, P$，$d=1, 2, \ldots, D$，v_{id}^{new} 代表的是粒子更新後的速度，v_{id}^{old} 代表的是粒子更新前的速度。w 是慣性權重[35]， 設定為 0.9。c_1 與 c_2 是學習因子，兩者的值通常是相等的，一般皆設定為 2。$rand_1$ 和 $rand_2$ 是介於 0 和 1 之間均勻分佈的隨機變數。$\left(p_{id}-x_{id}\right)$ 是粒子本身最佳值的位置與粒子目前所在位置之間的距離。$\left(p_{gd}-x_{id}\right)$ 是搜尋空間中群體最佳值的位置與粒子目前所在位置之間的距離。x_{id}^{new} 代表的是粒子更新後的位置，x_{id}^{old} 代表的是粒子更新前的位置。在更新公式中，$S\left(v_{id}^{new}\right)$ 代表的是一個 Sigmoid 的函數，由於本研究是將 v_{id}^{new} 視為 x_{id}^{new} 是否變更的機率，因此必須將其經由 $S\left(v_{id}^{new}\right)$ 的轉換，使其成為一個[0,1]之間的數值，以此來判斷 x_{id}^{new} 中的特徵選擇是否需要改變。

2.3 禁忌搜尋法

禁忌搜尋法(Tabu Search, TS)是在 1990 年由 Glover 所發展出來的，它是一種萬用式啟發式(Meta-heuristic)演算法，可以用來解決組合的最佳化問題(Combinatorial Optimization Problem)，普遍應用在各種不同領域的問題上[4][5]。TS 是種具有彈性架構並具有記憶功能的一種演算法，它跟其他一般搜尋演算法在觀念上有所不同，它是一種 local neighborhood search，它可以允許新的解移往較差的解是希望避免落入區域最佳解。

為了防止在解空間無謂的重覆搜尋，TS 使用了兩種記憶方式是分別為短期記憶(Short-term memory)稱為禁忌名單(Tabu List, TL)，用來禁忌重覆搜尋過的解，並所以可用來限制解的移動方向。並加上了免禁準則(Aspiration Level, AL)來打破符合的解不會被禁忌名單所限制，兩者互相搭配下可以記錄跟導引整個 TS 的搜尋過程。除了短期記憶(Short-term memory)外，並可使用長期記憶(Long-term memory)，主要是使用了加強性(Intensification)仔細搜尋過之前所經過的有希望的區域與多樣性(Diversification)能夠廣泛搜尋不同的區域兩種方式。

若 S 為到目前最好的解，TL 為禁忌名單，k 表迭代次數。禁忌搜尋法可表示為四個主要部分[12]:

Step1 初始化:產生初始解 x，讓 $S=x$，$k=1$，TL 為空集合。

Step2 產生候選者名單:在 x 產生鄰近解(Neighbor solutions)集合後，隨機從其中挑選出候選者名單 $N(x)$。

Step3 移步:(a)若 $N(x)$為空集合則返回 **Step 2** 並重新產生候選者名單，否則在 $N(x)$中找出最好的解 y。(b)若 y 在 TL 中且不滿足免禁準則，則讓 $N(x)=N(x)-\{y\}$並返回 **Step 3**(a)。否則

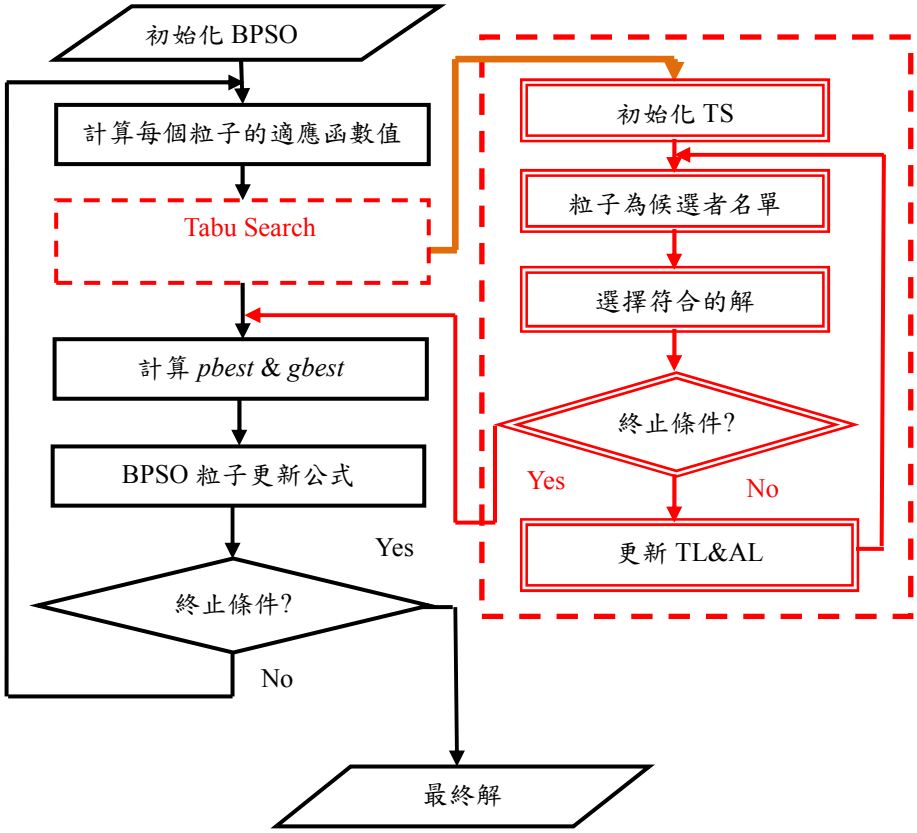

圖 1 BPSO-TS 流程圖

讓 $x=y$，並且 $S=y$ 若 y 比 S 還要好的解。

Step4 輸出:若滿足終止條件，則停止並輸出 S。反之，把 x 加入到 TL，即增加新解到 TL 的尾端若 TL 超過它原本所定義的大小，則移除最前端的解。讓 $k=k+1$ 並回到 **Step 2**。

圖 1 為本研究 BPSO-TS 的流程圖。TS 的候選者名單是以 BPSO 的粒子為主。因此 BPSO 的粒子數目會等同於 TS 的候選者名單數量。TS 的候選名單的格式，與 BPSO 粒子的位置 $X_p=\{X_1, X_2,\ldots, X_p\}$ 相同，代表基因微陣列的基因的特徵子集合。

3. 結果與討論

3.1 基因微陣列資料

本文所實驗的基因微陣列資料集來源 http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html [3]，並選用了 5 筆當作我們的研究資料集。這 5 筆基因微陣列資料分別為 Leukemia、Breast 2 class、Brain、Lymphoma 和 Prostate。微陣列基本格式如**表 1** 所示，列出了 5 個基因微陣列基本資料，包含了 Genes、Patients 和 Classes。Genes 表示原本基因微陣列所含的基因數量。Patients (Samples)則代表了樣本個數。Classes 則是基因微陣列是屬於幾類別的資料。舉例來說，Leukemia 是屬於兩類別的基因微陣列，而 Brain 是屬於多類別的基因微陣列(Classes >2)。

表 1 基因微陣列基本格式

Format / Dataset	Genes	Patients	Classes
Leukemia	3051	38	2
Breast 2 class	4869	78	2
Brain	5597	42	5
Lymphoma	4026	62	3
Prostate	6033	102	2

3.2 討論

表 2 中列出了經過 BPSO-TS 特徵選取過後所留下的基因特徵數量。**表 3** 則列出了文獻上對於這 5 筆基因微陣列其他方法的分類錯誤率與全體平均錯誤率與 BPSO-TS 所得到的錯誤率與全體平均錯誤率。粗體字則代表全部方法

中最低的分類錯誤率。

從**表 2** 中的結果得知，經過 BPSO-TS 特徵選取，這 5 筆基因微陣列的基因特徵數量全都減少了，舉例來說 Breast 2 class 從 4869 個基因減少到 971 個基因，Brain 從 5597 個基因減少到 433 個基因。所以若計算 Breast 2 class 基因微陣列時，計算分類錯誤率時是利用所選擇的 971 個基因來做分類的。這些由 BPSO-TS 所選擇基因扮演了很重要的角色在基因分類問題上，因為這些選擇的基因會影響到分類錯誤率。在選擇的基因組合中，若包含了大部分對分類有幫助的基因，則進行分類時分類錯誤率則會較低。若包含了大部分對分類沒幫助的基因，則會造成分類錯誤率提高。所以說為什麼進行分類時，特徵選取所選取的基因特別重要。

在 Leukemia 與 Lymphoma 這兩個基因微陣列得到一個重要的結果，在這筆基因微陣列資料集中基因特徵數量經過 BPSO-TS 後只剩下了一小部分的基因特徵數量但是分類錯誤率卻能得到 0.0，因此我們可以得知並非所有的基因特徵對分類都是具有幫助的。

在**表 3** 可以得知，本文實驗結果是與文獻[3]互相比較，包含文獻中的 4 種方法 Random Forest(s.e.=0)，Random Forest(s.e.=1)，Shrunken centroids(SC.s) and Nearest neighbor + variable selection(NN.vs)。在這 4 種方法中，5 個基因微陣列的平均分類錯誤率分別為 0.15、0.146、0.134、0.142，本研究所提出的方法的平均分類錯誤率則是 0.078，比其他 4 種方法都還要還低。BPSO-TS 在這 5 筆基因微陣列有 4 筆基因微陣列 Leukemia，Breast 2 class、Brain、和 Lymphoma 的分類錯誤率都較為優秀，分類錯誤率都比文獻中所提出的 4 種方法還要來的低。雖然 Prostate 在 BPSO-TS 分類錯誤率輸了 Random Forest(s.e.=0) 0.005。但是對於整體平均分類錯誤率來說，BPSO-TS 的效果還是極具競爭力。

BPSO 化擁有收斂速度快,易於實做且全域搜尋能力不錯等優點，但是若只單純使用 BPSO 於基因微陣列進行特徵選取的話，因為基因微陣列的特性是具有相當大量的基因，直接計算有極大機會落入區域最佳解。為了改良此缺點，我們加入了 TS 於 BPSO 內。TS 移步時，可以往移往較差的解，也就是說可以接受比目前現在還要差的解，這樣的特性可以讓 TS 有機會能跳脫出區域最佳解，進而有希望搜尋到全域最佳解。TS 最佳化 PSO 的粒子，增加粒子跳脫區域解的能力，進而搜尋到更為優良的解。

3.3 參數設定

TS 的參數包含了候選者名單大小與禁忌名單。一般來說候選者名單越大，越能增加其在解空間的搜尋範圍，因此能增加找到全域最佳解的機會。但是候選者名單越大，缺點則是其計算時間也越來越長，造成效率降低。為了避免計算時間的太過於冗長與效率低落，我們設定了 20 為候選者名單大小。而禁忌名單我們採用了 Glover 是建議魔術數字 7 為禁忌名單的大小。因為 TS 最佳化 BSPO 粒子所代表的特徵集合，只是希望能最佳化各粒子所以 TS 迭代次數不需過多，所以原則上設定為 30 次。

BPSO 的參數設定雖然較少, 但適當的給予參數, 仍然對其搜尋最佳化的結果有很大的幫助, 尤其對於 w 、c_1 及 c_2 這三個參數的設定特為重要，假使調整過小，則粒子的移動距離將會縮小，此舉雖然仍可找到較佳的解答，但所花費的運算時間也相對的會有些微的提升；若是調控過當則可能會造成粒子移動的距離過大而提早收斂，導致無法獲得較佳的結果，因此一個適當的參數調控將有助於粒子族群最佳化尋得較佳的結果。BPSO 的粒子的數目是跟 TS 候選者名單相同都設定為 20，因為 TS 是最佳化 BPSO 的粒子所以設定的數量為相同。$w=0.9$ 以及 $c_1=c_2=2$，BPSO 迭代次數則設定為 50 次。

4. 結論

基因微陣列的技術已應用在於人類的日常生活中，如何運用合適的演算法來幫助分析基因微陣列是重要的課題。本文提出以 BPSO-TS 的方法，BPSO 的迭代過程中加入了 TS，TS 最佳化了 BPSO 的粒子，TS 針對 BPSO 產生的特徵子集合再給予最佳化，所產生的特徵集合以 1-NN 搭配 LOOCV 作為分類器來計算分類錯誤率。經過實驗得知，本文所提出的方法在 5 個基因微陣列中，有 4 個資料集在都要比文獻中所提出的方法的分類錯誤率還要來的低。在全體平均分類錯誤率上也獲得了 0.078，比文獻另外四種方法整體上的分類錯誤率 0.15、0.146、0.134、0.142 都還要來的低，證明了 BPSO-TS 在此類問題是有不錯效果的。未來研究方向希望能利用各種演算法的優點，互相搭配或加以改良，期望在基因微陣列資料集上還有更不錯的結果。

參考文獻

[1] Cover, T. and P. Hart, "Nearest neighbor pattern classification," I***EEE Transactions on Information Theory***, vol. 13, pp. 21-27, 1967.

[2] Cover, T. M. and J. M. Van Campenhout, "On the possible orderings in the measurement selection problem," ***IEEE Trans. Systems, Man, and Cybernetics***, vol. 7, pp. 657-661, 1977.

[3] Diaz-Uriarte, R. and S. Alvarez de Andres, "Gene selection and classification of microarray data using random forest," ***BMC Bioinformatics***, vol. 7, p. 3, 2006.

[4] Glover, F., "Tabu Search-Part I," ***ORSA Journal on Computing***, vol. 1, pp. 190-206, 1989.

[5] Glover, F., "Tabu Search-Part Ⅱ," ***ORSA Journal on Computing***, vol. 2, pp. 4-32, 1990.

[6] Kennedy, J. and R. Eberhart, "Particle swarm optimization," ***IEEE International Conference on Neural Networks***, vol. 4, pp. 1942-1948, 1995.

[7] Kennedy, J. and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," ***IEEE International Conference on Systems, Man, and Cybernetics***, 1997, pp. 4104-4108.

[8] Oh, I.-S., J.-S. Lee, and B.-R. Moon, "Hybrid Genetic Algorithms for Feature Selection," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, vol. 26, pp. 1424-1437, 2004.

[9] Tahir, M. A., A. Bouridane, and F. Kurugollu, "Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier," ***Pattern Recognition Letters***, vol. 28, pp. 438-446, 2007.

[10] Tan, S., "An effective refinement strategy for KNN text classifier," ***Expert Systems with Applications***, vol. 30, pp. 290-298, 2006.

[11] Wang, X., J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," ***Pattern Recognition Letters***, vol. 28, pp. 459-471, 2007.

[12] Zhang, H. and G. Sun, "Feature selection using tabu search method," ***Pattern Recognition***, vol. 35, pp. 701-711, 2002.

表 2 基因微陣列基因選擇數量

Method / Dataset	Original # Genes	Random Forest # Genes (s.e. = 0)	Random Forest # Genes (s.e. = 1)	SC.s # Genes	NN.vs # Genes	BPSO-TS # Genes
Leukemia	3051	2	2	82	512	296
Breast 2 class	4869	14	14	31	88	971
Brain	5597	22	9	4177	1834	433
Lymphoma	4026	73	58	2796	15	198
Prostate	6033	18	2	4	7	706

Legends: (1) Original # Genes: original number of genes for dataset. (2) Random Forest # Genes: number of genes selected for Random Forest with s.e. = 0. (3) Random Forest # Genes: number of genes selected for Random Forest with s.e. = 1. (4) SC.s # Genes: number of genes selected for shrunken centroids with minimization of error and minimization of features if ties. (5) NN.vs # Genes: number of genes selected for nearest neighbor with variable selection. (6) BPSO-TS # Genes: number of genes selected for BPSO-TS.

表 3 基因微陣列分類錯誤率

Method / Dataset	Random Forest # Error (s.e. = 0)	Random Forest # Error (s.e. = 1)	SC.s # Error	NN.vs # Error	BPSO-TS # Error
Leukemia	0.087	0.075	0.062	0.056	**0.0**
Breast 2 class	0.337	0.332	0.326	0.337	**0.239**
Brain	0.216	0.216	0.159	0.194	**0.086**
Lymphoma	0.047	0.042	0.033	0.04	**0.0**
Prostate	**0.061**	0.064	0.089	0.081	0.067
Average	0.15	0.146	0.134	0.142	**0.078**

Legends: (1) Random Forest # Error: classification error rate of Random Forest with s.e. = 0. (2) Random Forest # Error: classification error rate of Random Forest with s.e. = 1. (3) SC.s # Error: classification error rate of shrunken centroids with minimization of error and minimization of features if ties. (4) NN.vs # Error: classification error rate of nearest neighbor with variable selection. (5) TS-CGA # Error: classification error rate of BPSO-TS.