# BPSO-SVM 應用單一核苷酸多型性預測乳癌危險性

# Risk Prediction for Breast Cancer Apply Single Nucleotide Polymorphism using BPSO-SVM

莊麗月 [a], 吳國銓 [b], 張學偉 [c], 楊正宏 [d, e*]

Chuang Li-Yeh [a], Kuo-Chuan Wu [b], Hsueh-Wei Chang [c], and Cheng- Hong Yang [d, e*]

[a] 義守大學化學工程系

[b] 國立高雄應用科技大學資訊工程系

[c] 高雄醫學大學生物醫學暨環境生物系

[d] 稻江科技暨管理學院網路系統學系

[e] 國立高雄應用科技大學電子工程系

[*] 通訊作者: 楊正宏, chyang@cc.kuas.edu.tw

**摘要**

本文利用資料探勘針對乳癌資料集進行特徵選取及分類問題。依據本文收集之乳癌及非乳癌樣本，我們嘗試以文獻中可能與乳癌相關之單核苷酸多型性，使用粒子族群最佳化作為特徵選取及參數最佳化，並以支持向量機為分類方法，進行乳癌危險性預測，並與其他分類方法做比較。結果顯示，本研究方法對於乳癌有 74%以上預測正確率，其正確率不但優於其他分類方法，且能有效挑選出重要性較高的單核苷酸多型性。

**關鍵字**：資料探勘、乳癌、粒子族群最佳化、支持向量機、單核苷酸多型性

## 1、前言

醫學研究指出，導致女性死亡的癌症疾病中，乳癌(breast cancer)僅次於支氣管癌(lung cancer)[1]。而乳癌的發生率在各國間逐漸增長，雖然乳癌的死亡率於部份地區有稍微減少的驅勢，但乳癌仍是各國女性常見癌症中最主要的健康問題[2-4]。部份研究顯示，乳癌的產生可能由基因元素(genetic element)的互相影響和各種可能的環境因素所引起，另外種族的特性也是扮演產生乳癌危險性的重要角色[1]。

人類基因體計畫(The human genome project, HGP)的進行，其目的是為完全解讀人類遺傳圖譜，破解人類基因密碼，最終得以解讀基因體核苷酸序列，並鑑別所有人類基因功能。目前，已步入後基因體世代，主要工作是利用序列探索該資訊涵義。其中，分析人類基因序列變異如研究單核苷酸多型性(Single Nucleotide Polymorphisms, SNPs) 對於人類疾病關聯性分析(Association with Linkage Disequilibrium)，再加上最近研究顯示 SNP 為人類相關性疾病研究之重要指標[5, 6]。因此，利用 SNP 資料來研究遺傳流行病學的複雜疾病成為近來相當熱門領域。透過生物資訊學分析 SNP 在正常個體與病患間的關聯性，這種人類多基因遺傳病變的研究不僅可找出影響骨質疏鬆症、糖尿病或乳癌等常見重大疾病的致病基因，更可針個人體質不同實施個人化之疾病預防與治療。

SNP 是由 E. Lander 於 1996 年所提出，被稱為「第三代 DNA 遺傳標記」，在同一物種而不同個體間基因體內某單一核苷酸不相同，其在群體間分佈的對偶基因比率至少大於 1%，換句話說在 50 個個體中，該差異位點至少要出現 1 次以上(因為一個人有雙套對偶基因)，如此才不會將偶發突變位點當成 SNP。SNP 為人體基因中單一個核苷酸改變所造成，故被視為人體基因多樣性主要原因之一。每個人基因組並不完全相同是造成基因差異表現的原因，除了一段基因在染色體上排列位置的改變、插入或刪除核苷酸序列的突變或變異外，不同個體間 SNP 亦是一重要影響因子。

本研究收集 550 個乳癌及非乳癌樣本，其中包含年齡及七個可能與乳癌有相關 SNP，共八個屬性。以現有樣本利用資料探勘之分類問題，達到預測效果。為分

析所有樣本各屬性與乳癌之關聯性或重要性，亦加入特徵選取方法挑選出較重要屬性。因此本文利用粒子族群最佳化(Particle Swarm Optimization, PSO)作為特徵選取方法，以支持向量機(Support Vector Machine, SVM)作為特徵選取後分類方法，並將支持向量機重要參數 $C$、$\gamma$ 之設定設計於粒子族群最佳化中，使粒子族群最佳化達到特徵選取及參數最佳化之目的。最後，我們利用 Weka (成熟應用於資料探勘軟體)[7]進行驗證及比較。以下將詳述本文的研究方法：粒子族群最佳化演算法及支持向量機。

## 2、研究方法

### 2.1、粒子族群最佳化

粒子族群最佳化[8]由 Eberhart 和 Kennedy 兩位學者提出，藉著觀察鳥群和魚群在自然界中覓食習性，而引發構想設計出一套最佳化演算法。然而為使粒子族群最佳化能解離散問題，由原本實數編碼改為二進制編碼，稱二進制粒子族群最佳化(Binary Particle Swarm Optimization, BPSO)[9]。在粒子族群最佳化中，每個粒子均視為一個解，這些粒子和基因演算法中染色體具有相同意義。粒子族群最佳化和基因演算法最大差異在於演化方式，基因演算法主要是基於母代，藉由選擇、交配、突變進行世代替換而產生較佳的子代。不同於基因演算法，每個粒子在各自搜尋解經驗中，個體最佳經驗稱之為 *pBest*，而在所有粒子中最佳經驗稱之為 *gBest*。根據這兩種經驗來決定飛行速度 $V$ 及移動方向，並決定所在位置 $X$。假設 $N_{dim}$ 為問題空間維度(即搜尋空間 $\Re^{N_{dim}}$)，有 $P$ 個粒子，每個粒子有其位置 $X_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$ 及速度 $V_i = \{v_{i1}, v_{i2}, \ldots, v_{id}\}$，其中 $i = 1, 2, \ldots, P$，$d = 1, 2, \ldots, N_{dim}$，$x_{id} \in \{0, 1\}$，$v_{id} \in [-v_{max}, v_{max}]$。其更新公式如下：

$$V_{id}^{new} = w \cdot V_{id}^{old} + c_1 \cdot rand_1(\,) \cdot (pBest - X_{id}^{old}) + c_2 \cdot rand_2(\,) \cdot (gBest - X_{id}^{old}) \quad (1)$$

$$S(V_{id}^{new}) = \frac{1}{1 + e^{-V_{id}^{new}}} \quad (2)$$

$$if\ (rand_{id} < S(V_{id}^{new})) \quad than \quad X_{id}^{new} = 1, \quad else \quad X_{id}^{new} = 0 \quad (3)$$

其中 $V_{id}^{new}$ 代表粒子更新後速度，$V_{id}^{old}$ 代表粒子更新前速度。$X_{id}^{new}$ 代表粒子更新後位置，$X_{id}^{old}$ 則代表粒子更新前位置。$(pBest - X_{id}^{old})$ 是粒子本身最佳位置和粒子當前所在位置之間距離，$(gBest - X_{id}^{old})$ 是截至目前迭代為止之粒子最佳位置和粒子當前所在位置之間距離。$w$ 為慣性權重[10]，一般介於 0.4 到 0.9 之間。$c_1$ 和 $c_2$ 為學習因數，其範圍為 0~4，一般均設定為 2。$rand_1(\,)$、$rand_2(\,)$、$rand_{id}$ 均為介於 0 至 1 之間均勻分佈的隨機變數。在更新公式裡，$S(\,)$ 為一個 Sigmoid 函數，粒子移動速度 $V$ 為粒子位置 $X$ 改變為 1 或 0 的機率，因此經由 Sigmoid 函數之轉換後，使 $V$ 介於 0 到 1 之間的機率數值，若使 $v_{max} = 6$ 可使 $v_i$ 介於 0.9975 至 0.0025 之間，再藉由隨機產生之亂數 $rand_{id}$ 來判斷粒子位置是否改變，如上述式子(3)所示。

### 2.2、支持向量機

支持向量機由 Vapnik 於 1995 年所提出，是一種基於統計學習理論結構風險最小化的機器學習方法。其原先概念是針對二類別分類問題，在資料空間裡找出對分類較好區分的支持向量，以利在高維度特徵空間中建構一個超平面，此超平面可將類別之間間隔最大化，得以正確地將二類別的資料進行區分，產生出較好之分類正確率[11]。

假設 $m$ 筆訓練資料$(x_i, y_i)$，其中 $i = 1, 2, \ldots, m$，$x_i \in \Re^n$，$y_i \in \{+1, -1\}$。若有一超平面 $w \cdot x_i + b$ 能將訓練資料區分開來，則落於此超平面上所有 $x$ 必須滿足 $w \cdot x_i + b = 0$，並與超平面之間具有最短距離，相對類別間擁有最大的距離。而本文利用可快速將類別之間的間隔最大化演算法，Kernel-Adatron 演算法(Kernel-Adatron algorithm, KA) [12]。此方法是由 Adatron 演算法與核函數(Kernel function)組合而成，在支持向量機訓練過程中，藉由核函數將原始資料投射至高維度特徵空間中，以不斷調整類別間之間隔，保證最後能收斂至最佳化。其中本文採用之核函數為輻狀基底函數(Radial basis function, RBF)。

### 2.3、粒子族群於特徵選取及參數最佳化

本研究主要提出以粒子族群最佳化找出最佳之支持向量機參數 $C$、$\gamma$ 及特徵集合，利用支持向量機所計算出預測正確率作為適應函數值。以下詳細介紹本研究的方法：方法之流程及架構、資料正規化、粒子編碼、

族群初始化和適應函數。

### 2.3.1、流程及架構

以粒子族群最佳化找出最佳特徵子集合及支持向量機參數 $C$、γ，其流程簡述如下：

***第一部份 - 資料處理***

1) 資料正規化：取得資料後利用公式(4)將資料轉換成[0, 1]之間的值。
2) K-Fold：隨機將資料平均分配至 K 個部份，其中一個部份為測試資料，其餘部份則為訓練資料。

***第二部份 - 演算法***

3) 族群初始化：隨機產生粒子的位置及速度。
4) 編碼：將粒子位元進行解碼，粒子前四十個位元中，分別表示 $C$，γ 各二十個位元，其解碼方式如公式(5)所示；而其餘位元為特徵數，其解碼方式即當位元為 0 時表示此特徵未選取，反之則表示此特徵被選取。
5) 計算適應函數：以支持向量機之正確率作為適應函數值。
   5.1) Kernel-Adatron 演算法：利用訓練資料，配合 Kernel-Adatron 演算法找出兩類別間最大的間隔。
   5.2) 預測正確率：利用測試資料進行測試，當所有測試資料分別得到類別時，判斷是否分類正確，最後將分類正確的資料除以所有測試資料數量。
6) 更新 *pBest* 及 *gBest*：粒子目前位置之適應值，與本身及群體最佳值比較，若當前解比本身最佳解好，則當前解為 *pBest*。若此 *pBest* 為所有群體最佳解，則當前解為 *gBest*。
7) 更新粒子目前位置：利用公式(1)、(2)、(3)更新粒子之位置及速度。
8) 停止條件：當迭代次數達設定次數，或當適應值達100%則停止；否則跳到步驟 3，直到符合停止條件為止。
9) 最佳參數及特徵：經過粒子族群最佳化後，會得到一組最佳化解，其中包含 $C$、γ 及特徵子集合。

### 2.3.2、資料正規化

將原始資料內容藉由公式(4)轉換成[0, 1]之間的值，其主要目的在於可避免資料屬性的數值範圍過大，將資料控制在一定範圍內，另一目的是避免分類器在計算上之困難。一般在正規化後，支持向量機的辨識率會提升[13]。

$$v' = \frac{v - \min_a}{\max_a - \min_a} \tag{4}$$

其中，$v$：原始資料、$v'$：正規化後資料、$\max_a$：原始資料之最大值、$\min_a$：原始資料之最小值。

### 2.3.3、粒子編碼

由於考慮支持向量機在輻狀基底函數核心函數投射至高維度空間有兩個主要參數，$C$ 及 γ[14]，因此我們在編碼上除了資料特徵以外，另外還加入了這兩個參數，如 Figure 1 所示。

| | | |
|---|---|---|
| $X_c^1 \ldots X_c^i \ldots X_c^{n_c}$ | $X_r^1 \ldots X_r^i \ldots X_r^{n_r}$ | $X_f^1 \ldots X_f^i \ldots X_f^{n_f}$ |

Figure 1 粒子編碼示意圖

$X_c^1 \sim X_c^{n_c}$ 為參數 $C$ 之編碼；$X_r^1 \sim X_r^{n_r}$ 為參數 γ 之編碼；$X_f^1 \sim X_f^{n_f}$ 則為特徵之編碼。其中 $n_c$、$n_r$ 是參數編碼位元的長度，在本文各設為 20 位元。而 $n_f$ 則是特徵數的位元長度，依據資料的形態而有所變動。另外在解碼的方面，在參數部份以式子(5)做解碼的動作。而特徵部份位元為 1 則代表此特徵被選取，為 0 則不被選取(亦即刪除特徵)，以上編碼方式參考[13]。

$$p = \min_p + \frac{\max_p - \min_p}{2^l - 1} \times d \tag{5}$$

其中 $p$：參數編碼後型態、$min_p$：參數設定最小值、$max_p$：參數設定最大值、$d$：二進制位元轉十進制、$l$：位元字串的長度。

### 2.3.4、族群初始化

依設定的族群數 $P$ 及編碼長度 $l$，利用隨機方式產生 $P$ 個粒子(即 $P$ 個解)，產生之位元字串由{0, 1}所組成，每個粒子之初始速度 $V$ 為隨機產生[0, 1]間的數值。

### 2.3.4、適應函數

利用每個粒子位置，取出資料的特徵子集合及支持向量機之參數，以訓練資料訓練出支持向量機模型，利用測試資料進行預測，最後預測正確率作為每個粒子的適應函數值，其式子如下所示。

$$fitness(x_{id}) = Accuracy_{SVM} \quad (6)$$

## 3、結果與討論

### 3.1、實驗描述

本研究實驗資料經高雄醫學大學人體試驗委員會認可使用[15]。內含 220 個病理學上證實為罹患乳癌的女性病患(平均年齡為 53.3±11.7 歲)，334 個來自身體例行檢查或一般小手術的非乳癌女性(平均年齡為 44.3±13.0 歲)，共 554 個樣本。樣本屬性分別為年齡及 7 個 SNP (如 Table 1 所示)共 8 個特徵。而 SNP 基因型(genotype)為字母型態，Table 2 為資料型態轉換表。

本研究以個人電腦：Intel Core 2 Quad 2.40 GHz CPU、2048 MB RAM、Windows XP 作業系統，並採用 JAVA JDK 6.0 開發環境做為實驗平台。方法則以支持向量機進行乳癌危險性預測，並利用粒子族群最佳化進行特徵選取及支持向量機之參數最佳化，透過 K-Fold 交叉驗證法(K-Fold cross validation)[16]進行正確率評估。K-fold 交叉驗證法將資料以隨機方式平均分成 K 個部份，將 K 個部份中每一部份獨立做為測試集合，而其餘 K-1 個部份做為訓練集合，當訓練集合訓練出支持向量機的雛型時，再利用測試集合進行評估，如此交叉驗證後即可獲得正確率之評估標準。一般 K 值依資料樣本數為依據，若資料樣本數過小，K 設定愈大愈好，如此可使訓練樣本數變多[17]，K 值設定為 10。

### 3.2 正確率評估

本研究採用醫學診斷最常使用的評估方式，分別為：陽性猜中率(Positive hit rate)，即敏感度(Sensitivity)、陰性猜中率(Negative hit rate)，即特異度(Specificity)及正確率(Accuracy rate)。若正確預測出有病稱真陽性(True Positive, *TP*)，然而，當預測沒病但實際上有病則稱偽陰性(False Negative, *FN*)。相對地，若正確預測出沒病稱真陰性(True Negative, *TN*)，當預測有病但實際上沒病則稱假陽性(False Positive, *FP*)。在資料探勘領域裡，一般正確率算法為：

$$Accuracy\ rate = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

其可評估分類器的正確率[7]。對於分類器而言以 *TP*、*FP* 較為重要[18]，而敏感度與特異度則是區分分類器對於有病或沒病的效果。敏感度是正確預測有病的比例，其公式為 $P(T^+|D^+) = TP/(TP+FN)$。特異度是正確預測沒病的比例，其公式為 $P(T^-|D^-) = TN/(TN+FP)$。

### 3.3、結果

本研究利用 Weka 中現有五種分類器，包含最近鄰居法(Nearest-Neighbor, NN)、C4.5 演算法、隨機森林(Random forest, RF)、貝氏分類器(Naïve Bayes, NB)及序列最小優化演算法(Sequential Minimal Optimization, SMO)，以 K-Fold 交叉驗證法驗證該分類器對於本文資料集的正確率，並與本研究方法比較。Table 3 為本研究結果，平均預測正確率為 74.73% ± 4.24，表示本研究利用年齡及七個 SNP，即可獲得 74%左右預測女性乳癌發生的危險性之準確率。由 Table 4 可知，除本研究方法以外，其餘方法均未超過 70%。其中，SMO 及 KA 之參數 C 和 γ 分別手動設定為 $2^4$ 和 $2^{-4}$，然而經由粒子族群最佳化之特徵選取及參數最佳化後，可將正確率提升 10%以上。

Table 2 資料型態轉換表

| Value | Genotype |
|---|---|
| 1 | AA、CC |
| 0 | AT、AG、CT |
| -1 | TT、GG |

Table 1 SNP 資料型態

| SNP No. | Chr. | Gene (location) | SNP rs# | Genotype | | |
|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 |
| 1 | 12 | CD4 (intron 3) | rs12812942 | AA | AT | TT |
| 2 | 17 | CCR7 (intron 1) | rs3136685 | AA | AG | GG |
| 3 | 2 | CXCR4 (I124I) | rs2228014 | CC | CT | TT |
| 4 | 10 | CXCL12 (3'UTR) | rs1801157 | AA | AG | GG |
| 5 | 6 | VEGF (C936T) | rs3025039 | CC | CT | TT |
| 6 | 16 | MMP2 (T460T) | rs2287074 | AA | AG | GG |
| 7 | 12 | KITLG (intro 1) | rs10506957 | CC | CT | TT |

Table 3 BPSO 於特徵選取及參數最佳化之實驗結果

| Fold | Samples | BPSO algorithm for feature selection and optimization parameter | | | | | |
|---|---|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Accuracy | Optimized *C* | Optimized γ | features |
| 1 | 55 | 0.714286 | 0.794118 | 0.763636 | 2201.03 | 8.372551 | 6 |
| 2 | 55 | 0.733333 | 0.700000 | 0.709091 | 4006.994 | 14.981 | 6 |
| 3 | 55 | 0.700000 | 0.771429 | 0.745455 | 302.5831 | 15.79585 | 5 |
| 4 | 55 | 0.750000 | 0.627907 | 0.654545 | 4077.81 | 2.328212 | 3 |
| 5 | 55 | 0.916667 | 0.720930 | 0.763636 | 1555.21 | 2.490861 | 4 |
| 6 | 55 | 0.714286 | 0.823529 | 0.781818 | 2601.646 | 10.84777 | 5 |
| 7 | 55 | 0.857143 | 0.780488 | 0.800000 | 1756.123 | 13.52724 | 3 |
| 8 | 55 | 0.812500 | 0.692308 | 0.727273 | 119.8872 | 0.76004 | 6 |
| 9 | 55 | 0.875000 | 0.765957 | 0.781818 | 594.9403 | 9.748901 | 2 |
| 10 | 59 | 0.727273 | 0.750000 | 0.745763 | 3262.34 | 8.212954 | 2 |
| Average ± SD | | 0.766667 ± 0.078602 | 0.740099 ±0.057772 | 0.747292 ± 0.042380 | | | 4.2 ± 1.62 |

Table 4 各分類器應用於乳癌資料集之預測結果

**Legends**: (1) NN: Nearest-Neighbor; (2) RF: Random forest; (3) NB: Naïve Bayes; (4) KA: Kernel-Adatron algorithm; (5) SMO: Sequential Minimal Optimization; (6) BPSO-KA: our propose approach.

| 分類器 | 敏感度 | 特異度 | 正確率 |
|---|---|---|---|
| NN | 0.455399 | 0.639296 | 0.568592 |
| C4.5 | 0.451613 | 0.630435 | 0.570397 |
| RF | 0.482353 | 0.640625 | 0.592058 |
| NB | 0.531646 | 0.656566 | 0.620939 |
| SMO | 0.585586 | 0.650113 | 0.637184 |
| KA | 0.618182 | 0.627255 | 0.626354 |
| BPSO-KA | 0.766667 | 0.740099 | 0.747292 |

**3.4、討論**

在資料探勘領域裡，我們利用 Weka 中五種分類器及本研究使用的 KA 演算法(包含兩種支持向量機，SMO 及 KA)進行比較。其結果顯示，兩種支持向量機的演算法能獲得較佳的預測正確率。然而在支持向量機中，最重要兩個參數 *C* 和 γ，會影響支持向量機預測的結果，文獻[14]建議 $C = (2^{-2}, 2^{-1}, \ldots, 2^{12})$、$\gamma = (2^{-10}, 2^{-9}, \ldots, 2^{4})$，因此本研究利用粒子族群最佳化演算法搜尋 *C* 和 γ 的最佳值，其搜尋範圍分別為 $C = [2^{-2}, 2^{12}]$及 $\gamma = [2^{-10}, 2^{4}]$。另一方面，在本資料集雖然 SMO 演算法正確率比 KA 演算法好，但 KA 演算法除了容易實現外，其快速、健全且正確率與 SMO 相差不大，本研究考量 BPSO 在不斷搜尋解的過程中，其耗費時間過長，因此本文使用 KA 演算法，做為支持向量機最大間隔的訓練方法。

粒子族群最佳化的優點除了演化方式簡單且易實現外，其搜尋範圍廣且快速收斂。其快速收斂的特性可彌補高運算量的支持向量機，而搜尋範圍廣能使特徵選取及參數最佳化獲得更有效的搜尋。此外，粒子族群最佳化演算法有部份重要的參數，包括族群數 *P*、慣性權重 *w*、學習因子 $c_1$、$c_2$ 以及迭代次數，其中族群數若設定太大會造成運算時間過於冗長，反之則無法在解空間找到最佳解，文獻[8]已證明 20~40 可獲得較好的結果，而本研究中族群數 *P* 設定為 40；*w*、$c_1$、

$c_2$ 的參數則是影響粒子族群最佳化的收斂效果，各設定為 $w = 0.8$，$c_1 = c_2 = 2$，若設定過大會造成粒子移動的速度過快，導致無法找到最佳解；反之若設定過小，則使粒子移動過慢，使得若要找到最佳解則需花費冗長的運算時間[10]。最後迭代次數設為 30，主要本研究利用支持向量機，高運算量分類器做為正確率的評估標準，且在 30 次迭代裡已可獲得不錯的效果，因此利用粒子族群最佳化做為研究方法。

在資料集裡利用既有的樣本，年齡及七個與乳癌可能相關的 SNP，以特徵選取方法選出這些重要特徵。這 8 個特徵，對於特徵選取問題而言，只有 $2^8 = 256$ 種組合可輕易解決。然而在不同的特徵組合及不同 $C$、$\gamma$ 之設定，亦會獲得不同的結果，因此本文參考文獻[13]染色體的設計，將 $C$、$\gamma$ 設定為搜尋解的空間，以最佳化演算法取代人工對於參數的設定。在本文所獲得之預測正確率，乃利用有限樣本以及與乳癌有關聯性的 SNP，進行機器學習訓練與測試所獲得。其結果可供生物學家參考，倘若配合臨床實驗證明、乳癌資訊相關的搜集及更多可用的樣本，可使本研究方法更強健、穩固且更可靠。

## 4、結論

本研究以七個 SNP 做為實驗，目的在於驗證所提出之方法能有效獲得較佳預測能力及 SNP 挑選。結果顯示，本研究方法有 74%以上預測正確率，且優於其他方法，我們希望此項成果可以供往後醫學預測乳癌，生物學家對於乳癌危險性預測或 SNP 挑選的使用。針對未來研究方向，將持續與生物學家合作，利用有效的機器學習方法進行其他疾病的預測或取得更多 SNP 資料，挑選出對人類疾病具有幫助及意義的資訊。

## 參考文獻


[1]Hsiao, W.-C., Young, K.-C., Lin, S.-L., Lin P.-W., "Estrogen receptor-alpha polymorphism in a Taiwanese clinical breast cancer population: a case-control study," Breast Cancer Research, vol. 6, pp. R180 - R186, 2004.

[2]Hortobagyi, G. N., "Treatment of Breast Cancer," The New England Journal of Medicine, vol. 339, pp. 974-984, 1998.

[3]Sasco, A. J., "Epidemiology of breast cancer: an environmental disease?," APMIS, vol. 109, pp. 321-32, 2001.

[4]Zhang, L., Zhang, Z., Yan, W., "Single nucleotide polymorphisms for DNA repair genes in breast cancer patients," Clinica Chimica Acta, vol. 359, pp. 150-155, 2005.

[5]Gray, I. C., Campbell, D. A., Spurr, N. K., "Single nucleotide polymorphisms as tools in human genetics," Human Molecular Genetics, vol. 9, pp. 2403-2408, 2000.

[6]Shastry, B. S., "SNP alleles in human disease and evolution," Journal of Human Genetics, vol. 47, pp. 561-566, 2002.

[7]Witten, I. H., Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2 ed. San Francisco: Morgan Kaufmann, 2005.

[8]Kennedy, J., Eberhart, R., Shi, Y., Swarm intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001.

[9]Kennedy, J., Eberhart, R., "A discrete binary version of the particle swarm algorithm," in 1997 IEEE International Conference on Systems, Man, and Cybernetics. vol. 5 Orlando, FL, USA, 1997, pp. 4104-4108

[10]Shi, Y., Eberhart, R., "A modified particle swarm optimizer," in Proceedings of the IEEE International Conference on Evolutionary Computation, Anchorage, AK, USA, 1998, pp. 69-73.

[11]Cortes, C., Vapnik, V., "Support-vector networks," Machine Learning, vol. 20, pp. 273-297, 1995.

[12]Frieß, T.-T., Cristianini, N., Campbell, C., "The Kernel-Adatron algorithm: a fast and simple learning procedure for Support Vector machines," in Proc. 15th International Conference on Machine Learning: Morgan Kaufmann, pp. 188-196 1998.

[13]Huang, C.-L., Wang, C.-J., "A GA-based feature

selection and parameters optimization for support vector machines," Expert Systems with Applications, vol. 31, pp. 231-240, 2006.

[14]Hsu, C.-W., Lin, C.-J., "A comparison of methods for multiclass support vector machines," IEEE Transactions on Neural Networks, vol. 13, pp. 415-425, 2002.

[15]Lin, G.-T., Tseng, H.-F., Yang, C.-H., Hou, M.-F., Chuang, L.-Y., Tai, H.-T., Tai, M.-H., Cheng, Y.-H., Wen, C.-H., Liu, C.-S., Huang, C.-J., Wang, C.-L., Chang, H.-W., "Combinational polymorphisms of seven cxcl12-related genes are protective against breast cancer in Taiwan," OMICS: A Journal of Integrative Biology, vol. 12, pp. 1-8, 2009.

[16]Stone, M., "Cross-validatory choice and assessment of statistical predictions," Journal of the Royal Statistical Society, vol. 36, pp. 111-147, 1974.

[17]Salzberg, S. L., "On comparing classifiers: Pitfalls to avoid and a recommended approach," Data Mining and Knowledge Discovery, vol. 1, pp. 317-327, 1997.

[18]Woods, K., Bowyer, K. W., "Generating ROC curves for artificial neural networks," IEEE Transactions on Medical Imaging, vol. 16, pp. 329-337, 1997.